Math for ML

Week 4.2: Basics of Convex Optimization

By: Samuel Deng

Logistics & Announcements

Lesson Overview

Convexity. A property of *sets* and *functions* that affords us a lot of nice "linearity-like" properties.

Convex set. A convex is a set that has no holes. The line segment between any two points lies completely in the set.

Convex function. A function that is bowl-shaped. Between any two points, the line segment is above the function.

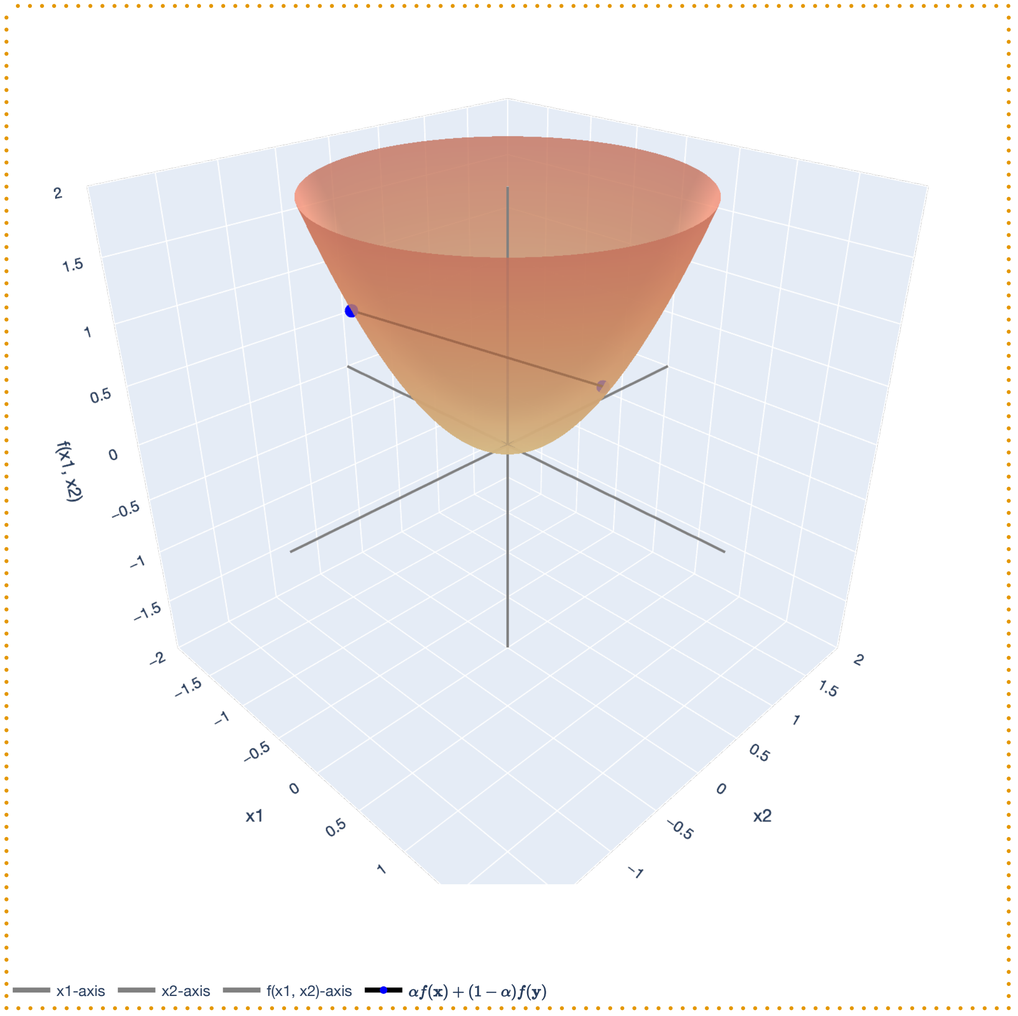
Convex optimization. Optimization problems with convex objectives and convex constraint sets. All local optima are global optima.

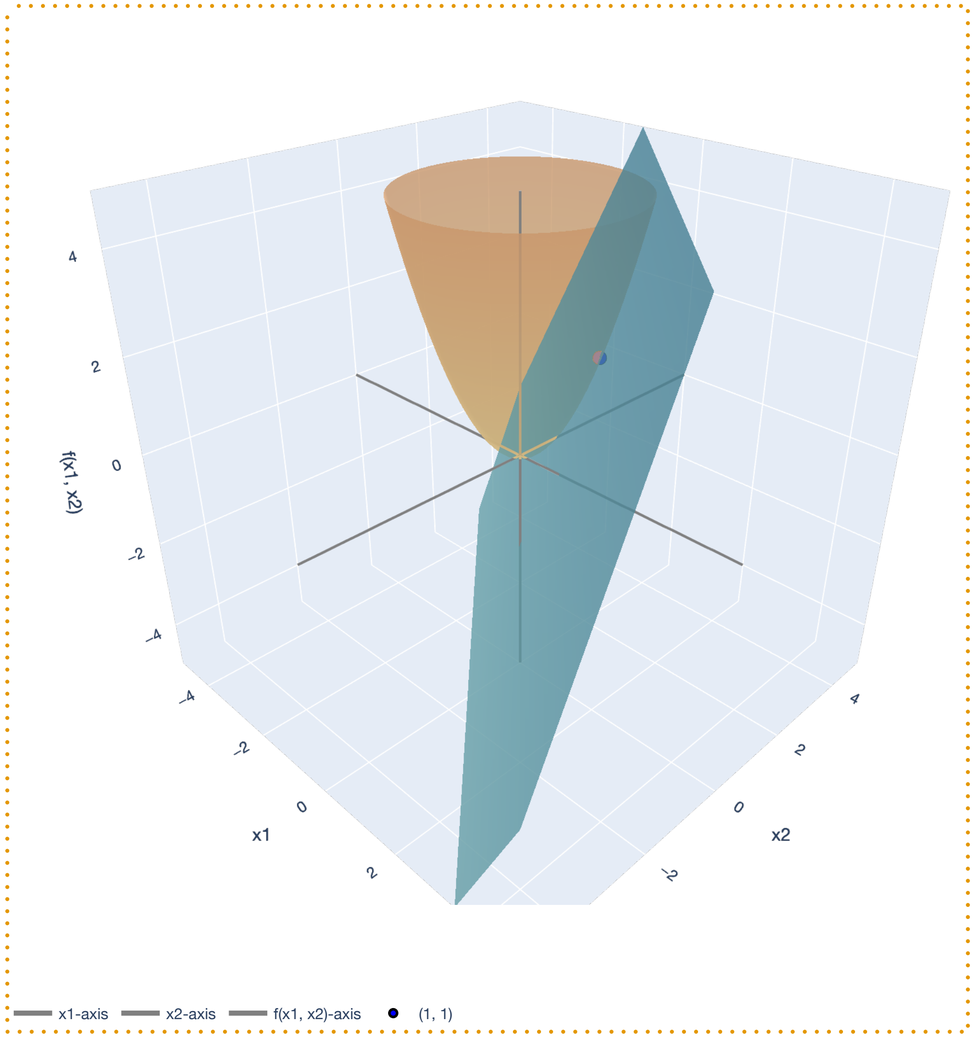
Gradient descent for convex problems. GD on convex functions is *guaranteed* to find a global min.

Gradient descent for OLS. We unite the two stories of this class and analyze GD applied to OLS!

Lesson Overview

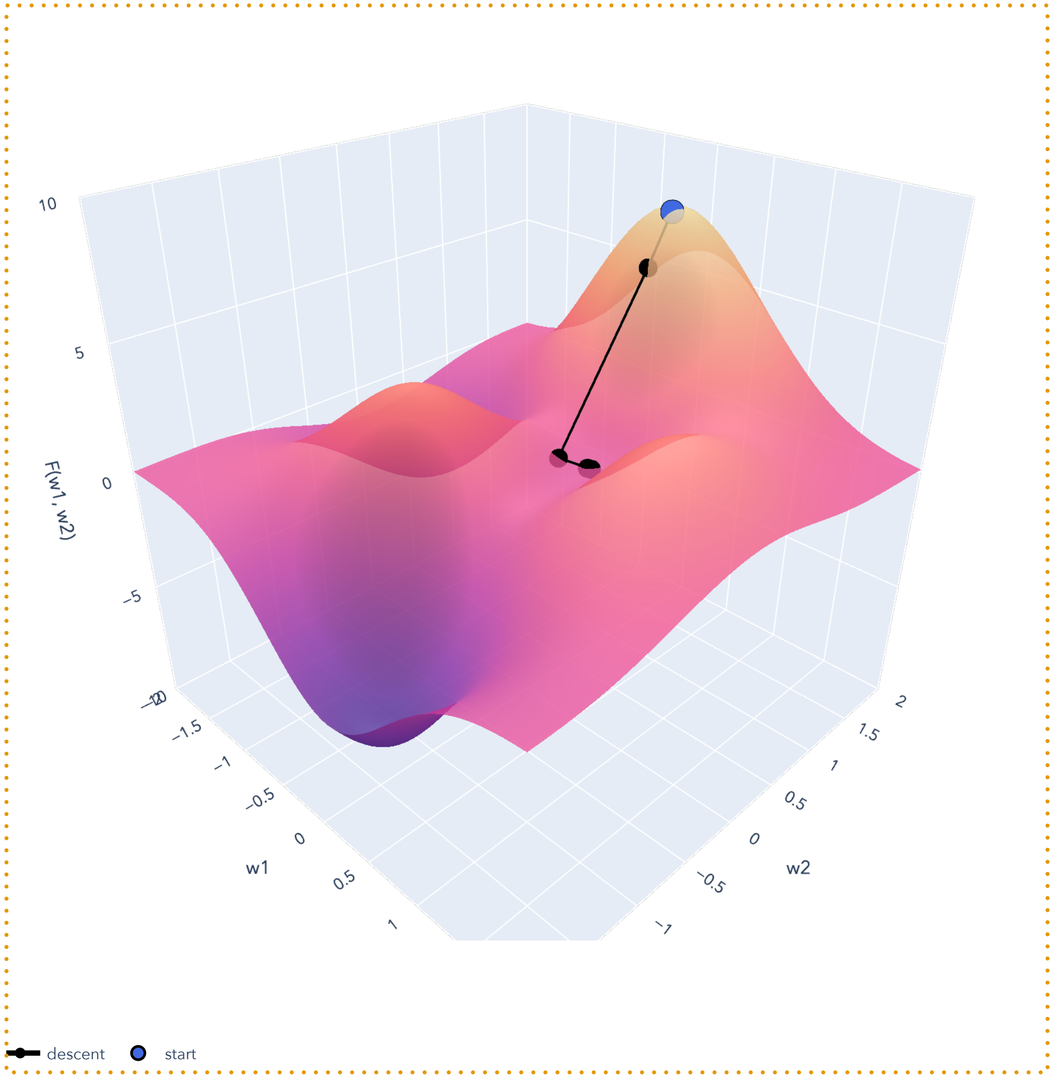
Big Picture: Least Squares

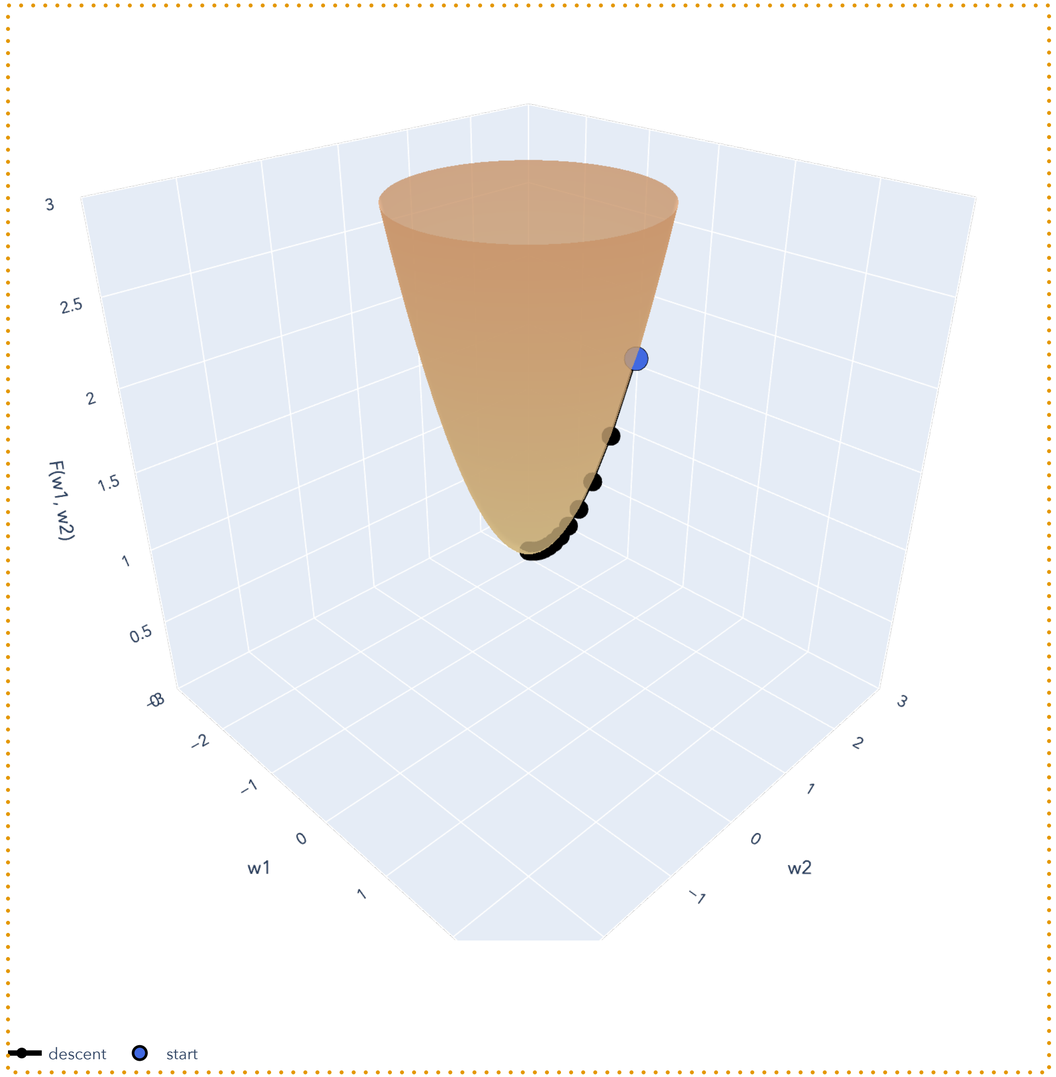




Lesson Overview

Big Picture: Gradient Descent





Convex Optimization

Motivation

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

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

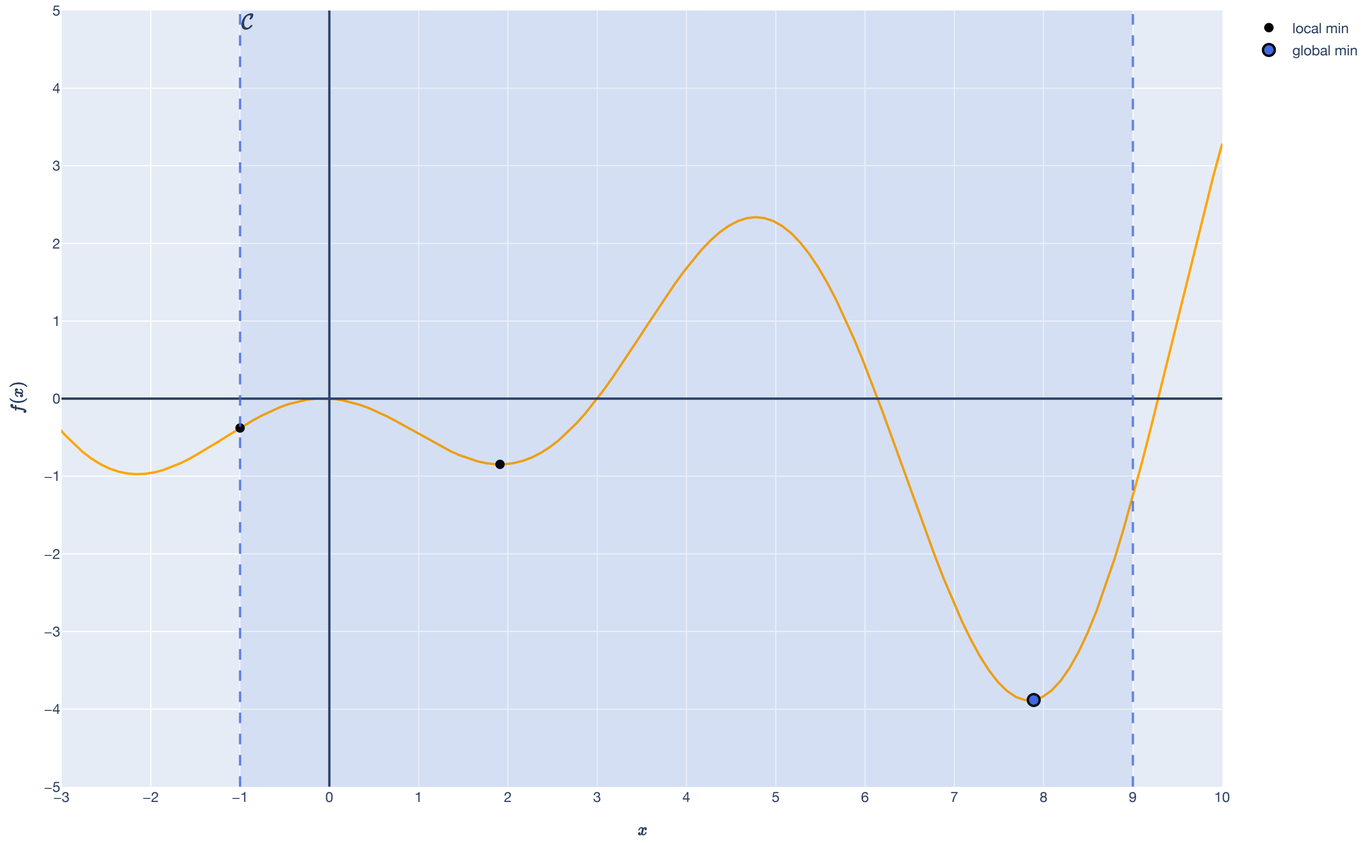
The optimal value is $f(\mathbf{x}^*)$. Our goal is to find $\mathbf{x}^*$ and $f(\mathbf{x}^*)$.

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

Global Minima

Local vs. global minima

Last lesson, we only developed methods for finding *local optima*.



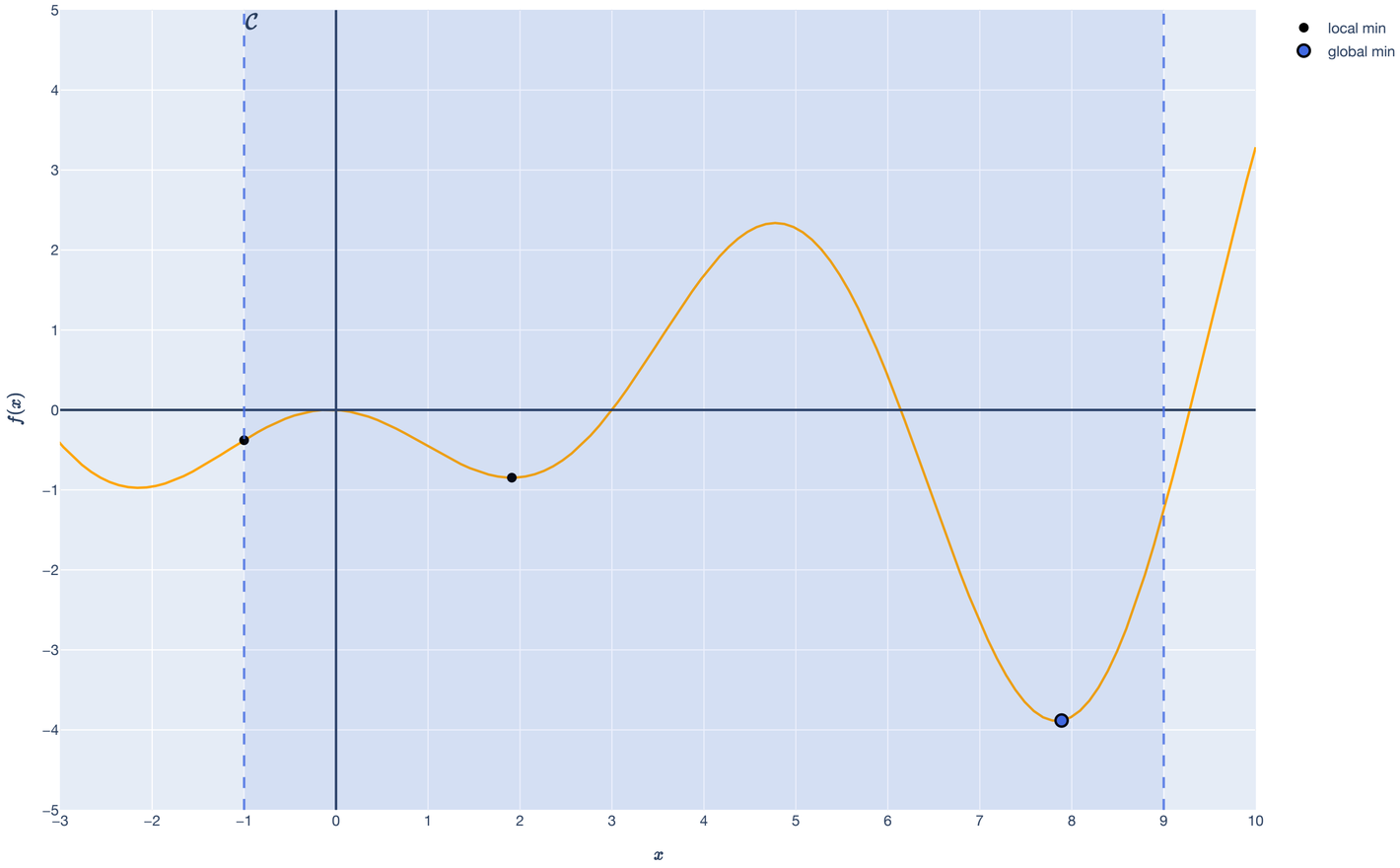
Types of Minima

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!*

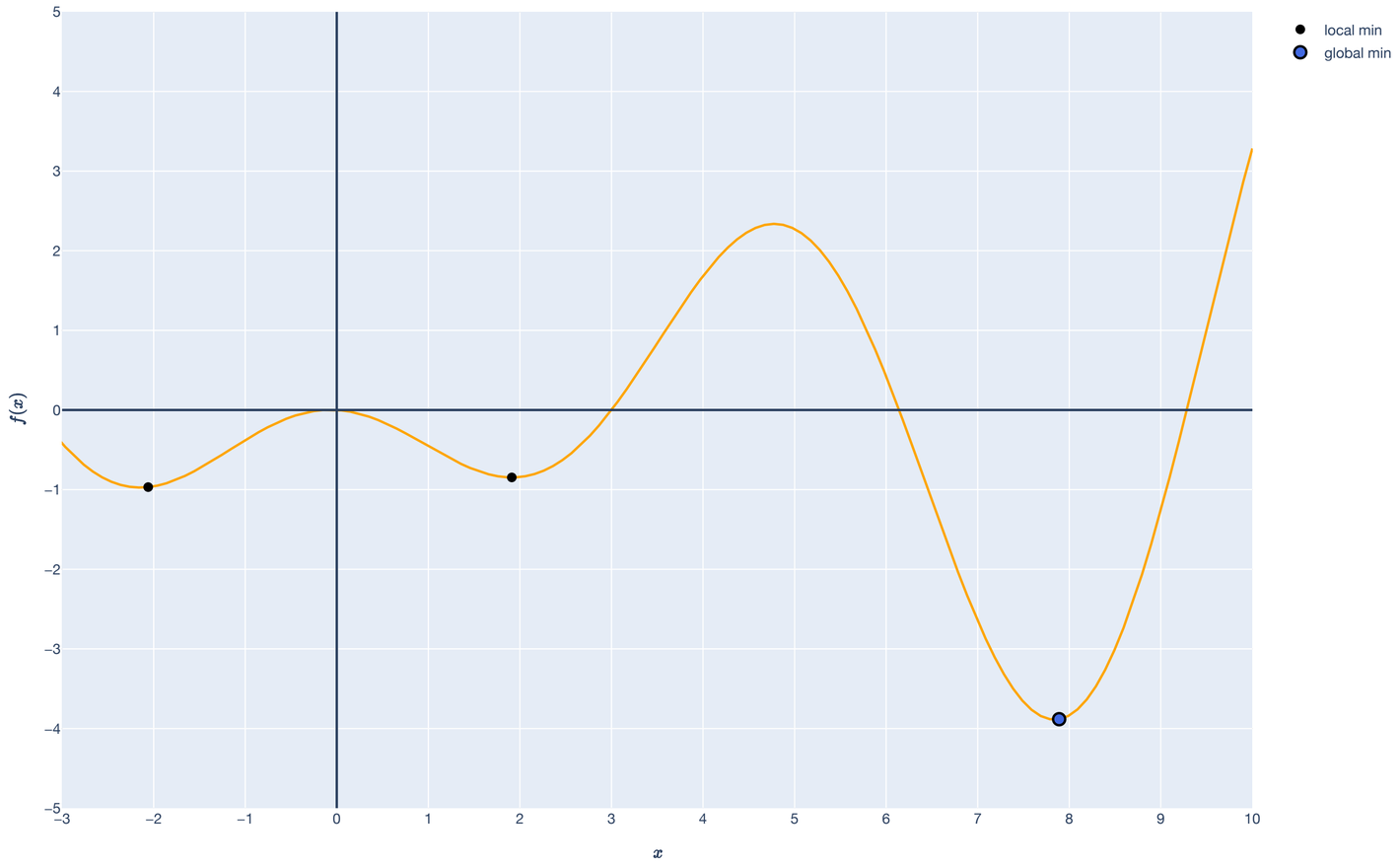


Convexity

Non-example ($d = 1$)

Functions that have many “hills/ valleys” are deceptive.

Local minima look like global minima when we’re sufficiently close.

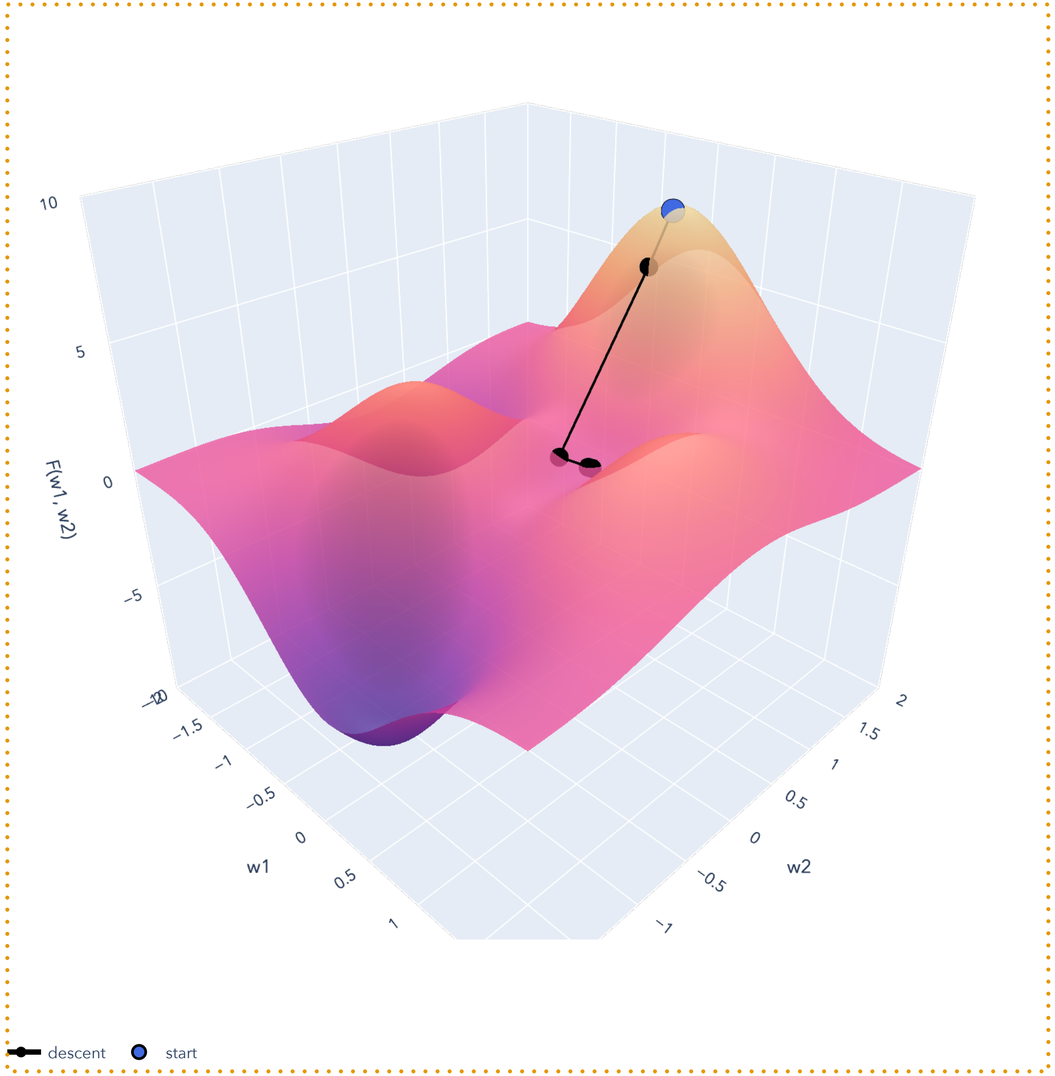


Convexity

Non-example ($d = 2$)

Functions that have many “hills/valleys” are deceptive.

Local minima look like global minima when we’re sufficiently close.

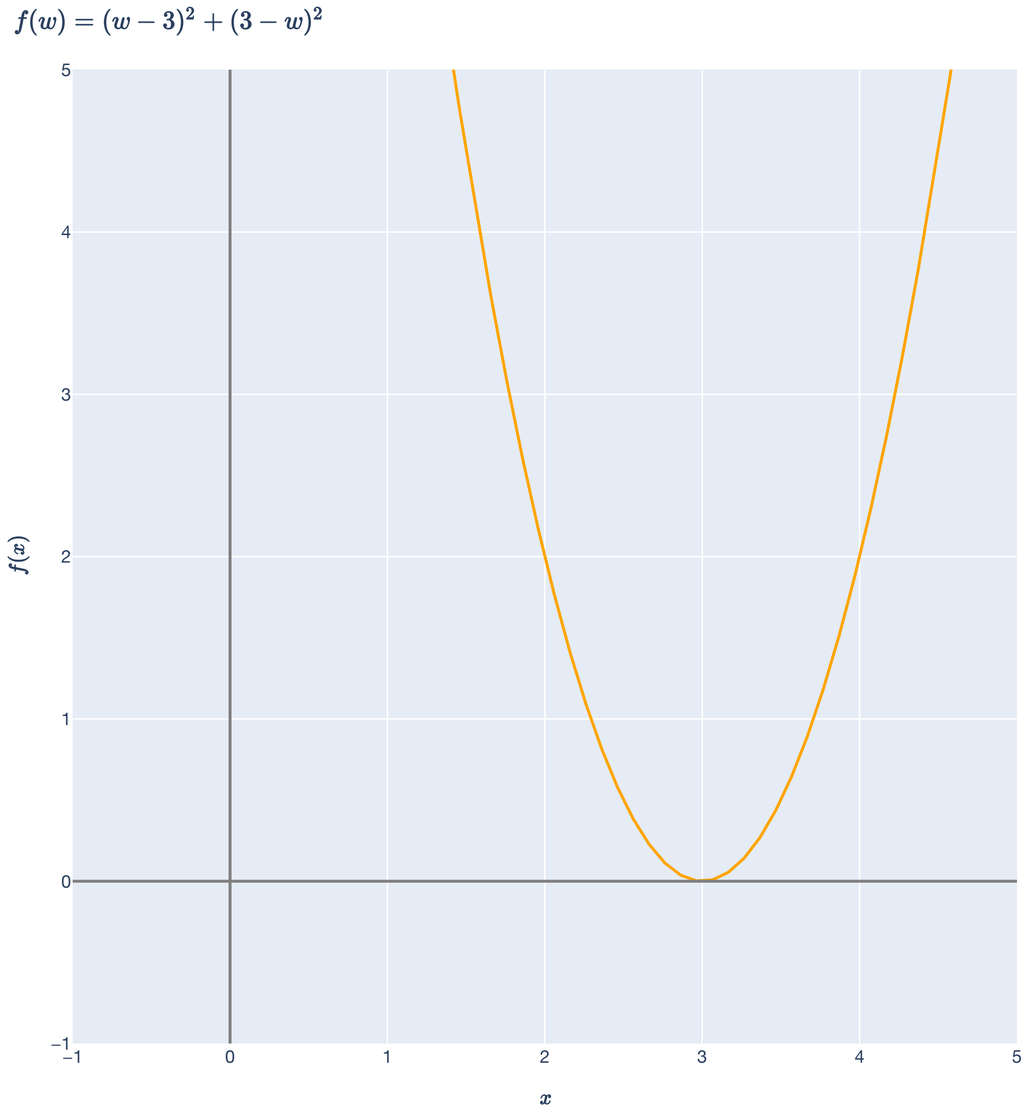


Convexity

Example ($d = 1$)

A *convex function* is a function that is "bowl-shaped."

Their local minima *are* global minima.

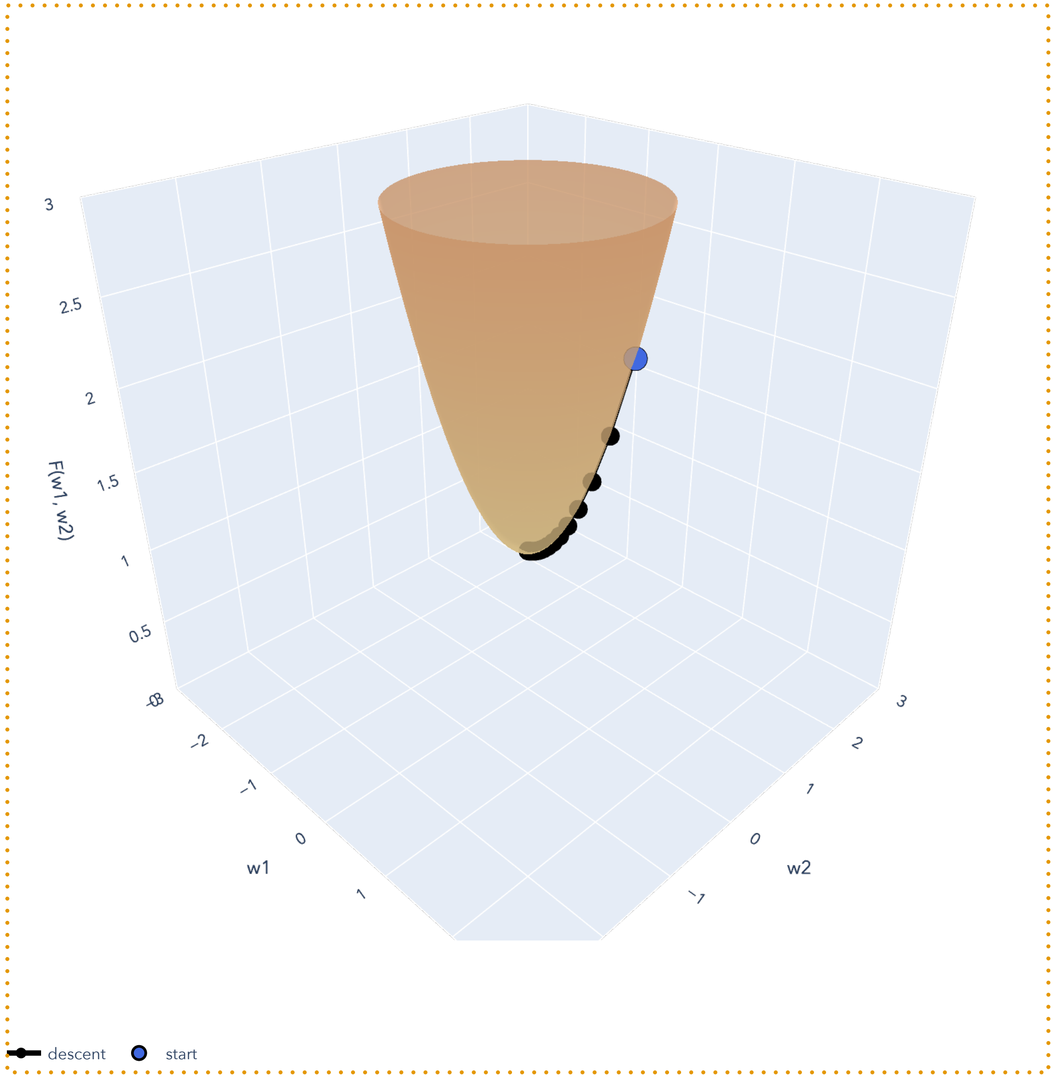


Convexity

Example ($d = 2$)

A *convex function* is a function that is "bowl-shaped."

Their local minima *are* global minima.



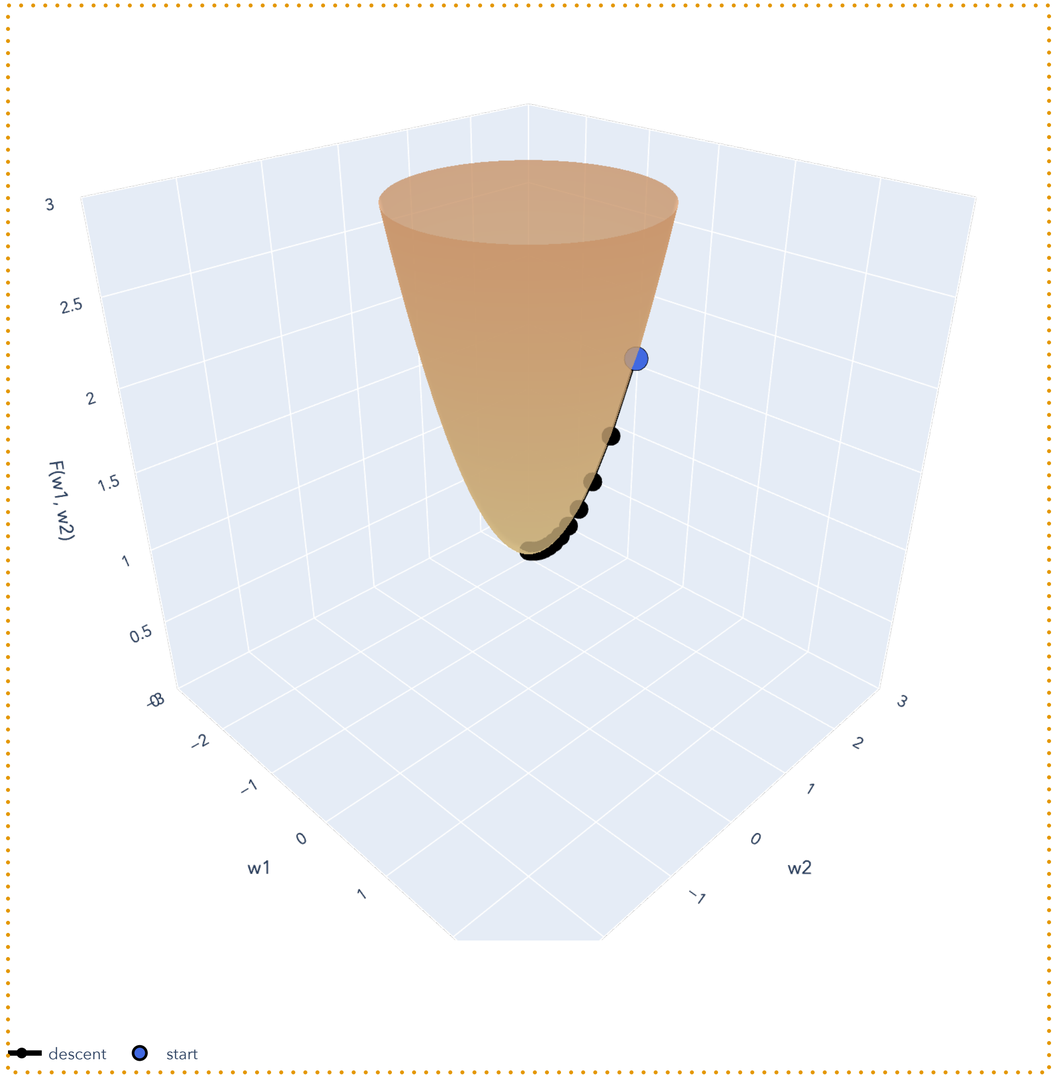
Convexity

Example ($d = 2$)

A *convex function* is a function that is "bowl-shaped."

Their local minima *are* global minima.

Goal: We will use gradient descent to solve convex optimization problems!



Convex Optimization Problem

Definition

A convex optimization problem (also known as *convex program*) is an optimization problem:

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

where $f(\mathbf{x})$ is a convex function and $\mathscr{C}$ is a convex set.

$f(\mathbf{x})$ is "bowl-shaped" and $\mathscr{C}$ has "no holes" or "gaps"

Convexity

Line segments

Line segments are very important to the study of convexity.

For any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the line segment between $\mathbf{x}$ and $\mathbf{y}$ is the set of points:

$$[\mathbf{x}, \mathbf{y}] := \{(1-\alpha)\mathbf{x} + \alpha\mathbf{y} : \alpha \in [0,1]\}$$

Sometimes, we'll denote the line segment as $[\mathbf{x}, \mathbf{y}]$.

Convexity

Line segments

Example. Line segment between $x = 1$ and $y = 3$.

Convexity

Line segments

Example. Line segment between $\mathbf{x} = (1,1)$ and $\mathbf{y} = (2,3)$.

Convex Sets

Intuition, Definition, and “Algebra”

Convex Sets

Idea

A *convex set* is a “set with no holes or gaps.”

We can draw a line between any two points and stay inside the set.

Convex Sets

Definition

A set $S \subseteq \mathbb{R}^d$ is a convex set if, for any $\mathbf{x}, \mathbf{y} \in S$, the point $(1-\alpha)\mathbf{x} + \alpha\mathbf{y} \in S$ for $\alpha \in [0,1]$.

That is, the line segment between any two points is completely in S.

Examples of Convex Sets

$\mathbb{R}^d$

Why is $\mathbb{R}^d$ a convex set?

Examples of Convex Sets

Line

Perhaps the most basic nontrivial example of a convex set is a *line*.

For any two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the line passing through $\mathbf{x}$ and $\mathbf{y}$ is the set of all points

$$(1 - \alpha)\mathbf{x} + \alpha\mathbf{y},$$

for *any* $\alpha \in \mathbb{R}$.

Examples of Convex Sets

Hyperplane

A hyperplane is the set of points

$$\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} = b\},$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are fixed, and $\mathbf{w} \neq \mathbf{0}$.

Why is this convex?

Examples of Convex Sets

Halfspace

A halfspace is the set of points

$$\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} \leq b\},$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are fixed, and $\mathbf{w} \neq \mathbf{0}$.

Why is this convex?

Examples of Convex Sets

Neighborhoods

The neighborhood centered at $\mathbf{c} \in \mathbb{R}^d$ with radius $\delta > 0$ is the set:

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

Why is this convex?

Closure of Convex Sets

The "Algebra" of Convex Sets

We can combine convex sets by using operations that preserve convexity:

Intersection. The *intersection* of (possibly infinite) convex sets is convex.

Scalar multiplication. If $C \subseteq \mathbb{R}^d$ is a convex set, then so is

$$\alpha C := \{\alpha \mathbf{x} : \mathbf{x} \in C\} \text{ for } \alpha \in \mathbb{R}.$$

Translation. If $C \subseteq \mathbb{R}^d$ is a convex set, then so is

$$C + \mathbf{a} := \{\mathbf{x} + \mathbf{a} \in \mathbb{R}^d : \mathbf{x} \in C\} \text{ for any } \mathbf{a} \in \mathbb{R}^d.$$

See *Boyd and Vandenberghe* Section 2.3 for reference and more rules.

Convex Functions

Intuition, Definition, and “Algebra”

Convex Function

Idea

A convex function is a function that is “bowl-shaped.”

All line segments through any two points lie above the function.

If differentiable, all tangents are *below* the function.

Convex Function

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and for any scalar $\alpha \in \mathbb{R}$ with $0 \leq \alpha \leq 1$,

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha)f(\mathbf{y}) .$$

That is, the (secant) line segment between any two points lies *above* the function.

Concave functions are negative convex functions.

Convex Function

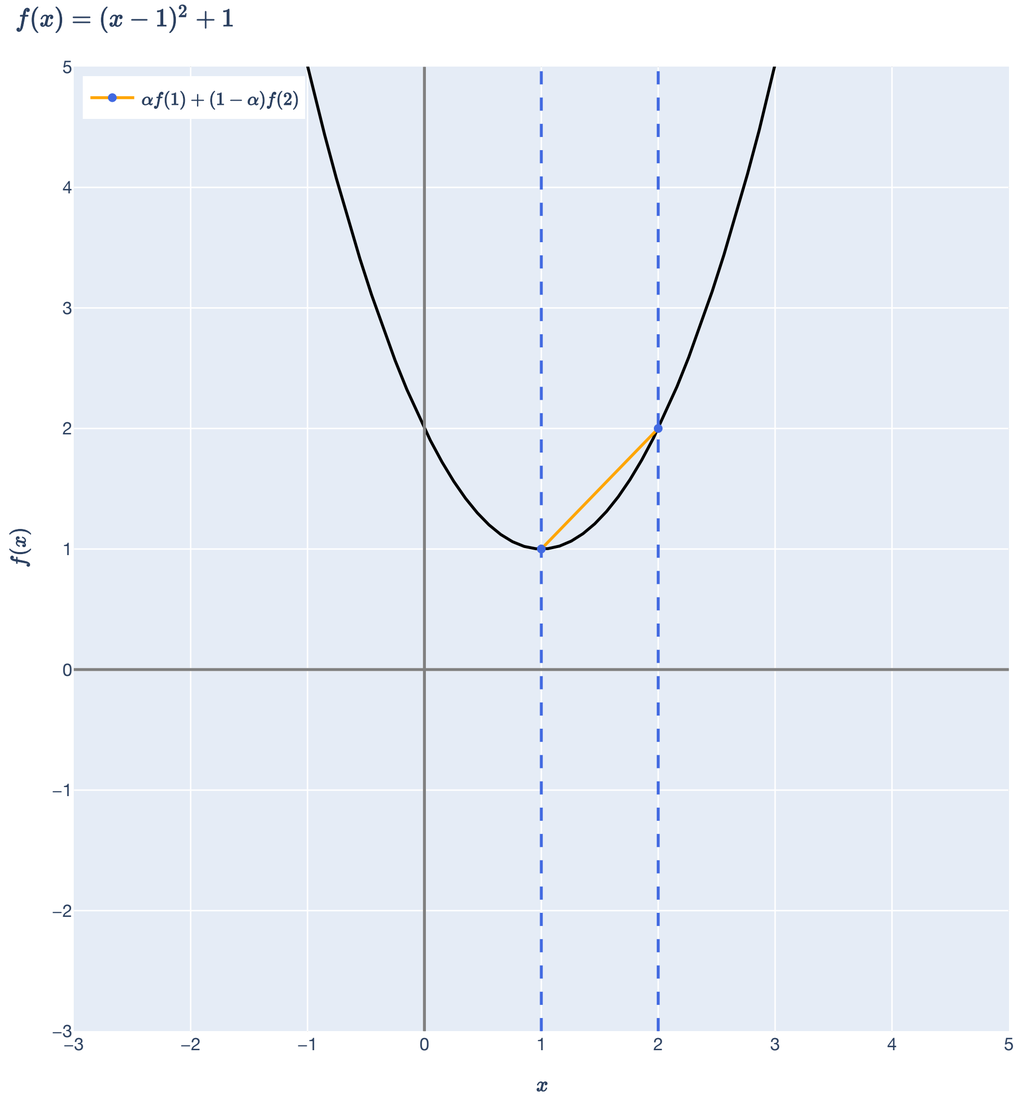
Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and for any scalar $\alpha \in \mathbb{R}$ with $0 \leq \alpha \leq 1$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})\,.$$

That is, the (secant) line segment between any two points lies *above* the function.

Concave functions are negative convex functions.



Convex Function

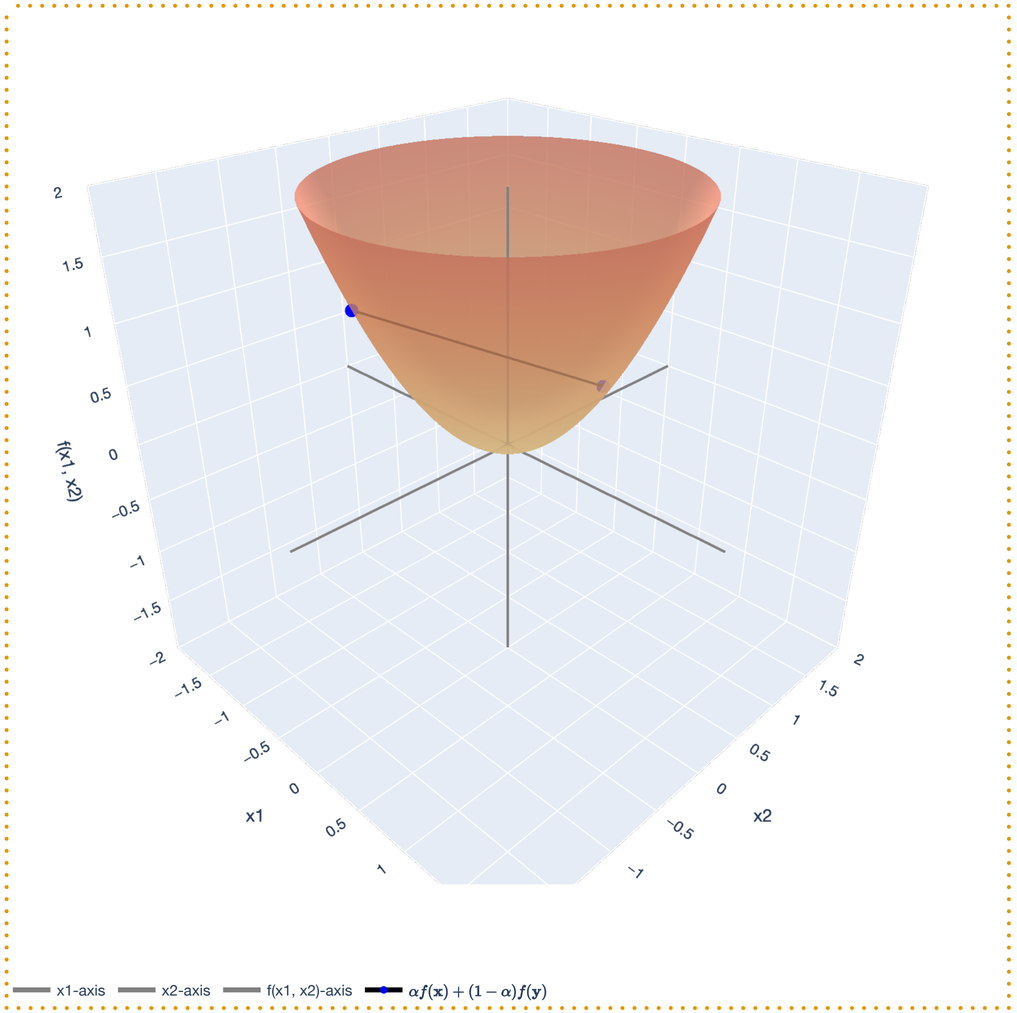
Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and for any scalar $\alpha \in \mathbb{R}$ with $0 \leq \alpha \leq 1$,

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})\,.$$

That is, the (secant) line segment between any two points lies *above* the function.

Concave functions are negative convex functions.



Convex Functions

First-Order Definition of Convexity

A *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \,.$$

The linearization at any $\mathbf{x}$ lies *below* the function.

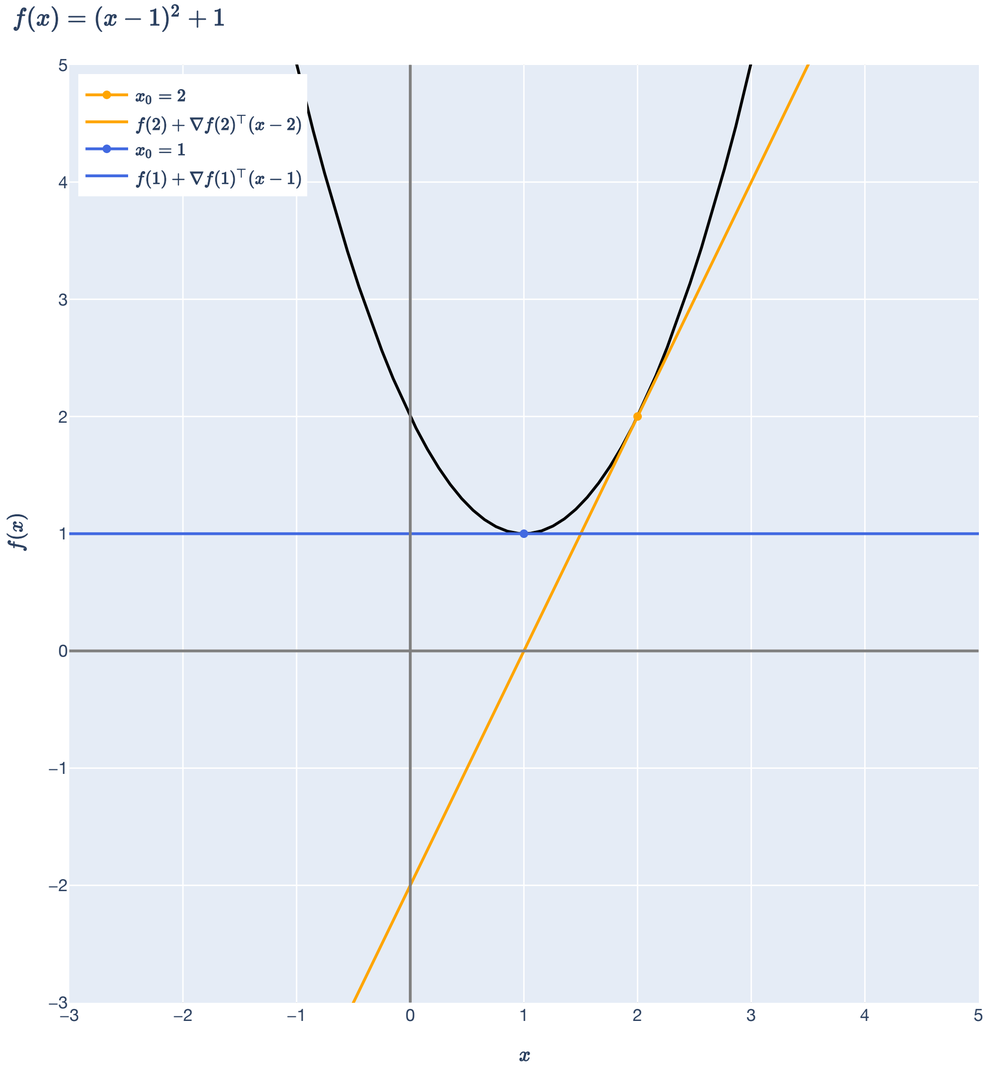
Convex Functions

First-Order Definition of Convexity

A *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \,.$$

The linearization at any $\mathbf{x}$ lies *below* the function.



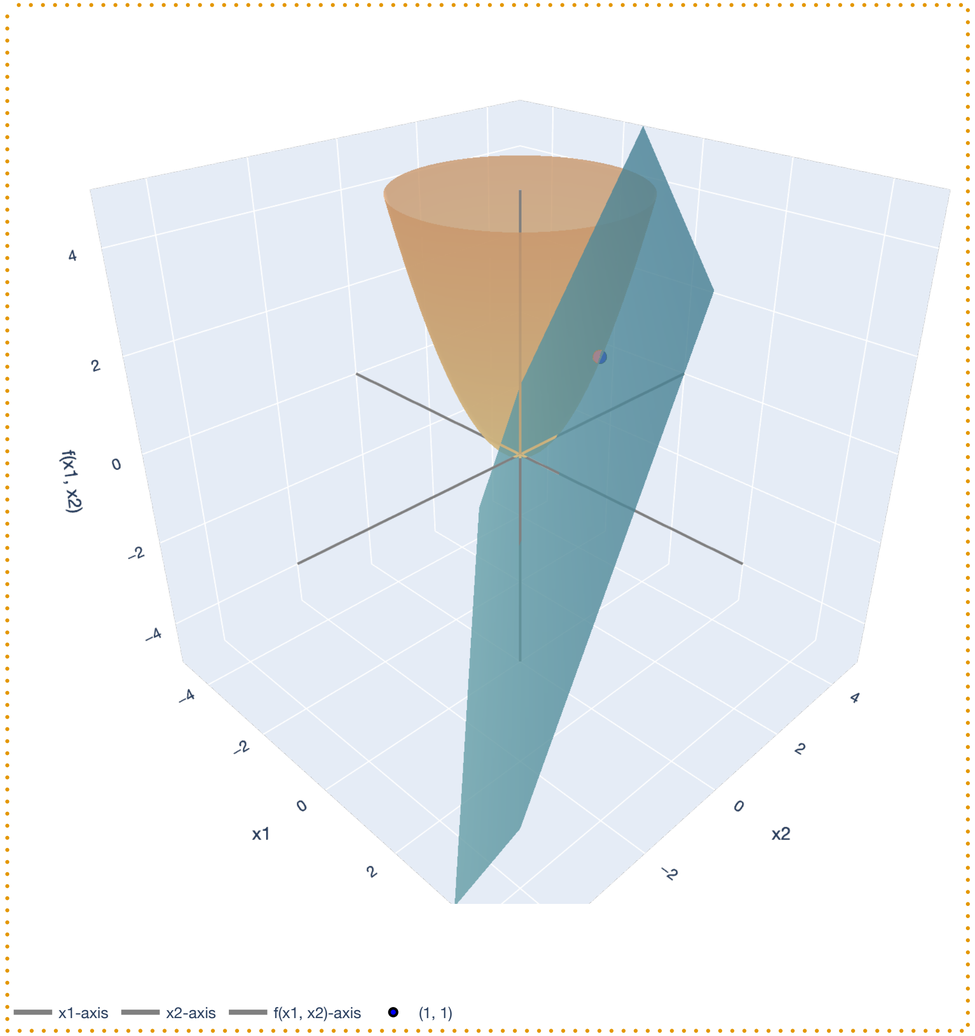
Convex Functions

First-Order Definition of Convexity

A *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \,.$$

The linearization at any $\mathbf{x}$ lies *below* the function.



Convex Functions

Second-Order Definition of Convexity

A *twice-differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function if, for any $\mathbf{x} \in \mathbb{R}^d$, the Hessian $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ is positive semidefinite:

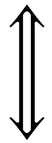
$$\mathbf{d}^\top \nabla^2_{\mathbf{x}} f(\mathbf{x}) \mathbf{d} \geq 0 \text{ for all } \mathbf{d} \in \mathbb{R}^d.$$

The function has a nonnegative “second derivative.”

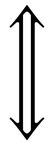
Convex Functions

Three characterizations

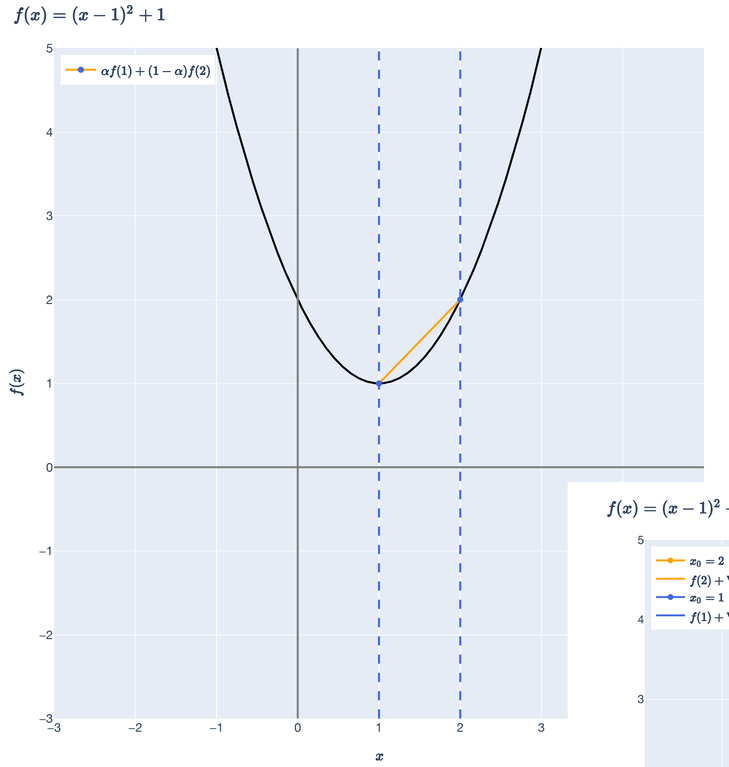
$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})\,.$$

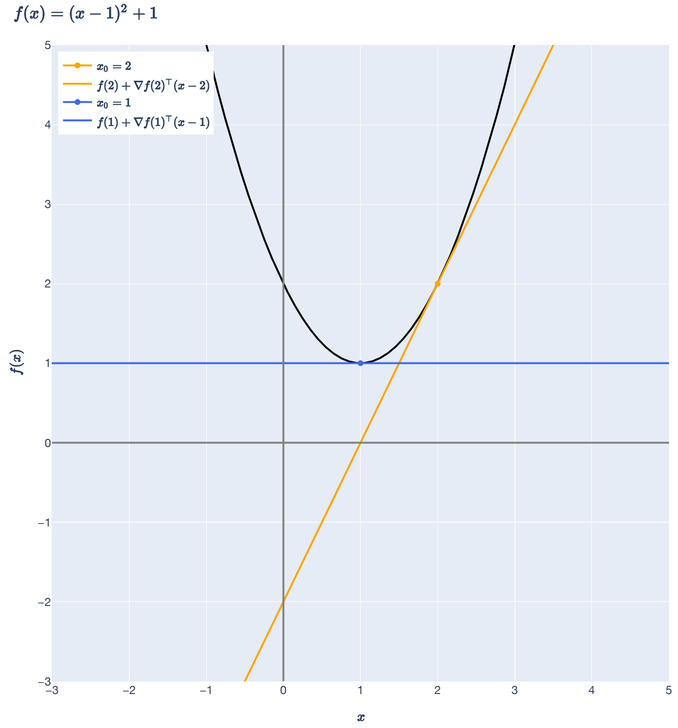


If differentiable: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})^{\top}(\mathbf{y} - \mathbf{x})$



If twice-differentiable: $\mathbf{d}^{\top} \nabla^2_{\mathbf{x}} f(\mathbf{x}) \mathbf{d} \geq 0$ for all $\mathbf{d} \in \mathbb{R}^d$.





Examples of Convex Functions

Quadratic Functions

Always keep this canonical "bowl-shaped" example $f : \mathbb{R} \to \mathbb{R}$ in mind:

$$f(x) = x^2$$

Examples of Convex Functions

Quadratic Forms

More generally, always keep quadratic forms $f : \mathbb{R}^d \to \mathbb{R}$ in mind:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} \text{ for symmetric } d \times d \text{ matrix } \mathbf{A}.$$

Examples of Convex Functions

Affine Functions

Let $\mathbf{w} \in \mathbb{R}^d$ be some vector and let $b \in \mathbb{R}$ be some scalar.

$$f(\mathbf{x}) := \mathbf{w}^\top \mathbf{x} + b.$$

Examples of Convex Functions

Other examples of convex functions on $\mathbb{R}$

Exponential. e^{ax} is convex for any $a \in \mathbb{R}$.

Powers. x^a is convex on $(0,\infty)$ for any $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.

Powers of absolute values. $|x|^p$ is convex on $\mathbb{R}$, for any $p \geq 1$.

Logarithm. $\log x$ is concave on $(0,\infty)$.

Negative entropy. $x \log x$ is convex on $(0,\infty)$, or convex on $[0,\infty)$ if we define $0 \log 0 := 0$.

Examples of Convex Functions

Other examples of convex functions on $\mathbb{R}^d$

Norms. Any norm $\|\cdot\|$ on $\mathbb{R}^d$ is convex. This includes the *Euclidean/ℓ_2 norm:*

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{d} x_i^2}$$

Max function. The function $f(\mathbf{x}) := \max\{x_1, \ldots, x_d\}$ is convex.

Log-sum-exp. The function $f(\mathbf{x}) := \log\left(e^{x_1} + \ldots + e^{x_d}\right)$ is convex.

Closure of Convex Functions

The "Algebra" of Convex Functions

We can also combine convex functions with operations that preserve convexity:

Nonnegative weighted sum. If $f_1, \ldots, f_n$ convex, then $g(\mathbf{x}) := \lambda_1 f_1(\mathbf{x}) + \ldots + \lambda_n f_n(\mathbf{x})$ is convex.

Extends to infinite sums and integrals.

Pre-composition with affine function. If f is convex, so is $f(\mathbf{A}\mathbf{x} + \mathbf{b})$.

Maximum. If $f_1, \ldots, f_n$ are convex, then $g(\mathbf{x}) := \max\{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$ is convex.

Extends to pointwise supremum.

See *Boyd and Vandenberghe* Section 3.2 for comprehensive reference.

Verifying Convexity

In order of preference…

To verify that $f : \mathbb{R}^d \to \mathbb{R}$ is convex:

1. Construct function from known convex functions (e.g. exponential, affine, etc.) and closure properties.

2. *If differentiable/twice-differentiable:* Use first-order or second-order equivalent definitions of convexity.

3. Restrict to a line: $f : C \to \mathbb{R}$ is convex if and only if, for every $\mathbf{x}, \mathbf{y} \in C$, if the function $g(\alpha) := f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y})$ is convex for $\alpha \in [0,1]$.

4. Directly verify using the definition of convexity: $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$.

Convex Optimization

Local minima are global minima

Convex Optimization

Optimality condition

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

where f is a *convex function* and $\mathscr{C}$ is a *convex set*.

The most important property of these optimization problems is:

All local minima are global minima!

Convex Optimization

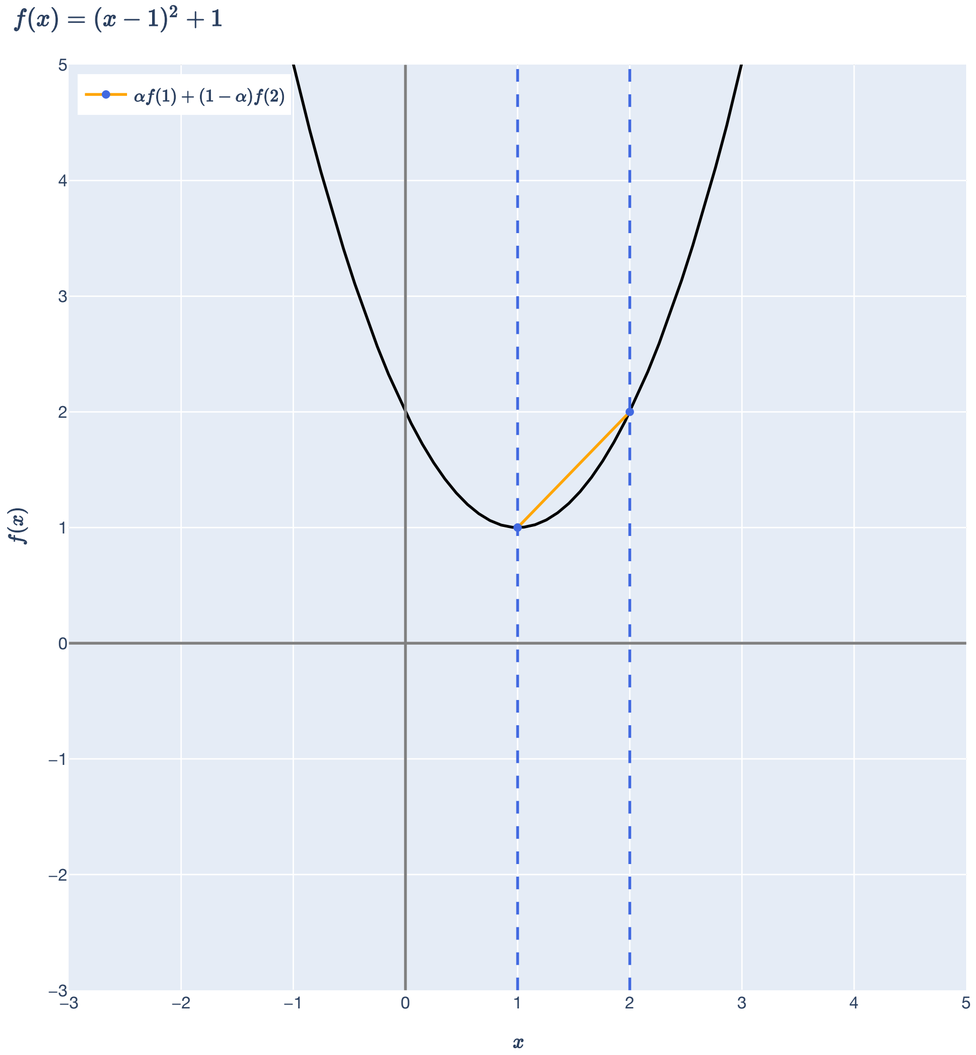
Optimality condition

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

where f is a *convex function* and $\mathscr{C}$ is a *convex set*.

The most important property of these optimization problems is:

All local minima are global minima!



Convex Optimization

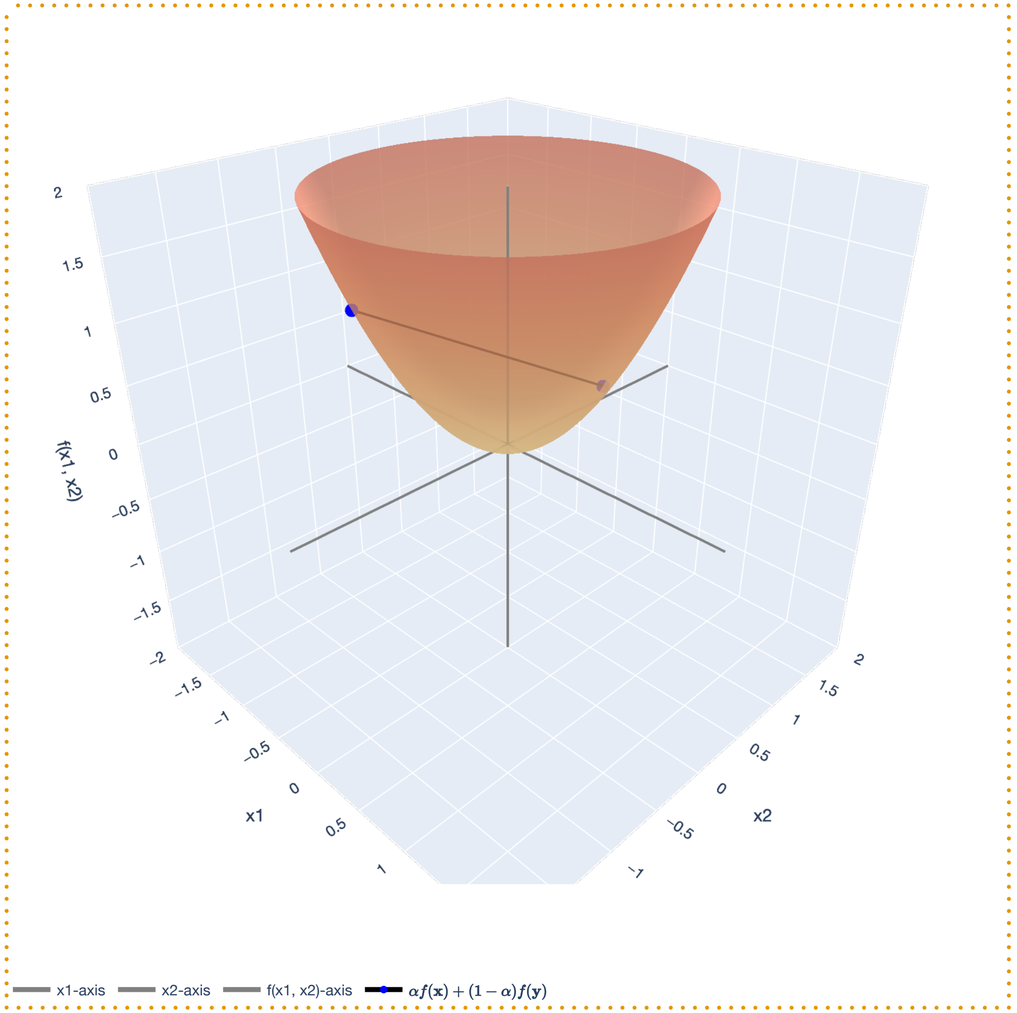
Optimality condition

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

where f is a *convex function* and $\mathscr{C}$ is a *convex set*.

The most important property of these optimization problems is:

All local minima are global minima!



Convex Optimization

Main Optimality Theorem

Theorem (Optimality for convex optimization). For a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex set $\mathscr{C} \subseteq \mathbb{R}^d$, consider the optimization problem:

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

Then, if $\mathbf{x}^* \in C$ is a *local minimum,* it must also be a *global minimum*:

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

Convex Optimization

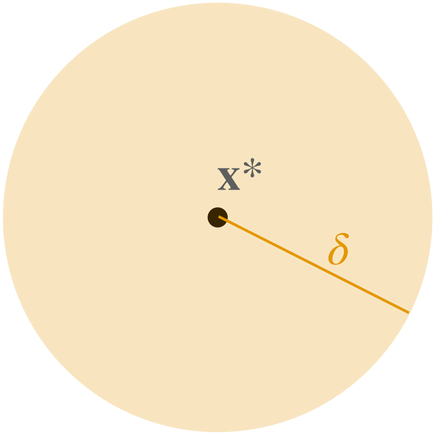
Step 1: Use definition of local minimum

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

Because $\mathbf{x}^*$ is a local minimum, there is a neighborhood $B_\delta(\mathbf{x}^*)$ around $\mathbf{x}^*$ such that

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

This allows us to move in all (*feasible*) directions from $\mathbf{x}^*$.



Convex Optimization

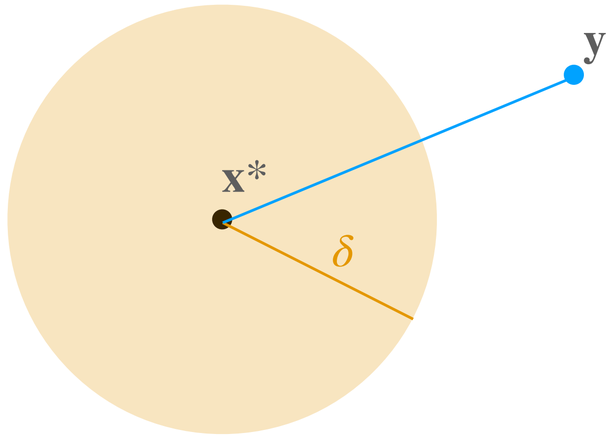
Step 2: Consider line segment to another point

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

From Step 1, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{C} \cap B_\delta(\mathbf{x}^*)$.

Choose *any* $\mathbf{y} \in \mathscr{C}$, not necessarily in $B_\delta(\mathbf{x}^*)$, and consider the line segment $[\mathbf{x}^*, \mathbf{y}]$ defined by:

$$[\mathbf{x}^*, \mathbf{y}] := \{(1-\alpha)\mathbf{x}^* + \alpha\mathbf{y} : \alpha \in [0,1]\}.$$



Convex Optimization

Step 3: Take a small step in line segment direction

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

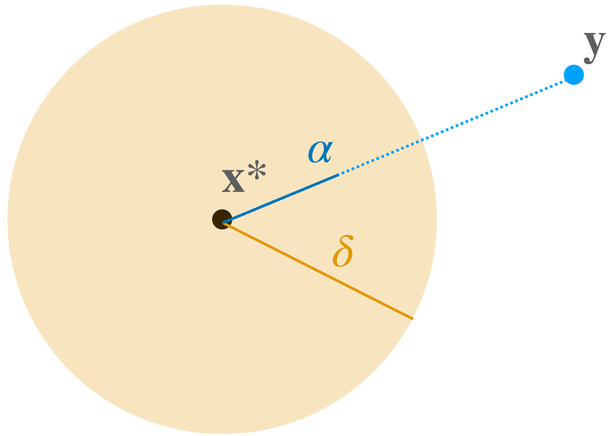
From Step 1, we got a neighborhood, $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{C} \cap B_\delta(\mathbf{x}^*)$.

From Step 2, we got the line segment:

$$[\mathbf{x}^*, \mathbf{y}] := \{(1-\alpha)\mathbf{x}^* + \alpha\mathbf{y} : \alpha \in [0,1]\}.$$

For $\alpha < \delta$ (sufficiently small), we're still in the neighborhood, so:

$$f(\mathbf{x}^*) \leq f((1-\alpha)\mathbf{x}^* + \alpha\mathbf{y}).$$



Convex Optimization

Step 4: Use convexity to extrapolate *outside* of the neighborhood

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

For $\alpha < \delta$ (sufficiently small), we're still in the neighborhood, so:

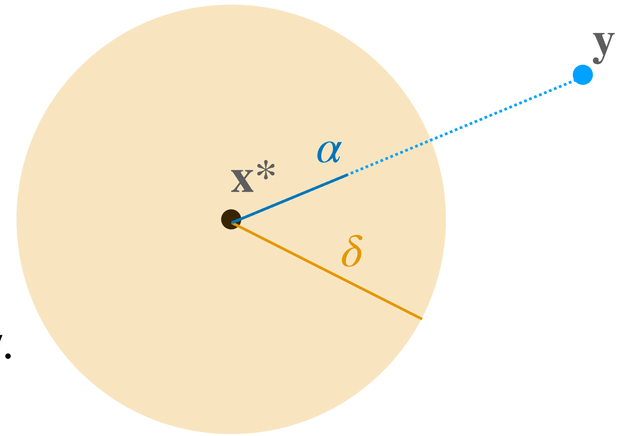
$$f(\mathbf{x}^*) \leq f((1-\alpha)\mathbf{x}^* + \alpha\mathbf{y}).$$

Using the definition of convexity,

$$\begin{aligned} f(\mathbf{x}^*) &\leq f((1-\alpha)\mathbf{x}^* + \alpha\mathbf{y}) \\ &\leq (1-\alpha)f(\mathbf{x}^*) + \alpha f(\mathbf{y}) \end{aligned}$$

Rearranging, we get:

$f(\mathbf{x}^*) \leq f(\mathbf{y})$, where we chose $\mathbf{y} \in \mathscr{C}$ arbitrarily.



Convex Optimization

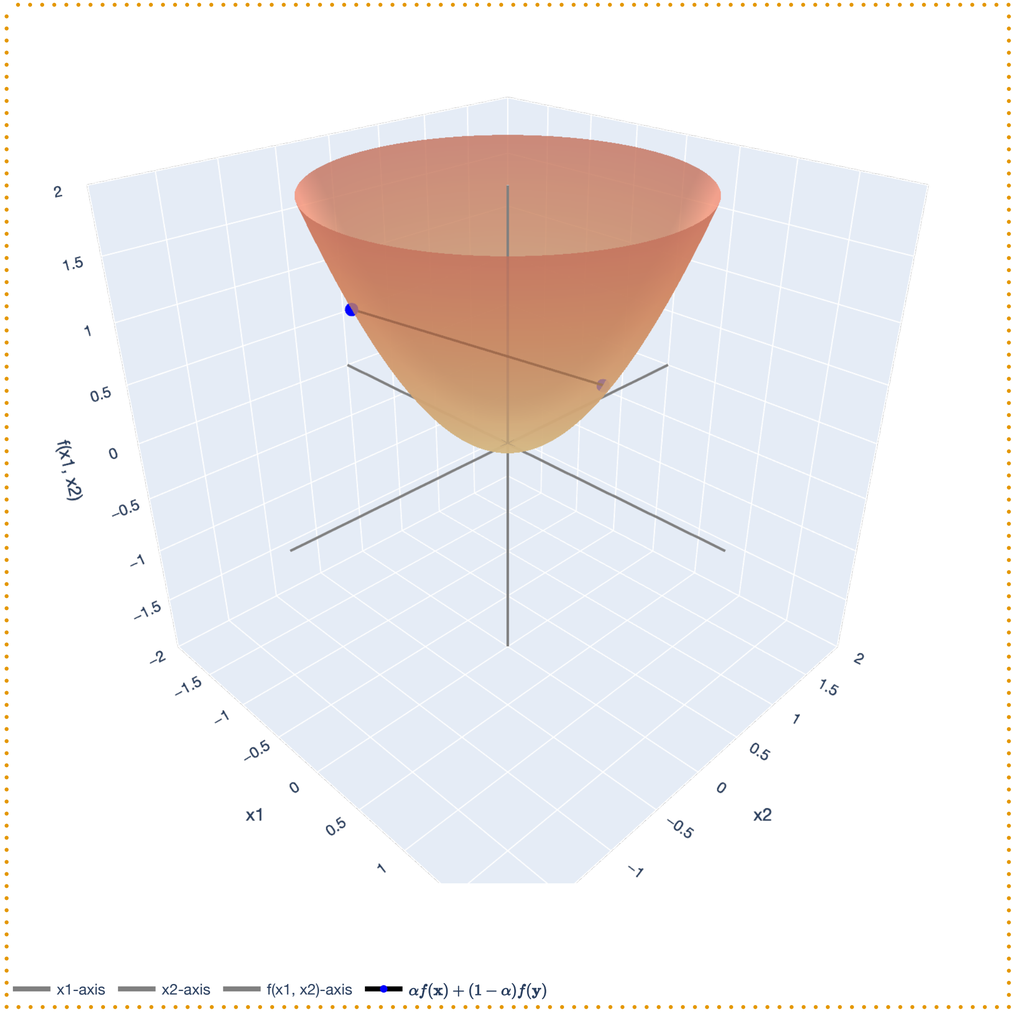
Main Optimality Theorem

Theorem (Optimality for convex optimization). For a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex set $\mathscr{C} \subseteq \mathbb{R}^d$, consider the optimization problem:

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

Then, if $\mathbf{x}^* \in C$ is a *local minimum,* it must also be a *global minimum*:

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



Convex Optimization

Optimality Theorem for Differentiable Functions

Theorem (Optimality for convex optimization for differentiable functions). For a convex, *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex set $\mathscr{C} \subseteq \mathbb{R}^d$

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

Then, $\mathbf{x}^* \in \mathscr{C}$ is a global minimum if and only if:

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0 \text{ for all } \mathbf{x} \in \mathscr{C}.$$

Convex Optimization

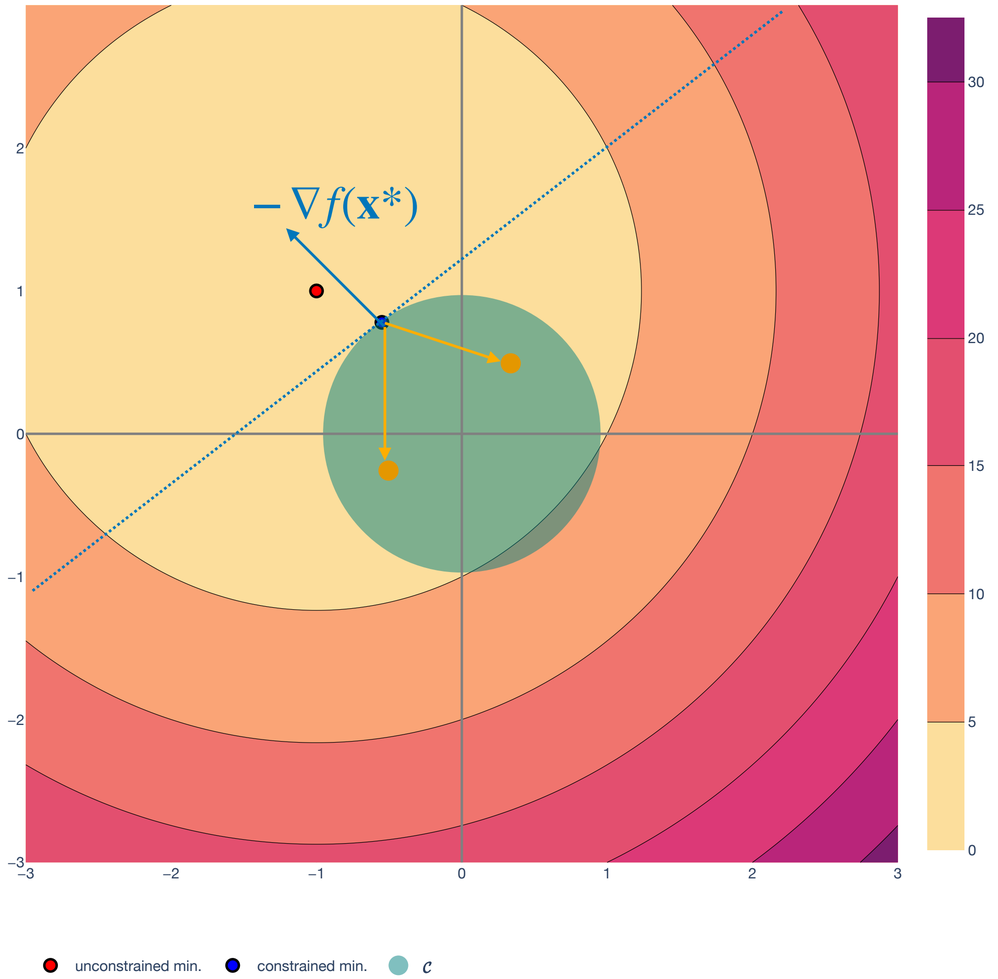
Optimality Theorem for Differentiable Functions

Theorem (Optimality for convex optimization for differentiable functions). For a convex, *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ and a convex set $\mathscr{C} \subseteq \mathbb{R}^d$

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

Then, $\mathbf{x}^* \in \mathscr{C}$ is a global minimum if and only if:

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0 \text{ for all } \mathbf{x} \in \mathscr{C}.$$



Gradient Descent and Convexity

Theorem Statement and Proof

Types of Minima

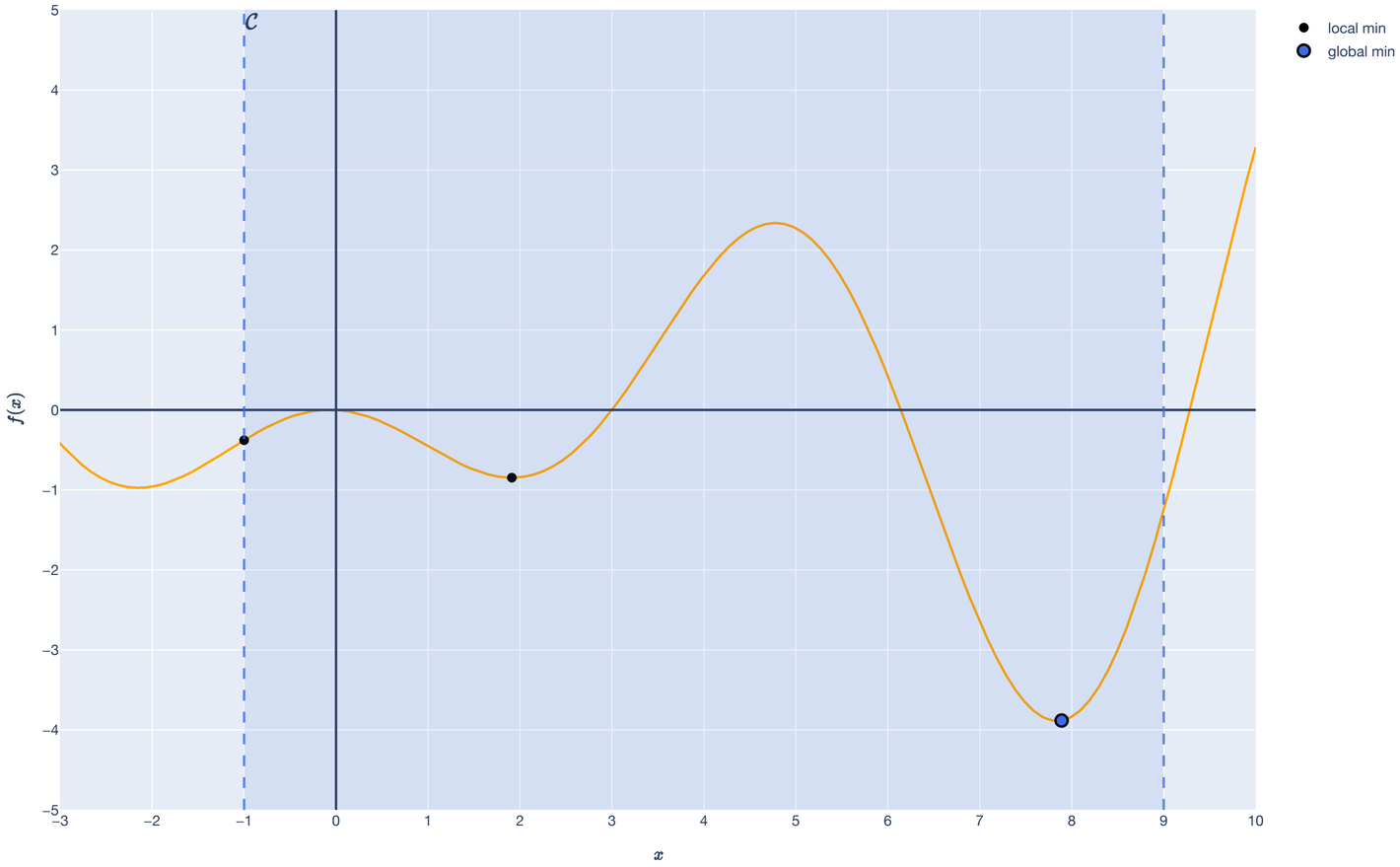
Big picture

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.

Often hard to do analytically!



Gradient Descent

Algorithm

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

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

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

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

Gradient Descent

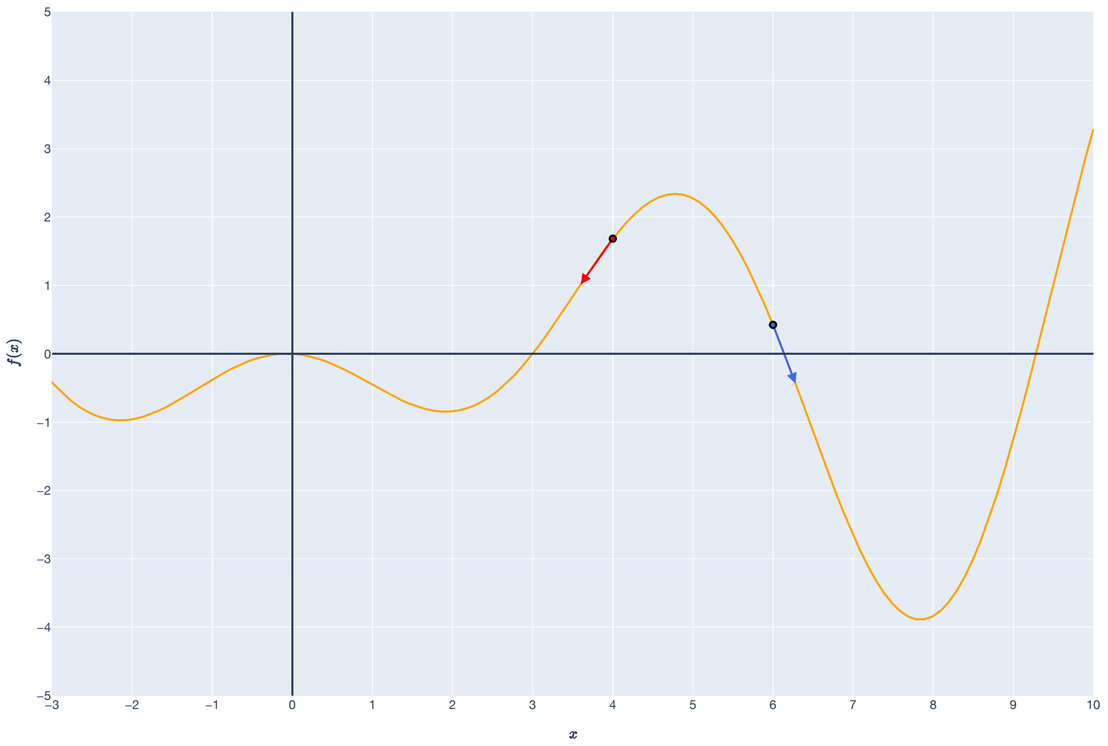
Theorem 1: Descent Lemma (Formal)

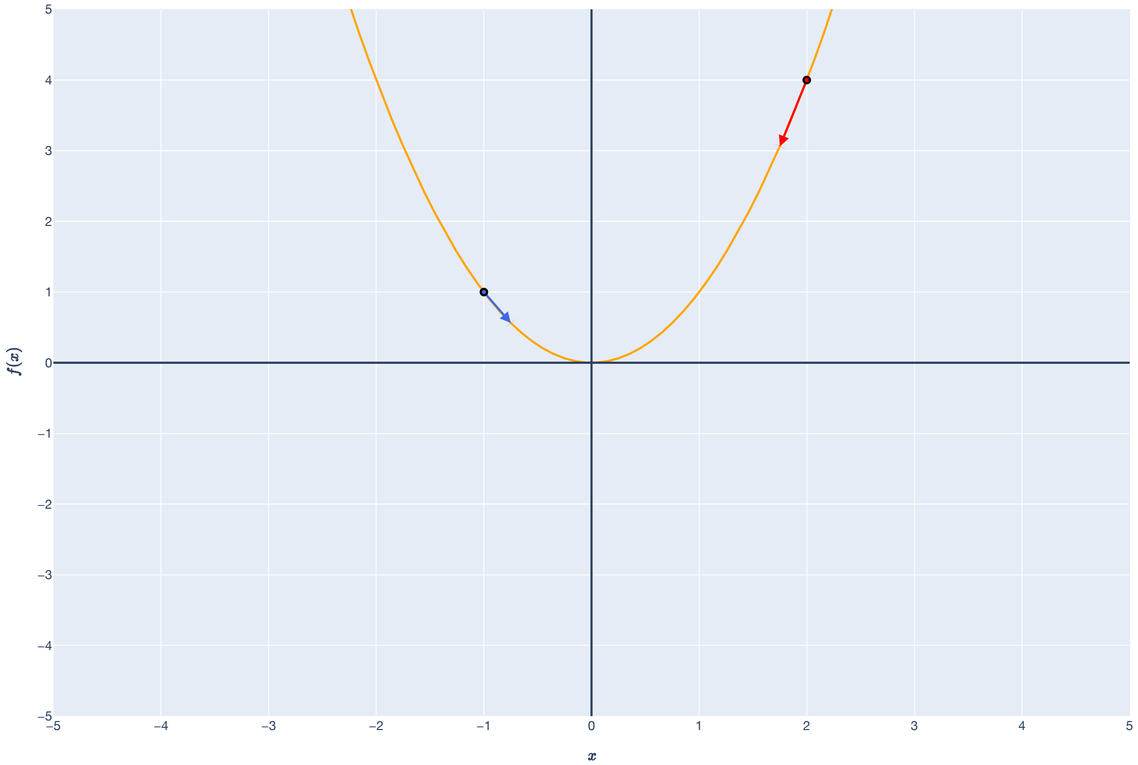
Theorem (Descent Lemma). If $f \in \mathscr{C}^2$ and is β-smooth, then with $\eta = 1/\beta$, for any $\mathbf{w} \in \mathbb{R}^d$,

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

Gradient Descent

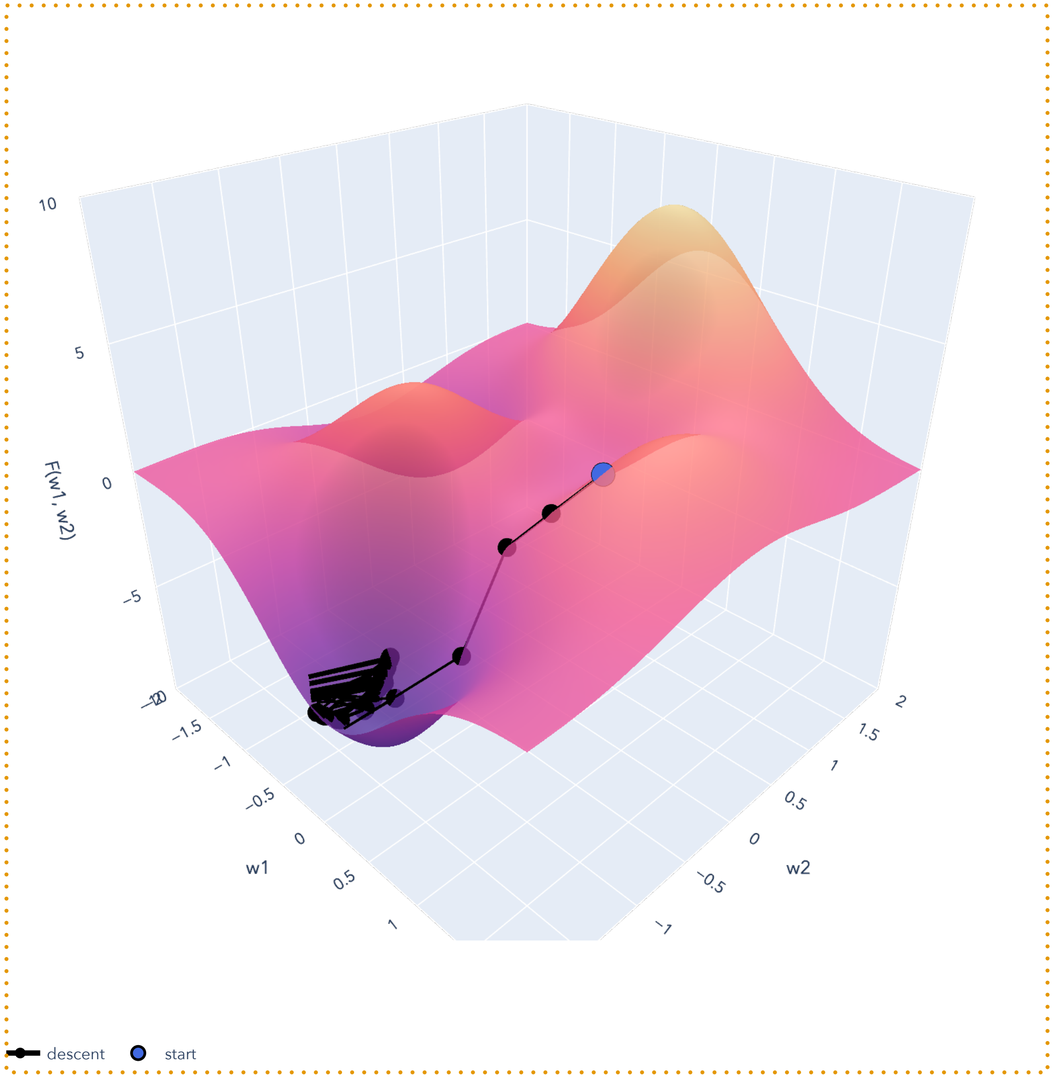
Behavior for $d = 1$ “Bowl-shaped” Functions

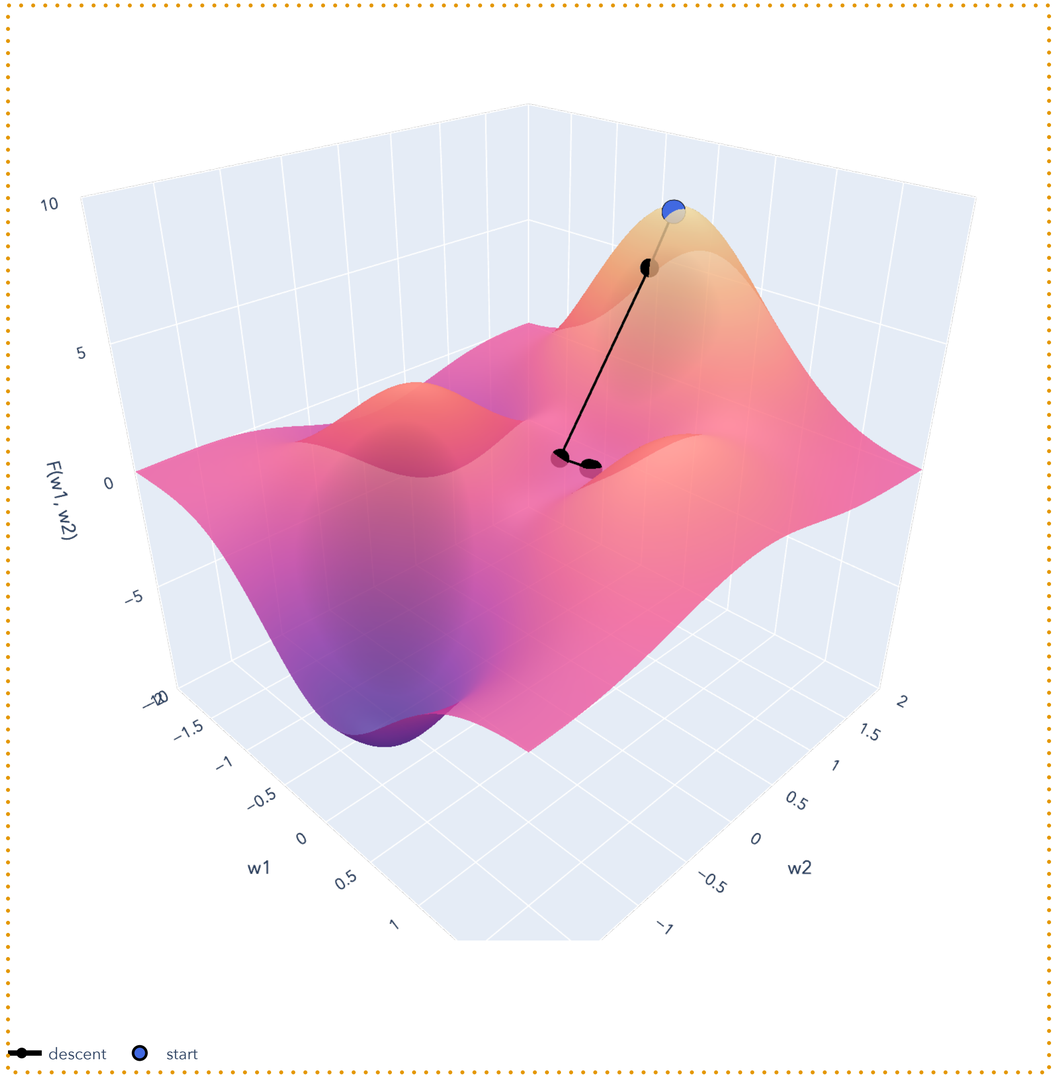




Gradient Descent

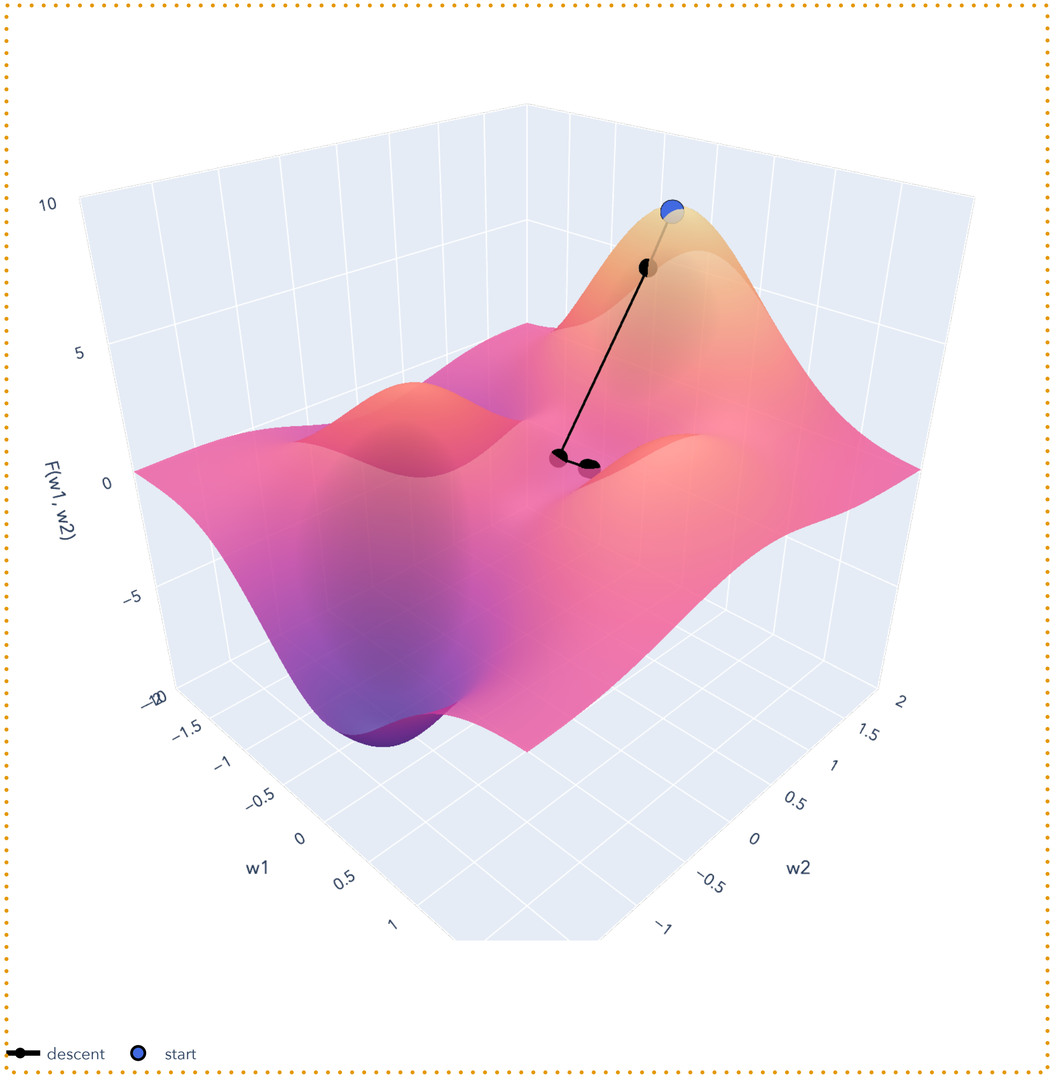
Behavior for $d = 2$ "Bowl-shaped" Functions

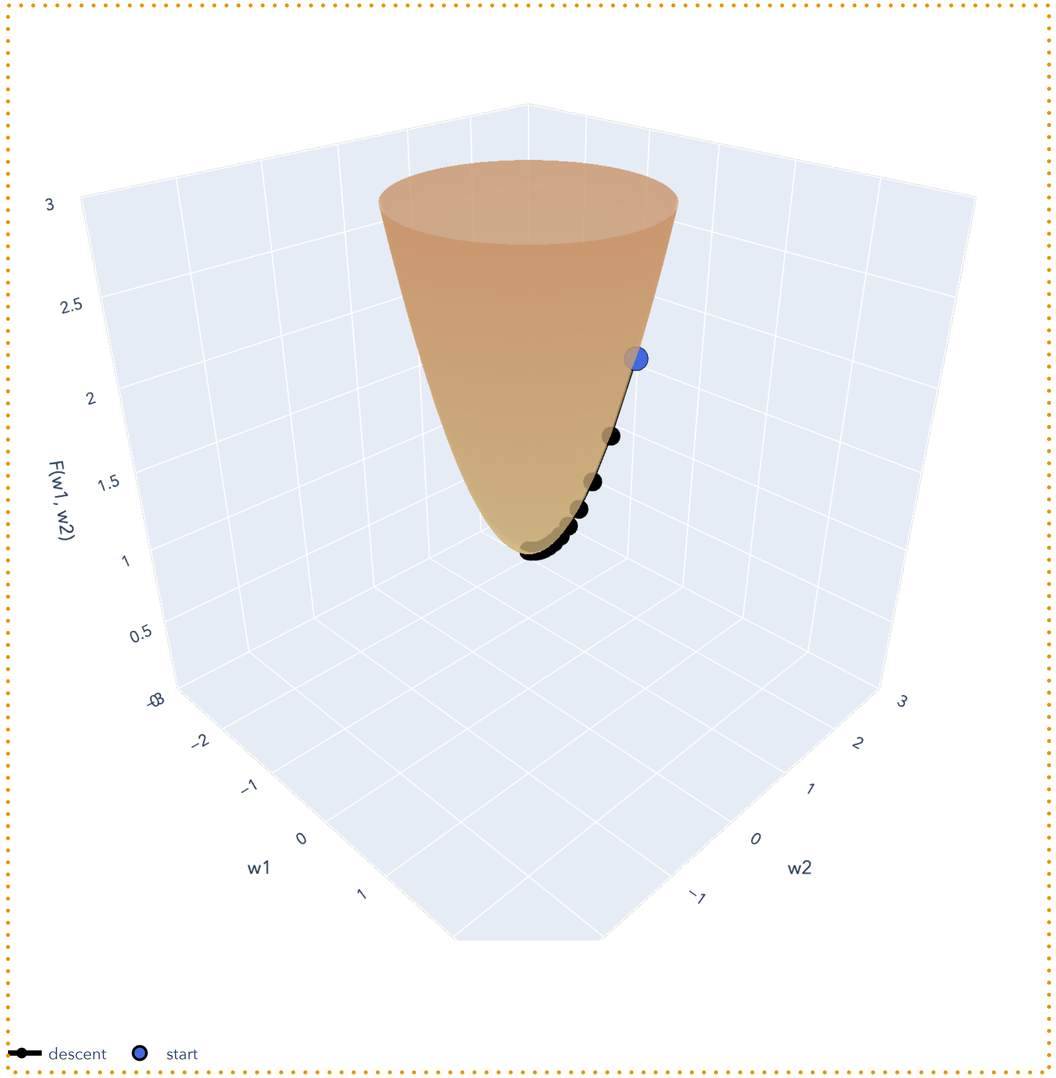




Gradient Descent

Behavior for $d = 2$ "Bowl-shaped" Functions





Gradient Descent

Theorem 1: Descent Lemma (Formal)

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

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

This theorem does NOT guarantee that we'll reach a global *minimum!*

Gradient Descent

Theorem 2: GD on convex, smooth functions

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

If we run gradient descent with step size $\eta = \frac{1}{\beta}$ and initial point $\mathbf{x}_0 \in \mathbb{R}^d$,

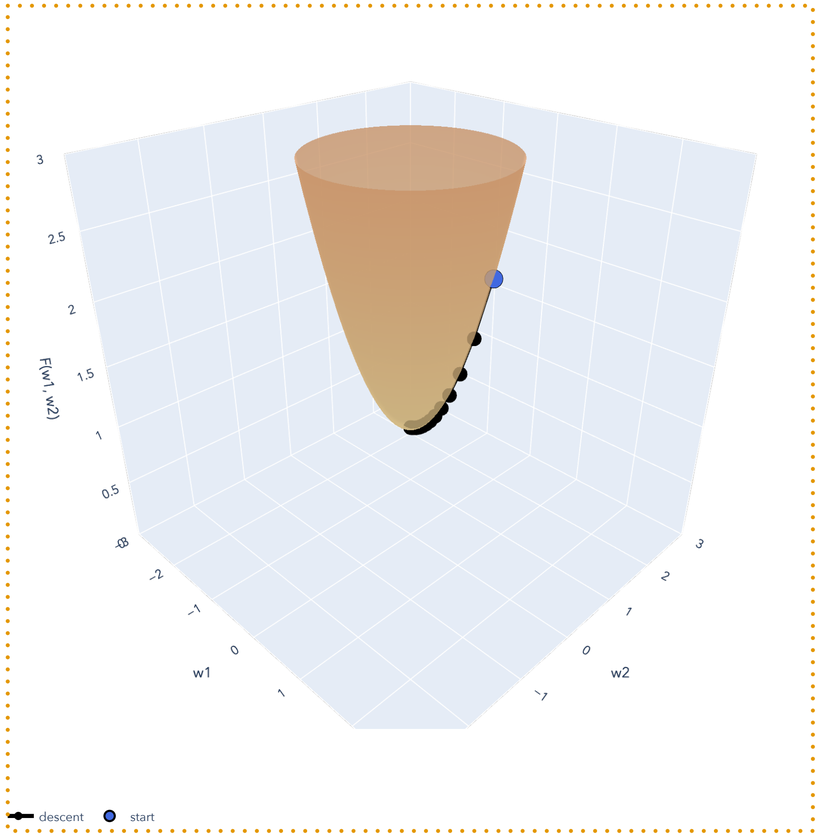
$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right),$$

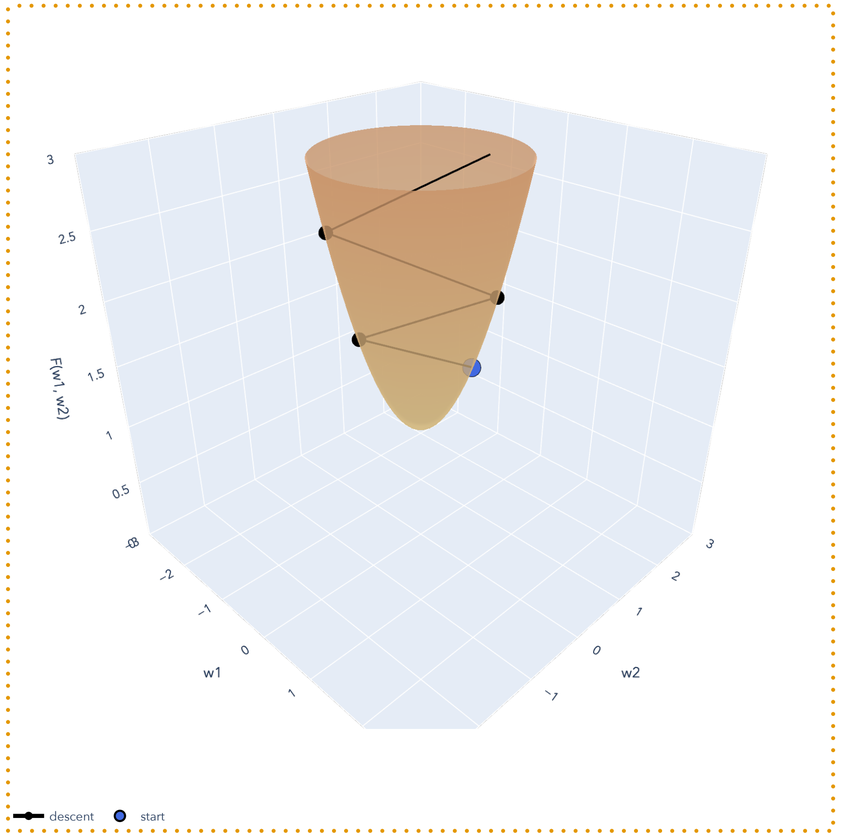
after T iterations of our algorithm.

Gradient Descent

Theorem 2: GD on convex, smooth functions

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

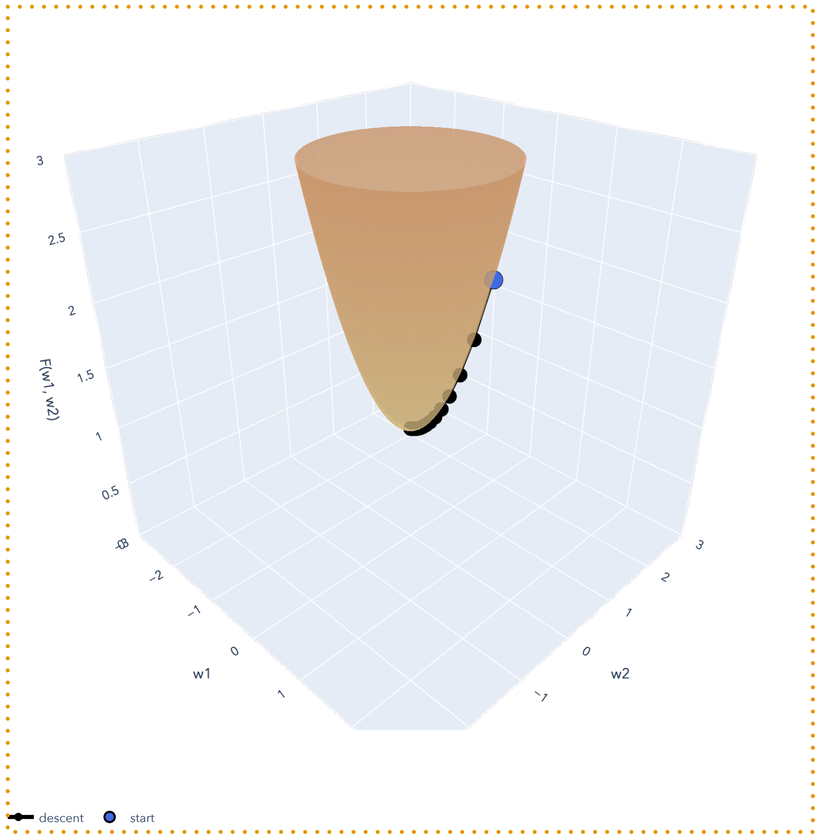


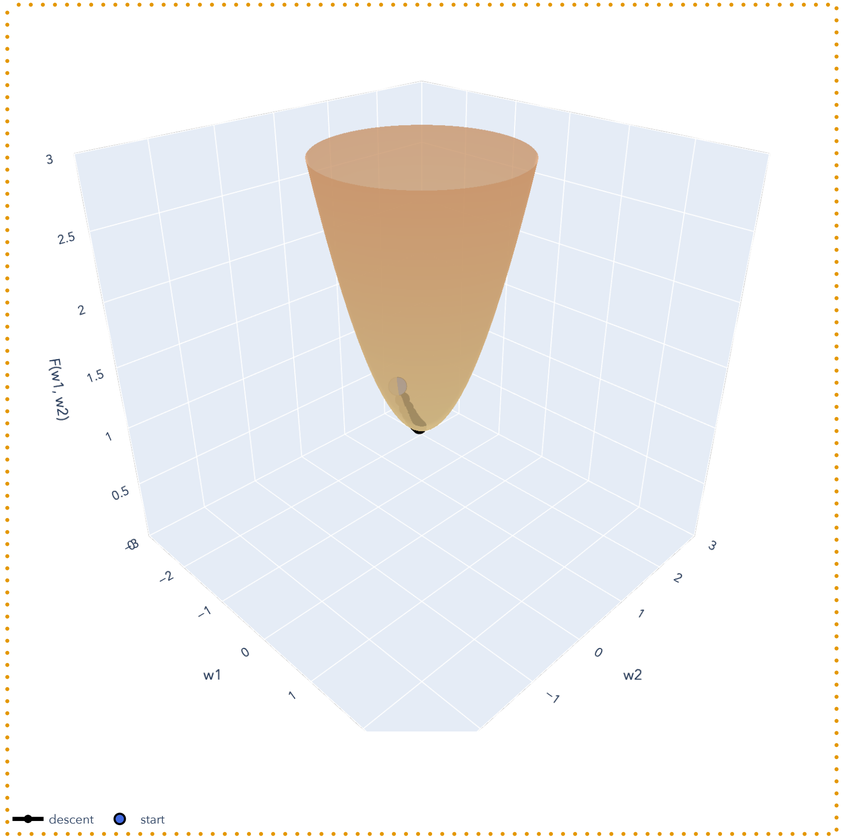


Gradient Descent

Theorem 2: GD on convex, smooth functions

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





Gradient Descent

Proof of GD Theorem for Convex, β-smooth functions

We want to show:

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right), \text{ after } T \text{ iterations of GD.}$$

Descent lemma. For any iteration $t = 1,2,\ldots,T$,

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

First-order definition of convexity. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + f(\mathbf{x}) \leq f(\mathbf{y}).$$

Gradient Descent

Step 1: Define "potential function"

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

Fix the optimal $\mathbf{x}^* \in \mathbb{R}^d$. Consider the "potential" function $\Phi : \mathbb{R}^d \to \mathbb{R}$:

$$\Phi(\mathbf{x}) = \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}^*\|^2.$$

This tracks our distance from the minimum, $\mathbf{x}^*$. At $\mathbf{x}_{t-1}$, our potential is:

$$\Phi(\mathbf{x}_{t-1}) = \frac{\beta}{2}\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2, \text{ where we chose } \eta = 1/\beta.$$

Gradient Descent

Step 2: Analyze drop in potential from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

Make sure that the the potential "drops" by a positive amount in each step: $\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) \geq 0$.

Analyze this quantity, plugging in the GD step: $\mathbf{x}_t = \mathbf{x}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{x}_{t-1})$.

$$\begin{aligned}
\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) &= \frac{\beta}{2}\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{\beta}{2}\|\mathbf{x}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{x}_{t-1}) - \mathbf{x}^*\|^2 \\
&= \frac{\beta}{2}\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{\beta}{2}\left(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{2}{\beta}(\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1}) + \frac{1}{\beta^2}\|\nabla f(\mathbf{x}_{t-1})\|^2\right) \\
&= (\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1}) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2.
\end{aligned}$$

Gradient Descent

Step 3: Deal with $(\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1})$ using first-order def. of convexity

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

For any $\mathbf{x}_{t-1} \in \mathbb{R}^d$ and $\mathbf{x}^* \in \mathbb{R}^d$,

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}^* - \mathbf{x}_{t-1}) + f(\mathbf{x}_{t-1}) \leq f(\mathbf{x}^*).$$

Rearranging, we get a *lower bound*:

$$\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{x}^*) \geq f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*)$$

Gradient Descent

Step 2: Analyze drop in potential from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

Make sure that the the potential "drops" by a positive amount in each step: $\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) \geq 0$.

Analyze this quantity, plugging in the GD step: $\mathbf{x}_t = \mathbf{x}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{x}_{t-1})$.

$$\begin{aligned}
\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) &= \frac{\beta}{2}\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{\beta}{2}\|\mathbf{x}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{x}_{t-1}) - \mathbf{x}^*\|^2 \\
&= \frac{\beta}{2}\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{\beta}{2}\left(\|\mathbf{x}_{t-1} - \mathbf{x}^*\|^2 - \frac{2}{\beta}(\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1}) + \frac{1}{\beta^2}\|\nabla f(\mathbf{x}_{t-1})\|^2\right) \\
&= (\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1}) - \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2. \\
&\geq f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*)
\end{aligned}$$

Gradient Descent

Step 4: Deal with $(1/2\beta)\|\nabla f(\mathbf{x}_{t-1})\|^2$ using descent lemma

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

By descent lemma for β-smooth functions:

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

Rearranging, we can lower bound:

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

Gradient Descent

Step 5: Lower bound drop in potential

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

Make sure that the the potential "drops" by a positive amount in each step: $\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) \geq 0$.

Analyze this quantity, plugging in the GD step: $\mathbf{x}_t = \mathbf{x}_{t-1} - \frac{1}{\beta}\nabla f(\mathbf{x}_{t-1})$.

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

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

$$= \underbrace{(\mathbf{x}_{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}_{t-1})}_{\geq f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*)} \underbrace{- \frac{1}{2\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2}_{\geq f(\mathbf{x}_t) - f(\mathbf{x}_{t-1})} \geq \underbrace{f(\mathbf{x}_{t-1}) - f(\mathbf{x}^*) + f(\mathbf{x}_t) - f(\mathbf{x}_{t-1})}_{= f(\mathbf{x}_t) - f(\mathbf{x}^*)}$$

Gradient Descent

Step 5: Lower bound drop in potential

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

The "drop in potential" is at least $f(\mathbf{x}_t) - f(\mathbf{x}^*)$.

$$\Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) \geq f(\mathbf{x}_t) - f(\mathbf{x}^*)$$

This means our potential always drops by a positive amount if we're not yet at the minimum!

Gradient Descent

Step 6: Sum up and telesecope

Goal: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{\beta}{2T}\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_T - \mathbf{x}^*\|^2\right)$, after T iterations of GD.

$$\sum_{t=1}^{T} \Phi(\mathbf{x}_{t-1}) - \Phi(\mathbf{x}_t) \geq \sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*)$$

Simplify the left-hand side $\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T) \geq \sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*)$ by telescoping sum.

Simplify the right-hand side $\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_T) \geq \sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq T(f(\mathbf{x}_T) - f(\mathbf{x}^*))$ by bounding $f(\mathbf{x}_t) \geq f(\mathbf{x}_T)$.

By the definition of potential $\Phi(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$, we proved our claim:

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

Gradient Descent

Theorem 2: GD on convex, smooth functions

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

If we run gradient descent with step size $\eta = \frac{1}{\beta}$ and initial point $\mathbf{x}_0 \in \mathbb{R}^d$,

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

after T iterations of our algorithm.

Gradient descent always *eventually reaches minimum for convex functions!*

Gradient Descent and OLS

“Uniting” our two main stories

Gradient Descent and OLS

Verifying OLS fits our theorem

We just need to $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ to be $\mathscr{C}^2$, β-smooth, and convex.

1. $\mathscr{C}^2$. Hessian is $\nabla^2 f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}$.

2. **β-smooth.** Recall the definition: $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq \beta$. Satisfied as long as:

$$\lambda_{\max}(\mathbf{X}^\top\mathbf{X}) \leq \beta/2.$$

3. **Convex.** Can use definition, first-order definition, or second-order definitions.

Gradient Descent and OLS

Uniting our two stories

Theorem (GD applied to OLS). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ be fixed. Let the maximum eigenvalue $\lambda_{\max}$ of $\mathbf{X}^\top \mathbf{X}$ satisfy $\lambda_{\max} \leq \beta/2$. Let $\mathbf{w}^*$ be a (global) minimizer of $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, satisfying:

$$\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2 \leq \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \text{ for all } \mathbf{w} \in \mathbb{R}^d.$$

After T iterations of gradient descent with step size $\eta = 1/\beta$ and initial point $\mathbf{w}_0 \in \mathbb{R}^d$:

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

Gradient Descent

Algorithm for OLS

What does gradient descent look like for OLS? Recall the objective function and its gradient:

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

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

Gradient Descent

Algorithm for OLS

$\nabla f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}$, so the gradient descent algorithm for OLS is:

Make an initial guess $\mathbf{w}_0$.

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

Compute: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta\mathbf{X}^\top\left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$.

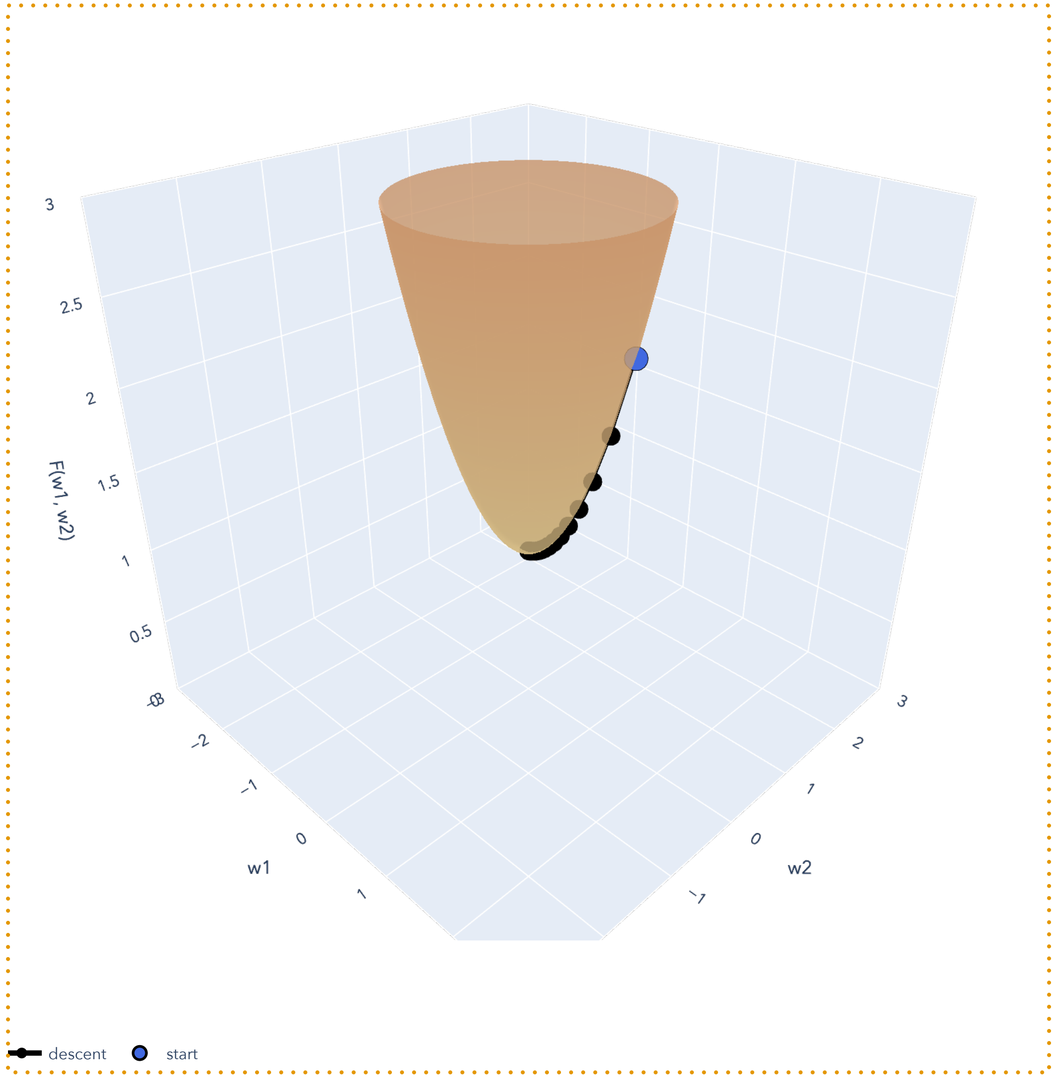
Gradient Descent

Algorithm for OLS

Make an initial guess $\mathbf{w}_0$.

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

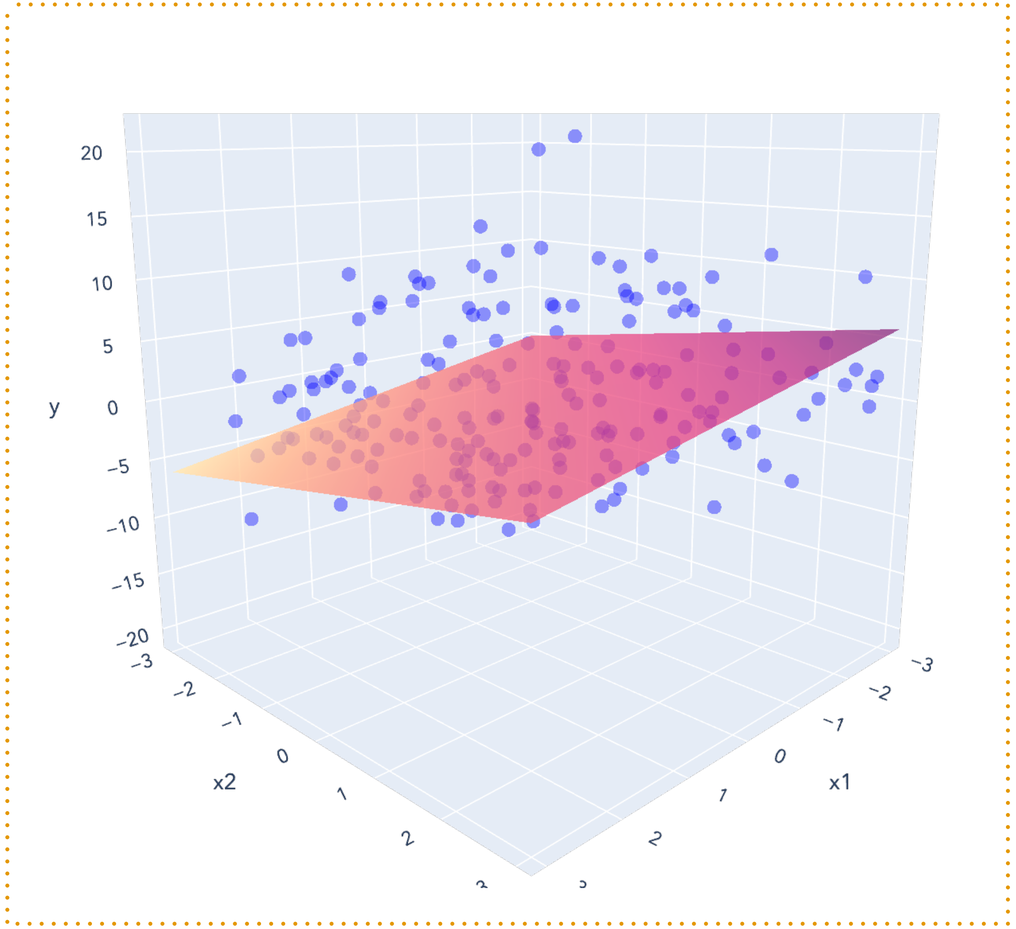
Compute: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta\mathbf{X}^\top\left(\mathbf{X}\mathbf{w} - \mathbf{y}\right)$.

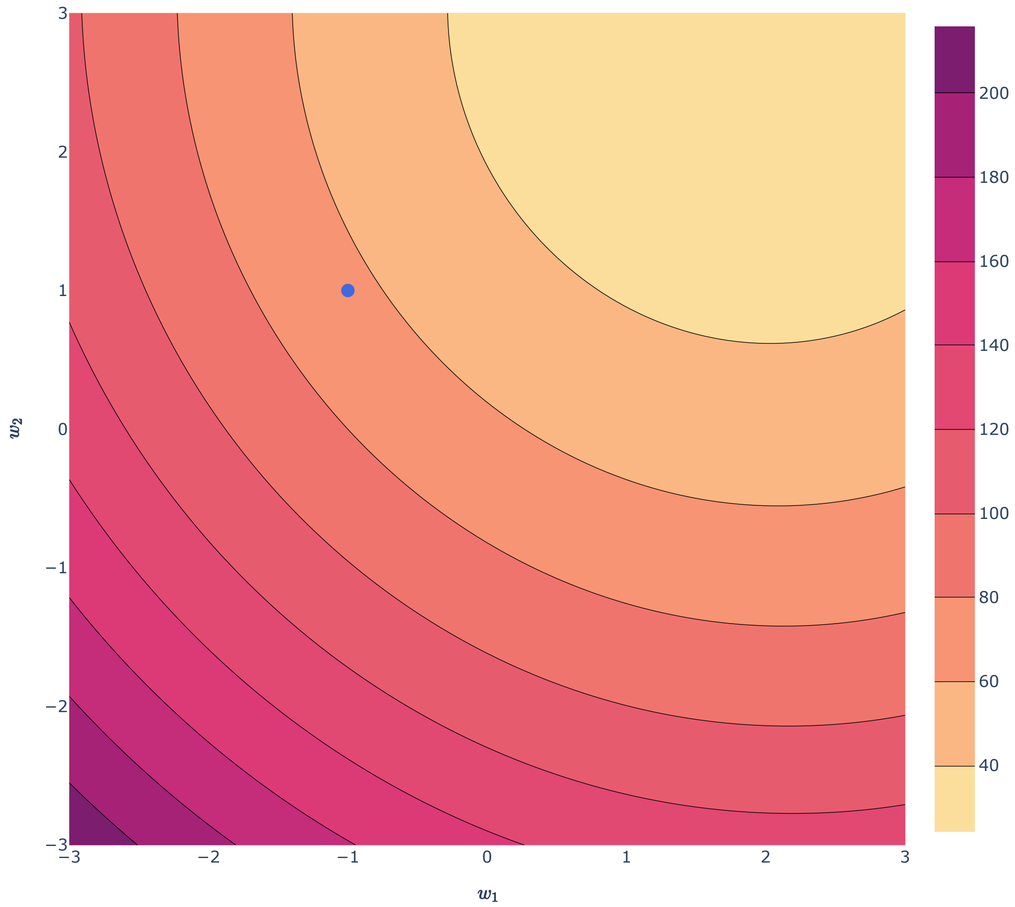


Gradient Descent

Synthetic Dataset ($T = 0$)

$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta \mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)$$

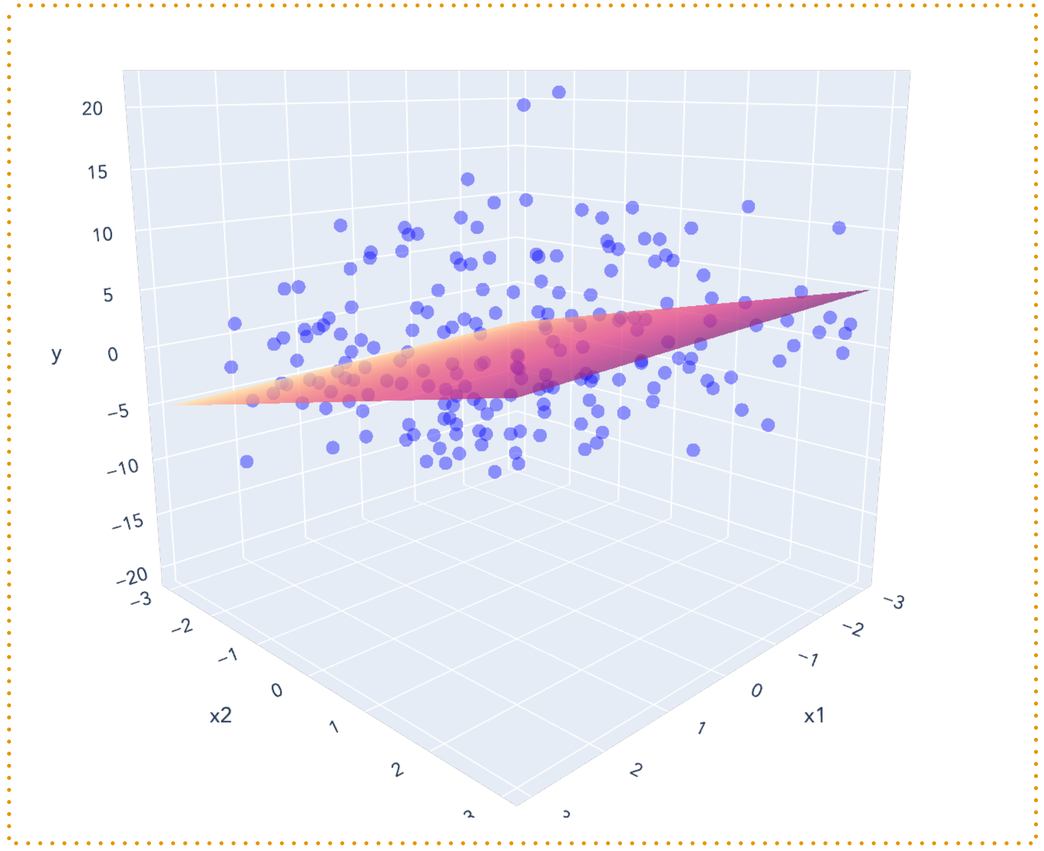


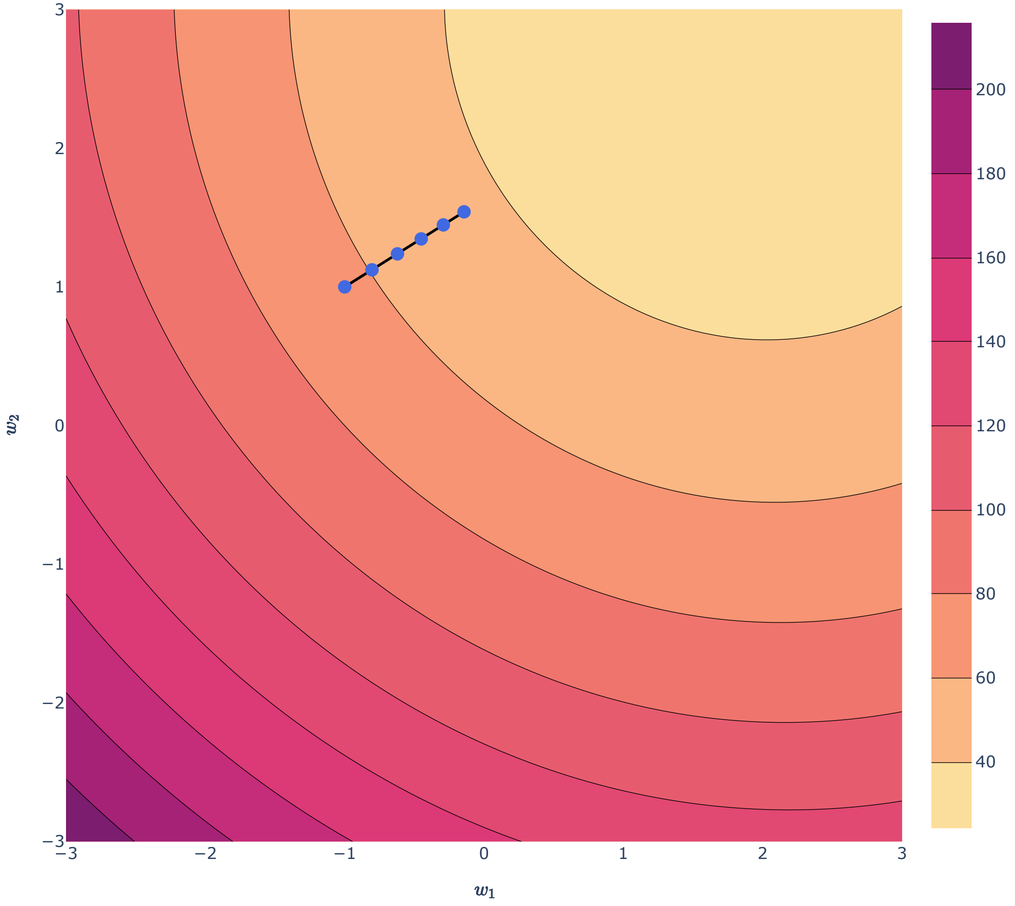


Gradient Descent

Synthetic Dataset ($T = 5$)

$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta \mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)$$

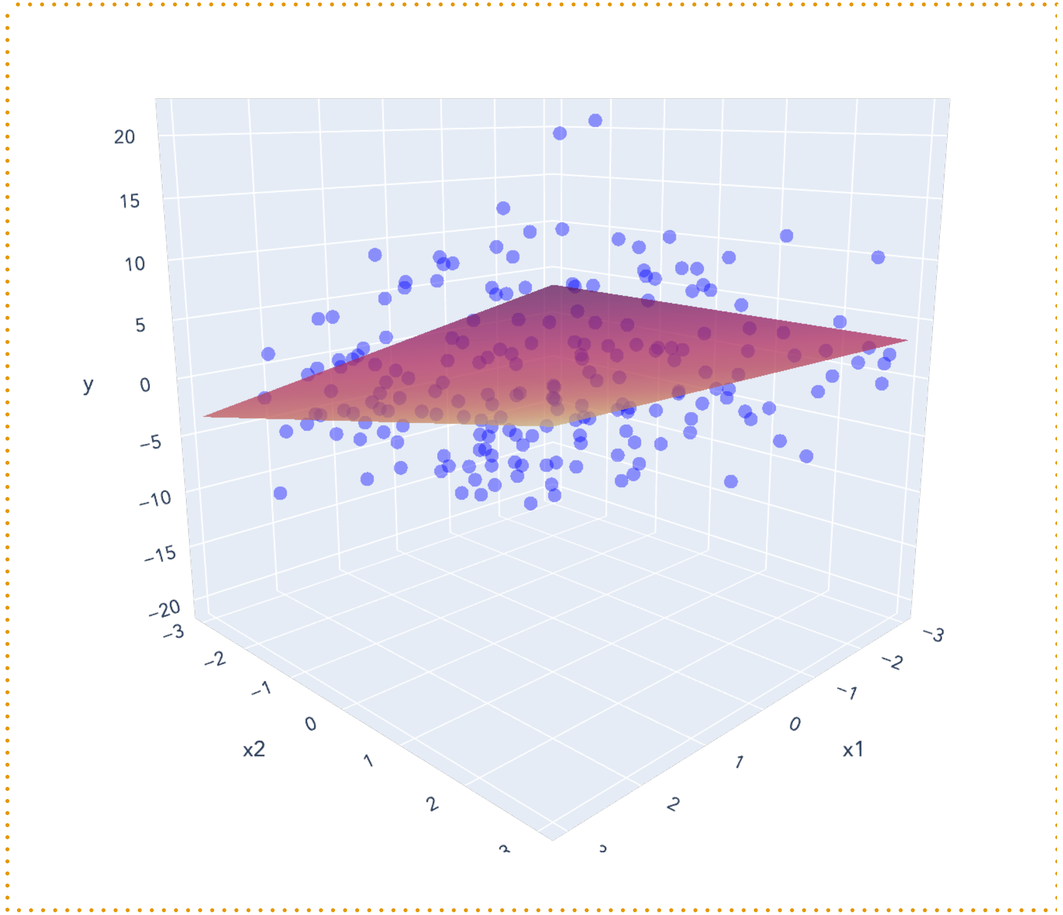


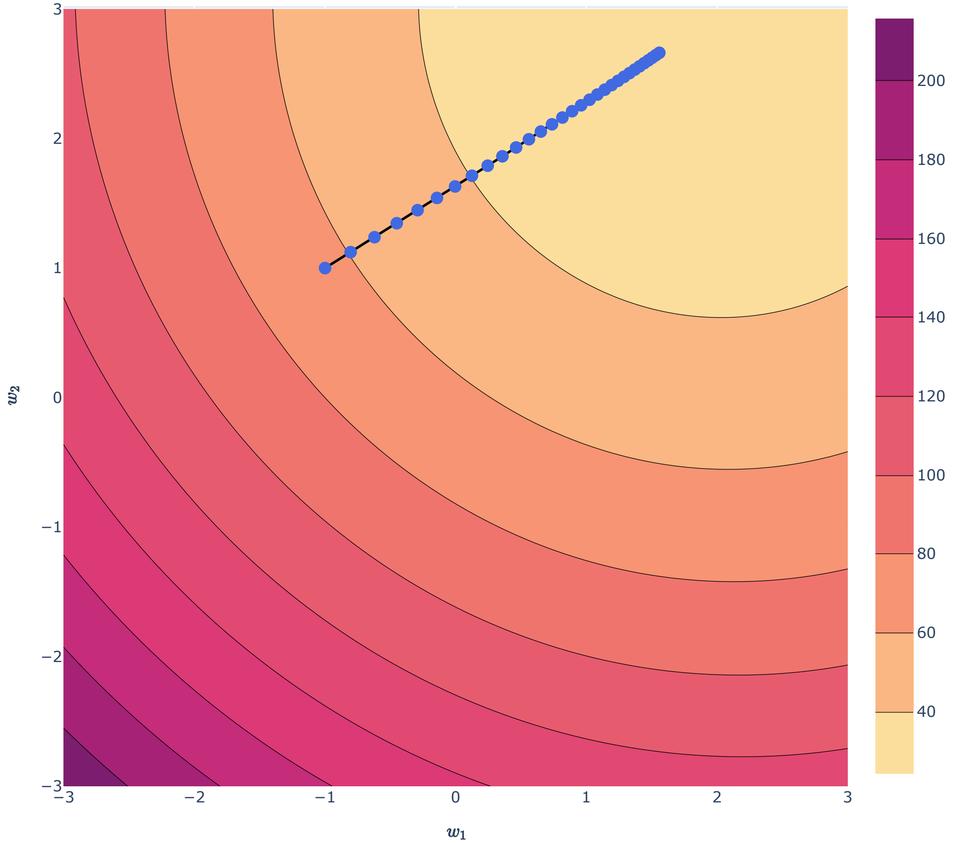


Gradient Descent

Synthetic Dataset ($T = 30$)

$$\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta \mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)$$





Solving OLS iteratively vs. analytically

Why use GD instead of the normal equations?

Solving the normal equations directly $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ takes

$$O(d^2 n + d^3) \text{ operations.}$$

Running gradient descent $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - 2\eta \mathbf{X}^\top \left(\mathbf{X}\mathbf{w} - \mathbf{y} \right)$ for T steps takes

$$O(Tdn) \text{ operations.}$$

Recap

Lesson Overview

Convexity. A property of *sets* and *functions* that affords us a lot of nice "linearity-like" properties.

Convex set. A convex is a set that has no holes. The line segment between any two points lies completely in the set.

Convex function. A function that is bowl-shaped. Between any two points, the line segment is above the function.

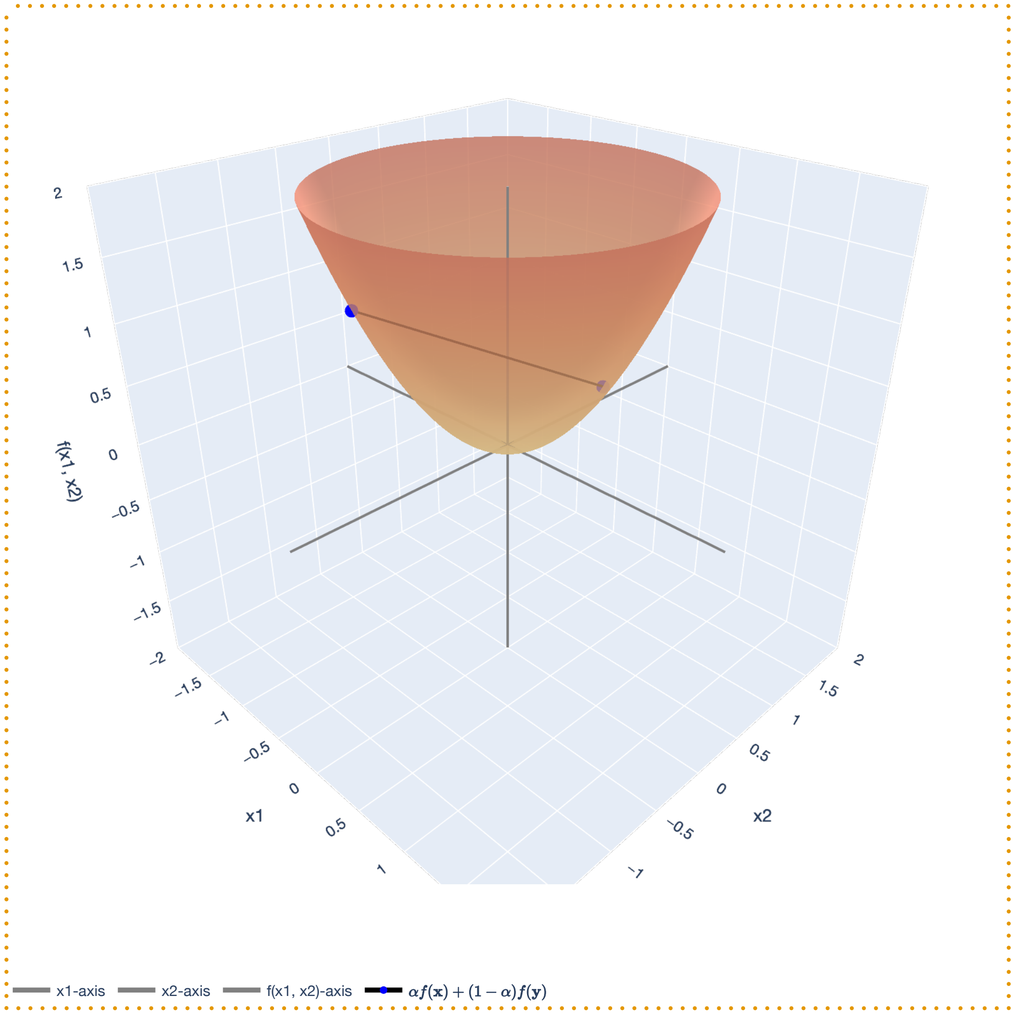
Convex optimization. Optimization problems with convex objectives and convex constraint sets. All local optima are global optima.

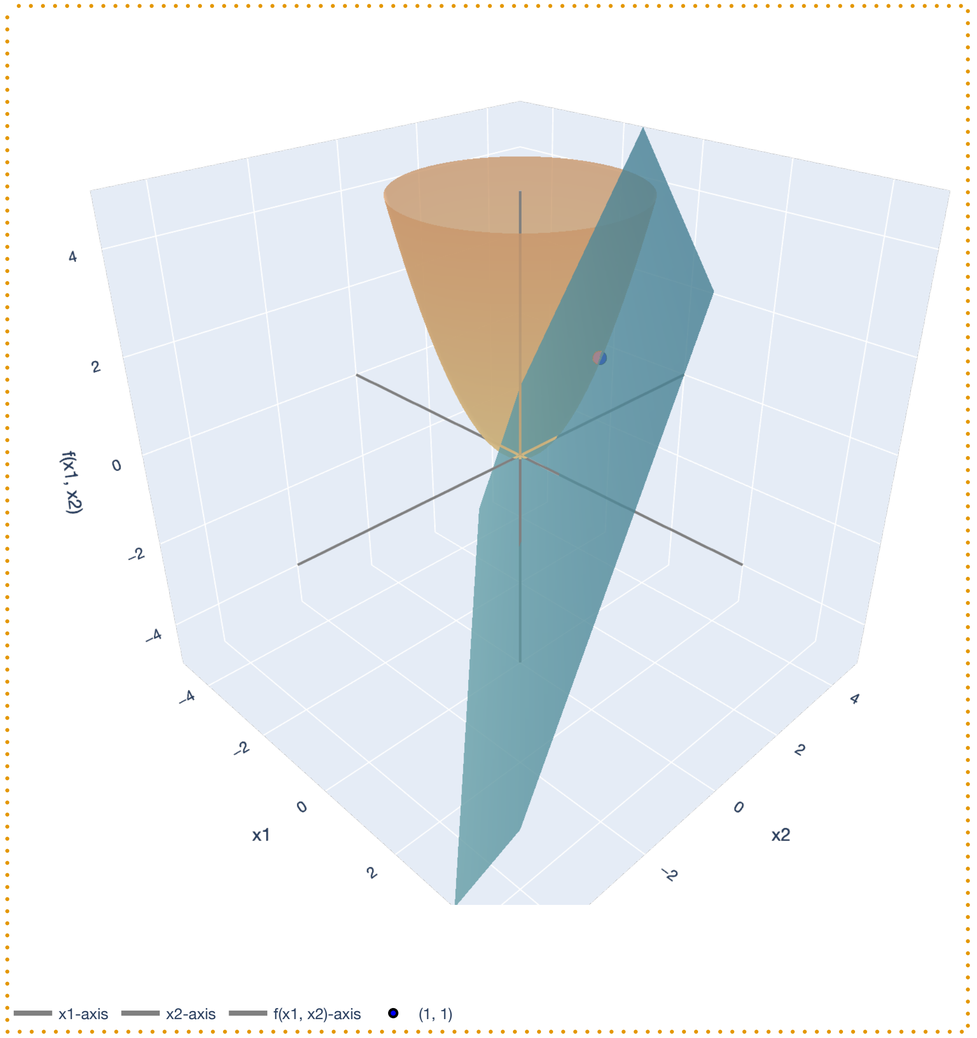
Gradient descent for convex problems. GD on convex functions is *guaranteed* to find a global min.

Gradient descent for OLS. We unite the two stories of this class and analyze GD applied to OLS!

Lesson Overview

Big Picture: Least Squares





Lesson Overview

Big Picture: Gradient Descent