MATHS

Optimization in Deep Learning: Concepts & Convexity

TEAM MEMBER

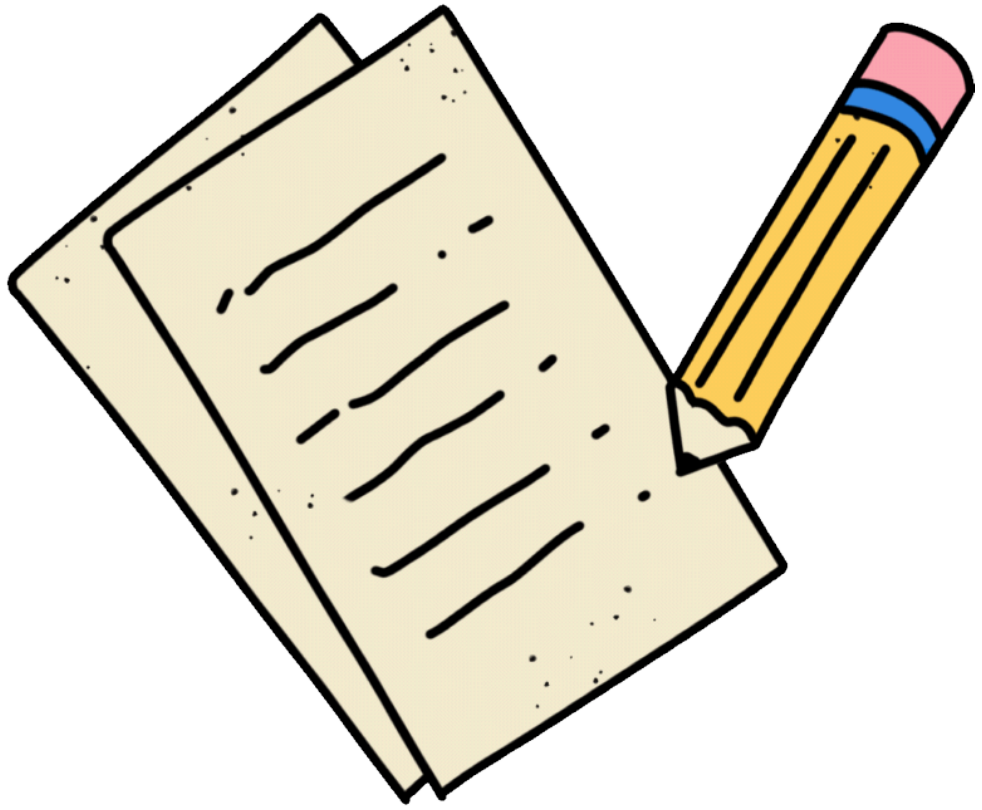
- Nguyễn Minh Tuấn - 2470574
- Nguyễn Trần Phước - 2470576
- Lê Trí Quyền - 2470740
- Lê Quang Trung - 2470746
- Trần Hoàng Nguyên - 2470739

THINK

PAIR

SHARE

TABLE CONTENT

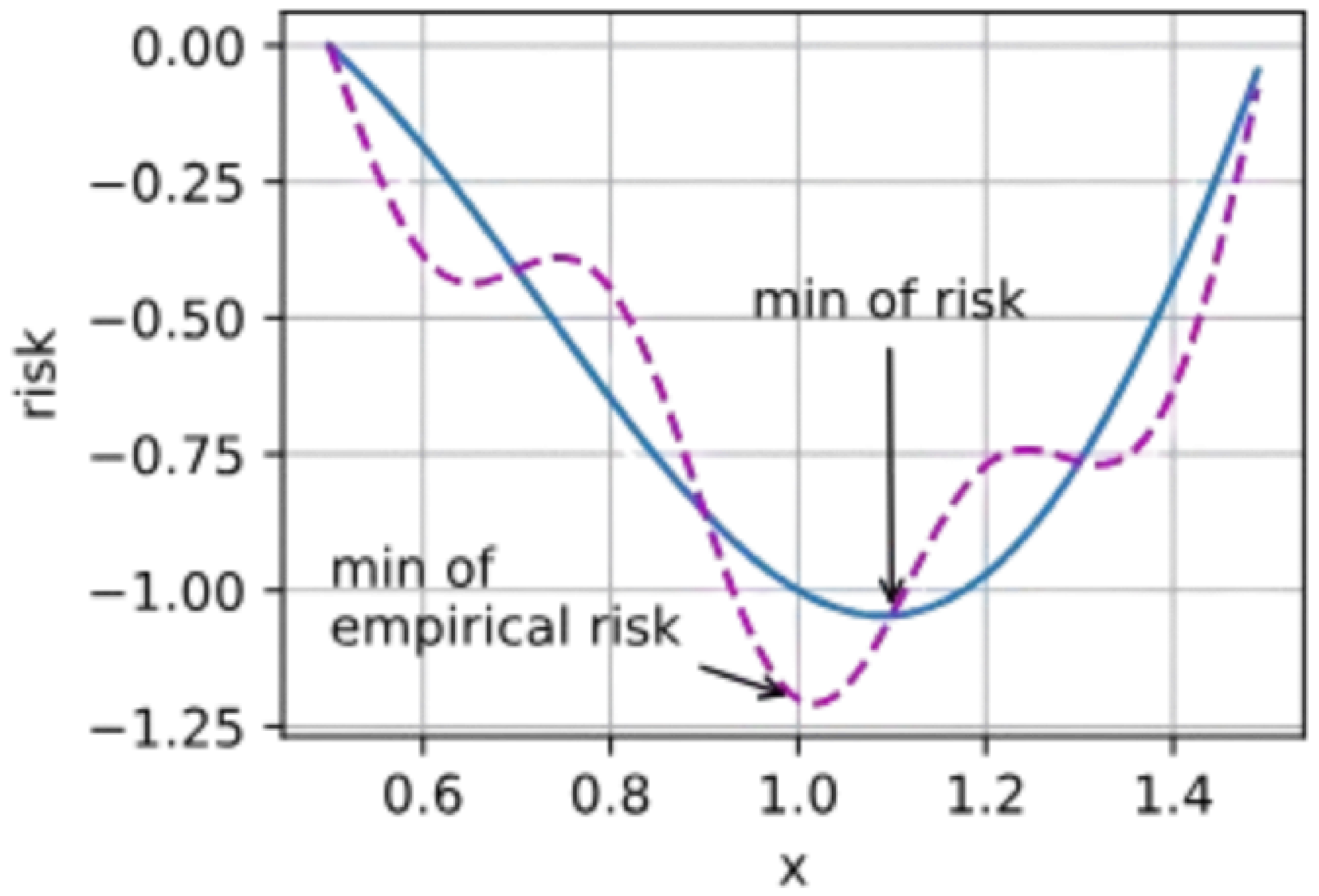


OPTIMIZATION

- Minimize the loss function
- Reduce the training error

DEEP LEARNING

- Finding a suitable model, given a finite amount of data
- Reduce the generalization error



OPTIMIZATION ISSUES

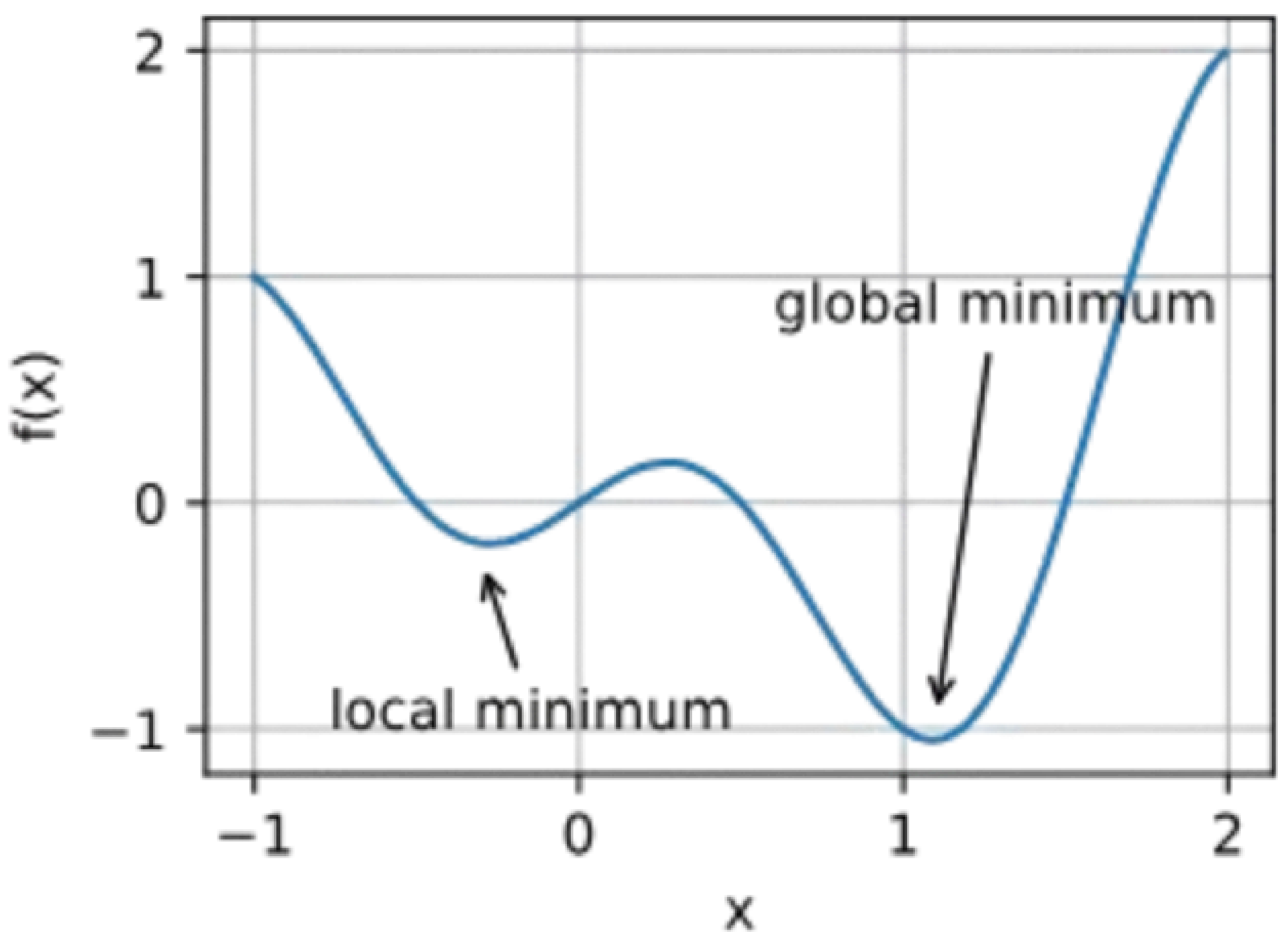
LOCAL MINIMA

For any objective function f(x), if the value of f(x) at x is smaller than the values of f(x) at any other points in the vicinity of x, then f(x) could be a local minimum.
If the value of f(x) at x is the minimum of the objective function over the entire domain, then f(x) is the global minimum.

OPTIMIZATION ISSUES

LOCAL MINIMA

The optimization problems may have many local minima.



OPTIMIZATION ISSUES

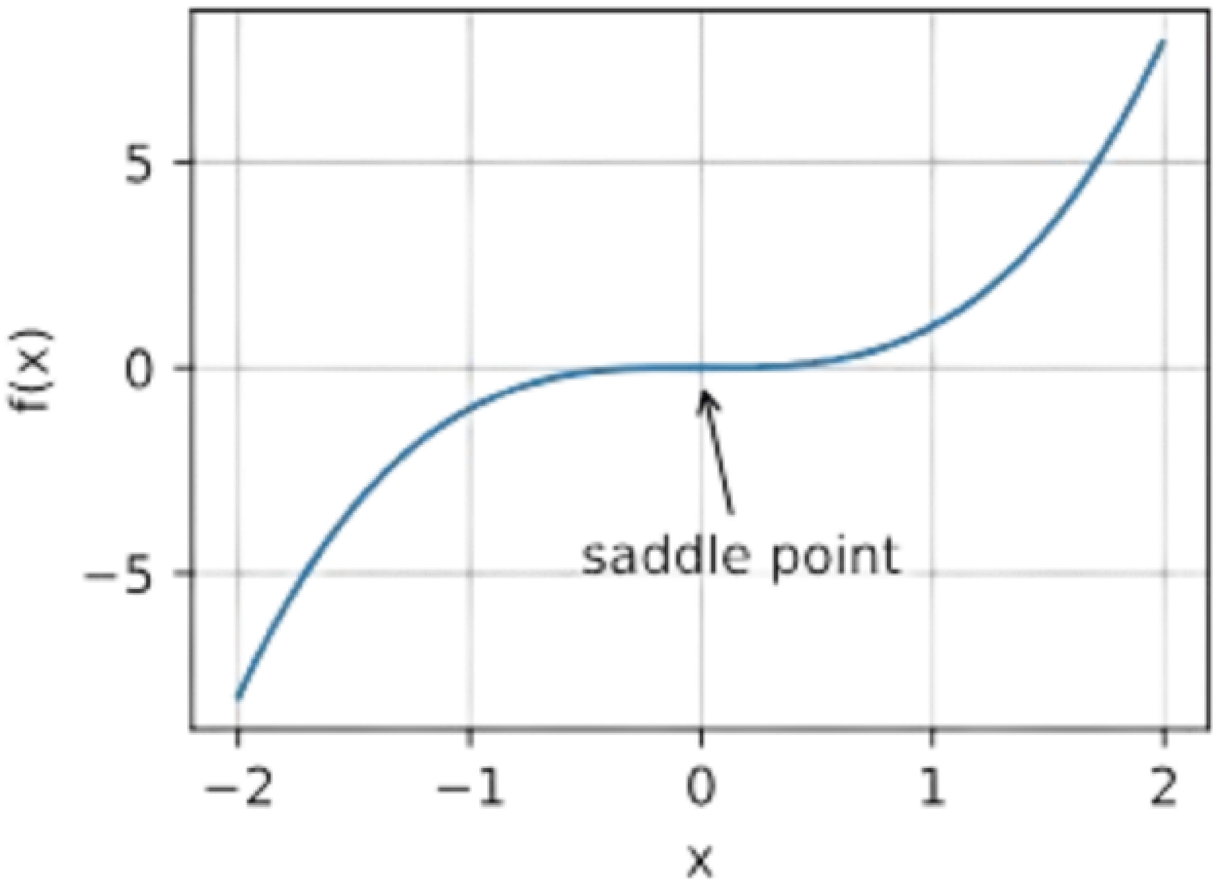
SADDLE POINTS

A saddle point is any location where all gradients of a function vanish but which is neither a global nor a local minimum.
Consider the function $f(x)=x^3$. Its first and second derivative vanish for $x=0$. Optimization might stall at this point, even though it is not a minimum.

OPTIMIZATION ISSUES

SADDLE POINTS

The problem may have even more saddle points, as generally the problems are not convex.

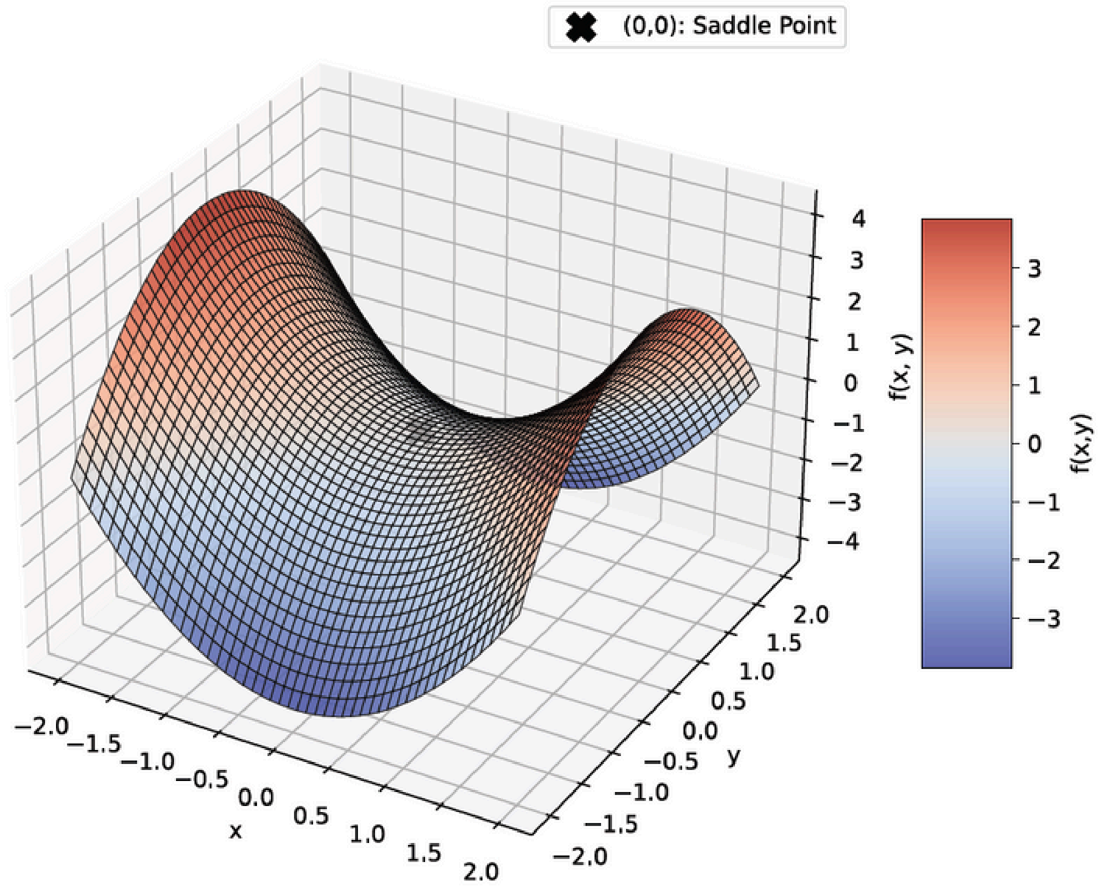


OPTIMIZATION ISSUES

SADDLE POINTS

Example: f(x,y) = x² - y²

Surface of $f(x, y) = x^2 - y^2$ (Classic Saddle Point)



1. Find Critical Points: The first step is to find where the gradient $\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$ is zero.

- $\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2 - y^2) = 2x$
- $\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(x^2 - y^2) = -2y$

Set the gradient components to zero:

- $2x = 0 \implies x = 0$
- $-2y = 0 \implies y = 0$ So, the only critical point is $(0, 0)$.

2. The Hessian Matrix $H(x, y)$: The Hessian matrix is $H(x, y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$.

- $\frac{\partial^2 f}{\partial x^2}$: Differentiate $\frac{\partial f}{\partial x} = 2x$ with respect to x:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(2x) = 2$$

- $\frac{\partial^2 f}{\partial y^2}$: Differentiate $\frac{\partial f}{\partial y} = -2y$ with respect to y:

$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(-2y) = -2$$

- $\frac{\partial^2 f}{\partial x \partial y}$: Differentiate $\frac{\partial f}{\partial y} = -2y$ with respect to x:

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(-2y) = 0$$

- $\frac{\partial^2 f}{\partial y \partial x}$: Differentiate $\frac{\partial f}{\partial x} = 2x$ with respect to y:

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(2x) = 0$$

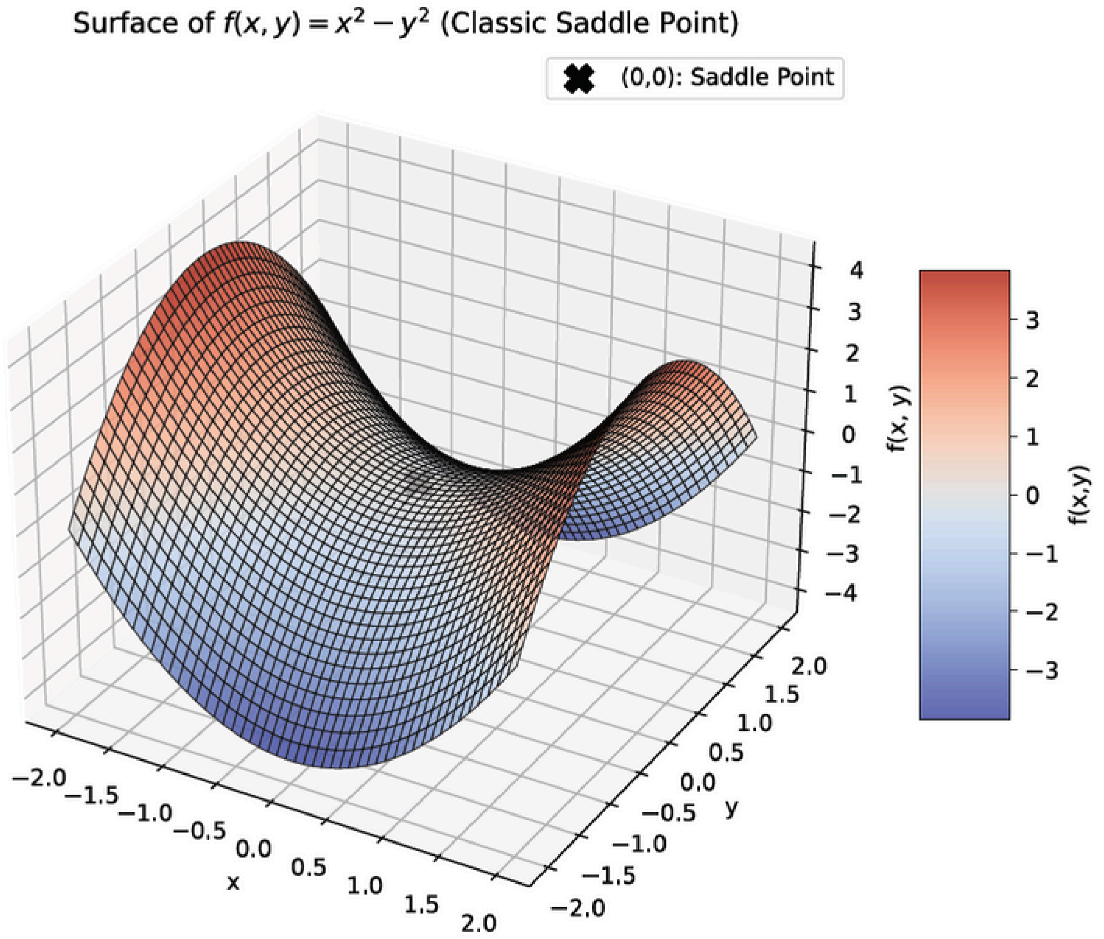
So, the Hessian matrix (which is constant for this function) is:

$$H(x, y) = H(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

OPTIMIZATION ISSUES

SADDLE POINTS

Example: f(x) = x^2 - y^2



3. Eigenvalues for $H = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$**:** We solve $\det(H - \lambda I) = 0$.

- $H - \lambda I = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2-\lambda & 0 \\ 0 & -2-\lambda \end{pmatrix}$
- $\det(H - \lambda I) = (2-\lambda)(-2-\lambda) - (0)(0) = (2-\lambda)(-2-\lambda)$
- Set the determinant to zero: $(2-\lambda)(-2-\lambda) = 0$
- This gives the eigenvalues:
  - $2 - \lambda = 0 \implies \lambda_1 = 2$
  - $-2 - \lambda = 0 \implies \lambda_2 = -2$

4. Classification of the Critical Point $(0,0)$**:** Since the eigenvalues are $\lambda_1 = 2$ (positive) and $\lambda_2 = -2$ (negative), the critical point $(0,0)$ is a **saddle point**.

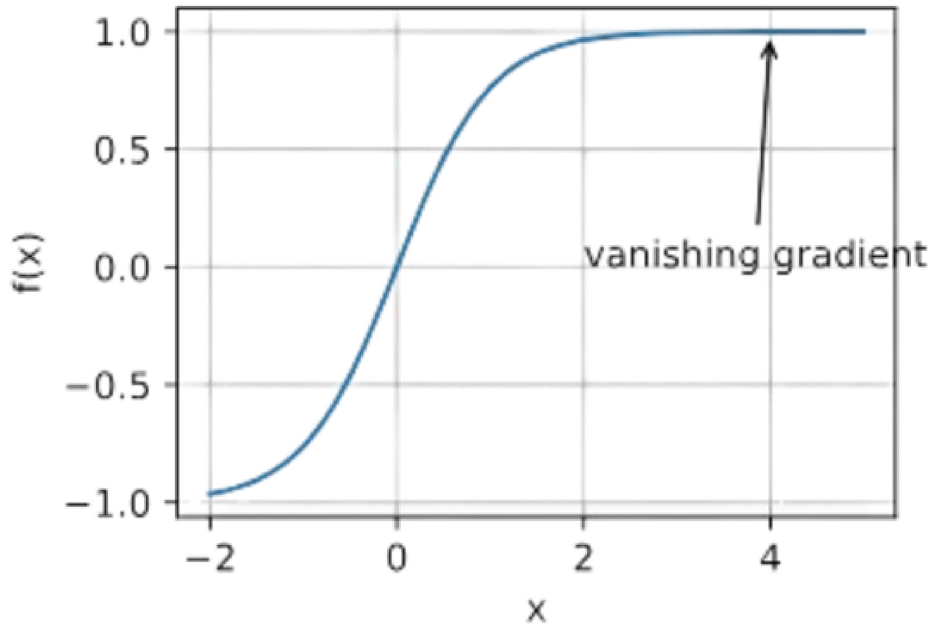
OPTIMIZATION ISSUES

VANISHING GRADIENTS

In a neural network, learning happens by adjusting the network's weights based on the error (or loss) calculated at the output. The backpropagation algorithm computes the gradient (which is essentially the derivative of the loss function with respect to each weight) and uses this gradient to update the weights. The vanishing gradient problem occurs when these gradients become extremely small as they are propagated backward from the output layers to the earlier layers of the network

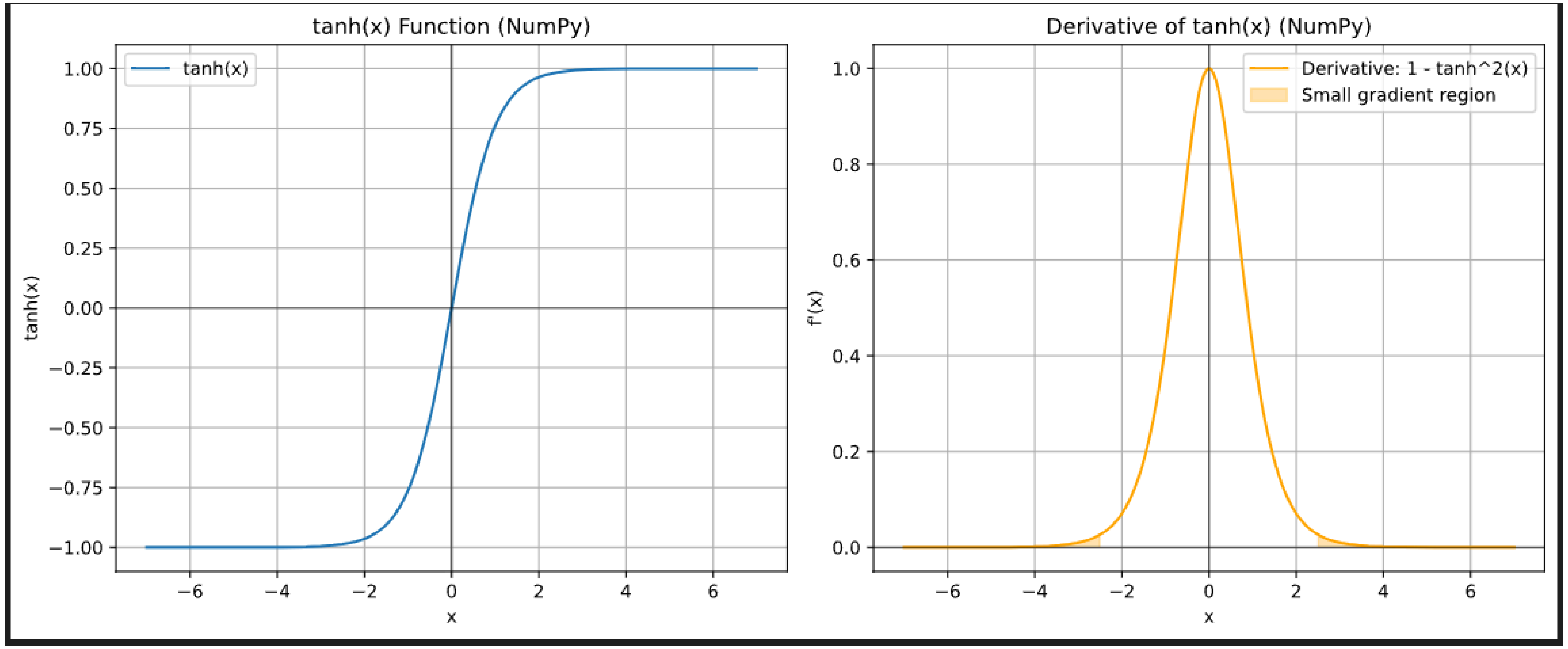
For instance, assume that we want to minimize the function f(x)=tanh(x) and we happen to get started at x=4. The gradient of f is close to nil. More specifically, f’(x)=1-tanh^2(x) and thus f’(4)=0.0013.

Vanishing gradients can cause optimization to stall.



OPTIMIZATION ISSUES

VANISHING GRADIENTS

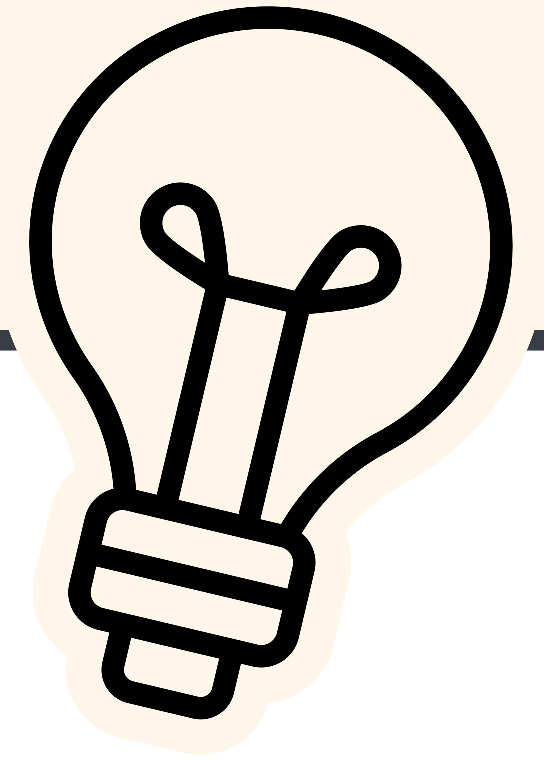


OPTIMIZATION ISSUES

VANISHING GRADIENTS

- Slow Training: The most immediate effect is that the network trains very slowly, as the earlier layers are not updated effectively.
- Poor Performance: If the early layers don't learn, the network cannot learn complex features and will likely perform poorly on the task it's being trained for.
- Inability to Train Deep Networks: Vanishing gradients historically made it very difficult to train very deep neural networks effectively.

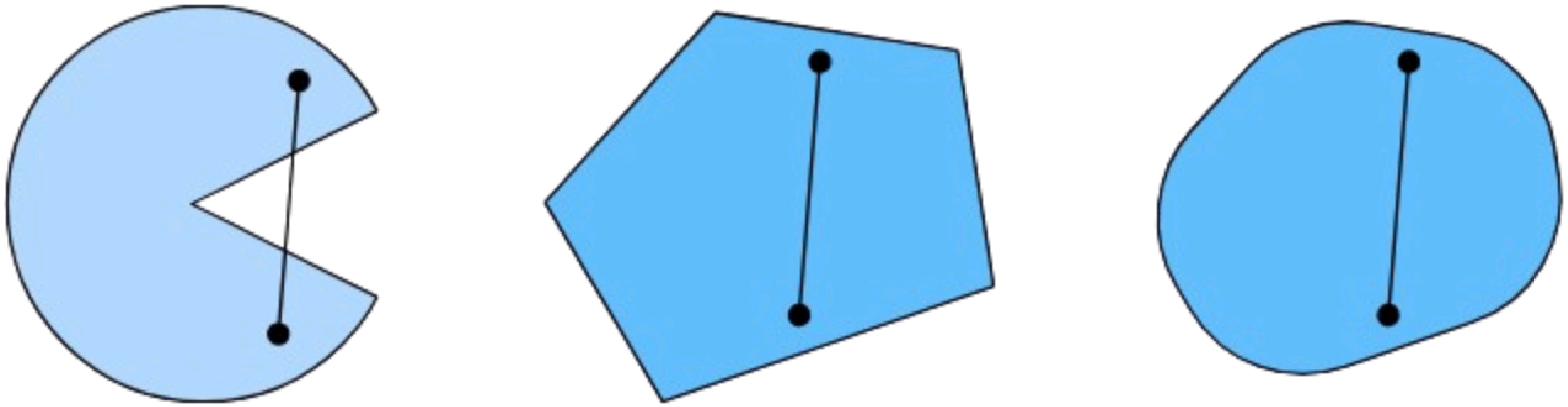
CONVEXITY



- Convexity plays a vital role in the design of optimization algorithms.
- This is largely due to the fact that it is much easier to analyze and test algorithms in such a context.

CONVEX SETS

Sets are the basis of convexity. Simply put, a set X in a vector space is convex if for any $\mathbf{a},\mathbf{b} \in X$ the line segment connecting $\mathbf{a}$ and $\mathbf{b}$ is also in X.

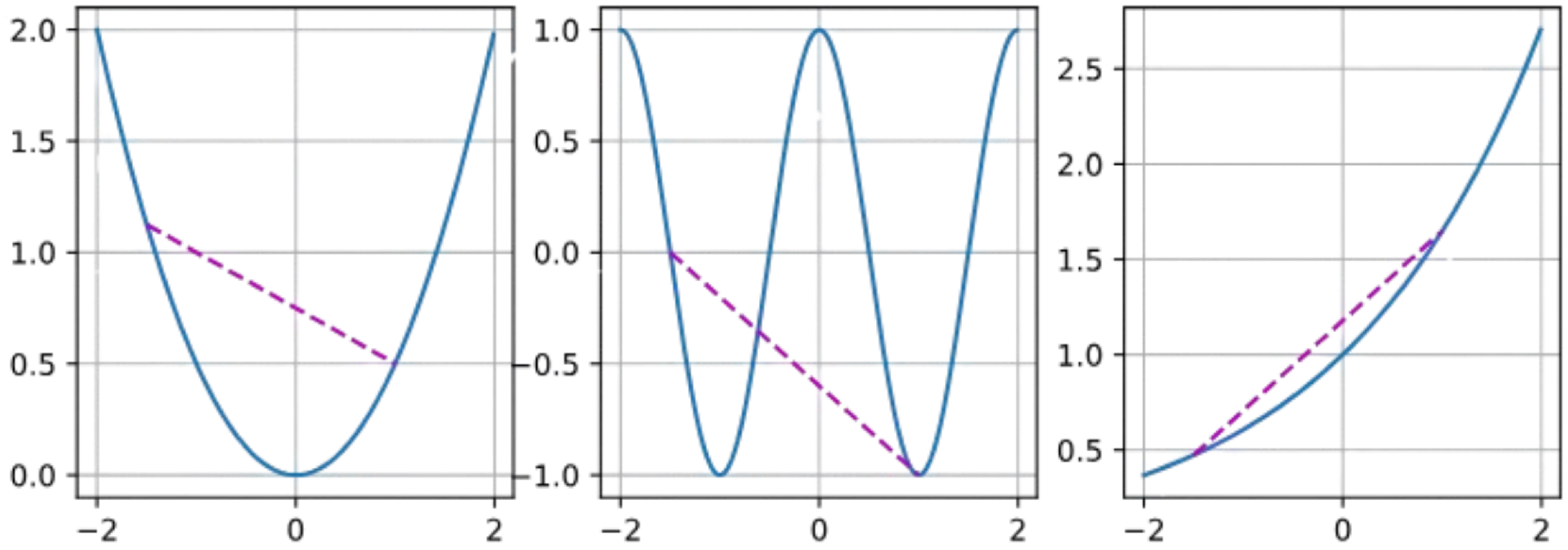


The first set is nonconvex and the other two are convex.

CONVEX FUNCTIONS

Now that we have convex sets we can introduce convex functions f. Given a convex set X, a function f:X→R is convex if for all x,x′∈X and for all λ∈[0,1] we have

$$\lambda f(x) + (1-\lambda) f(x') \geq f(\lambda x + (1-\lambda)x').$$



KEY PROPERTY

Local Minima Are Global Minima

Consider a convex function f defined on a convex set $\mathcal{X}$. Suppose that $x^* \in \mathcal{X}$ is a local minimum: there exists a small positive value p so that for $x \in \mathcal{X}$ that satisfies $0 < |x - x^*| \le p$ we have $f(x^*) < f(x)$.

Assume that the local minimum x^* is not the global minimum of f: there exists $x' \in \mathcal{X}$ for which $f(x') < f(x^*)$. There also exists $\lambda \in [0, 1)$ such as $\lambda = 1 - \frac{p}{|x^* - x'|}$ so that $0 < |\lambda x^* + (1 - \lambda)x' - x^*| \le p$.

However, according to the definition of convex functions, we have

$$\begin{aligned} f(\lambda x^* + (1-\lambda)x') &\le \lambda f(x^*) + (1-\lambda) f(x') \\ &< \lambda f(x^*) + (1-\lambda) f(x^*) \\ &= f(x^*), \end{aligned} \tag{12.2.5}$$

which contradicts with our statement that x^* is a local minimum. Therefore, there does not exist $x' \in \mathcal{X}$ for which $f(x') < f(x^*)$. The local minimum x^* is also the global minimum.

For instance, the convex function $f(x) = (x-1)^2$ has a local minimum at $x = 1$, which is also the global minimum.

CONSTRAINT

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad c_i(x) \leq 0,\ i = 1, \ldots, m$$

Assumptions:

- $f(x)$: convex objective function
- $c_i(x)$: convex inequality constraints

CONSTRAINT

Lagrangian

$$L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_i \alpha_i c_i(\mathbf{x}) \quad \text{with } \alpha_i \geq 0$$

- $\mathbf{x}$ is the **optimization variable** (a vector).
- $f(\mathbf{x})$ is the **objective function** that we want to minimize.
- $c_i(\mathbf{x}) \leq 0$ are the **inequality constraint functions**.
- $\alpha_i \geq 0$ are the **Lagrange multipliers** (also called **dual variables**) associated with each constraint.

CONSTRAINT

Penalty

$$\min_x \; f(x) + r \cdot \sum_{i=1}^{m} \max(0, c_i(x))^2$$

Where:

- $r > 0$ is a penalty parameter
- If x **violates** the constraint ⇒ high penalty
- If x is **feasible** ⇒ zero penalty

Q&A

WELL DONE