COMP 545: Advanced topics in optimization
From simple to complex ML systems

Lecture 2

Overview

$$\min_{x} \quad f(x)$$

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

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

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

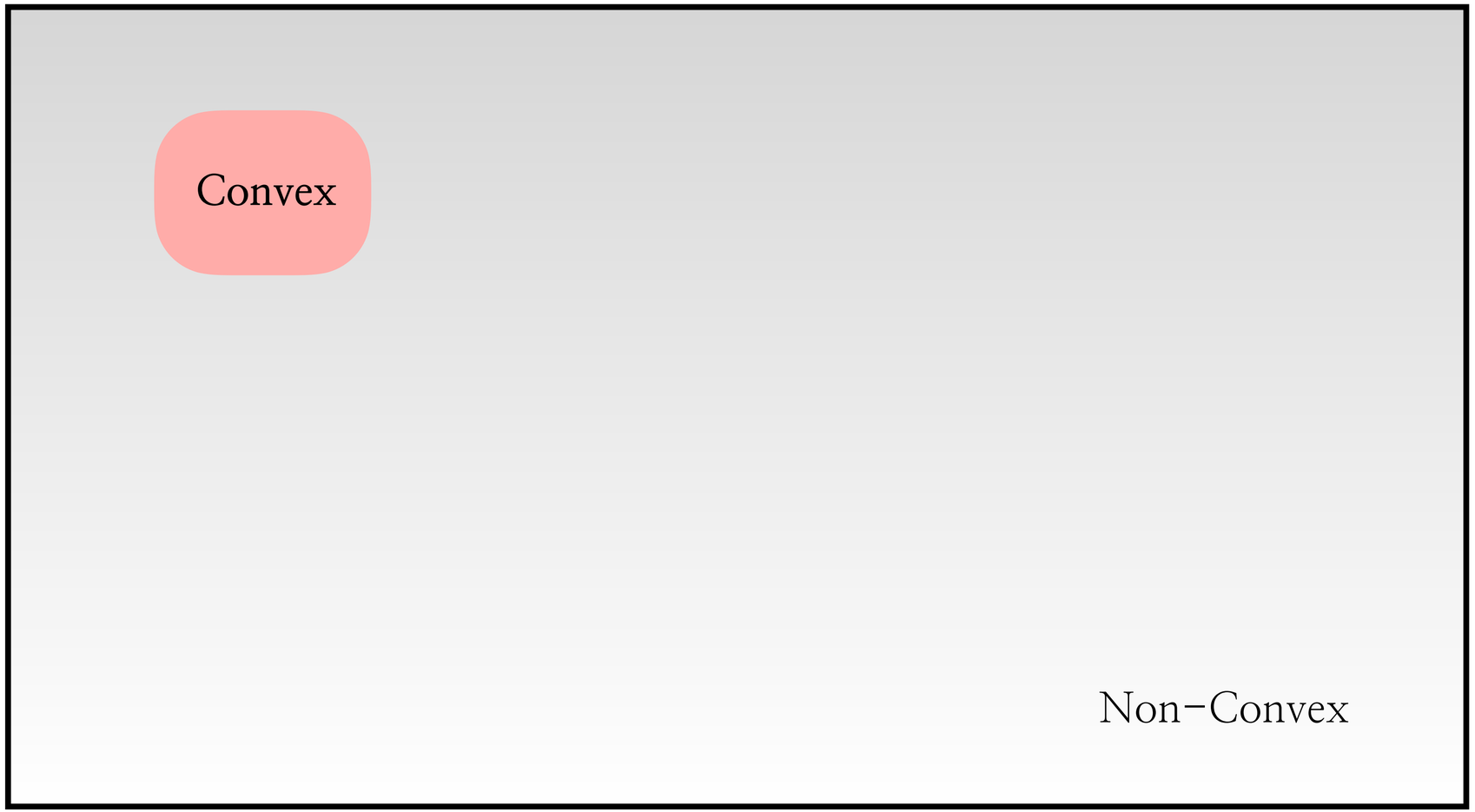
Convex vs. non-convex optimization

(Naive interpretation of) Space of optimization problems

Convex vs. non-convex optimization

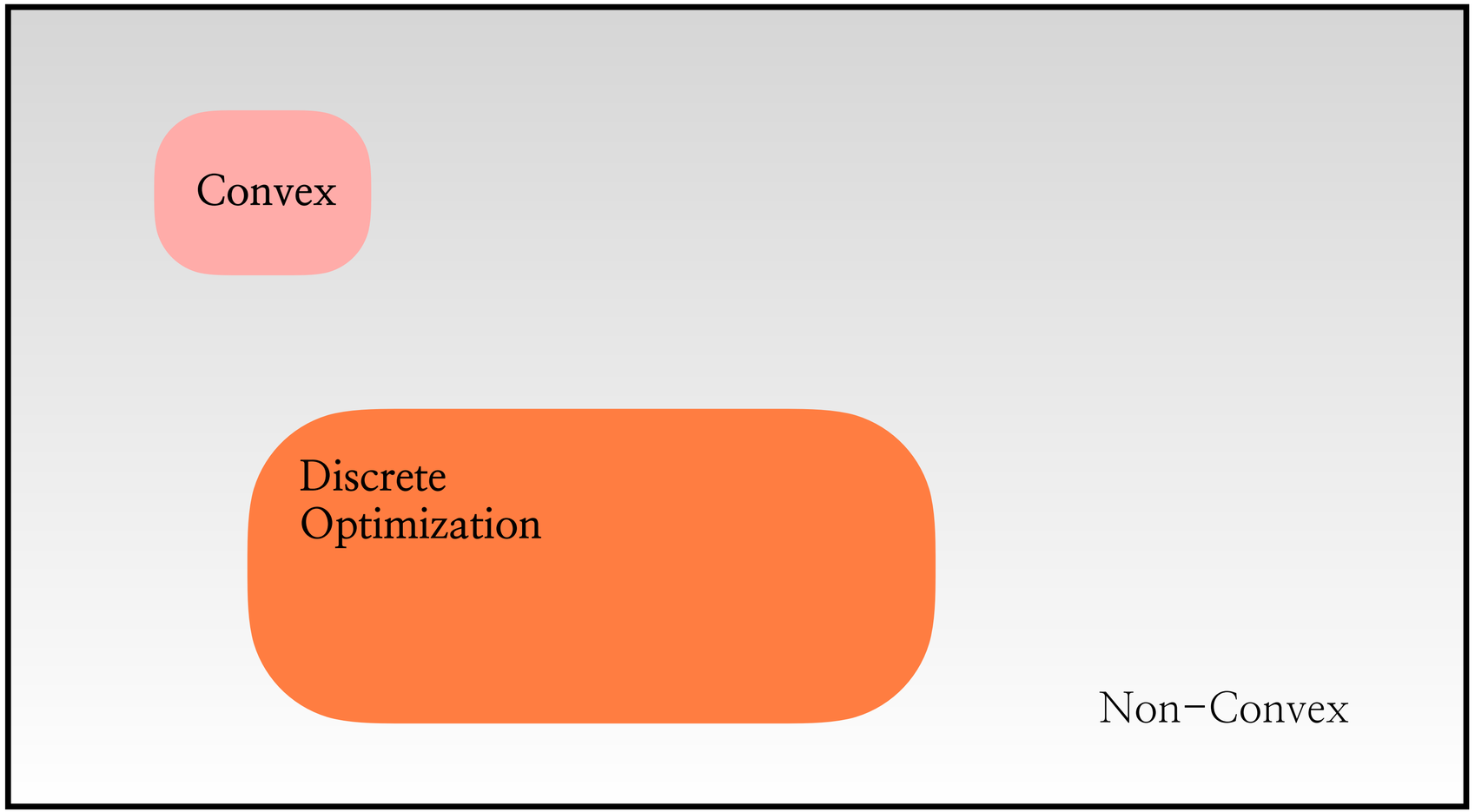
(Naive interpretation of) Space of optimization problems

Convex vs. non-convex optimization



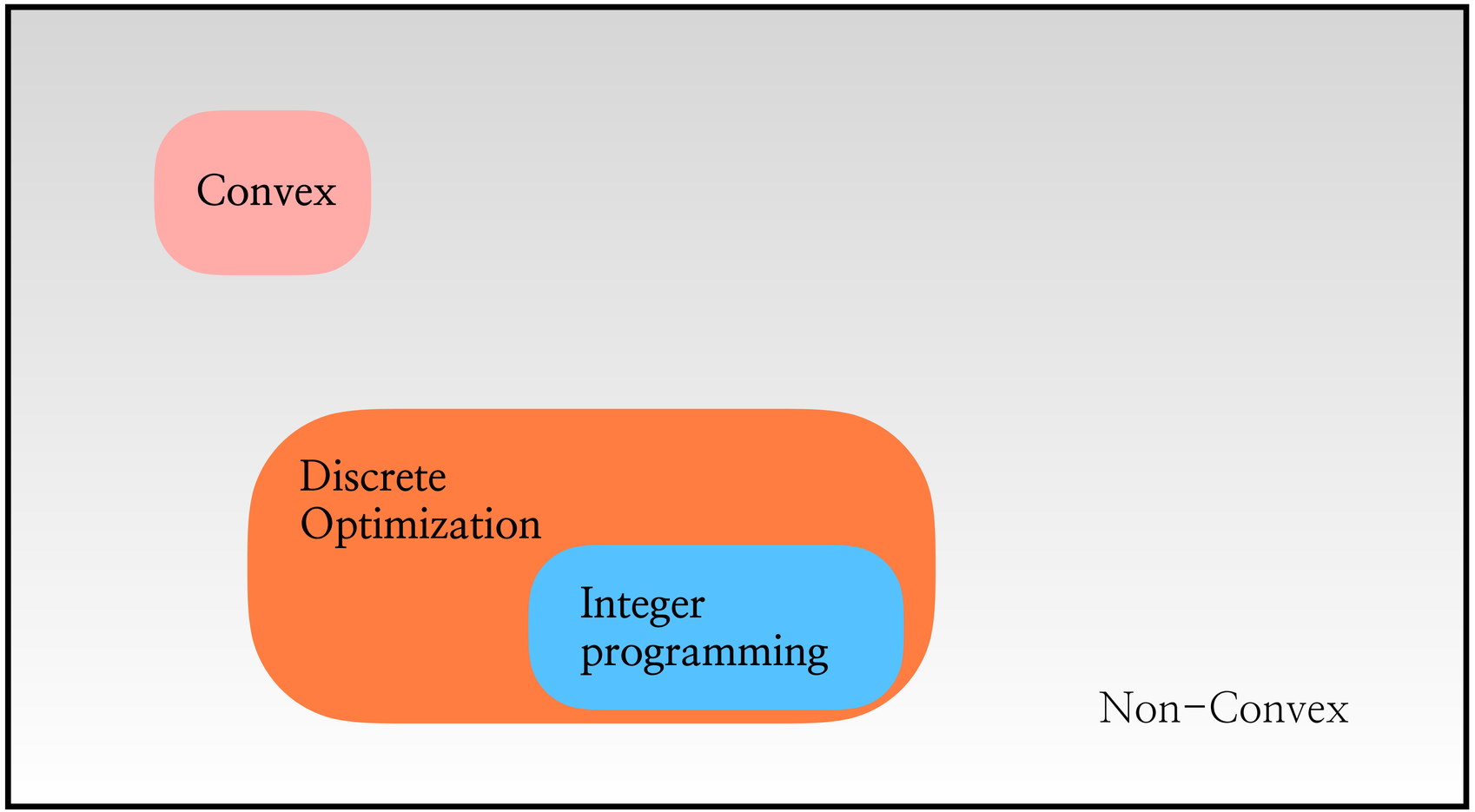
(Naive interpretation of) Space of optimization problems

Convex vs. non-convex optimization



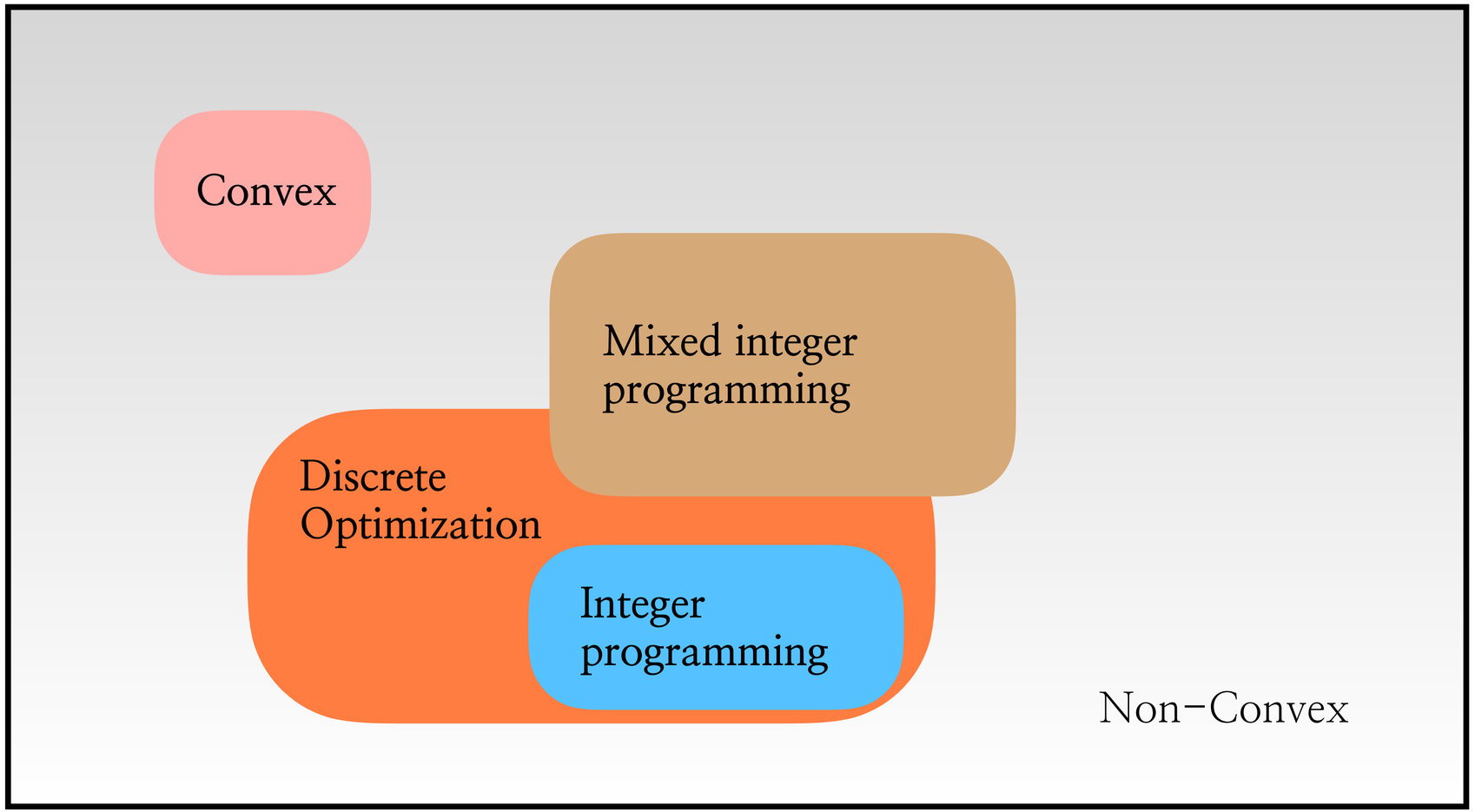
(Naive interpretation of) Space of optimization problems

Convex vs. non-convex optimization



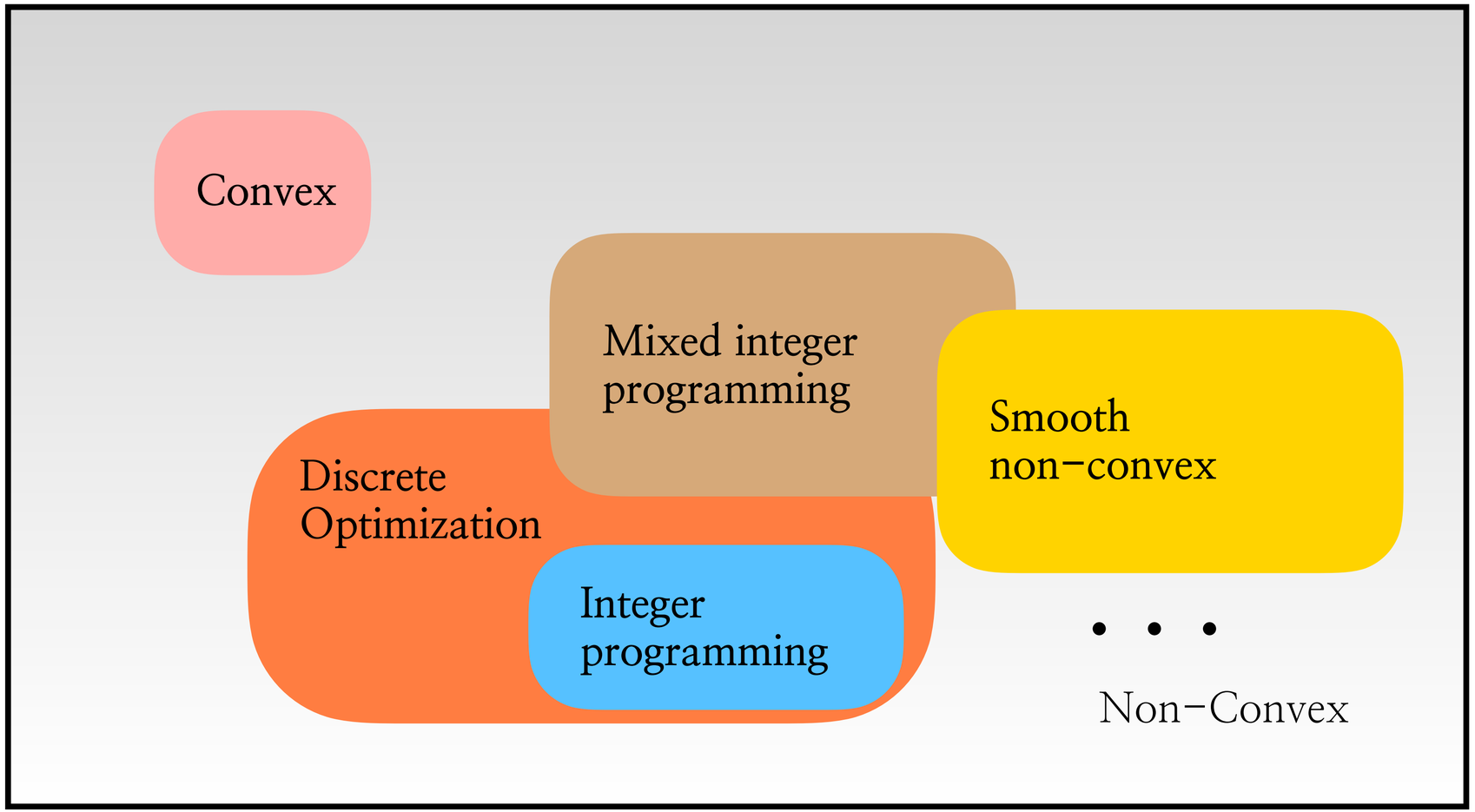
(Naive interpretation of) Space of optimization problems

Convex vs. non-convex optimization



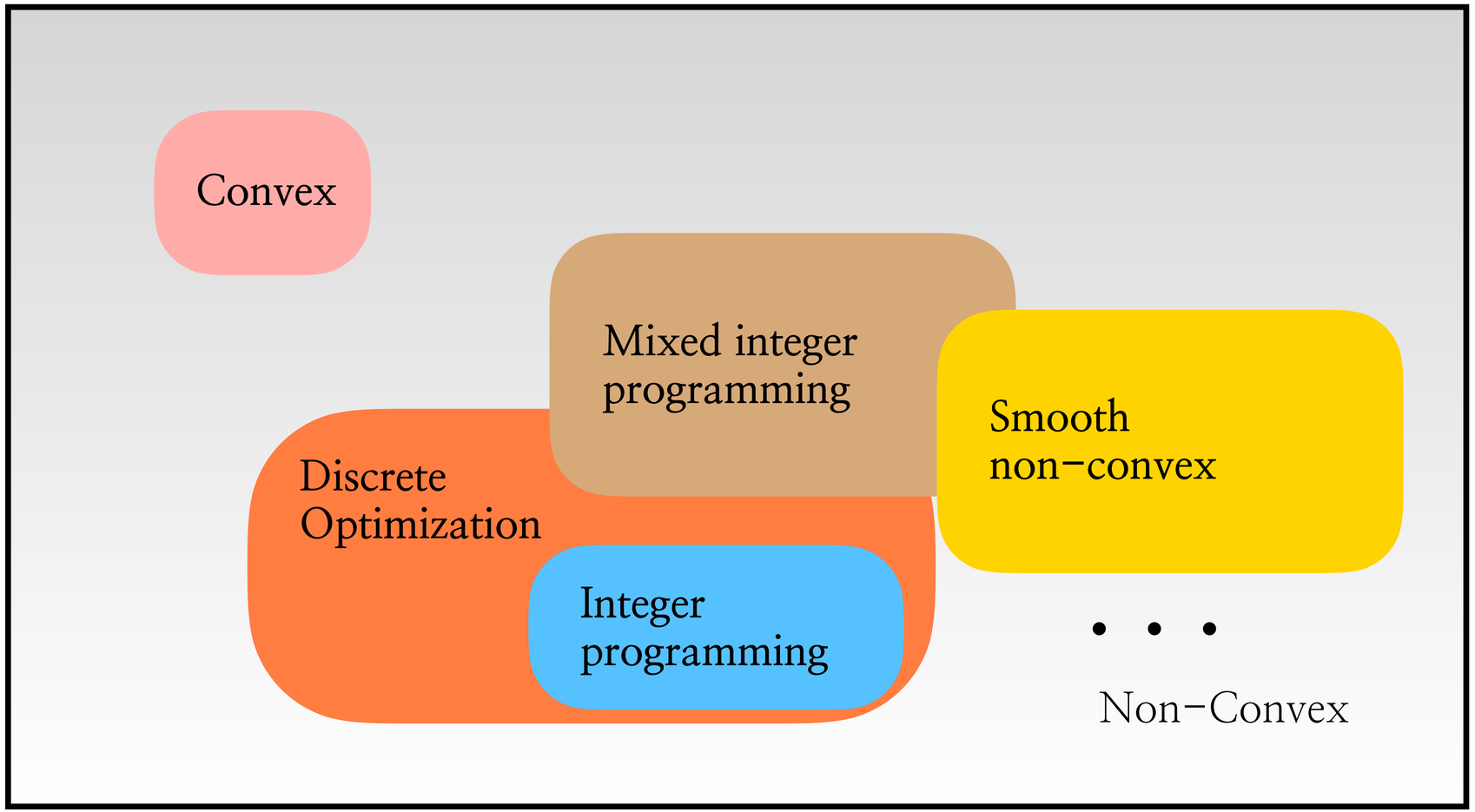
(Naive interpretation of) Space of optimization problems

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**

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

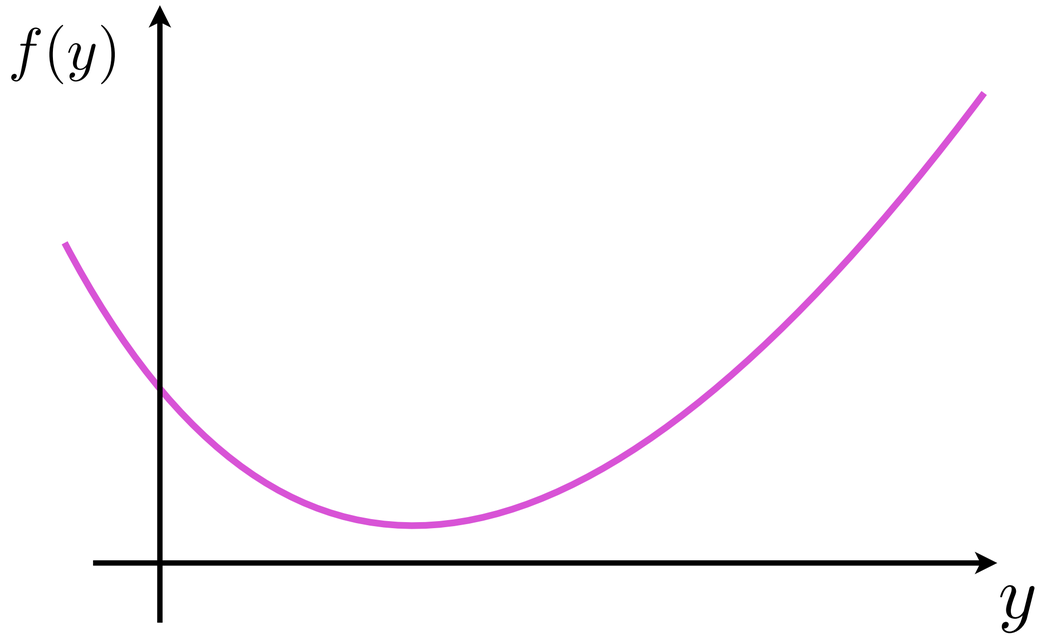
- Intuition: generate a sequence of points $\{f(x+\epsilon_i), \epsilon_i\}$, and compute the limit as $\epsilon_i \to 0$

Derivatives and gradients

- Definition of a **derivative**

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

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

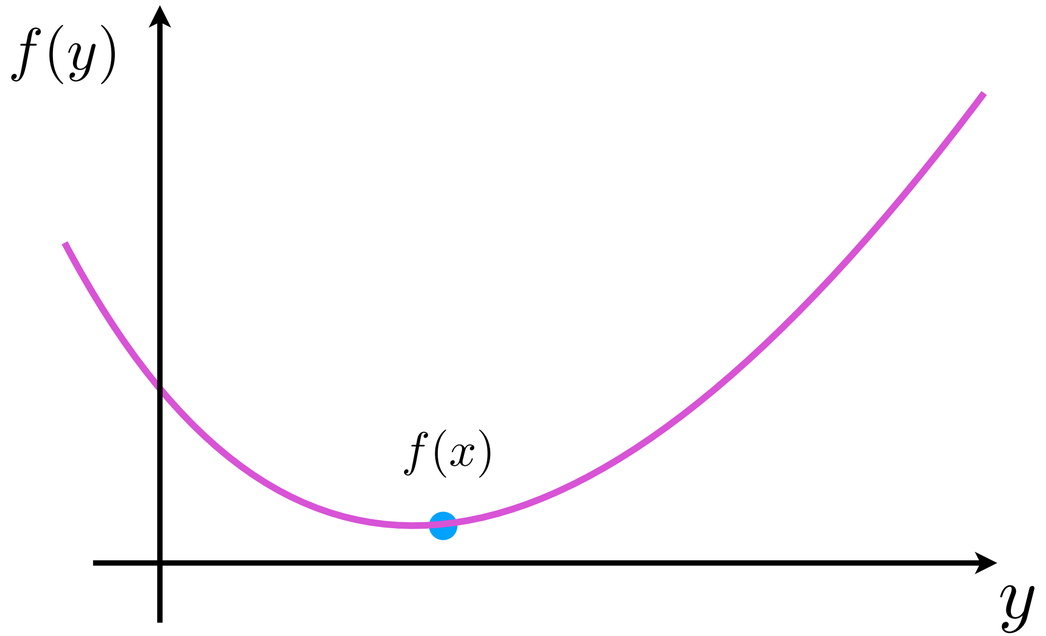


Derivatives and gradients

- 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 $\epsilon_i \to 0$

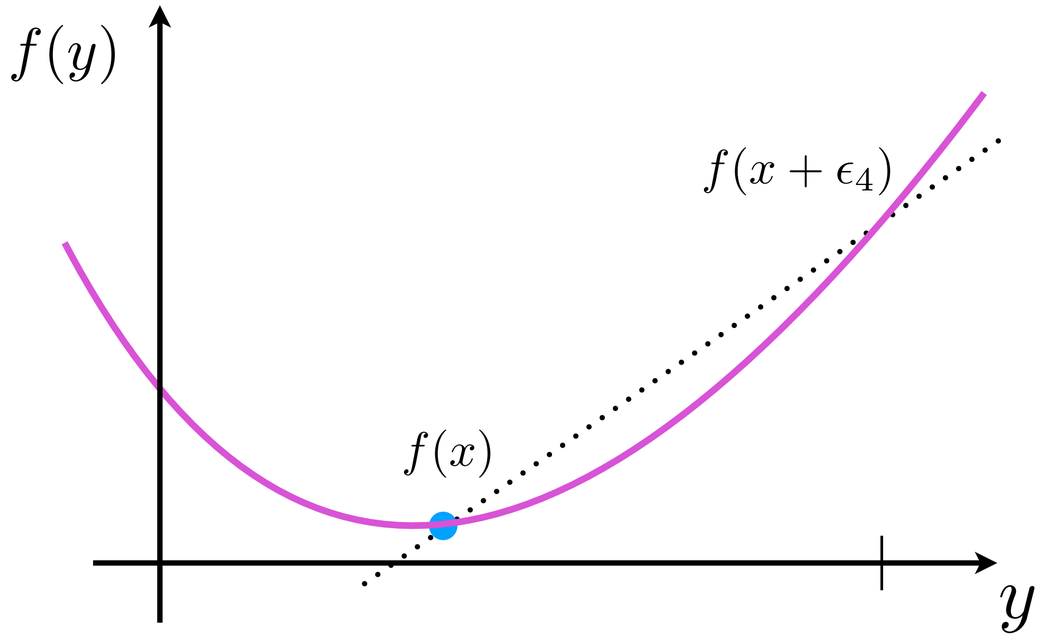


Derivatives and gradients

- 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 $\epsilon_i \to 0$

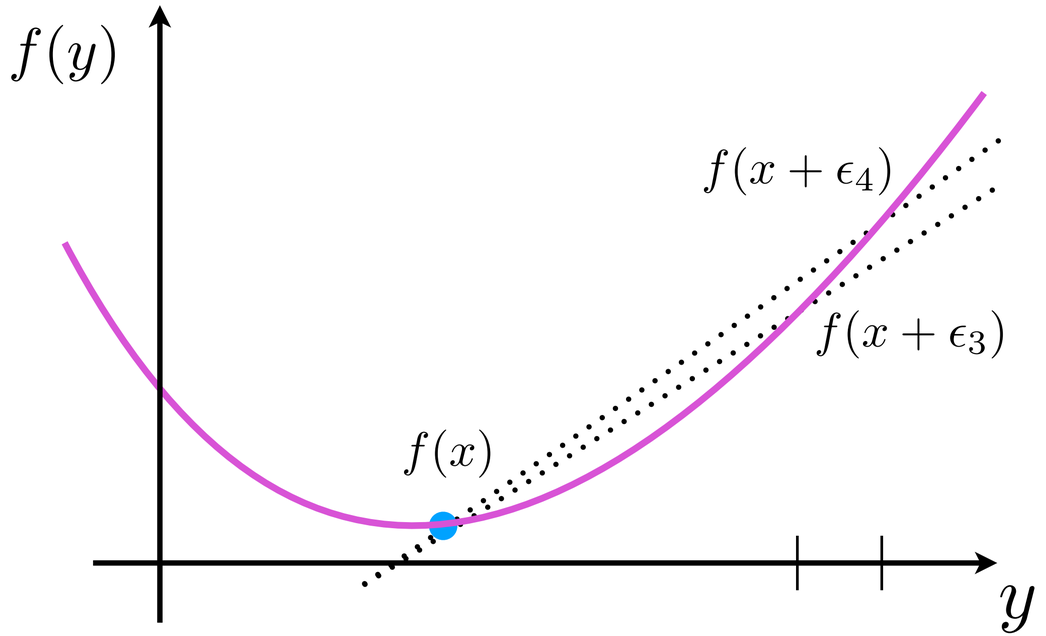


Derivatives and gradients

- 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 $\epsilon_i \to 0$

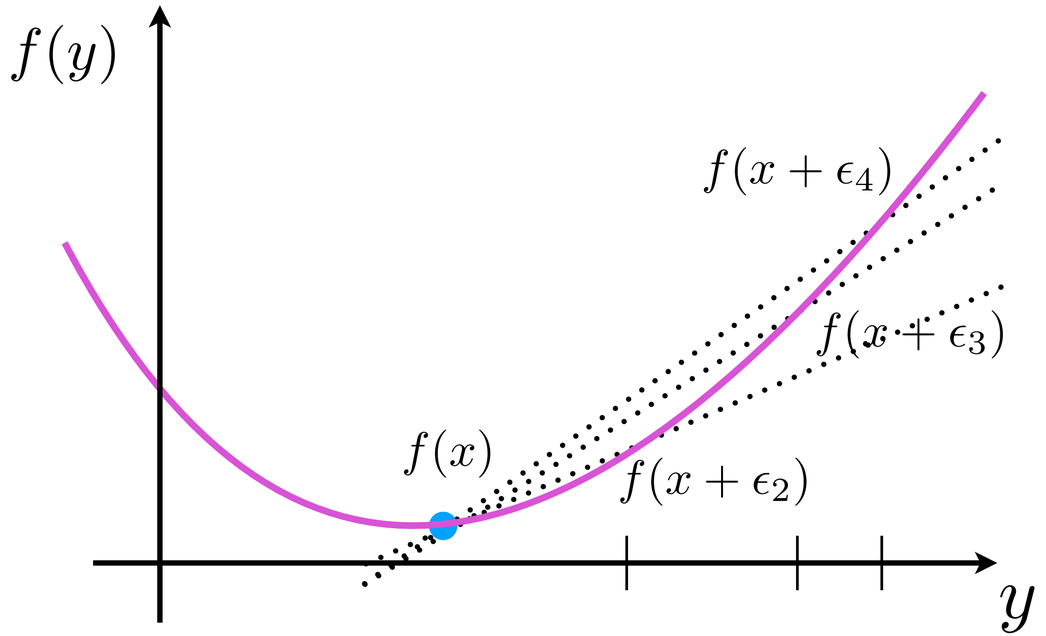


Derivatives and gradients

- 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 $\epsilon_i \to 0$

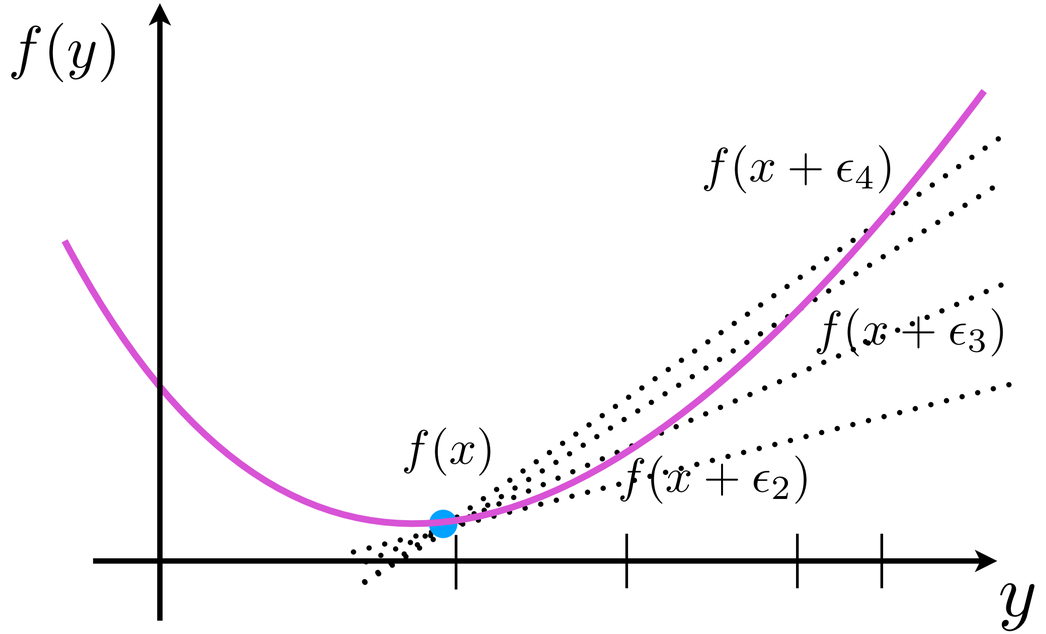


Derivatives and gradients

- 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 $\epsilon_i \to 0$

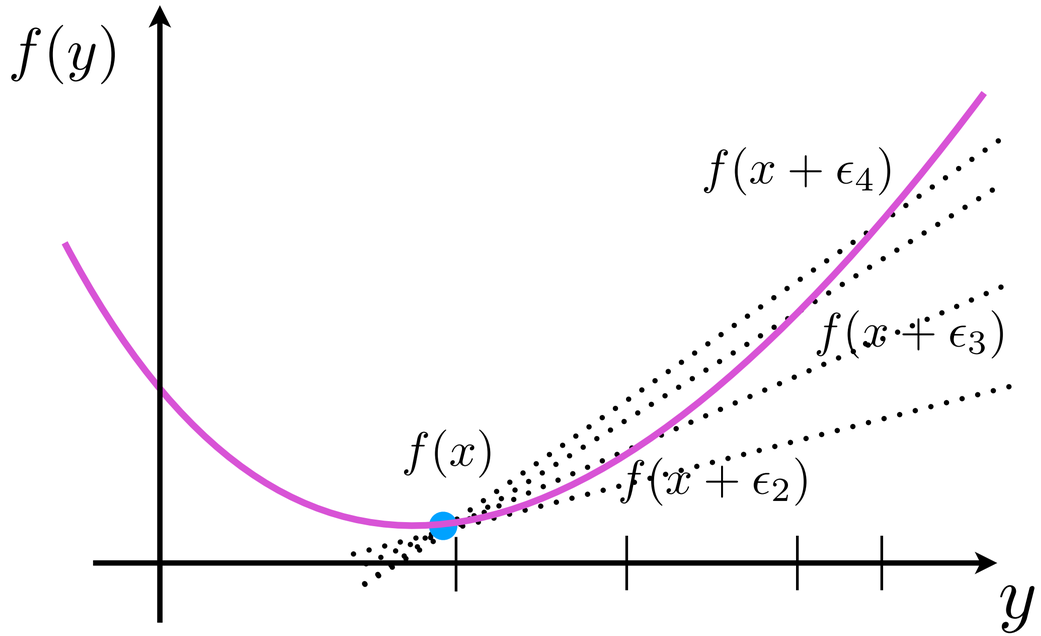


Derivatives and gradients

- 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 $\epsilon_i \to 0$



Represents the rate that function changes

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**

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

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)

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

- Generalizes the notion of gradient to multiple-output functions

Derivatives and gradients

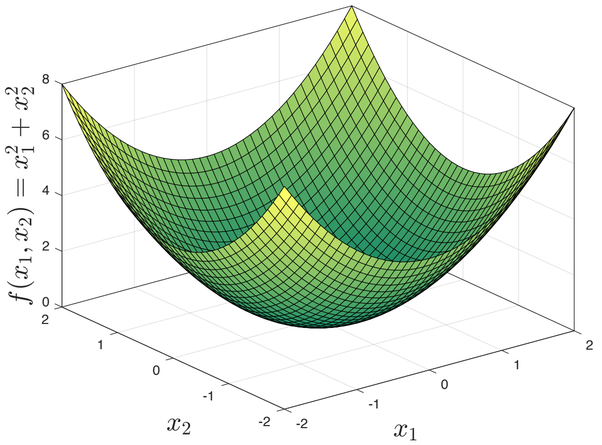
- **Hessian** matrix

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

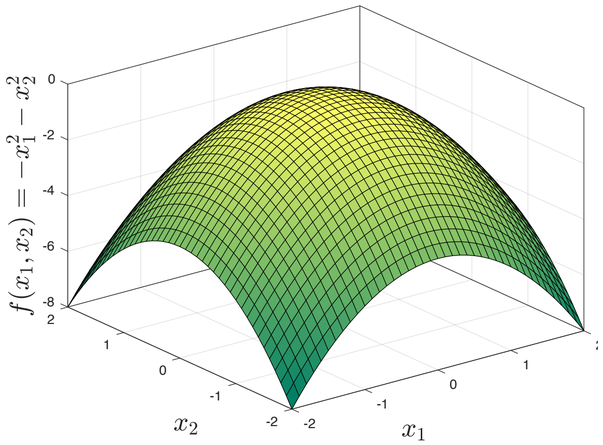
Derivatives and gradients

- **Hessian** matrix

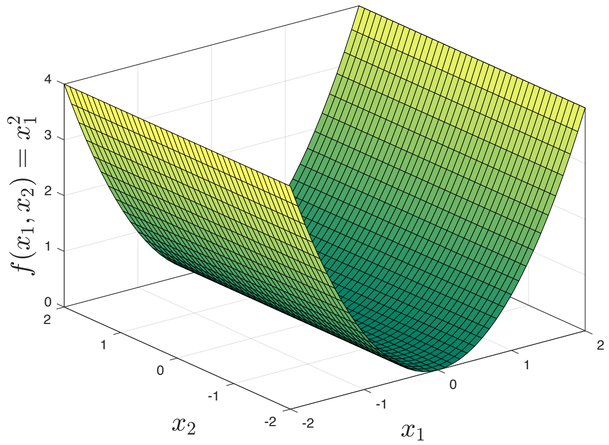
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial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



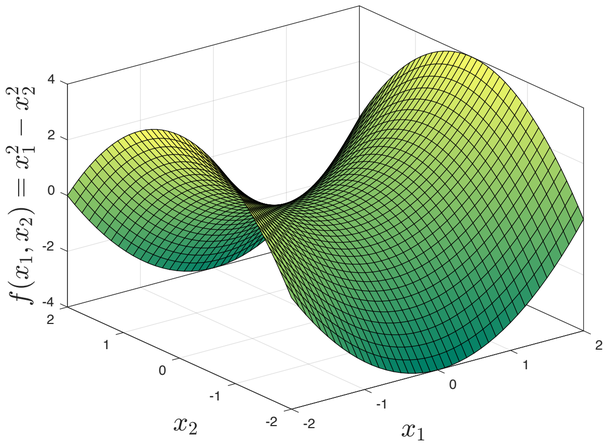
$\nabla^2 f(\cdot) \succ 0$



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



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

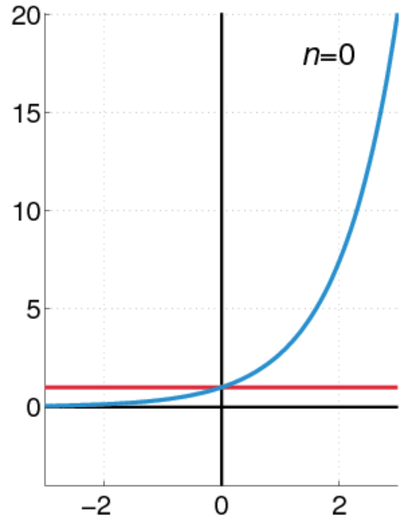


Indefinite

Taylor's expansion

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

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

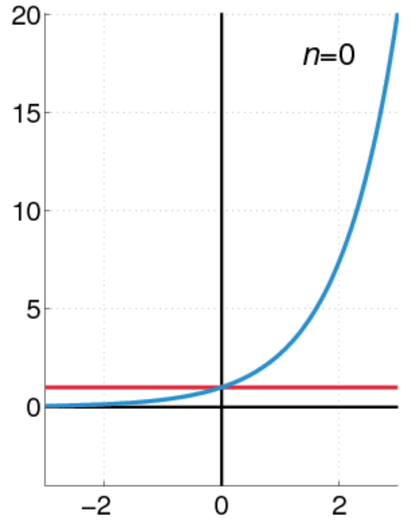


Example:
exponential function

Taylor's expansion

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

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

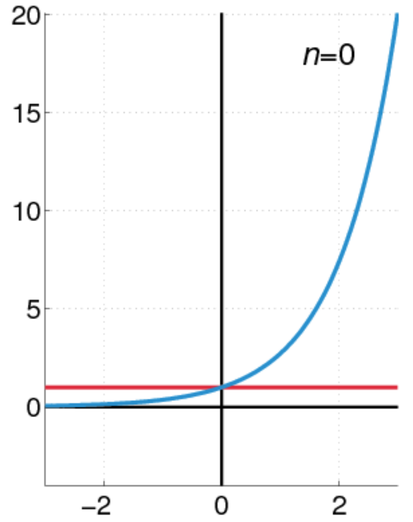


Example:
exponential function

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

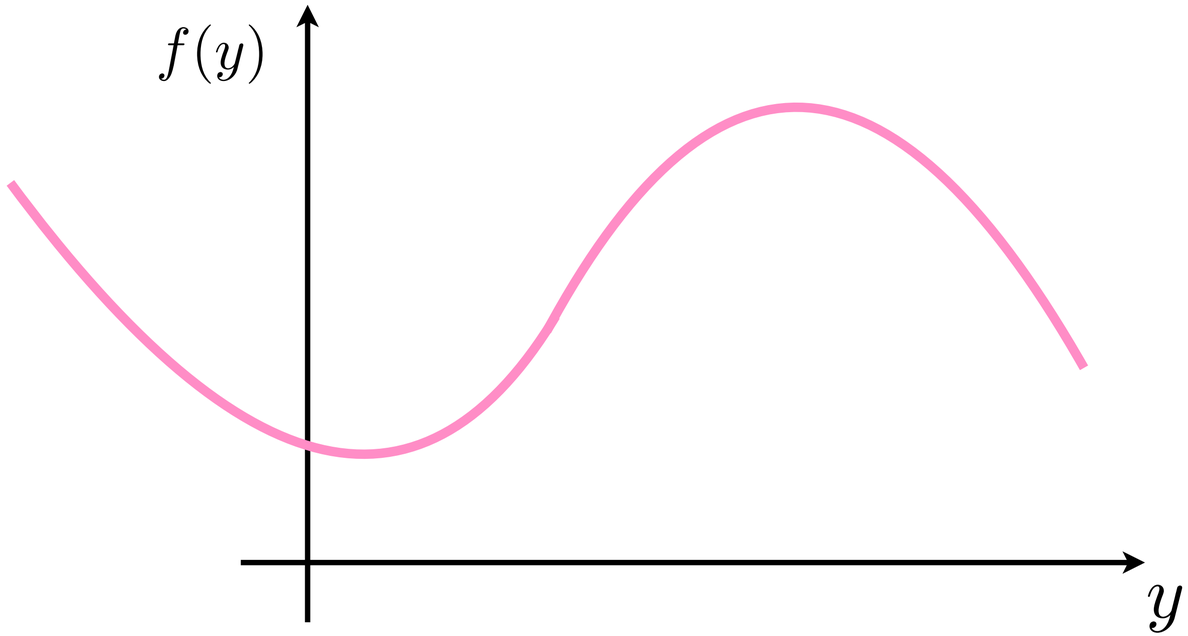
- First-order Taylor's approximation

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

Taylor's expansion

- First-order Taylor's approximation

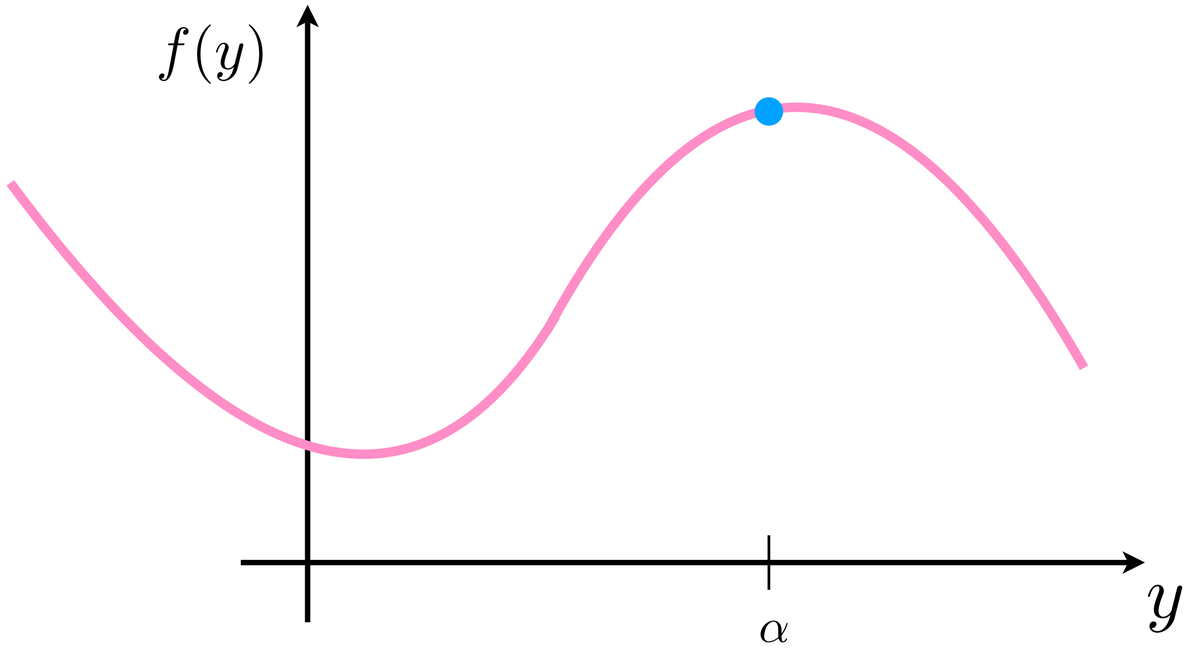
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$



Taylor's expansion

- First-order Taylor's approximation

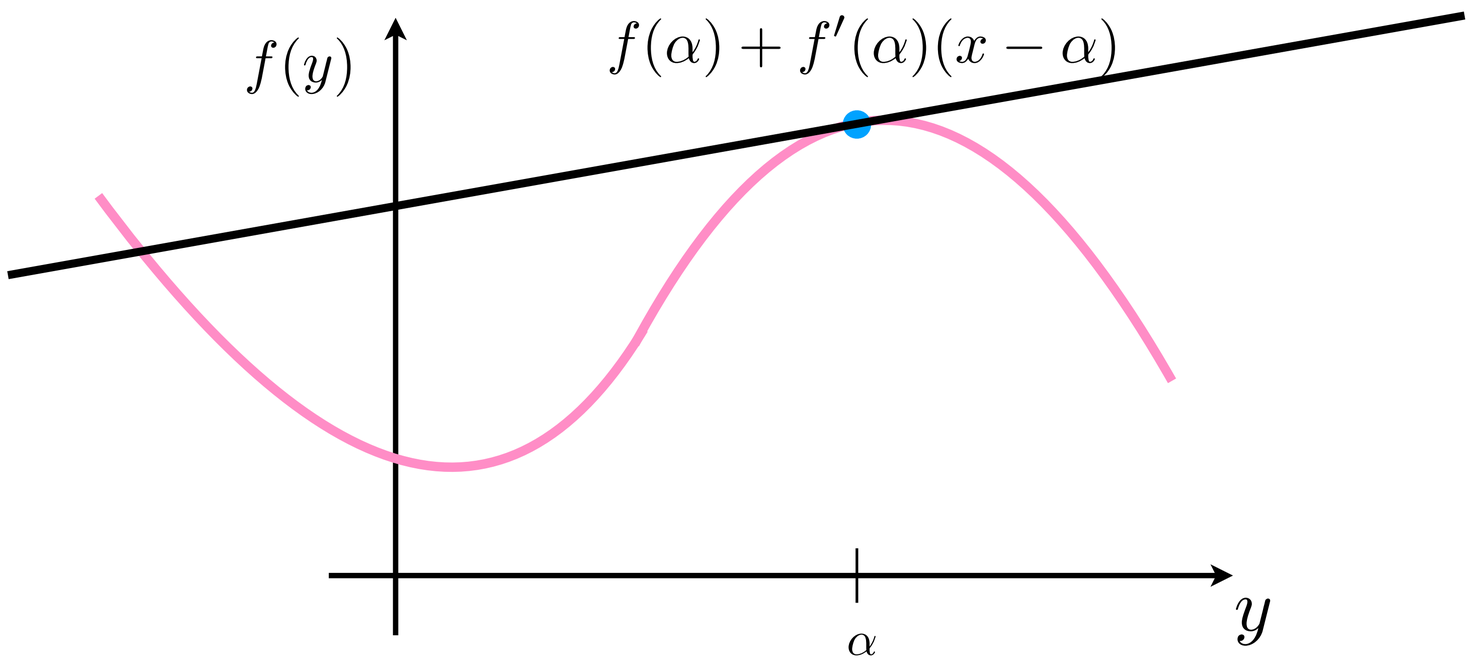
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$



Taylor's expansion

- First-order Taylor's approximation

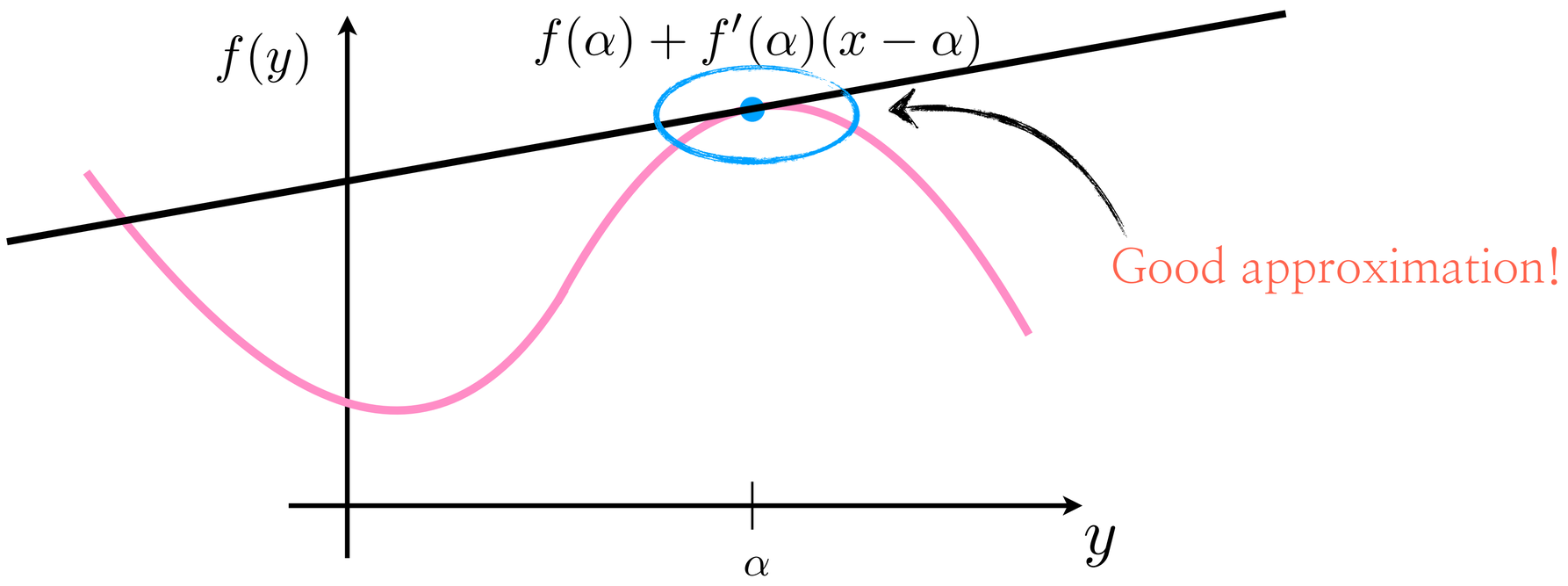
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$



Taylor's expansion

- First-order Taylor's approximation

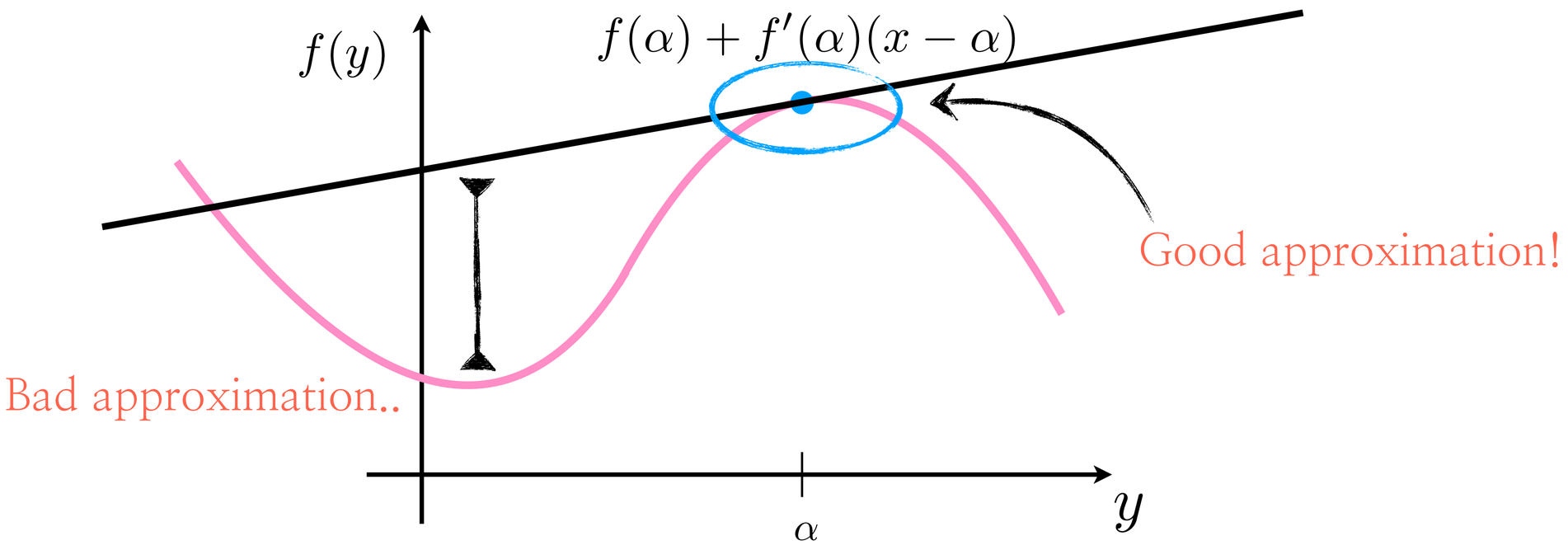
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle \, , \alpha \in \mathbb{R}^p$$



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

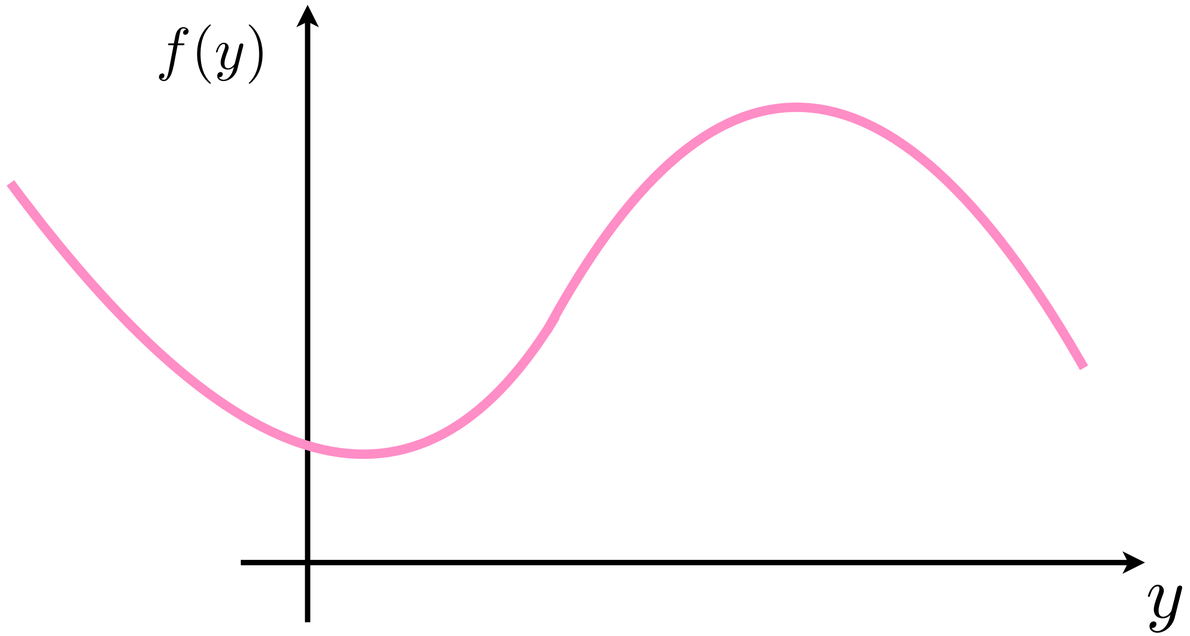
- Second-order Taylor's approximation

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

Taylor's expansion

- Second-order Taylor's approximation

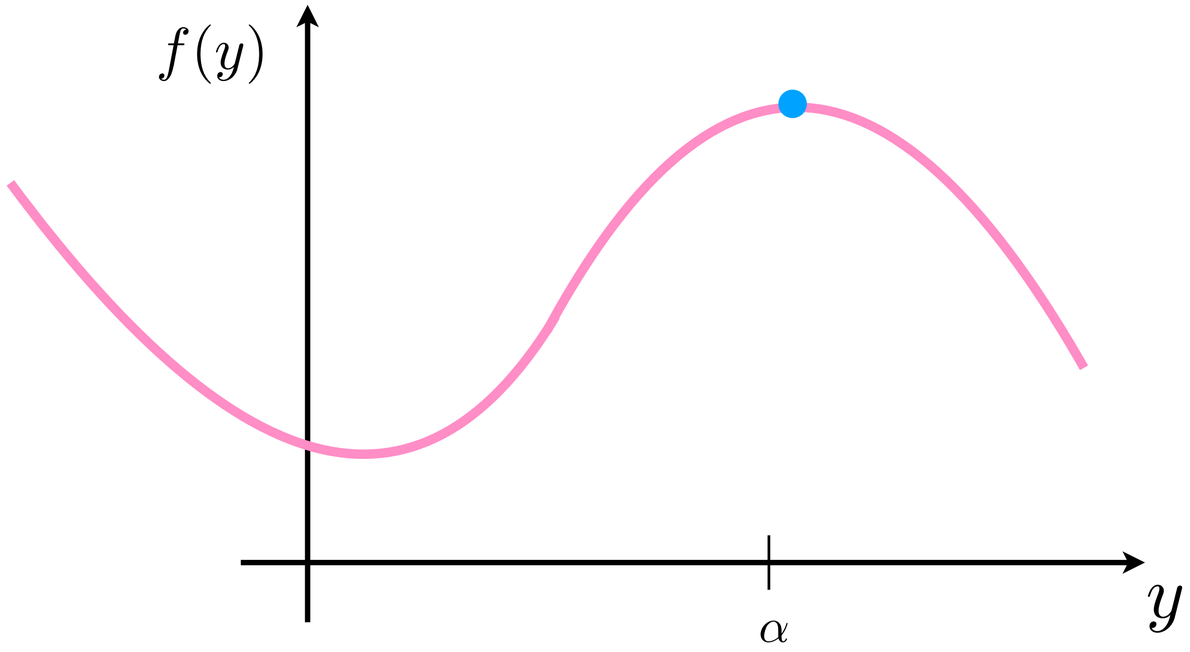
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \tfrac{1}{2} \left\langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \right\rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

- Second-order Taylor's approximation

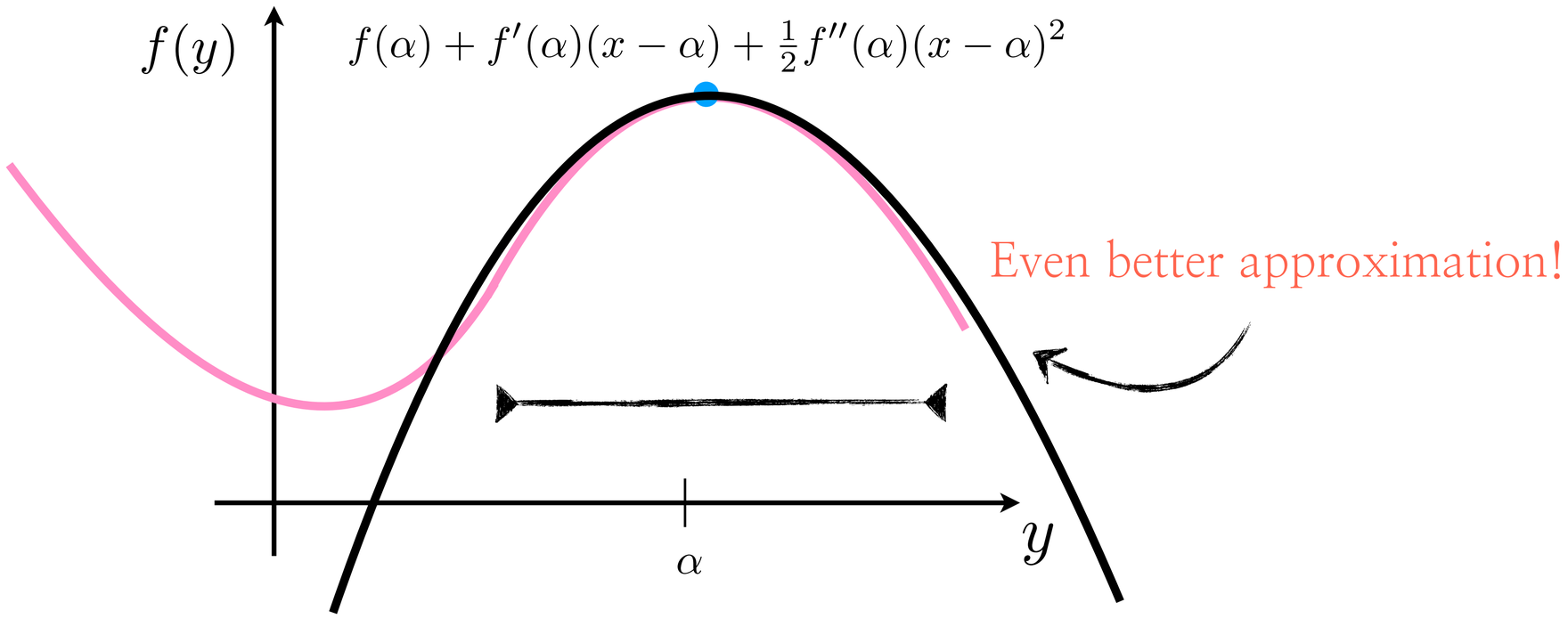
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \tfrac{1}{2} \left\langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \right\rangle, \alpha \in \mathbb{R}^p$$



Taylor's expansion

- Second-order Taylor's approximation

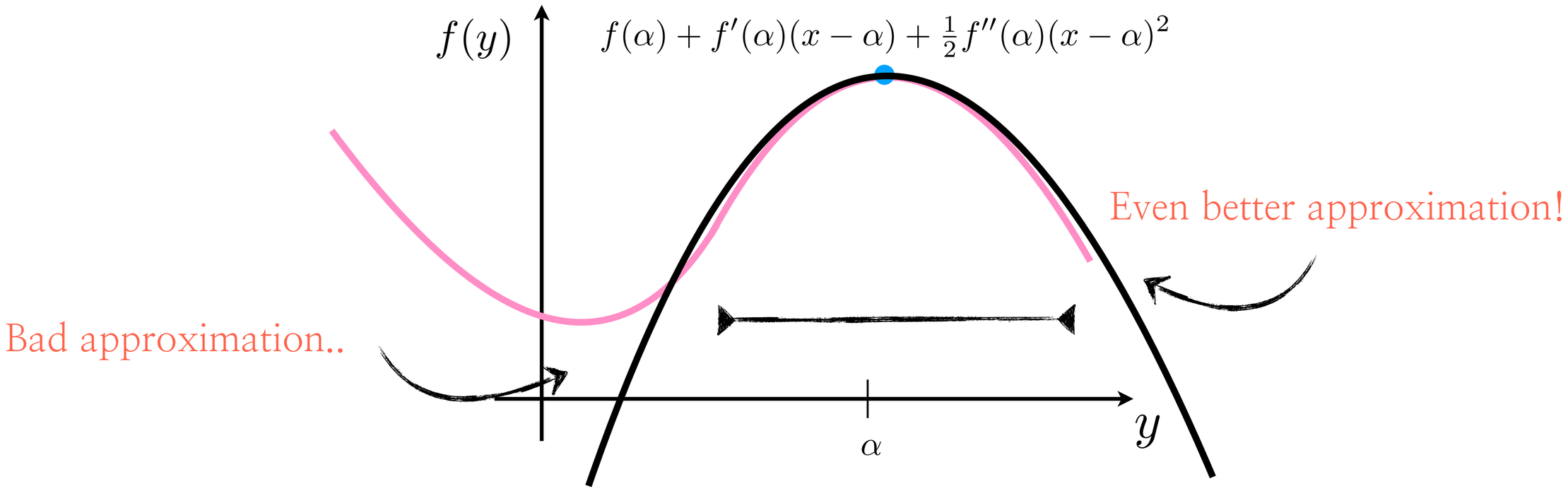
$$f : \mathbb{R}^p \to \mathbb{R} \quad \Big| \quad f(x) \approx f(\alpha) + \langle \nabla f(\alpha), x - \alpha \rangle + \tfrac{1}{2} \left\langle \nabla^2 f(\alpha)(x - \alpha), x - \alpha \right\rangle, \alpha \in \mathbb{R}^p$$



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

- 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)

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

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!

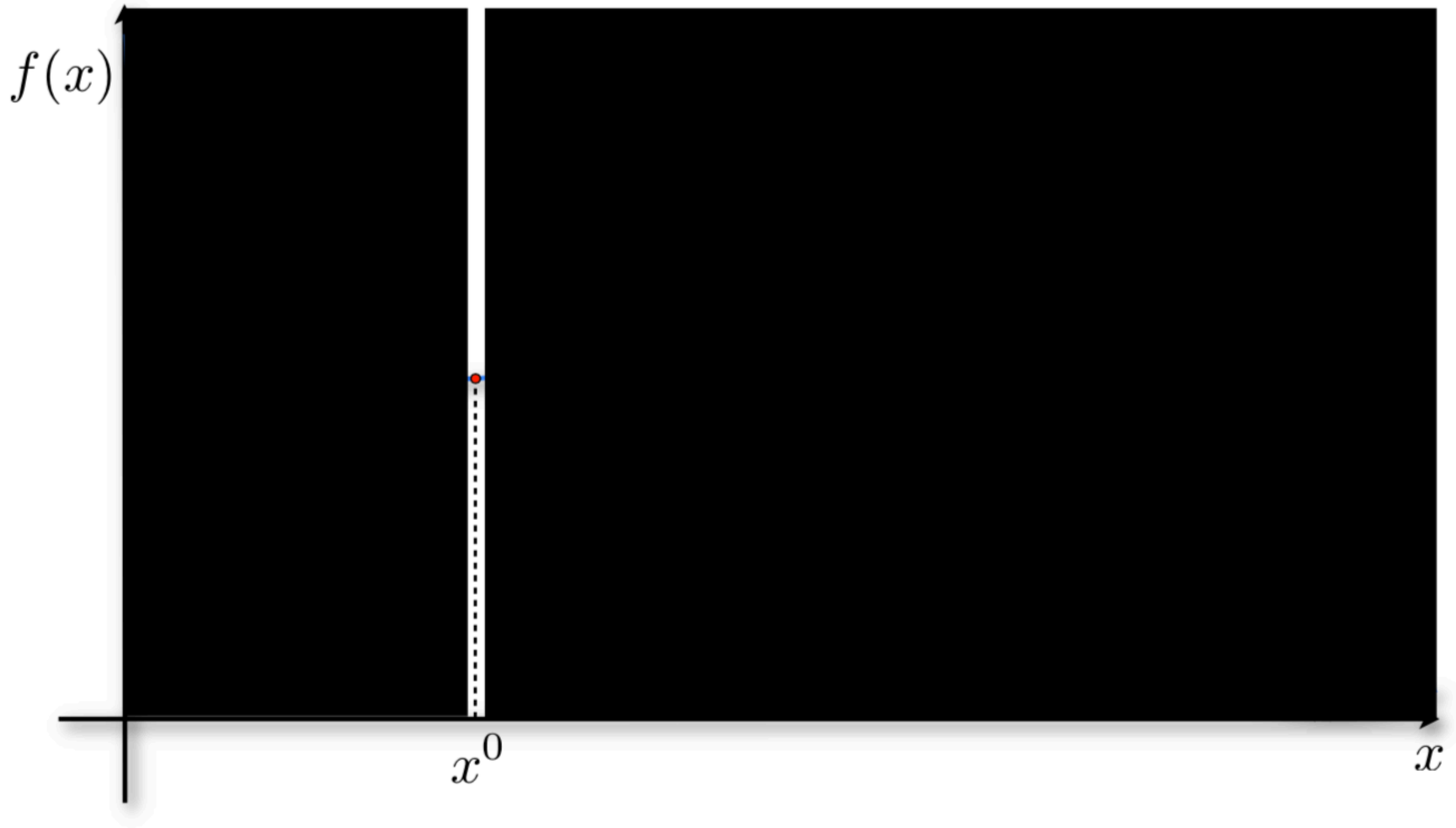
Weirdest function ever
(but differentiable)

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

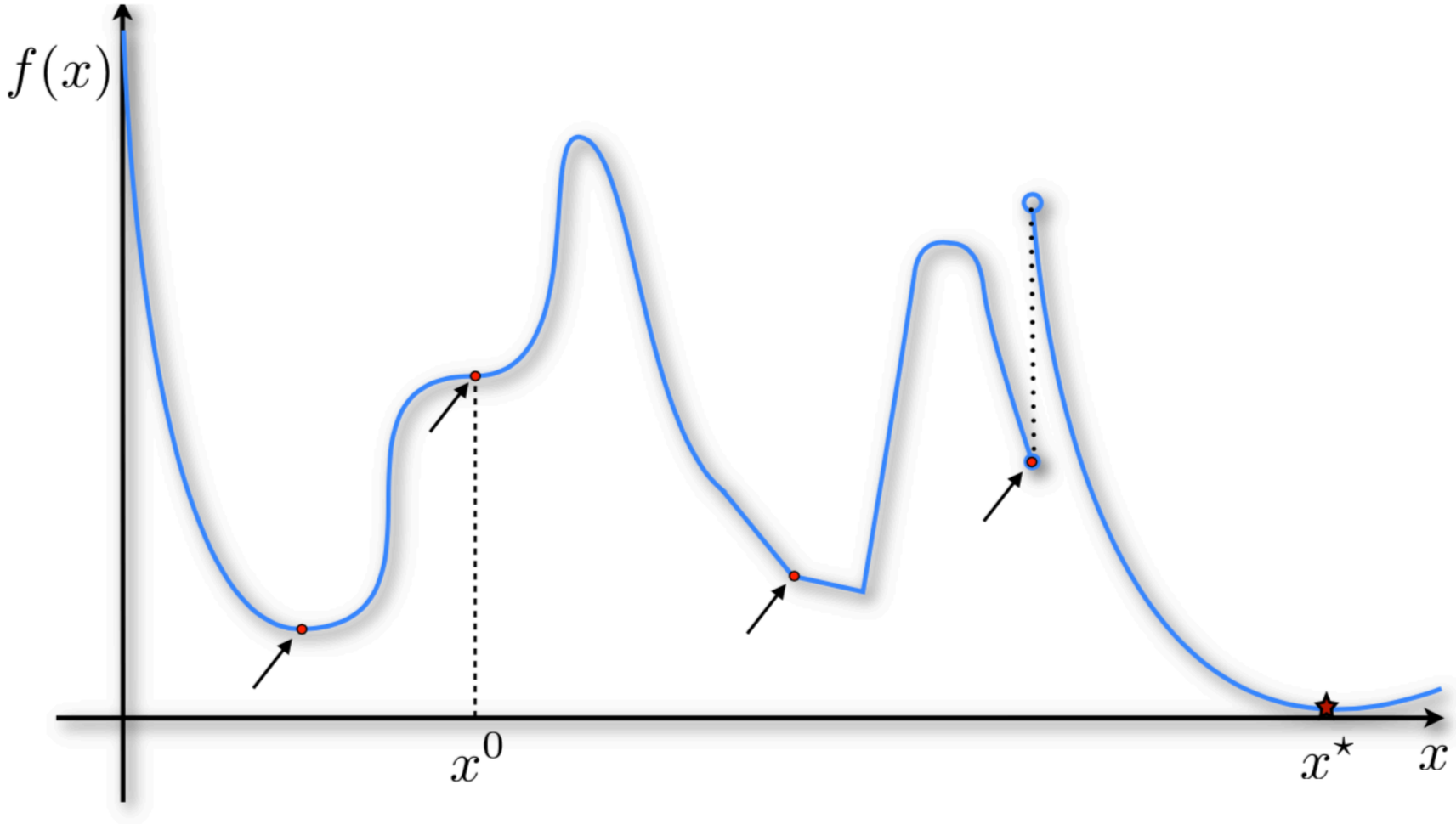
Agnostic optimization

Demo

Agnostic optimization



Agnostic optimization



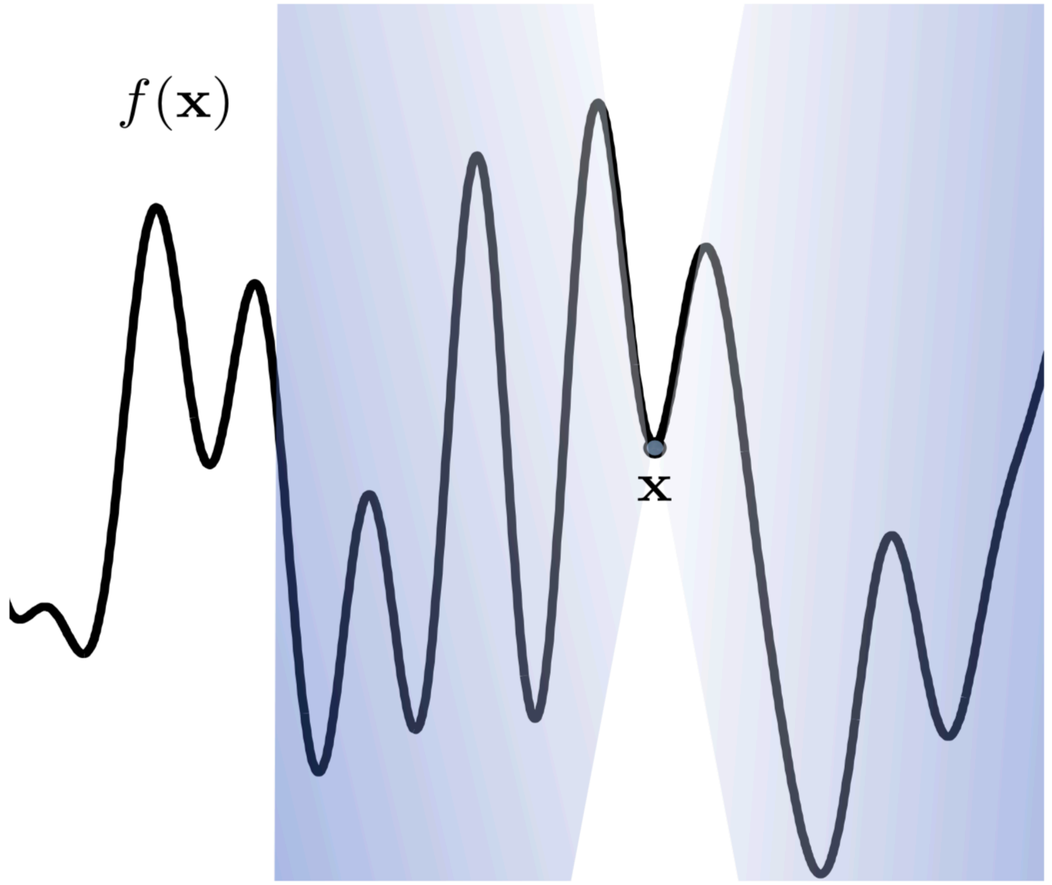
Lipschitz conditions

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

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

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

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

- Another important one:

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

Interpretation?

Lipschitz conditions

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

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

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

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

$$\vdots \quad \vdots$$

- Another important one:

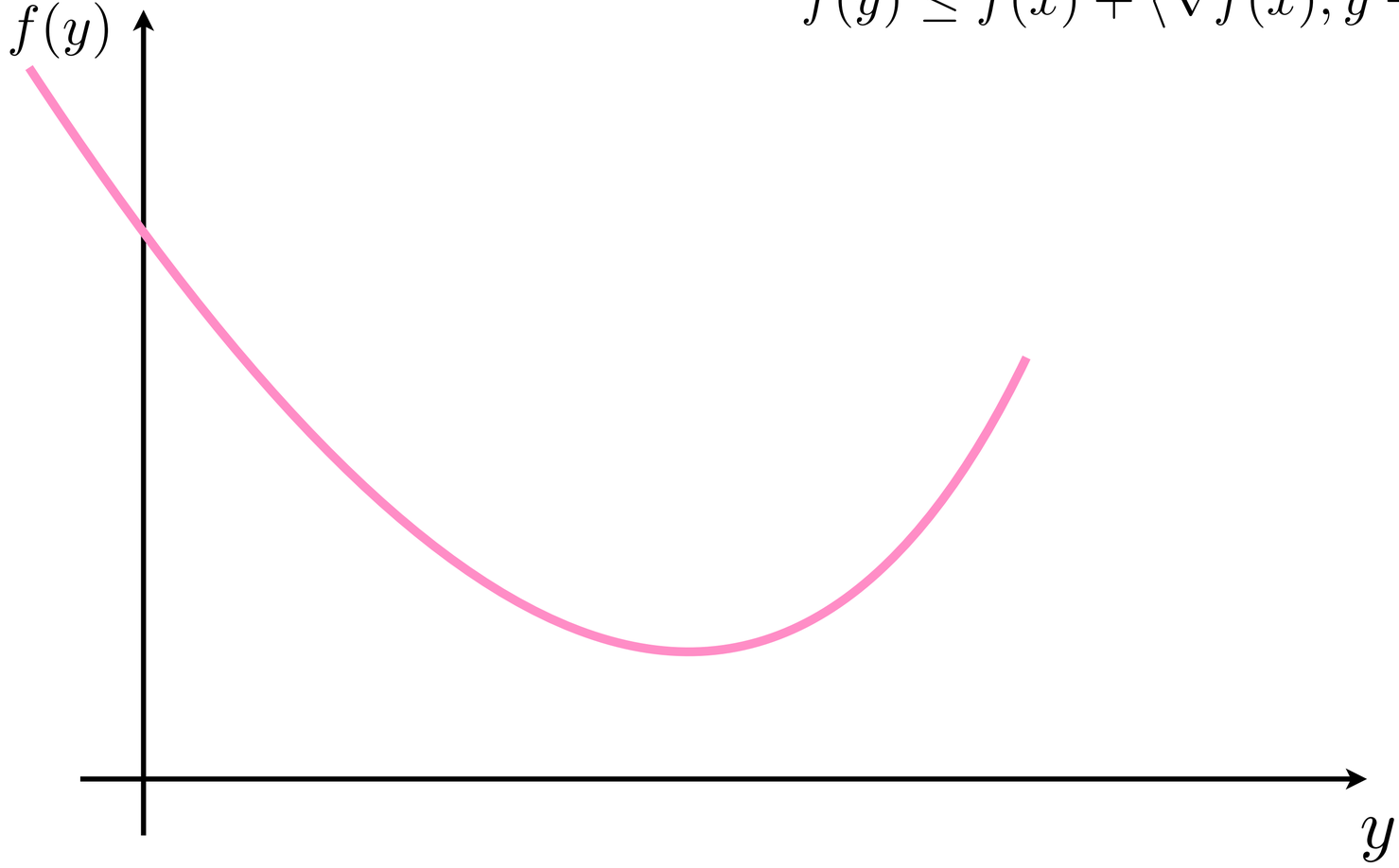
$$\nabla^2 f(x) \preceq LI$$

Interpretation?

Lipschitz conditions

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

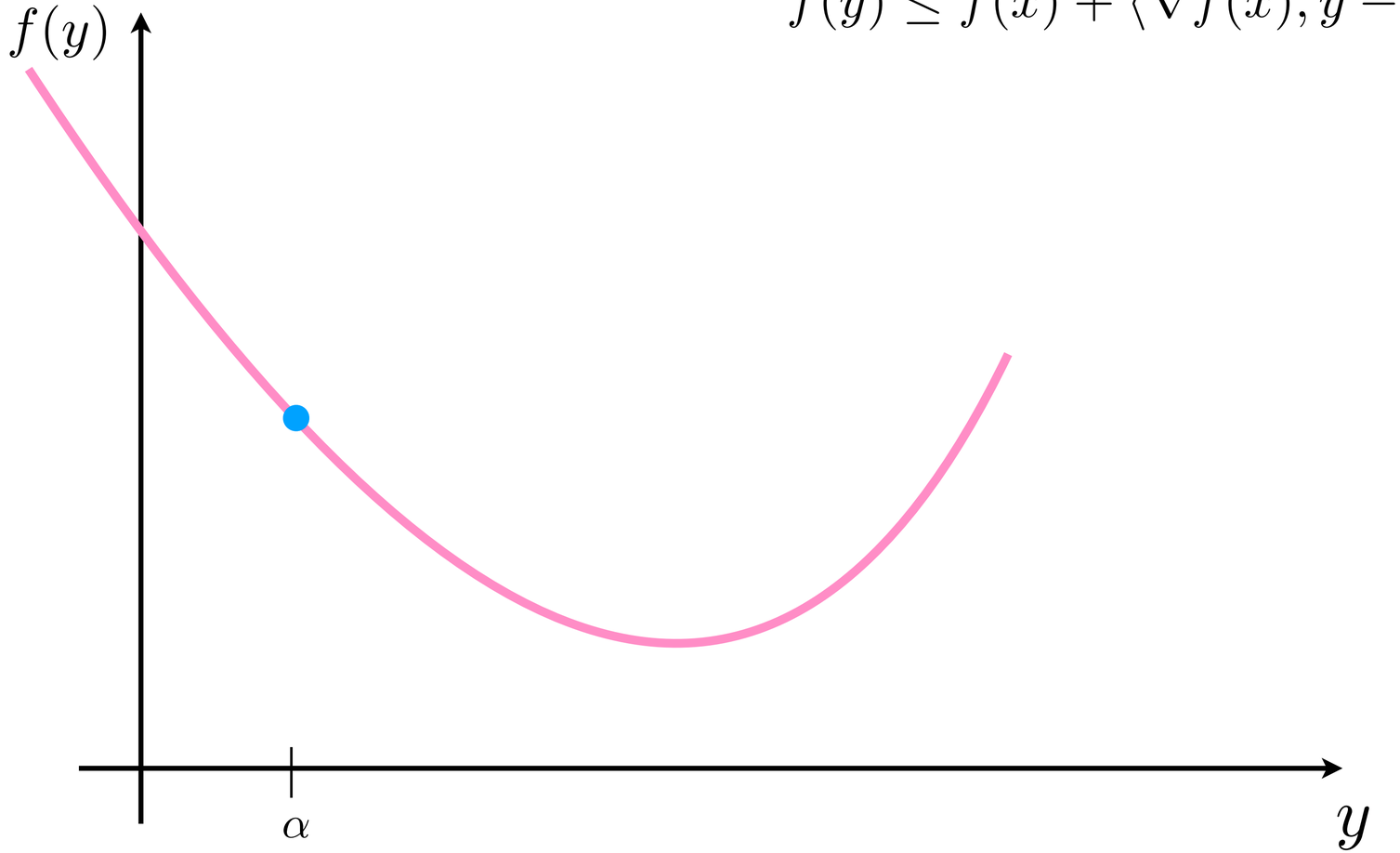
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

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

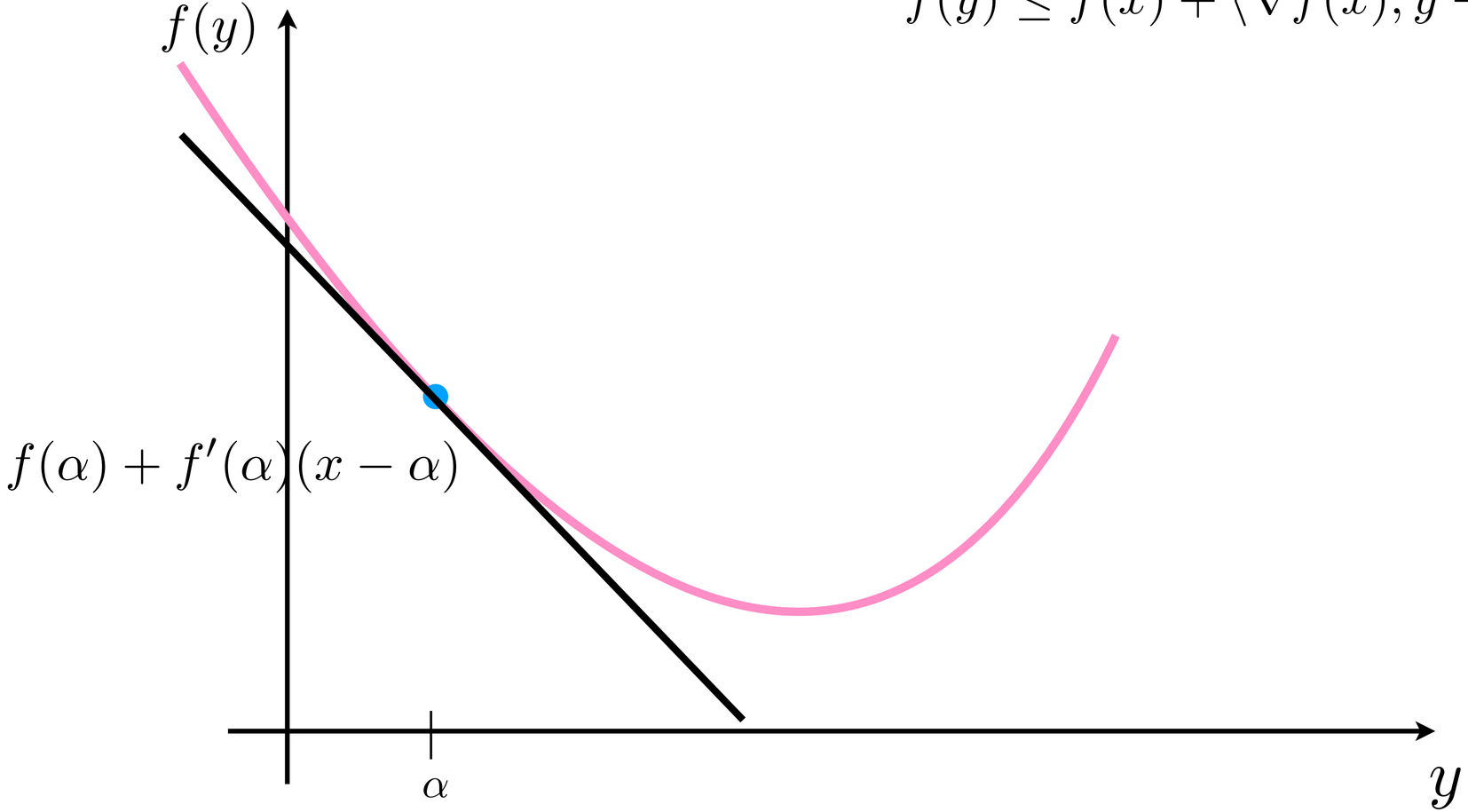
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

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

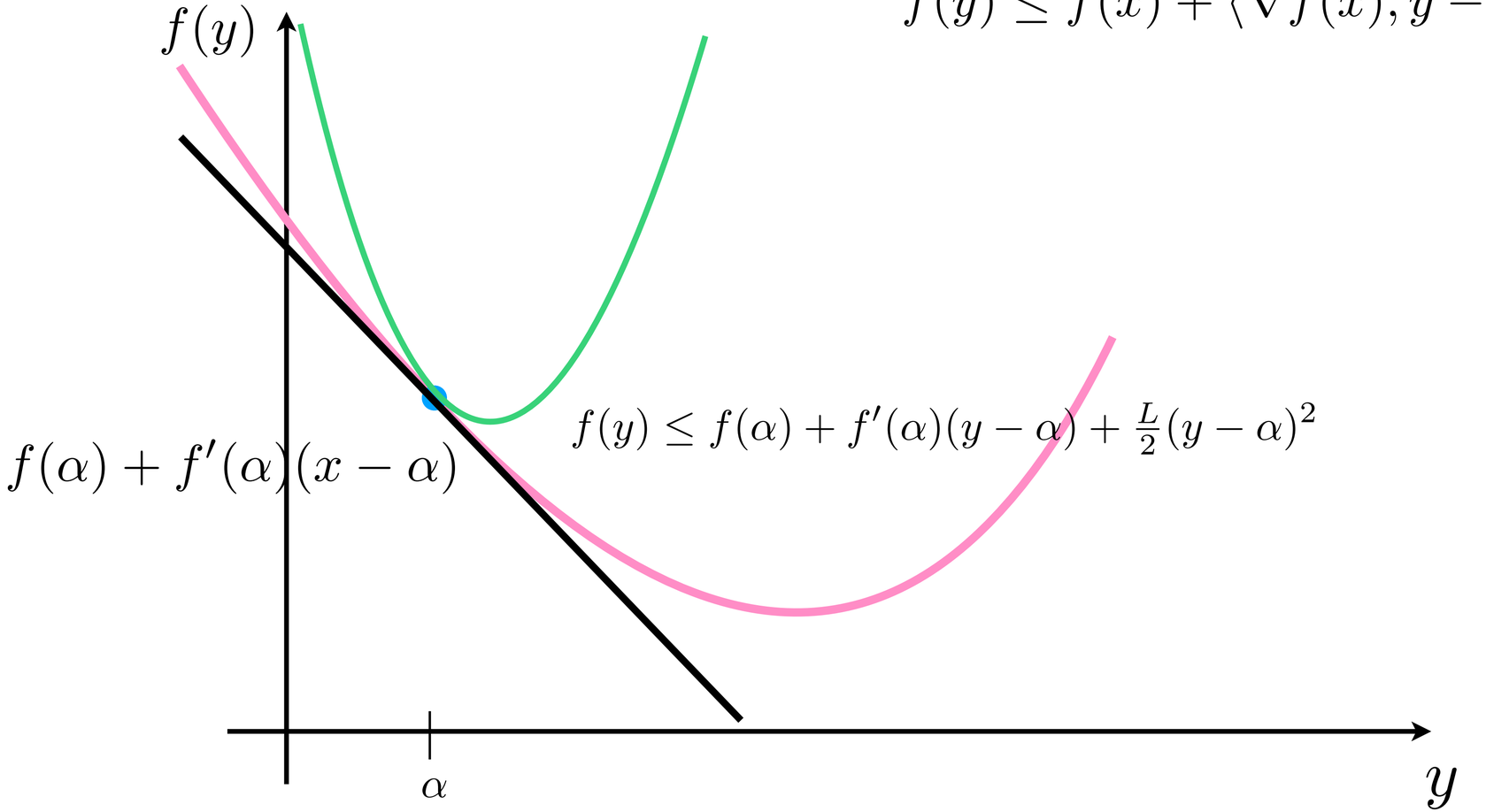
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2$$



Lipschitz conditions

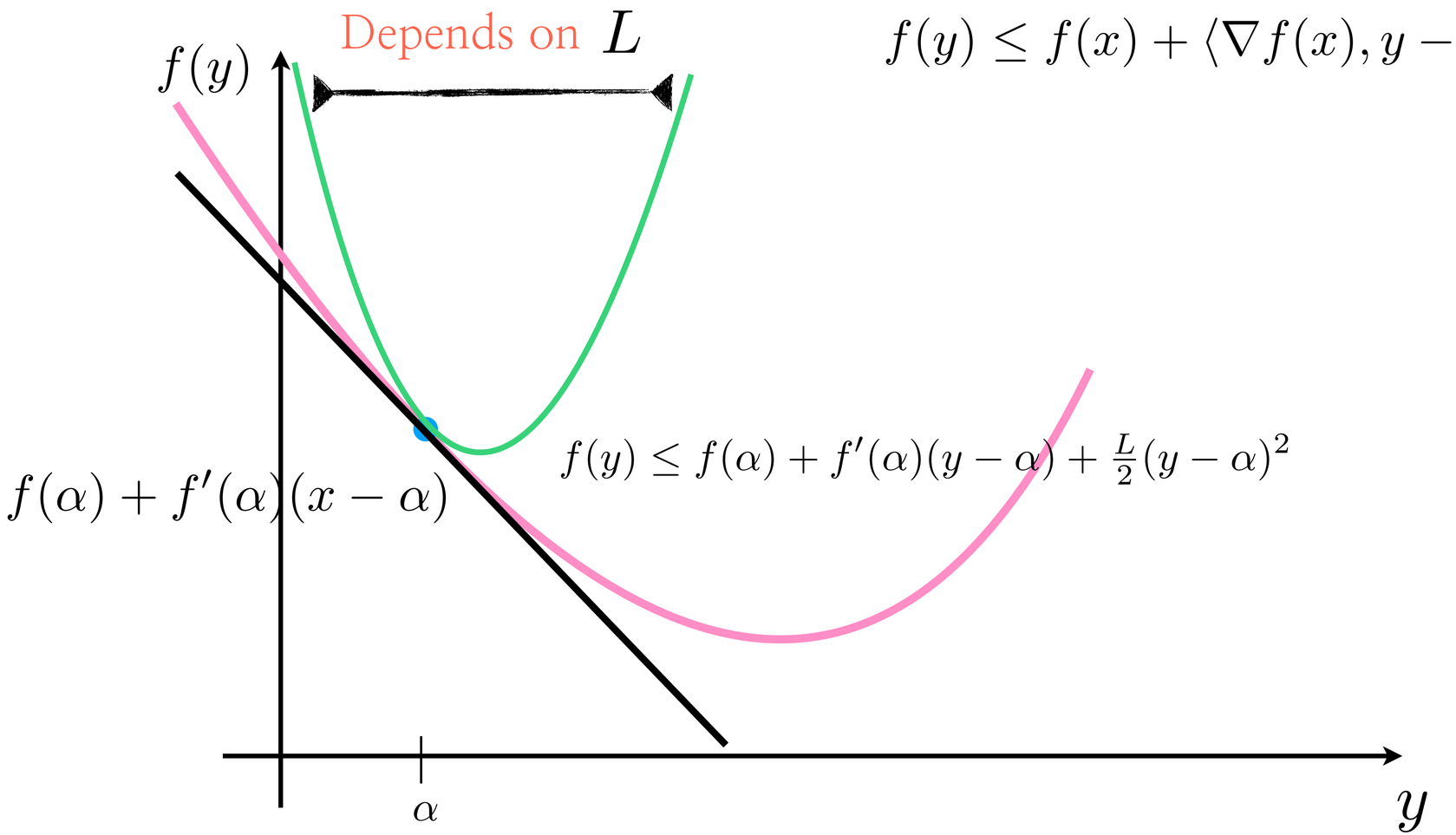
- Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

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



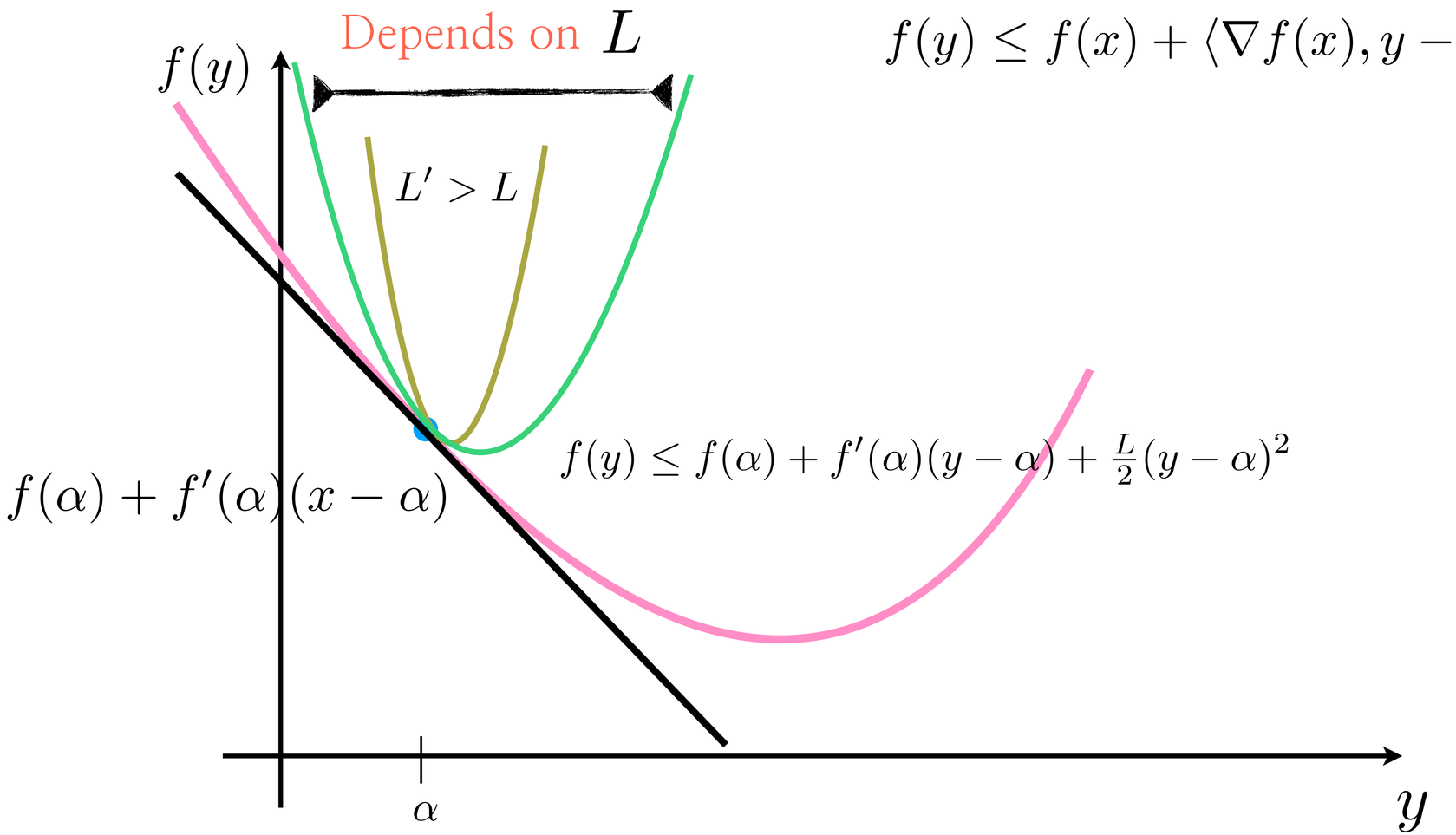
Lipschitz conditions

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



Lipschitz conditions

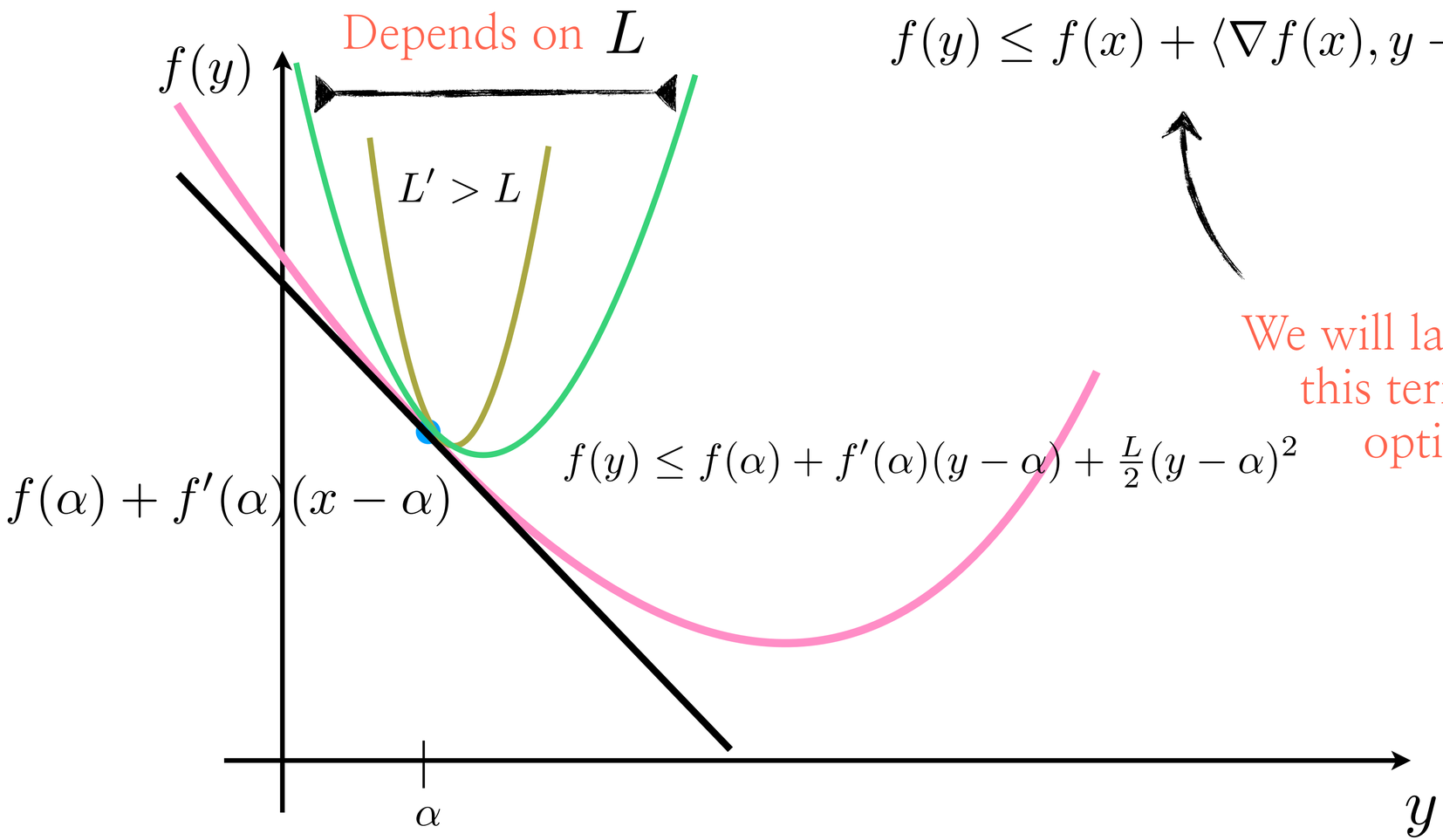
- Equivalent characterizations: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$



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

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:

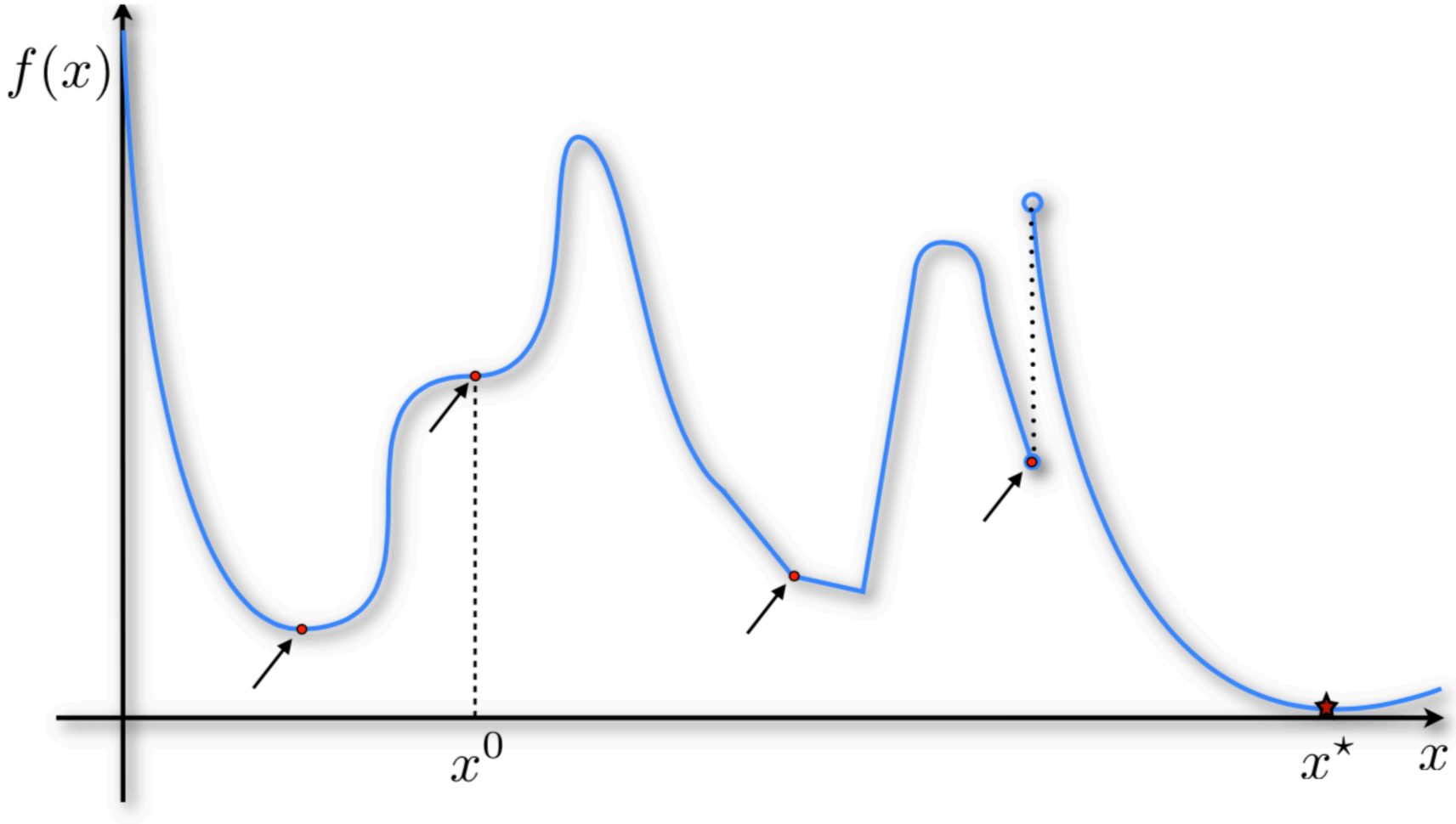
$$\nabla^2 f(x) \preceq LI \Rightarrow \|\nabla^2 f(x)\|_2 \leq \|LI\|_2 \Rightarrow \|\nabla^2 f(x)\|_2 \leq L$$

- Then:

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

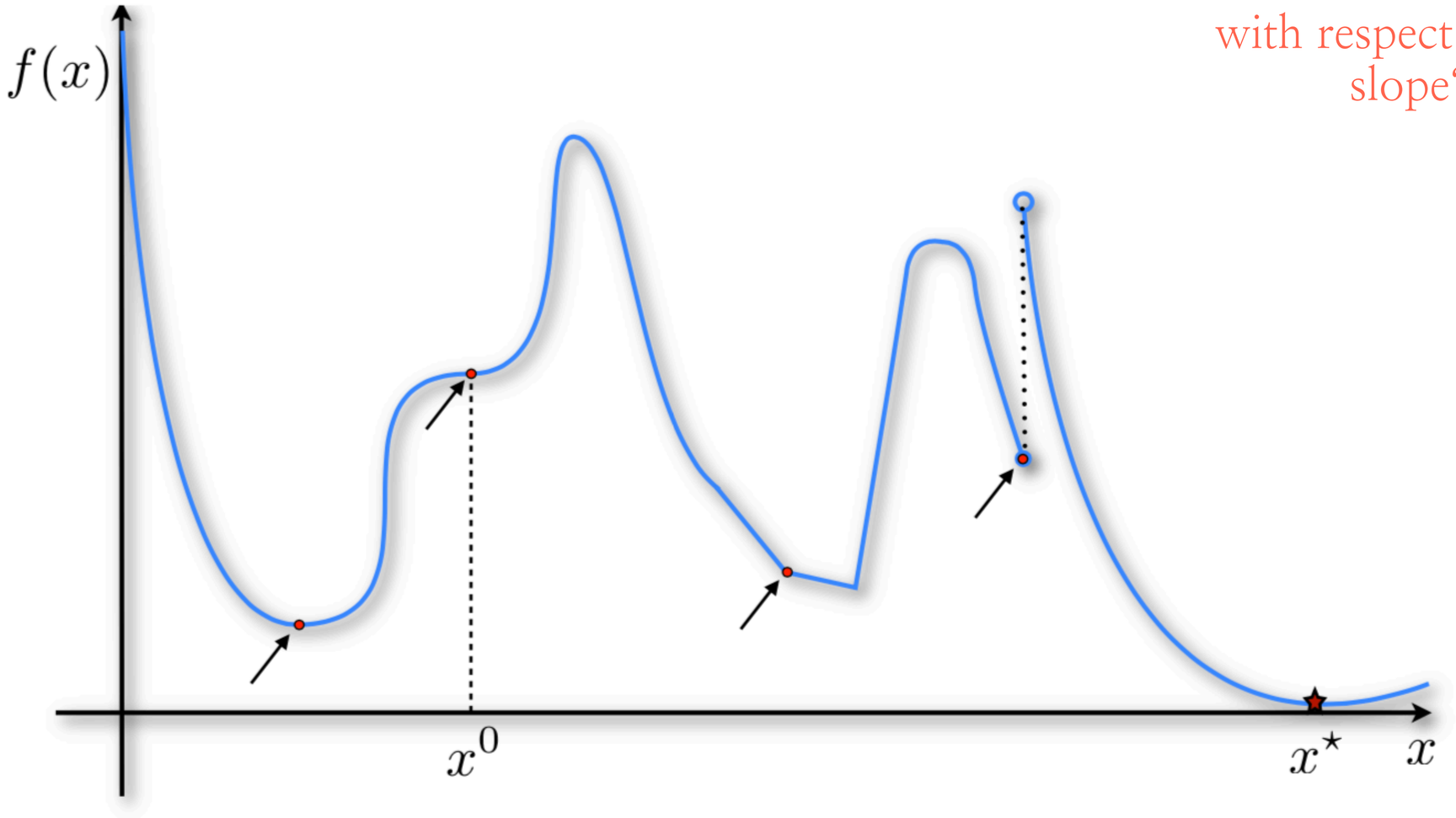
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2$$

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

(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?

Necessary

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

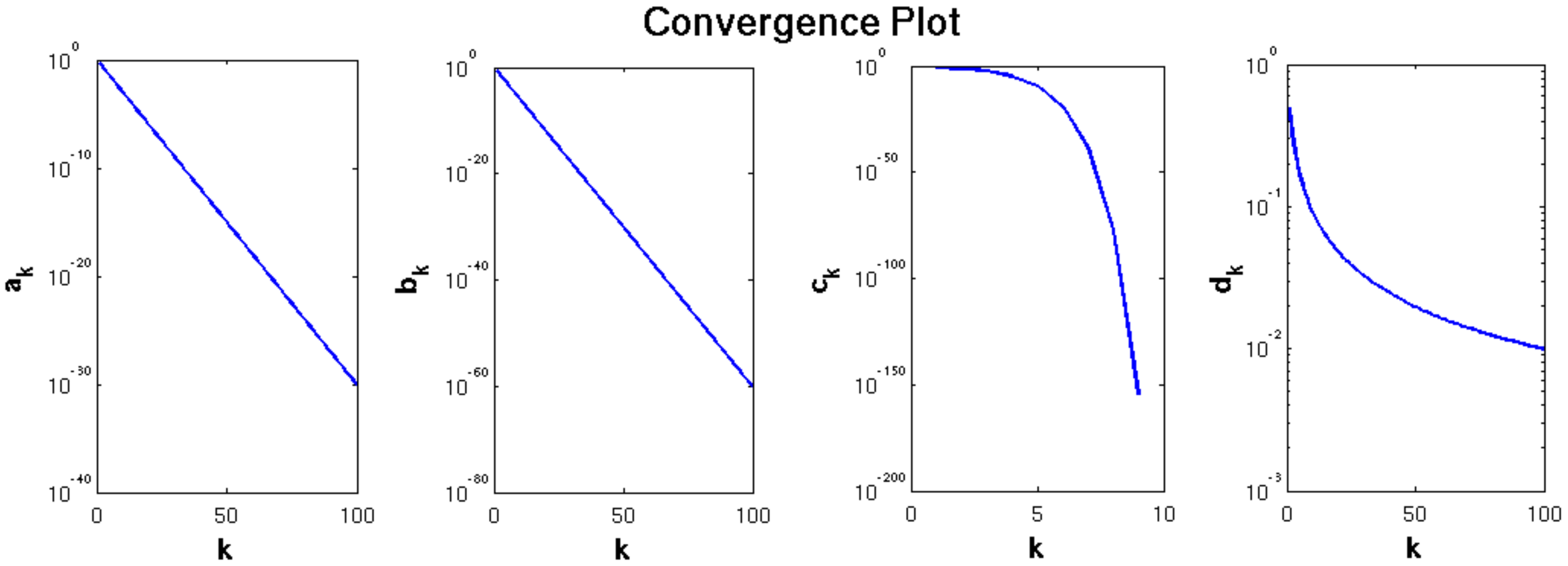
- 2^{nd}–order optimality condition: $\nabla f(\widehat{x}) = 0$ and $\nabla^2 f(\widehat{x}) \succeq 0$

First convergence result

Whiteboard

Convergence rates 101

(Source: Wikipedia)



$O(\log 1/\varepsilon)$

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

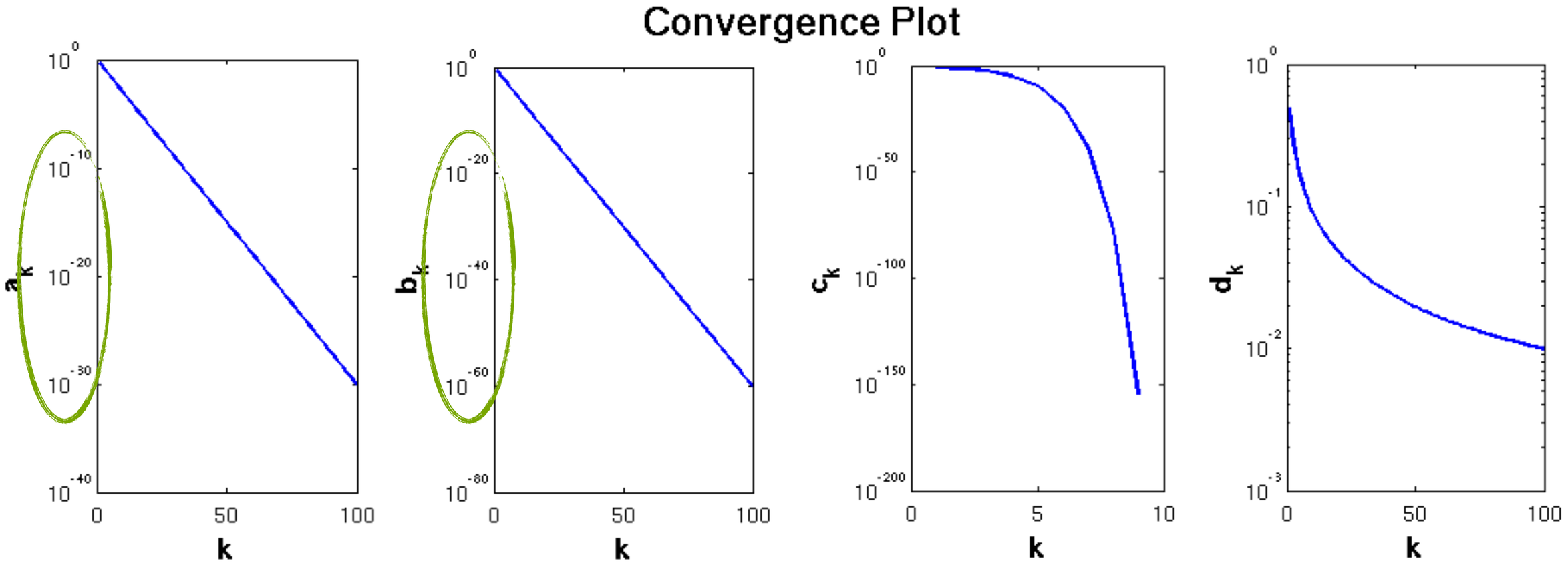
$O\left(\log\log(1/\varepsilon)\right)$

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

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

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.,

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

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

"But these are actually convex!"

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!"

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

Demo

“What does convexity bring onto the table?”

Convex functions

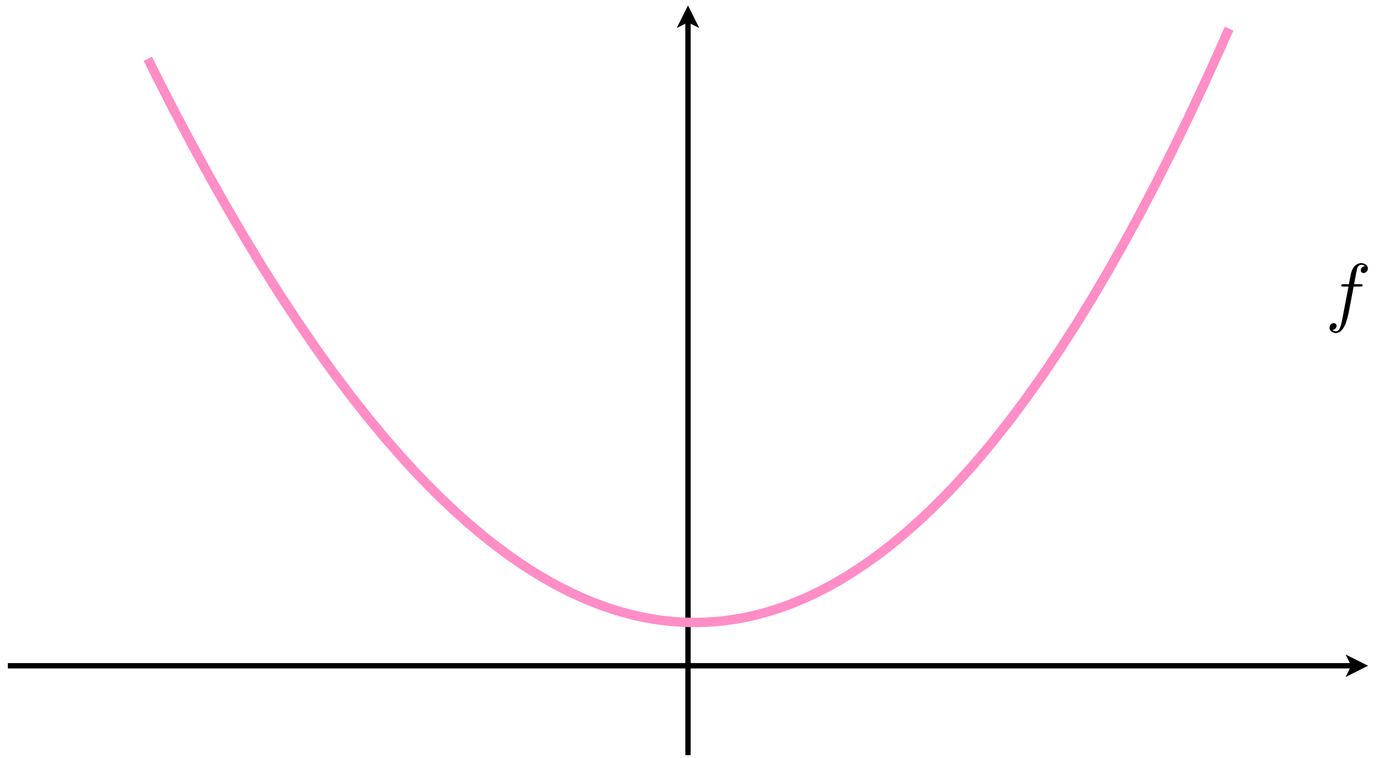
- General definition:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y),\ \forall a \in [0,1]$$

Convex functions

- General definition:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \ \forall a \in [0,1]$$

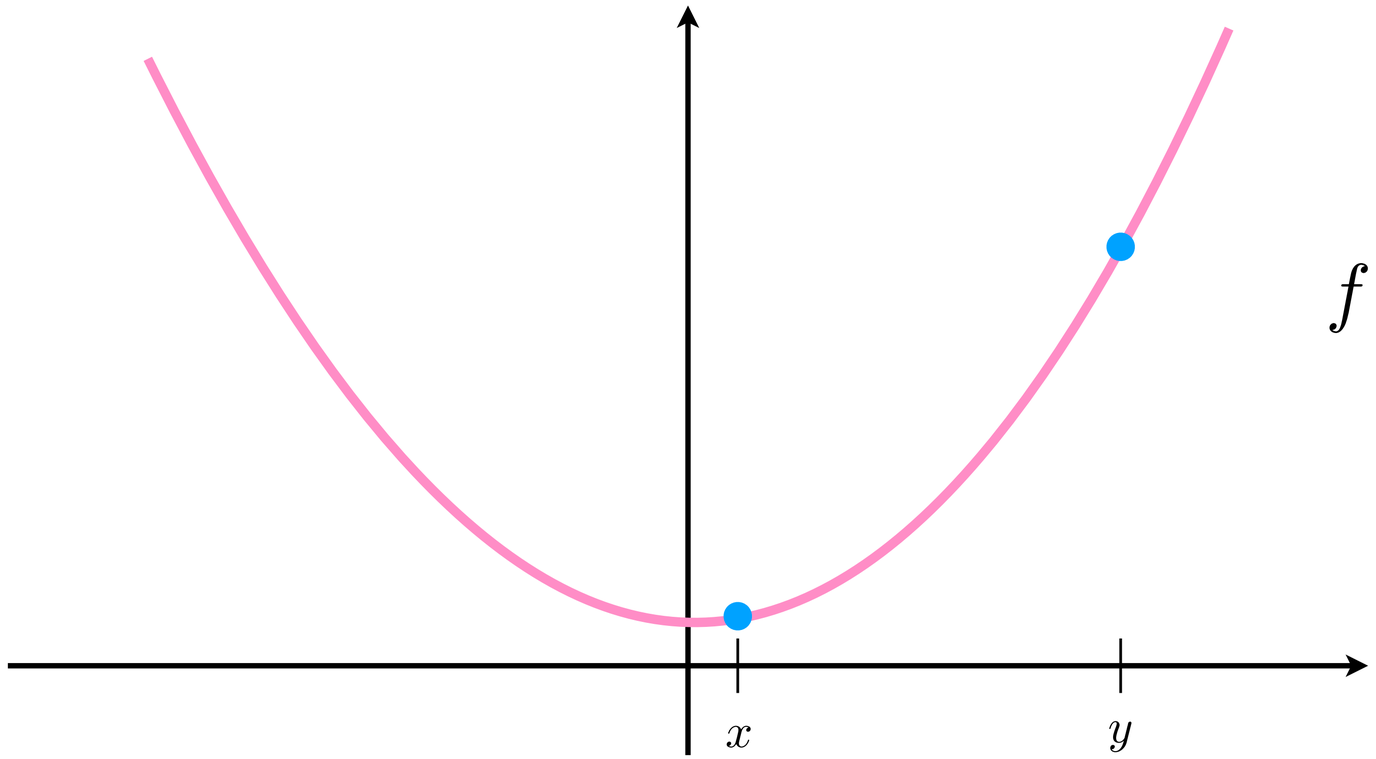


$$f(x) = x^2 + \beta$$

Convex functions

- General definition:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \ \forall a \in [0,1]$$

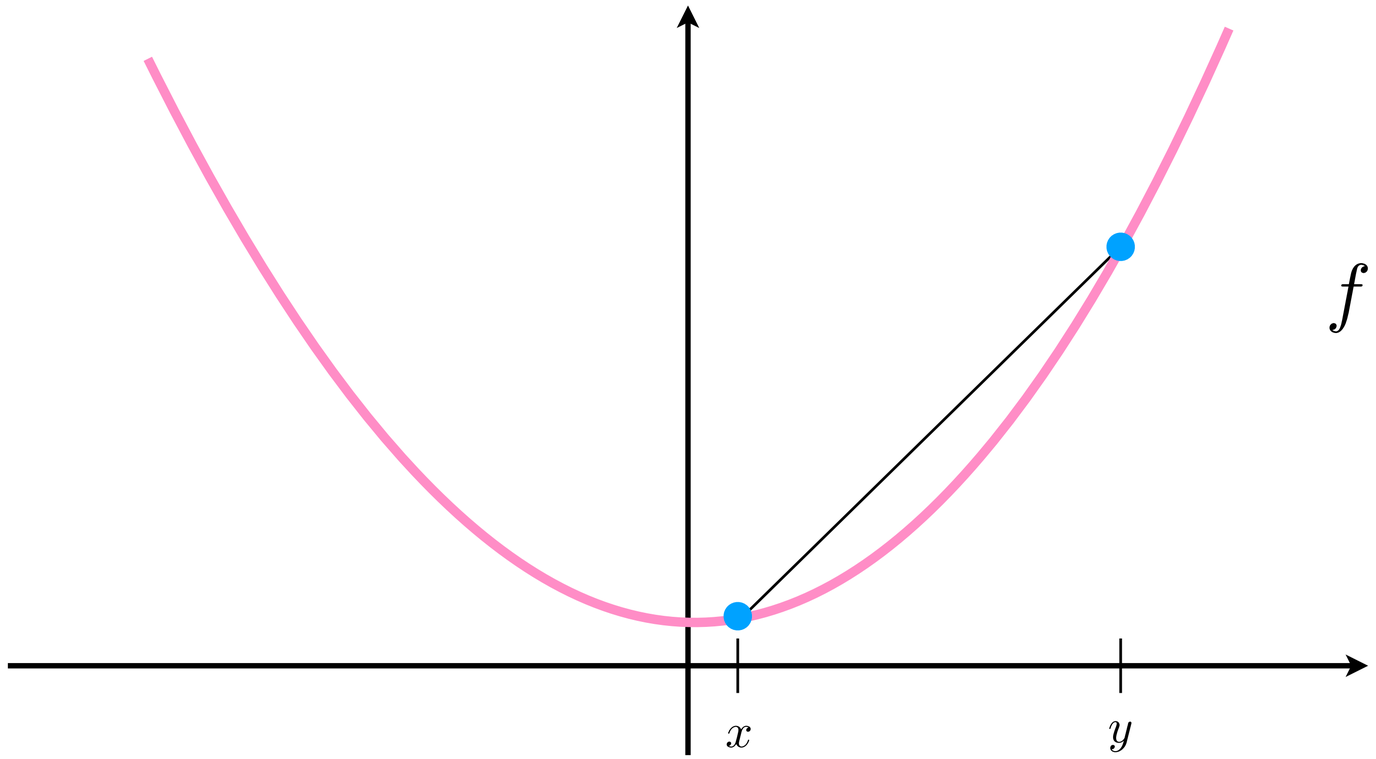


$$f(x) = x^2 + \beta$$

Convex functions

- General definition:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \ \forall a \in [0,1]$$

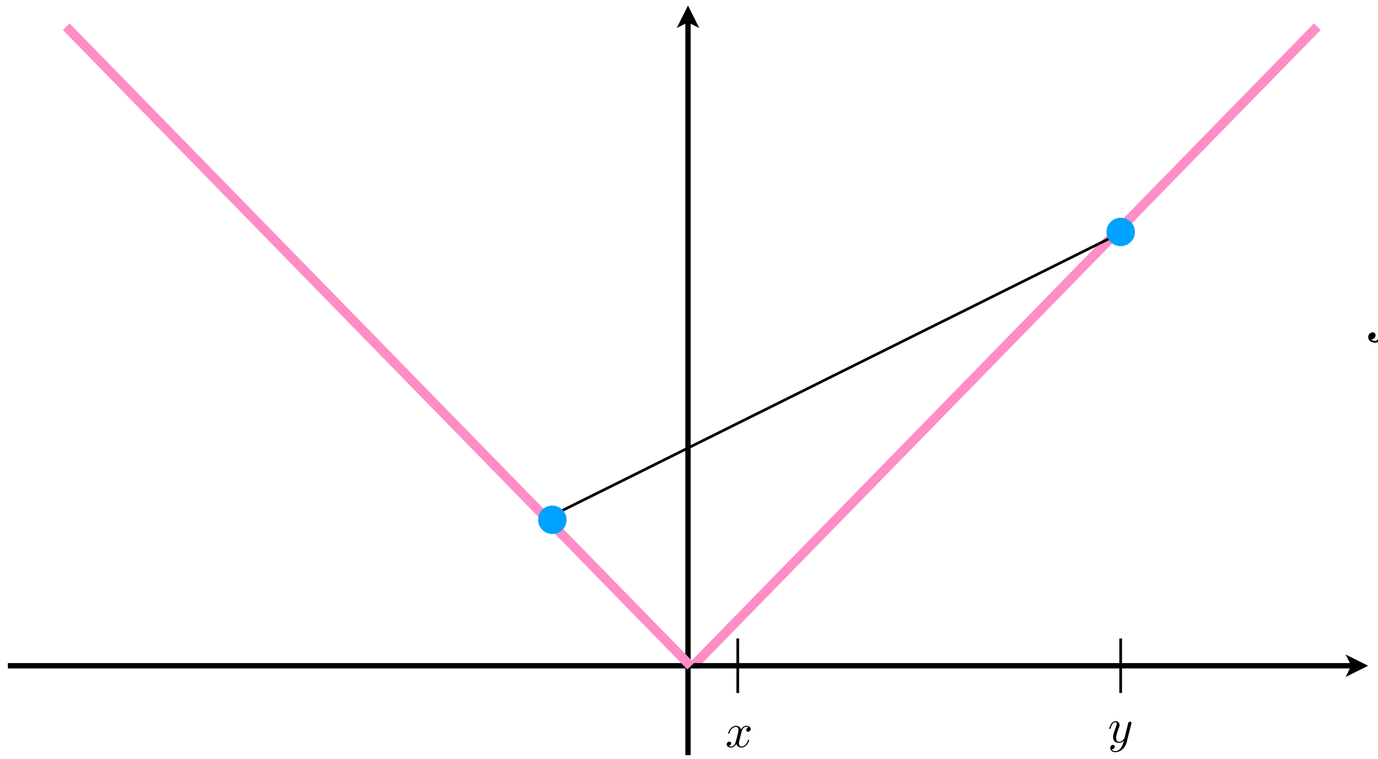


$$f(x) = x^2 + \beta$$

Convex functions

- General definition:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \ \forall a \in [0,1]$$



$$f(x) = |x|$$

Convex functions

- Examples:

Function	Example	Attributes
ℓ_p vector norms, $p \geq 1$	$\|\mathbf{x}\|_2, \|\mathbf{x}\|_1, \|\mathbf{x}\|_\infty$	convex
ℓ_p matrix norms, $p \geq 1$	$\|\mathbf{X}\|_* = \sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i$	convex
Square root function	$\sqrt{x}$	concave, nondecreasing
Maximum of functions	$\max\{x_1, \ldots, x_n\}$	convex, nondecreasing
Minimum of functions	$\min\{x_1, \ldots, x_n\}$	concave, nondecreasing
Sum of convex functions	$\sum_{i=1}^n f_i$, f_i convex	convex
Logarithmic functions	$\log(\det(\mathbf{X}))$	concave, assumes $\mathbf{X} \succ 0$
Affine/linear functions	$\sum_{i=1}^n X_{ii}$	both convex and concave
Eigenvalue functions	$\lambda_{max}(\mathbf{X})$	convex, assumes $\mathbf{X} = \mathbf{X}^T$

Convex functions

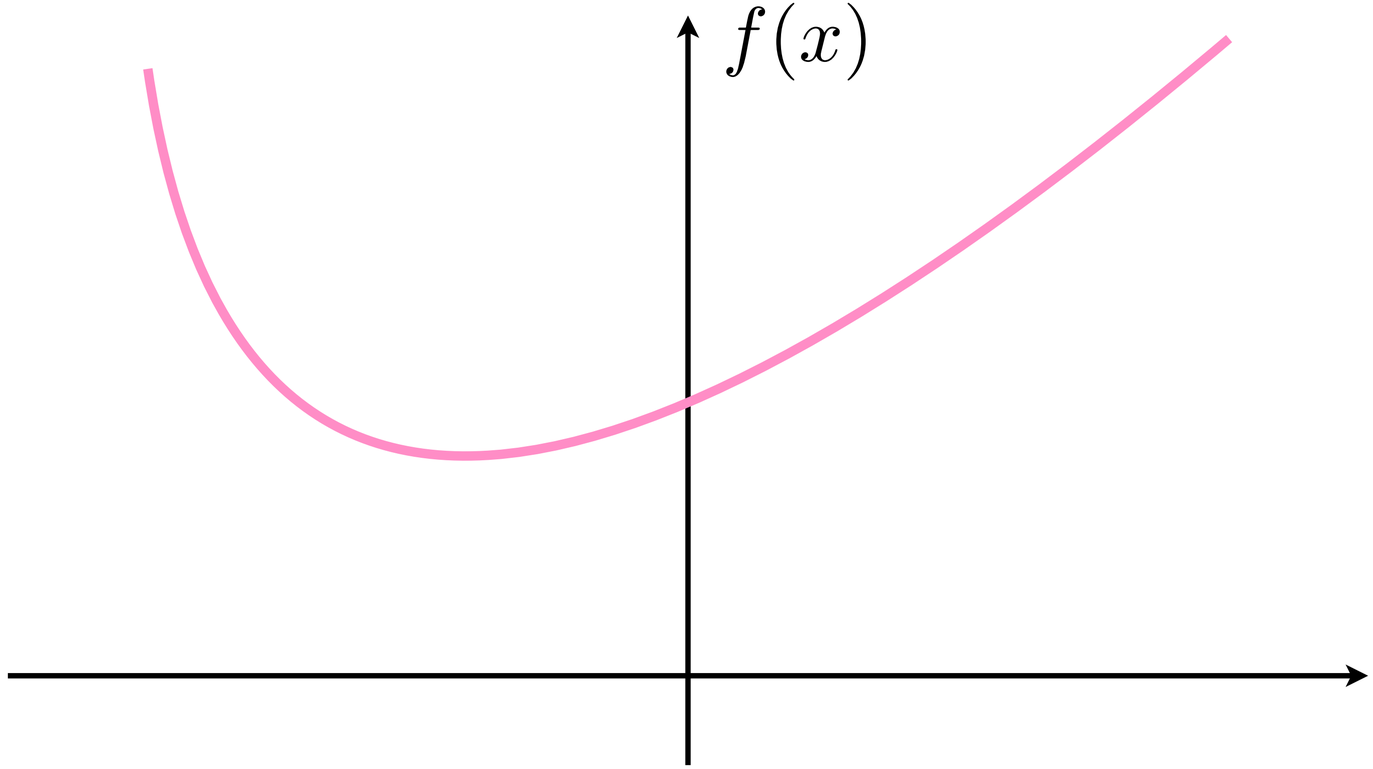
- Alternative (more practical) definitions of convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \,, \quad \forall x, y$$

Convex functions

- Alternative (more practical) definitions of convexity

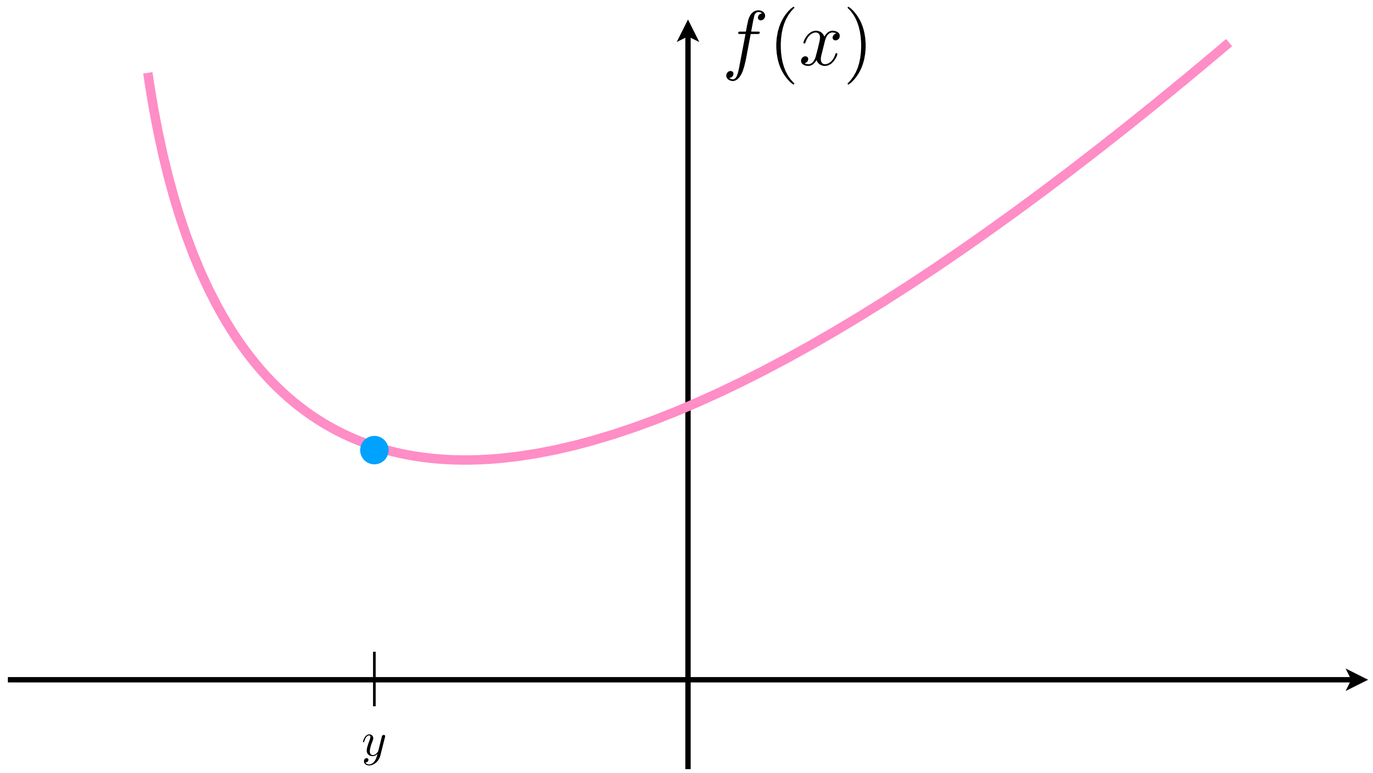
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \,, \quad \forall x, y$$



Convex functions

- Alternative (more practical) definitions of convexity

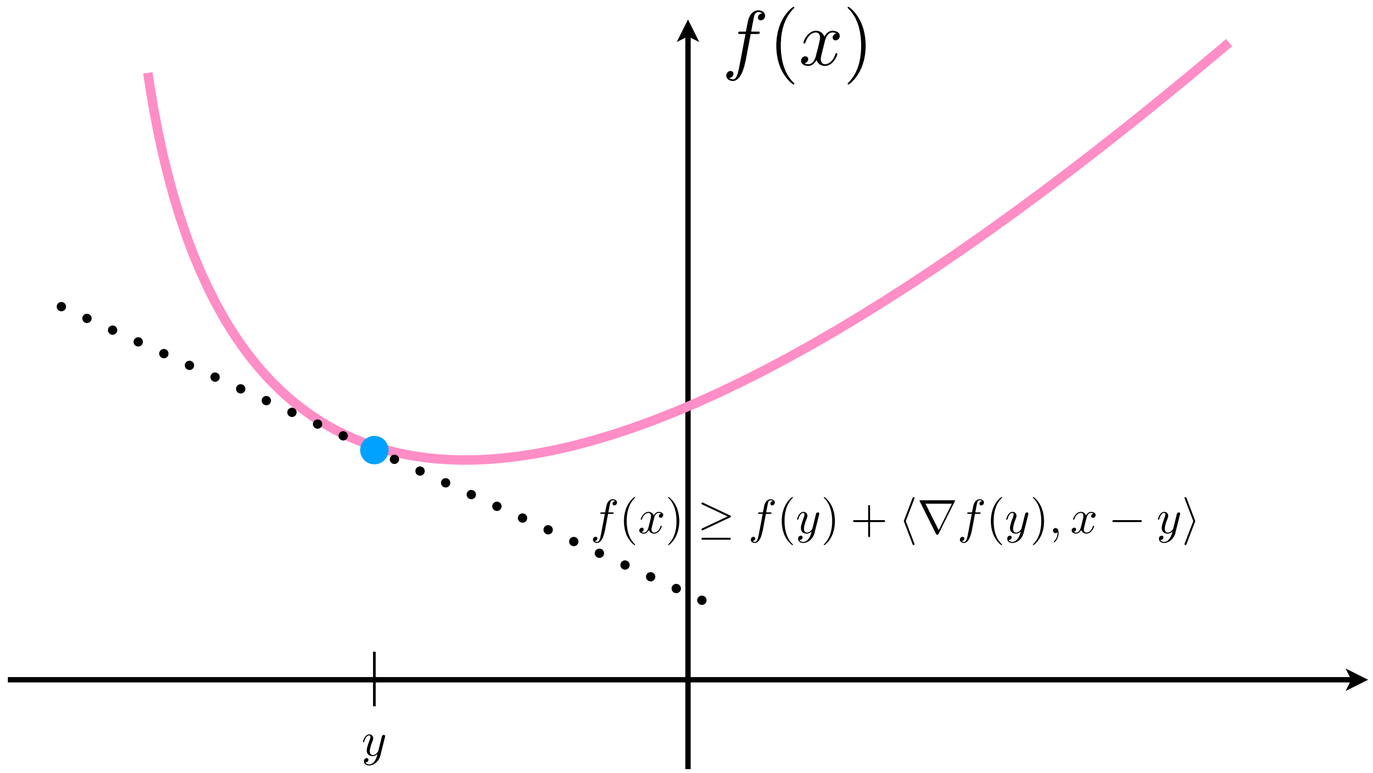
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \,, \quad \forall x, y$$



Convex functions

- Alternative (more practical) definitions of convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \,, \quad \forall x, y$$



Convex functions

- Alternative (more practical) definitions of convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \,, \quad \forall x, y$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \ \forall x, y$$

$$\nabla^2 f(x) \succeq 0, \ \forall x$$

(Assuming the function is twice differentiable)

Convex functions

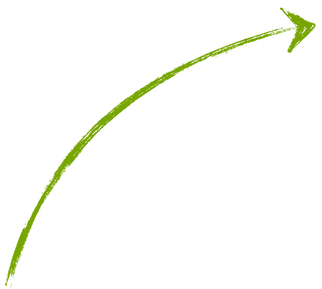
- Alternative (more practical) definitions of convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \ \forall x, y$$

$$\nabla^2 f(x) \succeq 0, \ \forall x$$

(Assuming the function is twice differentiable)



Any interpretations?

Convex functions

- Key consequences of convexity

“Any stationary point is a global minimum”

Proof:

Convex functions

- Key consequences of convexity

“Any stationary point is a global minimum”

Proof: Assume a stationary point $x^\star$. This implies $\nabla f(x^\star) = 0$

Convex functions

- Key consequences of convexity

"Any stationary point is a global minimum"

Proof: Assume a stationary point $x^\star$. This implies $\nabla f(x^\star) = 0$

By convexity:

$$f(x) \geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle = f(x^\star), \ \forall x$$

- This is what makes convex optimization preferable.

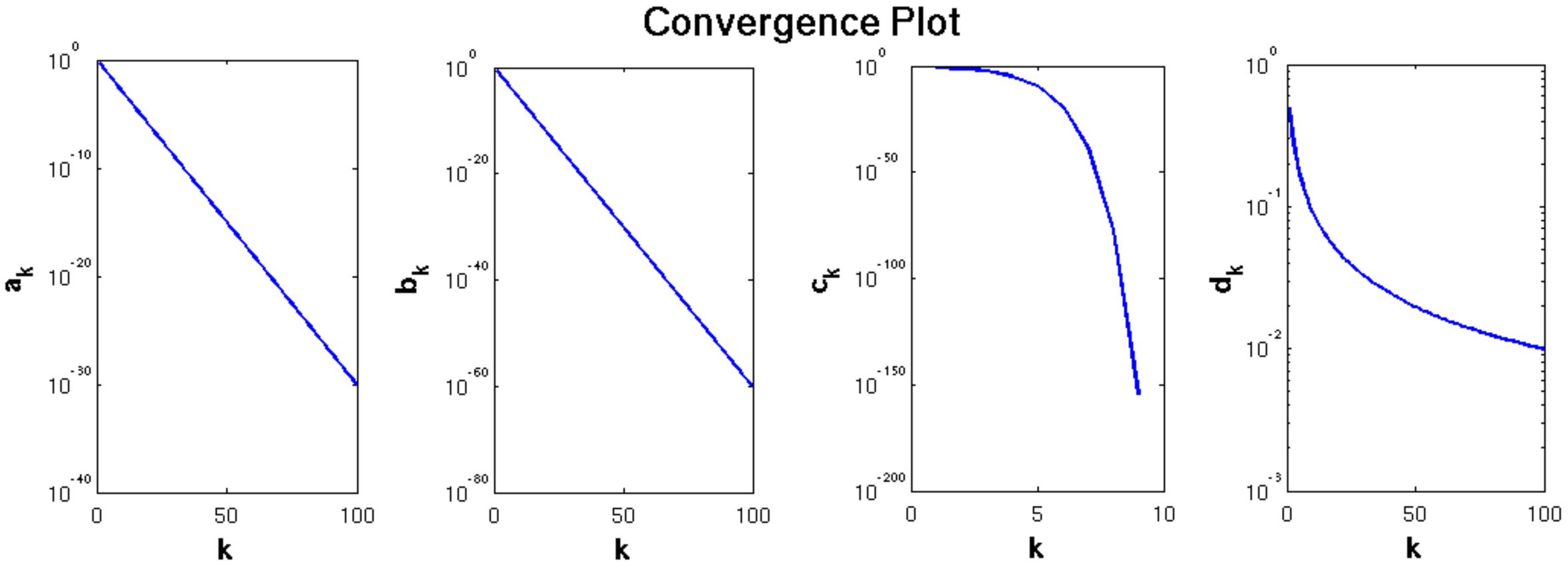
(Any local solution is actually global –
this does not mean that convex optimization is necessarily tractable!)

Does convexity improve guarantees?

Whiteboard

Convergence rates 101

(Source: Wikipedia)



$O(\log 1/\varepsilon)$

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

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

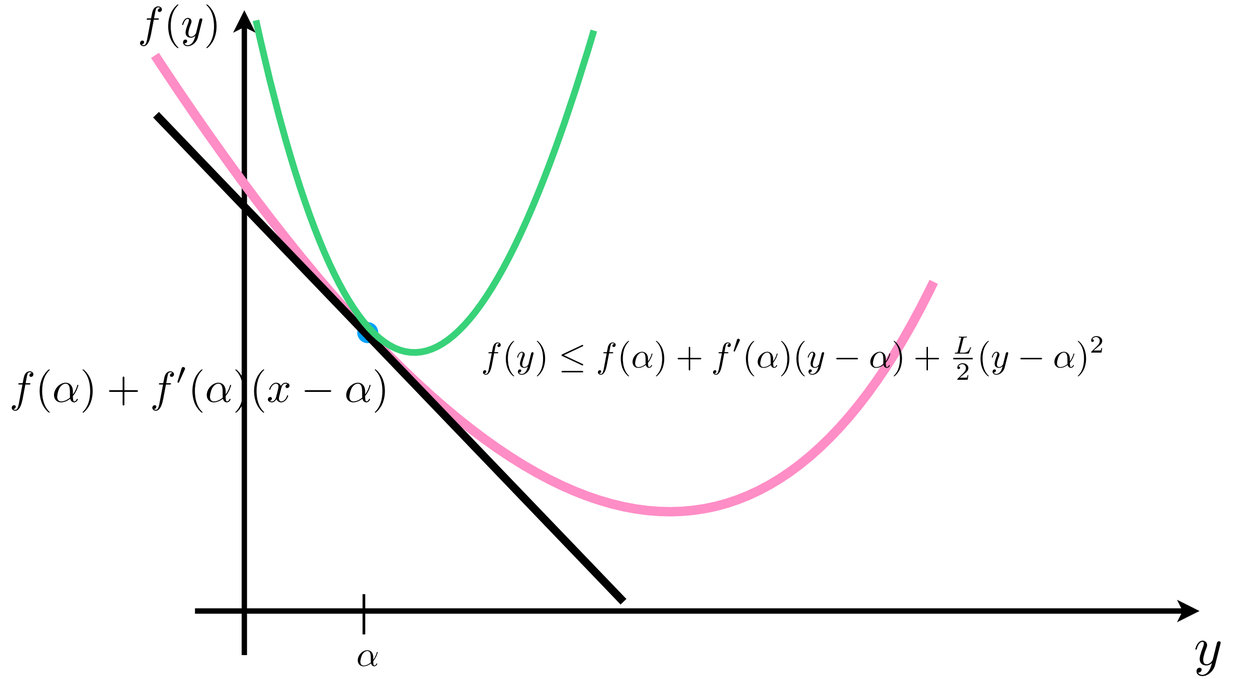
$O(1/\varepsilon^2),\ O(1/\varepsilon),\ O(1/\sqrt{\varepsilon})$

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

Can we achieve a better performance?

- Strong convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \ \forall x, y$
  - Strong convexity parameter: $\mu > 0$

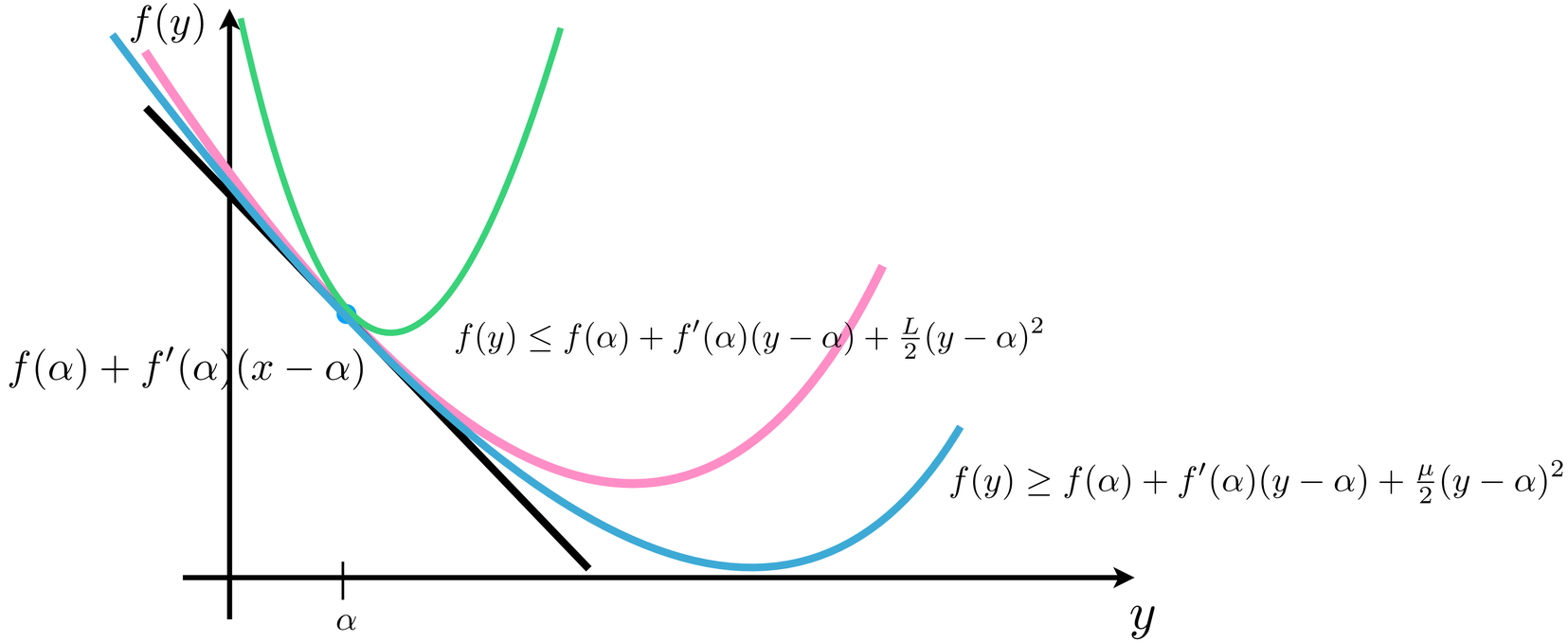
"Strong convexity implies that f should be steep enough to make progress"



Can we achieve a better performance?

- Strong convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \ \forall x, y$
  - Strong convexity parameter: $\mu > 0$

"Strong convexity implies that f should be steep enough to make progress"



Strong convexity

- Equivalent characterizations: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \ \forall x, y$

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

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

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

$$\vdots \qquad \vdots$$

Strong convexity

- Equivalent characterizations: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \ \forall x, y$

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

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

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

$$\vdots \qquad \vdots$$

- Another important one:

Interpretation?

$$\nabla^2 f(x) \succeq \mu I$$

Strong convexity

- Equivalent characterizations: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \ \forall x, y$

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

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

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

$$\vdots \qquad \vdots$$

- Another important one:

Interpretation?

$$\nabla^2 f(x) \succeq \mu I$$

- What if we also have Lipschitz continuous gradients?

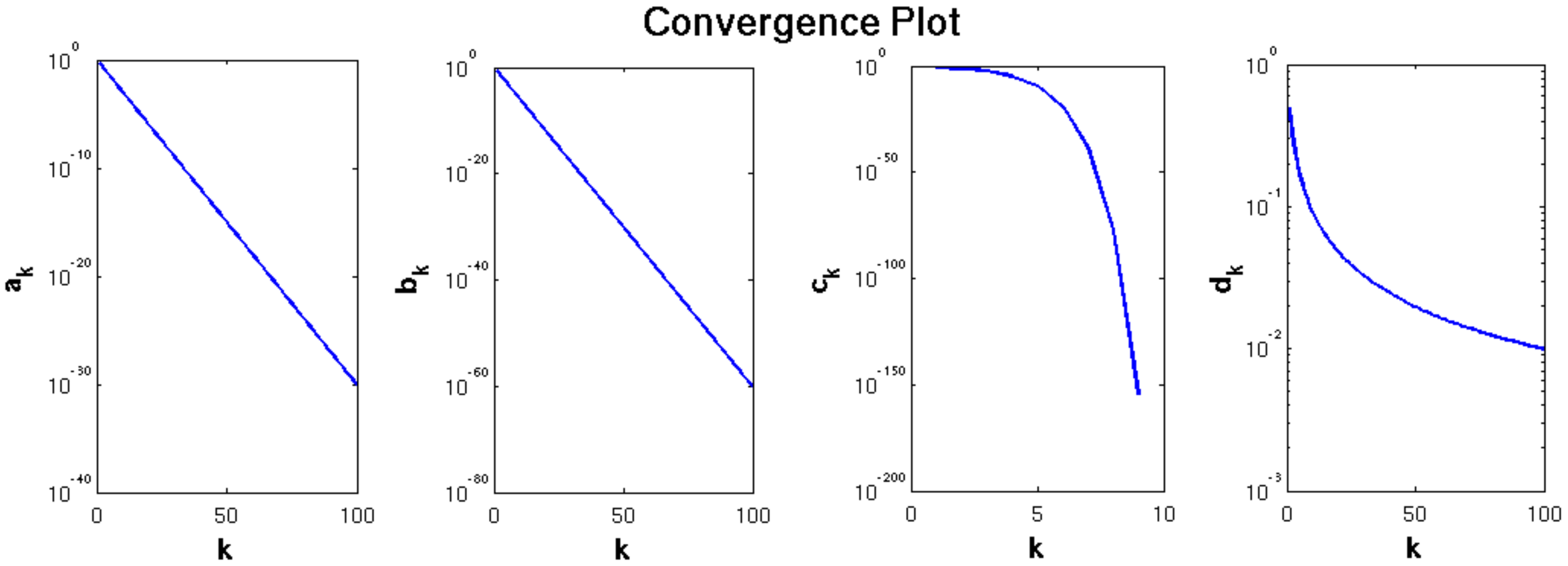
$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L}\|x - y\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

What is the gain?

Whiteboard

Convergence rates 101

(Source: Wikipedia)



$O(\log 1/\varepsilon)$

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

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

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

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

What should be our expectations: Lower bounds

- For objectives with Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

(Under these assumptions, and using only gradients, we cannot achieve better than $o\left(\frac{1}{t^2}\right)$)

What should be our expectations: Lower bounds

- For objectives with Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

(Under these assumptions, and using only gradients, we cannot achieve better than $o\left(\frac{1}{t^2}\right)$)

- In addition, for objectives that are strongly convex:

$$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$

(The case we described has near optimal exponent, but does not involve the square root of κ)

What should be our expectations: Lower bounds

- For objectives with Lipschitz continuous gradients:

$$f(x_t) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(t+1)^2}$$

(Under these assumptions, and using only gradients, we cannot achieve better than $o\left(\frac{1}{t^2}\right)$)

- In addition, for objectives that are strongly convex:

$$\|x_t - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - x^\star\|_2^2 \qquad \kappa := \frac{L}{\mu}$$

(The case we described has near optimal exponent, but does not involve the square root of κ)

- In future lectures: acceleration techniques that achieves these rates

Convex optimization

Demo

Are there other, more powerful, global assumptions?

- Remember, our analysis is based on: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

(and on $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2$ when we talk about convex functions)

Are there other, more powerful, global assumptions?

- Remember, our analysis is based on: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

(and on $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2$ when we talk about convex functions)

- Polyak–Lojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \xi(f(x) - f(x^\star)), \quad \forall x, \quad \text{for some } \xi > 0$$

(Any thoughts about what this implies?)

Are there other, more powerful, global assumptions?

- Remember, our analysis is based on: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

(and on $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2$ when we talk about convex functions)

- Polyak–Lojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \xi(f(x) - f(x^\star)), \quad \forall x, \quad \text{for some } \xi > 0$$

(Any thoughts about what this implies?)

- Using PL inequality + Lipschitz gradient continuity:

$$f(x_t) - f(x^\star) \leq \left(1 - \frac{\xi}{L}\right)^t (f(x_0) - f(x^\star))$$ Whiteboard

Are there other, more powerful, global assumptions?

- Remember, our analysis is based on: $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y$

(and on $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2$ when we talk about convex functions)

- Polyak–Lojasiewicz (PL) inequality

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \xi(f(x) - f(x^\star)), \quad \forall x, \quad \text{for some } \xi > 0$$

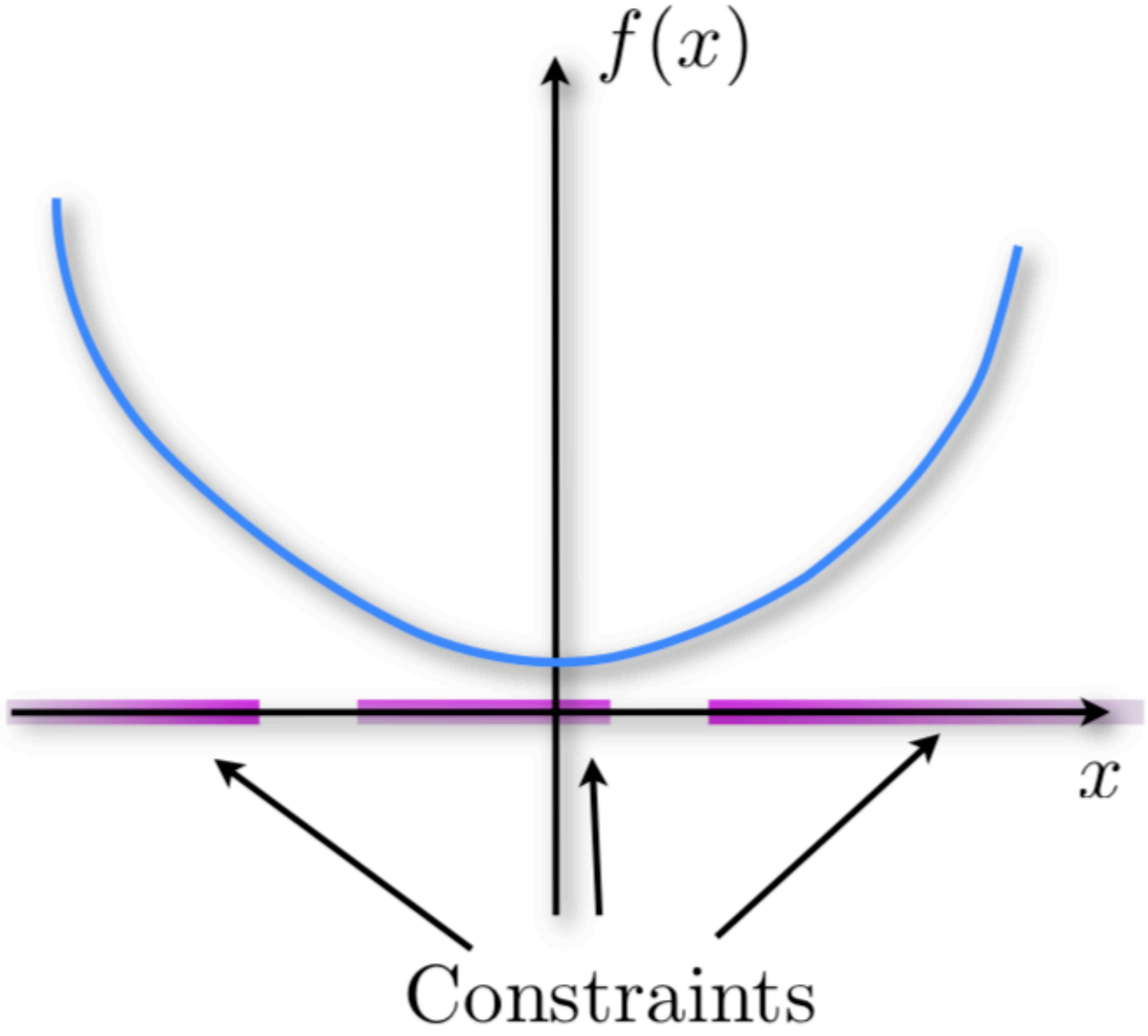
(Any thoughts about what this implies?)

- Using PL inequality + Lipschitz gradient continuity:

$$f(x_t) - f(x^\star) \leq \left(1 - \frac{\xi}{L}\right)^t (f(x_0) - f(x^\star))$$ Whiteboard

- Does not use **convexity**: holds for invex functions (stationary = global)

Convex optimization is not only about the objective



- Back to the first slide:

$$\min_{x} \quad f(x)$$

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

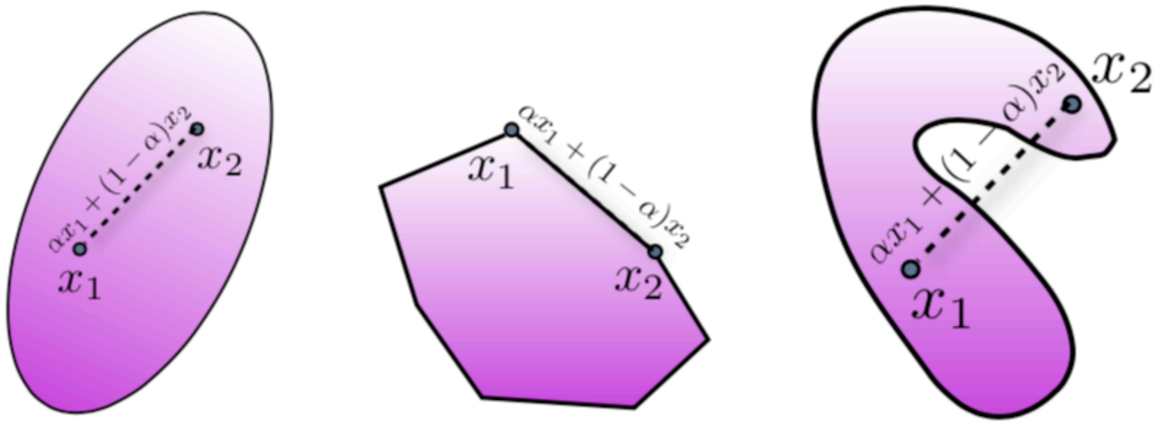
(We will worry about this in the lectures to follow!)

Convex sets

$\mathcal{C} \subseteq \mathbb{R}^p$ is convex if $\forall x_1, x_2 \in \mathcal{C}$, it holds $\forall \alpha \in [0,1] : \alpha x_1 + (1-\alpha) x_2 \in \mathcal{C}$

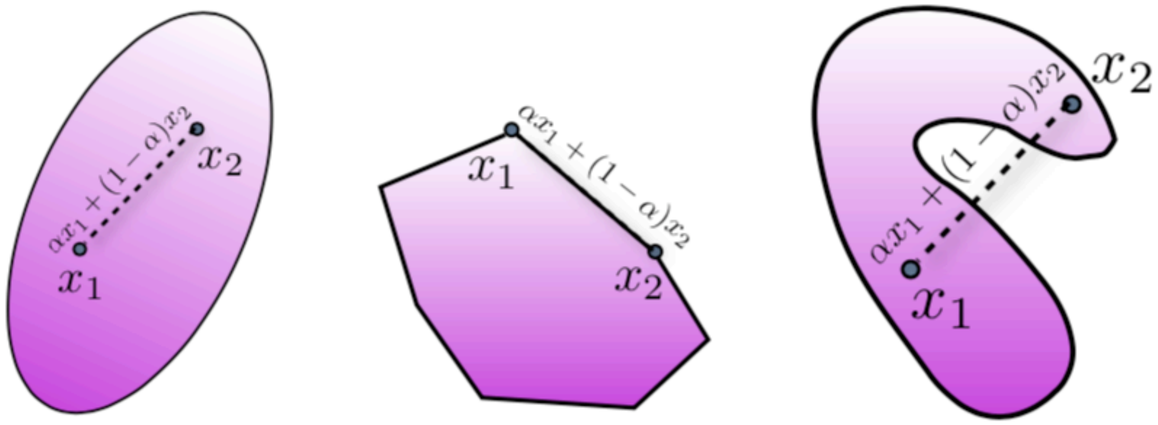
Convex sets

$\mathcal{C} \subseteq \mathbb{R}^p$ is convex if $\forall x_1, x_2 \in \mathcal{C}$, it holds $\forall \alpha \in [0,1] : \alpha x_1 + (1-\alpha) x_2 \in \mathcal{C}$



Convex sets

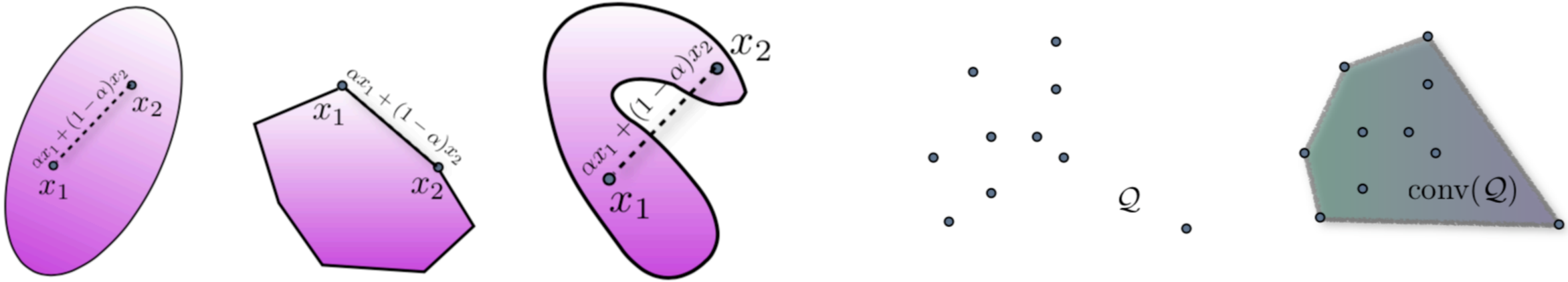
$\mathcal{C} \subseteq \mathbb{R}^p$ is convex if $\forall x_1, x_2 \in \mathcal{C}$, it holds $\forall \alpha \in [0,1] : \alpha x_1 + (1-\alpha) x_2 \in \mathcal{C}$



– Convex hull of points: $\text{conv}(\mathcal{V}) = \left\{ \sum_{i=1}^{|\mathcal{V}|} \alpha_i x_i \;:\; \sum_{i=1}^{|\mathcal{V}|} \alpha_i = 1,\ \alpha_i \geq 0, x_i \in \mathcal{V} \right\}$

Convex sets

$\mathcal{C} \subseteq \mathbb{R}^p$ is convex if $\forall x_1, x_2 \in \mathcal{C}$, it holds $\forall \alpha \in [0,1] : \alpha x_1 + (1-\alpha) x_2 \in \mathcal{C}$



– Convex hull of points: $\text{conv}(\mathcal{V}) = \left\{ \sum_{i=1}^{|\mathcal{V}|} \alpha_i x_i \ : \ \sum_{i=1}^{|\mathcal{V}|} \alpha_i = 1, \ \alpha_i \geq 0, x_i \in \mathcal{V} \right\}$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

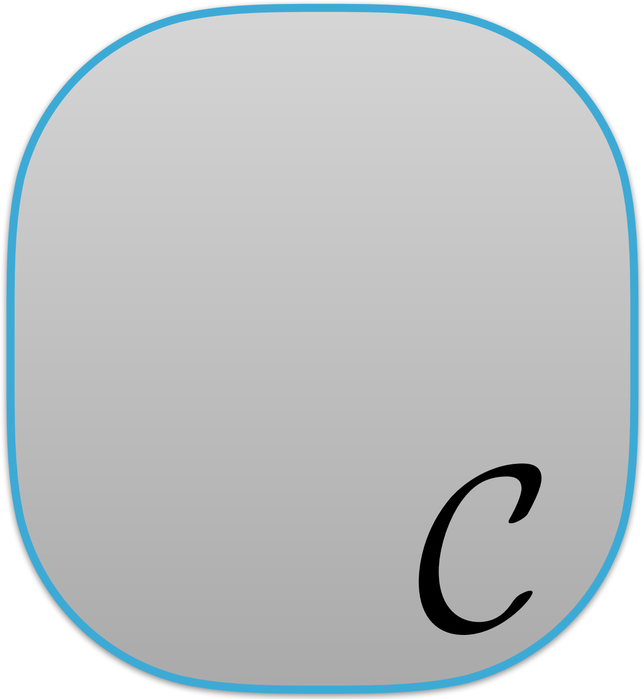
$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

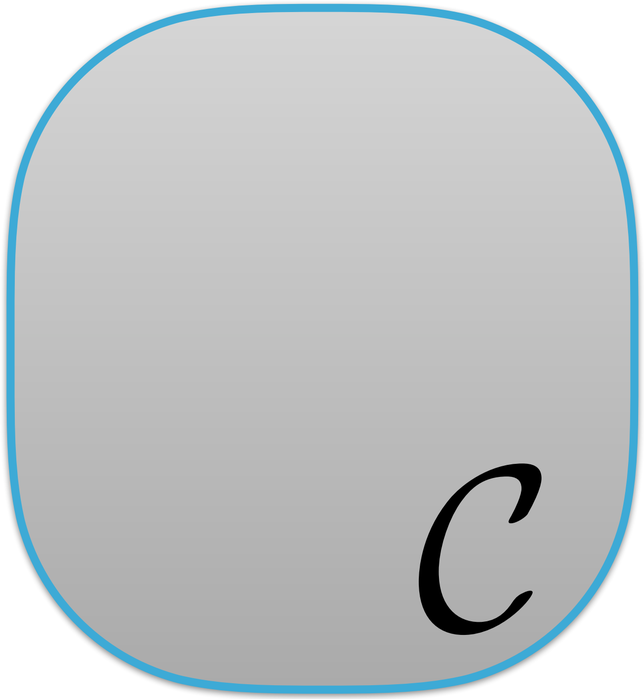
$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

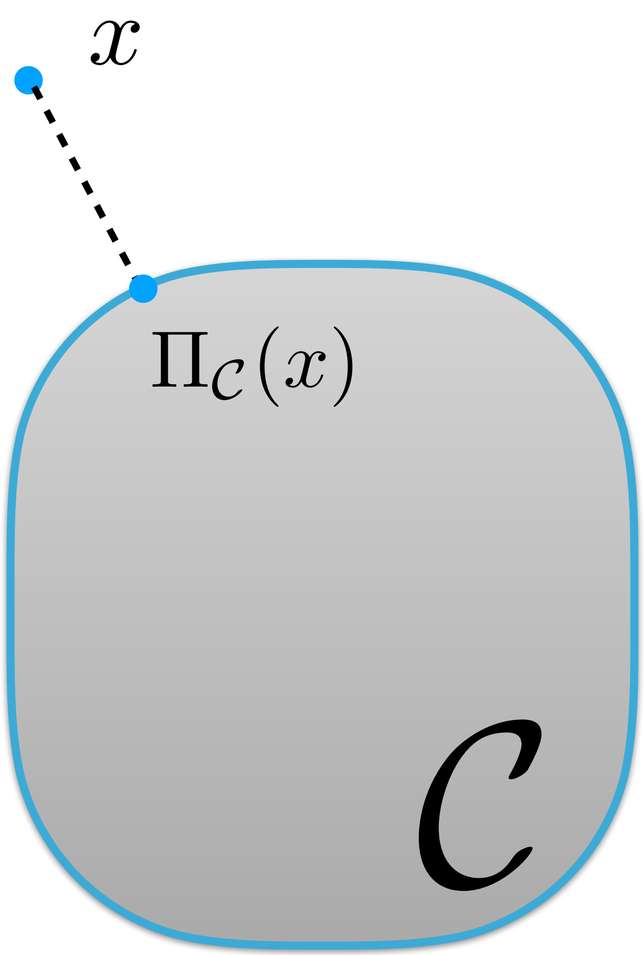
$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

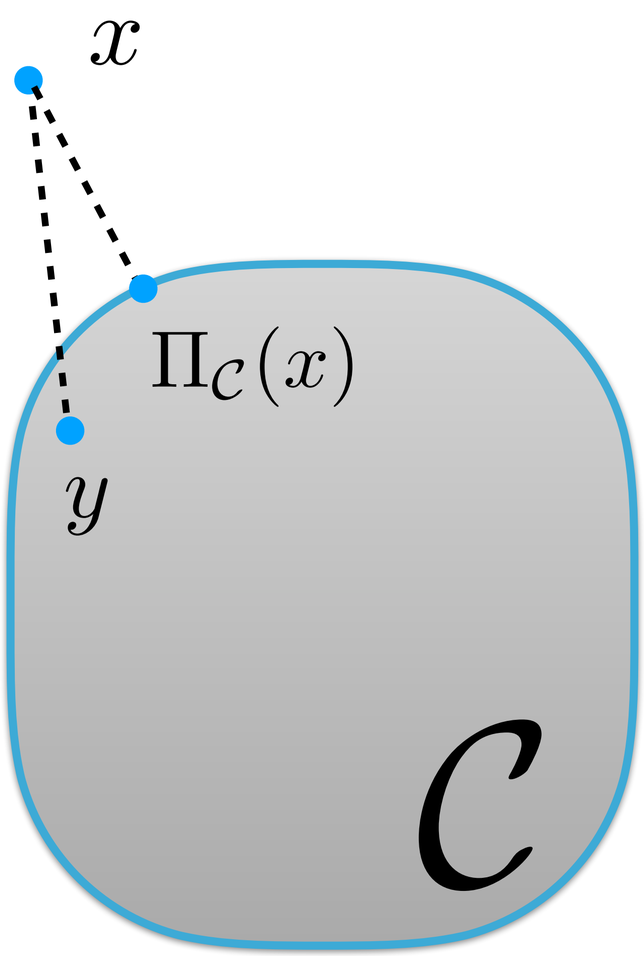
$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

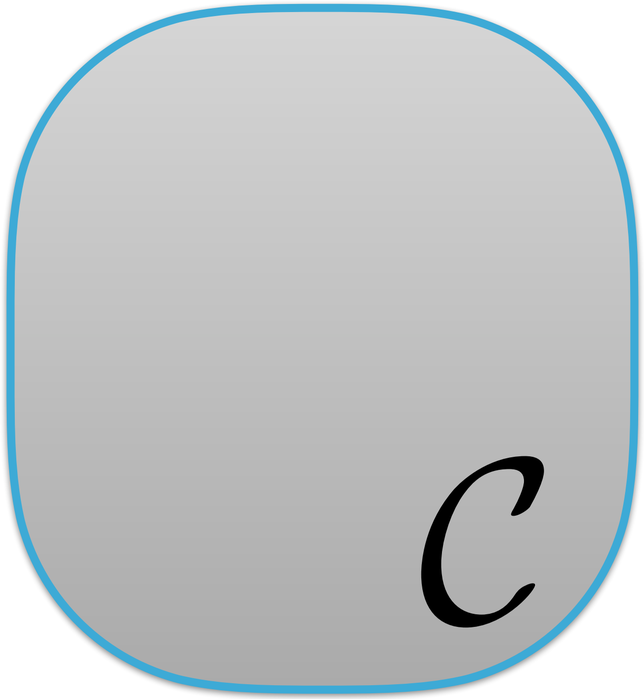
(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

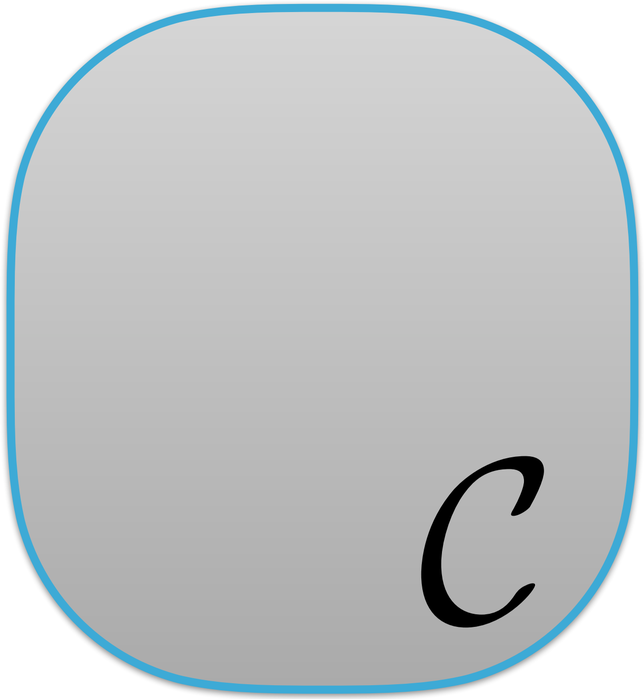
(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

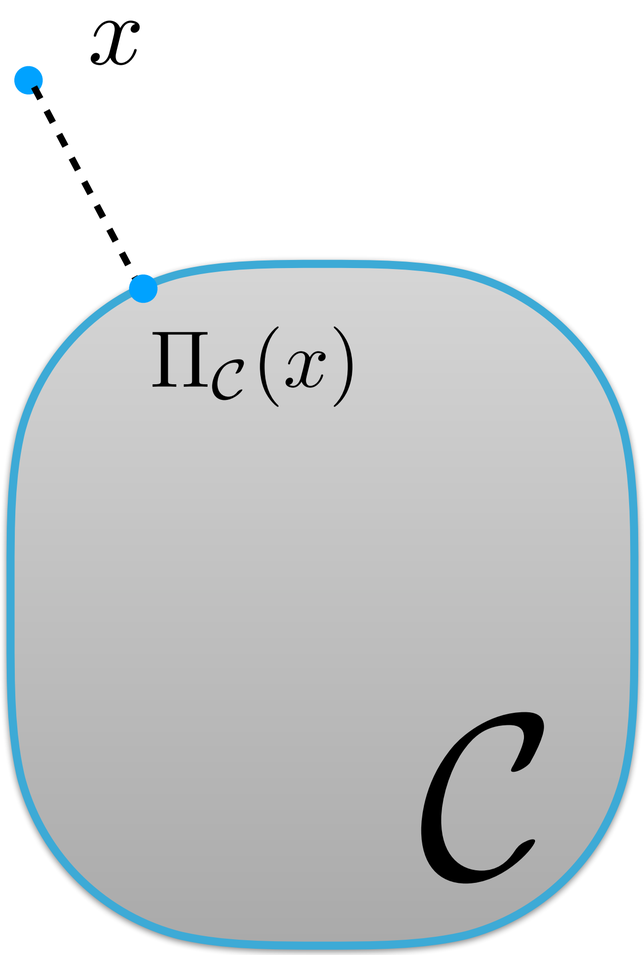
(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

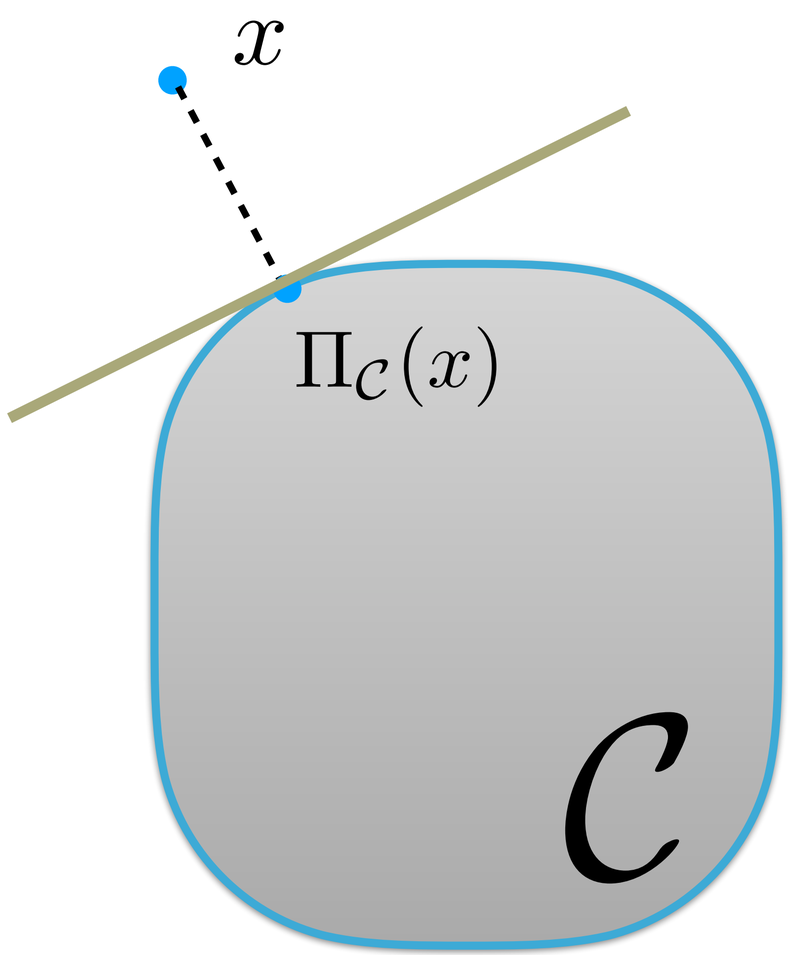


Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

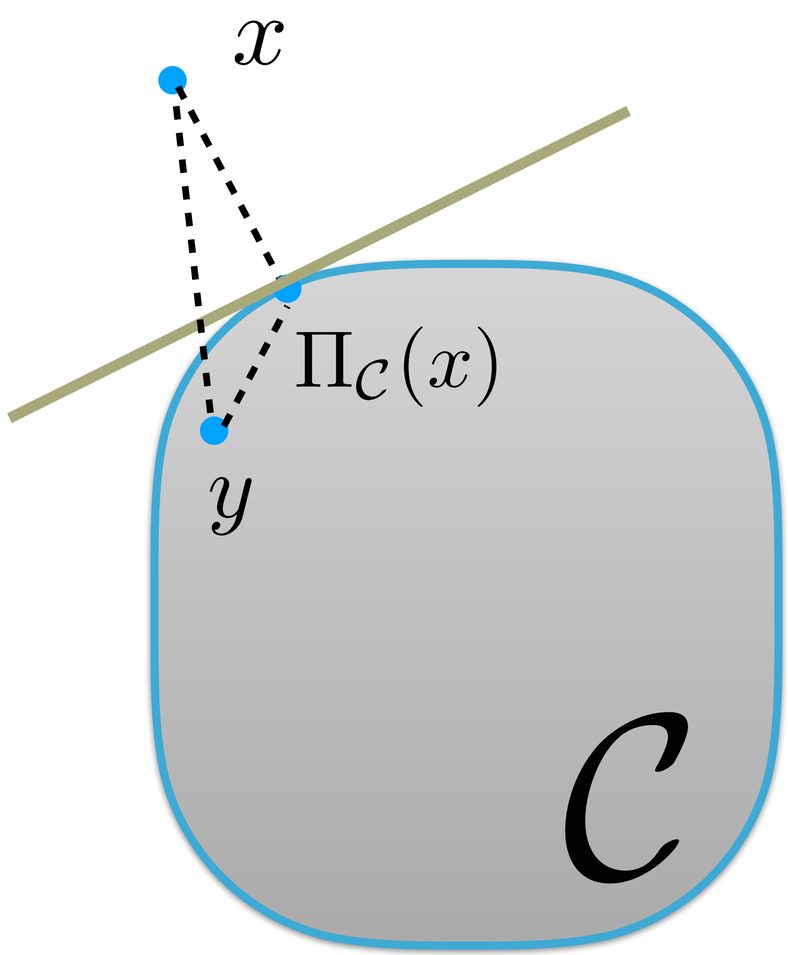
$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

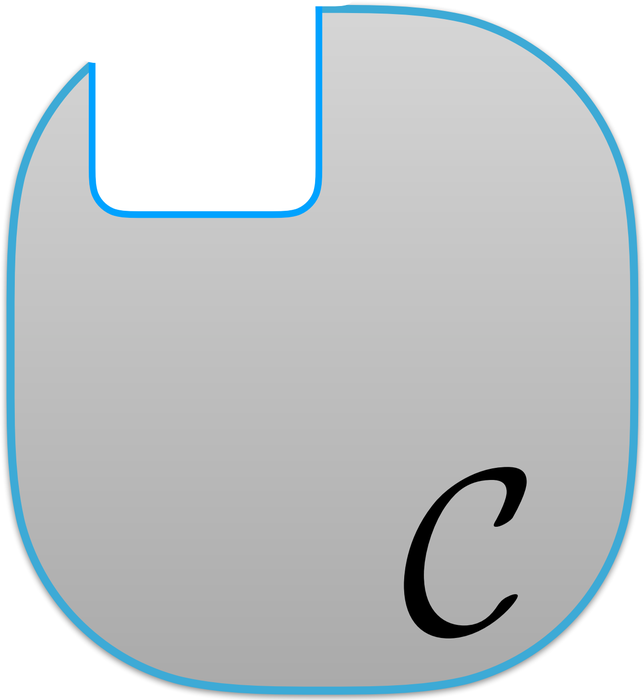
(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$



Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

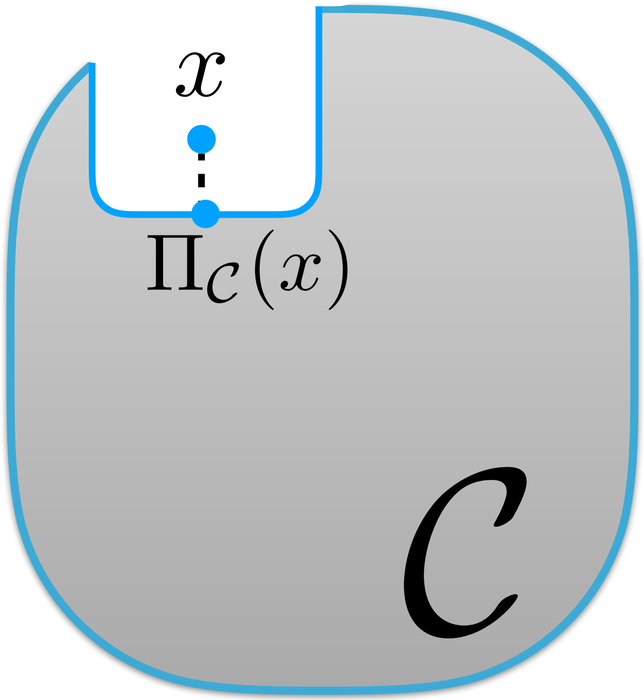
(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$

- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

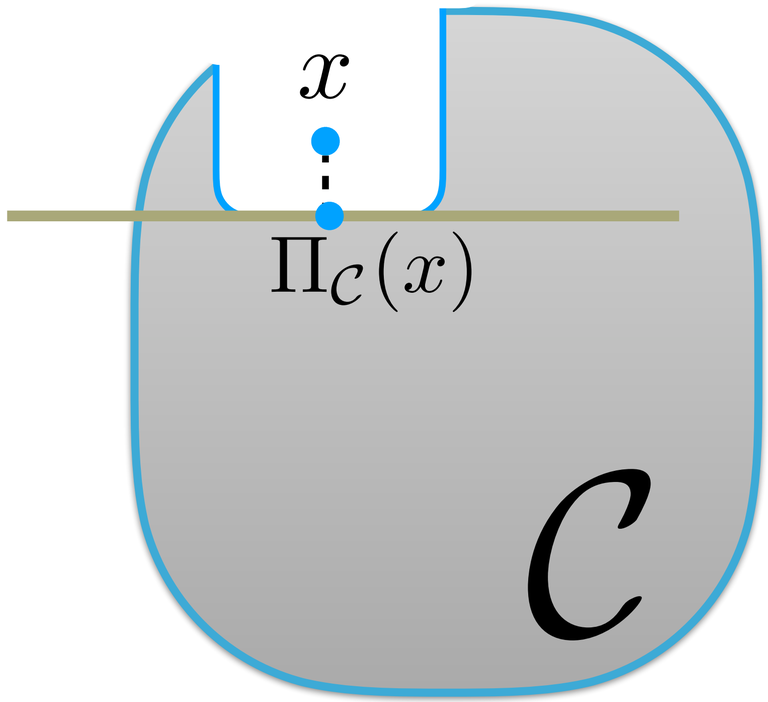


Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

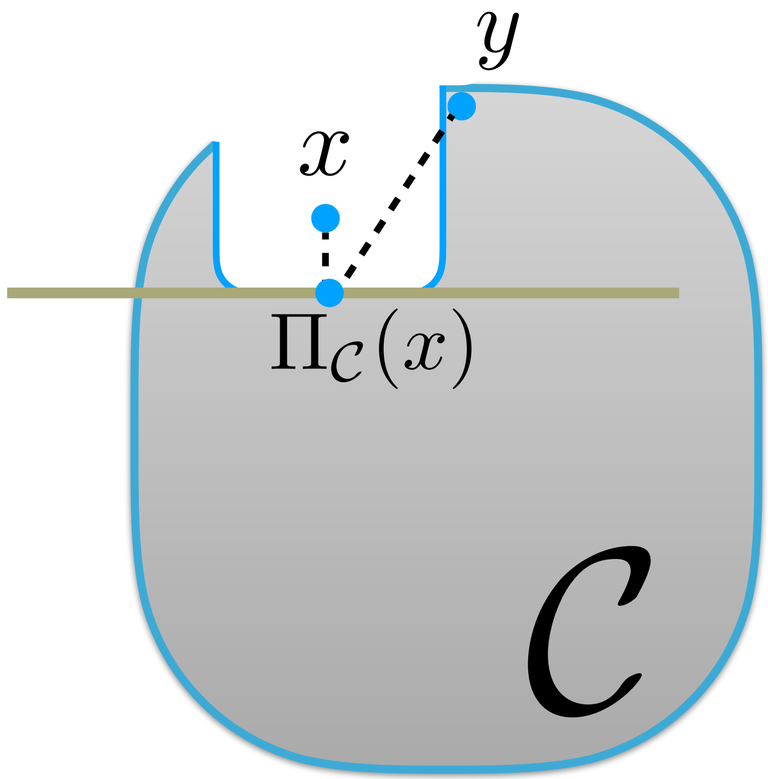
$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

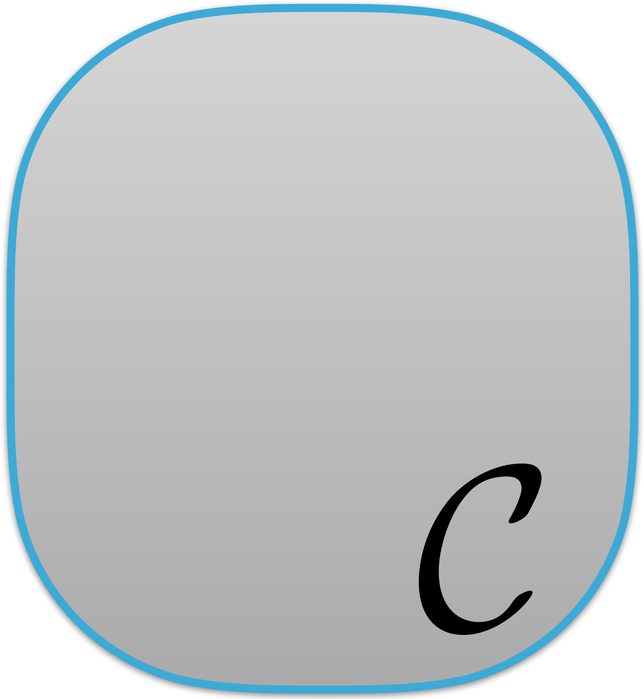
$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

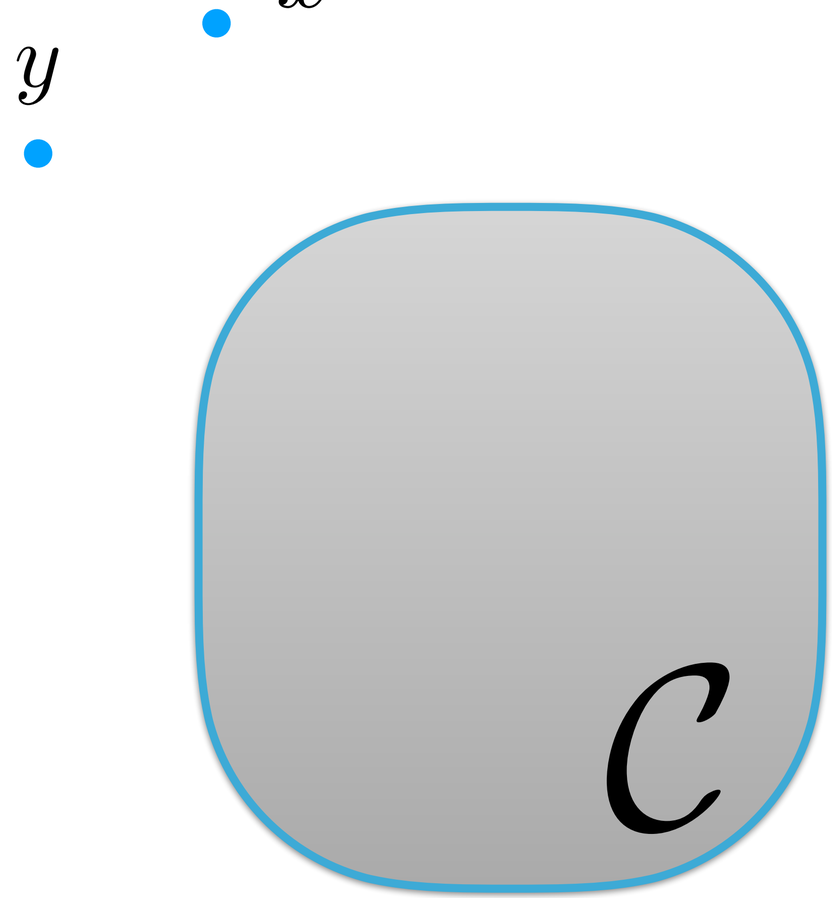
$$\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|_2 \leq \|x - y\|_2, \ \forall x, y$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

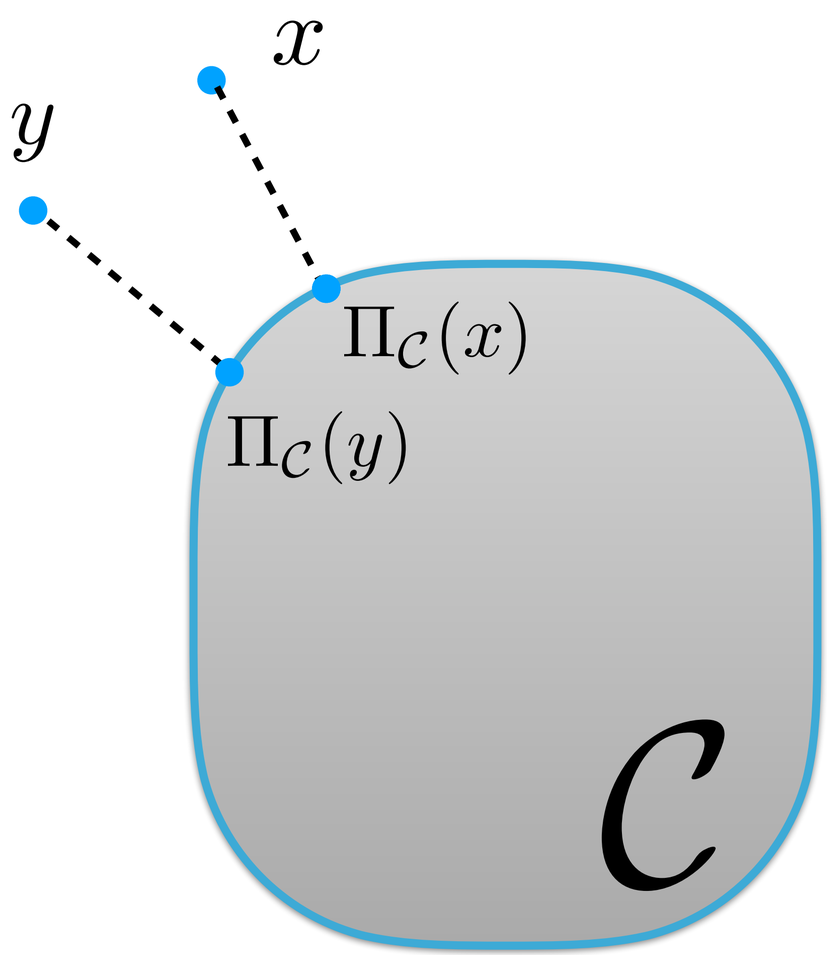
$$\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|_2 \leq \|x - y\|_2, \ \forall x, y$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

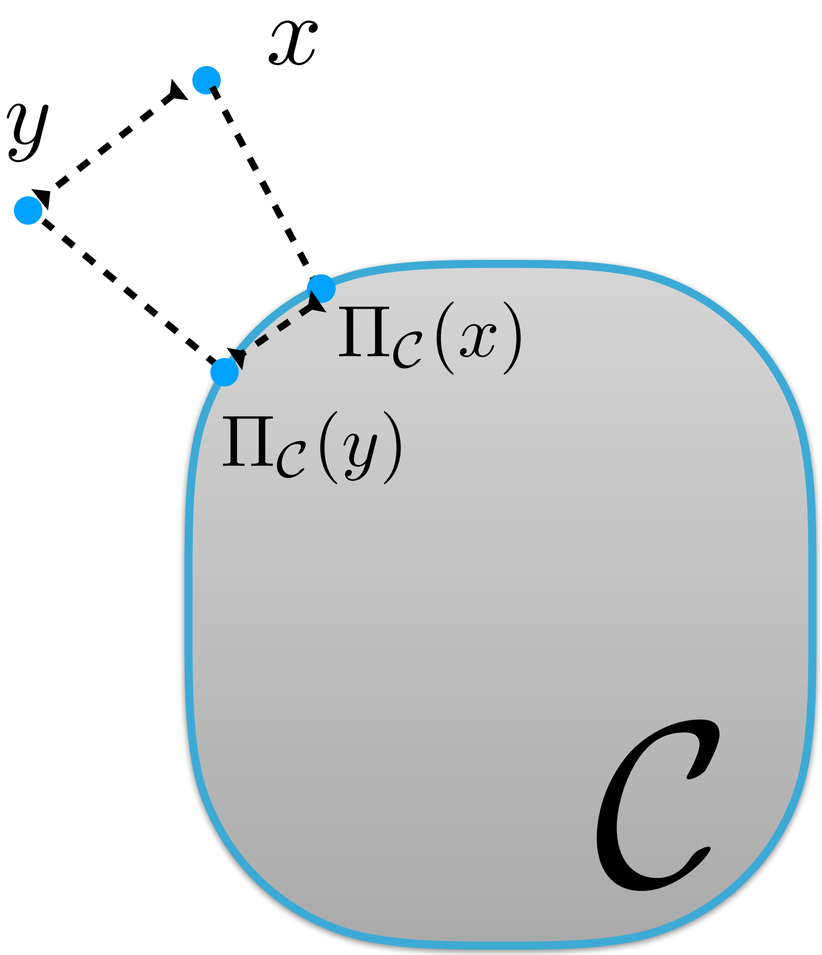
$$\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|_2 \leq \|x - y\|_2, \ \forall x, y$$

Projections onto convex sets

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_2^2$$

(The use of Euclidean norm is arbitrary and often depends on the application)

$$\Pi_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|_1$$



- Key properties of convex sets

$$\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \forall y \in \mathcal{C}, \forall x$$

$$\langle \Pi_{\mathcal{C}}(x) - y, \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \forall y \in \mathcal{C}, \forall x$$

$$\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|_2 \leq \|x - y\|_2, \ \forall x, y$$

Projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

Projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- A two-step procedure:
  1. $\widetilde{x} = x_t - \eta \nabla f(x_t)$
  2. $x_{t+1} = \Pi_{\mathcal{C}} \left(\widetilde{x} \right)$

Projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- A two-step procedure:
  1. $\widetilde{x} = x_t - \eta \nabla f(x_t)$
  2. $x_{t+1} = \Pi_{\mathcal{C}} \left(\widetilde{x} \right)$

Demo

Projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- What about its convergence guarantees? Do we lose much by projecting?

Projected gradient descent

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- What about its convergence guarantees? Do we lose much by projecting?

Whiteboard

But wait; didn't we consider proj. GD before?

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

But wait; didn't we consider proj. GD before?

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- Yes, in the case of exact sparse linear regression:

$$\mathcal{C} = \{ x \in \mathbb{R}^p \ : \ \|x\|_0 \leq k \}$$

Is this set convex? If no or yes, why?

But wait; didn't we consider proj. GD before?

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- Yes, in the case of exact sparse linear regression:

$$\mathcal{C} = \{ x \in \mathbb{R}^p \ : \ \|x\|_0 \leq k \}$$

Is this set convex? If no or yes, why?

- But we observed that, despite non-convexity, it works just fine..

(Thus, a different analysis is needed, depending on the problem at hand)

But, constrained optimization can be hard..

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- There are constrained problems where we need exponentially many bits even to describe the solution..

But, constrained optimization can be hard..

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- There are constrained problems where we need exponentially many bits even to describe the solution..

- We considered cases where the constraints are simple:
Operations Research is an area where multiple, difficult constraints appear

But, constrained optimization can be hard..

$$x_{t+1} = \Pi_{\mathcal{C}} \left(x_t - \eta \nabla f(x_t) \right)$$

- There are constrained problems where we need exponentially many bits even to describe the solution..

- We considered cases where the constraints are simple:
Operations Research is an area where multiple, difficult constraints appear

- Prof. Richard Tapia is teaching a course on constrained convex opt.

Convex optimization: Is it a technology?

- Yes, we know about it more than most other areas of optimization

Convex optimization: Is it a technology?

- Yes, we know about it more than most other areas of optimization
- Yes, there are off-the-shelf solvers available online

CVXOPT – https://cvxopt.org
CVXPy – http://www.cvxpy.org/
CVX – http://cvxr.com/cvx/
JuliaOpt – https://www.juliaopt.org/

Many optimizers in NN training can be applied to convex problems

TensorFlow – https://www.tensorflow.org/
PyTorch – https://pytorch.org/

Convex optimization: Is it a technology?

- Yes, we know about it more than most other areas of optimization
- Yes, there are off-the-shelf solvers available online

CVXOPT - https://cvxopt.org
CVXPy - http://www.cvxpy.org/
CVX - http://cvxr.com/cvx/
JuliaOpt - https://www.juliaopt.org/

Many optimizers in NN training can be applied to convex problems
TensorFlow - https://www.tensorflow.org/
PyTorch - https://pytorch.org/

- Why should we still care about convex optimization?

Several practical problems are actually convex
Many practical problems can be approximated by convex ones
If one doesn't understand convex opt., why even try understanding non-convex opt.?

Papers to review – due next Tuesday

(Select one of the following papers)

- "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition", Karimi et al., 2016.

(Focus on Sections 1,2,4,5 – and the corresponding appendix)

(Rule: as you read, think of extensions – feel free to find me for more discussions)

Conclusion

- Gradient descent has nice properties (even for non-convex problems)
- Further global assumptions lead to better convergence guarantees
- Convex optimization has nice properties (local = global)

..but can be intractable!

Next lecture

- Provide a taxonomy of alternative approaches
- Focus on more computational side and how we can accelerate opt. methods