Introduction to Machine Learning
CentraleSupélec Paris — Fall 2017

2. Elements of convex optimization

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

Why talk about optimization?

- Supervised ML: **empirical risk minimization**

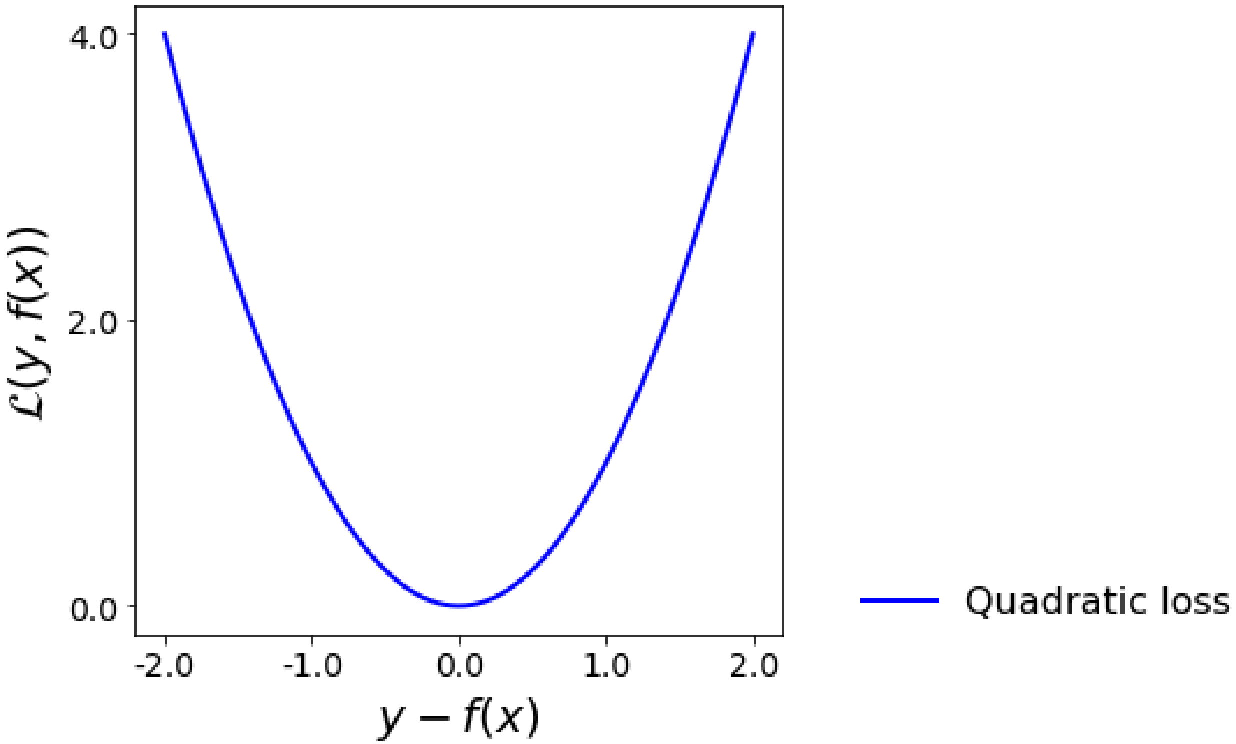
$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}(y^i, f(\boldsymbol{x}^i))$$

Why talk about optimization?

- Supervised ML: **empirical risk minimization**

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}(y^i, f(\boldsymbol{x}^i))$$

- **Quadratic loss** $\mathcal{L}(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$

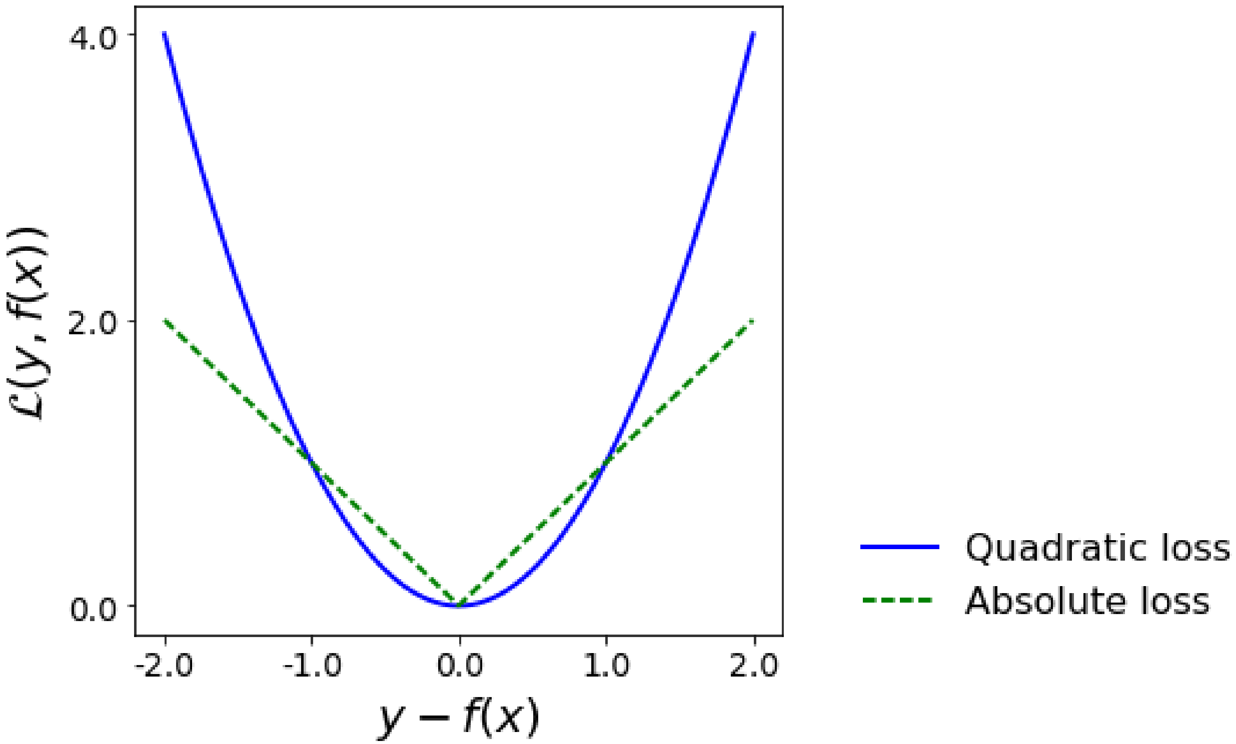


Why talk about optimization?

- Supervised ML: **empirical risk minimization**

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}(y^i, f(\boldsymbol{x}^i))$$

- **Absolute loss** $\mathcal{L}(y, f(\boldsymbol{x})) = |y - f(\boldsymbol{x})|$



Why talk about optimization?

- Supervised ML: **empirical risk minimization**

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}(y^i, f(\boldsymbol{x}^i))$$

- **0/1 loss**

$$\mathcal{L}(y, f(\boldsymbol{x})) = \begin{cases} 0 \text{ if } \; f(\boldsymbol{x}) = y \\ 1 \text{ otherwise} \end{cases}$$

Why talk about optimization?

- **Unsupervised machine learning** also involves minimizing functions.

  Examples:

  – **Dimensionality reduction:** find a set of m features, m<p, such that the data projected on these m features retains maximal information.

  – **Clustering:** find K groups of samples such that the between-groups variance is high and the within-group variance is small.

Learning objectives

- **Recognize** a convex optimization problem.
- Solve an **unconstrained convex optimization** problem
  - Exactly when possible
  - By **gradient descent** and a number of its variants.
- Solve a **quadratic program**
  - Formulate the **dual problem**
  - Write down **Karush-Kuhn-Tucker conditions.**
- Transform inequality constraints with **slack variables.**

Convex functions

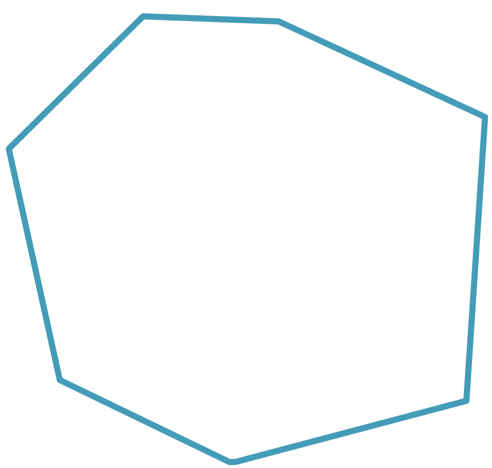
Convex set

$\mathcal{S} \subseteq \mathbb{R}^n$ is a **convex set** iff:

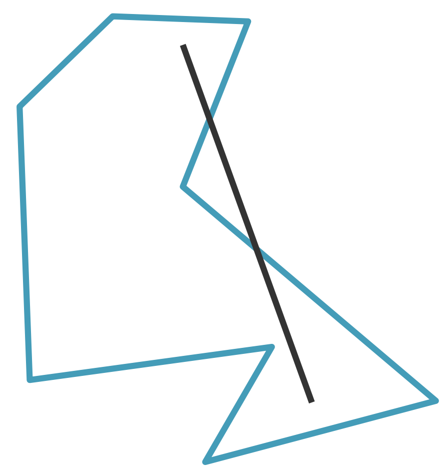
$$\boxed{t\boldsymbol{u} + (1-t)\boldsymbol{v} \in \mathcal{S}}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}$ and $0 \leq t \leq 1$

Line segments between 2 points of $\mathcal{S}$ lie entirely in $\mathcal{S}$.



Convex set of $\mathbb{R}^2$



Non-convex set of $\mathbb{R}^2$

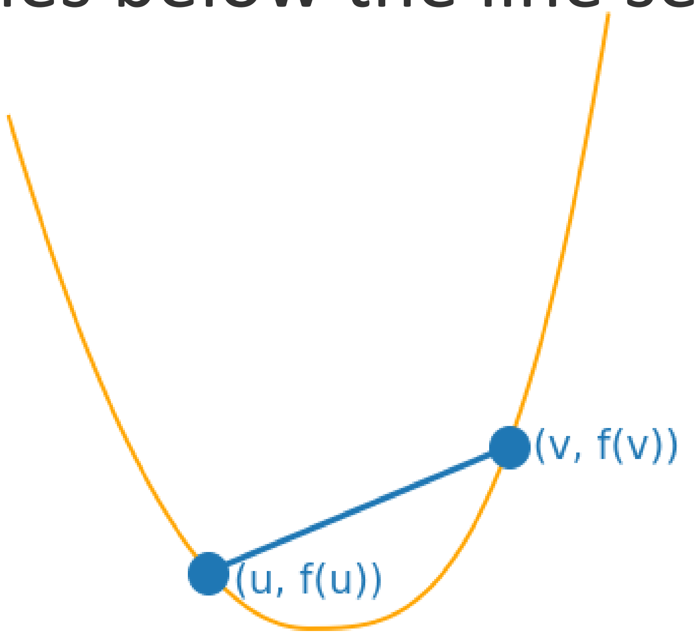
Convex function

$f : \mathbb{R}^n \to \mathbb{R}$ is **convex** iff:

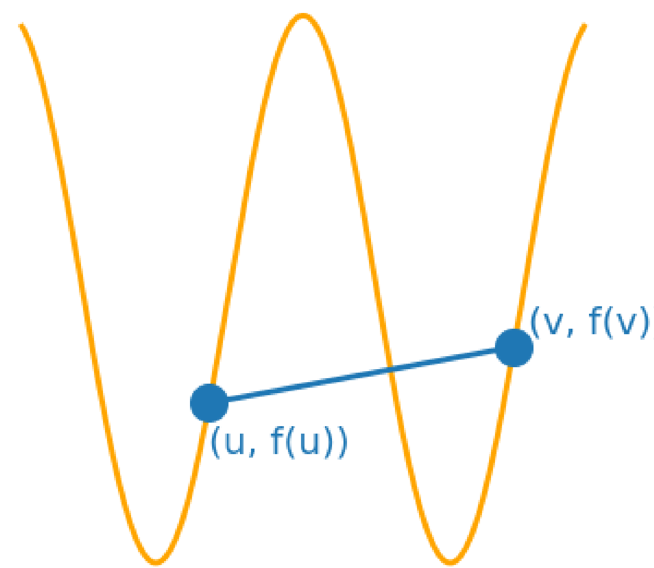
- its domain is a **convex set**
- $\boxed{f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq t f(\boldsymbol{u}) + (1-t) f(\boldsymbol{v})}$

  for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$ and $0 \leq t \leq 1$

f lies below the line segment joining f(**u**) and f(**v**).



Convex function of $\mathbb{R} \to \mathbb{R}$



Non-convex function of $\mathbb{R} \to \mathbb{R}$

Concave function

$f : \mathbb{R}^n \to \mathbb{R}$ is **concave** iff:

- its domain is a **convex set**
- $\boxed{f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \boxed{\geq} tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v})}$

  for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$ and $0 \leq t \leq 1$

f concave ⇔ -f convex

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $f : [0, 1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$
- $\boxed{f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3}$
- $f : [0,1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$
- $\boxed{f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3}$ **Yes!**
- $f : [0,1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $\boxed{f : [0, 1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3}$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3$
- $\boxed{f : [0,1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3}$ **No!**
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$

Are the following univariate functions convex?

- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 + 3$ Yes
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^2 - 3$ Yes
- $f : [0, 1] \cup [3, +\infty[\to \mathbb{R} \qquad u \mapsto 2u^2 - 3$ No
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$ No
- $f : \mathbb{R}_+ \to \mathbb{R} \qquad u \mapsto 2u^3 - 3$ Yes
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \max(u, 2)$ Yes
- $f : \mathbb{R} \to \mathbb{R} \qquad u \mapsto \sin(u)$ No

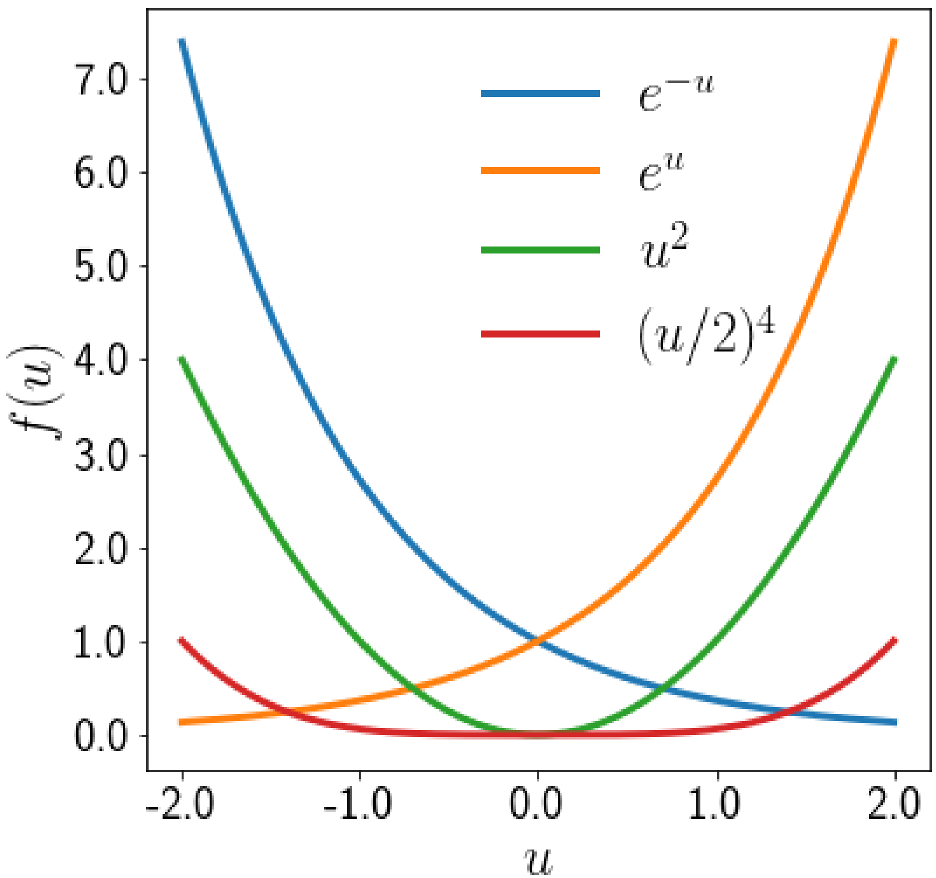
Univariate examples

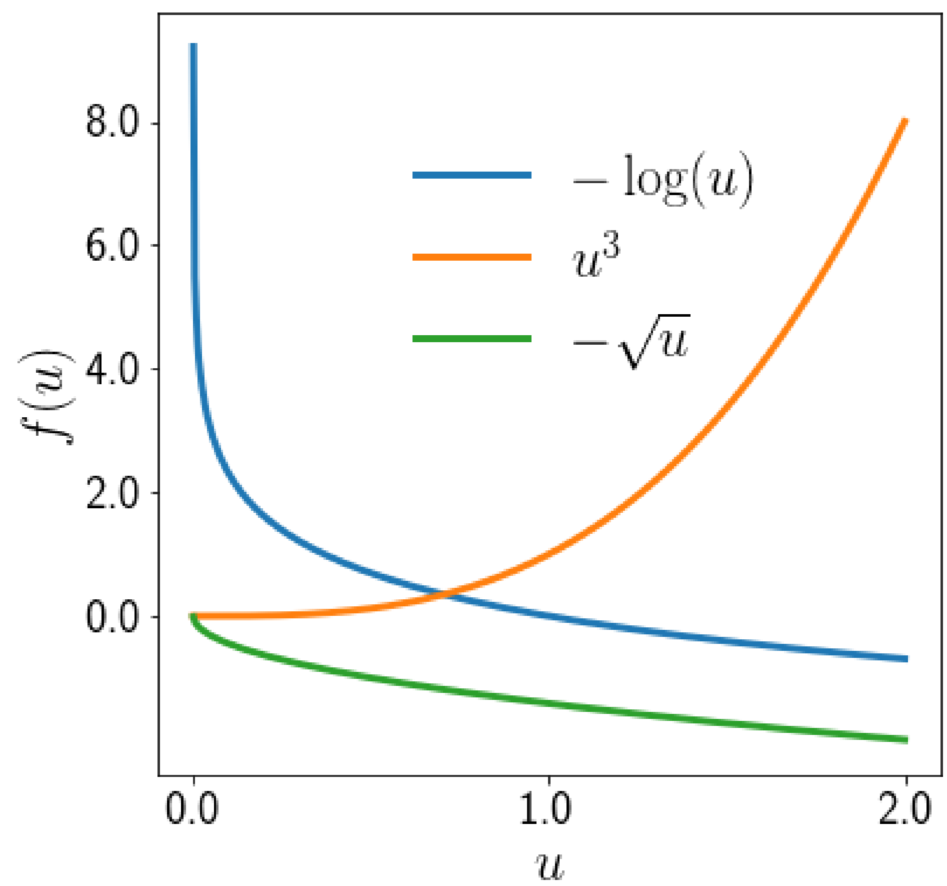
- **Exponential:** $u \mapsto e^{au} \ \forall a \in \mathbb{R}$
- **Logarithmic:** $u \mapsto -\log(au) \ \forall a > 0 \quad \text{on } \mathbb{R}_+^*$
- **Power functions:**

$u \mapsto u^a \ \forall a \geq 1 \text{ or } a \leq 0 \text{ on } \mathbb{R}_+^*$

$u \mapsto -u^a \ \forall 0 \leq a \leq 1 \text{ on } \mathbb{R}_+^*$

$u \mapsto u^a \ \forall a = 2n, n \in \mathbb{N}$





More examples

- **Affine functions** are both convex and concave

$$\boldsymbol{u} \mapsto \boldsymbol{a}^\top \boldsymbol{u} + b \;\; \boldsymbol{a} \in \mathbb{R}^n$$

More examples

- **Affine functions** are both convex and concave

$$\boldsymbol{u} \mapsto \boldsymbol{a}^\top \boldsymbol{u} + b \quad \boldsymbol{a} \in \mathbb{R}^n$$

- **Quadratic functions**

$$\boldsymbol{u} \mapsto \frac{1}{2}\boldsymbol{u}^\top Q \boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u} + c \quad \boxed{Q \succeq 0} \quad \boldsymbol{b} \in \mathbb{R}^n \quad c \in \mathbb{R}$$

?

Q **positive semi-definite**

More examples

- **Affine functions** are both convex and concave

$$\boldsymbol{u} \mapsto \boldsymbol{a}^\top \boldsymbol{u} + b \quad \boldsymbol{a} \in \mathbb{R}^n$$

- **Quadratic functions**

$$\boldsymbol{u} \mapsto \frac{1}{2}\boldsymbol{u}^\top Q\boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u} + c \quad \boxed{Q \succeq 0} \quad \boldsymbol{b} \in \mathbb{R}^n \quad c \in \mathbb{R}$$

Q **positive semi-definite**

$$\boldsymbol{u}^\top Q\boldsymbol{u} \geq 0 \qquad \forall\, \boldsymbol{u} \in \mathbb{R}^n$$

- All eigenvalues of Q are non-negative
- The bilinear form $\boldsymbol{u}, \boldsymbol{v} \mapsto \boldsymbol{v}^\top Q\boldsymbol{u}$ is an inner product
- Q is a Gram matrix of independent vectors $Q_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$
- Unique Cholesky decomposition $Q = LL^\top$

More examples

- **Affine functions** are both convex and concave

$$\boldsymbol{u} \mapsto \boldsymbol{a}^\top \boldsymbol{u} + b \quad \boldsymbol{a} \in \mathbb{R}^n$$

- **Quadratic functions**

$$\boldsymbol{u} \mapsto \frac{1}{2}\boldsymbol{u}^\top Q \boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u} + c \quad \boxed{Q \succeq 0} \quad \boldsymbol{b} \in \mathbb{R}^n \quad c \in \mathbb{R}$$

Q **positive semi-definite**

- **Lp norms**

$$\boldsymbol{u} \mapsto ||\boldsymbol{u}||_p = \left(\sum_{j=1}^{n} |u_j^p| \right)^{1/p}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** iff:

$$\forall\ \boldsymbol{u} \neq \boldsymbol{v} \in \mathrm{dom}(f),\ \ \forall\ 0 < t < 1$$

$$\boxed{f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) < tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v}).}$$

f is convex and has **greater curvature than a linear function.**

- $f : \mathbb{R}^n \to \mathbb{R}$ is **strongly convex of parameter m>0** iff:

$$f - \frac{m}{2}||\boldsymbol{u}||_2^2 \quad \text{is convex.}$$

f is convex and has curvature **as least as great as a quadratic function.**

strongly convex ⇒ strictly convex ⇒ convex

First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})$$

First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f}(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})$$

Gradient of f

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial u_1} \\ \frac{\partial f}{\partial u_2} \\ \cdots \\ \frac{\partial f}{\partial u_n} \end{pmatrix}$$

First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})}$$

?

First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})}$$

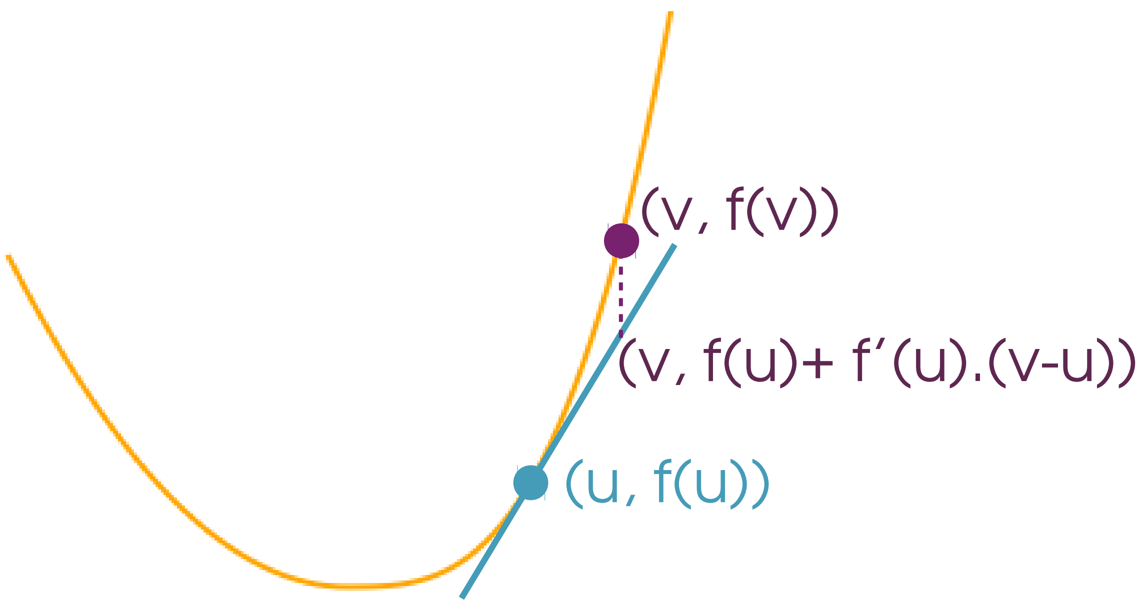
First-order Taylor expansion of f in u

First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})}$$

First-order Taylor expansion of f in u



First-order characterization

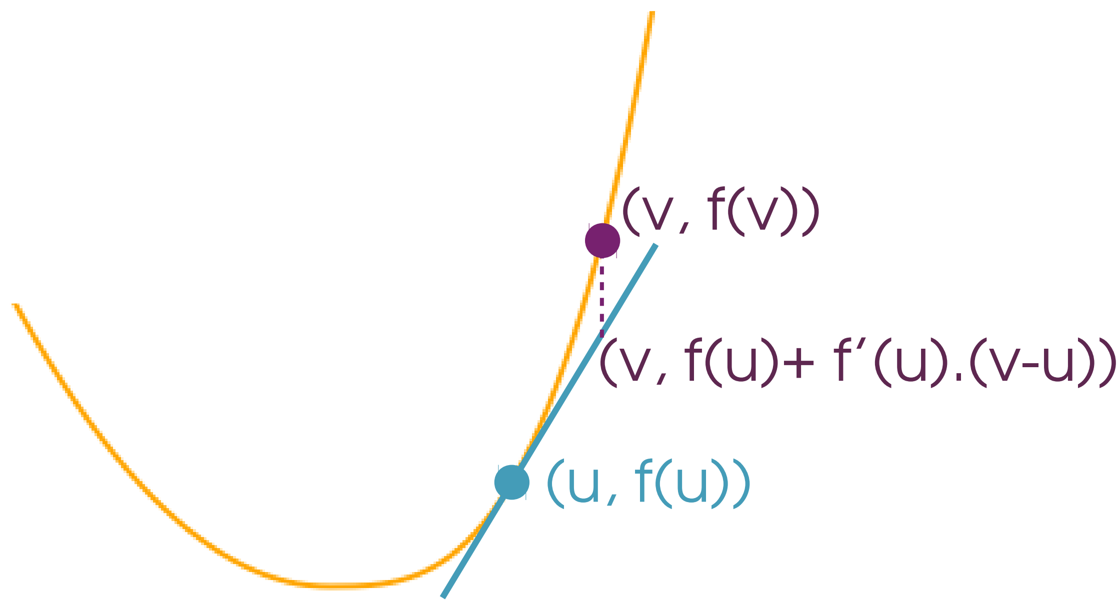
- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})}$$

First-order Taylor expansion of f in u

What does it mean if

$\nabla f(\boldsymbol{u}) = 0$ **?**



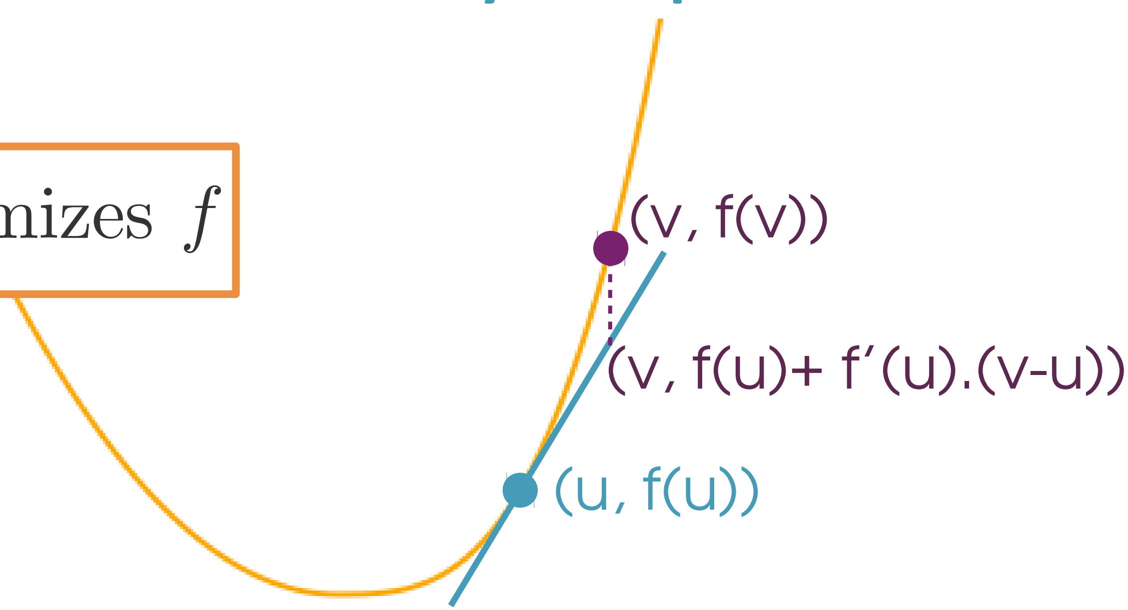
First-order characterization

- If f is **differentiable**, then f is **convex** if and only if:
  - its domain is a convex set
  - for all $\boldsymbol{u}, \boldsymbol{v} \in \mathrm{dom}(f)$

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \boxed{\nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u})}$$

First-order Taylor expansion of f in u

$$\boxed{\nabla f(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} \text{ minimizes } f}$$



Second-order characterization

- If f is **twice differentiable**, then f is **convex** iff:
  - its domain is a convex set
  - for all $\boldsymbol{u} \in \mathrm{dom}(f)$

$$\nabla^2 f(\boldsymbol{u}) \succeq 0$$

Second-order characterization

- If f is **twice differentiable**, then f is **convex** iff:
  - its domain is a convex set
  - for all $\boldsymbol{u} \in \mathrm{dom}(f)$

$$\boxed{\nabla^2 f}(\boldsymbol{u}) \succeq 0$$

Hessian of f

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial u_1^2} & \frac{\partial^2 f}{\partial u_1 u_2} & \cdots & \frac{\partial^2 f}{\partial u_1 u_n} \\ \frac{\partial^2 f}{\partial u_2 u_1} & \frac{\partial^2 f}{\partial u_2^2} & \cdots & \frac{\partial^2 f}{\partial u_2 u_n} \\ \cdots & & & \\ \frac{\partial^2 f}{\partial u_n u_1} & \frac{\partial^2 f}{\partial u_n u_2} & \cdots & \frac{\partial^2 f}{\partial u_n^2} \end{pmatrix}$$

Second-order characterization

- If f is **twice differentiable**, then f is **convex** iff:
  - its domain is a convex set
  - for all $\boldsymbol{u} \in \mathrm{dom}(f)$

$$\boxed{\nabla^2 f}(\boldsymbol{u}) \succeq 0$$

Hessian of f

- f has **positive curvature** in any point u.

Operations preserving convexity

- **Non-negative linear combination**

  If $f_1, f_2, \ldots, f_m$ convex and $a_1, a_2, \ldots, a_m \geq 0$ then $a_1 f_1 + a_2 f_2 + \cdots + a_m f_m$ is convex.

- **Pointwise maximization**

  If $f_1, f_2, \ldots, f_m$ convex, then

  $\boldsymbol{u} \mapsto \max\limits_{1,\ldots,m} f_k(\boldsymbol{u})$ is convex (also true for an infinite number of functions f_k).

- **Partial minimization**

  If $f : \mathbb{R}^n \to \mathbb{R}$ convex and C is a convex set, then

$(u_1, u_2, \ldots, u_{n-1}) \mapsto \min\limits_{v \in C} f(u_1, u_2, \ldots, u_{n-1}, v)$ is convex.

Convex optimization

Unconstrained convex optimization

Unconstrained convex optimization program/problem:

$$\min_{\boldsymbol{u} \in \mathrm{dom}(f)} \quad f(\boldsymbol{u})$$

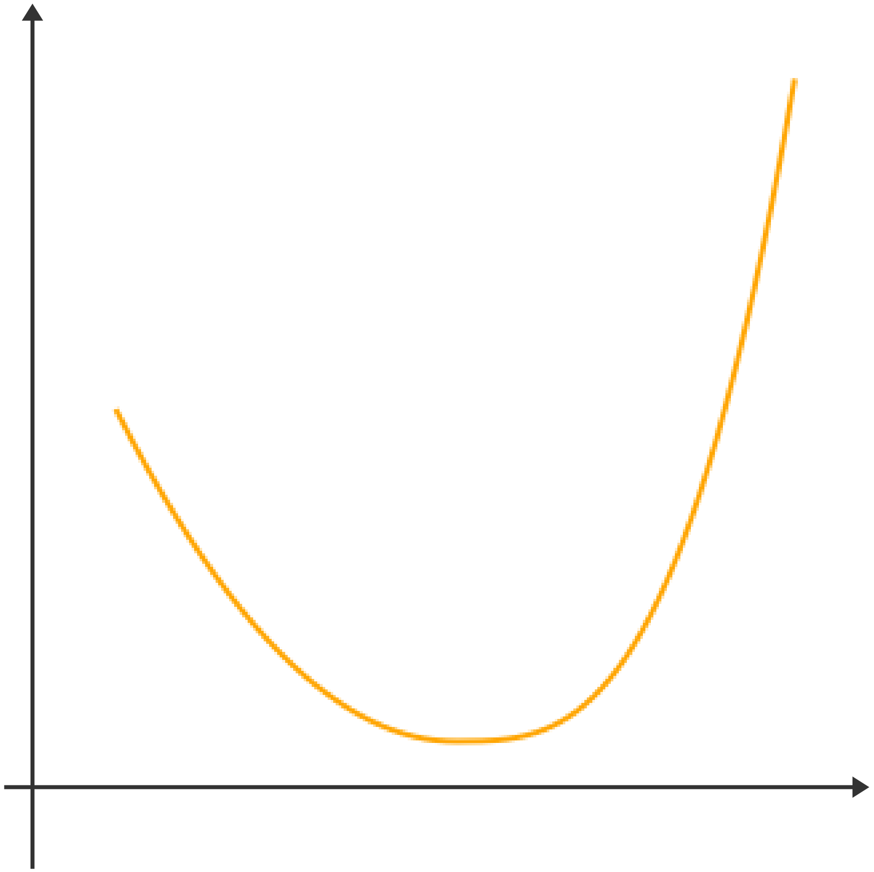
where f is convex.

Unconstrained convex optimization

Unconstrained convex optimization program/problem:

$$\min_{\boldsymbol{u} \in \mathrm{dom}(f)} \quad f(\boldsymbol{u})$$

where f is convex.

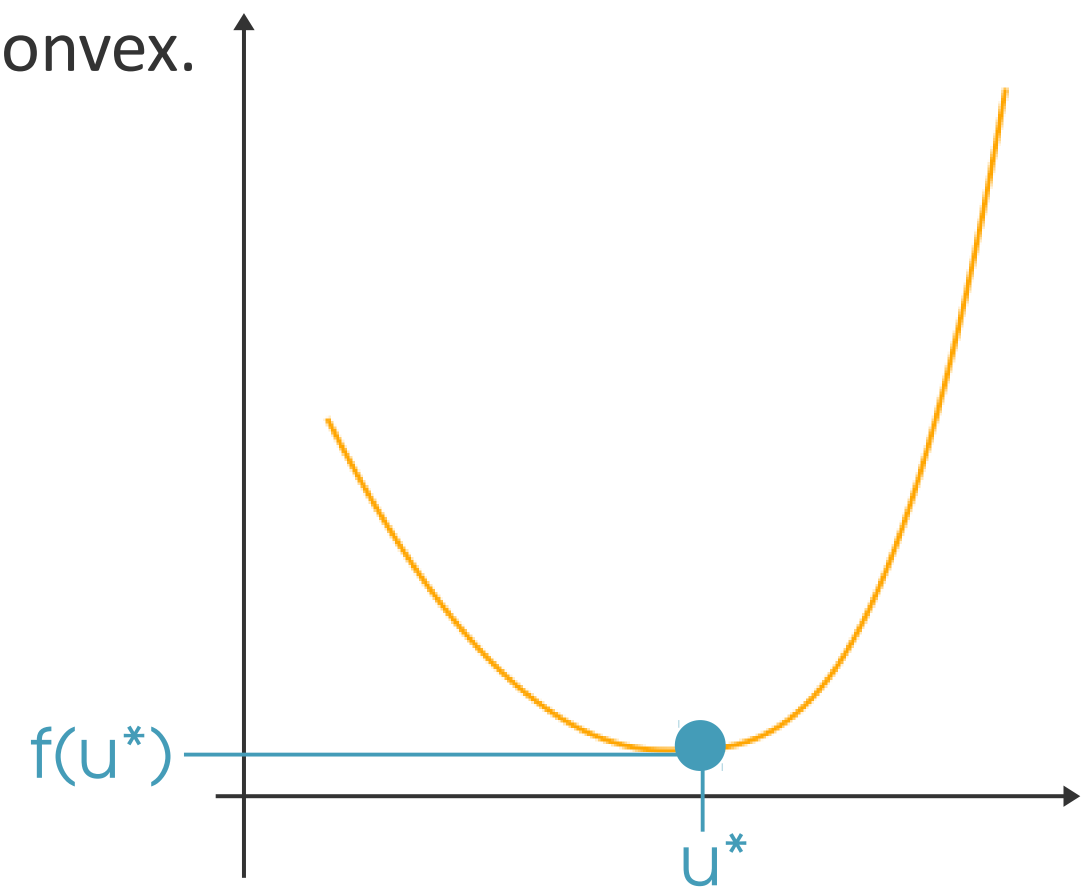


Unconstrained convex optimization

Unconstrained convex optimization program/problem:

$$\min_{\boldsymbol{u} \in \text{dom}(f)} \quad f(\boldsymbol{u})$$

where f is convex.



Constrained convex optimization

- **Convex optimization program/problem:**

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \\ \text{subject to } & g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ & h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$

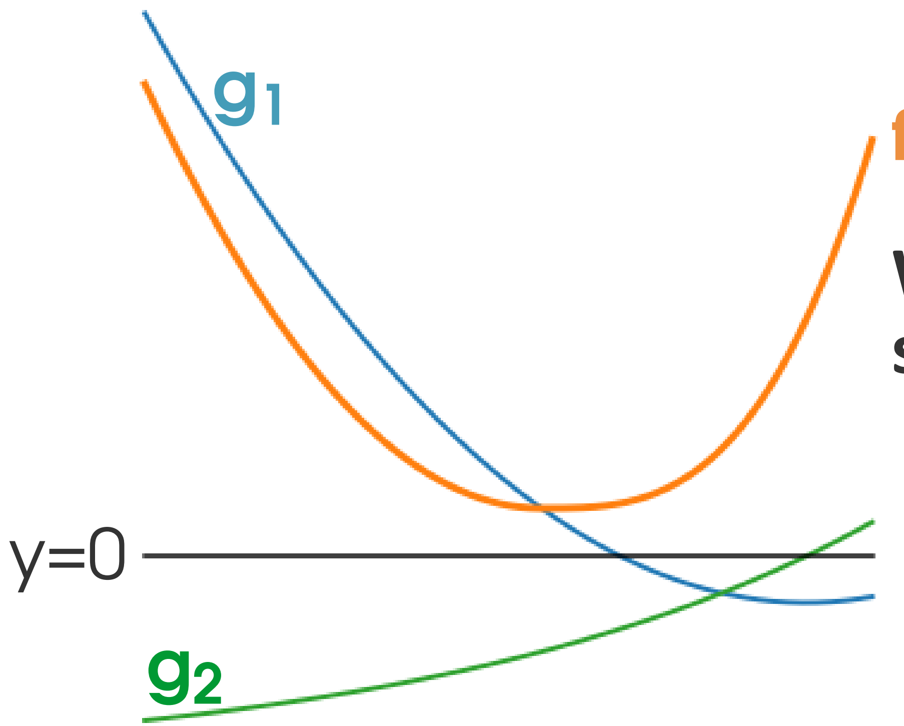
  - f is **convex**
  - $g_i, i = 1, \ldots, m$ are **convex**
  - $h_j, j = 1, \ldots, r$ are **affine** $h_j : \boldsymbol{u} \mapsto \boldsymbol{a}_j^\top \boldsymbol{u} + b_j$
  - D is the **common domain** of all the functions.

$$D = \operatorname{dom}(f) \cap \bigcap_{i=1}^{m} \operatorname{dom}(g_i) \bigcap_{j=1}^{r} \operatorname{dom}(h_j)$$

Constrained convex optimization

- **Convex optimization program/problem:**

$$
\begin{aligned}
&\min_{\boldsymbol{u}\in D} f(\boldsymbol{u}) \\
\text{subject to } & g_i(\boldsymbol{u}) \le 0, i = 1, \ldots, m \\
& h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r
\end{aligned}
$$

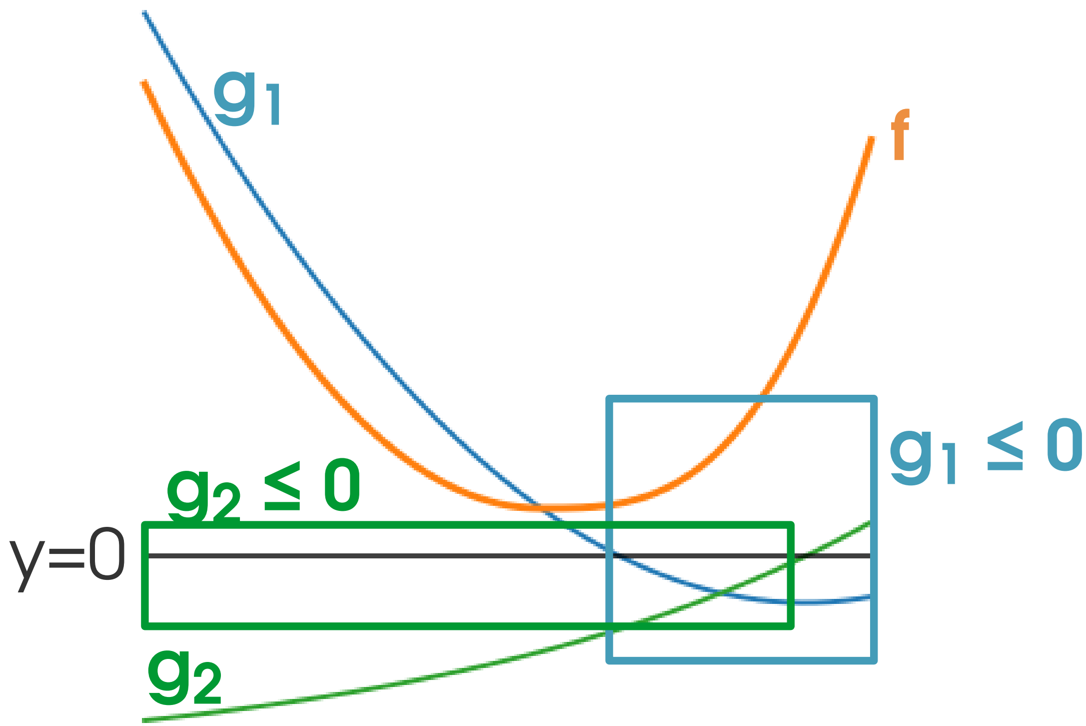


Where is the solution?

Constrained convex optimization

- **Convex optimization program/problem:**

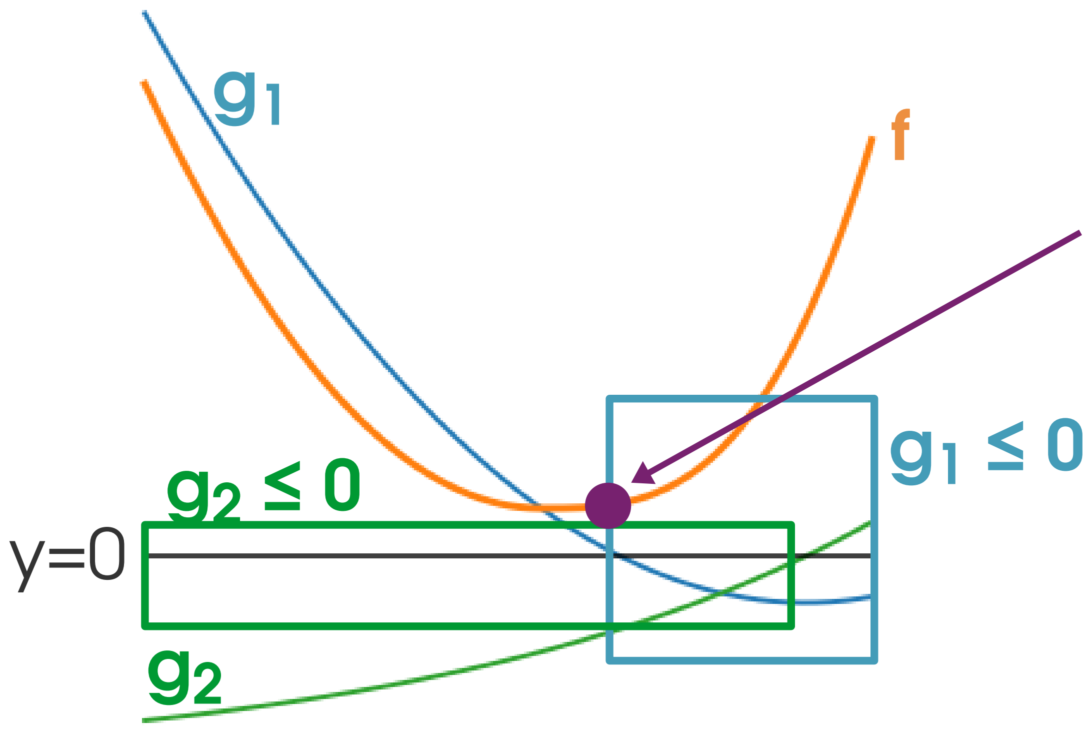
$$\boxed{\begin{aligned} &\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \\ \text{subject to } & g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ & h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$



Constrained convex optimization

- **Convex optimization program/problem:**

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \\ \text{subject to } & g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ & h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$



Constrained convex optimization

- **Convex optimization program/problem:**

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \\ \text{subject to } & g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ & h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$

  - f is the **objective function**
  - $g_i, i = 1, \ldots, m$ are the **inequality constraints**
  - $h_j, j = 1, \ldots, r$ are the **equality constraints**
  - $\boldsymbol{v} \in D$ that verifies all constraints is a **feasible point**
    $g_i(\boldsymbol{v}) \leq 0, i = 1, \ldots, m \text{ and } h_j(\boldsymbol{v}) = 0, j = 1, \ldots, r$
  - The set of all feasible points is the **feasible region**

$$\{\boldsymbol{v} : \boldsymbol{v} \in D; g_i(\boldsymbol{v}) \leq 0, i = 1, \ldots, m; h_j(\boldsymbol{v}) = 0, j = 1, \ldots, r\}$$

Constrained convex optimization

- **Convex optimization program/problem:**

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u}\in D} f(\boldsymbol{u}) \\ \text{subject to } &g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ &h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$

  - Assuming it exists, the solution f^*, that is to say, the minimum value of f over all feasible points, is the **optimal value (optimum)**
  - $\boldsymbol{u}$ feasible such that $f(\boldsymbol{u}) = f^*$ is called **optimal**, or a **minimizer** (it needs not be unique).
  - If u is feasible and $g_i(\boldsymbol{u}) = 0$ then g_i is **active** at u.

Local & global optima

For convex optimization problems,

local minima are global minima!

If **u** is feasible and minimizes f in a local neighborhood:

$f(\boldsymbol{u}) \leq f(\boldsymbol{v})$ for all feasible $\boldsymbol{v}$, $||\boldsymbol{u} - \boldsymbol{v}||_2^2 \leq \epsilon$

then **u** minimizes f globally.

Local & global optima

For convex optimization problems,

local minima are global minima!

If **u** is feasible and minimizes f in a local neighborhood:

$$f(\boldsymbol{u}) \leq f(\boldsymbol{v}) \text{ for all feasible } \boldsymbol{v},\ ||\boldsymbol{u}-\boldsymbol{v}||_2^2 \leq \epsilon$$

then **u** minimizes f globally.

Suppose $\bar{\boldsymbol{u}}$ is feasible and a local optimum of f: $\exists \epsilon > 0 : f(\bar{\boldsymbol{u}}) \leq f(\boldsymbol{v})$ for all feasible $\boldsymbol{v}$ such that $||\bar{\boldsymbol{u}} - \boldsymbol{v}|| \leq \epsilon$. Suppose $\boldsymbol{u}^*$ is a global optimum, $\boldsymbol{u}^* \neq \bar{\boldsymbol{u}}$. Then

$$f(\boldsymbol{u}^*) < f(\bar{\boldsymbol{u}}) \tag{1}$$

and

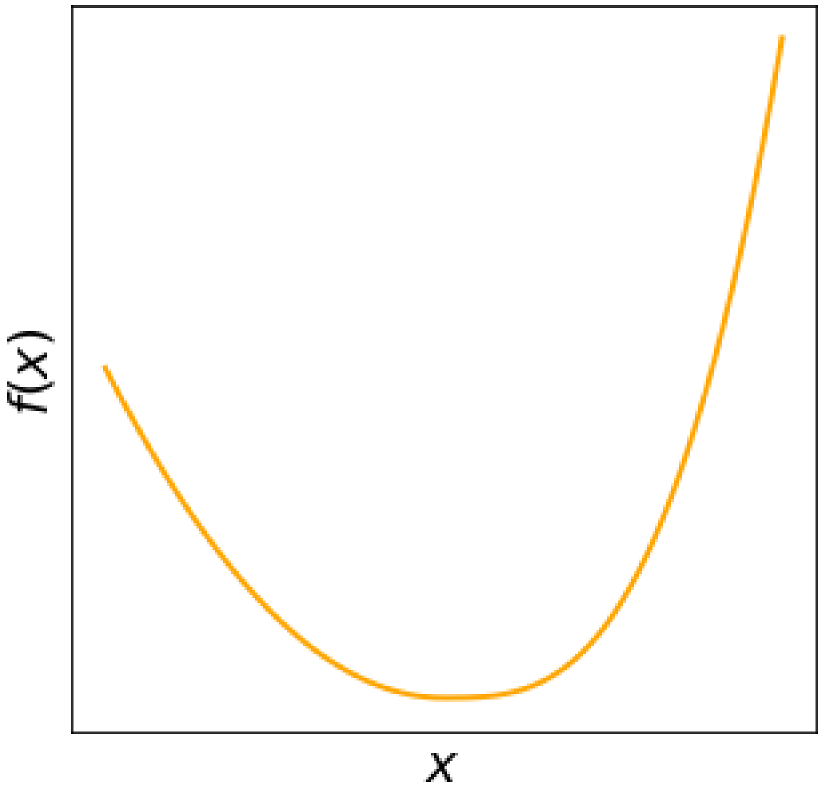
$$||\bar{\boldsymbol{u}} - \boldsymbol{u}^*|| > \epsilon \tag{2}$$

Consider $\boldsymbol{u} = (1-\lambda)\bar{\boldsymbol{u}} + \lambda \boldsymbol{u}^*$, where $\lambda = \frac{1}{2}\frac{\epsilon}{||\bar{\boldsymbol{u}} - \boldsymbol{u}^*||}$. Because of (2), $0 \leq \lambda < 1$. Because the feasible set is convex and both $\bar{\boldsymbol{u}}$ and $\boldsymbol{u}^*$ are feasible, $\boldsymbol{u}$ is feasible.

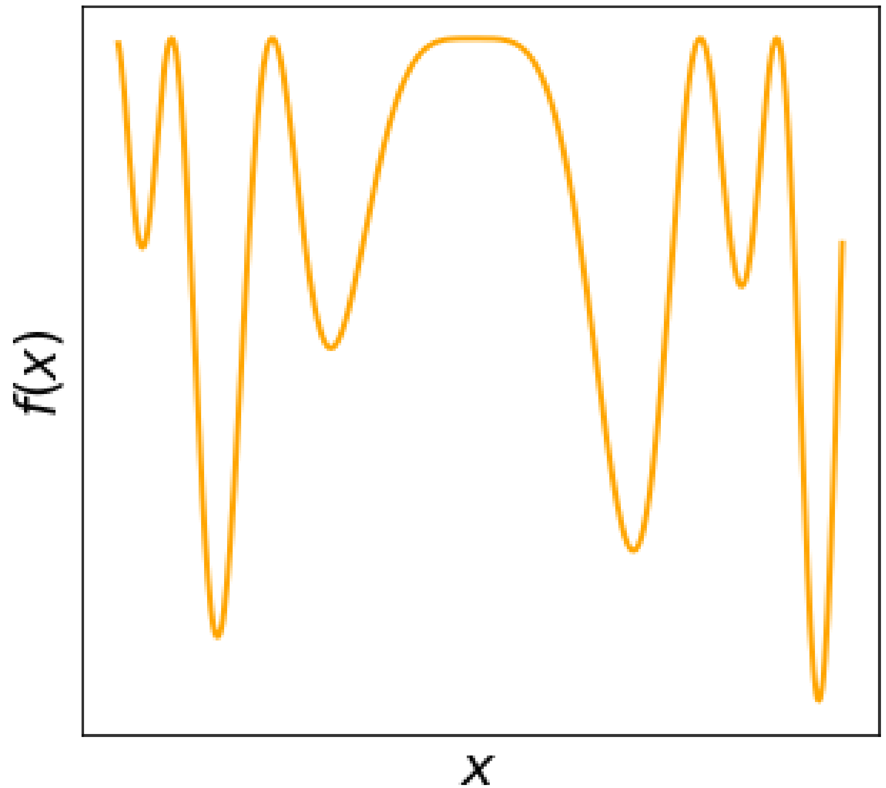
$||\boldsymbol{u} - \bar{\boldsymbol{u}}|| = ||(1-\lambda)\bar{\boldsymbol{u}} + \lambda \boldsymbol{u}^*|| = \lambda ||\boldsymbol{u}^* - \bar{\boldsymbol{u}}|| = \epsilon/2 < \epsilon$, hence $f(\boldsymbol{u}) \geq f(\bar{\boldsymbol{u}})$ (as $\bar{\boldsymbol{u}}$ is a local minimum).

But because f is convex, $f(\boldsymbol{u}) \leq (1-\lambda) f(\bar{\boldsymbol{u}}) + \lambda f(\boldsymbol{u}^) = f(\bar{\boldsymbol{u}}) + \lambda(f(\boldsymbol{u}^*) - f(\bar{\boldsymbol{u}})) < f(\bar{\boldsymbol{u}})$. (The last inequality comes from (1).) The contradiction implies that it is impossible that $\boldsymbol{u}^* \neq \bar{\boldsymbol{u}}$, hence the local optimum is also global.

Why talk about convex optimization?



convex



non-convex

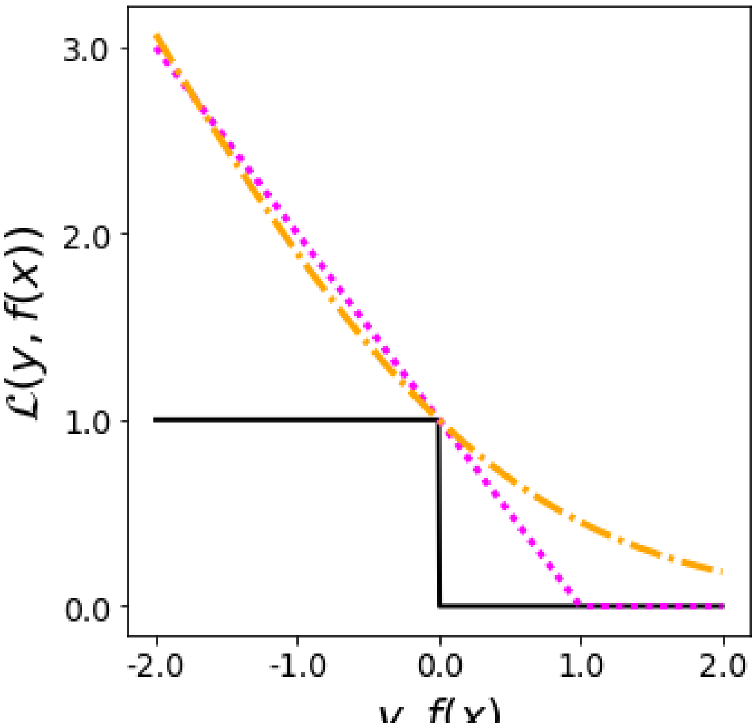
- Convex optimization is “easy”.
- We’ll often try to formulate ML problems as convex optimization problems.

Why talk about convex optimization?

- Supervised ML: **empirical risk minimization**

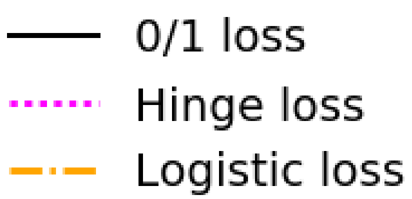
$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}(y^i, f(\boldsymbol{x}^i))$$

- **Losses for classification**



The 0/1 loss is non-convex.
We'll replace it with other losses.

$y \in \{-1, 1\}$



Unconstrained convex optimization

First-order characterization

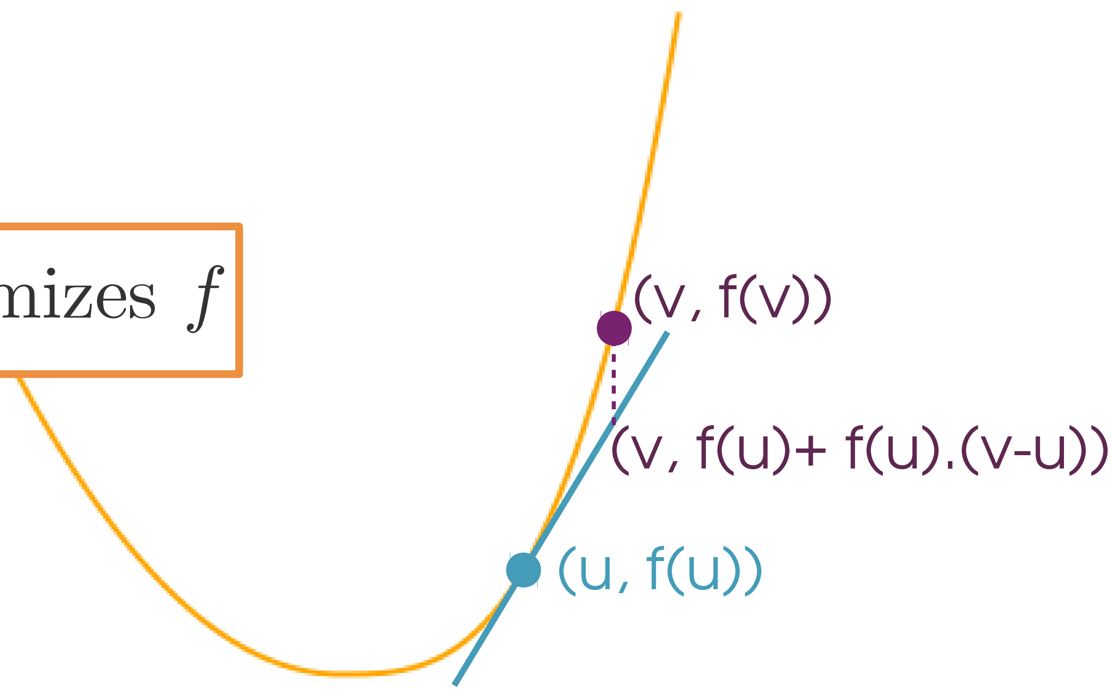
- Suppose f differentiable
- **Given the first-order characterization of convex functions, how can we solve** $\min_{\boldsymbol{u} \in \mathrm{dom}(f)} f(\boldsymbol{u})$ **?**

First-order characterization

- Suppose f differentiable
- **Given the first-order characterization of convex functions, how can we solve** $\min_{\boldsymbol{u} \in \mathrm{dom}(f)} f(\boldsymbol{u})$ **?**

REMINDER

$$\boxed{\nabla f(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} \text{ minimizes } f}$$

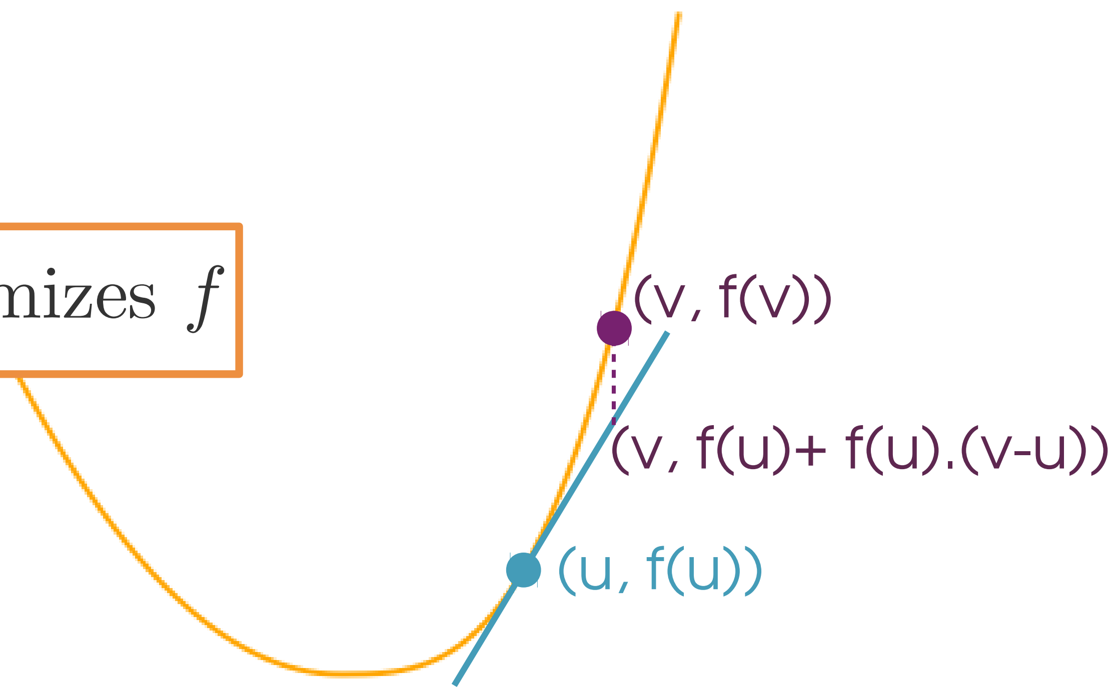


First-order characterization

- Suppose f differentiable
- **Given the first-order characterization of convex functions, how can we solve** $\min_{\boldsymbol{u} \in \mathrm{dom}(f)} f(\boldsymbol{u})$ **?**

Set the gradient of f to 0

$$\boxed{\nabla f(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} \text{ minimizes } f}$$



First-order characterization

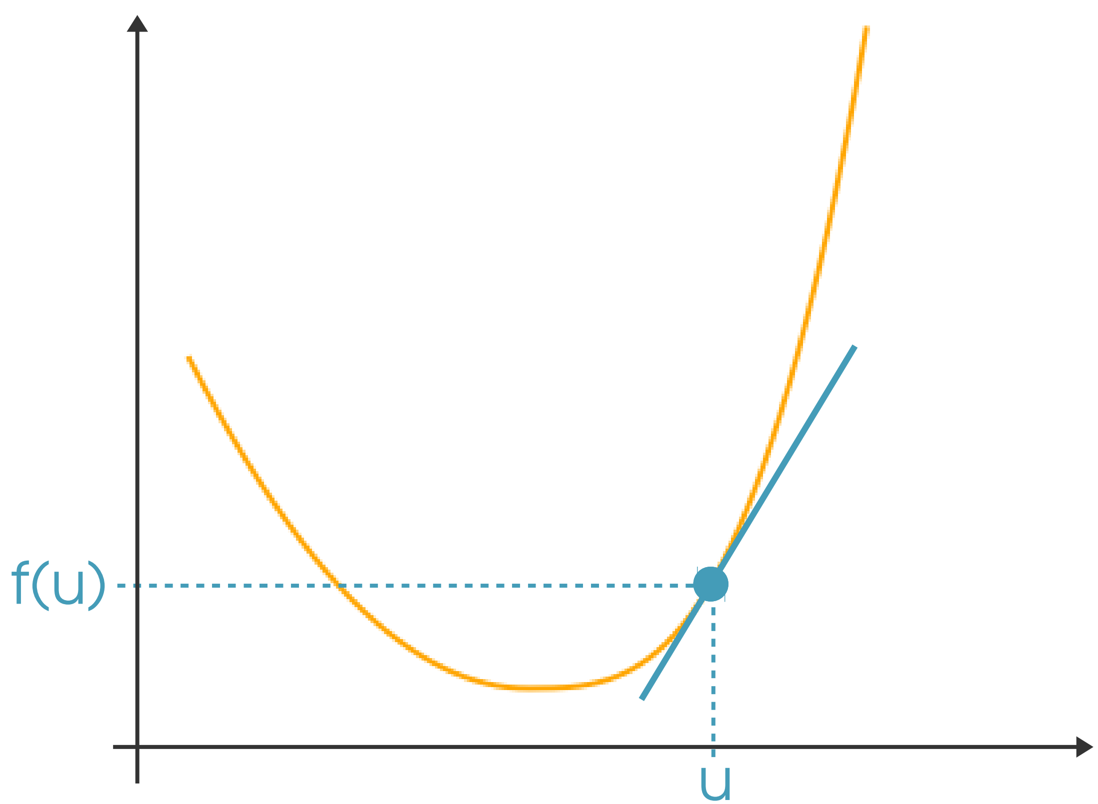
- Suppose f differentiable
- **Given the first-order characterization of convex functions, how can we solve** $\min_{\boldsymbol{u} \in \mathrm{dom}(f)} f(\boldsymbol{u})$ **?**

Set the gradient of f to 0

- But what if $\nabla f(\boldsymbol{u}) = 0$ cannot be solved?

Gradient descent

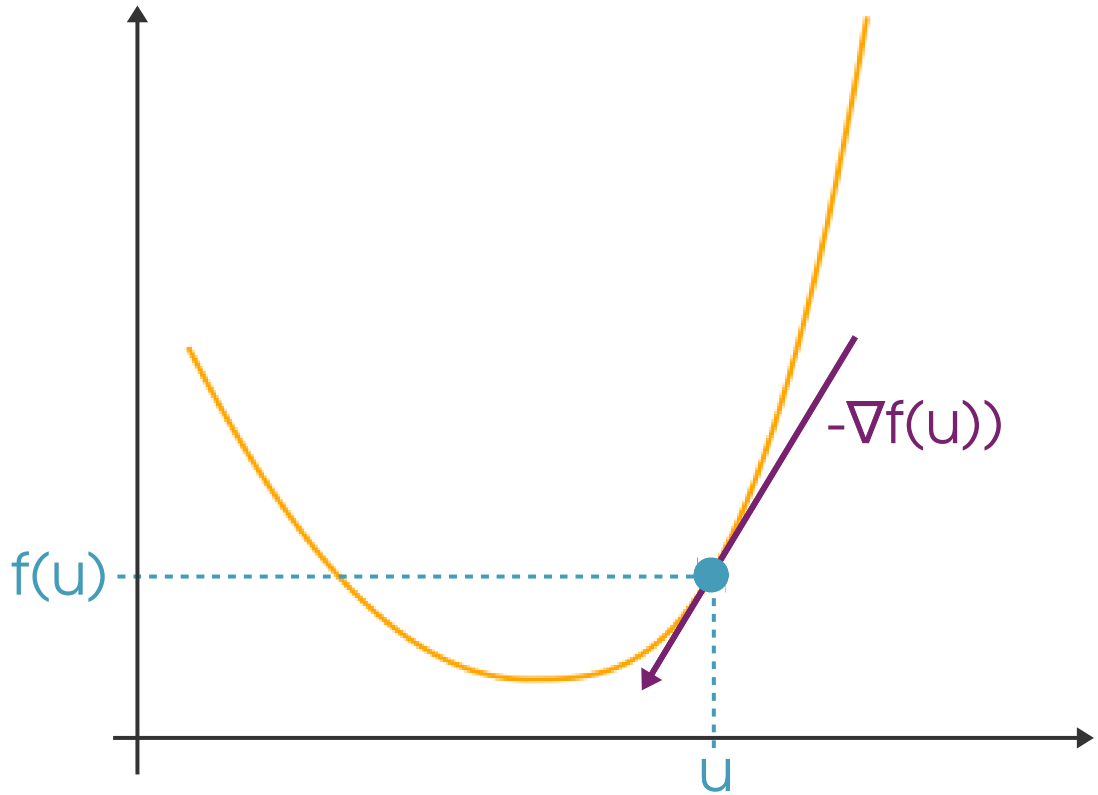
- Start from a random point **u**.
- **How do I get closer to the solution?**



Gradient descent

- Start from a random point **u**.
- **How do I get closer to the solution?**
- Follow the **opposite of the gradient.**

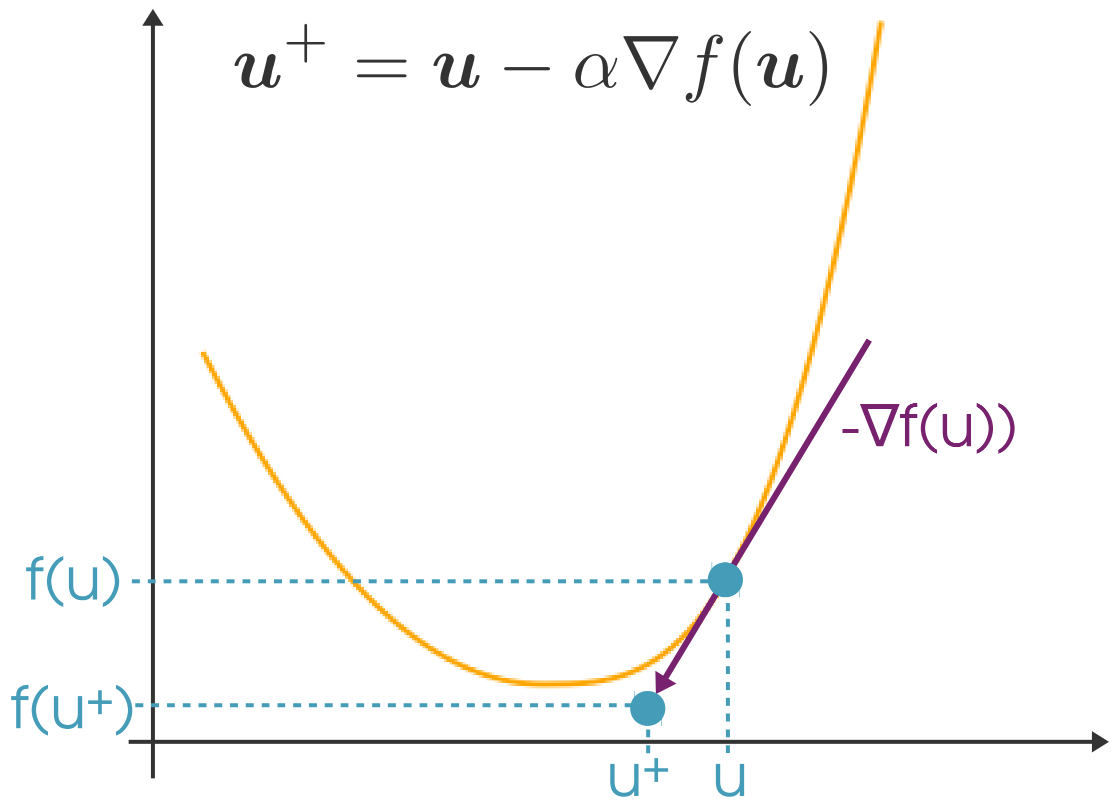
The gradient indicates the direction of steepest increase.



Gradient descent

- Start from a random point **u**.
- **How do I get closer to the solution?**
- Follow the **opposite of the gradient.**

The gradient indicates the direction of steepest increase.



Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$$

- Stop at some point

Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\alpha_k} \nabla f(\boldsymbol{u}^{(k-1)})$$

step size

- Stop at some point

stopping criterion

Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\alpha_k} \nabla f(\boldsymbol{u}^{(k-1)})$$

step size

- Stop at some point

stopping criterion

Usually: stop when $||\nabla f(\boldsymbol{u}^{(k)})||_2 < \epsilon \quad \epsilon = 10^{-\mathrm{sthg}}$

$$\nabla f(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u} \text{ minimizes } f$$

Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\alpha_k} \nabla f(\boldsymbol{u}^{(k-1)})$$

step size

  - **If the step size is too big,**

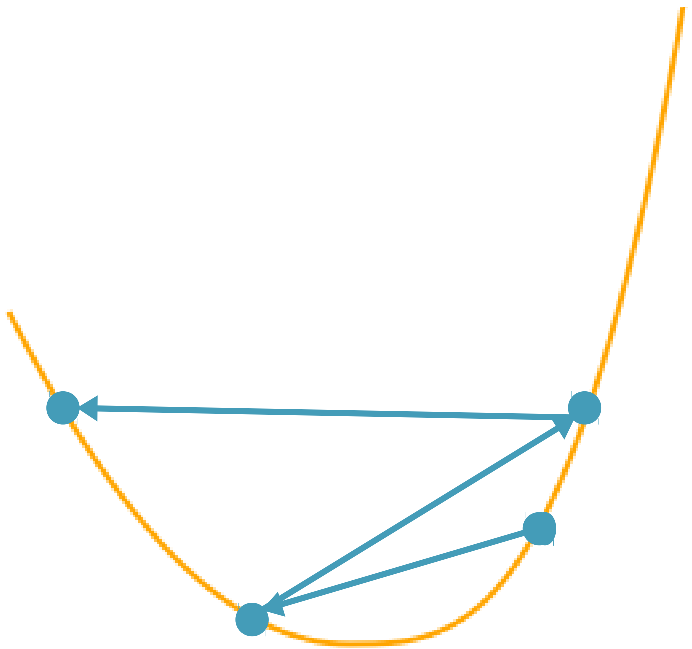
Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\alpha_k} \nabla f(\boldsymbol{u}^{(k-1)})$$

step size

  – **If the step size is too big,**



Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \underbrace{\boxed{\alpha_k}}_{\text{step size}} \nabla f(\boldsymbol{u}^{(k-1)})$$

  - **If the step size is too big,** the search might diverge

Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \underbrace{\boxed{\alpha_k}}_{\text{step size}} \nabla f(\boldsymbol{u}^{(k-1)})$$

  - **If the step size is too big,** t[illegible]earch might diverge
  - **If the step size is too small**

Gradient descent algorithm

- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \underbrace{\boxed{\alpha_k}}_{\text{step size}} \nabla f(\boldsymbol{u}^{(k-1)})$$

  - **If the step size is too big,** the search might diverge
  - **If the step size is too small,** the search might take a very long time

Gradient descent algorithm

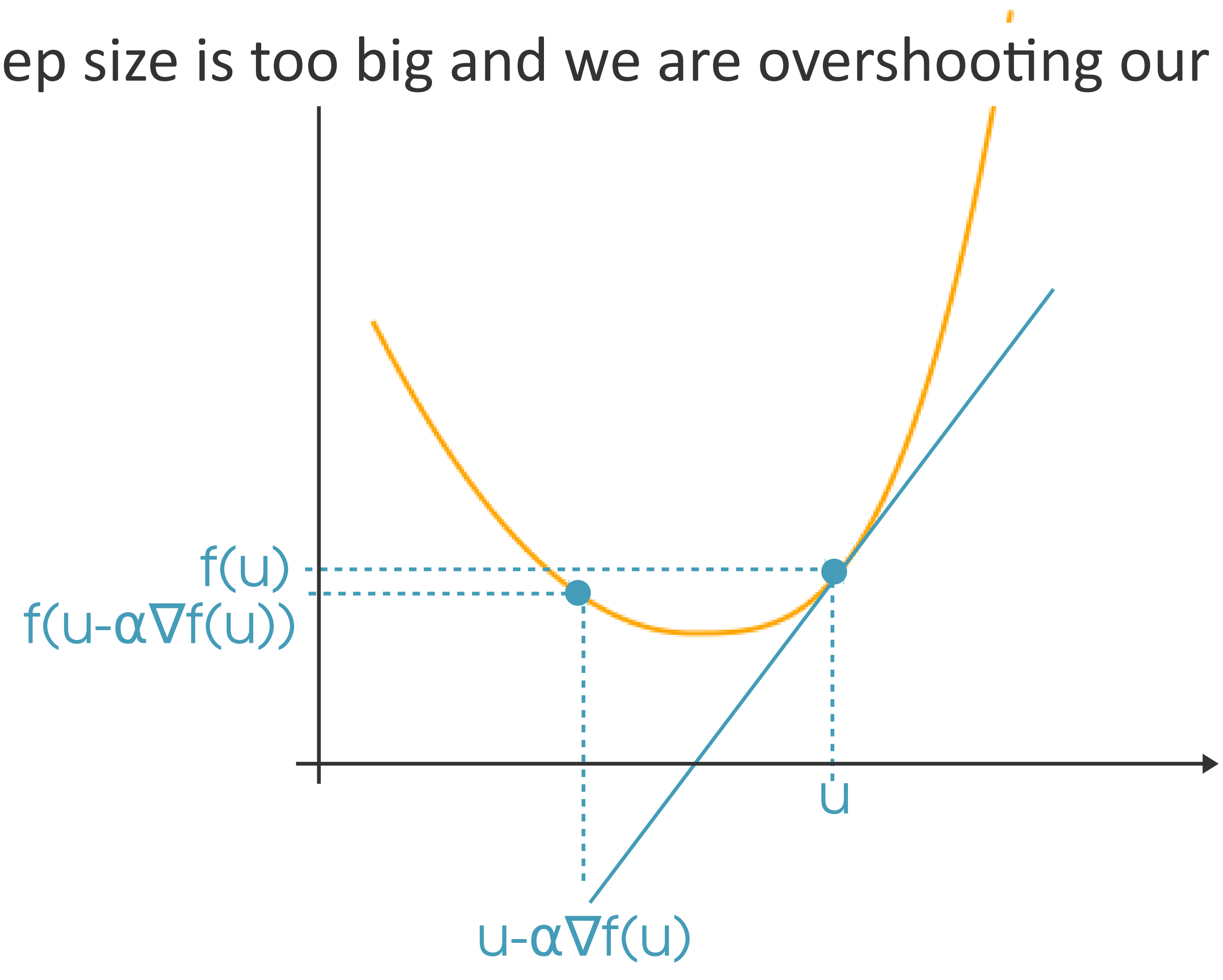
- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \underbrace{\boxed{\alpha_k}}_{\text{step size}} \nabla f(\boldsymbol{u}^{(k-1)})$$

  - **If the step size is too big,** the search might diverge
  - **If the step size is too small,** the search might take a very long time
  - Backtracking line search makes it possible to chose the step size adaptively.

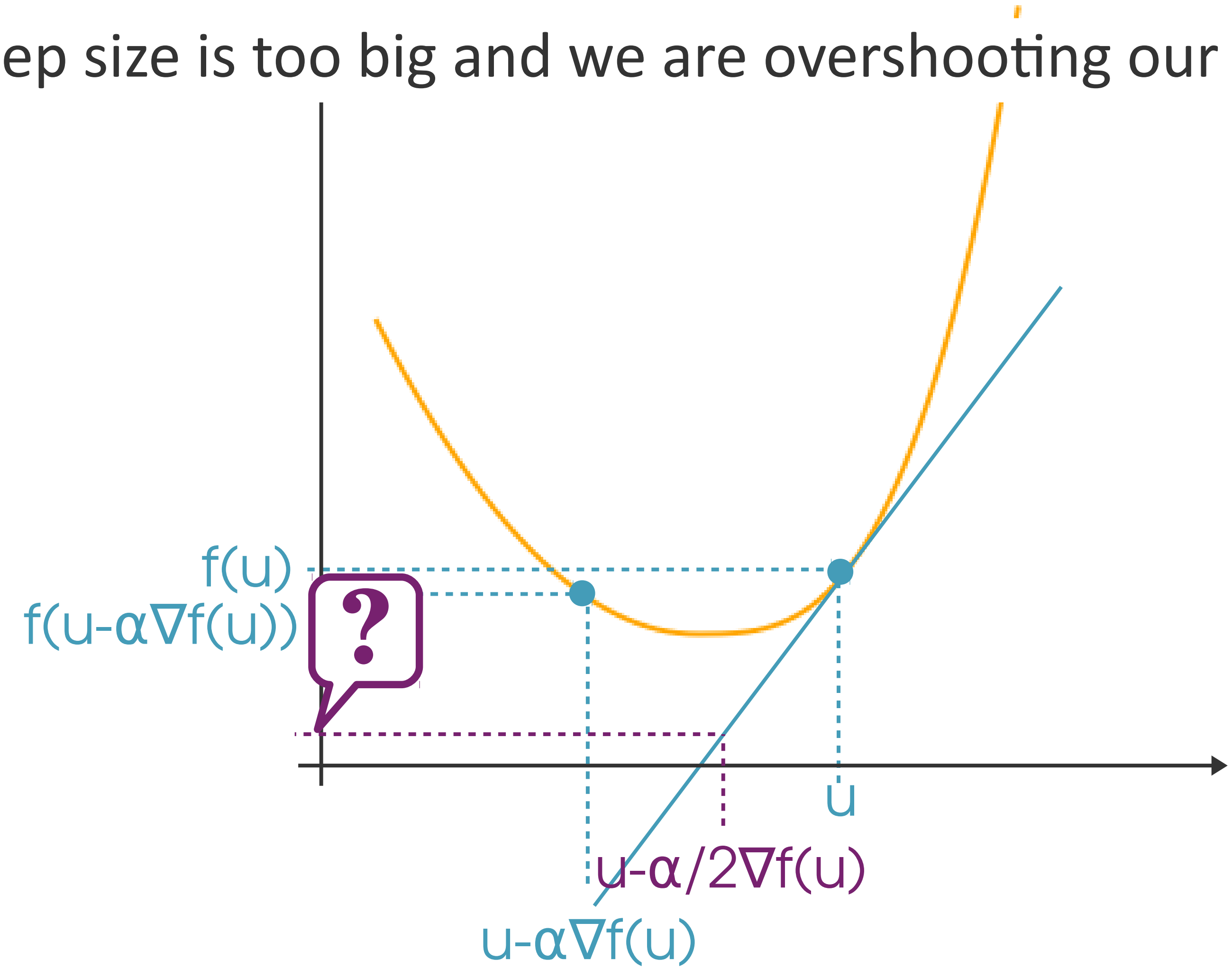
BLS: shrinking needed

The step size is too big and we are overshooting our goal.



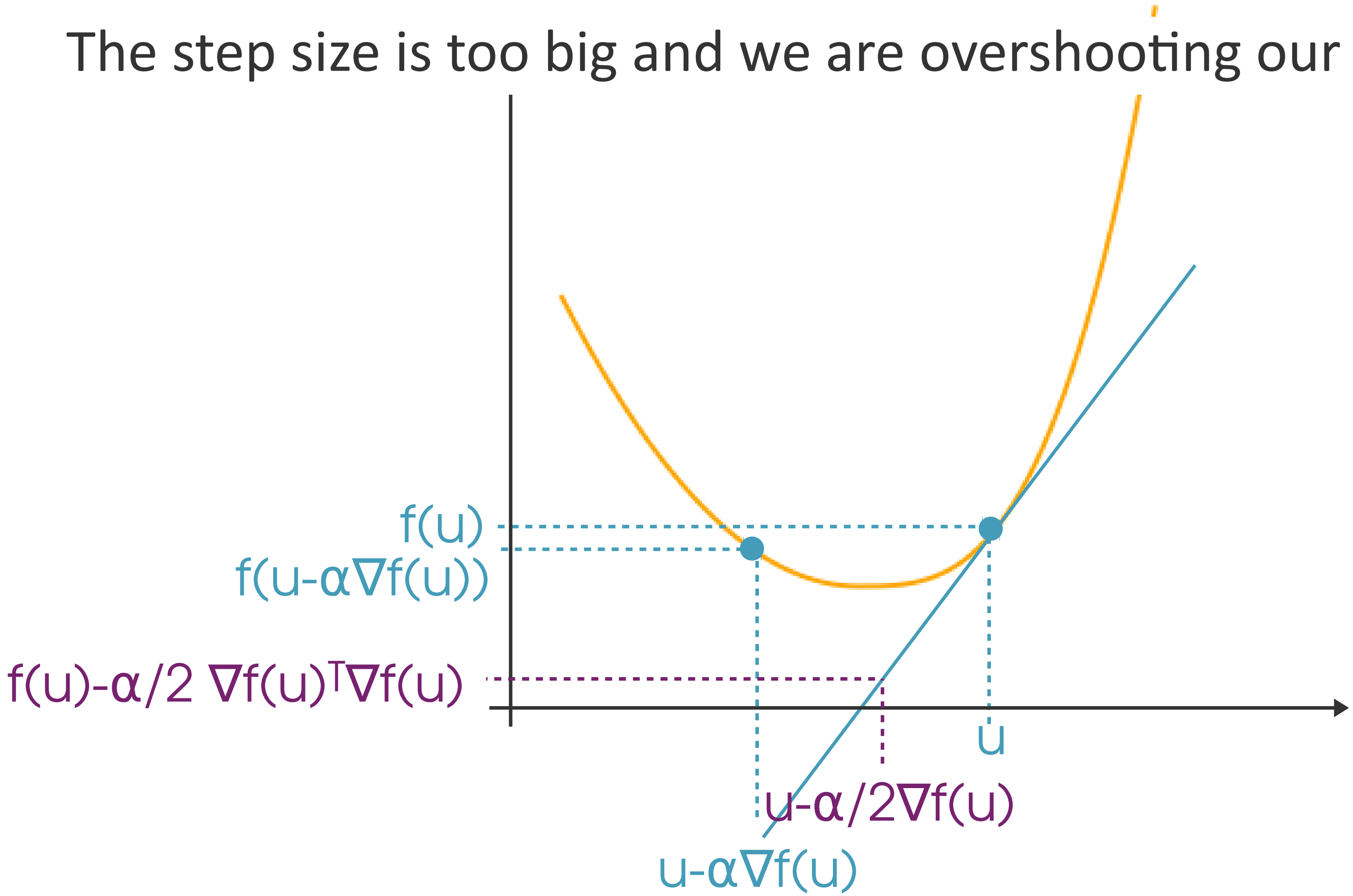
BLS: shrinking needed

The step size is too big and we are overshooting our goal.



BLS: shrinking needed

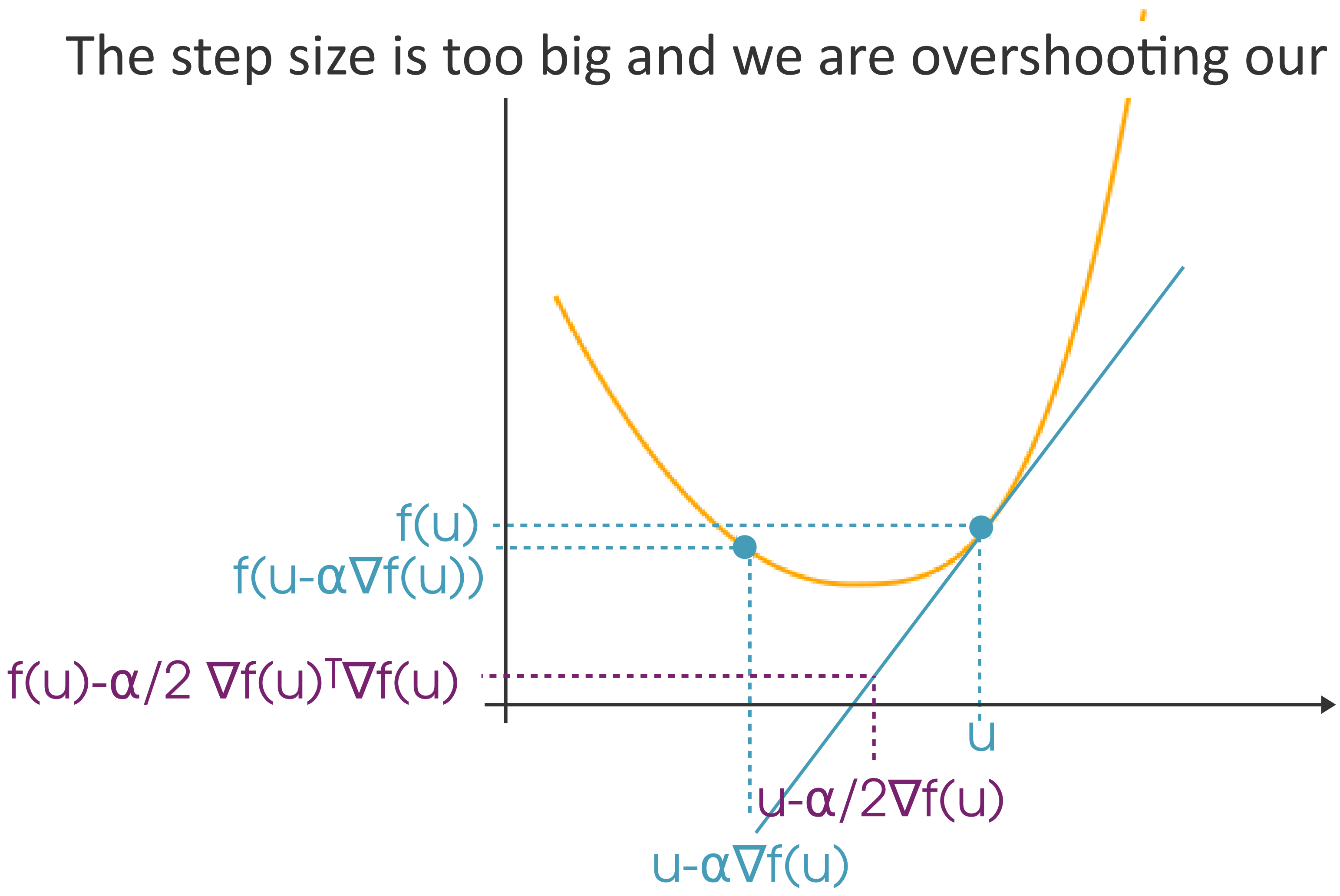
The step size is too big and we are overshooting our goal.



BLS: shrinking needed

$f(u-\alpha\nabla f(u)) \quad > \quad f(u)-\alpha/2\, \nabla f(u)^T \nabla f(u)$

The step size is too big and we are overshooting our goal.

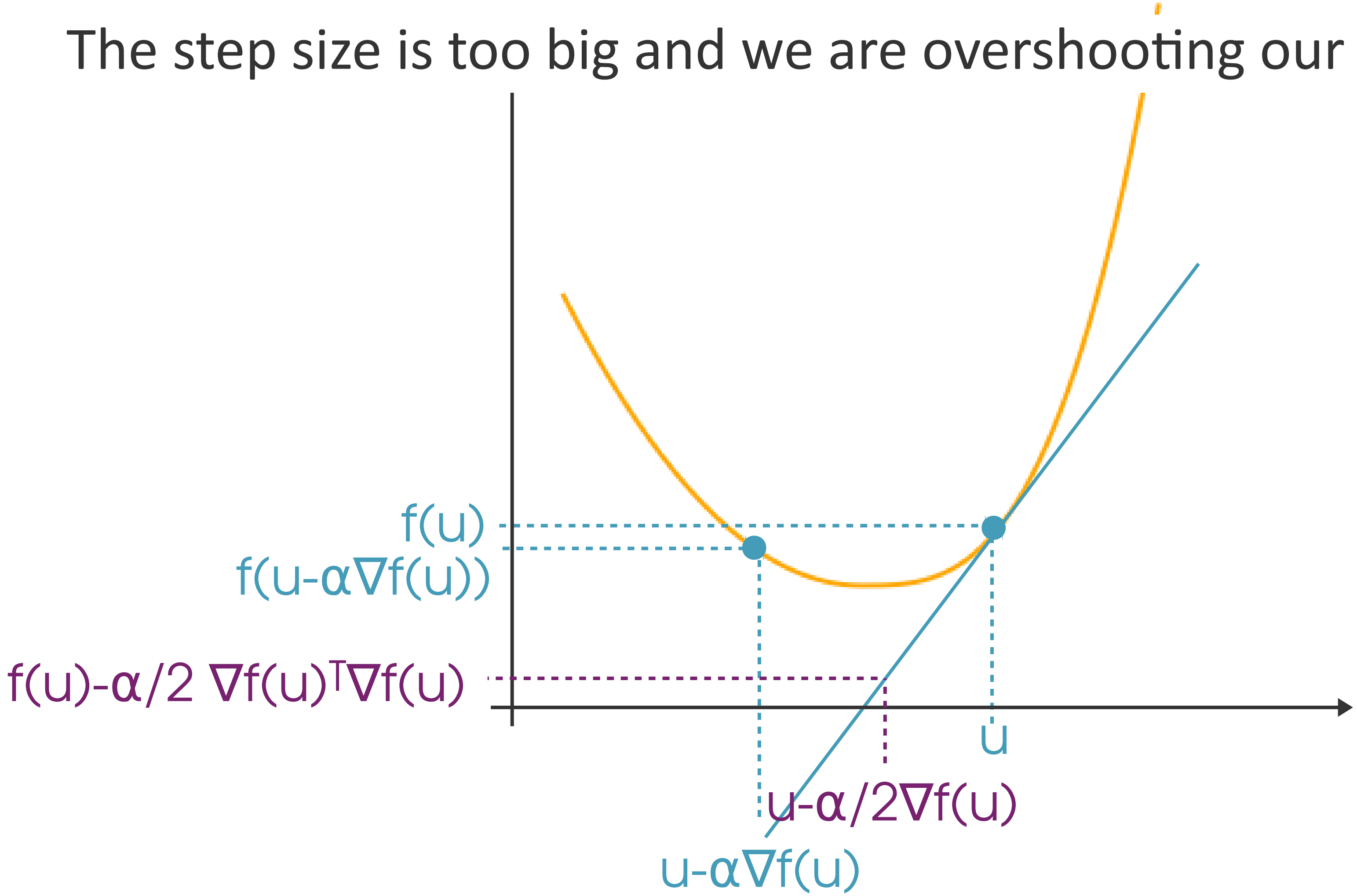


BLS: shrinking needed

$$f(\boldsymbol{u}^{(k-1)} - \alpha_{k-1}\nabla f(\boldsymbol{u}^{(k-1)})) > f(\boldsymbol{u}^{(k-1)}) - \frac{1}{2}\alpha_{k-1}||\nabla f(\boldsymbol{u}^{(k-1)})||_2^2$$

f(u-α∇f(u)) > f(u)-α/2 ∇f(u)^T∇f(u)

The step size is too big and we are overshooting our goal.

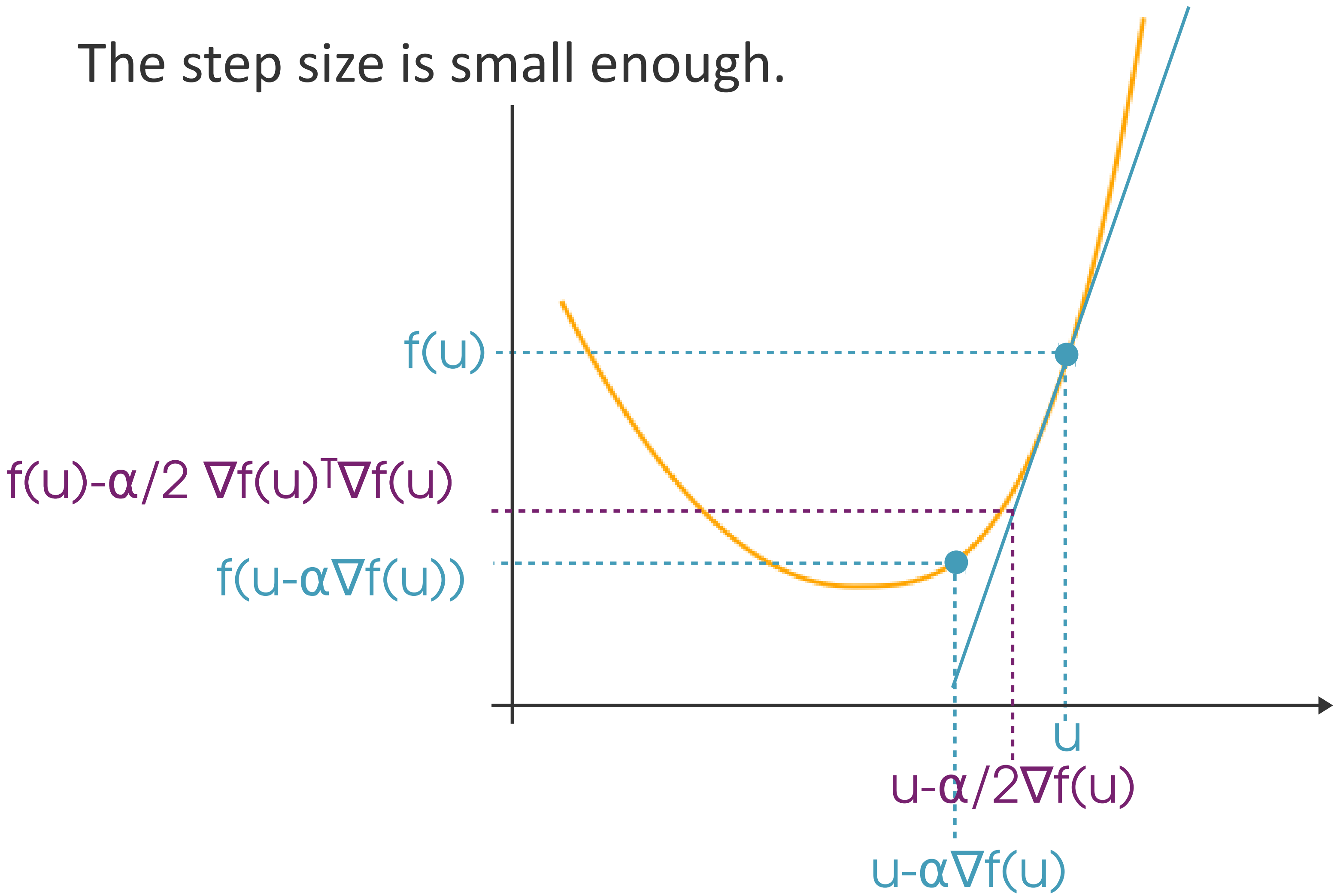


BLS: no shrinking needed

$$f(\boldsymbol{u}^{(k-1)} - \alpha_{k-1}\nabla f(\boldsymbol{u}^{(k-1)})) \leq f(\boldsymbol{u}^{(k-1)}) - \frac{1}{2}\alpha_{k-1}||\nabla f(\boldsymbol{u}^{(k-1)})||_2^2$$

$f(u-\alpha\nabla f(u)) \quad \leq \quad f(u)-\alpha/2\ \nabla f(u)^T\nabla f(u)$

The step size is small enough.



Backtracking line search

- **Shrinking parameter** $0 < \beta < 1$, **initial step size** α_0
- Choose an initial point $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
- Repeat for k=1, 2, 3, …
  - If $f(\boldsymbol{u}^{(k-1)} - \alpha_{k-1} \nabla f(\boldsymbol{u}^{(k-1)})) > f(\boldsymbol{u}^{(k-1)}) - \frac{1}{2}\alpha_{k-1} ||\nabla f(\boldsymbol{u}^{(k-1)})||_2^2$

    shrink the step size: $\alpha_k = \beta \alpha_{k-1}$
  - Else: $\alpha_k = \alpha_{k-1}$
  - Update: $\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$
- Stop when $||\nabla f(\boldsymbol{u}^{(k)})||_2 < \epsilon \quad \epsilon = 10^{-\mathrm{sthg}}$

Newton's method

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$$

- Suppose f is twice derivable
- Second-order Taylor's expansion:

$$f(\boldsymbol{v}) \approx f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u}) + \frac{1}{2}(\boldsymbol{v} - \boldsymbol{u})^\top \nabla^2 f(\boldsymbol{u})(\boldsymbol{v} - \boldsymbol{u})$$

- Minimize in **v** instead of in **u**

Newton's method

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$$

- Suppose f is twice derivable
- Second-order Taylor's expansion:

$$f(\boldsymbol{v}) \approx \underbrace{\boxed{f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^\top (\boldsymbol{v} - \boldsymbol{u}) + \frac{1}{2}(\boldsymbol{v} - \boldsymbol{u})^\top \nabla^2 f(\boldsymbol{u})(\boldsymbol{v} - \boldsymbol{u})}}_{g(\boldsymbol{v})}$$

- Minimize in **v** instead of in **u**

Newton's method

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$$

- Suppose f is twice derivable
- Second-order Taylor's expansion:

$$f(\boldsymbol{v}) \approx \underbrace{\boxed{f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^\top (\boldsymbol{v}-\boldsymbol{u}) + \frac{1}{2}(\boldsymbol{v}-\boldsymbol{u})^\top \nabla^2 f(\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})}}_{g(\boldsymbol{v})}$$

- Minimize in **v**: $\nabla_{\boldsymbol{v}} g(\boldsymbol{v}) = \nabla f(\boldsymbol{u}) + \nabla^2 f(\boldsymbol{u})(\boldsymbol{v}-\boldsymbol{u})$

$$\nabla_{\boldsymbol{v}} g(\boldsymbol{v}) = 0 \;\Rightarrow\; \boldsymbol{v}-\boldsymbol{u} = -\left(\nabla^2 f(\boldsymbol{u})\right)^{-1} \nabla f(\boldsymbol{u})$$

$$\boxed{\alpha_k = \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1}}$$

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1}} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})$ for δ_k
- **What is the new update rule?**

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \boxed{\left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1}} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})$ for δ_k
- **New update rule:**

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \delta_k$$

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})}$ for δ_k

This is a problem of the form $A\boldsymbol{x} - \boldsymbol{b} = 0 \quad A \succeq 0$

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})}$ for δ_k

This is a problem of the form $A\boldsymbol{x} - \boldsymbol{b} = 0$ $\boxed{A \succeq 0}$?

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})}$ for δ_k

This is a problem of the form $A\boldsymbol{x} - \boldsymbol{b} = 0$ $\boxed{A \succeq 0}$

Second-order characterization of convex functions

Newton CG (conjugate gradient)

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1} \nabla f(\boldsymbol{u}^{(k-1)})$$

- Computing the inverse of the Hessian is **computationally intensive.**
- Instead, compute $\nabla^2 f(\boldsymbol{u}^{(k-1)})$ and $\nabla f(\boldsymbol{u}^{(k-1)})$ and solve $\boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})}$ for δ_k

This is a problem of the form $A\boldsymbol{x} - \boldsymbol{b} = 0 \quad A \succeq 0$

Solve using the **conjugate gradient method.**

Conjugate gradient method

Solve $Ax - \boldsymbol{b} = 0 \quad A \succeq 0$

- Idea: build a set of **A-conjugate vectors** (basis of $\mathbb{R}^n$)

  $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} : \boxed{\boldsymbol{v}_i^\top A \boldsymbol{v}_j = 0} \; \forall i \neq j$

  – **Initialisation:** $\boldsymbol{v}_0 = \boldsymbol{r}_0 = b - A\boldsymbol{x}^{(0)}$

  – At step t:

  - **Update rule:** $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$
  - **residual** $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$
  - $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$

$$\boxed{\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}}$$

$$\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}$$

ensures

$$\boldsymbol{v}_t^\top A \boldsymbol{v}_{t-1} = 0$$

$$\boldsymbol{v}_t^\top A \boldsymbol{v}_i = 0 \;\; \forall i < t$$

  – **Convergence:**
  $\boldsymbol{r}_i^\top \boldsymbol{r}_j = 0 \;\; \forall i \neq j$ hence $\boxed{\boldsymbol{r}_n = 0}$

Conjugate gradient method

Prove $\boxed{\boldsymbol{v}_t^\top A \boldsymbol{v}_{t-1} = 0}$?

Given

- **Initialisation:** $\boldsymbol{v}_0 = \boldsymbol{r}_0 = b - A\boldsymbol{x}^{(0)}$
- At step t:
  - **Update rule:** $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$
  - **residual** $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$
  - $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$

and assuming $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$

$$\boxed{\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}}$$

$$\boxed{\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}}$$

Prove $\boldsymbol{v}_t^\top A\boldsymbol{v}_{t-1} = 0$

$$\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A\boldsymbol{v}_{t-1}}$$

$$\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}$$

Given

- **Initialisation:** $\boldsymbol{v}_0 = \boldsymbol{r}_0 = b - A\boldsymbol{x}^{(0)}$
- **Update rule:** $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$
- **residual** $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$
- $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$ and assuming $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$

By definition, $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$. Hence $\boldsymbol{r}_t = b - A(\boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1})$ and

$$\boldsymbol{r}_t = \boldsymbol{r}_{t-1} - \alpha_t A\boldsymbol{v}_{t-1}. \tag{1}$$

By definition, $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$ and $\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}$. Hence $\boldsymbol{v}_t^\top A\boldsymbol{v}_{t-1} = \boldsymbol{r}_t^\top A\boldsymbol{v}_{t-1} + \beta_t \boldsymbol{v}_{t-1}^\top A\boldsymbol{v}_{t-1}$ and therefore

$$\boldsymbol{v}_t^\top A\boldsymbol{v}_{t-1} = \boldsymbol{r}_t^\top A\boldsymbol{v}_{t-1} + \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}} \boldsymbol{v}_{t-1}^\top A\boldsymbol{v}_{t-1}.$$

Because $\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A\boldsymbol{v}_{t-1}}$,

$$\boldsymbol{v}_t^\top A\boldsymbol{v}_{t-1} = \boldsymbol{r}_t^\top A\boldsymbol{v}_{t-1} + \frac{1}{\alpha_t}\boldsymbol{r}_t^\top \boldsymbol{r}_t.$$

From (1), $A\boldsymbol{v}_{t-1} = \frac{1}{\alpha_t}(\boldsymbol{r}_{t-1} - \boldsymbol{r}_t)$, and therefore

$$\boldsymbol{v}_t^\top A\boldsymbol{v}_{t-1} = \frac{1}{\alpha_t}\boldsymbol{r}_t^\top (\boldsymbol{r}_{t-1} - \boldsymbol{r}_t) + \frac{1}{\alpha_t}\boldsymbol{r}_t^\top \boldsymbol{r}_t = \frac{1}{\alpha_t}\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0.$$

This is true if and only if we have shown that $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$.

Conjugate gradient method

Prove $\boxed{\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0}$ **and conclude the proof**

Given

- **Initialisation:** $\boldsymbol{v}_0 = \boldsymbol{r}_0 = b - A\boldsymbol{x}^{(0)}$
- At step t:
  - **Update rule:** $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$
  - **residual** $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$
  - $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$

$$\boxed{\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}}$$

$$\boxed{\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}}$$

Prove $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$

$$\alpha_t = \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}$$

$$\beta_t = \frac{\boldsymbol{r}_t^\top \boldsymbol{r}_t}{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}$$

Given

- **Initialisation:** $\boldsymbol{v}_0 = \boldsymbol{r}_0 = b - A\boldsymbol{x}^{(0)}$
- **Update rule:** $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \alpha_t \boldsymbol{v}_{t-1}$
- **residual** $\boldsymbol{r}_t = \boldsymbol{b} - A\boldsymbol{x}^{(t)}$
- $\boldsymbol{v}_t = \boldsymbol{r}_t + \beta_t \boldsymbol{v}_{t-1}$

From (1),

$$\begin{aligned}\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} &= \boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1} - \alpha_t \left(A\boldsymbol{v}_{t-1}\right)^\top \boldsymbol{r}_{t-1} \\ &= \boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1} - \frac{\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}} \boldsymbol{v}_{t-1}^\top A^\top \boldsymbol{r}_{t-1} = \left(\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}\right)\left(1 - \frac{\boldsymbol{v}_{t-1}^\top A^\top \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}\right).\end{aligned}$$

Because $A \succeq 0$, $A = A^\top$ and hence

$$\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = \left(\boldsymbol{r}_{t-1}^\top \boldsymbol{r}_{t-1}\right)\left(1 - \frac{\boldsymbol{v}_{t-1}^\top A \boldsymbol{r}_{t-1}}{\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1}}\right). \tag{2}$$

- If $t = 1$, because $\boldsymbol{r}_0 = \boldsymbol{v}_0$, $\boldsymbol{r}_1^\top \boldsymbol{r}_0 = 0$. From the previous proof, we now have that $\boldsymbol{v}_1^\top A \boldsymbol{v}_0 = 0$.
- For $t = 2$, by definition $\boldsymbol{v}_{t-1} = \boldsymbol{r}_{t-1} + \beta_{t-1}\boldsymbol{v}_{t-2}$ and we can replace $\boldsymbol{r}_{t-1}$ accordingly to find that

  $$\boldsymbol{v}_{t-1}^\top A \boldsymbol{r}_{t-1} = \boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-1} - \beta_{t-1}\boldsymbol{v}_{t-1}^\top A \boldsymbol{v}_{t-2}.$$

  Because we have shown that $\boldsymbol{v}_1^\top A \boldsymbol{v}_0 = 0$, we conclude from (2) that $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$.
- We can iterate this procedure to show alternatively that $\boldsymbol{v}_t^\top A \boldsymbol{v}_{t-1} = 0$ and $\boldsymbol{r}_t^\top \boldsymbol{r}_{t-1} = 0$ for all values of t.

Quasi-Newton methods

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \delta_k \qquad \boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})}\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})$$

- What if the Hessian is unavailable / expensive to compute at each iteration?
- **Approximate the inverse Hessian:** $\delta_k = \boxed{W_{k-1}} \nabla f(\boldsymbol{u}^{(k-1)})$

  update $W_k \approx (\nabla^2 f(\boldsymbol{u}^{(k)}))^{-1}$ iteratively
- **Conditions:**
  - $W_k \succeq 0$
  - **Secant equation:** $\boxed{\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v}) \approx \nabla^2 f(\boldsymbol{u})(\boldsymbol{u} - \boldsymbol{v})}$

    1st order Taylor applied to ∇f

$$\Rightarrow W_k \left(\nabla f(\boldsymbol{u}^{(k)}) - \nabla f(\boldsymbol{u}^{(k-1)}) \right) = \boldsymbol{u}^{(k)} - \boldsymbol{u}^{(k-1)}$$

- **Initialization:** Identity $W_0 = I_n$

Quasi-Newton methods

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \delta_k \qquad \boxed{\nabla^2 f(\boldsymbol{u}^{(k-1)})}\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})$$

- What if the Hessian is unavailable / expensive to compute at each iteration?
- **Approximate the inverse Hessian:** $\delta_k = \boxed{W_{k-1}} \nabla f(\boldsymbol{u}^{(k-1)})$

  update $W_k \approx (\nabla^2 f(\boldsymbol{u}^{(k)}))^{-1}$ iteratively

- **BFGS:** Broyden-Fletcher-Goldfarb-Shanno

$$W_k = W_{k-1} - \frac{\boldsymbol{s}_k \boldsymbol{y}_k^\top W_{k-1} + W_{k-1} \boldsymbol{y}_k \boldsymbol{s}_k^\top}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle} + \left(1 + \frac{\langle \boldsymbol{y}_k, W_k \boldsymbol{y}_k \rangle}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle}\right) \frac{\boldsymbol{s}_k \boldsymbol{s}_k^\top}{\langle \boldsymbol{y}_k, \boldsymbol{s}_k \rangle}$$

$$\boldsymbol{s}_k = \boldsymbol{u}^{(k)} - \boldsymbol{u}^{(k-1)} \qquad \boldsymbol{y}_k = \nabla f(\boldsymbol{u}^{(k)}) - \nabla f(\boldsymbol{u}^{(k-1)})$$

- **L-BFGS:** Limited memory variant

  Do not store the full matrix W_k.

Stochastic gradient descent

- For $f : \boldsymbol{u} \mapsto \sum_{i=1}^{m} h_i(\boldsymbol{u})$

- **Gradient descent:**

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \sum_{i=1}^{m} \nabla h_i(\boldsymbol{u}^{(k-1)})$$

- **Stochastic gradient descent:**

$$\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla h_{\boxed{i_k}}(\boldsymbol{u}^{(k-1)})$$

  - **Cyclic**: cycle over 1, 2, …, m, 1, 2, …, m, …
  - **Randomized:** chose i_k uniformely at random in {1, 2, …, m}.

Coordinate Descent

- For $f : \mathbb{R}^n \to \mathbb{R}$
$$\boldsymbol{u} \mapsto g(\boldsymbol{u}) + \sum_{i=1}^{n} h_i(u_i)$$
  - g: convex and differentiable
  - h_i: convex ⇒ the **non-smooth** part of f is **separable.**
- Minimize **coordinate by coordinate:**
  - Initialisation: $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
  - For k=1, 2, ...: $u_1^{(k)} \in \arg\min_{\underline{u_1}} f(\boxed{u_1}, u_2^{(k-1)}, \ldots, u_n^{(k-1)})$

$$u_2^{(k)} \in \arg\min_{\underline{u_2}} f(u_1^{(k)}, \boxed{u_2}, \ldots, u_n^{(k-1)})$$

$$\ldots$$

$$u_n^{(k)} \in \arg\min_{\underline{u_n}} f(u_1^{(k)}, u_2^{(k)}, \ldots, \boxed{u_n})$$

Coordinate Descent

- For $f : \mathbb{R}^n \to \mathbb{R}$
$$\boldsymbol{u} \mapsto g(\boldsymbol{u}) + \sum_{i=1}^{n} h_i(u_i)$$
  - g: convex and differentiable
  - h_i: convex ⇒ the **non-smooth** part of f is **separable.**
- Minimize **coordinate by coordinate:**
  - Initialisation: $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
  - For k=1, 2, …:
$$\boxed{u_1^{(k)}} \in \underset{\underline{u_1}}{\arg\min}\, f(\boxed{u_1}, u_2^{(k-1)}, \ldots, u_n^{(k-1)})$$
$$\boxed{u_2^{(k)}} \in \underset{\underline{u_2}}{\arg\min}\, f(\boxed{u_1^{(k)}}, \boxed{u_2}, \ldots, u_n^{(k-1)})$$
$$\ldots$$
$$u_n^{(k)} \in \underset{\underline{u_n}}{\arg\min}\, f(\boxed{u_1^{(k)}, u_2^{(k)}, \ldots,} \boxed{u_n})$$

Coordinate Descent

- For $f : \mathbb{R}^n \to \mathbb{R}$
$$\boldsymbol{u} \mapsto g(\boldsymbol{u}) + \sum_{i=1}^{n} h_i(u_i)$$
  - g: convex and differentiable
  - h_i: convex ⇒ the **non-smooth** part of f is **separable.**
- Minimize **coordinate by coordinate:**
  - Initialisation: $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$
  - $\boxed{\text{For k=1, 2, …:}}$
$$u_1^{(k)} \in \arg\min_{u_1} f(u_1, u_2^{(k-1)}, \ldots, u_n^{(k-1)})$$
$$u_2^{(k)} \in \arg\min_{u_2} f(u_1^{(k)}, u_2, \ldots, u_n^{(k-1)})$$

Variants:

– re-order the coordinates randomly

– Proceed by blocks of coordinates (2 or more at a time)

Summary: Unconstrained convex optimization

If f is **differentiable**

- Set its gradient to zero
- If hard to solve: **gradient descent** $\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)} - \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$

Setting the learning rate:

- **Backtracking Line Search** (adapt heuristically to avoid "overshooting")
- **Newton's method:** Suppose f **twice differentiable**
  - $\alpha_k = \left(\nabla^2 f(\boldsymbol{u}^{(k-1)})\right)^{-1}$
  - If the Hessian is hard to invert, compute $\delta_k = \alpha_k \nabla f(\boldsymbol{u}^{(k-1)})$ by solving $\nabla^2 f(\boldsymbol{u}^{(k-1)})\delta_k = \nabla f(\boldsymbol{u}^{(k-1)})$ by the **conjugate gradient method**
  - If the Hessian is hard to compute, approximate the inverse Hessian with a **quasi-Newton method** such as **BFGS** (**L-BFGS**: less memory)

- If f is separable: **stochastic gradient descent**
- If the non-smooth part of f is separable: **coordinate descent.**

Constrained convex optimization

Constrained convex optimization

- **Convex optimization program/problem:**

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u}\in D} f(\boldsymbol{u}) \\ \text{subject to } &g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ &h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$

  - f is **convex**
  - $g_i, i = 1, \ldots, m$ are **convex**
  - $h_j, j = 1, \ldots, r$ are **affine** $h_j : \boldsymbol{u} \mapsto \boldsymbol{a}_j^\top \boldsymbol{u} + b_j$

  - The **feasible set** is convex

$$\mathcal{C} = \{\boldsymbol{v} : \boldsymbol{v} \in D; g_i(\boldsymbol{v}) \leq 0, i = 1, \ldots, m; h_j(\boldsymbol{v}) = 0, j = 1, \ldots, r\}$$

Lagrangian

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u})$$

$$\text{subject to } g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m$$
$$h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r$$

- **Lagrangian:** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$

$$\boldsymbol{u}, \boxed{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mapsto f(\boldsymbol{u}) + \sum_{i=1}^{m} \boxed{\alpha_i} g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \boxed{\beta_i} h_j(\boldsymbol{u})$$

= Lagrange multipliers
= dual variables

Lagrange dual function

- Lagrangian: $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}$

$$\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mapsto f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_i h_j(\boldsymbol{u})$$

- **Lagrange dual function:**

$$Q : \qquad \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}$$

$$\boldsymbol{\alpha}, \boldsymbol{\beta} \mapsto \inf_{\boldsymbol{u} \in D} L(\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Infimum = the greatest value x
such that x ≤ L(u, α, β)

- Q is **concave** (independently of the convexity of f)

Lagrange dual function

- $Q: \quad \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$

$$\boldsymbol{\alpha}, \boldsymbol{\beta} \mapsto \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_i h_j(\boldsymbol{u})$$

- Q is **concave** (independently of the convexity of f)

Consider $(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1), (\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2)$ and $0 \leq \lambda \leq 1$. Set $\boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2$ and $\boldsymbol{\beta} = \lambda \boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2$.
$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \boldsymbol{\alpha}^\top \boldsymbol{g}(\boldsymbol{u}) + \boldsymbol{\beta}^\top \boldsymbol{h}(\boldsymbol{u})$,
where $\boldsymbol{g}(\boldsymbol{u}) = (g_1(\boldsymbol{u}), g_2(\boldsymbol{u}), \ldots g_m(\boldsymbol{u}))$ and $\boldsymbol{h}(\boldsymbol{u}) = (h_1(\boldsymbol{u}), h_2(\boldsymbol{u}), \ldots h_r(\boldsymbol{u}))$. Hence

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \lambda \boldsymbol{\alpha}_1^\top \boldsymbol{g}(\boldsymbol{u}) + (1-\lambda)\boldsymbol{\alpha}_2^\top \boldsymbol{g}(\boldsymbol{u}) + \lambda \boldsymbol{\beta}_1^\top \boldsymbol{h}(\boldsymbol{u}) + (1-\lambda)\boldsymbol{\beta}_2^\top \boldsymbol{h}(\boldsymbol{u}) \\ &= \inf_{\boldsymbol{u} \in D} \lambda \left(f(\boldsymbol{u}) + \boldsymbol{\alpha}_1^\top \boldsymbol{g}(\boldsymbol{u}) + \boldsymbol{\beta}_1^\top \boldsymbol{h}(\boldsymbol{u}) \right) + (1-\lambda) \left(f(\boldsymbol{u}) + \boldsymbol{\alpha}_2^\top \boldsymbol{g}(\boldsymbol{u}) + \boldsymbol{\beta}_2^\top \boldsymbol{h}(\boldsymbol{u}) \right). \end{aligned}$$

This last equality holds because $f(\boldsymbol{u}) = \lambda f(\boldsymbol{u}) + (1-\lambda) f(\boldsymbol{u})$. Hence

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) &\geq \lambda \inf_{\boldsymbol{u} \in D} \left(f(\boldsymbol{u}) + \boldsymbol{\alpha}_1^\top \boldsymbol{g}(\boldsymbol{u}) + \boldsymbol{\beta}_1^\top \boldsymbol{h}(\boldsymbol{u}) \right) + (1-\lambda) \inf_{\boldsymbol{u} \in D} \left(f(\boldsymbol{u}) + \boldsymbol{\alpha}_2^\top \boldsymbol{g}(\boldsymbol{u}) + \boldsymbol{\beta}_2^\top \boldsymbol{h}(\boldsymbol{u}) \right) \\ &\geq \lambda Q(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) + (1-\lambda) Q(\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2). \end{aligned}$$

$\blacksquare$

Lagrange dual function

- The dual function gives a **lower bound** on our solution

Let $p^* = \min_{\boldsymbol{u} \in \boxed{\mathcal{C}}} f(\boldsymbol{u})$ ← **feasible set**

Then for any $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ $\quad \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots \alpha_m \geq 0$

$\boldsymbol{\beta} \in \mathbb{R}^r$

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$$

Consider any feasible point $\boldsymbol{u}$. Then $g_1(\boldsymbol{u}) \leq 0, g_2(\boldsymbol{u}) \leq 0, \ldots g_m(\boldsymbol{u}) \leq 0$ and $h_1(\boldsymbol{u}) = h_2(\boldsymbol{u}) = \cdots = h_r(\boldsymbol{u}) = 0$. By definition, $L(\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j h_j(\boldsymbol{u})$. Because the second term is negative ($\alpha_i \geq 0$, $g_i(\boldsymbol{u}) \leq 0$) and the third one is zero ($h_j(\boldsymbol{u}) = 0$), $L(\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq f(\boldsymbol{u})$ for every feasible point $\boldsymbol{u}$. Hence $\inf_{\boldsymbol{u} \in D} L(\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\boldsymbol{u} \in \mathcal{C}} f(\boldsymbol{u})$.

Weak duality

- $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$ for any $\boldsymbol{\alpha} \in \mathbb{R}^m_+, \boldsymbol{\beta} \in \mathbb{R}^r$
- What is the **best lower bound** on p* we can get?

Weak duality

- $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$ for any $\boldsymbol{\alpha} \in \mathbb{R}_+^m, \boldsymbol{\beta} \in \mathbb{R}^r$
- What is the **best lower bound** on p* we can get?

$$\boxed{\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\text{subject to } \ \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_m \geq 0 \end{aligned}}$$

Weak duality

- $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$ for any $\boldsymbol{\alpha} \in \mathbb{R}_+^m, \boldsymbol{\beta} \in \mathbb{R}^r$
- What is the **best lower bound** on p* we can get?

$$\boxed{\begin{aligned}&\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\text{subject to } \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_m \geq 0\end{aligned}}$$

Lagrange dual problem

Weak duality

- $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$ for any $\boldsymbol{\alpha} \in \mathbb{R}^m_+, \boldsymbol{\beta} \in \mathbb{R}^r$
- What is the **best lower bound** on p* we can get?

$$\boxed{\begin{aligned} &\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\text{subject to } \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_m \geq 0 \end{aligned}}$$

Lagrange dual problem

- Optimal values **α*, β*** of **α, β** are called **dual optimal** or **optimal Lagrange multipliers.**
- Original optimization problem = **primal**
- The dual is a **convex optimization problem** (even if the primal is not!)

Weak duality

- Let d* be the solution to the dual problem

$$d^* = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$\text{subject to } \alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_m \geq 0$$

- Because for every dual admissible **α, β** $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq p^*$

$$\boxed{d^* \leq p^*}$$

Weak duality
(always holds)

Strong duality & Slater's conditions

- **Strong duality:** $\boxed{d^* = p^*}$
  - Does not hold in general
  - But often holds for convex optimization problems

Strong duality & Slater's conditions

- **Strong duality:** $\boxed{d^* = p^*}$
  - Does not hold in general
  - But often holds for convex optimization problems
- **Constraint qualifications:** conditions under which strong duality holds (in addition to convexity)

Strong duality & Slater's conditions

- **Strong duality:** $\boxed{d^* = p^*}$
  - Does not hold in general
  - But often holds for convex optimization problems
- **Constraint qualifications:** conditions under which strong duality holds (in addition to convexity)
- In particular: **Slater's conditions:**
  - If the primal is **convex** and there exists at least one **strictly feasible point** (i.e. the inequalities hold strictly), then strong duality holds
  - Strict inequalities only need to hold for **non-affine** constraints.

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ [strong duality]

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*) \qquad \text{[definition of inf]}$$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} \qquad \text{[definition of inf]}$$

What is the sign of this expression?

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \quad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \quad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} \quad \text{[definition of inf]}$$

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$
\begin{aligned}
f(\boldsymbol{u}^*) &= Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) && \text{[strong duality]} \\
&= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) && \text{[definition of Q]} \\
&\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} && \text{[definition of inf]} \\
&\leq f(\boldsymbol{u}^*).
\end{aligned}
$$

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} \qquad \text{[definition of inf]}$$

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$\leq f(\boldsymbol{u}^*).$$

- Hence ?

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} \qquad \text{[definition of inf]}$$

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$\leq f(\boldsymbol{u}^*).$$

- Hence all above inequalities are equalities

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \qquad \text{[strong duality]}$$

$$= \inf_{\boldsymbol{u} \in D} f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}) \qquad \text{[definition of Q]}$$

$$\leq f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)} \qquad \text{[definition of inf]}$$

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$\leq f(\boldsymbol{u}^*).$$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$ [strong duality]

$$= \inf_{\boldsymbol{u}\in D} \overbrace{\boxed{f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u})}}^{L(\boldsymbol{u}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}$$ [definition of Q]

$$= f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)}$$ [definition of inf]

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$= f(\boldsymbol{u}^*).$$

- $f(\boldsymbol{u}^*) = L(\boldsymbol{u}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \inf_{\boldsymbol{u}\in D} L(\boldsymbol{u}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$ [strong duality]

$$= \inf_{\boldsymbol{u} \in D} \overbrace{f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u})}^{L(\boldsymbol{u}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}$$ [definition of Q]

$$= f(\boldsymbol{u}^*) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)$$ [definition of inf]

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$= f(\boldsymbol{u}^*).$$

- $f(\boldsymbol{u}^*) = L(\boldsymbol{u}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \inf_{\boldsymbol{u} \in D} L(\boldsymbol{u}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$

$$\boxed{\nabla_{\boldsymbol{u}} L(\boldsymbol{u}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = 0}$$

stationarity

Karush-Kuhn-Tucker conditions

- Suppose f, g_i, h_j differentiable + strong duality

$$f(\boldsymbol{u}^*) = Q(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$ [strong duality]

$$= \inf_{\boldsymbol{u}\in D} \underbrace{\boxed{f(\boldsymbol{u}) + \sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u})}}^{L(\boldsymbol{u}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}$$ [definition of Q]

$$= f(\boldsymbol{u}^*) + \boxed{\sum_{i=1}^{m} \alpha_i^* g_i(\boldsymbol{u}^*)} + \boxed{\sum_{j=1}^{r} \beta_j^* h_j(\boldsymbol{u}^*)}$$ [definition of inf]

u* feasible ⇒ $h_j = 0$

u* feasible ⇒ $g_i \leq 0$

α* feasible ⇒ $\alpha_i \geq 0$

$$= f(\boldsymbol{u}^*).$$

complementary slackness

- $\boxed{\alpha_i^* g_i(\boldsymbol{u}^*) = 0}$

$$\alpha_i > 0 \Rightarrow g_i(\boldsymbol{u}^*) = 0$$
$$g_i(\boldsymbol{u}^*) < 0 \Rightarrow \alpha_i = 0$$

Karush-Kuhn-Tucker conditions

- Let's sum up all of our conditions:
  - **Primal feasibility:** $g_i(\boldsymbol{u}^*) \leq 0 \qquad i = 1, \ldots, m$
    $h_j(\boldsymbol{u}^*) = 0 \qquad j = 1, \ldots, r$
  - **Dual feasibility:** $\alpha_i^* \geq 0 \qquad i = 1, \ldots, m$
  - **Complementary slackness:** $\alpha_i^* g_i(\boldsymbol{u}^*) = 0 \qquad i = 1, \ldots, m$
  - **Stationarity:**

$$\nabla_u f(\boldsymbol{u}^*) + \sum_{i=1}^{m} \alpha_i^* \nabla_u g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* \nabla_u h_j(\boldsymbol{u}^*) = 0.$$

Karush-Kuhn-Tucker conditions

- Let's sum up all of our conditions:
  - **Primal feasibility:** $g_i(\boldsymbol{u}^*) \leq 0 \qquad i = 1, \ldots, m$

    $h_j(\boldsymbol{u}^*) = 0 \qquad j = 1, \ldots, r$
  - **Dual feasibility:** $\alpha_i^* \geq 0 \qquad i = 1, \ldots, m$
  - **Complementary slackness:** $\alpha_i^* g_i(\boldsymbol{u}^*) = 0 \qquad i = 1, \ldots, m$
  - **Stationarity:**

$$\nabla_u f(\boldsymbol{u}^*) + \sum_{i=1}^{m} \alpha_i^* \nabla_u g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* \nabla_u h_j(\boldsymbol{u}^*) = 0.$$

Karush-Kuhn-Tucker (KKT) conditions

Karush-Kuhn-Tucker conditions

- Let's sum up all of our conditions:
  - **Primal feasibility:** $g_i(\boldsymbol{u}^*) \leq 0 \qquad i = 1, \ldots, m$
    $h_j(\boldsymbol{u}^*) = 0 \qquad j = 1, \ldots, r$
  - **Dual feasibility:** $\alpha_i^* \geq 0 \qquad i = 1, \ldots, m$
  - **Complementary slackness:** $\alpha_i^* g_i(\boldsymbol{u}^*) = 0 \qquad i = 1, \ldots, m$
  - **Stationarity:**

$$\nabla_u f(\boldsymbol{u}^*) + \sum_{i=1}^{m} \alpha_i^* \nabla_u g_i(\boldsymbol{u}^*) + \sum_{j=1}^{r} \beta_j^* \nabla_u h_j(\boldsymbol{u}^*) = 0.$$

Karush-Kuhn-Tucker (KKT) conditions

- For **convex optimization problems,** any (**u, α, β**) that verify the KKT conditions are **optimal.**

Karush-Kuhn-Tucker conditions

- For **convex optimization problems,** any (**u, α, β**) that verify the KKT conditions are **optimal.**

Consider $\bar{\boldsymbol{u}}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}$ that verify the KKT conditions.

Primal feasibility implies that $\bar{\boldsymbol{u}}$ is feasible.

$L(\boldsymbol{u}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$ is convex in $\boldsymbol{u}$. Indeed $L(\boldsymbol{u}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = f(\boldsymbol{u}) + \sum_{i=1}^{m} \bar{\alpha}_i g_i(\boldsymbol{u}) + \sum_{j=1}^{r} \bar{\beta}_j h_j(\boldsymbol{u})$, and f and g_i are convex, h_j is affine, and *dual feasibility* implies $\bar{\alpha}_i \geq 0$.

Because $L(\boldsymbol{u}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$ is convex in $\boldsymbol{u}$, *stationarity* implies that $\bar{\boldsymbol{u}}$ minimizes $L(\boldsymbol{u}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$, hence $Q(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = L(\boldsymbol{u}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$. Because of *complementary slackness*, $\sum_{i=1}^{m} \bar{\alpha}_i g_i(\boldsymbol{u}) = 0$, and because of *primal feasibility*, $\sum_{j=1}^{r} \bar{\beta}_j h_j(\boldsymbol{u}) = 0$. Therefore, $Q(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = f(\bar{\boldsymbol{u}})$.

Let p^* be the optimal value of the primal, and d^* the optimal value of the dual. By definition of the optimal, $Q(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) \leq d^*$. By weak duality, $d^* \leq p^*$. Therefore $f(\bar{\boldsymbol{u}}) = Q(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) \leq d^* \leq p^*$. Because p^* is the optimal value of the primal, this implies $f(\bar{\boldsymbol{u}}) = p^*$ and hence all the above inequalities are equalities: $f(\bar{\boldsymbol{u}}) = p^* = d^* = Q(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$. Hence $\bar{\boldsymbol{u}}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}$ are optimal.

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

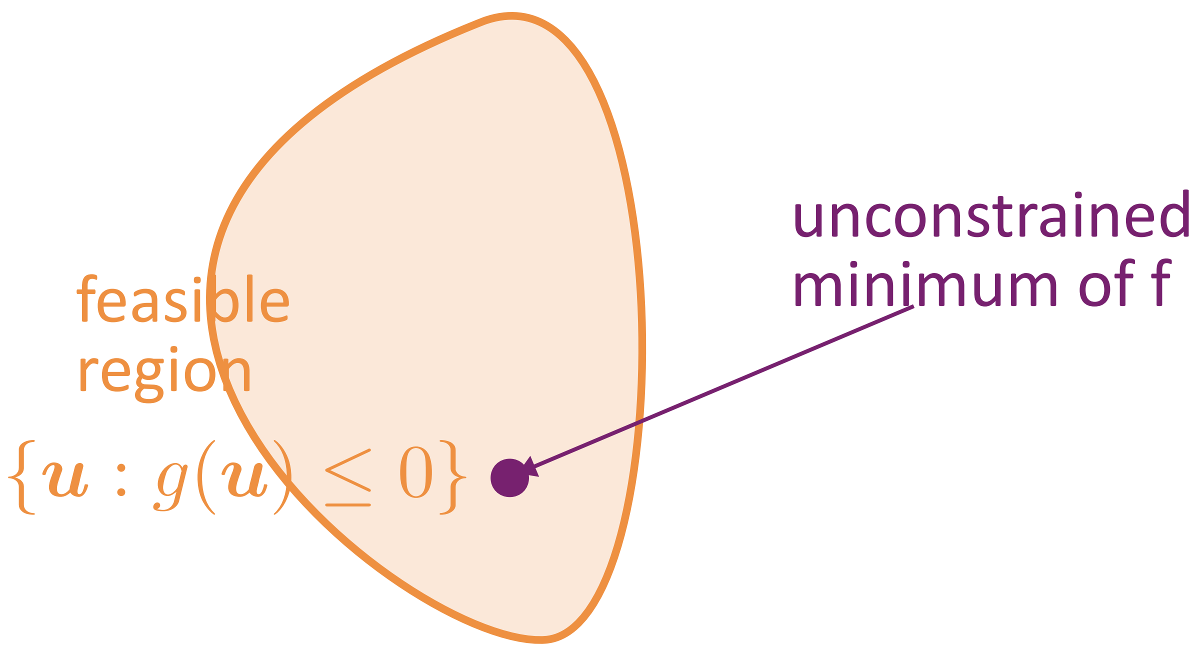
Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

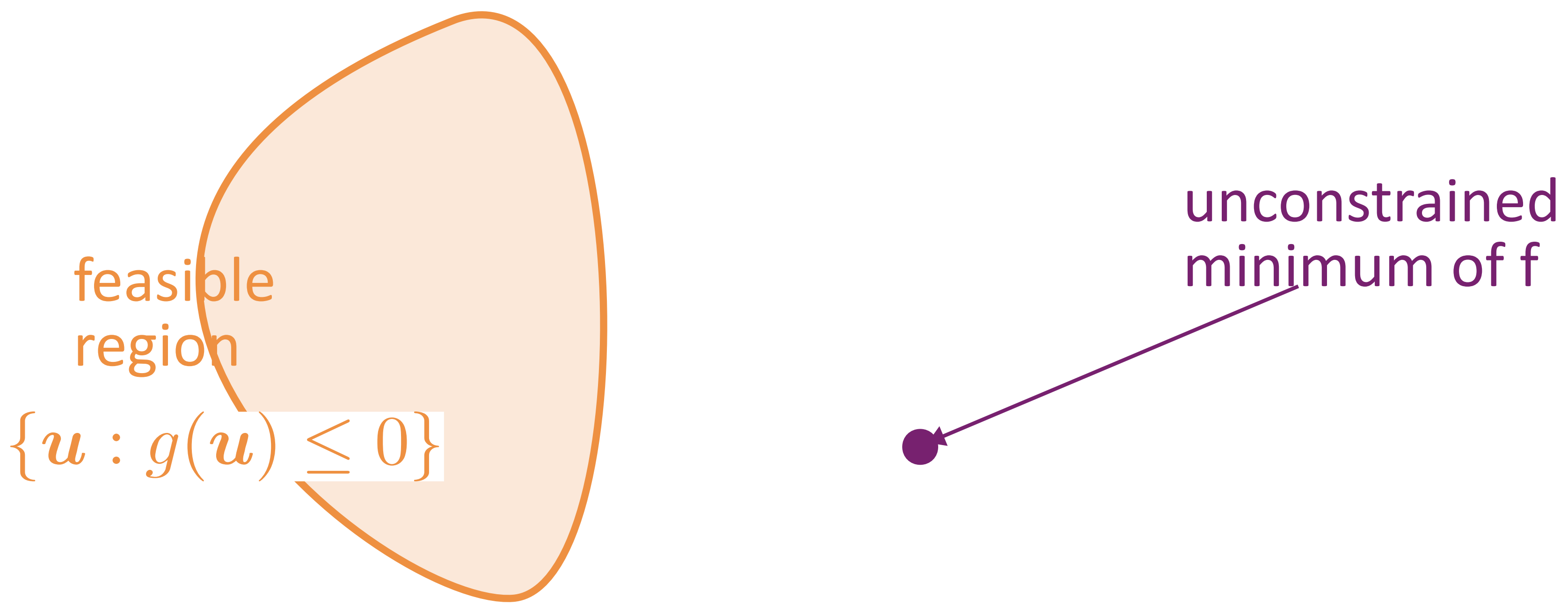
- **Case 1:** the unconstrained minimum lies in the feasible region



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \le 0$$

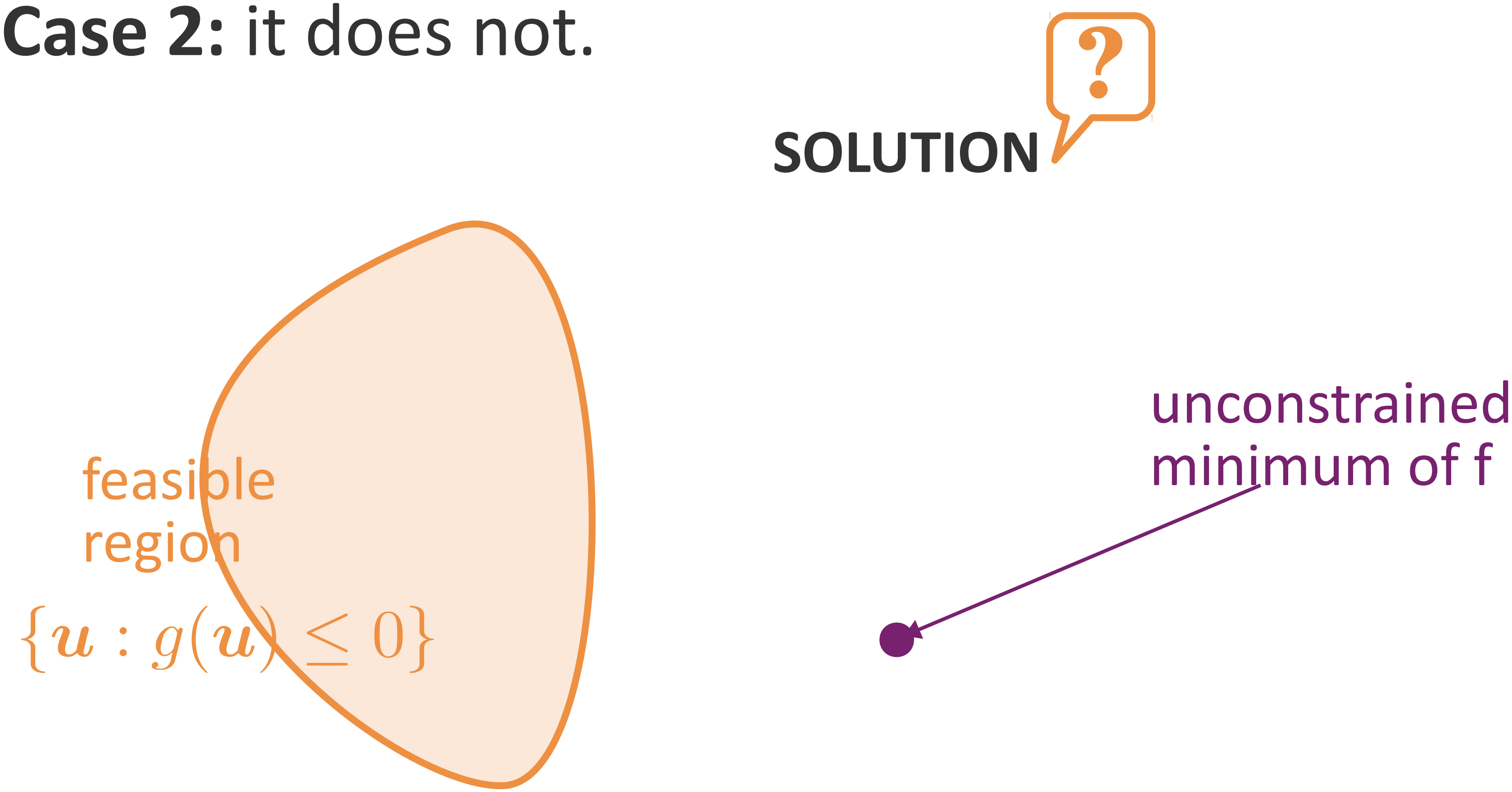
- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not.



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not.

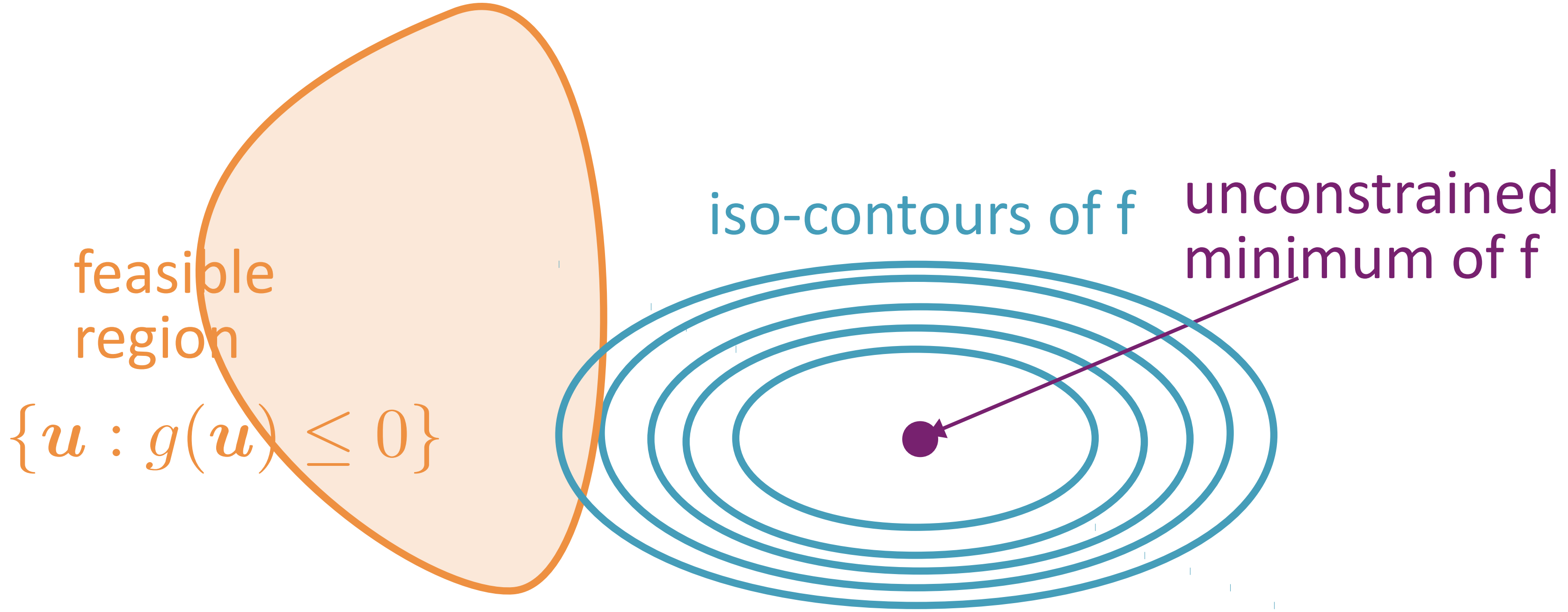


Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not.

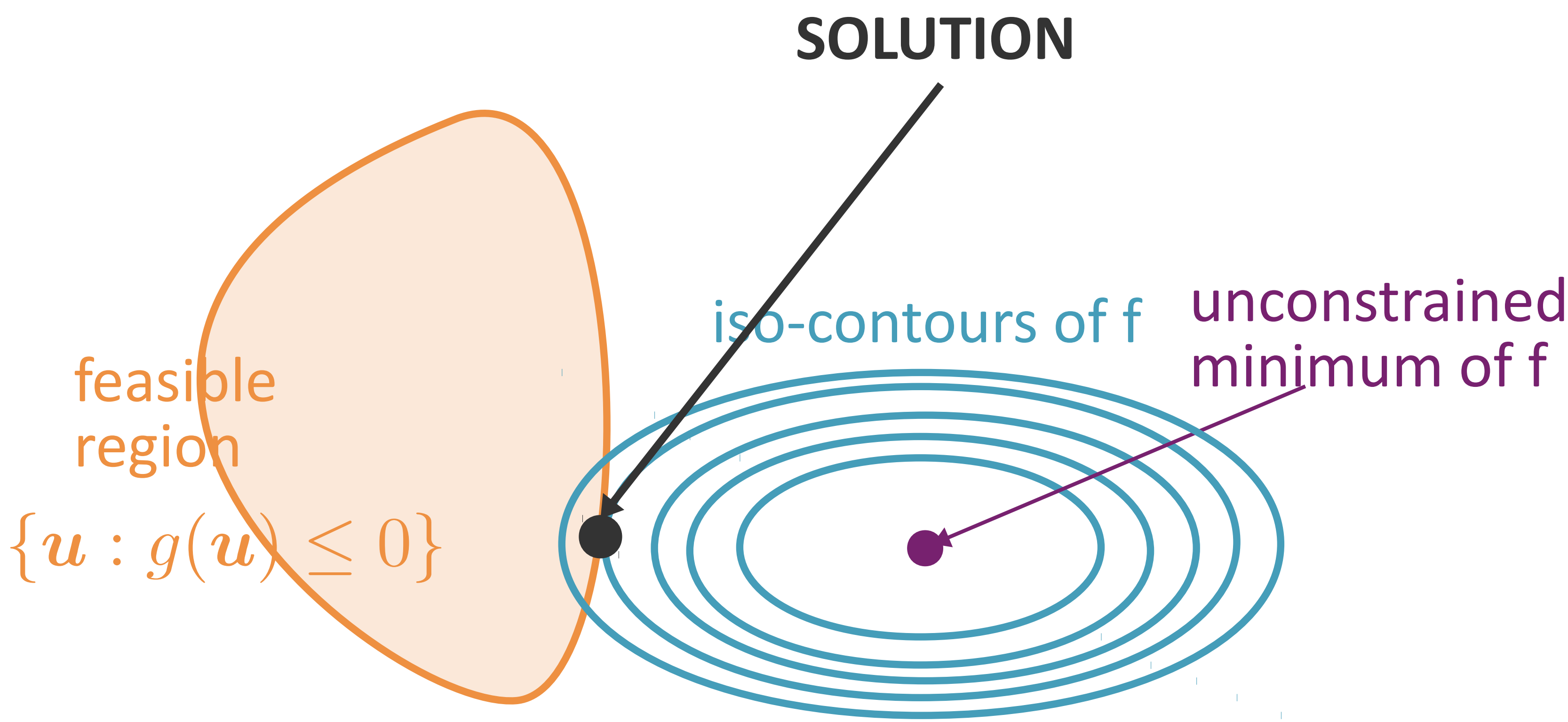
SOLUTION



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not.

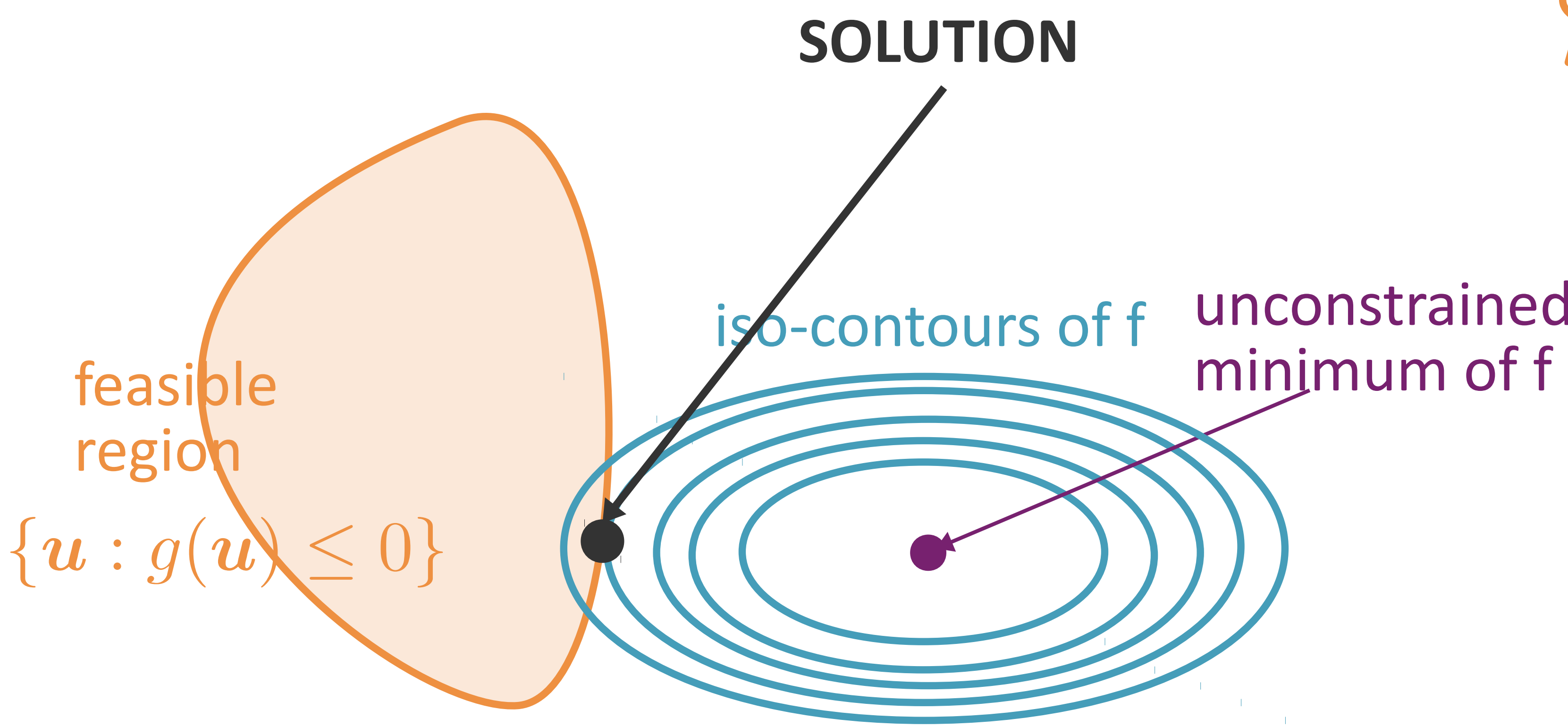


Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not.

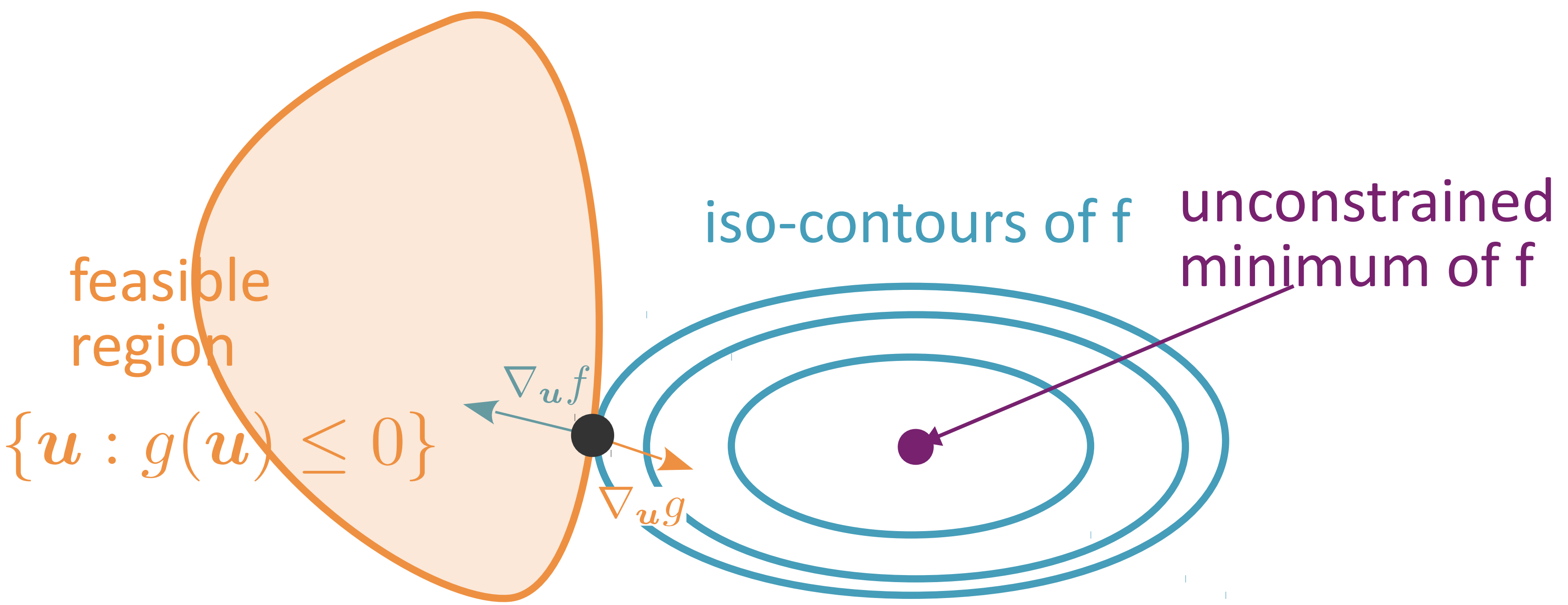
What can you say about the gradients of f and g?



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

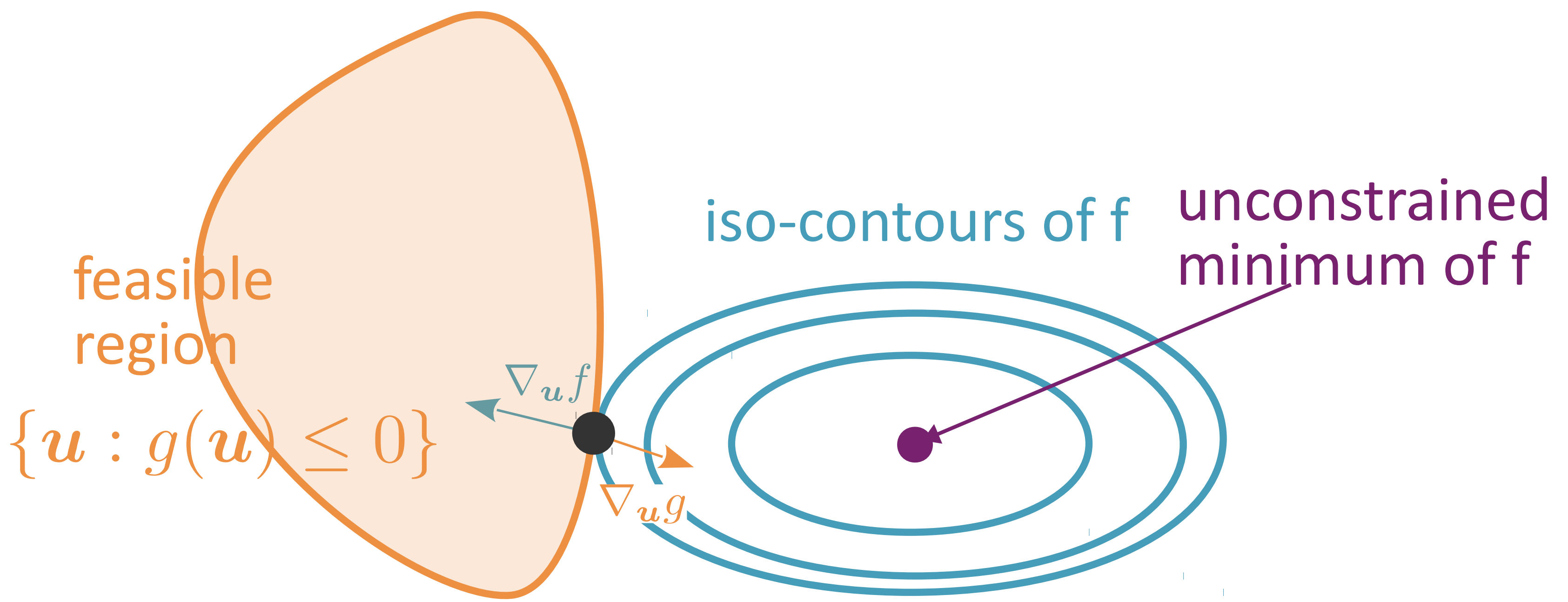
- **Case 1:** the unconstrained minimum lies in the feasible region
- **Case 2:** it does not. The solution lies where the iso-contours of f meet the feasible region.



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

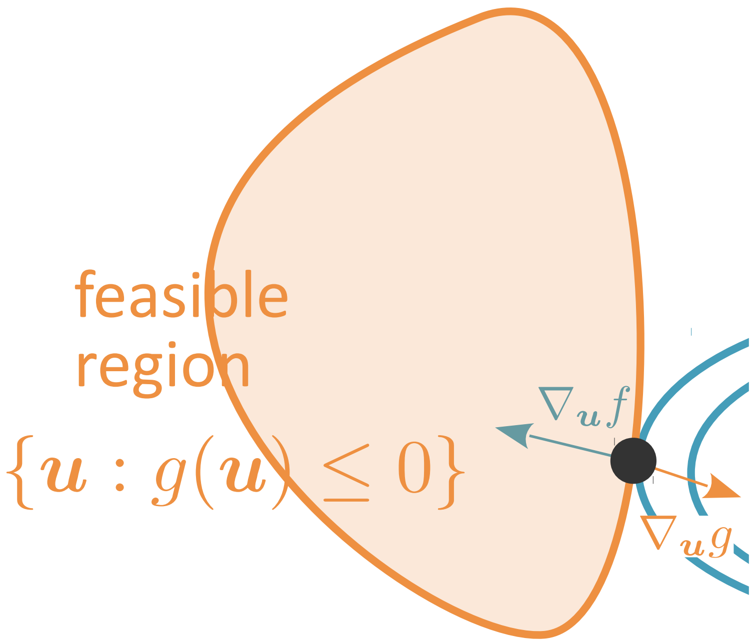
- **Case 1:** the unconstrained minimum lies in the feasible region $\nabla f(\boldsymbol{u}^*) = 0$ and $g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** it does not. The solution lies where the iso-contours of f meet the feasible region.



Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region $\nabla f(\boldsymbol{u}^*) = 0$ and $g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** it does not. The solution lies where the iso-contours of f meet the feasible region.



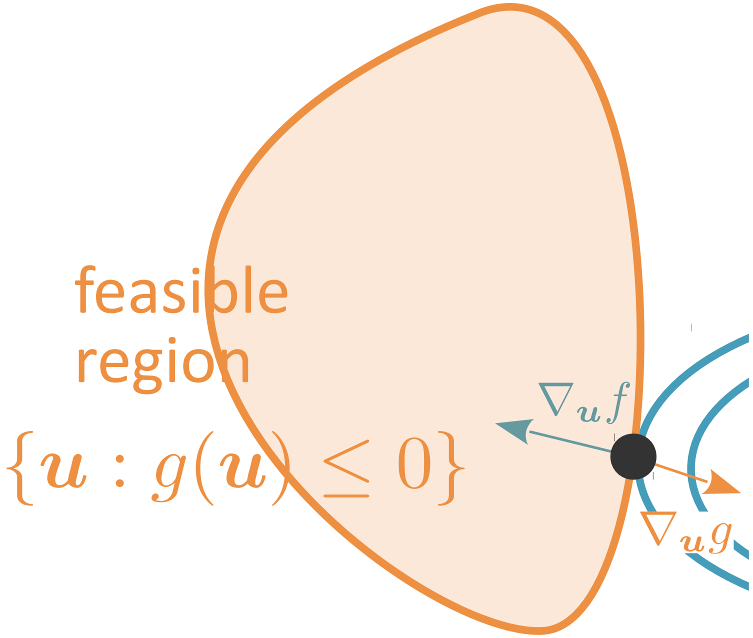
- The gradients of f and g are parallel, of opposite directions

$$\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$$

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** the unconstrained minimum lies in the feasible region $\nabla f(\boldsymbol{u}^*) = 0$ and $g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** it does not. The solution lies where the iso-contours of f meet the feasible region.



- The gradients of f and g are parallel, of opposite directions

$$\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$$

- The solution lies at the border of the feasible space

$$g(\boldsymbol{u}^*) = 0$$

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** $\nabla f(\boldsymbol{u}^*) = 0$ and $g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** $\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$

  and $g(\boldsymbol{u}^*) = 0$
- Can be summarized as: $\nabla f(\boldsymbol{u}^*) + \alpha \nabla g(\boldsymbol{u}^*) = 0$

  and $\alpha g(\boldsymbol{u}^*) = 0$
  - Either $\alpha = 0$ and $g(\boldsymbol{u}^*) \leq 0$ (case 1)
  - or $g(\boldsymbol{u}^*) = 0$ (case 2).

Geometric interpretation

$$\min_{\boldsymbol{u}\in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** $\nabla f(\boldsymbol{u}^*) = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** $\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$

  and $g(\boldsymbol{u}^*) = 0$
- Can be summarized as: $\boxed{\nabla f(\boldsymbol{u}^*) + \alpha \nabla g(\boldsymbol{u}^*) = 0}$?

  and $\alpha g(\boldsymbol{u}^*) = 0$
  - Either $\alpha = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$ (case 1)
  - or $g(\boldsymbol{u}^*) = 0$ (case 2).

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** $\nabla f(\boldsymbol{u}^*) = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** $\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$

  and $g(\boldsymbol{u}^*) = 0$

  stationarity
- Can be summarized as: $\boxed{\nabla f(\boldsymbol{u}^*) + \alpha \nabla g(\boldsymbol{u}^*) = 0}$

  and $\alpha g(\boldsymbol{u}^*) = 0$
  - Either $\alpha = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$ (case 1)
  - or $g(\boldsymbol{u}^*) = 0$ (case 2).

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** $\nabla f(\boldsymbol{u}^*) = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** $\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$

  and $g(\boldsymbol{u}^*) = 0$

stationarity

- Can be summarized as: $\boxed{\nabla f(\boldsymbol{u}^*) + \alpha \nabla g(\boldsymbol{u}^*) = 0}$

  and $\boxed{\alpha g(\boldsymbol{u}^*) = 0}$?
  - Either $\alpha = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$ (case 1)
  - or $g(\boldsymbol{u}^*) = 0$ (case 2).

Geometric interpretation

$$\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \qquad \text{subject to } g(\boldsymbol{u}) \leq 0$$

- **Case 1:** $\nabla f(\boldsymbol{u}^*) = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$
- **Case 2:** $\nabla f(\boldsymbol{u}^*) = -\alpha \nabla g(\boldsymbol{u}^*) \qquad \alpha \geq 0$

  and $g(\boldsymbol{u}^*) = 0$

  stationarity
- Can be summarized as: $\boxed{\nabla f(\boldsymbol{u}^*) + \alpha \nabla g(\boldsymbol{u}^*) = 0}$

  and $\boxed{\alpha g(\boldsymbol{u}^*) = 0}$ complementary slackness
  - Either $\alpha = 0 \text{ and } g(\boldsymbol{u}^*) \leq 0$ (case 1)
  - or $g(\boldsymbol{u}^*) = 0$ (case 2).

Quadratic Programs

- Special case of convex optimization problems where
  - f is **quadratic**

$$f : \boldsymbol{u} \mapsto \frac{1}{2}\boldsymbol{u}^\top Q \boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u} + c \quad Q \succeq 0 \quad \boldsymbol{b} \in \mathbb{R}^n \quad c \in \mathbb{R}$$

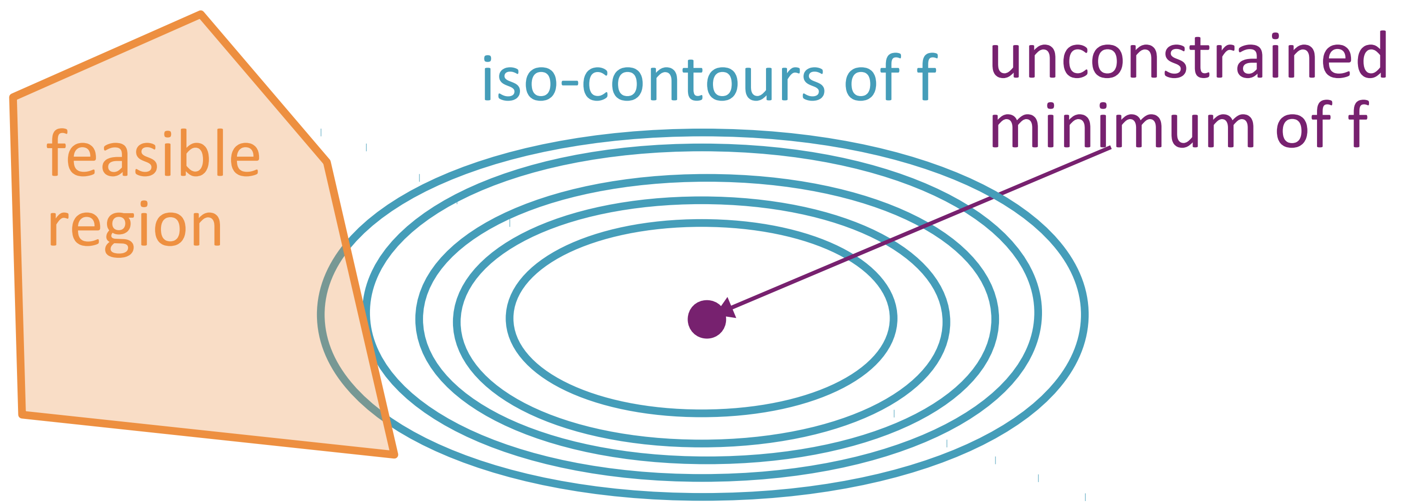
  - g_i and h_j are **affine** $\boldsymbol{u} \mapsto \boldsymbol{c}^\top \boldsymbol{u} + d \qquad \boldsymbol{c} \in \mathbb{R}^n \quad d \in \mathbb{R}$
- The feasible set is 

Quadratic Programs

- Special case of convex optimization problems where
  - f is **quadratic**

$$f : \boldsymbol{u} \mapsto \frac{1}{2}\boldsymbol{u}^\top Q\boldsymbol{u} + \boldsymbol{b}^\top \boldsymbol{u} + c \quad Q \succeq 0 \quad \boldsymbol{b} \in \mathbb{R}^n \quad c \in \mathbb{R}$$

  - g_i and h_j are **affine** $\boldsymbol{u} \mapsto \boldsymbol{c}^\top \boldsymbol{u} + d \qquad \boldsymbol{c} \in \mathbb{R}^n \quad d \in \mathbb{R}$
- The feasible set is a **polyhedron.**



Quadratic Programs

- Many methods can be used to solve QPs, for example
  - **Interior point** methods
  - **Active set** methods
- Many solvers implement them
  - CPLEX
  - CVXOPT
  - CGAL and more.

Slack variables

- Replace the inequality constraints

$$g_i(\boldsymbol{u}) \leq 0 \Leftrightarrow \exists\, s_i \geq 0 : g_i(\boldsymbol{u}) + s_i = 0$$

- s_i = **slack variable.**

$$\begin{aligned} &\min_{\boldsymbol{u} \in D} f(\boldsymbol{u}) \\ \text{subject to } & g_i(\boldsymbol{u}) + s_i = 0, i = 1, \ldots, m \\ & s_i \geq 0, i = 1, \ldots, m \\ & h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}$$

Summary

- We often try to formulate **machine learning** problems as **convex optimization** problems

$$\boxed{\begin{aligned} &\min_{\boldsymbol{u}\in D} f(\boldsymbol{u}) \\ \text{subject to } &g_i(\boldsymbol{u}) \leq 0, i = 1, \ldots, m \\ &h_j(\boldsymbol{u}) = 0, j = 1, \ldots, r \end{aligned}}$$

- If **f differentiable:** $\nabla f(\boldsymbol{u}) = 0 \Leftrightarrow \boldsymbol{u}$ minimizes f
- **Unconstrained** convex optimization problems can be solved by **gradient descent.**

  Flavors: Backtracking line search, Newton's methods, BFGS, stochastic gradient descent.

- **Constrained** convex optimization problems can be solved in **dual space** via the **Lagrangian.**

References

- *Convex optimization*. S. Boyd and L. Vandenberghe. https://web.stanford.edu/~boyd/cvxbook/
  - **Convex sets:** Chapter 2.1
  - **Convex functions:** Chapter 3.1.1 – 3.1.5, 3.2
  - **Convex optimization problems:** 4.1.1 – 4.1.2 + 4.2.2
  - **Unconstrained minimization:** 9.1.1 + 9.2 – 9.3 (gradient descent) + 9.5 (Newton)
  - **QP:** 4.4.1 + 5.1 (Lagrange) + 5.2 (Duality) + 5.3.2 (Slater) + 5.5.3 (KKT)
  - **Slack variables:** 4.1.3
  - Also see the Bibliography section at the end of each chapter.
- To go further
  - *Numerical Optimization.* J. Bonnans, J. Gilbert, C. Lemaréchal, C. Sagastizábal. **Quasi-Newton methods:** 4.3 – 4.4.
  - *Stochastic gradient descent tricks.* L. Bottou (2012). http://leon.bottou.org/publications/pdf/tricks-2012.pdf
  - *Coordinate Descent Algorithms*. S. Wright (2015). https://arxiv.org/abs/1502.04759

Homework

- By **Monday (Oct 2nd)**

  Visit `http://tinyurl.com/ma2823-2017`

  Download and read **the complete syllabus.**

  Set up your computer for the labs.

- By **Friday (Oct 6th)**

  Download, solve and turn in **HW 1.**

See you on Monday, 8:30am in Amphi sc.046!