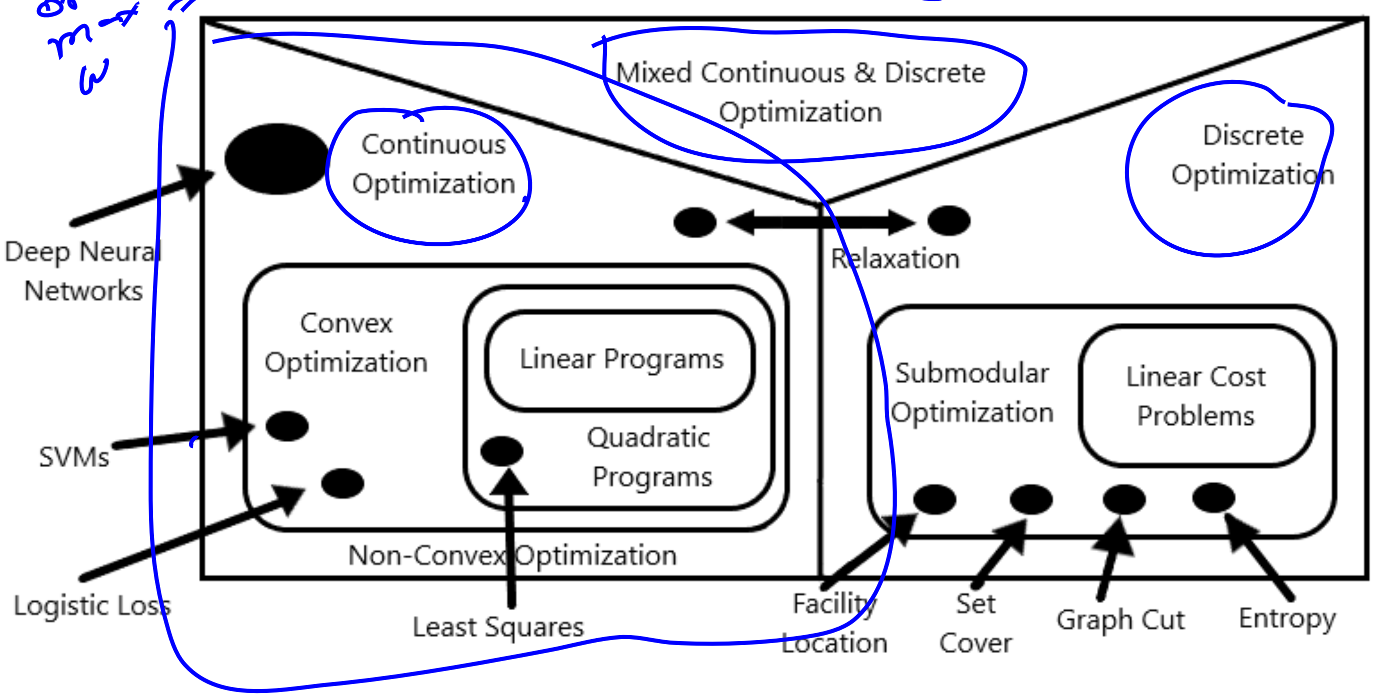
Numerical Optimization for Data Science

Dr. Phani Motamarri
Assistant Professor
Department of Computational and Data Sciences
Indian Institute of Science, Bangalore


Outline

Types of Optimization Problems

$\min_{w}$ or $\max_{w}$ $f(w)$ $w \in \mathbb{R}$ $w_1, w_2 \ldots w_n$ $w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n$



Example 1: Linear Least Squares Problem

The problem is to fit a straight line $\boxed{g(x) = w_1 + w_2 x}$ to a training set with observations $(x_1, x_2, \dots x_n)$ and corresponding estimated responses $(y_1, y_2, \dots y_n)$ using least squares.

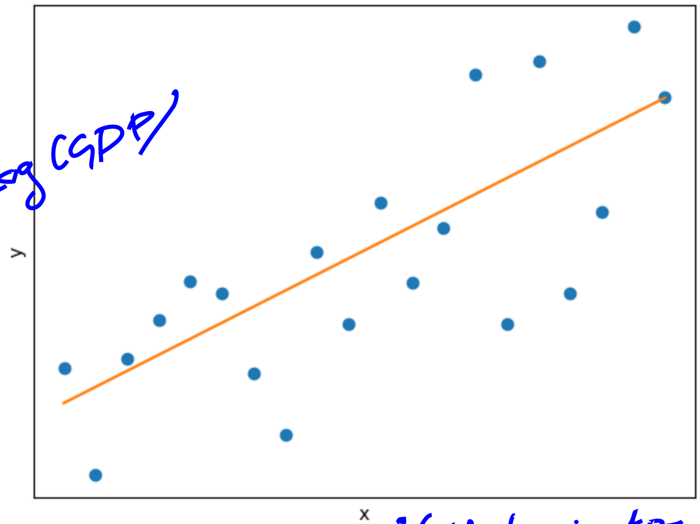
The objective function (or the loss function) to be minimized is:

$\|\varepsilon\|$

$$\boxed{f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(g(x_i) - y_i)^2}$$

$$f(w_1, w_2) = \frac{1}{2}\sum_{i=1}^{n}(w_1 + w_2 x_i - y_i)^2$$

log(GDP)



% urbanization

Figure: Least Squares Problem

$$\boxed{X^T X w = X^T y}$$

Example 2: Nonlinear Least Squares Problem

(w, b) $\quad g(x) = w_1 + w_2 x$

The problem is to fit $g(x) = \frac{1}{1+e^{-(wx+b)}}$ to a training set with observations $(x_1, x_2, \ldots x_n)$ and corresponding estimated responses $(y_1, y_2, \ldots y_n)$ using least squares.

The objective function (or the loss function) to be minimized is:

$$f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} (g(x_i) - y_i)^2$$

$$f(w, b) = \frac{1}{2}\sum_{i=1}^{n} \left(\frac{1}{1 + e^{-(wx_i+b)}} - y_i\right)^2$$

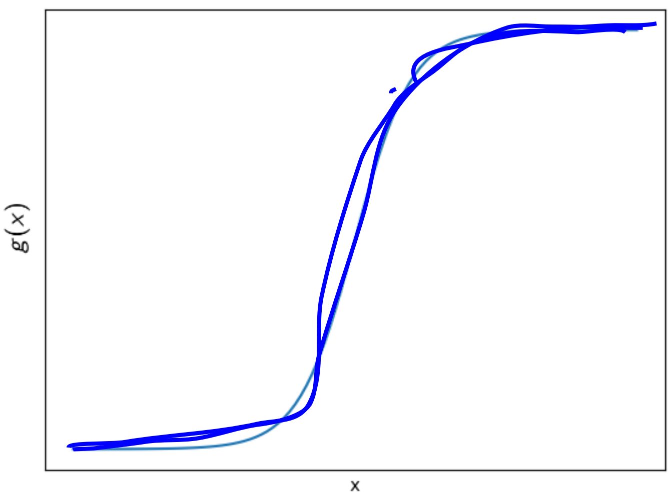


Figure: Sigmoid Function

Example 3: Transportation Problem

Objective function
to be minimized:

$$\boxed{f(\mathbf{x}) = \sum_{ij} c_{ij} x_{ij}}$$

subject to constraints $\sum_{ij} c_{ij} x_{ij}$

$$\sum_{j=1}^{12} x_{ij} \le a_i, \quad i = 1, 2$$

i F → j

$$\sum_{i=1}^{2} x_{ij} \ge b_j, \quad j = 1, ..., 12$$

$$\boxed{x_{ij} \ge 0}, \quad i = 1, 2 \quad j = 1, ..., 12$$

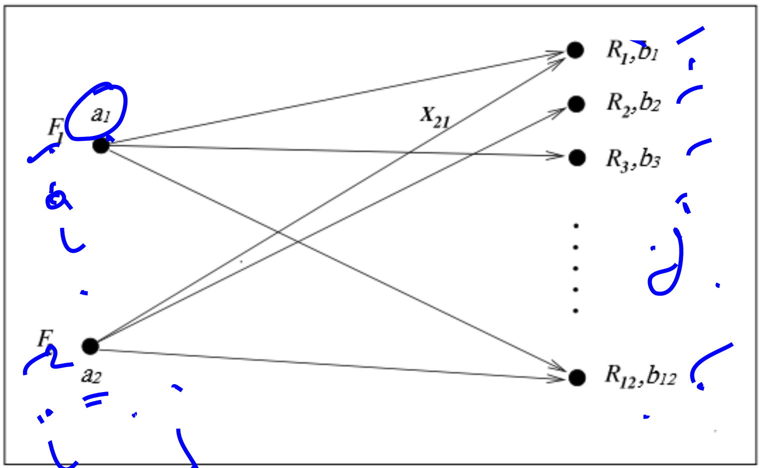


Figure: Transportation Problem

$F_i \to Factories, \quad R_i \to Retailers$

$a_i \to Capacity, \quad b_j \to Demand$

$c_{ij} \to Cost, \quad x_{ij} \to Product\ Shipped$

unit

Characterizing the solution to an Optimization Problem

- Global Minimizer
- Local Minimizer
- Strictly Local Minimizer

Global Minimizer

Global minimizer **of f:**

- A point $\mathbf{x}^*$ is a *global minimizer* if $\boxed{f(\mathbf{x}^*) < f(\mathbf{x})}$ for all $\mathbf{x}$ where $\mathbf{x}$ ranges over all of $\mathbb{R}^n$ or over the domain of interest.

$f(w,b)$

x^*

w, b

$\mathbb{R}$

x

$f(x^*)$

$f(x)$

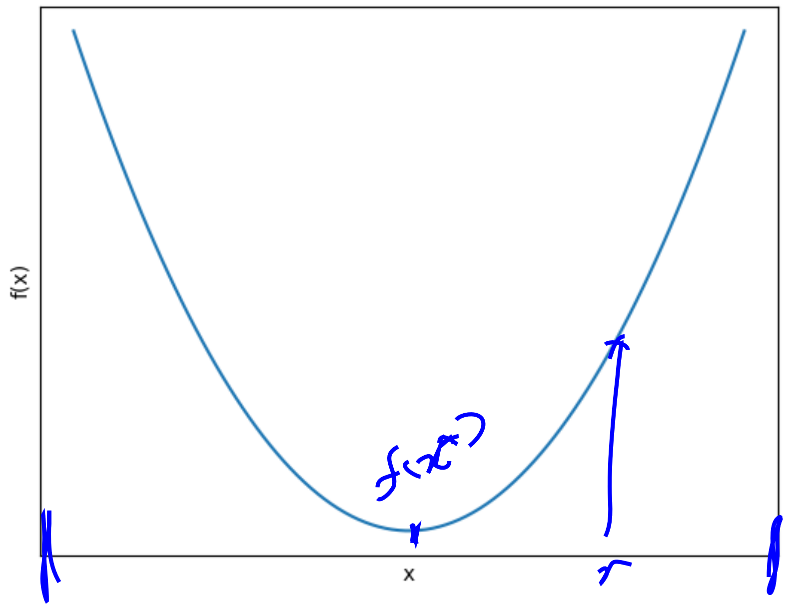


Figure: Global minimum

Local Minimizer

Local minimizer of f:

- A point $\mathbf{x}^*$ is a *local minimizer* (also called a *weak local minimizer*) if there is a region N of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in N$

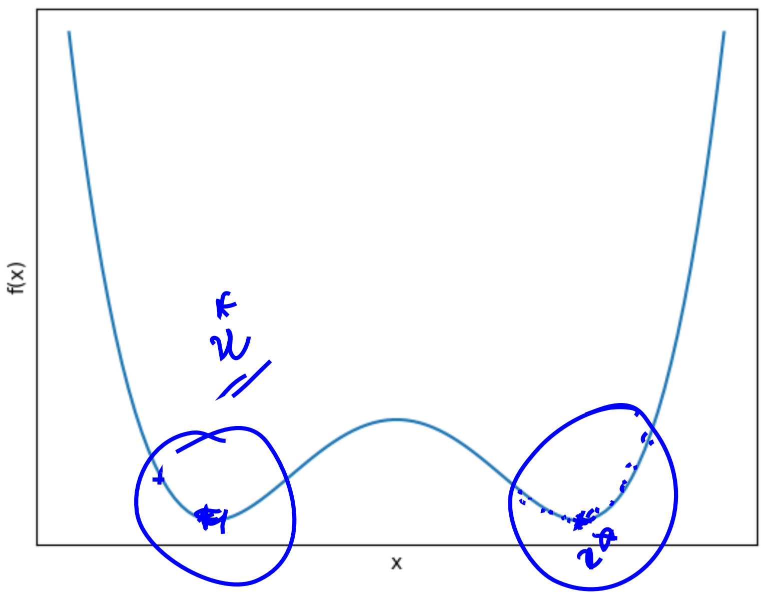


Figure: Local minima

$f(x^*) \leq f(x)$
$x \in N$

Strict Local Minimizer

Strict Local minimizer of f:

- A point $\mathbf{x}^*$ is a *strict local minimizer* (also called a *strong local minimizer*) if there is a region N of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in N$.

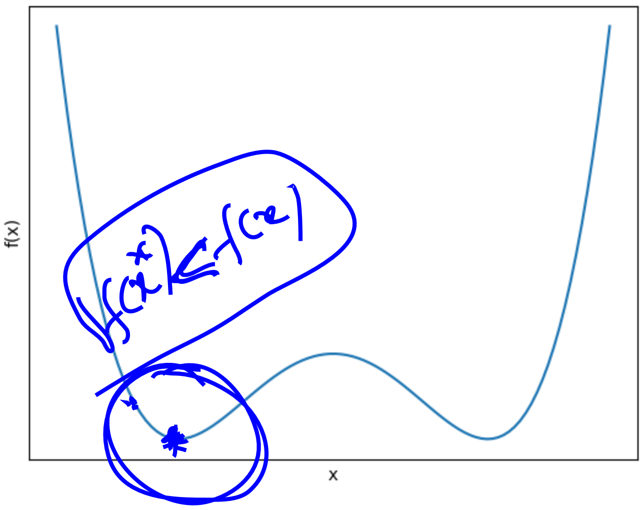


Figure: Strong Local minima

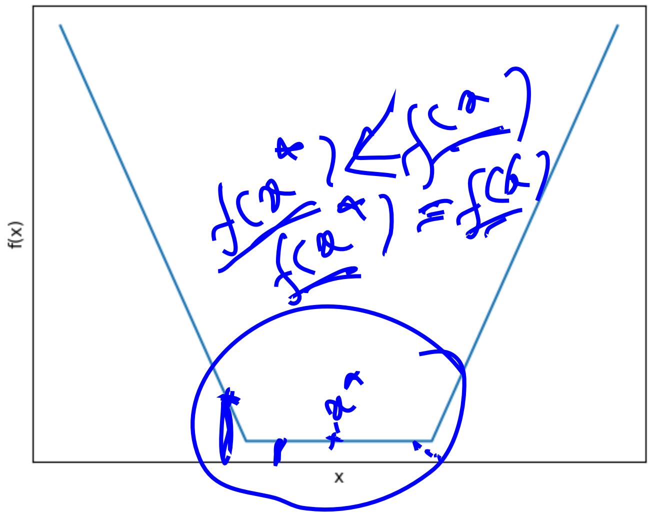
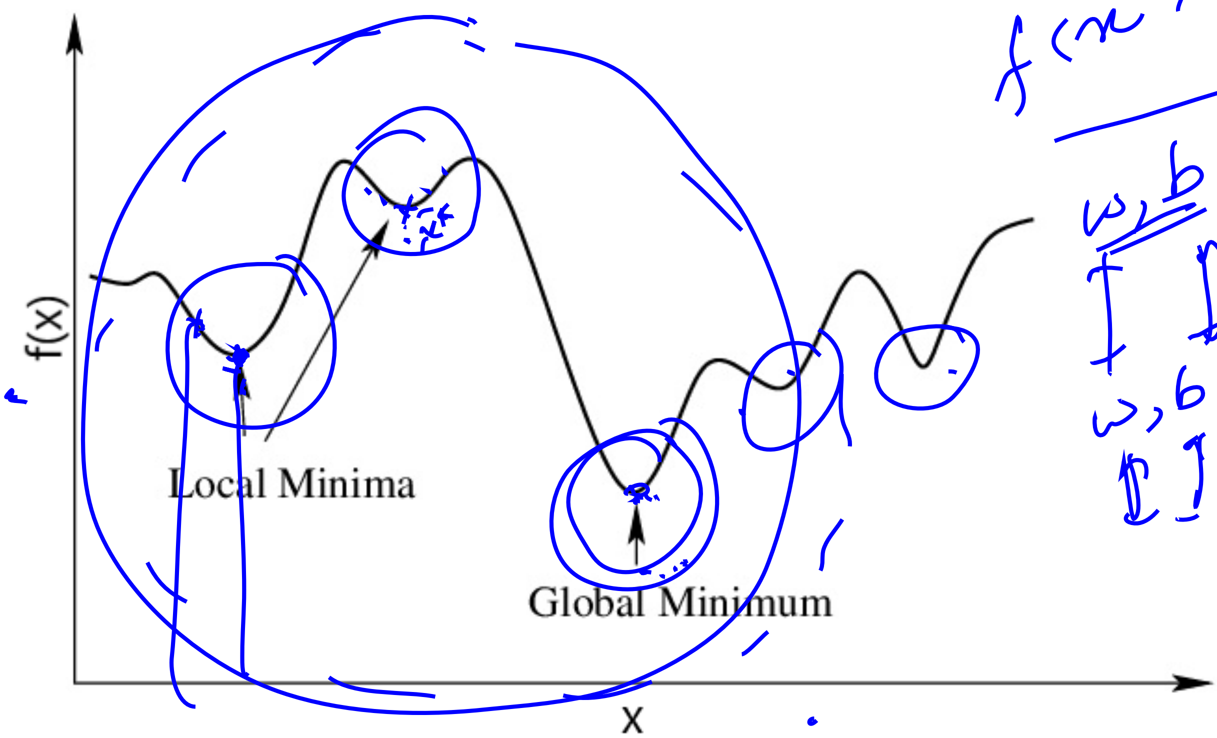


Figure: Weak Local minima

Characterizing the solution to an optimization problem



$f(x^*) \leq f(x)$

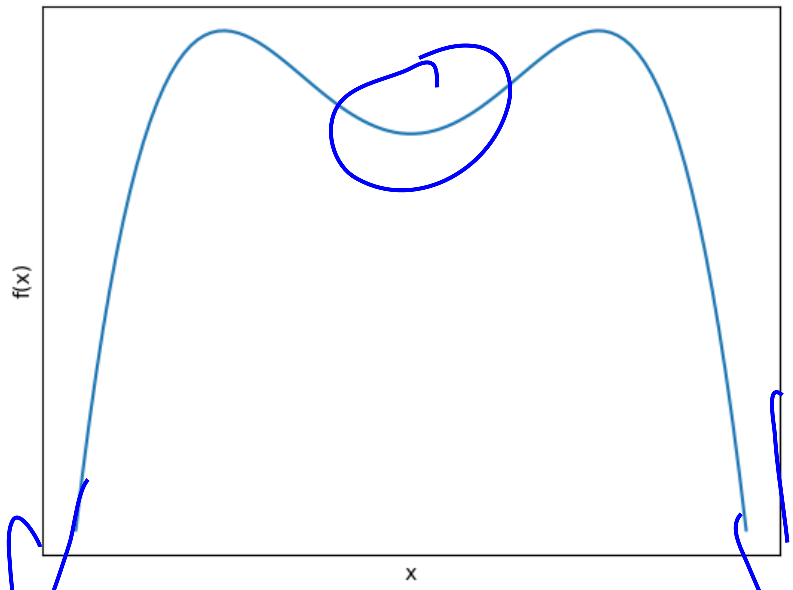
w, b

$\in \mathbb{R}$

Figure: Global and Local minima

Global and Local Minimum

- Every global minimum is also a local minimum
- It may not always be possible to find the global minimum by finding all local minima.



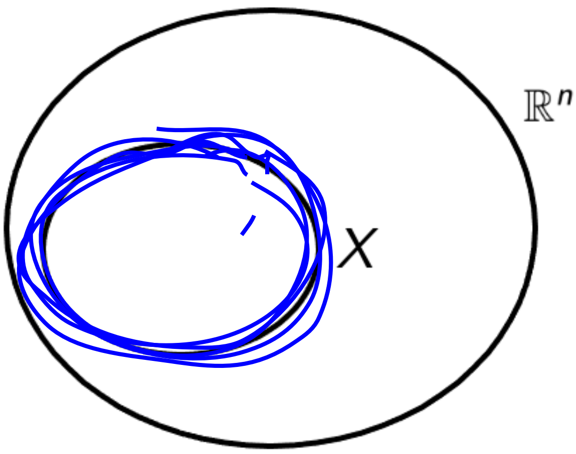
- f does not have a global minimum but has a local minimum

Optimization Problems

Optimization Problems:

Let $X \subseteq \mathbb{R}^n$ and $f : X \to \mathbb{R}$

- Constrained Optimization Problem : minimize $f(\mathbf{x})$ with respect to $\mathbf{x}$ such that $\mathbf{x} \in X$
- Unconstrained Optimization Problem : minimize $f(\mathbf{x})$ with respect to $\mathbf{x}$ such that $\mathbf{x} \in \mathbb{R}^n$



$f(x)$

Unconstrained Optimization

Unconstrained Optimization:

- Necessary Conditions - Conditions which are satisfied by every local minimum
- Sufficient Conditions - Conditions which guarantee a local minimum

Optimality Conditions:

- First-Order Necessary Conditions
- Second-Order Necessary Conditions
- Second-Order Sufficient Conditions
- Sufficient Optimality Conditions

First-Order Necessary Conditions

x^*

First-Order Necessary Conditions:

If $\mathbf{x}^*$ is a local minimizer and f is continuously differentiable in a region of $\mathbf{x}^*$ then $\nabla f(\mathbf{x}^*) = 0$

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

- In one-dimensional case, the condition is $f'(x^*) = 0$
- Here $f'(x^*) = \left.\dfrac{df}{dx}\right|_{x=x^*} = 0$
- And $\nabla f(\mathbf{x}^*) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}^*) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}^*) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}^*) \end{bmatrix} = 0$

$\frac{\partial f}{\partial x_1}$ $\quad x^*$ $\quad x^* \in \mathbb{R}^n$

$f'(x^*) = 0$

$n \times 1$

$\min (y_i - f(x_i))^2$

$\|\nabla f(x^*)\|_2$ $\quad x^*$

$A \Rightarrow B$

$f'(x^*) = 0$

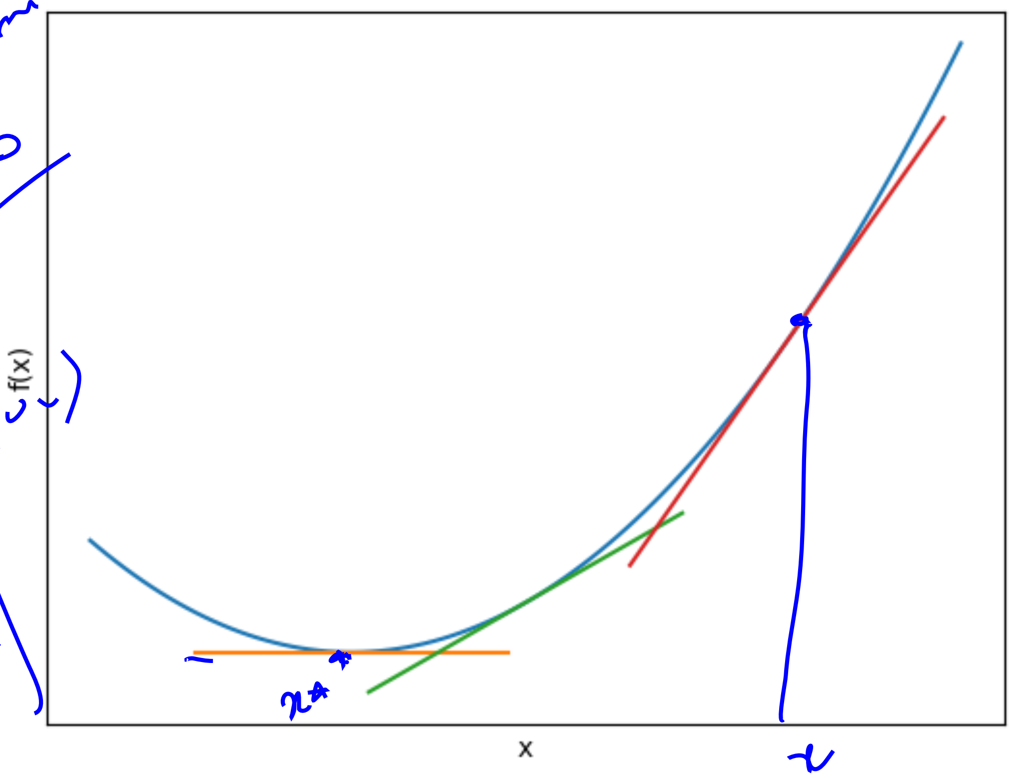
First-Order Necessary Conditions

$x^* \rightarrow$ local min

$f'(x^*) = 0$

$f(w_0, w_1, \ldots)$

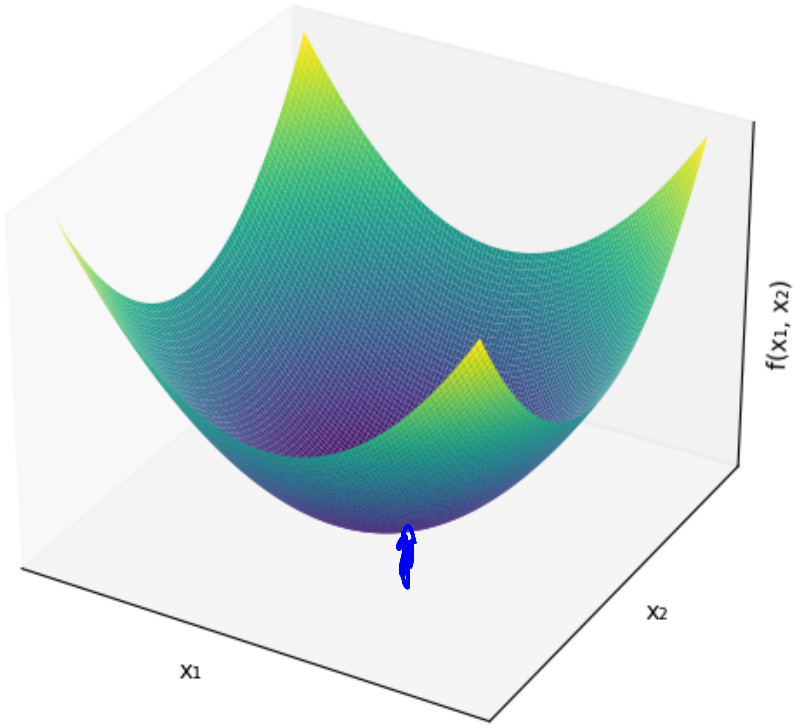
$x = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \end{pmatrix}$



$x = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}$

Figure: Tangents at different points on a curve

First-Order Necessary Conditions



$f(w, b)$

$x_1 \to w$

$x_2 \to b$

$\left.\frac{\partial f}{\partial x_1}\right|_{x_1^*} = 0$

$\left.\frac{\partial f}{\partial x_2}\right|_{x_2^*} = 0$

Figure: Gradient of $f(x_1, x_2)$ is 0 at the minimum

First-Order Necessary Conditions: Example

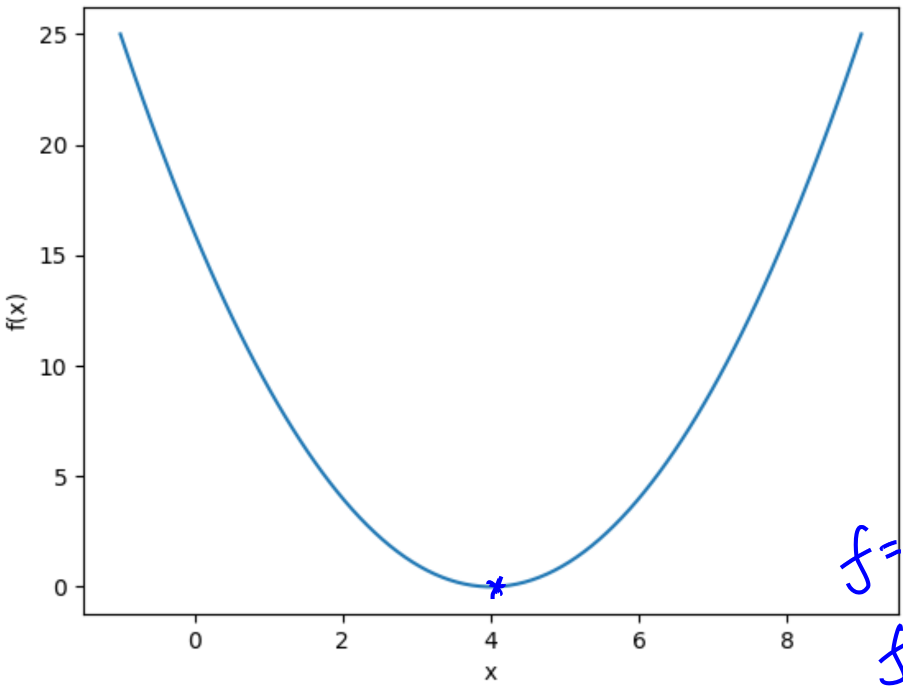


Figure: $f(x) = (x-4)^2$
$f'(4) = 0$

$x = 4$

$f = (x-4)^2$

$f' = 2(x-4)$

$f'(4) = 0$

First-Order Necessary Conditions: Example

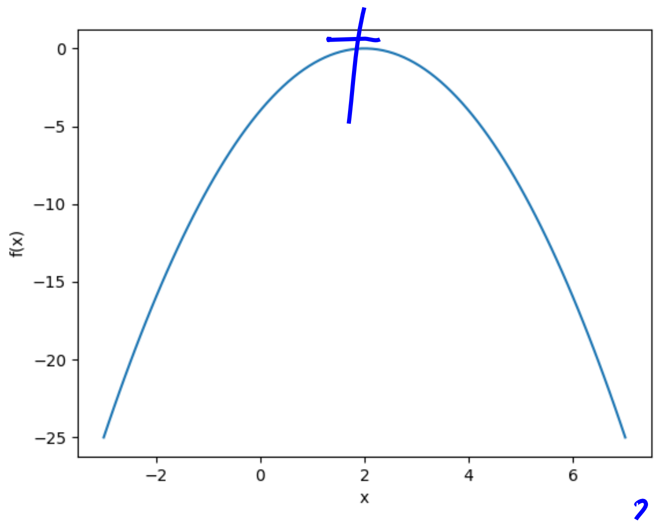


Figure: $f(x) = -(x-2)^2$
$f'(2) = 0$

$f'(x) = -2(x-2)$

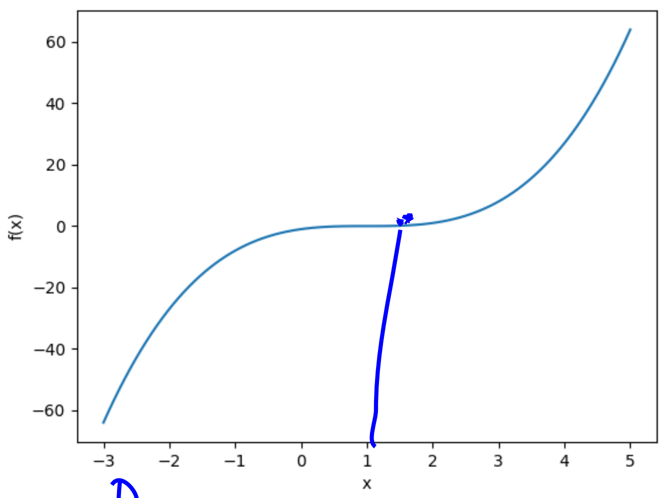


Figure: $f(x) = (x-1)^3$
$f'(1) = 0$

- Slope of a function or $f'(x)$ is zero at local minimum, local maximum and at saddle point.
- How to determine if the stationary point x for which $f'(x) = 0$, is a local minimum?

Second-Order Necessary Conditions

Second-Order Necessary Conditions:

If $\mathbf{x}^*$ is a local minimizer of f and $\nabla^2 f$ exists and is continuous in a region of $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite.

- In one-dimensional case, the conditions are $f'(x^*) = 0$ and $f''(x^*) \geq 0$
- A matrix M is positive semi-definite if $\mathbf{x}^T M \mathbf{x} \geq 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$
- Here $f''(x^*) = \left.\dfrac{d^2 f}{dx^2}\right|_{x=x^*}$
- And $M = \nabla^2 f(\mathbf{x}^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}^*) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}^*) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}^*) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}^*) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}^*) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}^*) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}^*) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}^*) \end{bmatrix}$

$x^T M x \geq 0$

$\lambda \geq 0$

Second-Order Necessary Conditions: Example

$A \Rightarrow B$
$\sim B \Rightarrow \sim A$

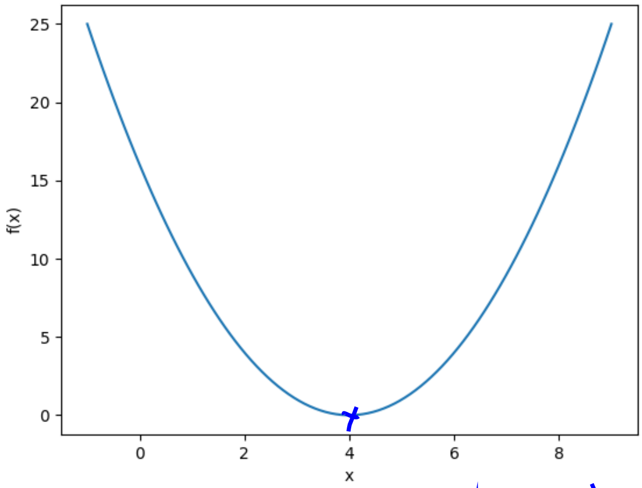


Figure: $f(x) = (x-4)^2$

$2(x-4)$
2

$f'(4) = 0$
$f''(4) = 2 \geq 0$
For a local minimum, we can verify
$\boxed{f''(x^*) \geq 0}$

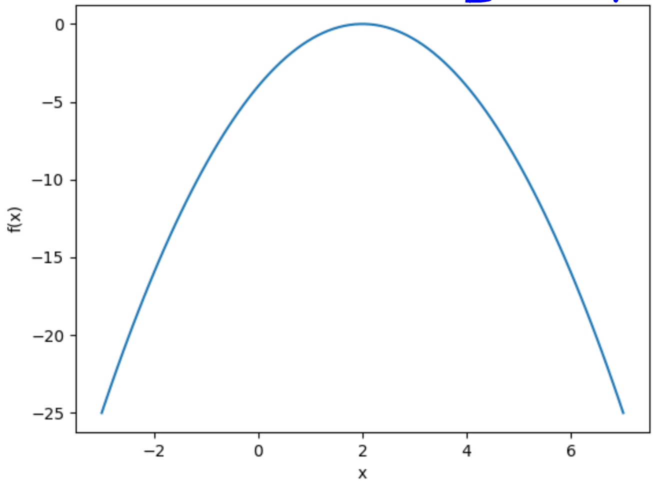


Figure: $f(x) = -(x-2)^2$
$f'(2) = 0$
$f''(2) = -2 \not\geq 0$
Since, $f''(x^*) \not\geq 0$, x^* is not a local minimum

$f(x)$ $\quad f(w, b)$

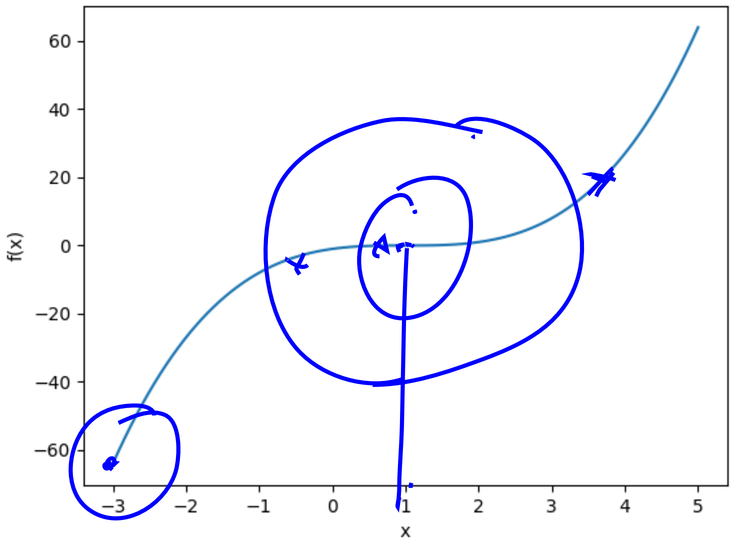
$$\nabla^2 f = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{bmatrix} \frac{\partial^2 f}{\partial w^2} & \frac{\partial^2 f}{\partial w \partial b} \\ \frac{\partial^2 f}{\partial b \partial w} & \frac{\partial^2 f}{\partial b^2} \end{bmatrix}$$

3×3

4×4

$n \times n$

Second-Order Necessary Conditions: Example



$f(x^*) < f(x)$

≥ 0

Figure: $f(x) = (x-1)^3$ $\quad f'(1) = 0 \quad f''(1) = 0 \geq 0$

- But $x^* = 1$ is not a local minimum
- The second-order necessary conditions are not sufficient.

Second-Order Sufficient Conditions

Second-Order Sufficient Conditions:

$\mathbf{x}^*$ is a strict local minimizer of f if $\nabla^2 f$ is continuous in a region of $\mathbf{x}^*$ and $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*)$ is positive definite.

- In one-dimensional case, the conditions are $f'(x^*) = 0$ and $f''(x^*) > 0$
- A matrix M is positive definite if $\mathbf{x}^T M\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$

$f'1x$

$\lambda_i \geq 0$

$\lambda_i > 0$

Note:

Second-order sufficient conditions:

- guarantee that the local minimum is strict and
- are not necessary. For example, $f(x) = (x-5)^6, x^* = 5$ is a strict local minimum but $f'(x^*) = f''(x^*) = 0$

Sufficient Optimality Conditions

Sufficient Optimality Conditions:

x^* is a local minimum if and only if the first non-zero element of the sequence $\{f^k(x^*)\}$ is positive and occurs at even positive k

- Here $f^k(x^*)$ is the k-th derivative of f at $x = x^*$
- $f^k(x^*) = \left.\dfrac{d^k f}{dx^k}\right|_{x=x^*}$

x^*

$f^{(1)}(x),\ f^{(2)}(x),\ f^{(3)}(x)$

$f^{(4)}(x),\ f^{(5)}(x)$

Sufficient Optimality Conditions: Example

$x = 0, x = 1$
$x = -1$

- Consider the function $f(x) = (x^2 - 1)^3$
- The stationary points are obtained by
  $f'(x) = 0 \implies 6x(x^2 - 1)^2 = 0 \implies f'(-1) = f'(0) = f'(1) = 0$
- Check second derivative: $f''(x) = 6(x^2 - 1)(5x^2 - 1)$ $\leftarrow 0$
  - $f''(0) = 6 > 0 \implies$ 0 is a strict local minimum
  - $f''(-1) = f''(1) = 0 \implies$ Check for higher derivatives
- Check third derivative: $f'''(x) = 24x(5x^2 - 3)$ $\leftarrow -1, 1$
  - $f'''(-1) = -48 < 0$ and $f'''(1) = 48 > 0$
  - Both -1 and 1 are saddle points as the first non-zero derivative is at odd positive k (here $k = 3$)

Sufficient Optimality Conditions: Example

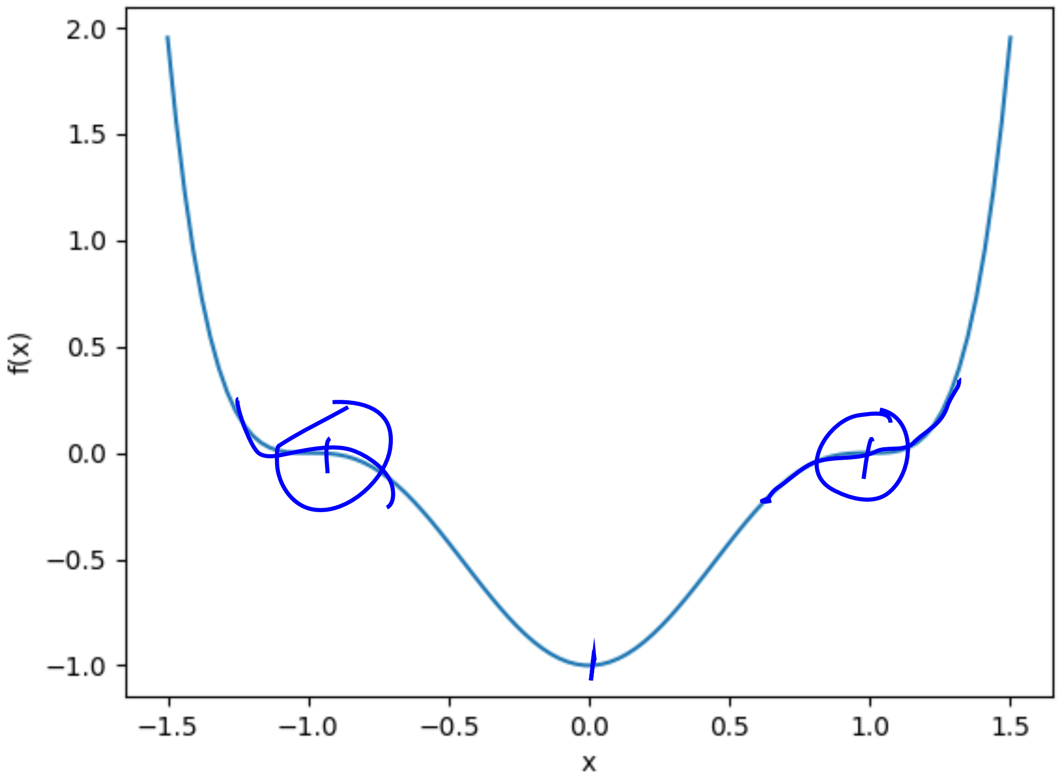


Figure: $f(x) = (x^2 - 1)^3$

Sufficient Optimality Conditions: Example

$f(x,y)$

$f(w,b)$

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix}$$

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- Consider the function $f(x_1, x_2) = x_1 exp(-x_1^2 - x_2^2)$
- $\nabla f = \begin{pmatrix} exp(-x_1^2 - x_2^2)(1 - 2x_1^2) \\ exp(-x_1^2 - x_2^2)(-2x_1 x_2) \end{pmatrix}$
- $\nabla f = 0 \implies \boxed{\mathbf{x}_1^* = (\frac{1}{\sqrt{2}}, 0)}$ and $\boxed{\mathbf{x}_2^* = (-\frac{1}{\sqrt{2}}, 0)}$
- $\nabla^2 f(\mathbf{x}_2^*) = \begin{pmatrix} 2\sqrt{2}exp(-\frac{1}{2}) & 0 \\ 0 & \sqrt{2}exp(-\frac{1}{2}) \end{pmatrix}$ is positive definite $\implies \mathbf{x}_2^*$ is a strict local minimum
- $\nabla^2 f(\mathbf{x}_1^*) = \begin{pmatrix} -2\sqrt{2}exp(-\frac{1}{2}) & 0 \\ 0 & -\sqrt{2}exp(-\frac{1}{2}) \end{pmatrix}$ is negative definite $\implies \mathbf{x}_1^*$ is a strict local maximum

Sufficient Optimality Conditions: Example

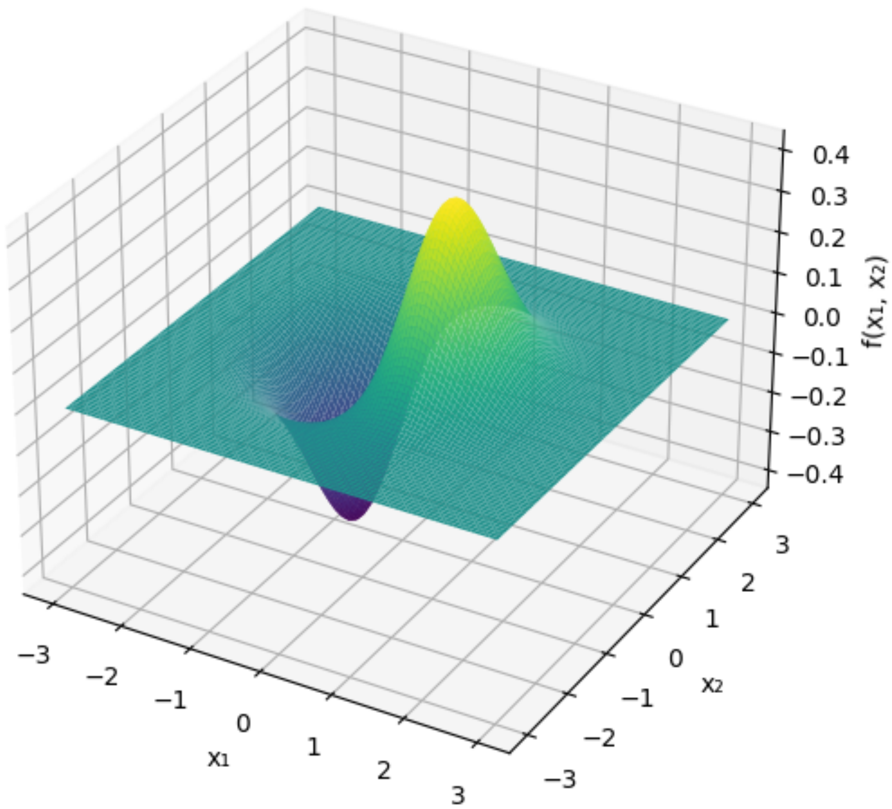


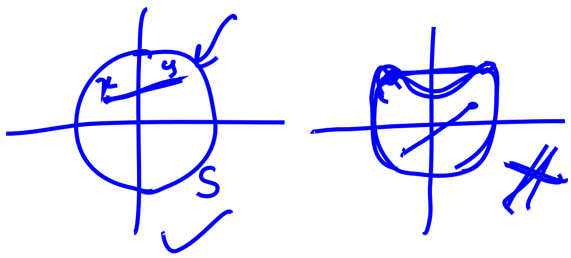
Figure: $f(x_1, x_2) = x_1 exp(-x_1^2 - x_2^2)$

Convexity

Convex Sets:

A set $S \in \mathbb{R}^n$ is a *convex set* if $\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in S$ for all $\mathbf{x} \in S$ and $\mathbf{y} \in S$ and $\alpha \in [0, 1]$

- Also means that the straight line segment connecting any two points in S lies entirely inside S.



Example :

- Unit sphere: $\{\mathbf{x} \in \mathbb{R}^3 \mid ||\mathbf{x}||_2 \leq 1\}$
- $\mathbb{R}^n$ itself is convex

Convex Functions:

A function f is a *convex function* if its domain S is a convex set and if for any two points $\mathbf{x}$ and $\mathbf{y}$ in S, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$ for all $\alpha \in [0, 1]$

Convex Functions: Example

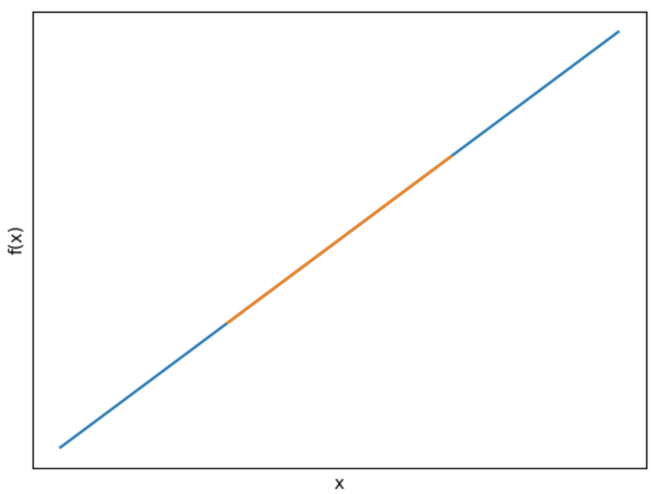


Figure: $f(x) = \alpha x + \beta$ is convex on $\mathbb{R}$ for all $\alpha, \beta \in \mathbb{R}$

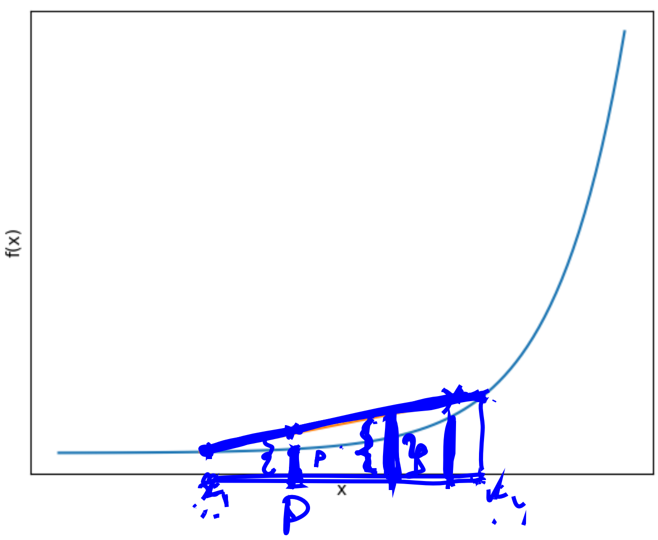


Figure: $f(x) = e^{\alpha x}$ is convex on $\mathbb{R}$ for all $\alpha \in \mathbb{R}$

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

Convex Functions: Example

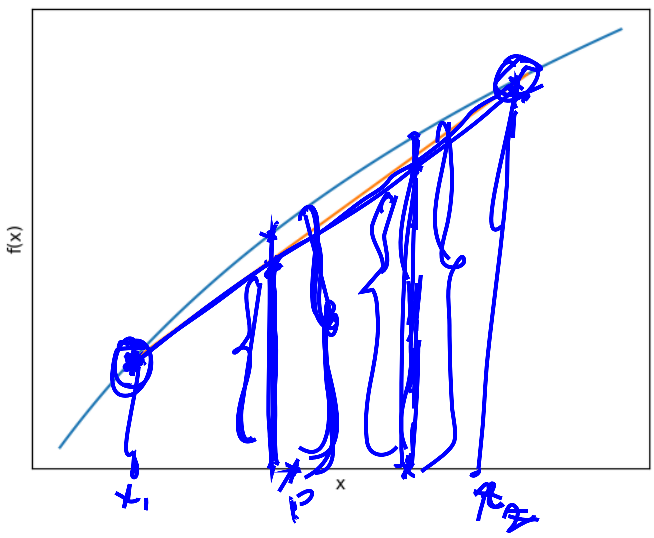


Figure: $f(x) = \log(\alpha x)$ is not convex on $\{x > 0, x \in \mathbb{R}\}$ for all $\alpha > 0$

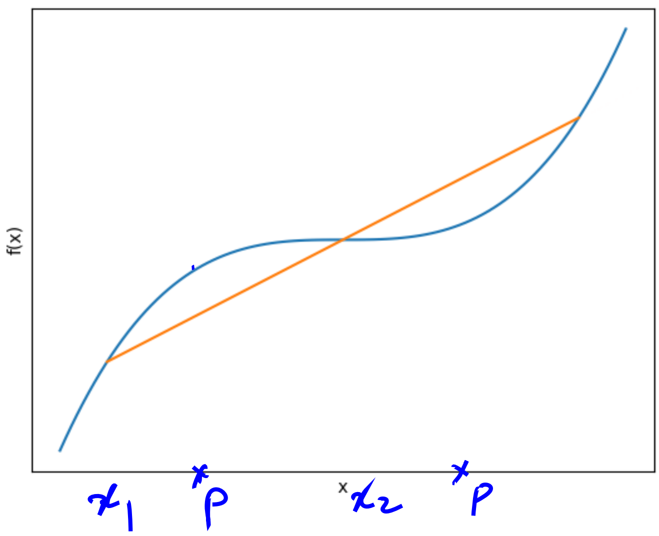


Figure: $f(x) = (x - \alpha)^3$ is not convex on $\mathbb{R}$ for all $\alpha \in \mathbb{R}$

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

Convex Functions

- Any local minimizer $\mathbf{x}^*$ is a global minimizer of f if f is convex.
- Further if f is differentiable, then any stationary point $\mathbf{x}^*$ is a global minimizer of f

$\underset{\sim}{x}$ $\underset{\sim}{y}$ $\alpha \underset{\sim}{x} + (1-\alpha)\underset{\sim}{y}$ α

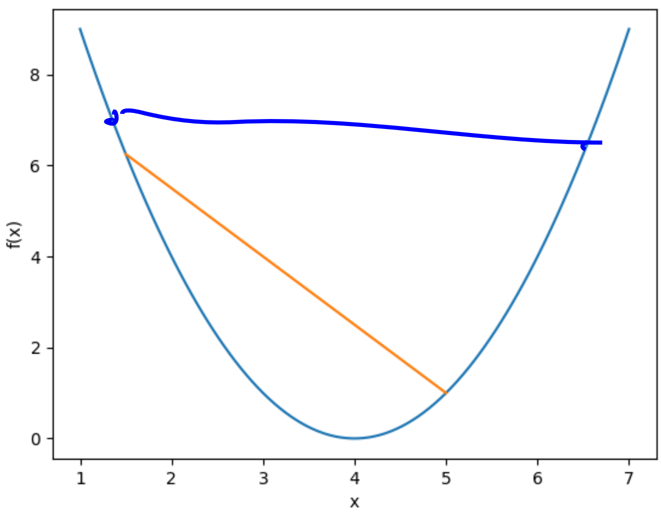


Figure: $f(x) = (x-4)^2$ is convex on $\mathbb{R}$ and $x^* = 4$ is both global and local minimizer of $f(x)$

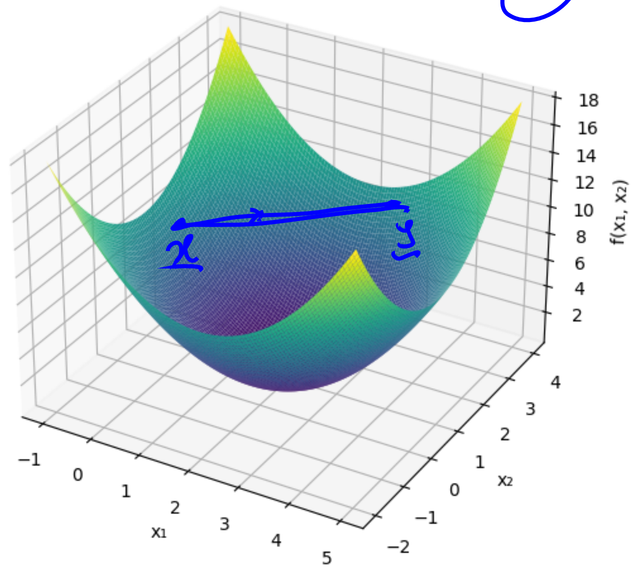


Figure: $f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 1)^2$ is convex on $\mathbb{R}^2$ and $x^* = (2, 1)$ is both global and local minimizer of $f(x_1, x_2)$

Unconstrained Minimization

Unconstrained Minimization Algorithm:

1. Initialize $\mathbf{x}_0$ and $i = 0$
2. If stopping condition is not satisfied, then continue, else stop
   1. Calculate $\mathbf{x}_{i+1}$ so that $f(\mathbf{x}_{i+1}) < f(\mathbf{x}_i)$
   2. Update i with $i + 1$
3. Final $\mathbf{x}_i$ is the local minimum $\mathbf{x}^*$ of $f(\mathbf{x})$

Following questions need to be answered:

- What stopping condition can be used?
- What is the speed of convergence?
- How to calculate $\mathbf{x}_{i+1}$?

Unconstrained Minimization

Stopping Conditions:

- $||\nabla f(\mathbf{x}_i)|| \leq \epsilon$
- $\frac{f(\mathbf{x}_i)-f(\mathbf{x}_{i+1})}{|f(\mathbf{x}_i)|} \leq \epsilon$

Speed of Convergence:

The sequence $\{\mathbf{x}_i\}$ converges to $\mathbf{x}^*$ with order p and convergence rate β if

$$\lim_{i \to \infty} \frac{||\mathbf{x}_{i+1} - \mathbf{x}^*||}{||\mathbf{x}_i - \mathbf{x}^*||^p} = \beta, \qquad \beta \in \mathbb{R}$$

- Convergence is faster for higher p
- Linear Convergence: $p = 1, 0 < \beta < 1$
- Quadratic Convergence: $p = 2, \beta > 0$

Unconstrained Minimization

Calculating $\mathbf{x}_{i+1}$ for the Unconstrained Minimization Algorithm:

- First Order Methods
  - Gradient Descent
  - Stochastic Gradient Descent
  - Mini-Batched Gradient Descent
  - Stochastic Average Gradient
  - Optimizers - AdaGrad, RMSProp & Adam
- Second Order Methods
  - Newton's Method
  - Quasi-Newton Method

Nonlinear Least Squares Problem

The
problem is to fit $g(x) = \frac{1}{1+e^{-(wx+b)}}$
to a training set with
observations $(x_1, x_2, \ldots x_n)$ and
corresponding estimated responses
$(y_1, y_2, \ldots y_n)$ using least squares.

The objective function (or the
loss function) to be minimized is:

$$f(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(g(x_i) - y_i)^2$$

$$f(w, b) = \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{1 + e^{-(wx_i+b)}} - y_i\right)^2$$

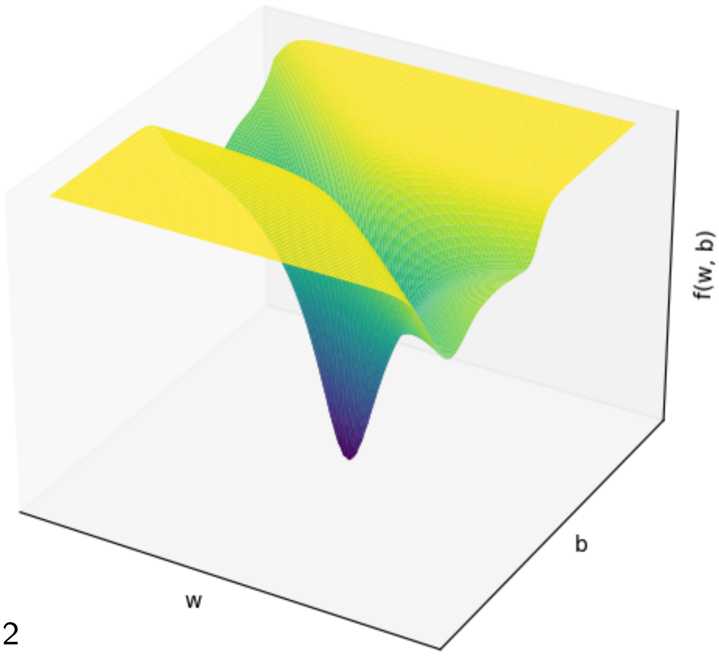


Figure: Surface plot of Loss Function $f(w, b)$

Gradient Descent

Descent Direction:

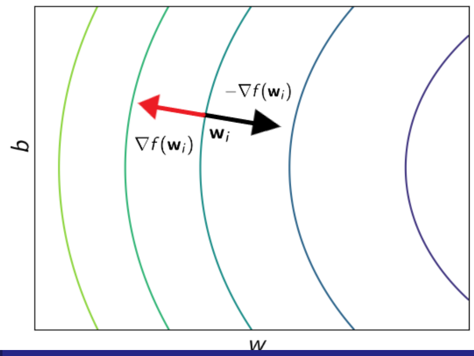
$\mathbf{d} \in \mathbb{R}^n$ is a *descent direction* of $f(\mathbf{w})$ at $\mathbf{w}^* \in \mathbb{R}^n$ if $f(\mathbf{w}^* + \eta\mathbf{d}) < f(\mathbf{w}^*)$ for all $\eta \in (0, \delta)$, $\delta > 0$

- In terms of minimization, $f(\mathbf{w}_{i+1}) < f(\mathbf{w}_i)$, where $\mathbf{w}_{i+1} = \mathbf{w}_i + \eta_i\mathbf{d}_i$
- $\eta_i =$ minimize $f(\mathbf{w}_i + \eta\mathbf{d}_i)$, $\eta > 0$
- $\eta_i =$ minimize $h(\eta)$, $\eta > 0$
- $\mathbf{d}_i = -\nabla f(\mathbf{w}_i)$ in gradient descent algorithm

Gradient Descent

Gradient Descent Algorithm:

1. Initialize $\mathbf{w}^0$ and $k = 0$
2. If stopping condition like $||\nabla f(\mathbf{w}^k)|| \leq \epsilon$ is satisfied, then stop, else continue
   1. $\mathbf{d}^k = -\nabla f(\mathbf{w}^k)$
   2. Calculate η^k along $\mathbf{d}^k$ so that $f(\mathbf{w}^k + \eta^k \mathbf{d}^k) < f(\mathbf{w}^k)$
   3. $\mathbf{w}^{k+1} = \mathbf{w}^k + \eta^k \mathbf{d}^k$
   4. Update k with $k + 1$
3. Final $\mathbf{w}^k$ is the local minimum $\mathbf{w}^*$ of $f(\mathbf{w})$



Gradient Descent

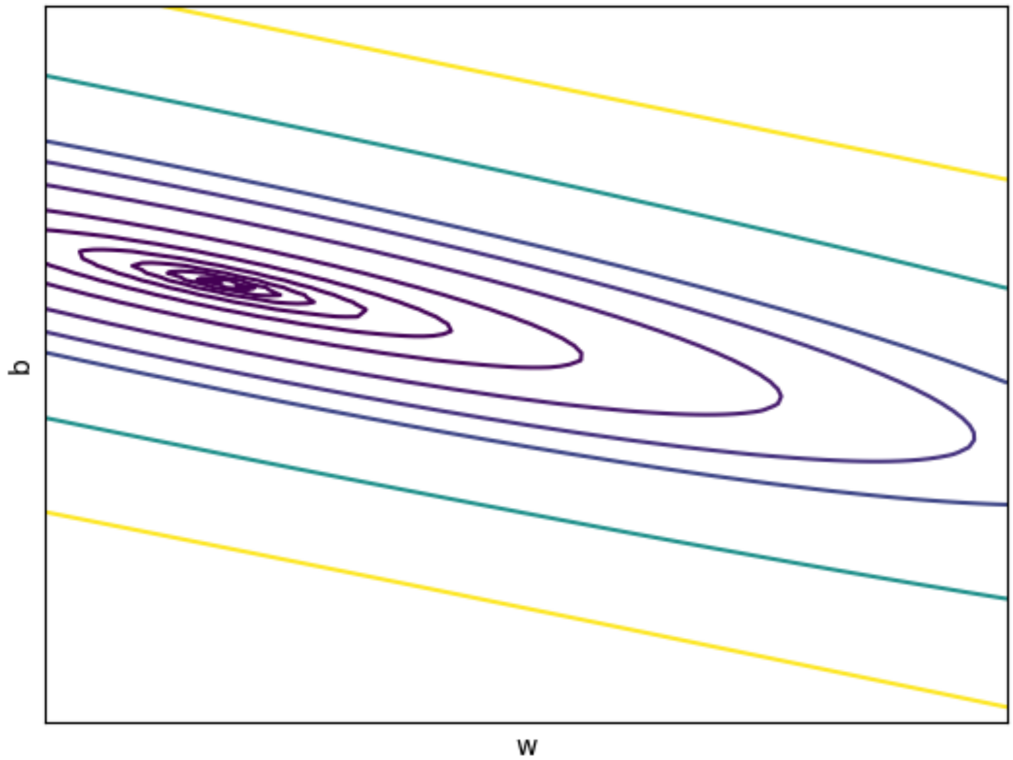


Figure: Contour plot of Loss function $f(w, b)$