COMP 414/514: Optimization – Algorithms, Complexity and Approximations

Lecture 2

Overview

- Different objective classes
- Different strategies within each problem
- Different approaches based on computational capabilities
- Different approaches based on constraints

$$\min_x \quad f(x)$$

$$\text{s.t.} \quad x \in \mathcal{C}$$

And, always having in mind applications in machine learning, AI and signal processing

Overview

- In the last lecture, we:
  - Introduced some very basic ideas from linear algebra

- In this lecture, we will:
  - Discuss briefly **smooth continuous optimization**
  - Introduce the important class of **convex optimization**
  - Discuss about **convergence rates** and some **lower bounds** on such rates

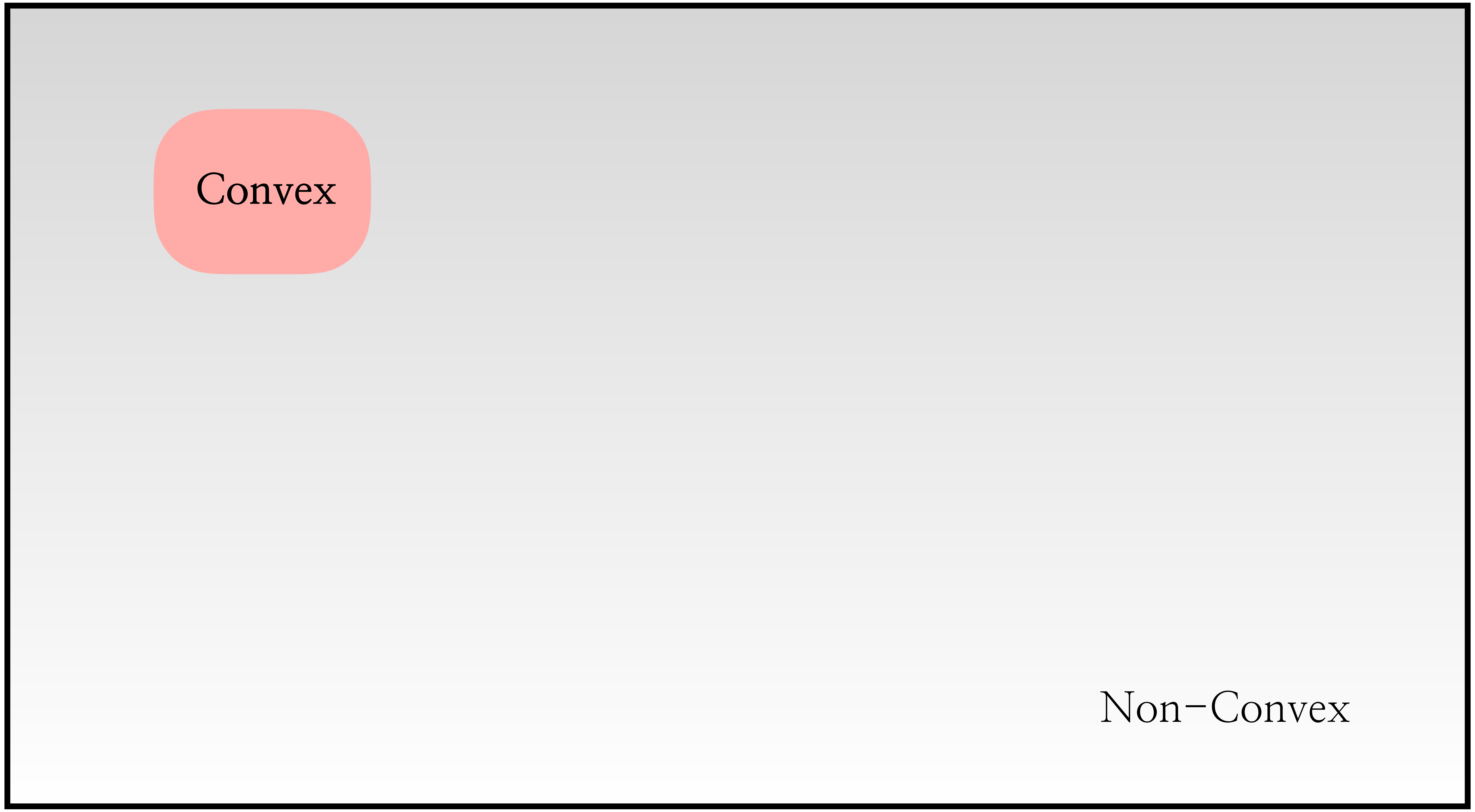
Convex vs. non−convex optimization

(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization

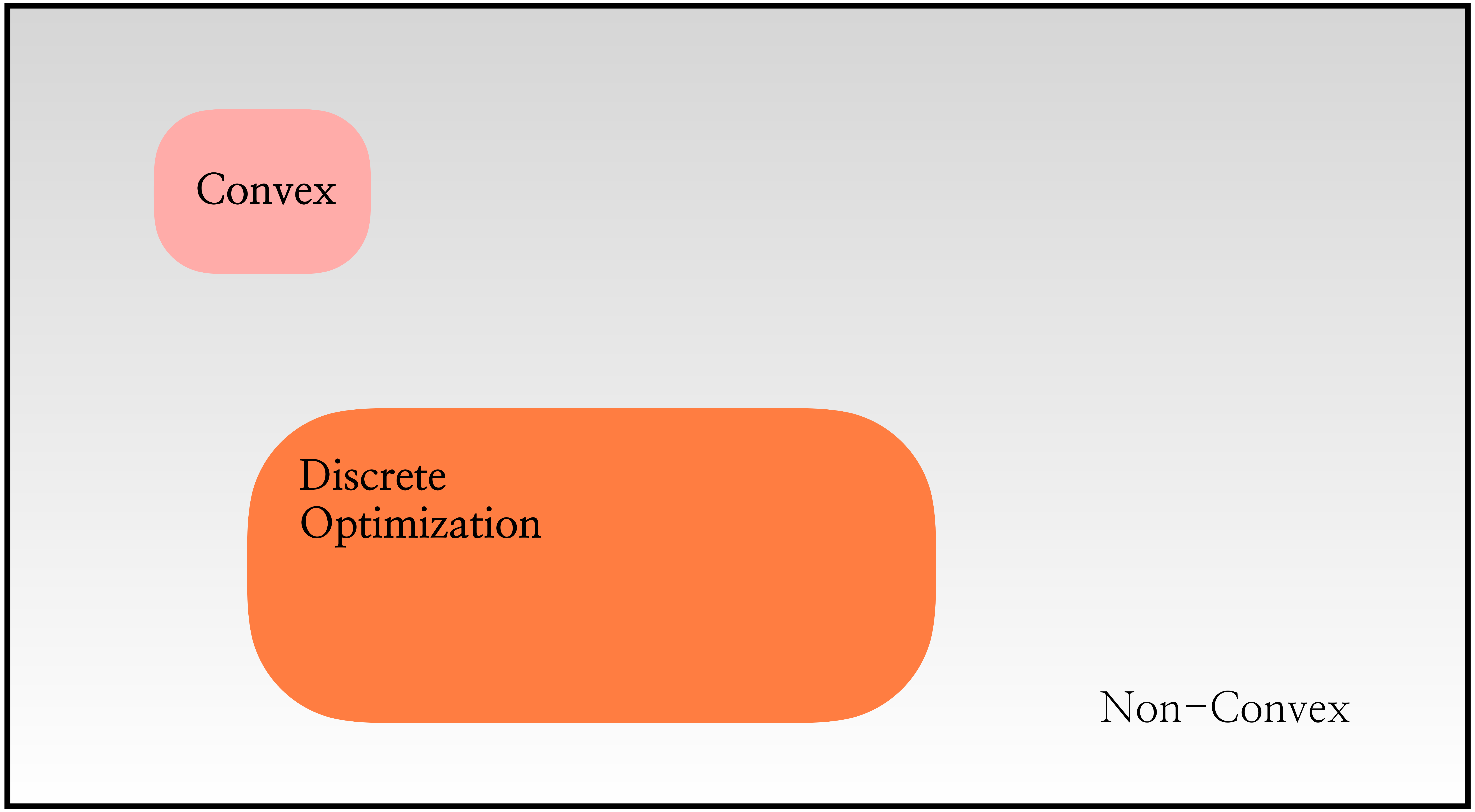
(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



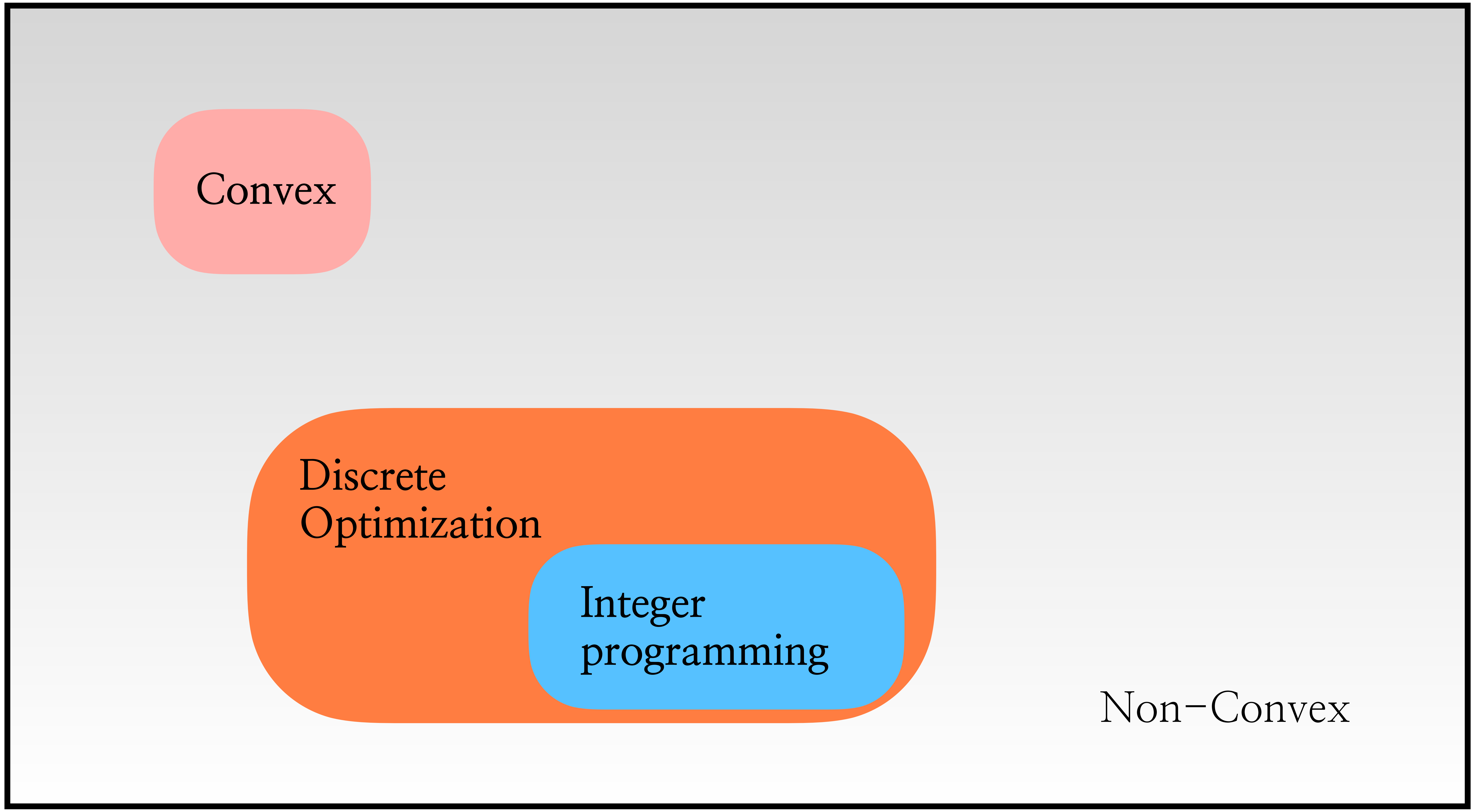
(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



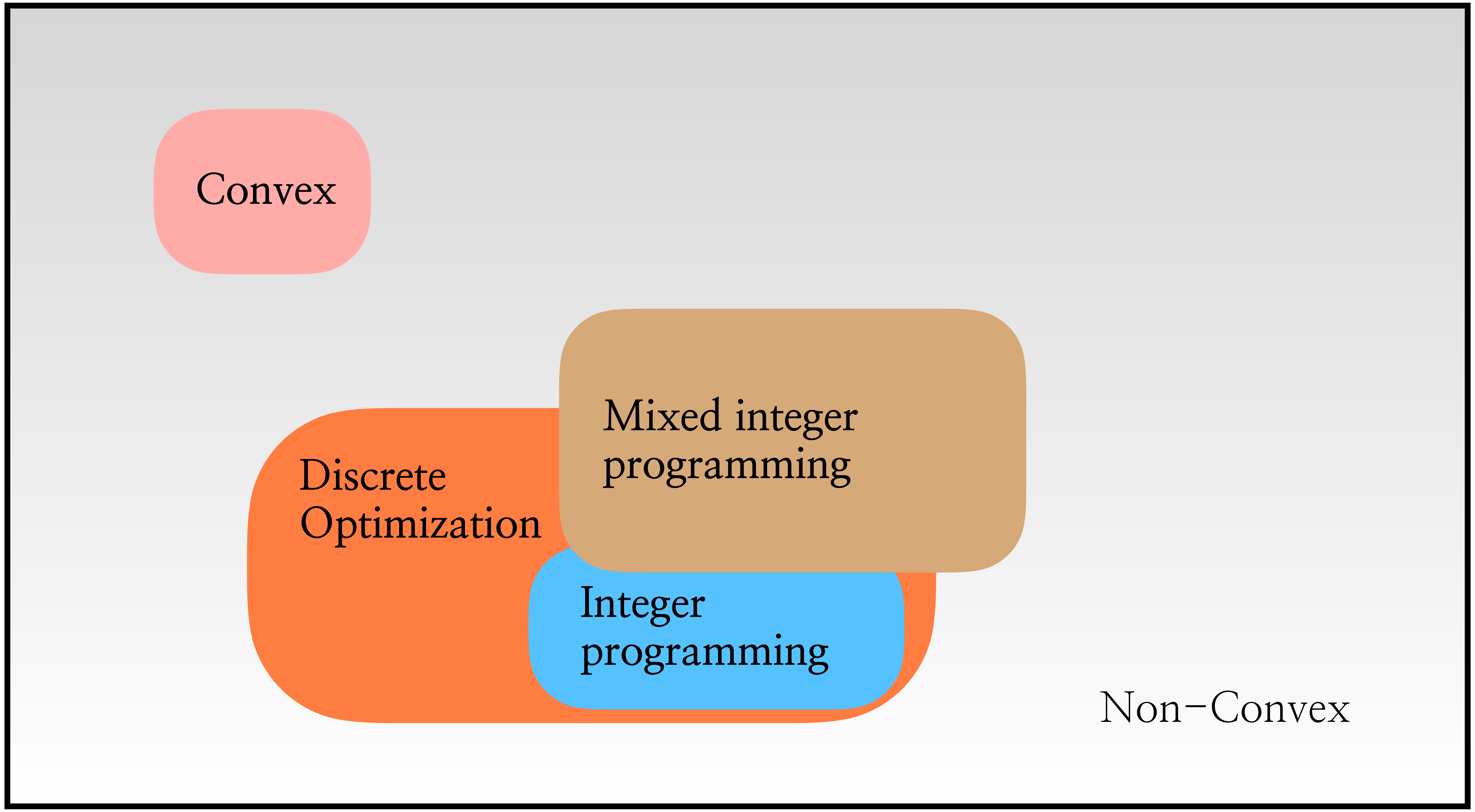
(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



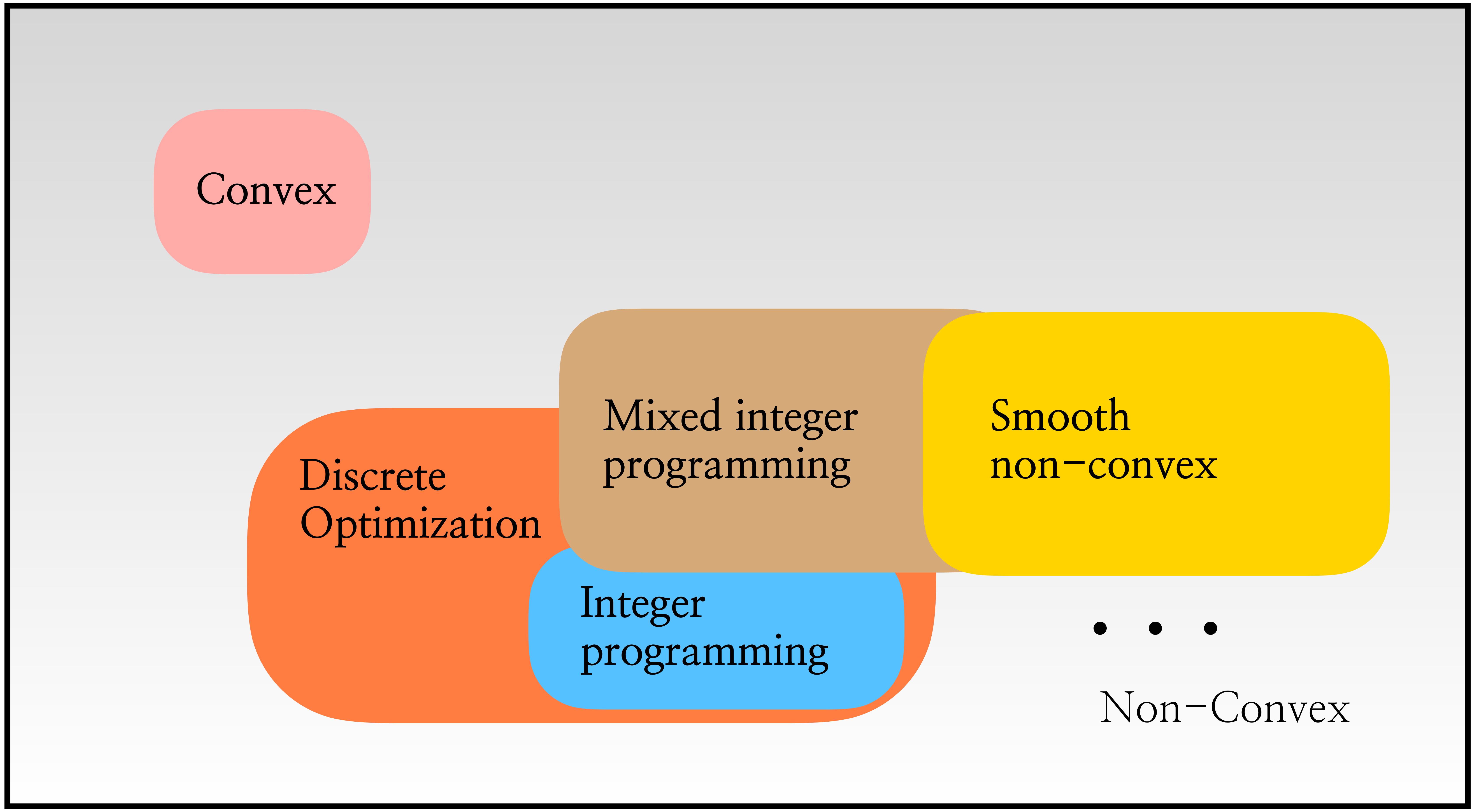
(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



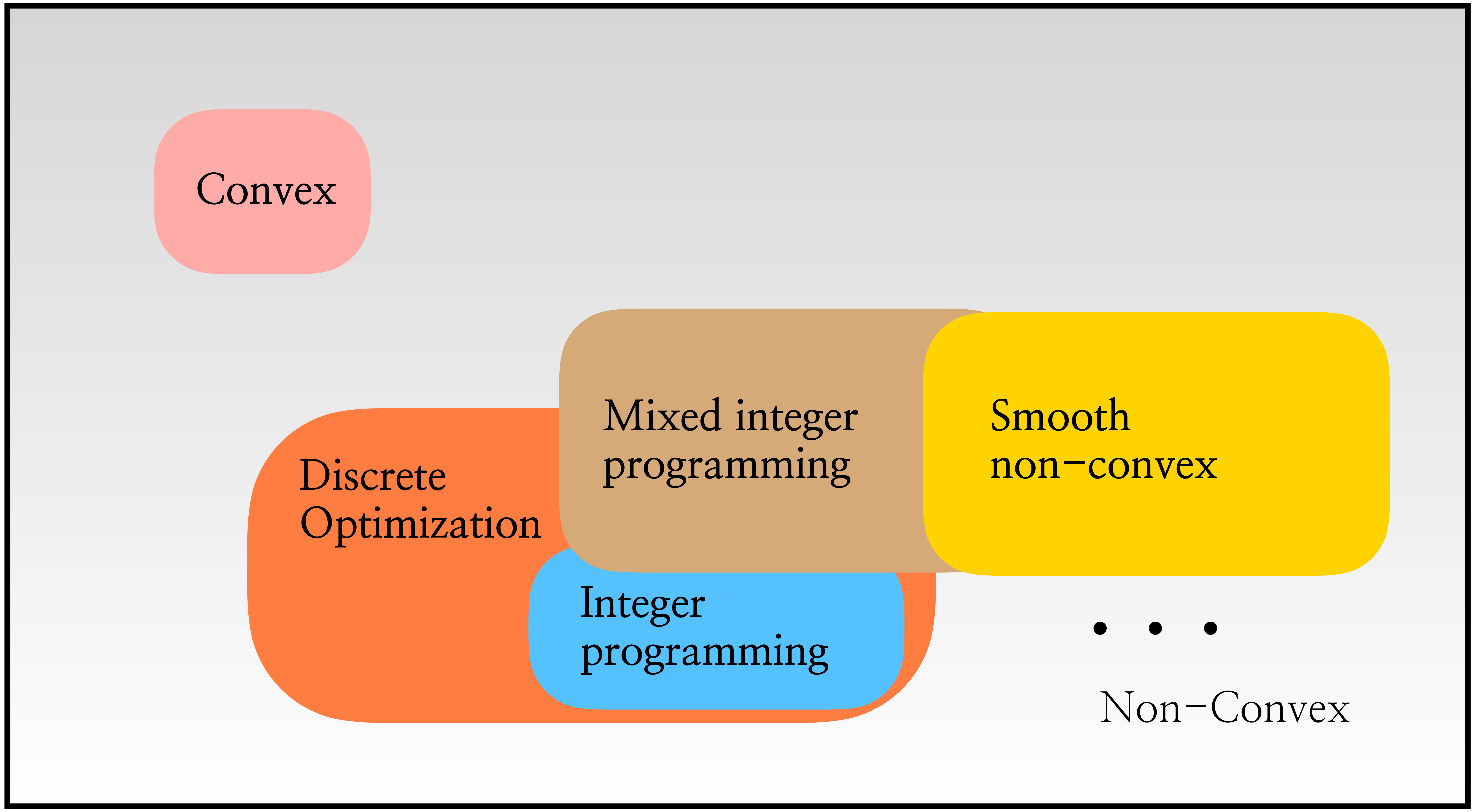
(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



(Naive interpretation of) Space of optimization problems

Convex vs. non−convex optimization



(Naive interpretation of) Space of optimization problems

Derivatives and gradients

- Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

Derivatives and gradients

- Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

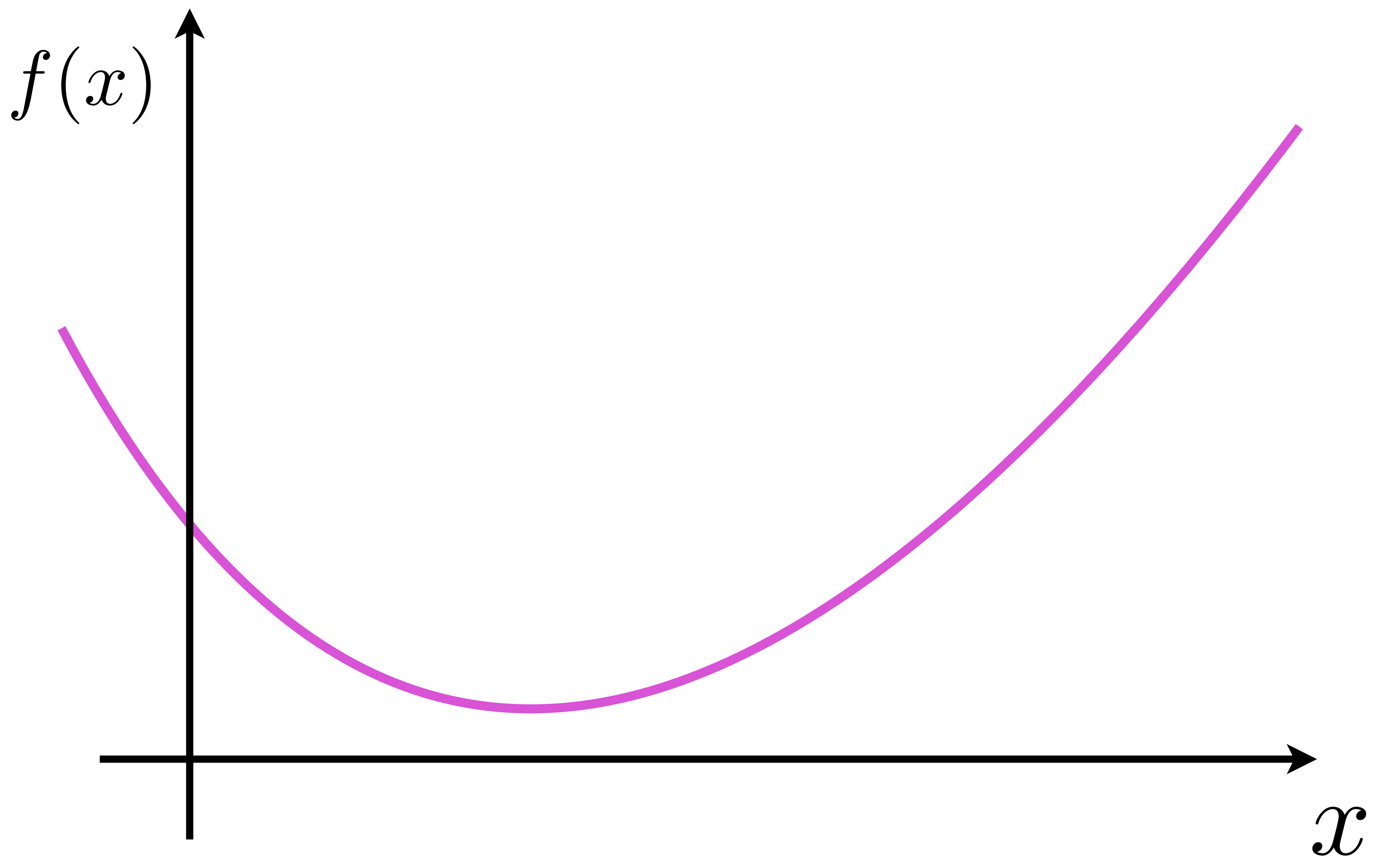
- Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \qquad \Bigg| \qquad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

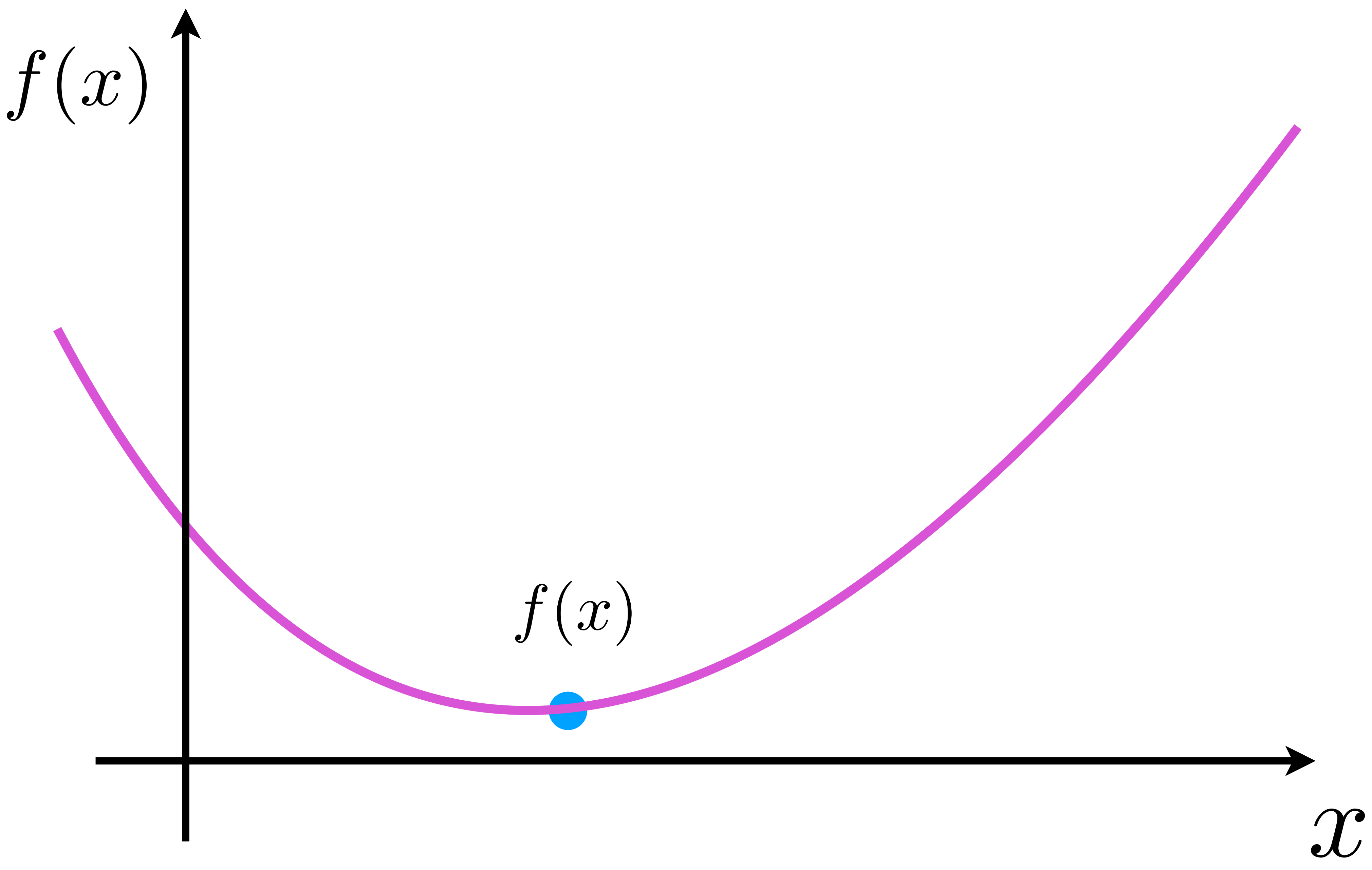


Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

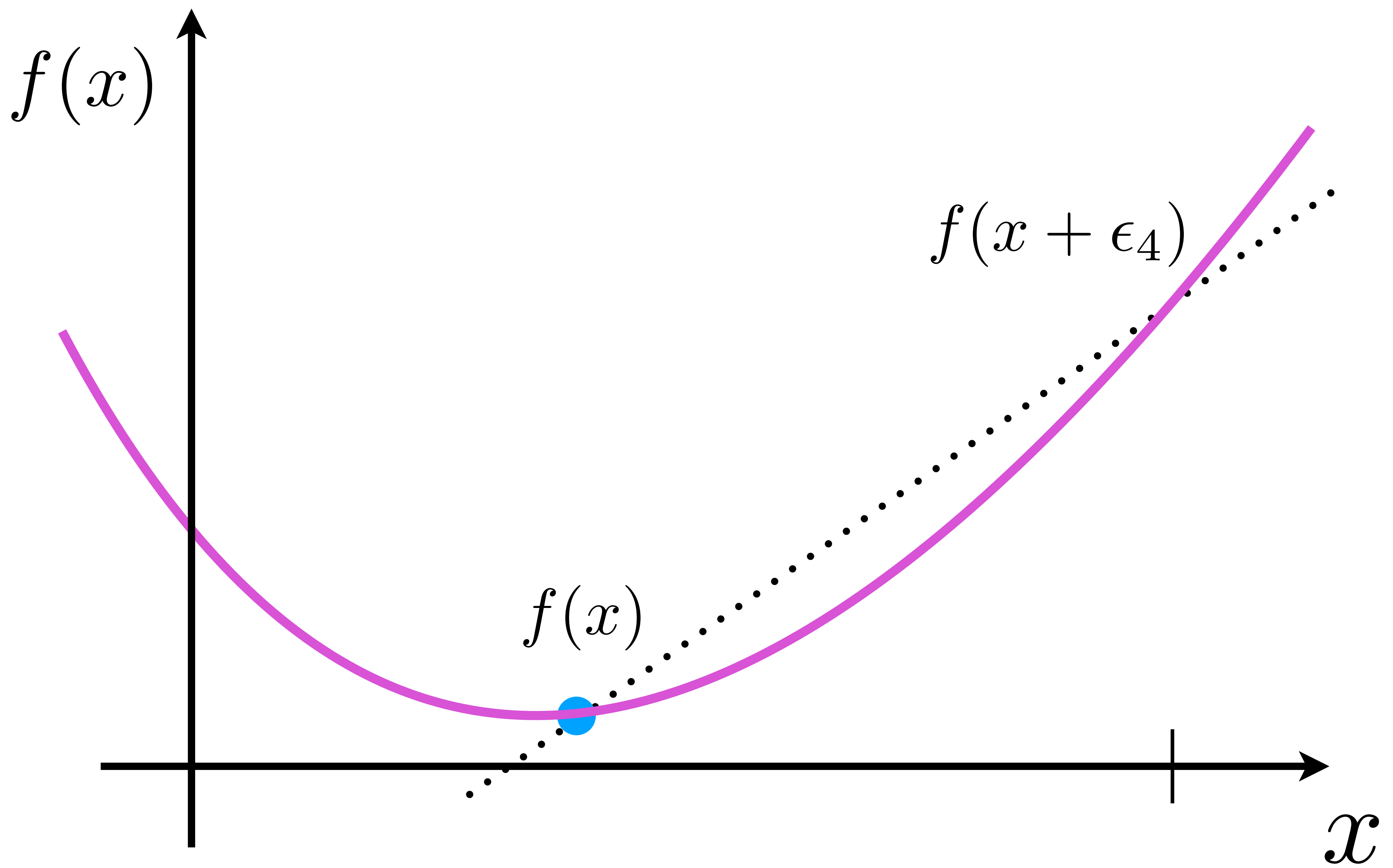


Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

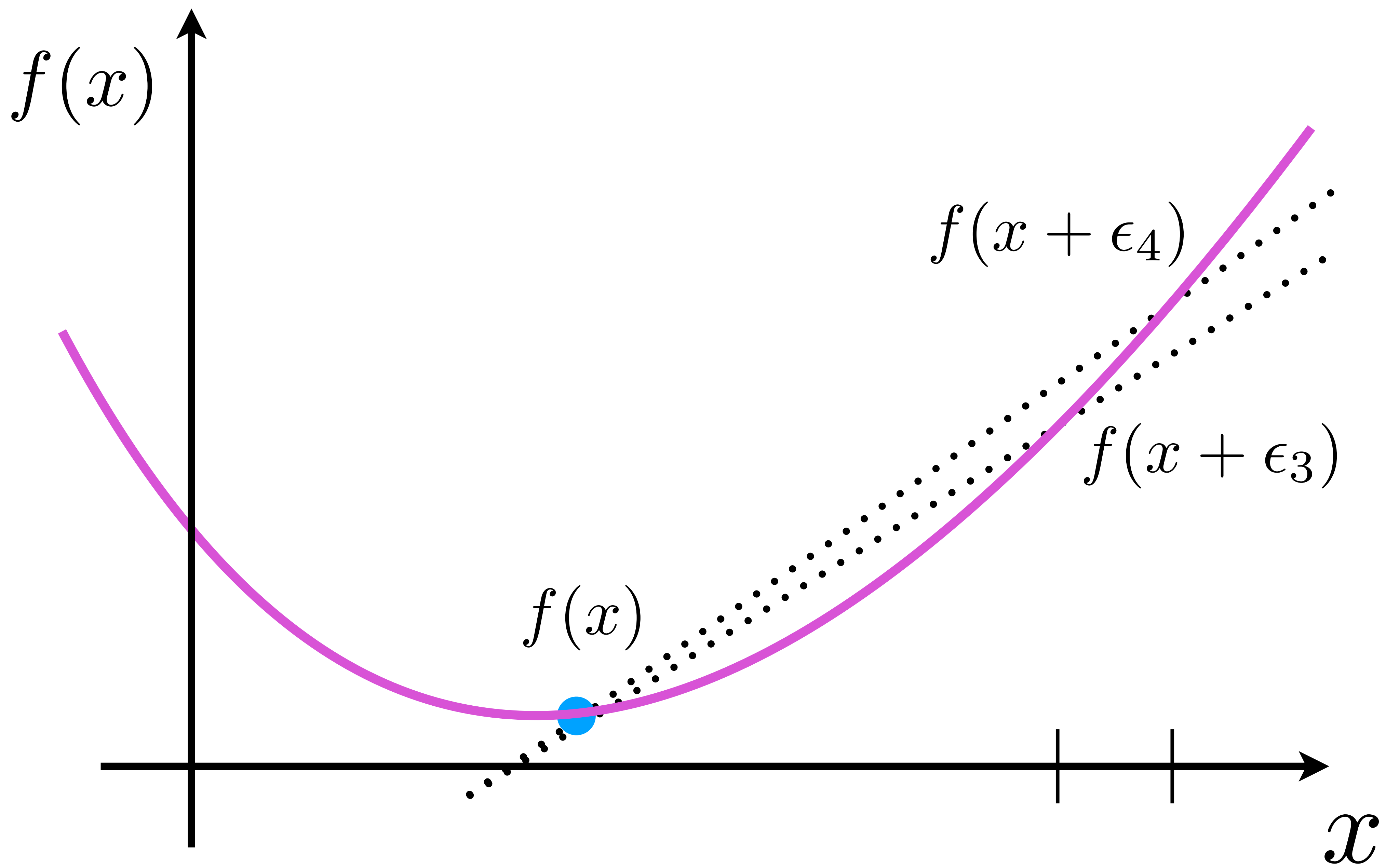


Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

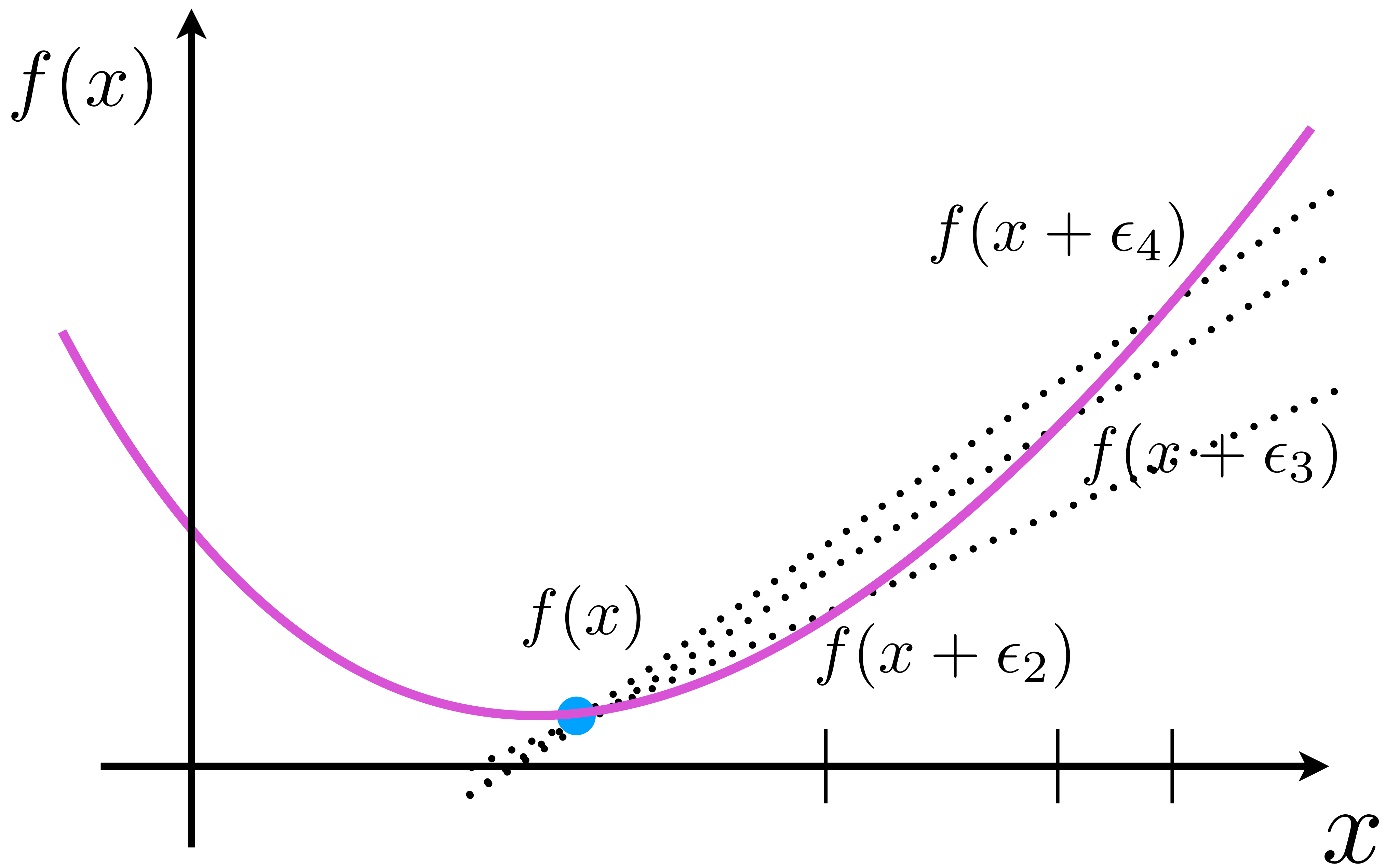


Derivatives and gradients

- Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

- Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

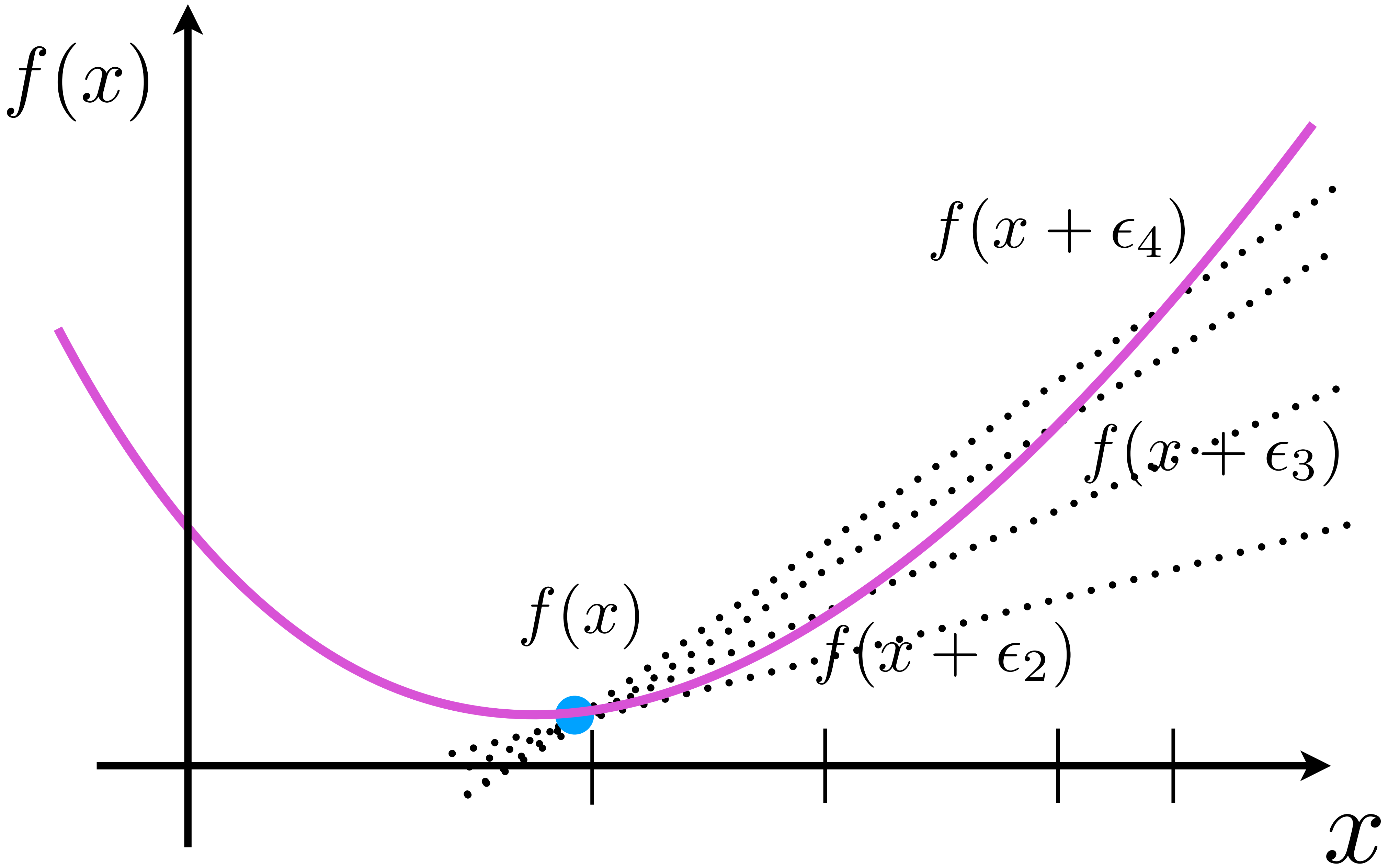


Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

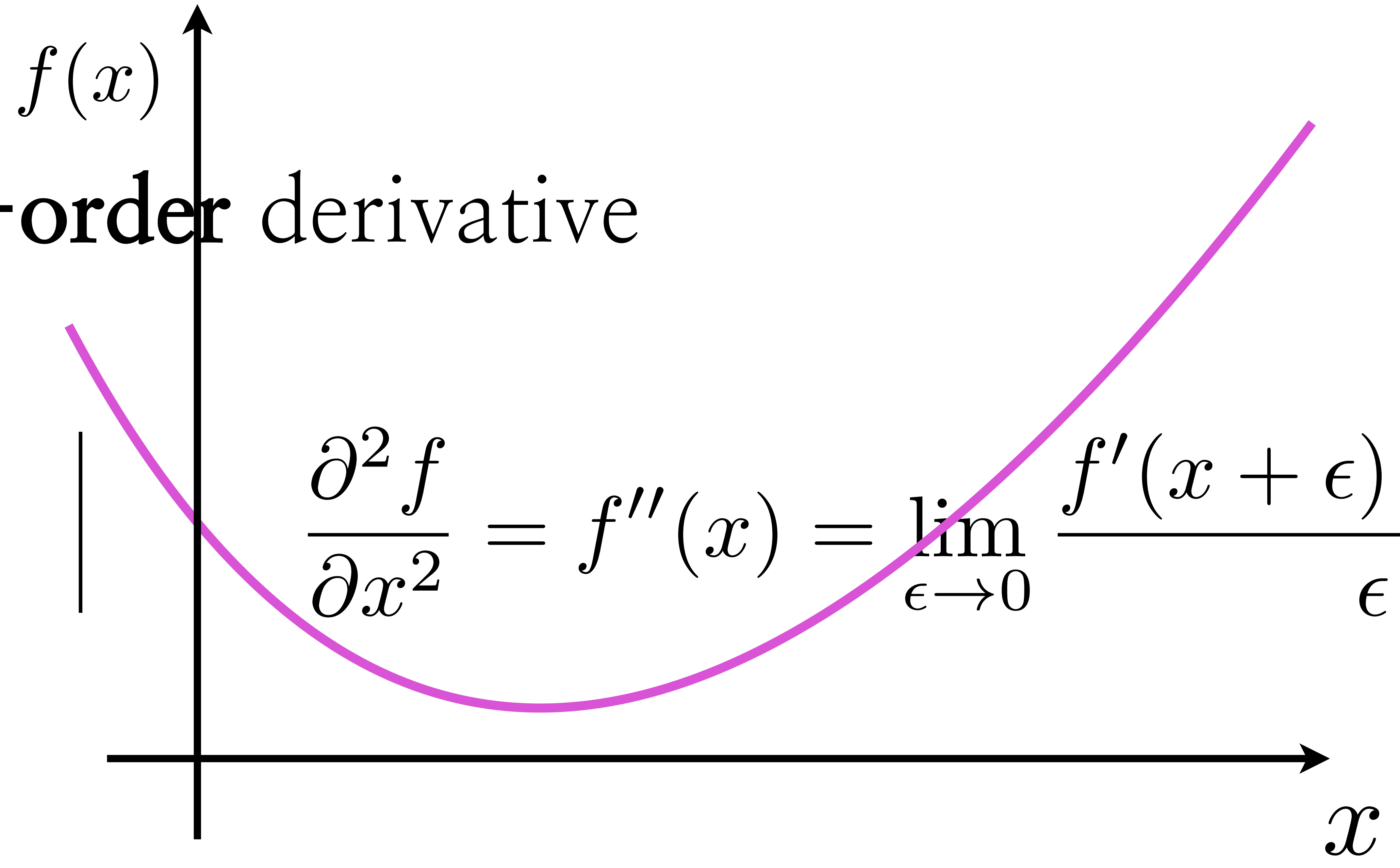


Derivatives and gradients

– Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

– Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$



– Definition of **second-order** derivative

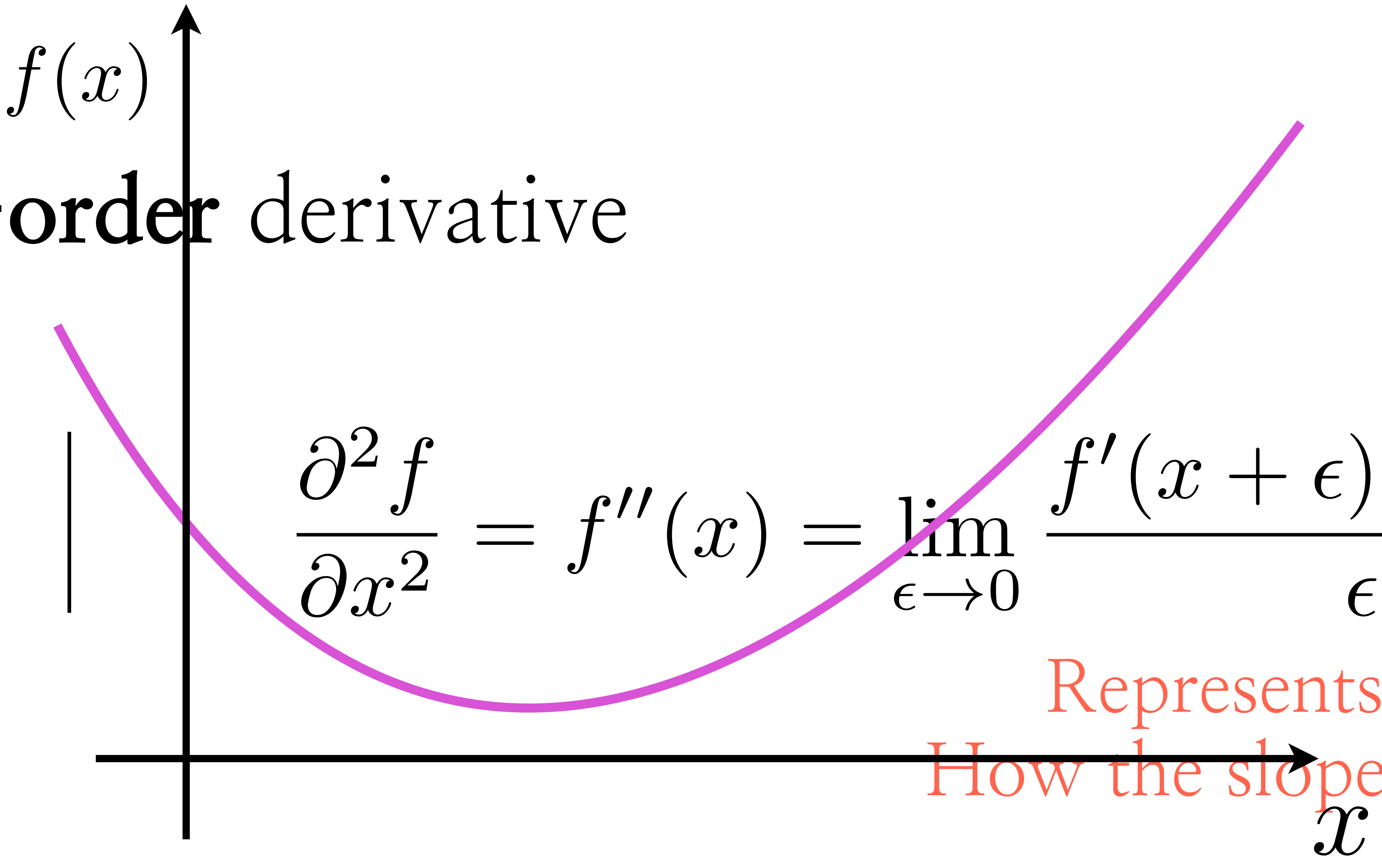
$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial^2 f}{\partial x^2} = f''(x) = \lim_{\epsilon \to 0} \frac{f'(x+\epsilon) - f'(x)}{\epsilon}$$

Derivatives and gradients

- Definition of a **derivative**

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial f}{\partial x} = f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

- Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$



- Definition of **second-order** derivative

$$f : \mathbb{R} \to \mathbb{R} \quad \Big| \quad \frac{\partial^2 f}{\partial x^2} = f''(x) = \lim_{\epsilon \to 0} \frac{f'(x+\epsilon) - f'(x)}{\epsilon}$$

Represents the local curvature:
How the slope of the function changes

Derivatives and gradients

- Generalization to multiple components: **gradient**

$$f : \mathbb{R}^p \to \mathbb{R} \qquad \Bigg| \qquad \nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} \in \mathbb{R}^p$$

where

$$\frac{\partial f}{\partial x_i} = \lim_{\epsilon \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + \epsilon, x_{i+1}, \ldots, x_p) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_p)}{\epsilon} = \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$$

Derivatives and gradients

- **Jacobian** matrix (relates to neural networks)

$$f : \mathbb{R}^p \to \mathbb{R}^m \qquad \bigg| \qquad Df(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_p} \end{bmatrix} \in \mathbb{R}^{m \times p}$$

- Generalizes the notion of gradient to multiple−output functions

Derivatives and gradients

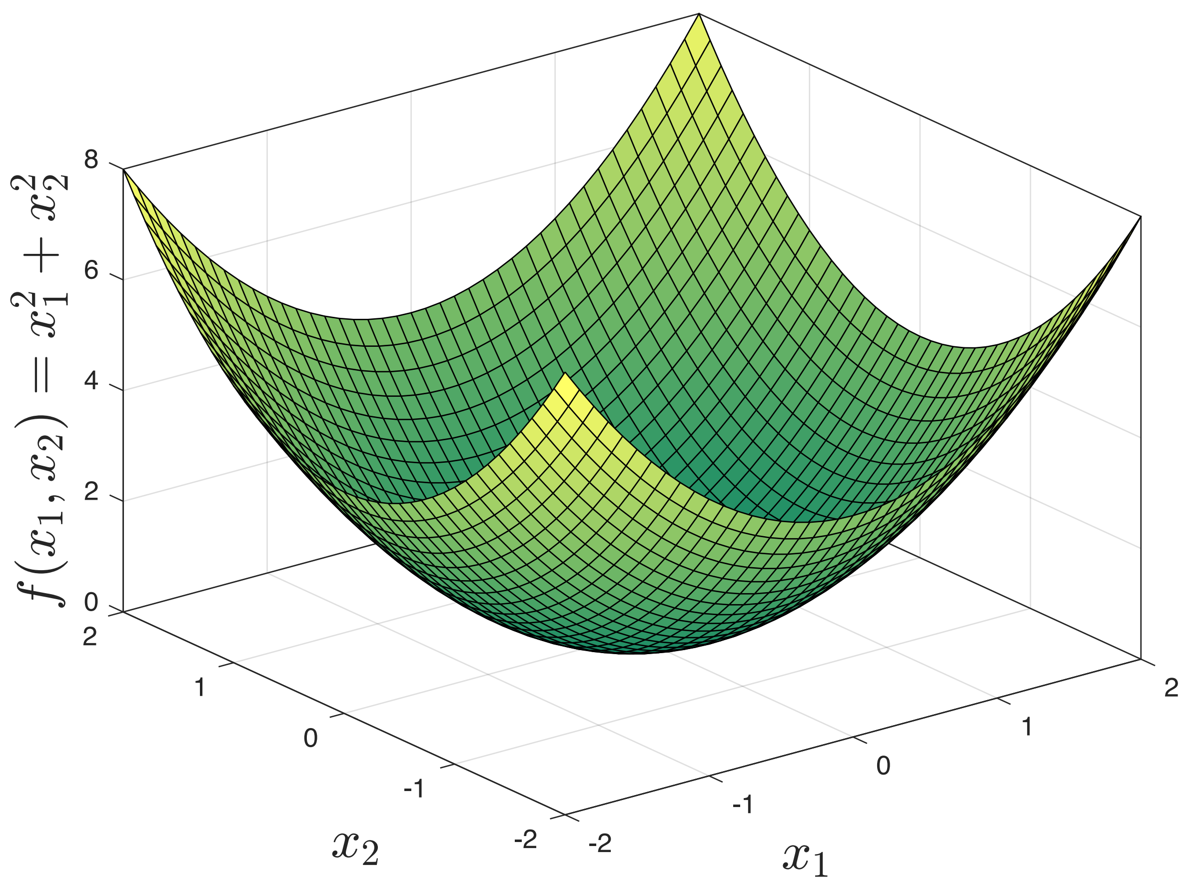
- **Hessian** matrix

$$f : \mathbb{R}^p \to \mathbb{R} \qquad \Bigg| \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_p x_1} & \frac{\partial^2 f}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_p^2} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

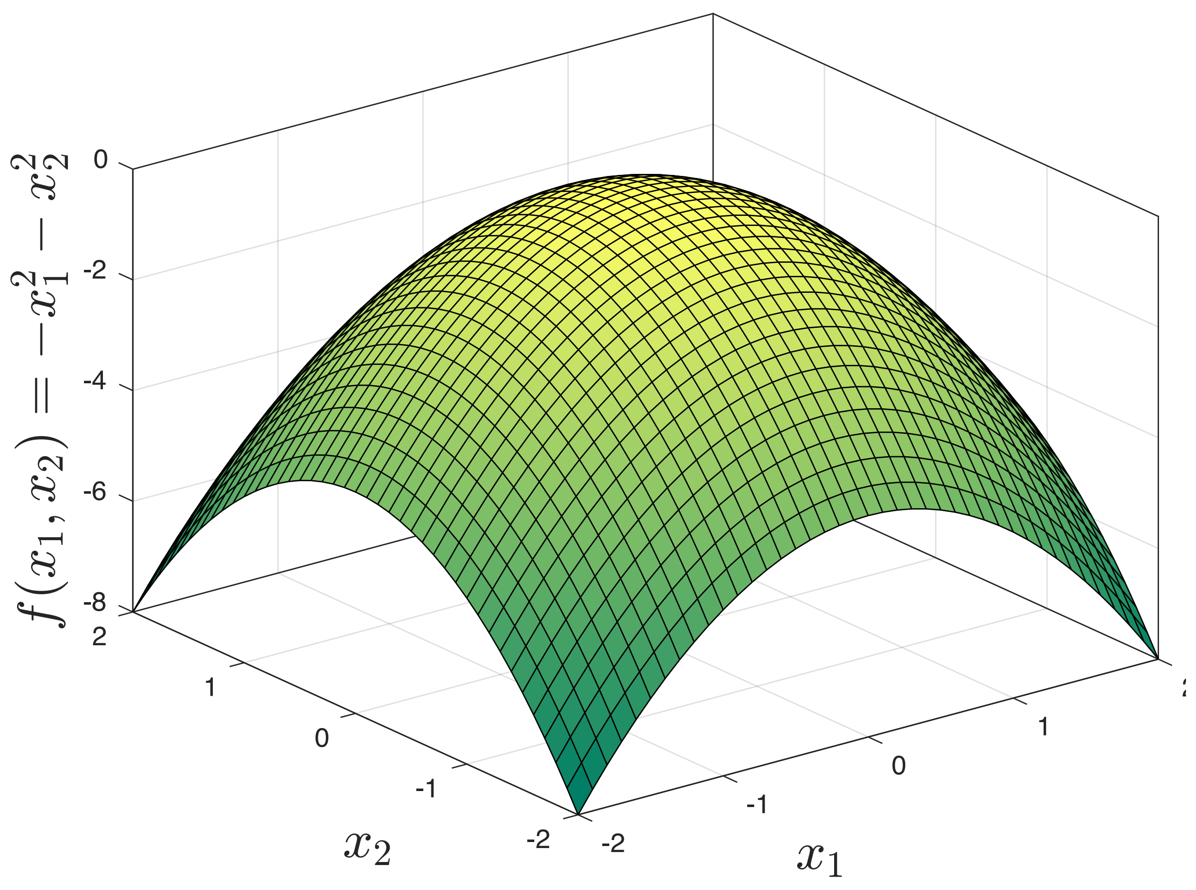
Derivatives and gradients

– **Hessian** matrix

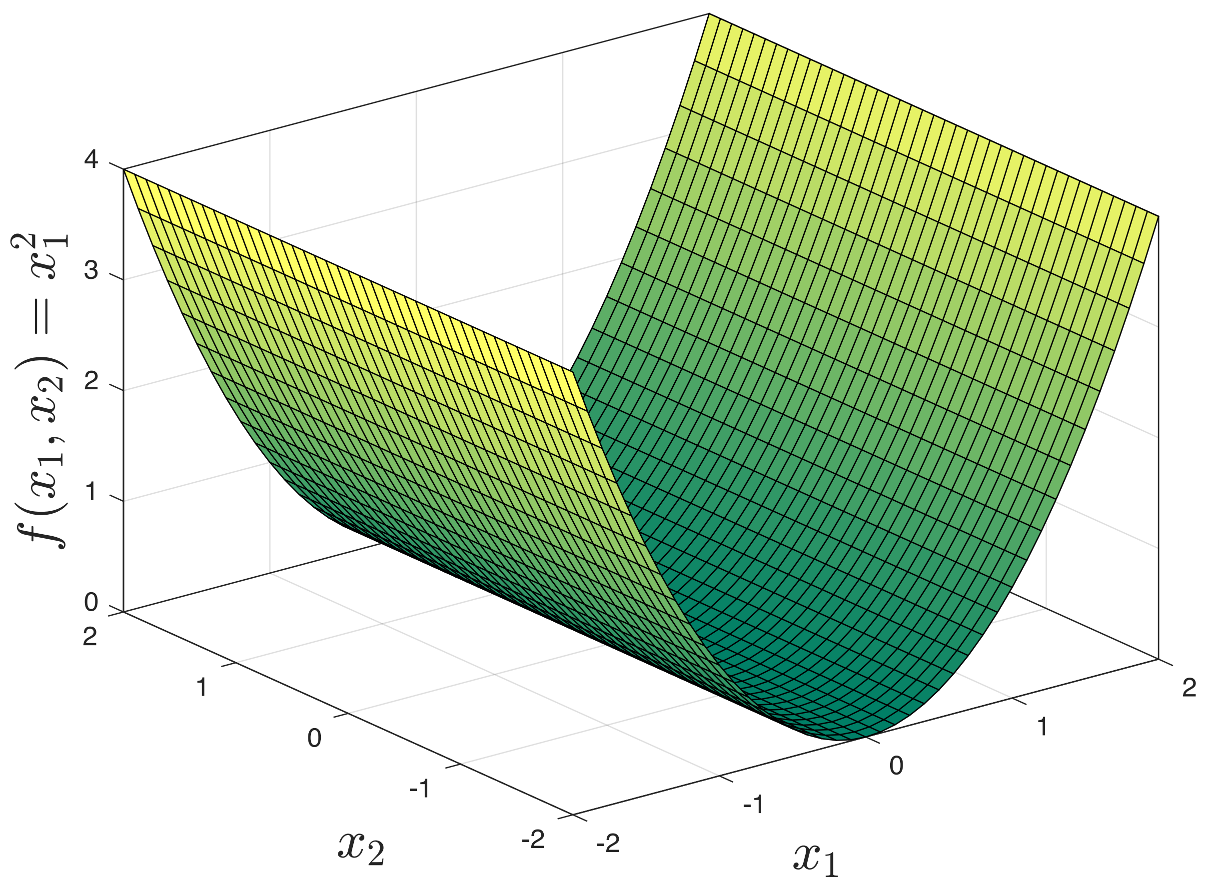
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_p x_1} & \frac{\partial^2 f}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_p^2} \end{bmatrix} \in \mathbb{R}^{p \times p}$$



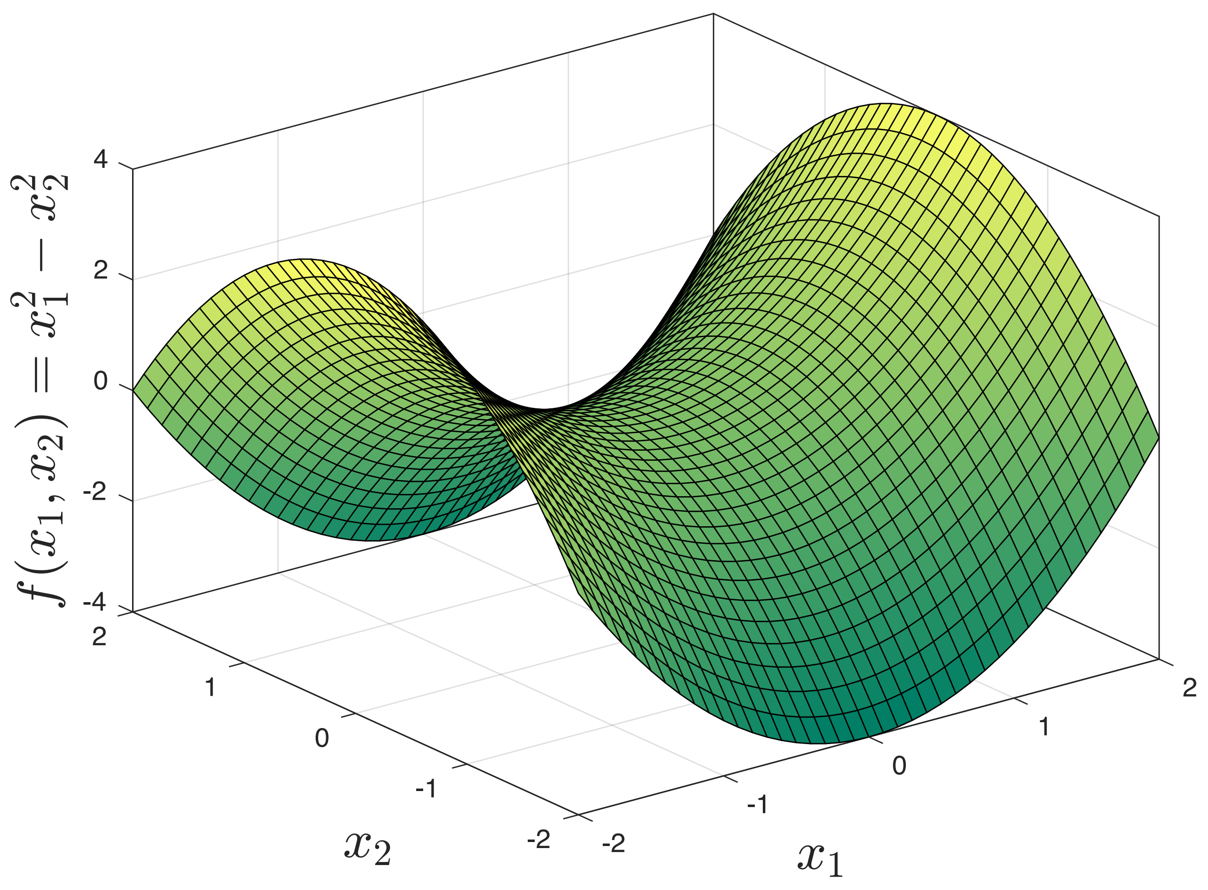
$\nabla^2 f(\cdot) \succ 0$



$\nabla^2 f(\cdot) \prec 0$



$\nabla^2 f(\cdot) \succeq 0$

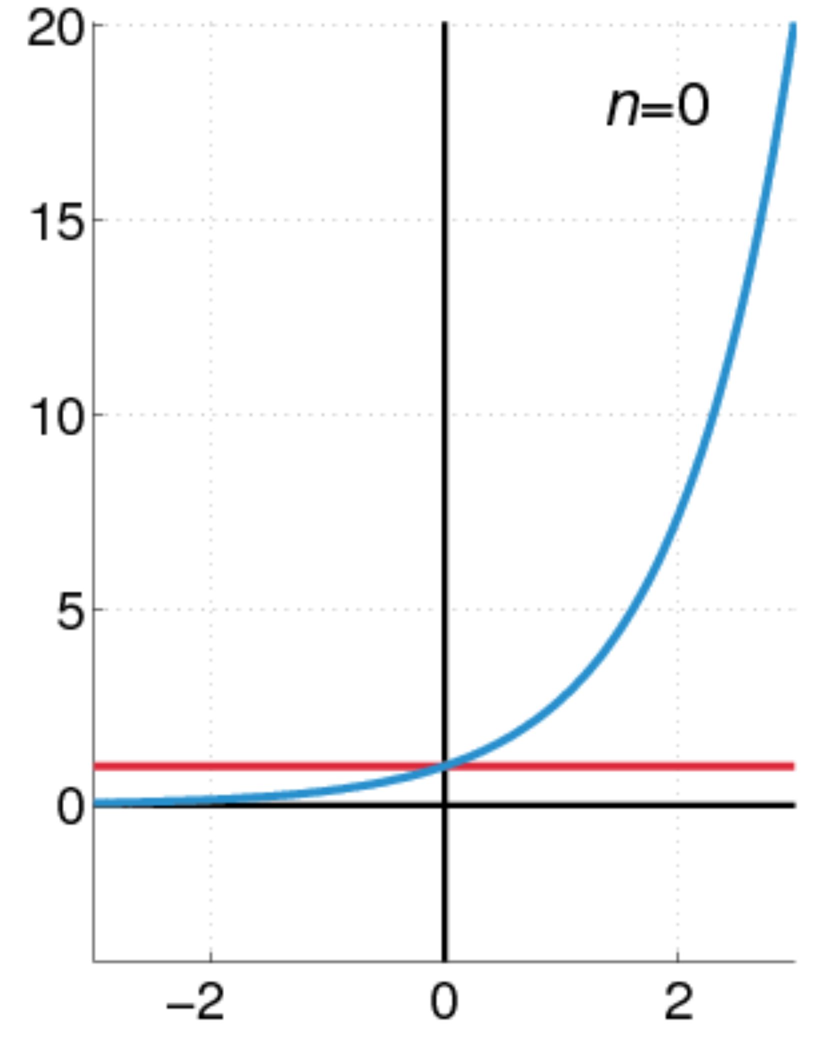


Indefinite

Taylor's expansion

– **Taylor's** expansion: used for (locally) approximating a function

$$f(x)\Big|_{x=\alpha} = f(\alpha) + f'(\alpha)(x-\alpha) + \frac{f''(a)}{2!}(x-\alpha)^2 + \cdots + \frac{f^{(n)}(\alpha)}{n!}(x-\alpha)^n + R_n$$

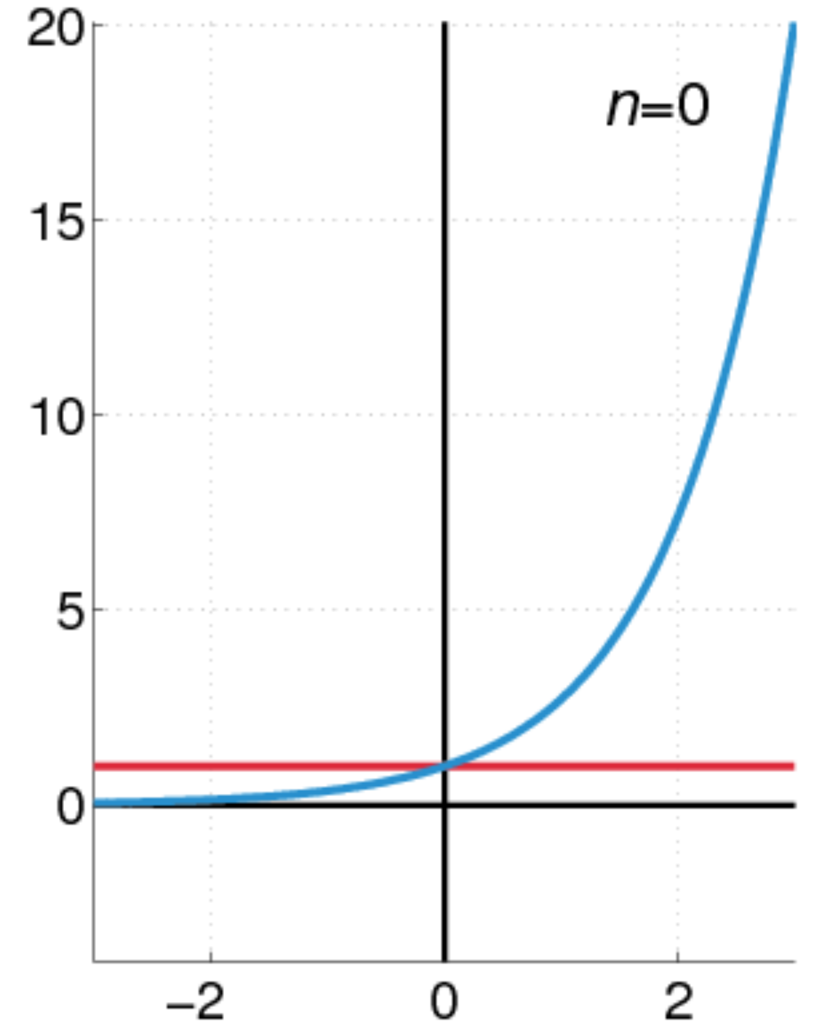


Example: exponential function

Taylor's expansion

– **Taylor's** expansion: used for (locally) approximating a function

$$f(x)\Big|_{x=\alpha} = f(\alpha) + f'(\alpha)(x-\alpha) + \frac{f''(a)}{2!}(x-\alpha)^2 + \cdots + \frac{f^{(n)}(\alpha)}{n!}(x-\alpha)^n + R_n$$

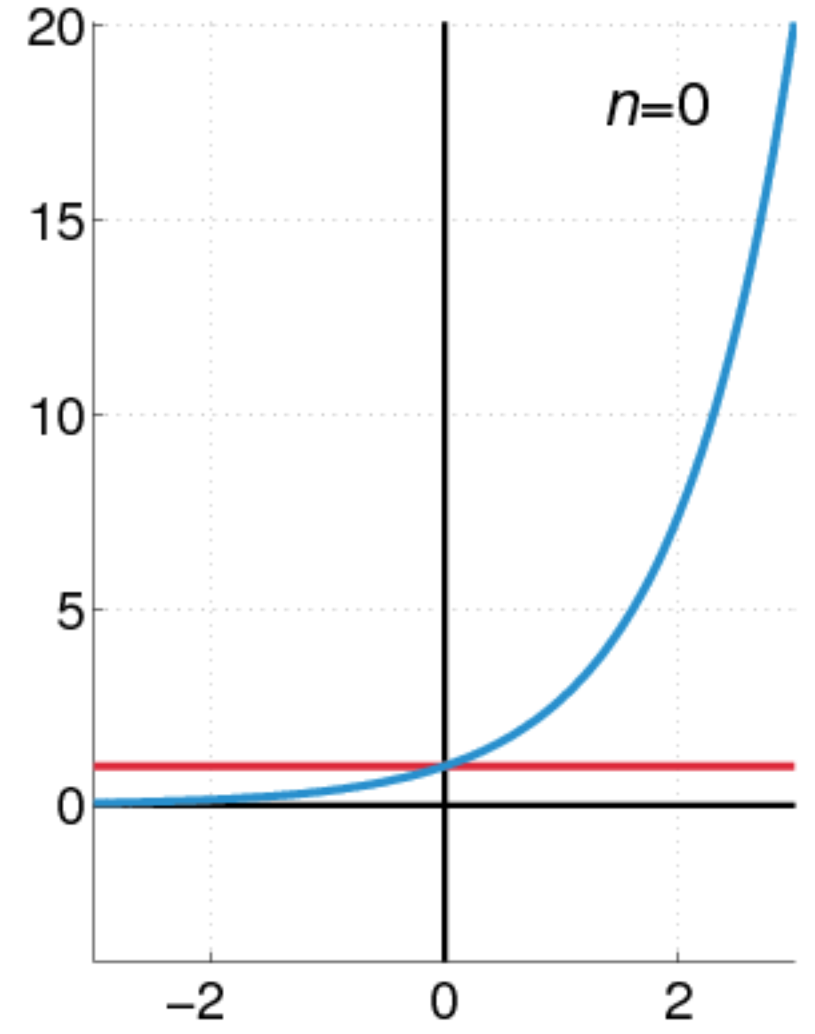


Example: exponential function

Taylor's expansion

– **Taylor's** expansion: used for (locally) approximating a function

$$f(x)\Big|_{x=\alpha} = f(\alpha) + f'(\alpha)(x-\alpha) + \frac{f''(\alpha)}{2!}(x-\alpha)^2 + \cdots + \frac{f^{(n)}(\alpha)}{n!}(x-\alpha)^n + R_n$$



Example: exponential function

– Key properties/assumptions:
  – Function f is differentiable as many times we'd like
  – Provides **(locally)** a good approximation of the function

Taylor's expansion

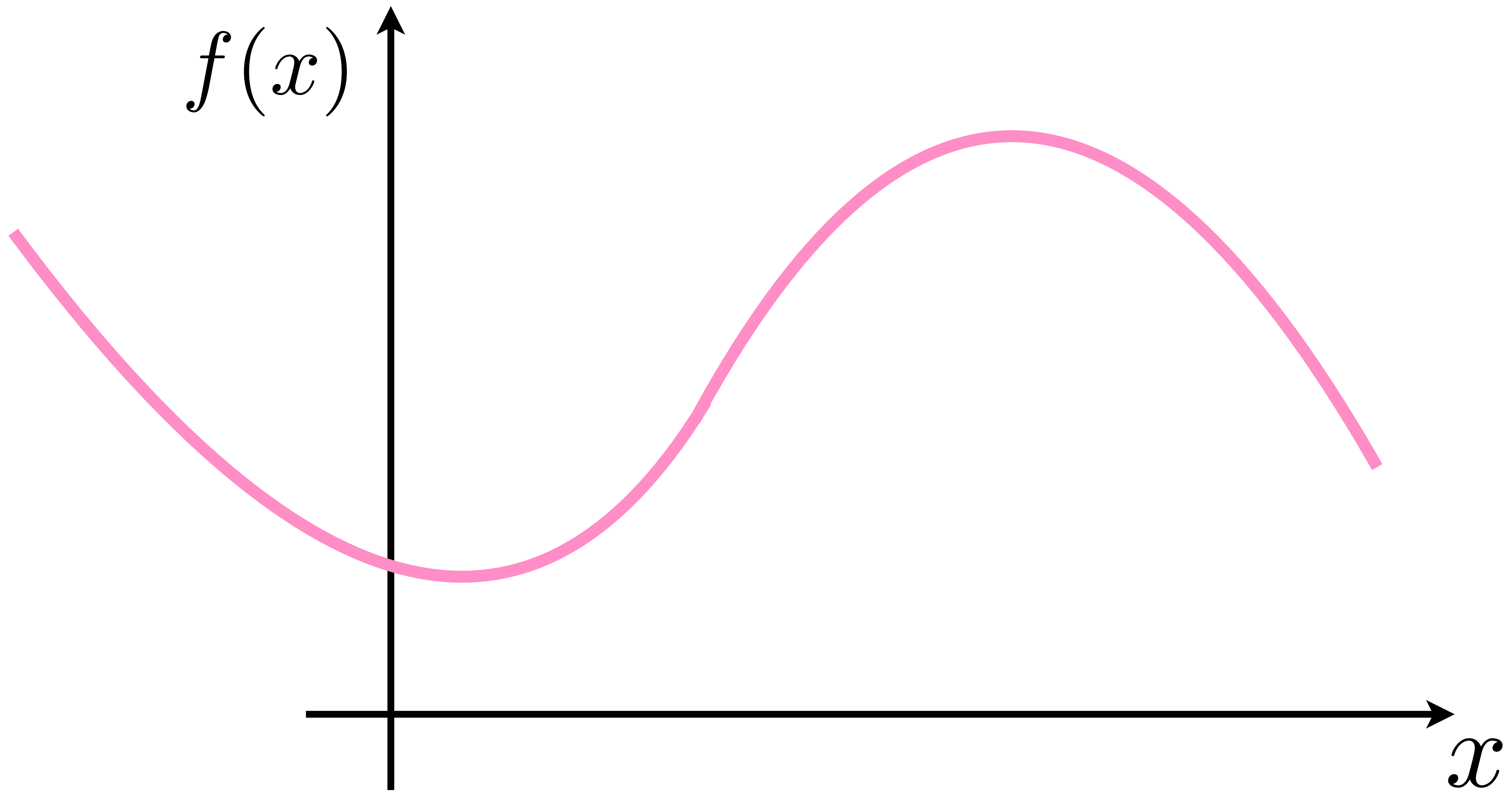
- First-order Taylor's approximation

$$f : \mathbb{R}^p \to \mathbb{R} \quad \Bigg| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$

Taylor's expansion

- First-order Taylor's approximation

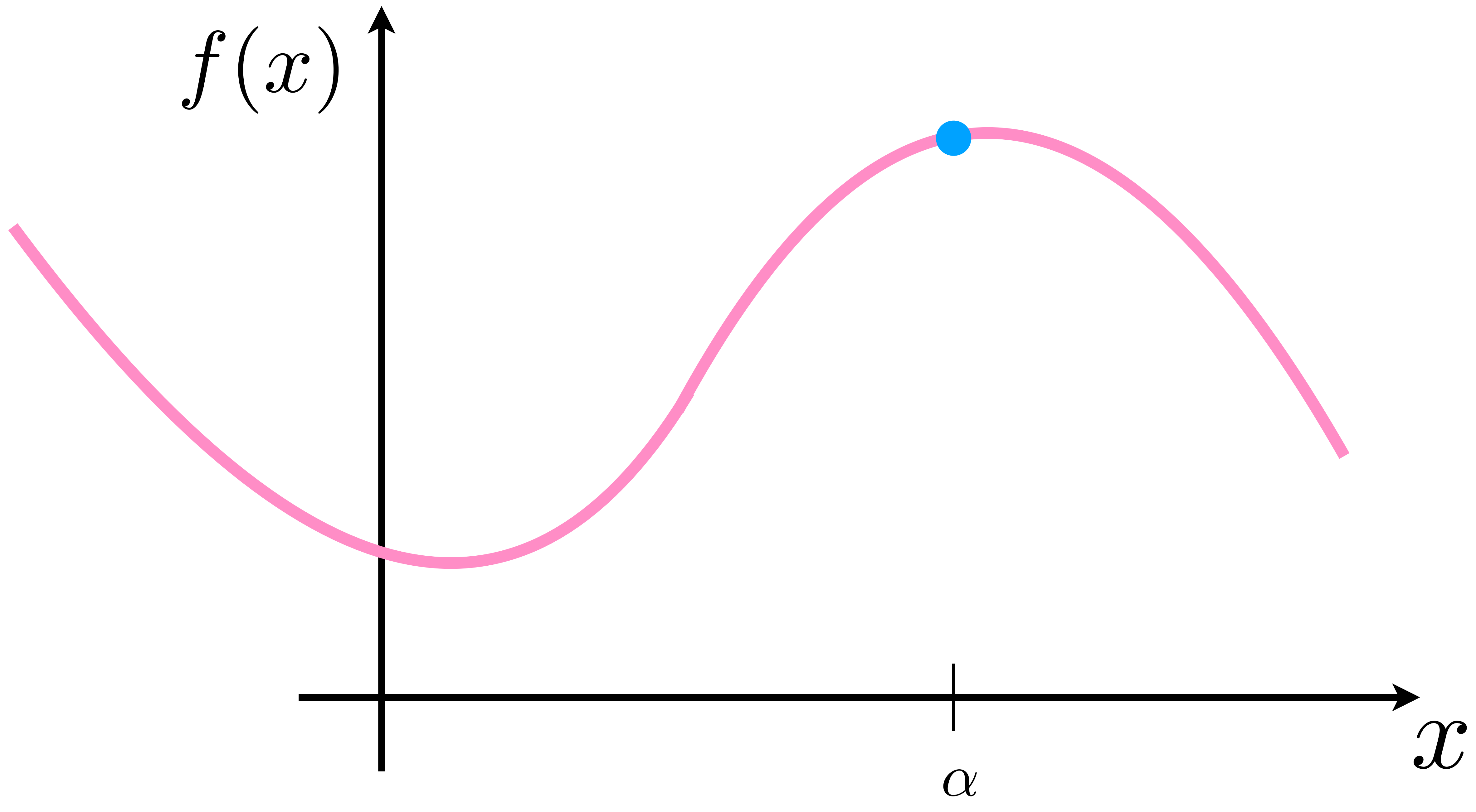
$$f : \mathbb{R}^p \to \mathbb{R} \qquad \Big| \qquad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

– First–order Taylor's approximation

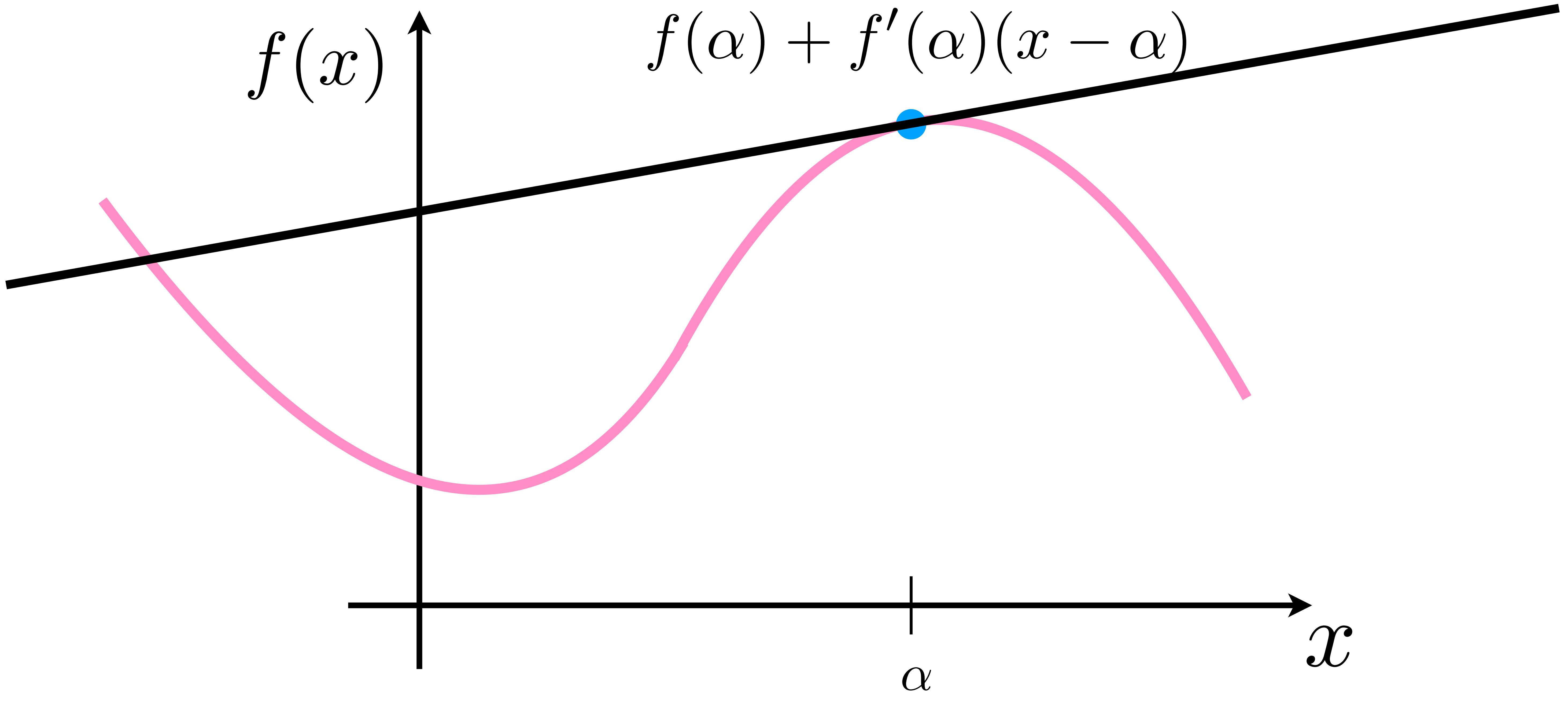
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle , \alpha \in \mathbb{R}^p$$



Taylor's expansion

- First−order Taylor's approximation

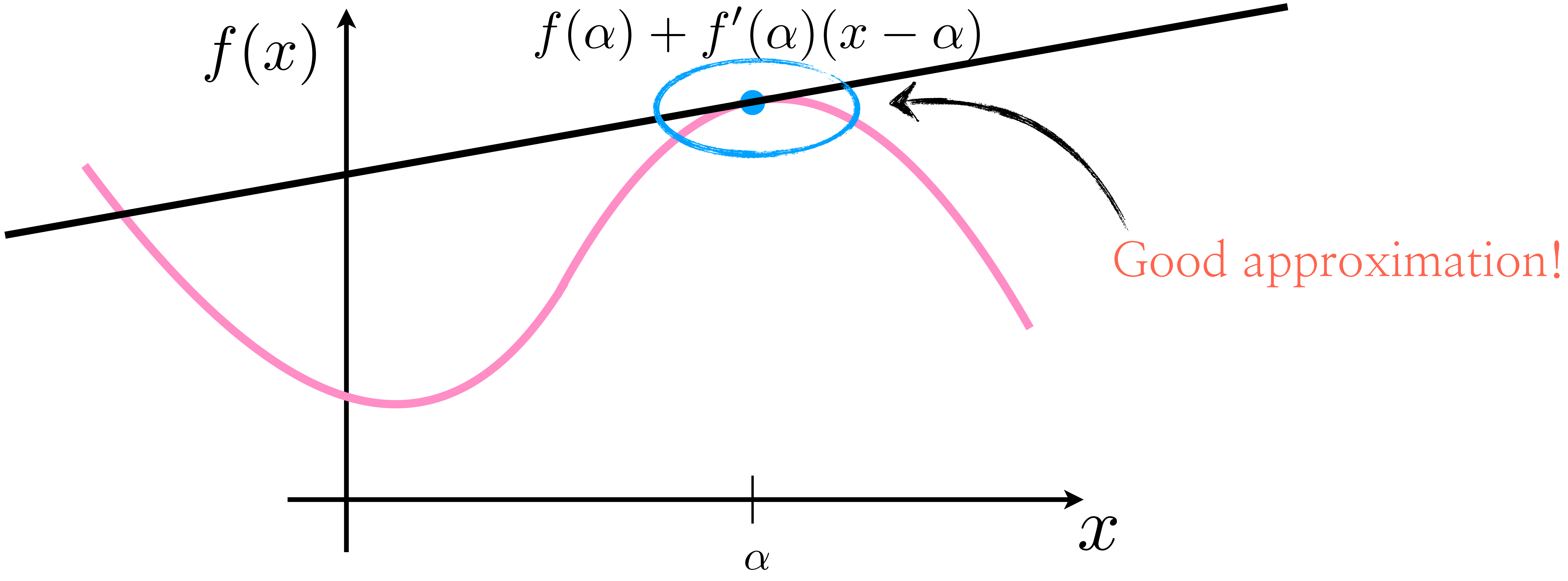
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$



Taylor's expansion

- First-order Taylor's approximation

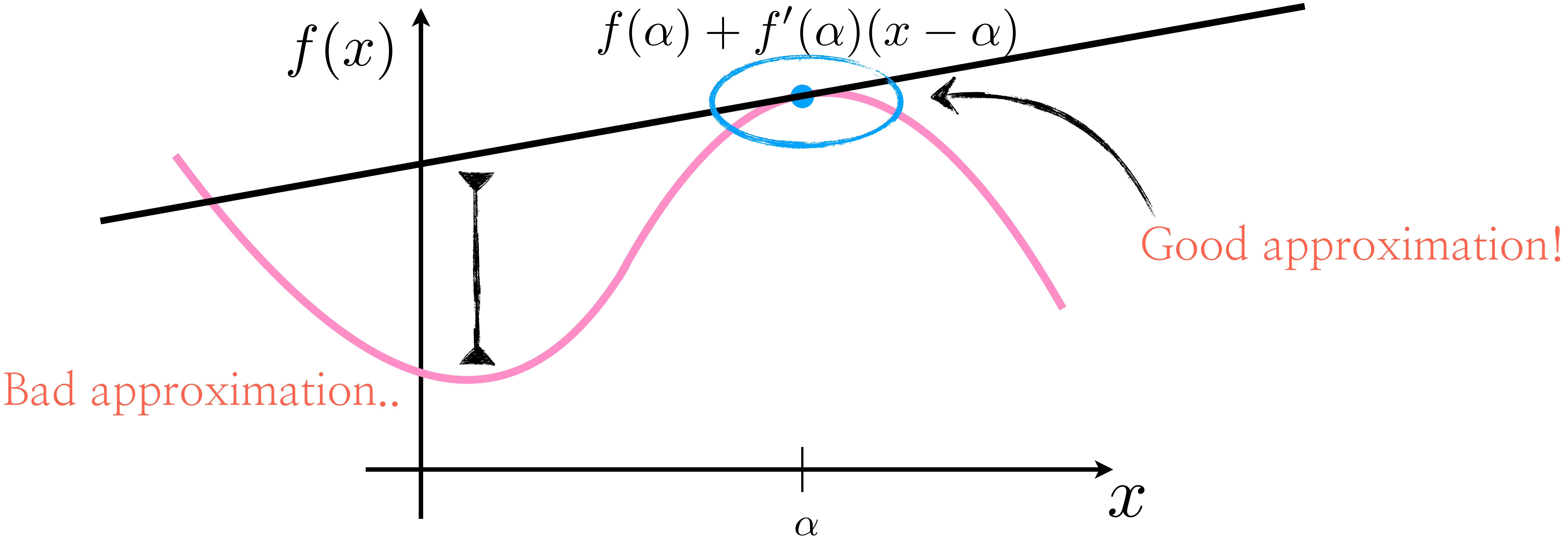
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

– First–order Taylor's approximation

$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

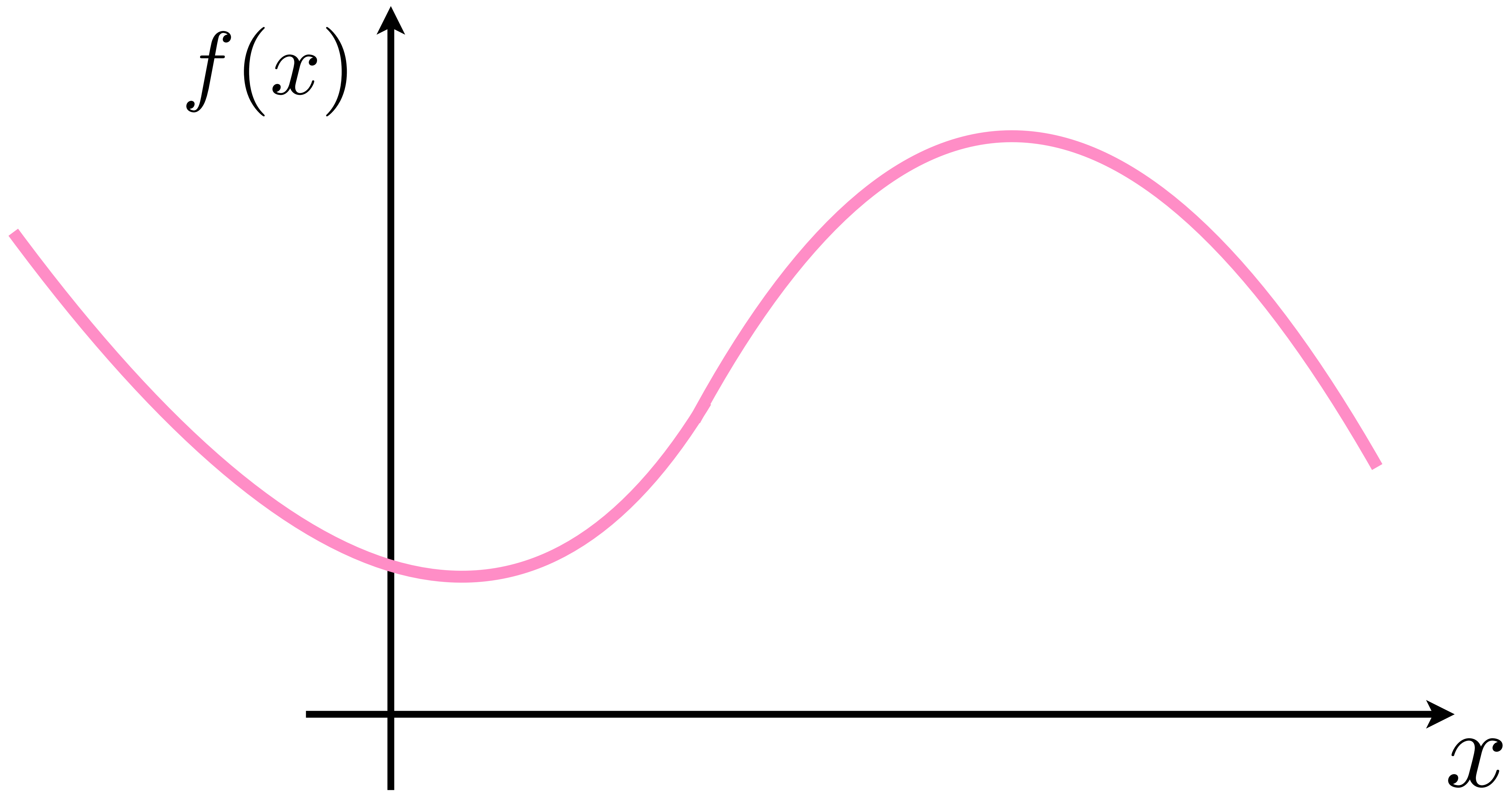
- Second-order Taylor's approximation

$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \frac{1}{2} \left\langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \right\rangle, \alpha \in \mathbb{R}^p$$

Taylor's expansion

- Second-order Taylor's approximation

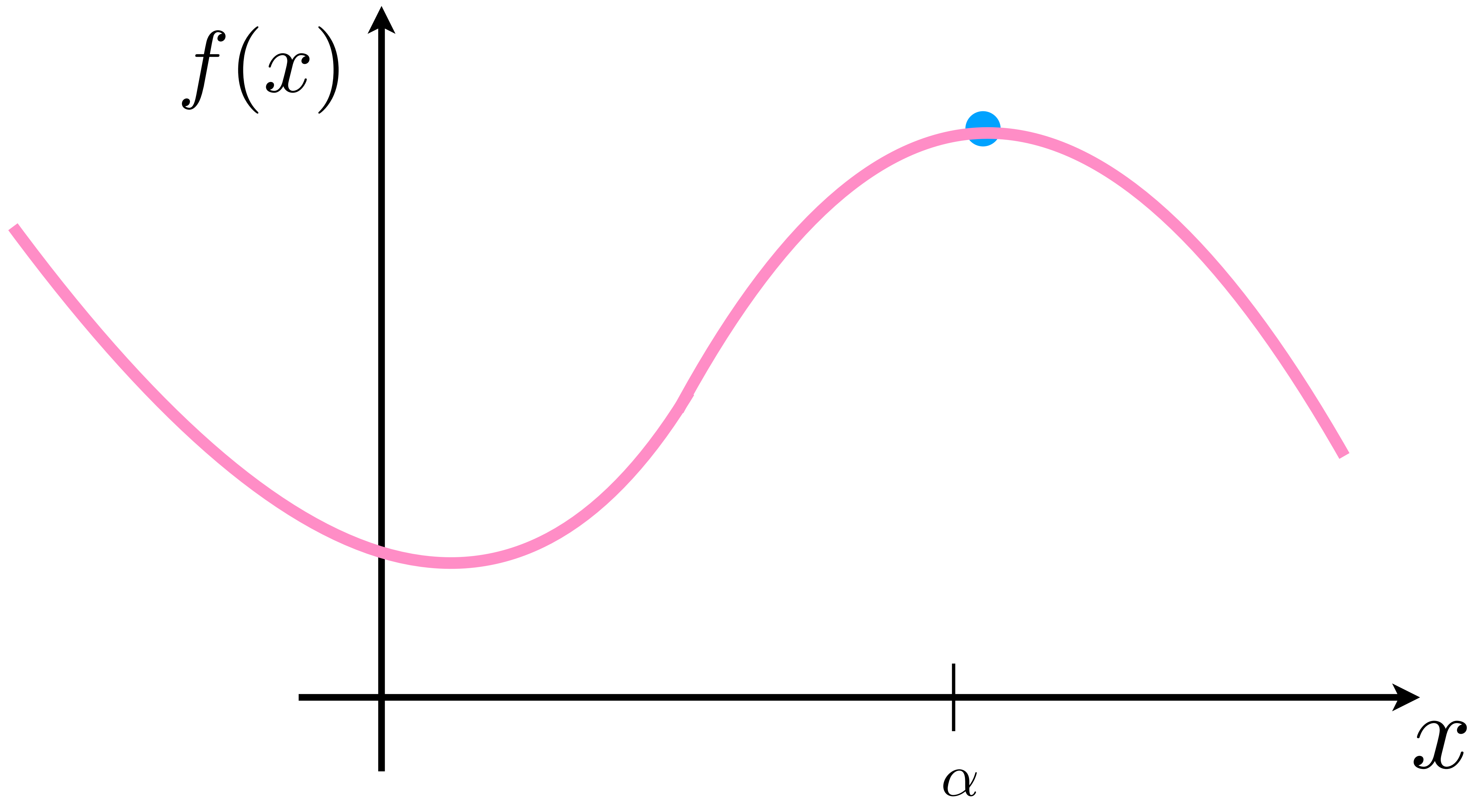
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \frac{1}{2} \left\langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \right\rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

– Second–order Taylor's approximation

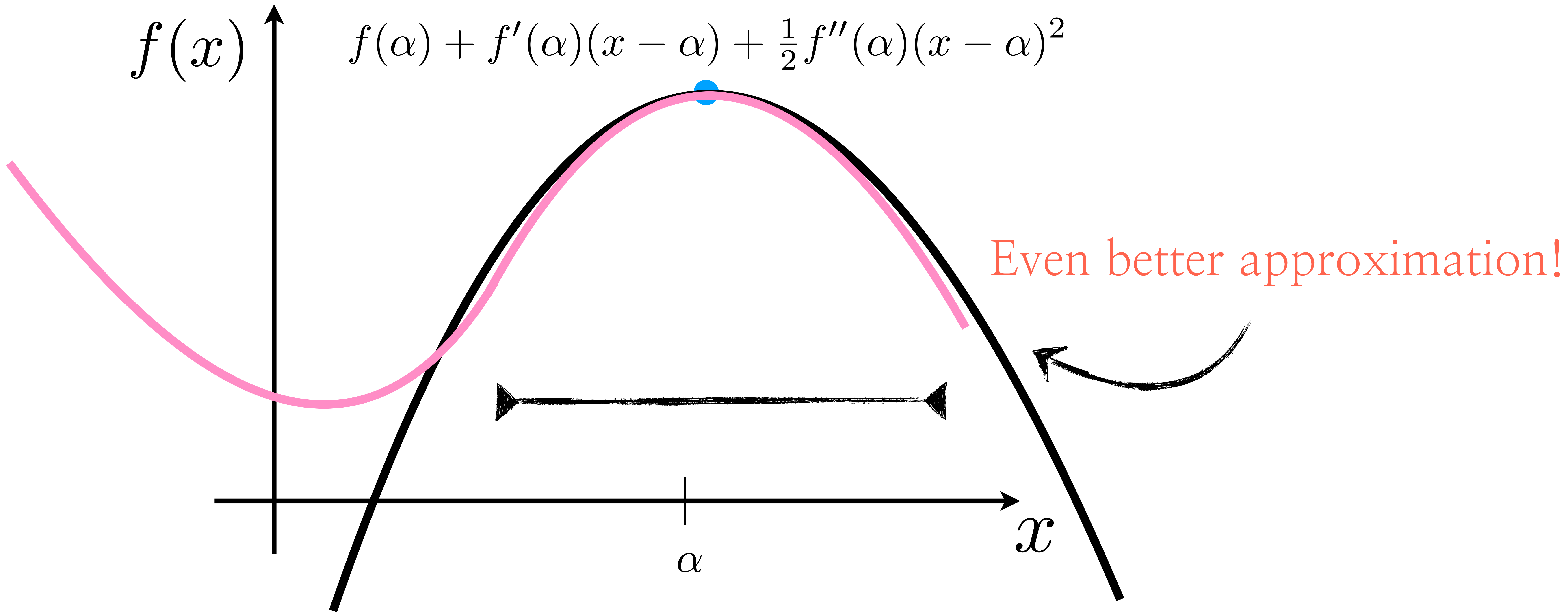
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \frac{1}{2} \langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

- Second–order Taylor's approximation

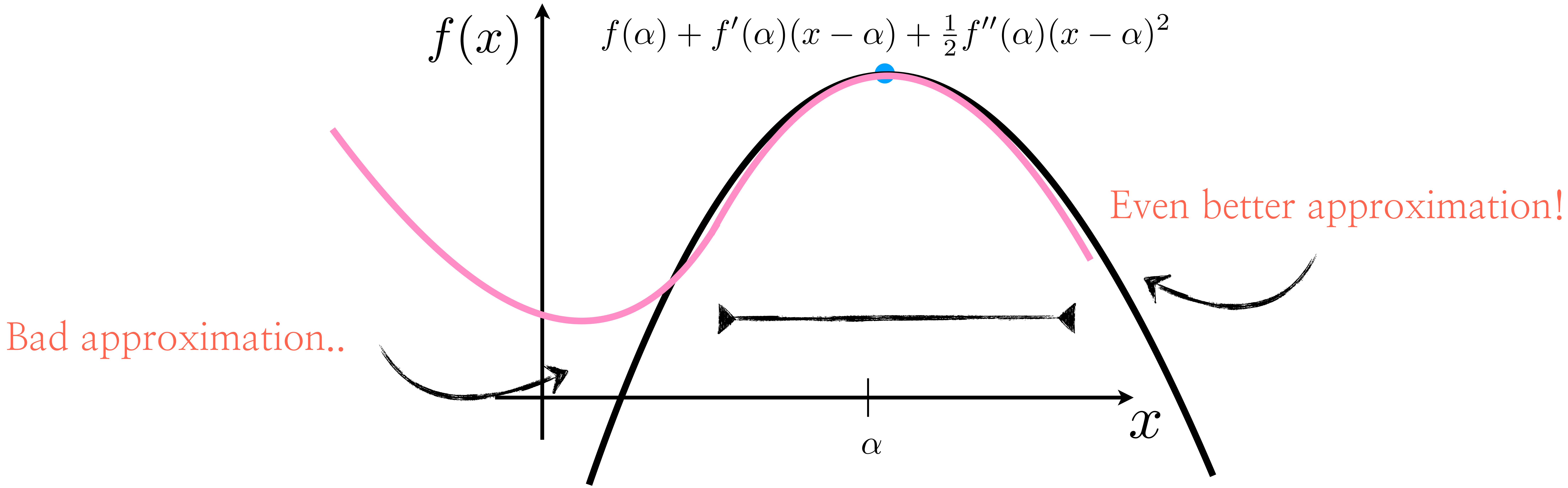
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \tfrac{1}{2} \langle \nabla^2 f(\alpha)(x-\alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

– Second–order Taylor's approximation

$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \frac{1}{2} \langle \nabla^2 f(\alpha)(x-\alpha), x - \alpha \rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

- Spoiler alert: "Why are all these useful?"
    - Often, we optimize a function through its local approximations
    - E.g., second order approximations are.. quadratic functions!

Taylor’s expansion

- Spoiler alert: “Why are all these useful?”
    - Often, we optimize a function through its local approximations
    - E.g., second order approximations are.. quadratic functions!

$$\min_x f(x)$$

Taylor's expansion

- Spoiler alert: "Why are all these useful?"
  - Often, we optimize a function through its local approximations
  - E.g., second order approximations are.. quadratic functions!

Weirdest function ever
(but differentiable)

$$\min_x f(x)$$

Taylor's expansion

- Spoiler alert: "Why are all these useful?"
  - Often, we optimize a function through its local approximations
  - E.g., second order approximations are.. quadratic functions!

Weirdest function ever
(but differentiable)

$$\min_x f(x) \longrightarrow \min_x \left\{ f(x_0) + \nabla f(x_0)^\top (x - x_0) + \tfrac{1}{2}(x - x_0)^\top \nabla f(x_0)(x - x_0) \right\}$$

Taylor's expansion

- Spoiler alert: "Why are all these useful?"
  - Often, we optimize a function through its local approximations
  - E.g., second order approximations are.. quadratic functions!

Weirdest function ever
(but differentiable)

$$\min_x f(x) \longrightarrow \min_x \{p^\top x + \tfrac{1}{2} x^\top H x\}$$

Taylor's expansion

- Spoiler alert: "Why are all these useful?"
  - Often, we optimize a function through its local approximations
  - E.g., second order approximations are.. quadratic functions!

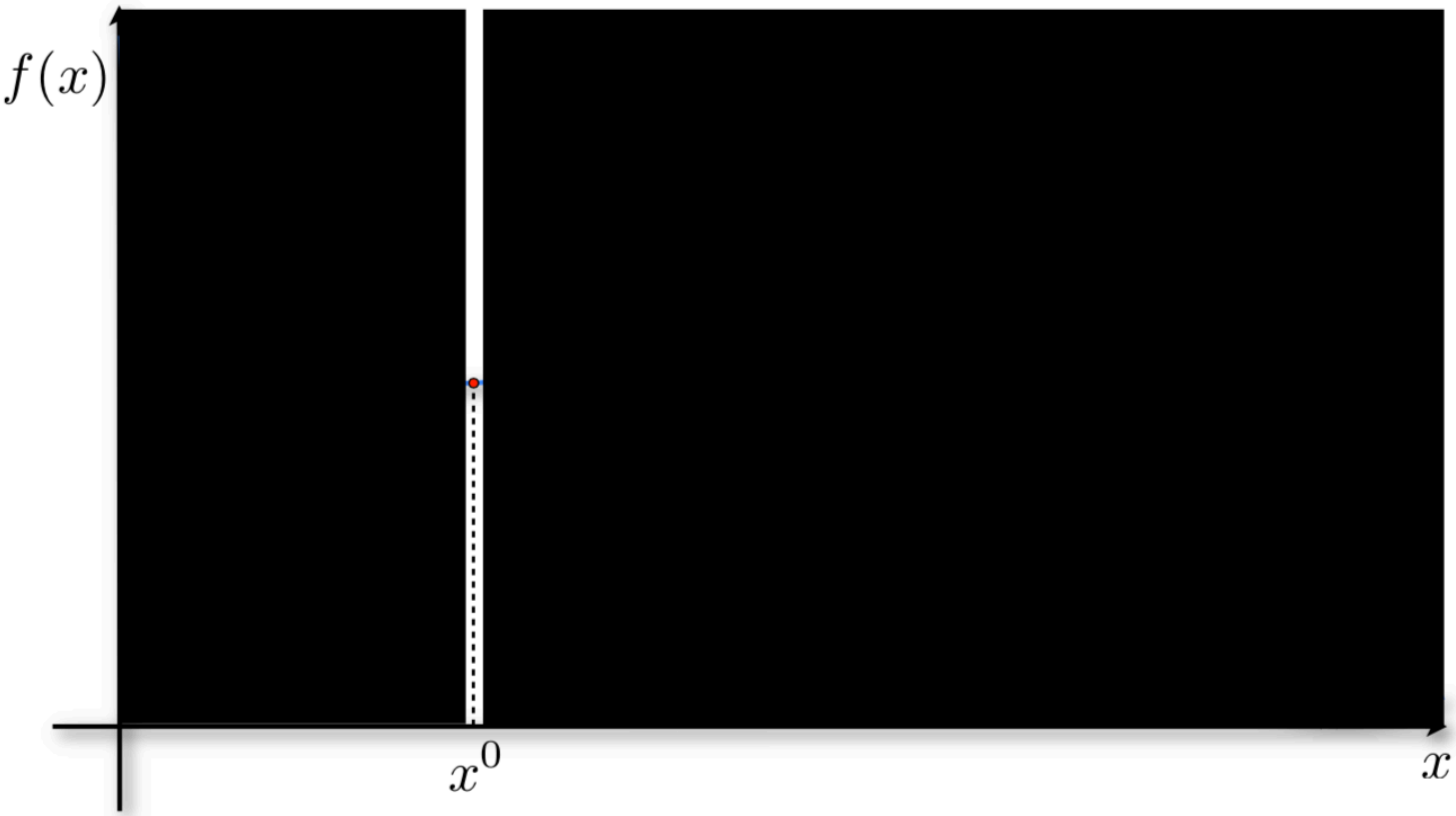
Weirdest function ever
(but differentiable)

$$\min_x f(x) \longrightarrow \min_x \{p^\top x + \tfrac{1}{2}x^\top H x\} \longrightarrow \dots$$

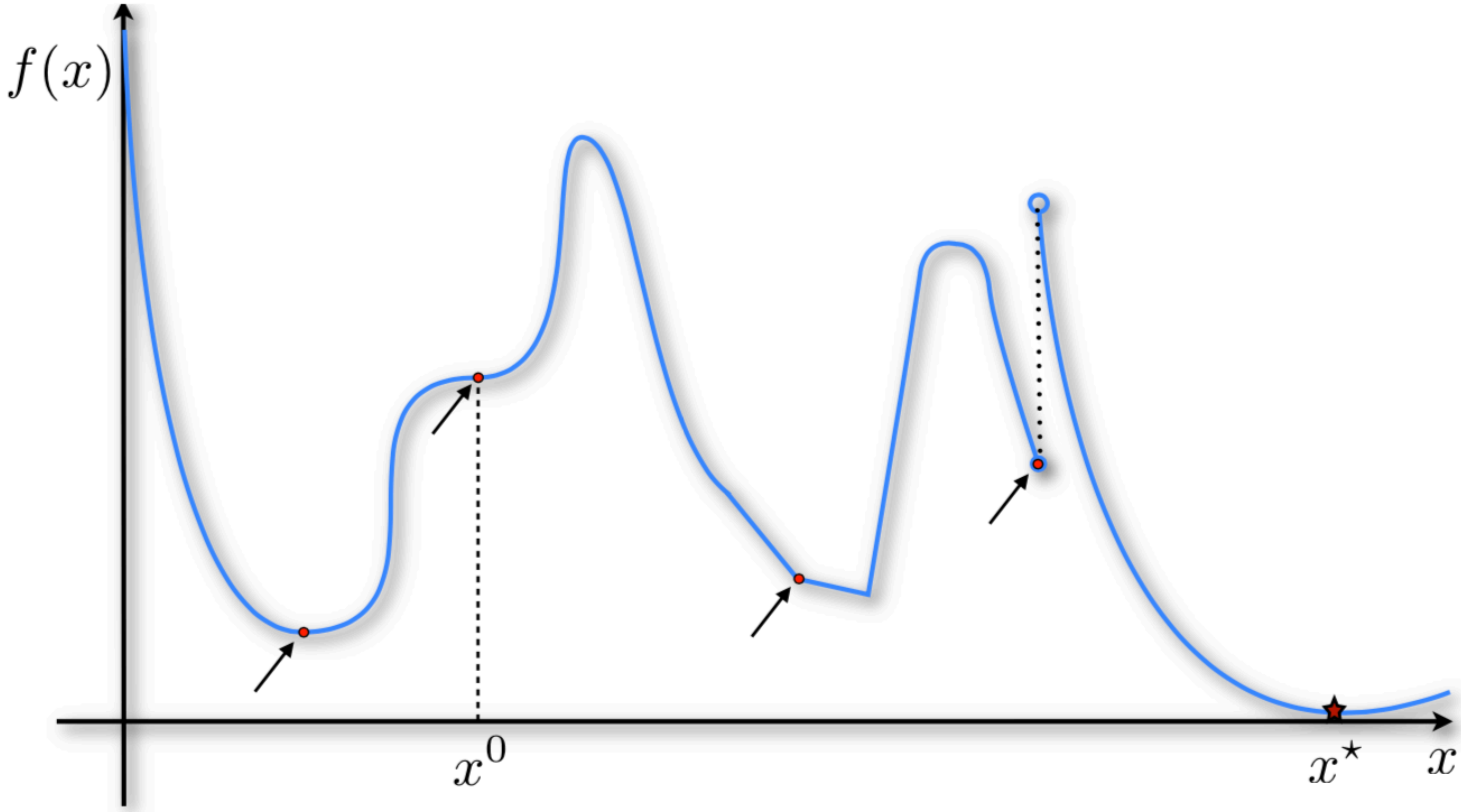
Agnostic optimization

Demo

Agnostic optimization



Agnostic optimization



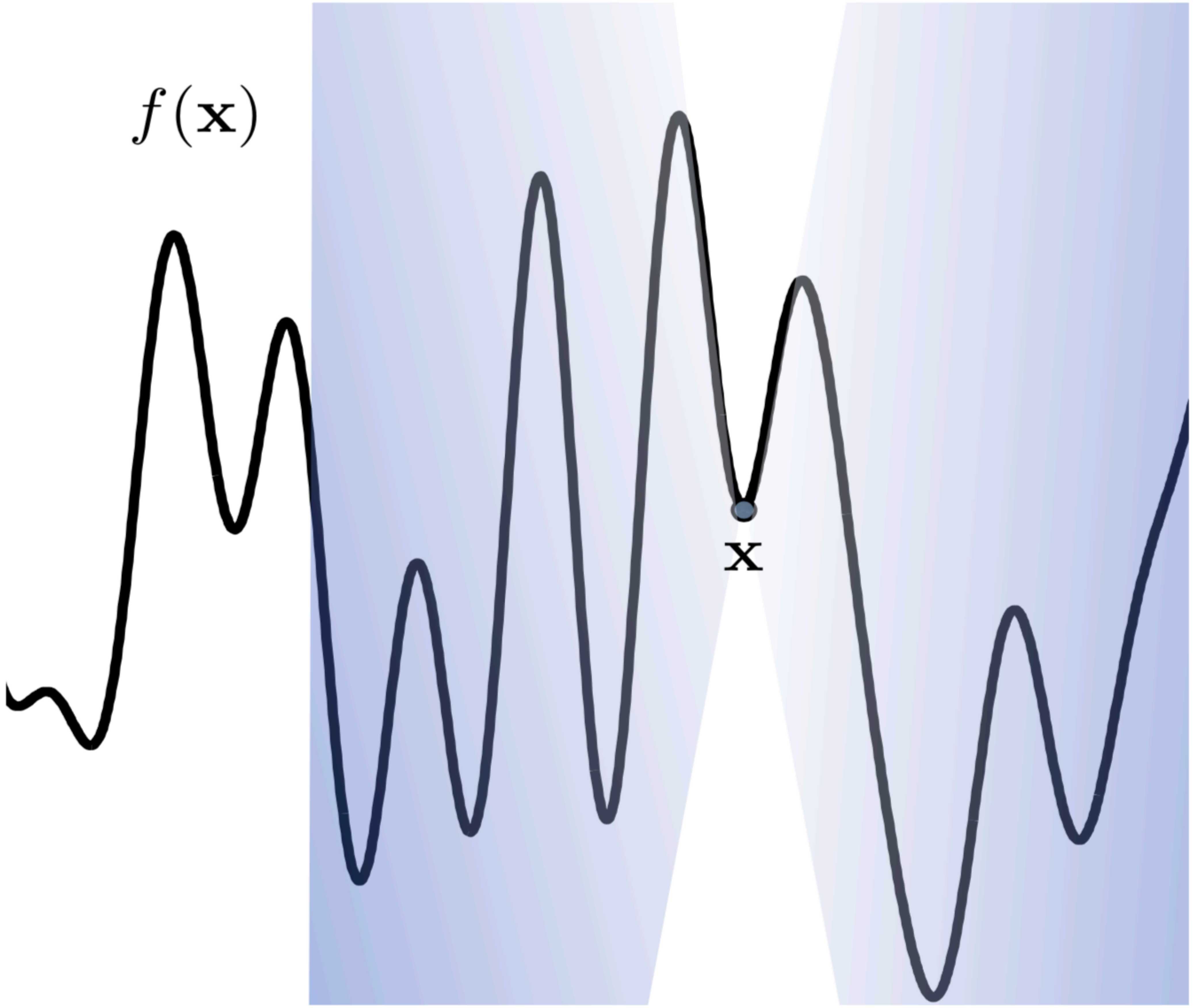
Lipschitz conditions

- **Lipschitz continuity**: $|f(x) - f(y)| \leq M\|x - y\|_2, \quad \forall x, y$

Lipschitz conditions

- **Lipschitz continuity**: $|f(x) - f(y)| \leq M\|x - y\|_2, \quad \forall x, y$

Interpretation via double cone



Function examples:

1. Absolute value
2. Trigonometric functions
3. Quadratics (..)

Lipschitz conditions

- **Lipschitz gradient continuity**: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

Lipschitz conditions

- **Lipschitz gradient continuity**: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$
  - Intuition + comparison with Lipschitz continuity:

    "Lipschitz continuity implies that f should not be too steep"

    "Lipschitz gradient continuity implies that changes in the slope of f should not happen suddenly"

Lipschitz conditions

- **Lipschitz gradient continuity**: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$
  - Intuition + comparison with Lipschitz continuity:

    "Lipschitz continuity implies that f should not be too steep"

    "Lipschitz gradient continuity implies that changes in the slope of f should not happen suddenly"

  - Example: Quadratics are not **globally** Lipschitz continuous

    (the function becomes arbitrarily steep as we approach infinity)

    but: $\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$

    for: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

Lipschitz conditions

- **Lipschitz gradient continuity**: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$
  - Intuition + comparison with Lipschitz continuity:

    "Lipschitz continuity implies that f should not be too steep"

    "Lipschitz gradient continuity implies that changes in the slope of f should not happen suddenly"

  - Example: Quadratics are not **globally** Lipschitz continuous

    (the function becomes arbitrarily steep as we approach infinity)

but: $\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$

for: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

Largest singular value

Lipschitz conditions

– Equivalent characterizations： $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2, \quad \forall x, y$

$$
\begin{aligned}
f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{L}{2}\|x-y\|_2^2 \\
f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\
\langle \nabla f(x) - \nabla f(y), x - y\rangle &\leq L\|x-y\|_2^2 \\
&\vdots \quad \vdots
\end{aligned}
$$

Lipschitz conditions

– Equivalent characterizations： $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

$$
\begin{aligned}
f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{L}{2}\|x - y\|_2^2 \\
f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\
\langle \nabla f(x) - \nabla f(y), x - y\rangle &\leq L\|x - y\|_2^2 \\
\vdots \quad & \quad \vdots
\end{aligned}
$$

– Another important one：

$$\nabla^2 f(x) \preceq LI$$

Lipschitz conditions

- Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{L}{2}\|x - y\|_2^2$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x - y\|_2^2$$

$$\vdots \qquad \vdots$$

- Another important one:

$$\nabla^2 f(x) \preceq LI$$

Interpretation?

Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

$$\boxed{f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|_2^2$$

$$\vdots \qquad \vdots$$

– Another important one:

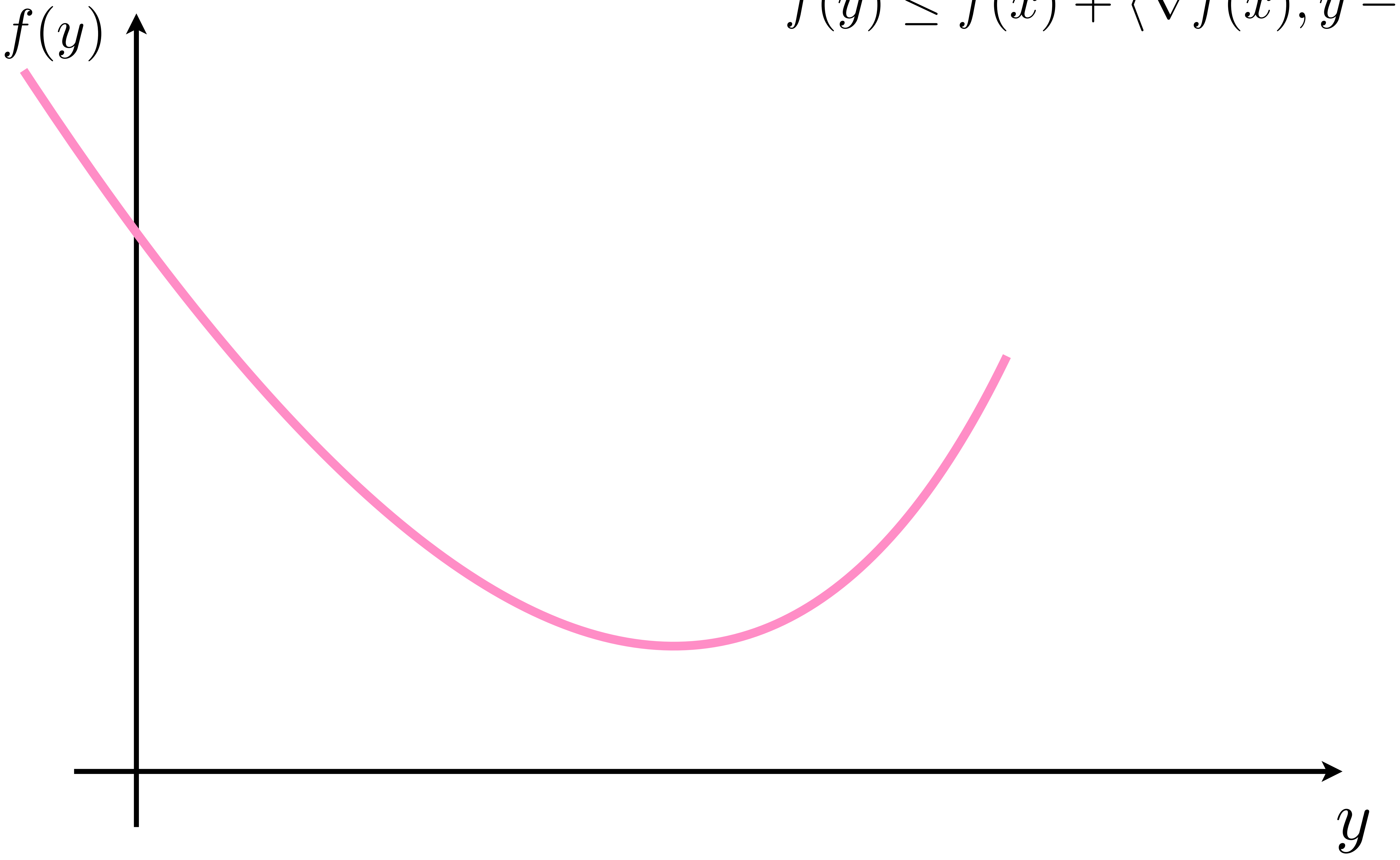
$$\nabla^2 f(x) \preceq LI$$

Interpretation?

Lipschitz conditions

- Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

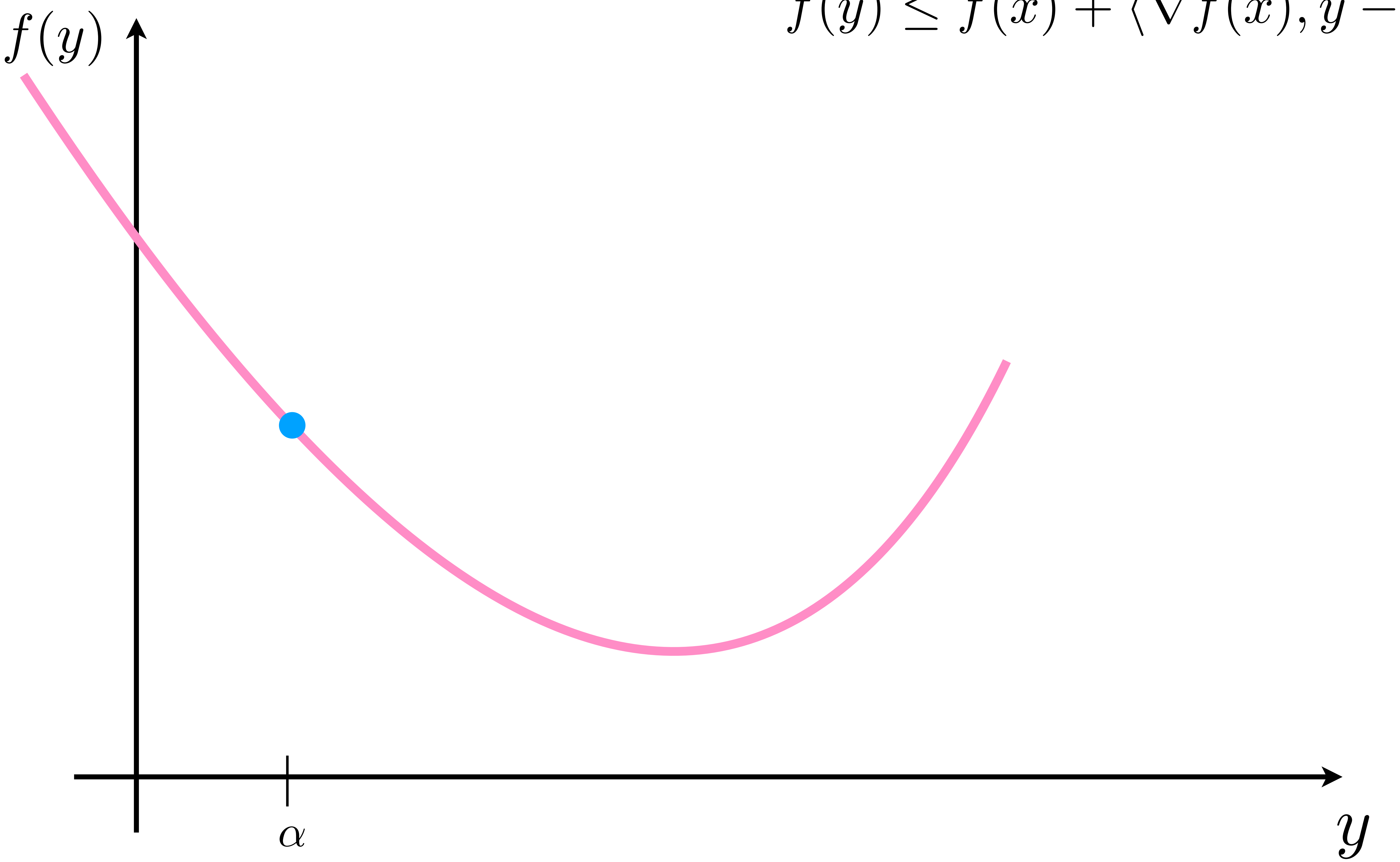
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2, \quad \forall x, y$

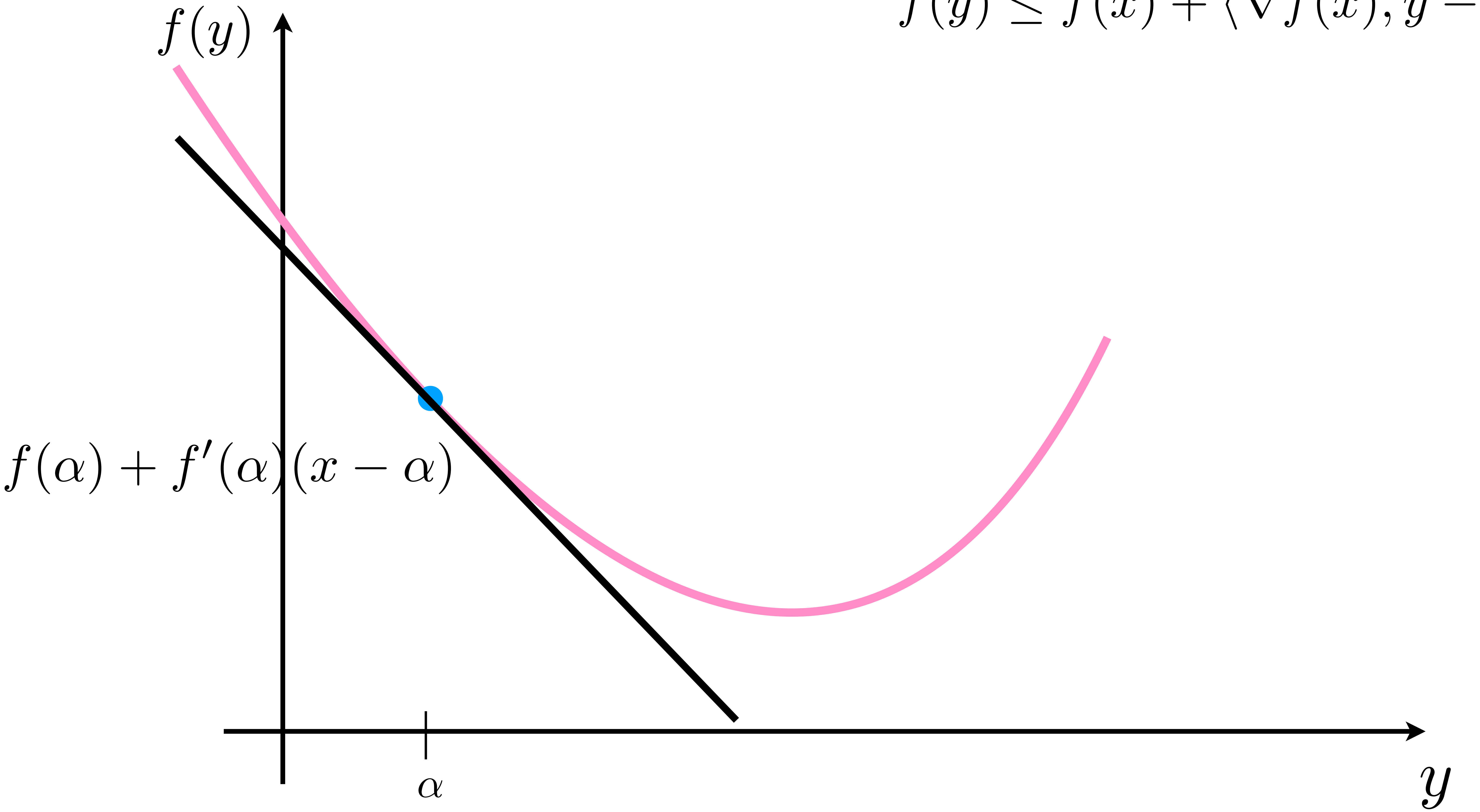
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x-y\|_2^2$$



Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

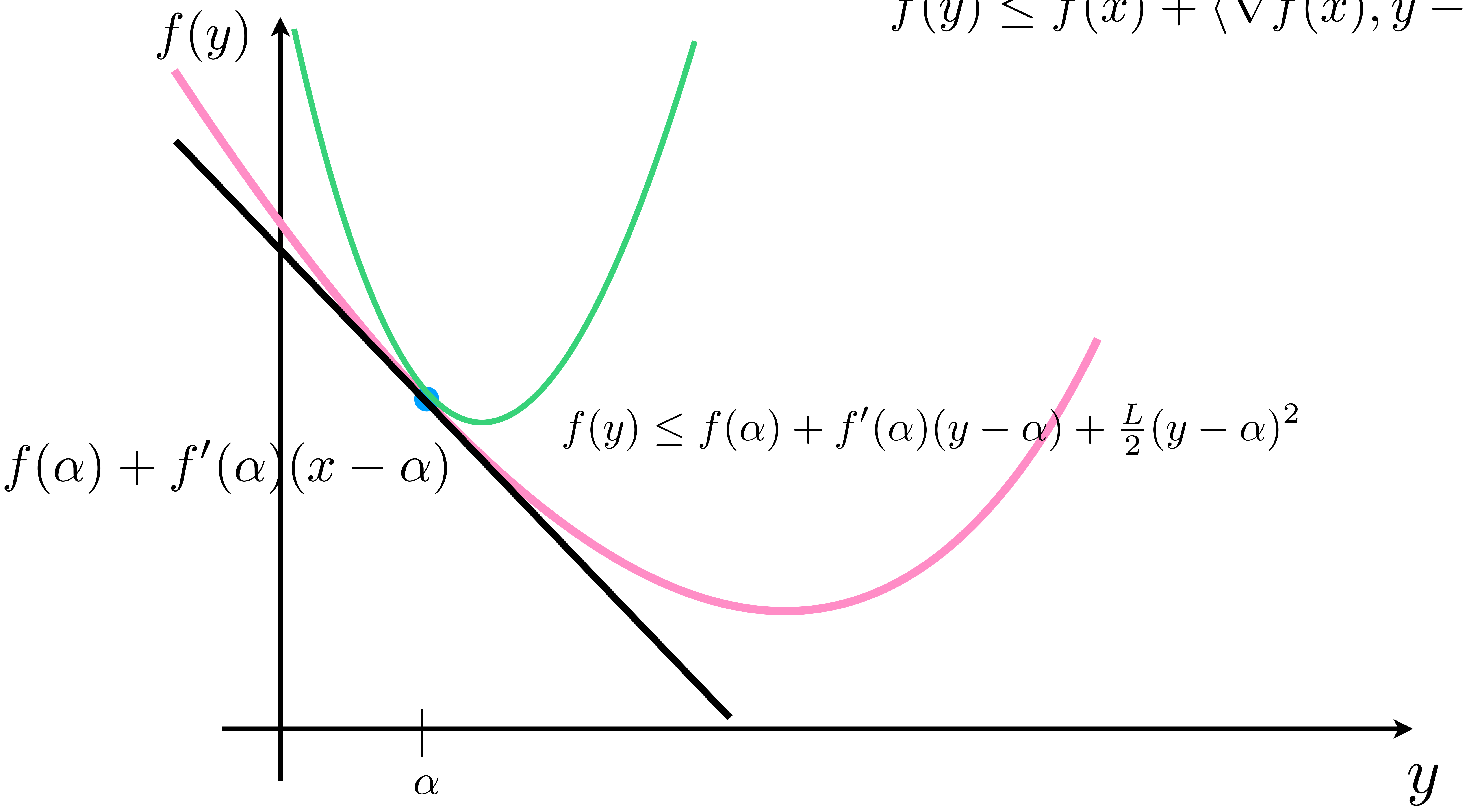
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

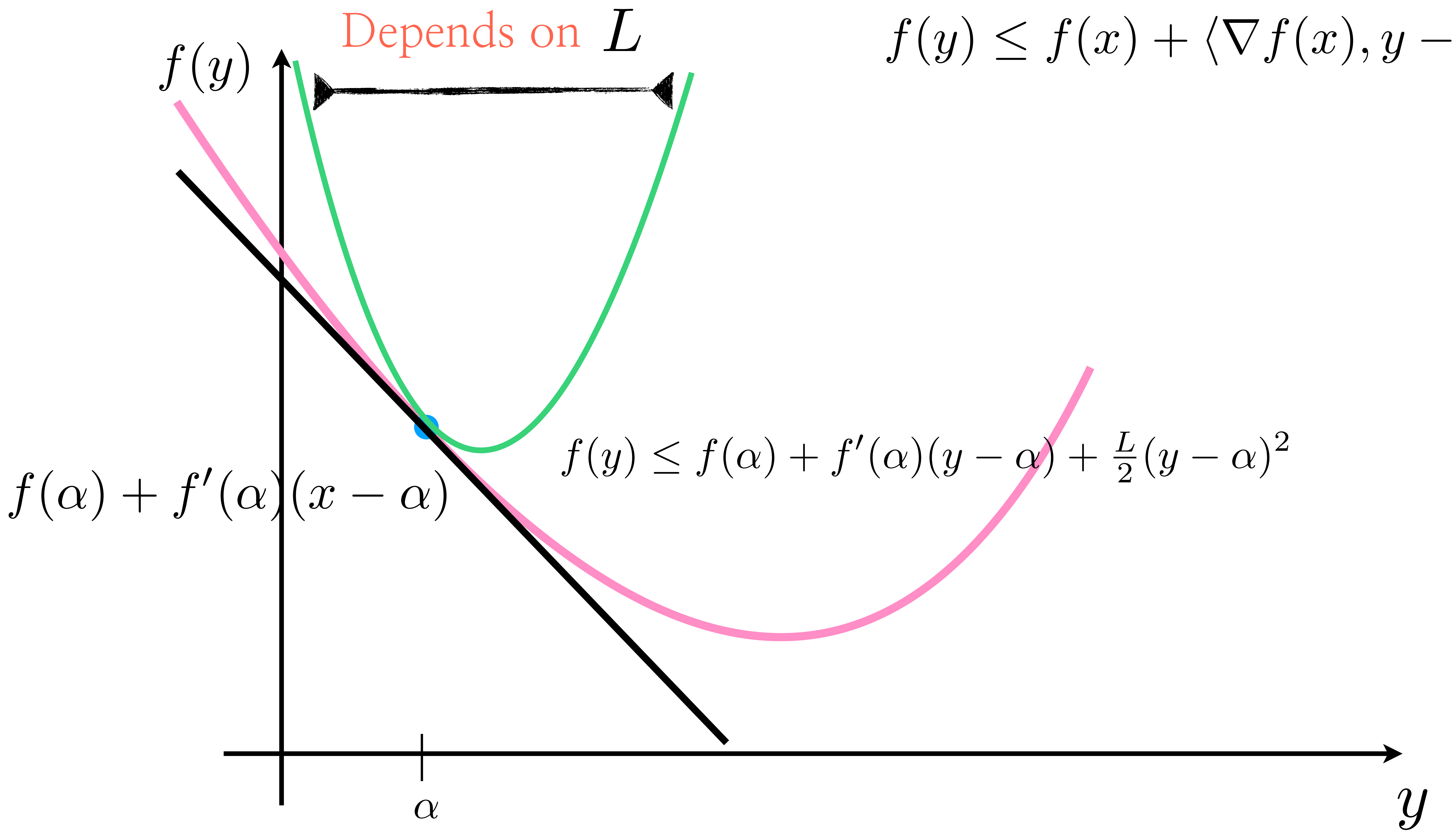
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

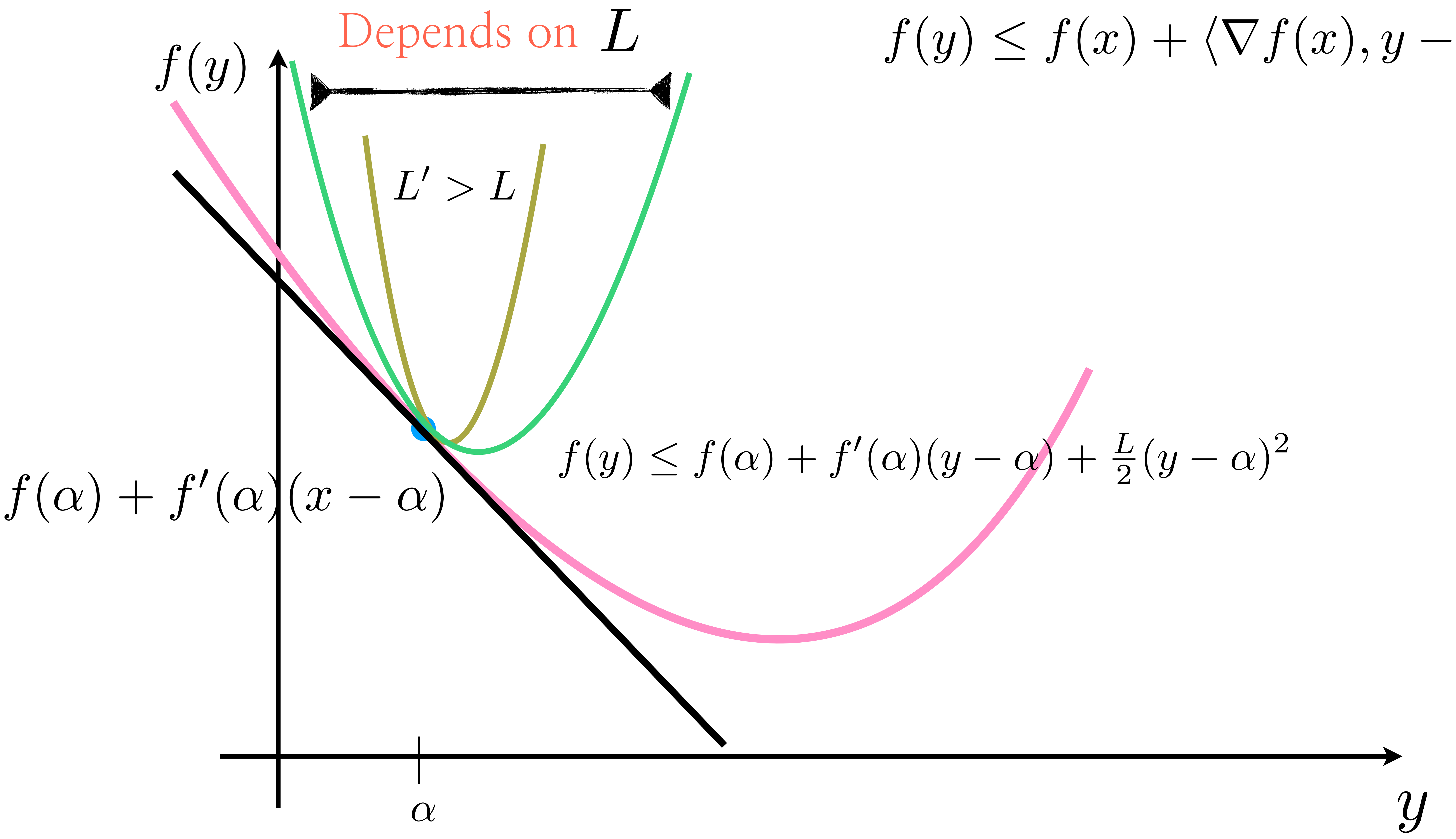
$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$



Lipschitz conditions

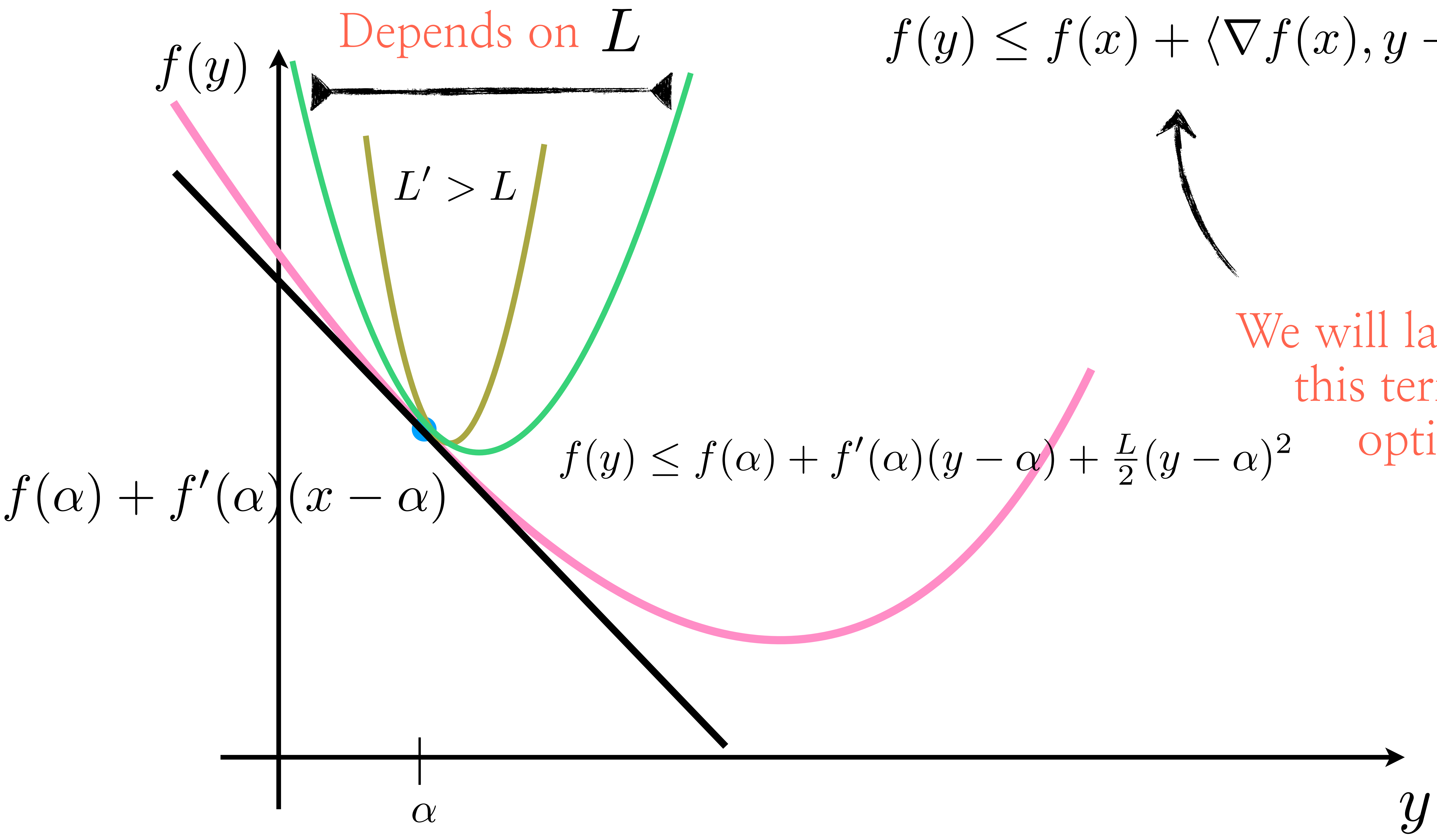
– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

– Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2, \quad \forall x, y$



$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x-y\|_2^2$

We will later on see how this term appear in optimization

Lipschitz conditions

- How does this relate to Taylor's expansion?

Lipschitz conditions

- How does this relate to Taylor's expansion?

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(z)(y - x), y - x \right\rangle$$

Lipschitz conditions

- How does this relate to Taylor's expansion?

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(z)(y - x), y - x \right\rangle$$

- From $\nabla^2 f(x) \preceq LI$, we have:

$$\nabla^2 f(x) \preceq LI \Rightarrow \|\nabla^2 f(x)\|_2 \leq \|LI\|_2 \Rightarrow \|\nabla^2 f(x)\|_2 \leq L$$

Lipschitz conditions

– How does this relate to Taylor's expansion?

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(z)(y - x), y - x \right\rangle$$

– From $\nabla^2 f(x) \preceq LI$, we have:

$$\nabla^2 f(x) \preceq LI \Rightarrow \|\nabla^2 f(x)\|_2 \leq \|LI\|_2 \Rightarrow \|\nabla^2 f(x)\|_2 \leq L$$

– Then:

$$\frac{1}{2} \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle \leq \frac{1}{2} \|\nabla^2 f(x)(y - x)\|_2 \cdot \|y - x\|_2 \leq \|\nabla^2 f(x)\|_2 \|y - x\|_2^2$$

Lipschitz conditions

- How does this relate to Taylor's expansion?

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(z)(y - x), y - x \right\rangle$$

- From $\nabla^2 f(x) \preceq LI$, we have:

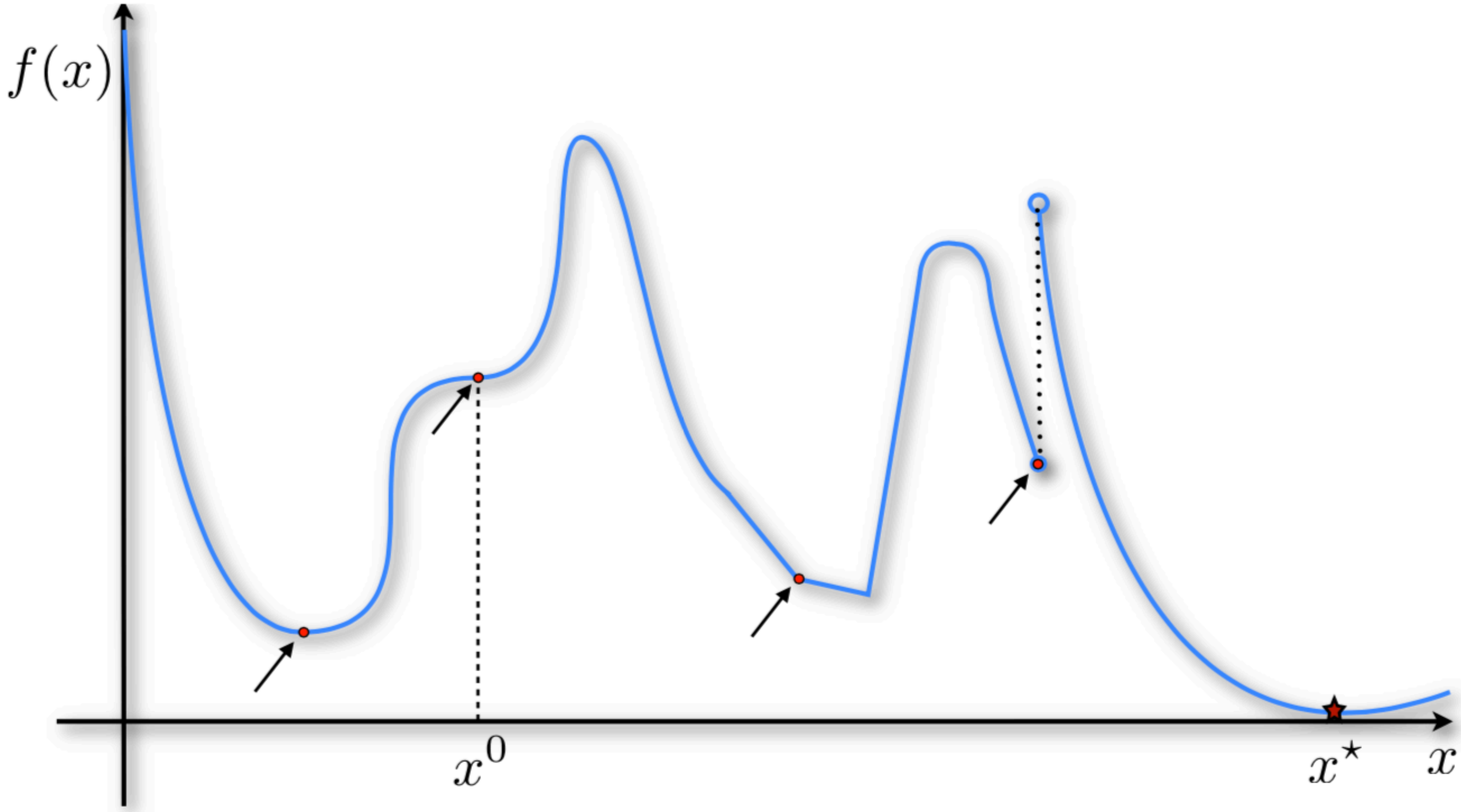
$$\nabla^2 f(x) \preceq LI \Rightarrow \|\nabla^2 f(x)\|_2 \leq \|LI\|_2 \Rightarrow \|\nabla^2 f(x)\|_2 \leq L$$

- Then:

$$\frac{1}{2} \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle \leq \frac{1}{2} \|\nabla^2 f(x)(y - x)\|_2 \cdot \|y - x\|_2 \leq \|\nabla^2 f(x)\|_2 \|y - x\|_2^2$$

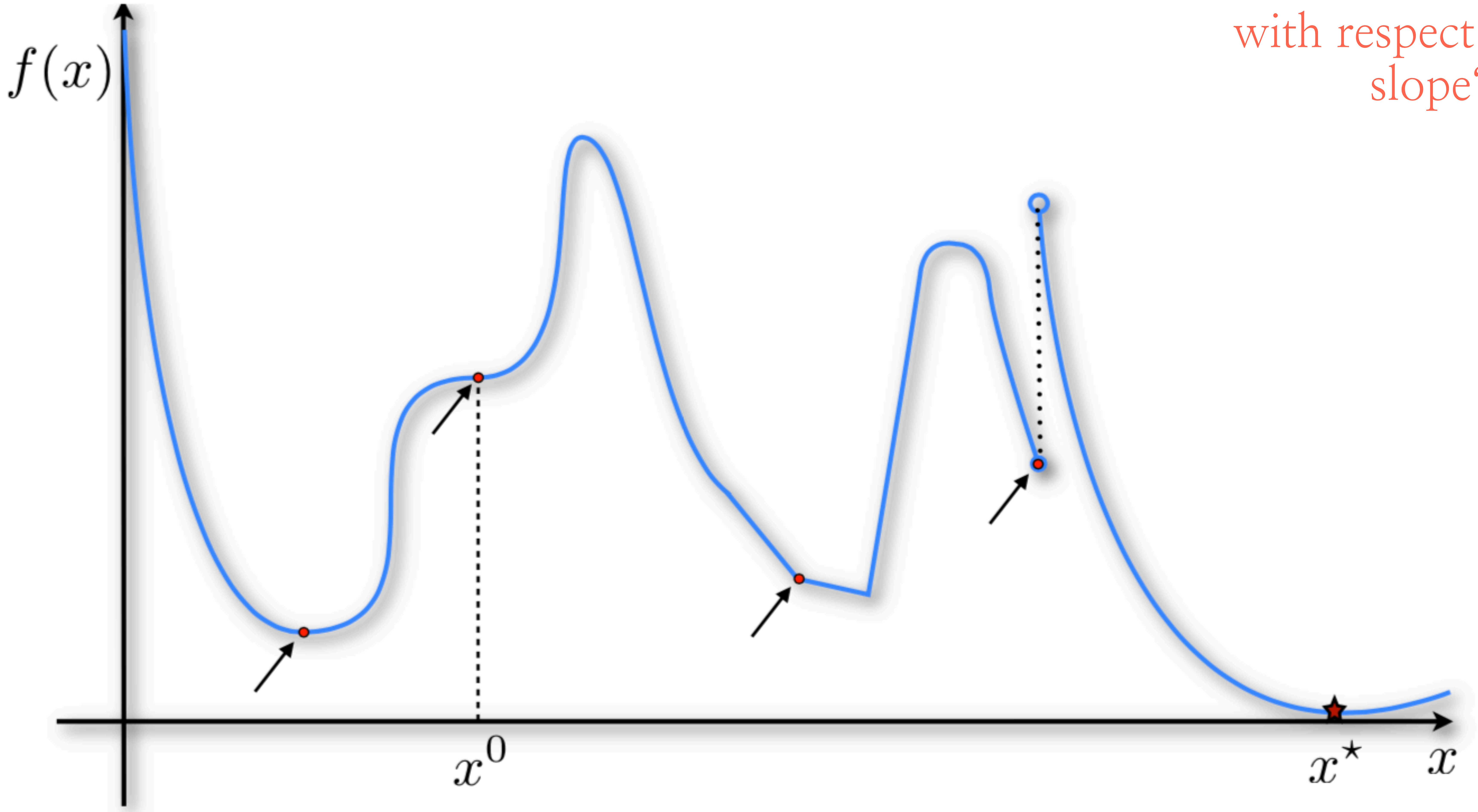
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2$$

Agnostic optimization



Agnostic optimization

What do you observe at local minima/maxima with respect to their slope?



Types of solutions

- Global minimizer $x^\star$:

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \ \forall x$

(If we could somehow check that, we are golden!)

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \ \forall x$

(If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$:

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x),\ \forall x$

(If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$: $f(\widehat{x}) \leq f(x),\ \forall x \in \mathcal{N}_{\widehat{x}}$

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \ \forall x$

  (If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$: $f(\widehat{x}) \leq f(x), \ \forall x \in \mathcal{N}_{\widehat{x}}$

- What is the meaning of strict inequality vs. inequality?

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \; \forall x$

(If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$: $f(\widehat{x}) \leq f(x), \; \forall x \in \mathcal{N}_{\widehat{x}}$

- What is the meaning of strict inequality vs. inequality?

- How do we recognize that a solution we have is a local (global) solution?

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \ \forall x$

(If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$: $f(\widehat{x}) \leq f(x), \ \forall x \in \mathcal{N}_{\widehat{x}}$

- What is the meaning of strict inequality vs. inequality?

- How do we recognize that a solution we have is a local (global) solution?

  - 1^{st}–order optimality condition: $\nabla f(\widehat{x}) = 0$

Types of solutions

- Global minimizer $x^\star$: $f(x^\star) \leq f(x), \ \forall x$

  (If we could somehow check that, we are golden!)

- Local minimizer $\widehat{x}$: $f(\widehat{x}) \leq f(x), \ \forall x \in \mathcal{N}_{\widehat{x}}$

- What is the meaning of strict inequality vs. inequality?

- How do we recognize that a solution we have is a local (global) solution?

  - 1^{st}–order optimality condition: $\nabla f(\widehat{x}) = 0$

  - 2^{nd}–order optimality condition: $\nabla f(\widehat{x}) = 0$ and $\nabla^2 f(\widehat{x}) \succeq 0$

Types of solutions

– Global minimizer $x^\star$: $f(x^\star) \le f(x), \ \forall x$

(If we could somehow check that, we are golden!)

– Local minimizer $\widehat{x}$: $f(\widehat{x}) \le f(x), \ \forall x \in \mathcal{N}_{\widehat{x}}$

– What is the meaning of strict inequality vs. inequality?

– How do we recognize that a solution we have is a local (global) solution?

Necessary

– 1^{st}–order optimality condition: $\nabla f(\widehat{x}) = 0$

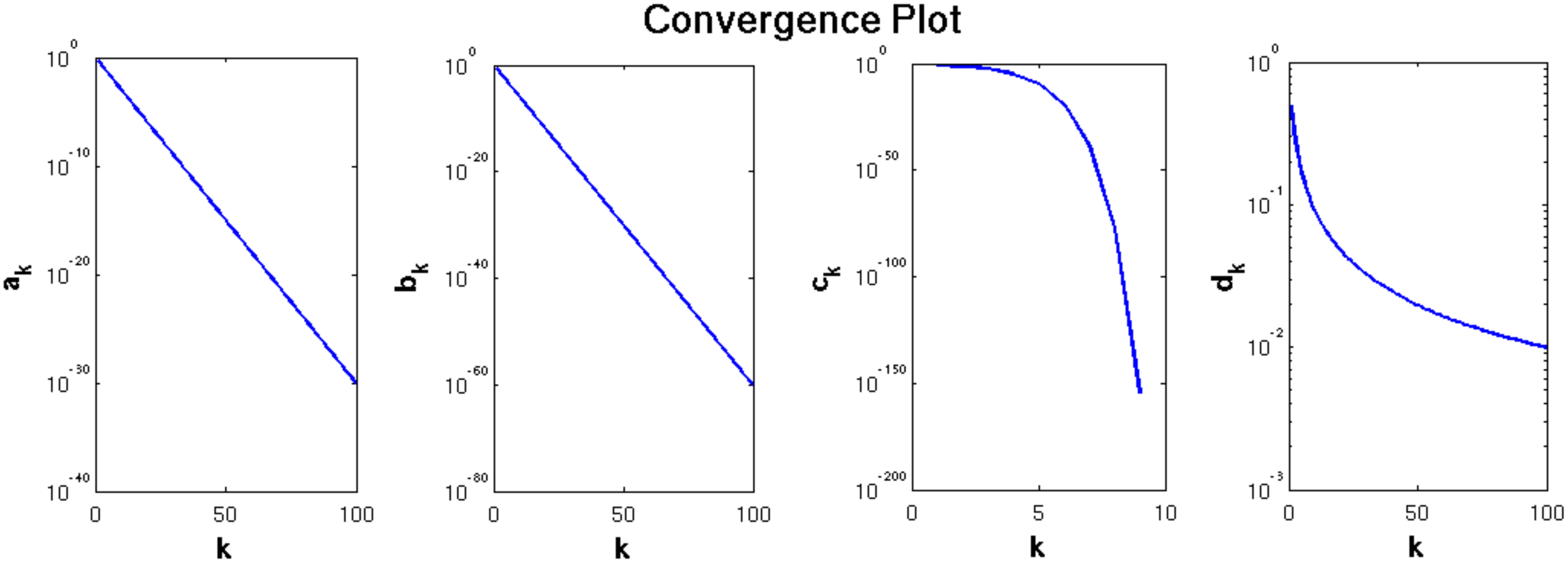
– 2^{nd}–order optimality condition: $\nabla f(\widehat{x}) = 0$ and $\nabla^2 f(\widehat{x}) \succeq 0$

First convergence result

Whiteboard

Convergence rates 101

(Source: Wikipedia)



$O(\log 1/\varepsilon)$

$O\left(\log\log(1/\varepsilon)\right)$

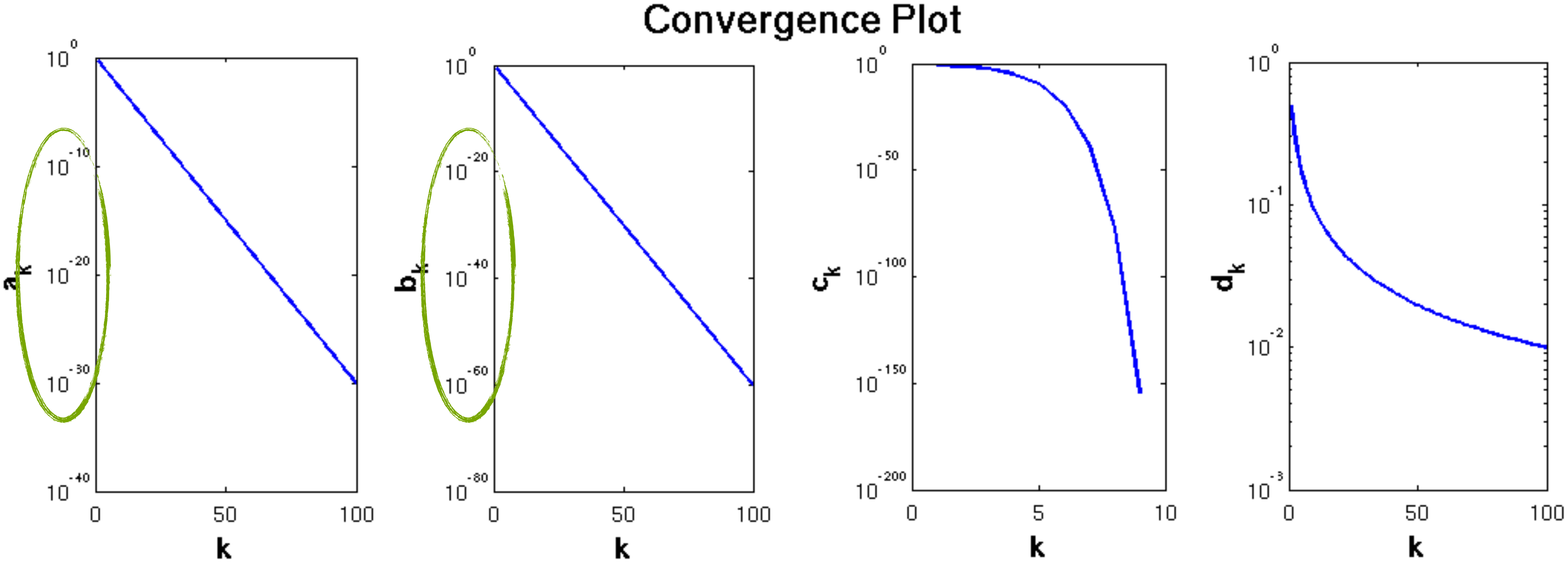
$O(1/\varepsilon^2),\ O(1/\varepsilon),\ O(1/\sqrt{\varepsilon})$

$q^k,\ q \in (0,1)$

$O(1/k^2), O(1/k), O(\sqrt(k))$

Convergence rates 101

(Source: Wikipedia)



$O(\log 1/\varepsilon)$

$q^k, \ q \in (0,1)$

$O\left(\log\log(1/\varepsilon)\right)$

$O(1/\varepsilon^2), \ O(1/\varepsilon), \ O(1/\sqrt{\varepsilon})$

$O(1/k^2), O(1/k), O(\sqrt(k))$

First convergence result

$$\min_{x \in \mathbb{R}^p} f(x)$$

"Assume the objective is has Lipschitz continuous gradients. Then, gradient descent:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

with step size

$$\eta = \frac{1}{L}$$

converges sublinearly to a stationary point; i.e.,

$$\min_t \|\nabla f(x_t)\|_2 \leq \sqrt{\frac{2L}{T+1}} \cdot \left(f(x_0) - f(x^\star)\right)^{1/2} = O\left(\frac{1}{\sqrt{T}}\right) \text{"}$$

First convergence result

- "But, which functions satisfy Lipschitz gradient continuity?"

First convergence result

- "But, which functions satisfy Lipschitz gradient continuity?"
  - Least-squares objectives: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$$

First convergence result

- "But, which functions satisfy Lipschitz gradient continuity?"
  - Least–squares objectives: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$$

  - Logistic regression objectives: $f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \alpha_i^\top x)\right)$

Whiteboard

First convergence result

- "But, which functions satisfy Lipschitz gradient continuity?"
  - Least−squares objectives: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$$

  - Logistic regression objectives: $f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \alpha_i^\top x)\right)$

Whiteboard

"But these are actually convex!"

First convergence result

- "But, which functions satisfy Lipschitz gradient continuity?"
  - Least−squares objectives: $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \|A^\top A\|_2 \cdot \|x - y\|_2$$

  - Logistic regression objectives: $f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-y_i \alpha_i^\top x)\right)$

Whiteboard

"But these are actually convex!"

  - Non−convex objective: $f(x) = x^2 + 3\sin^2(x)$

Demo

Conclusion

- We have set up background of smooth optimization
- We have provided the first convergence rate result, and defined different convergence rates that could be attainable

Conclusion

- We have set up background of smooth optimization
- We have provided the first convergence rate result, and defined different convergence rates that could be attainable

Next lecture

- Brief introduction to convex optimization and related topics