Optimization for Machine Learning

M2 IA App., 2023-2024

Today:

Logistics
Introduction to optimization
First optimization problems

Logistics

Course webpage: https://www.lamsade.dauphine.fr/~croyer/teachOML.html

My email: clement.royer@lamsade.dauphine.fr

8 sessions:

Sep. 18	8H30 - 11H45	(Lecture)
Sep. 21	8H30 - 11H45	(Lecture + Tutorial)
Sep 25	8H30 - 11H45	(Tutorial + Lecture)
Sep 28	8H30 - 11H45	(Lecture + Tutorial)
Oct 2	8H30 - 11H45	(Tutorial + Lecture)
Oct 5	10H15 - 13H30	(Lecture + Tutorial)
Nov 6	8H30 - 11H45	(Tutorial + Lecture)
Nov 9	10H15 - 13H30	(Tutorial)

Exam: Dec. 15, 2pm - 4pm
60% of the grade

Course project 40% of the grade

Introduction to optimization

(1) About optimization

→ Tool for modeling decision making, wide range of applications

Def (informal): Optimize = Taking the best choice out of a set of alternatives

Mathematically: Find the alternative that yields the lowest/highest value of a given numerical function.

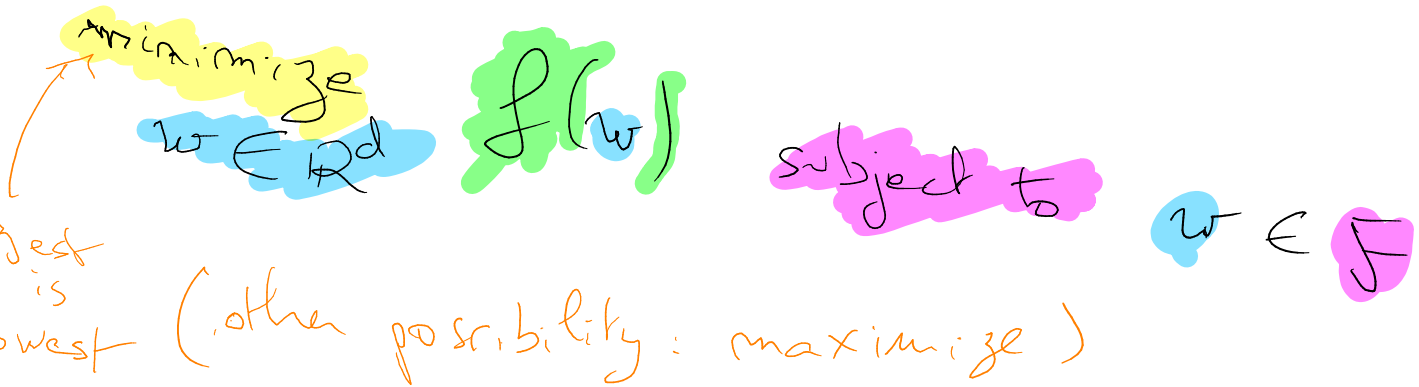
Ex) Power flow systems
Portfolio allocation
Vaccine dispatch

Data Science / ML context

Optimization problem is built using (a large amount of) data
→ Specific challenges
→ Specific algorithms

(2) Anatomy of an optimization problem

Optimization problem for this course



$f(\cdot): \mathbb{R}^d \to \mathbb{R}$ objective function
(quantifies how good an alternative is)

$w \in \mathbb{R}^d$: vector of decision variables
(define the alternative)

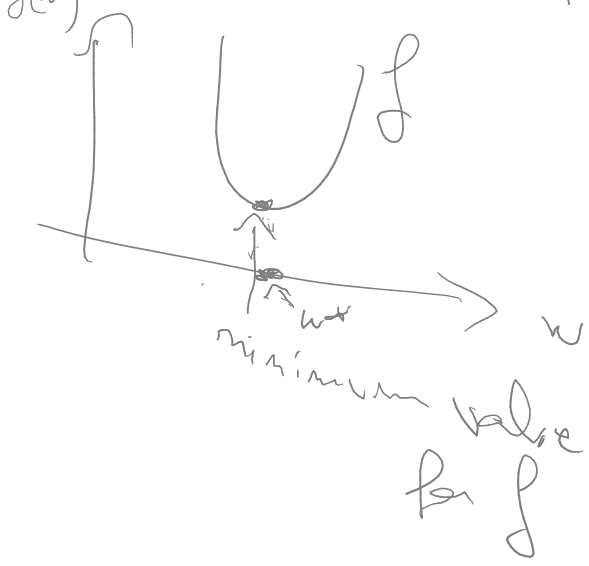
$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^d$$

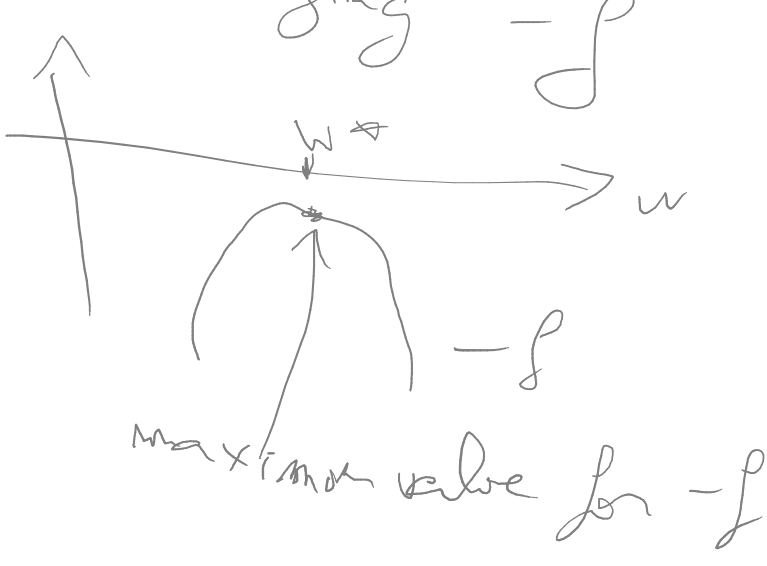
$F \subseteq \mathbb{R}^d$: feasible set
(describes conditions that the alternative must satisfy)

$F = \mathbb{R}^d \Rightarrow$ unconstrained problem

$F \subsetneq \mathbb{R}^d \Rightarrow$ constrained problem

NB: minimizing a function f is equivalent to maximizing $-f$





Solutions of an optimization problem

Focus: $(P)\begin{cases} \text{minimize} \\ w \in \mathbb{R}^d \end{cases} f(w)$

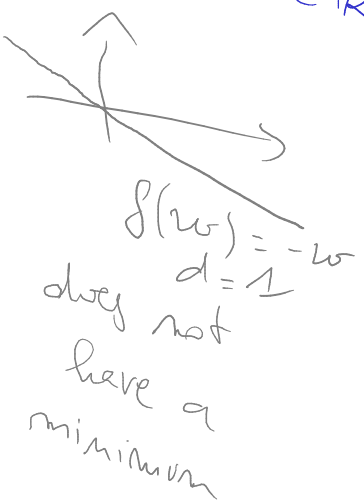
(unconstrained problem)

Def: $w^* \in \mathbb{R}^d$ is a solution of (P) if

$$f(w) \geq f(w^*) \quad \forall w \in \mathbb{R}^d$$

w^* is called a global minimum of (P) (or of f).

- $\operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$: set of solutions of (P) (subset of $\mathbb{R}^d$, can be empty)
- $\min_{w \in \mathbb{R}^d} f(w)$: optimal / minimal value for (P)

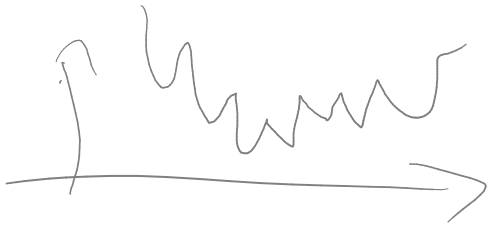


If (P) has a solution w^*, then $\min_{w \in \mathbb{R}^d} f(w) = f(w^*)$

If not, $\min_{w \in \mathbb{R}^d} f(w) = -\infty$

(unbounded problem)

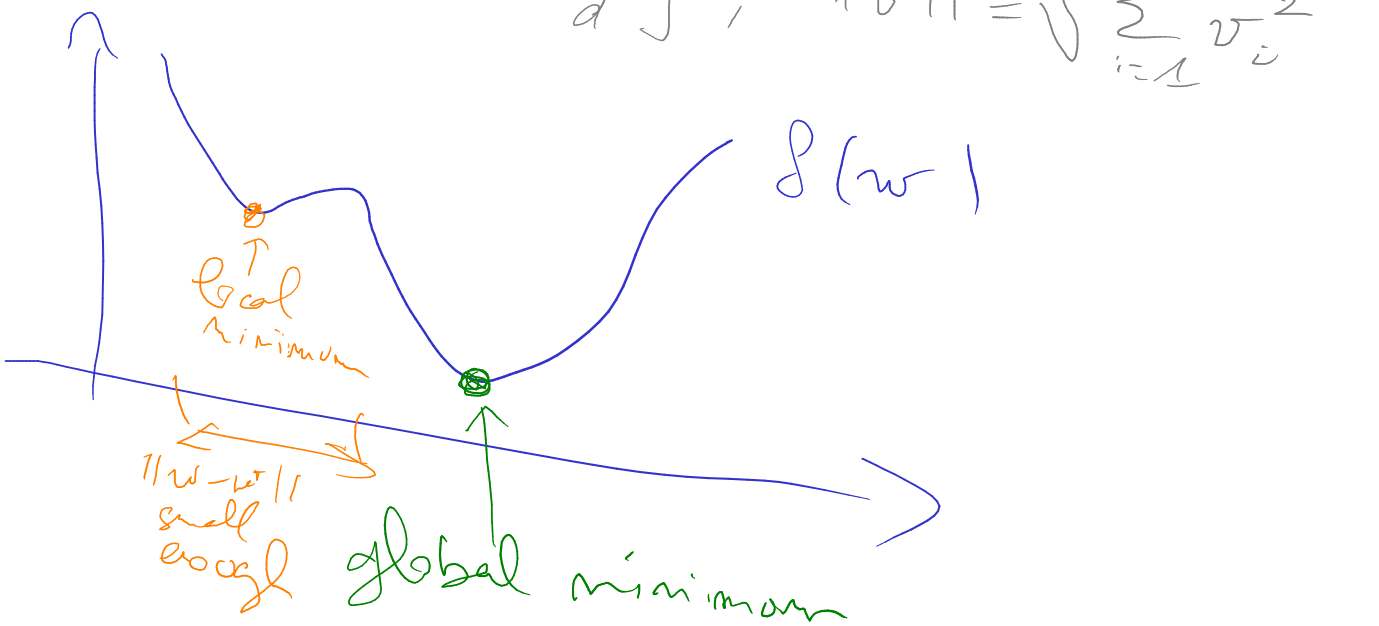
NB: Finding global minima is hard



$\Rightarrow$ Easier concept: local minima

Def: $w^* \in \mathbb{R}^d$ is a local solution of (P) (or a local minimum of f) if $\forall w \in \mathbb{R}^d$ such that $\|w - w^*\|$ is small enough, $f(w) \geq f(w^*)$

$\forall v \in \mathbb{R}^d$, $v = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix}$, $\|v\| = \sqrt{\sum_{i=1}^{d} v_i^2}$



$\hookrightarrow$ In general, finding local minima is (still!) hard

BUT There exist optimization problems for which we can find local minima numerically (we can design and implement algorithms that converge towards local minima)

(3) Optimality conditions

↳ Definition of local and global minima are hard to check in general (require to look at infinitely many points)

⇒ Approach: Using regularity properties of the objective function, replace defs. of local/global minima by conditions that we can compute in finite time. (aka optimality conditions)

Def: $f: \mathbb{R}^d \to \mathbb{R}$ is C^1 if it is continuous and $\forall w \in \mathbb{R}^d$, $\forall v \in \mathbb{R}^d$,

$$f(v) \approx f(w) + \nabla f(w)^T (v - w)$$

when $\|v - w\|$ is small

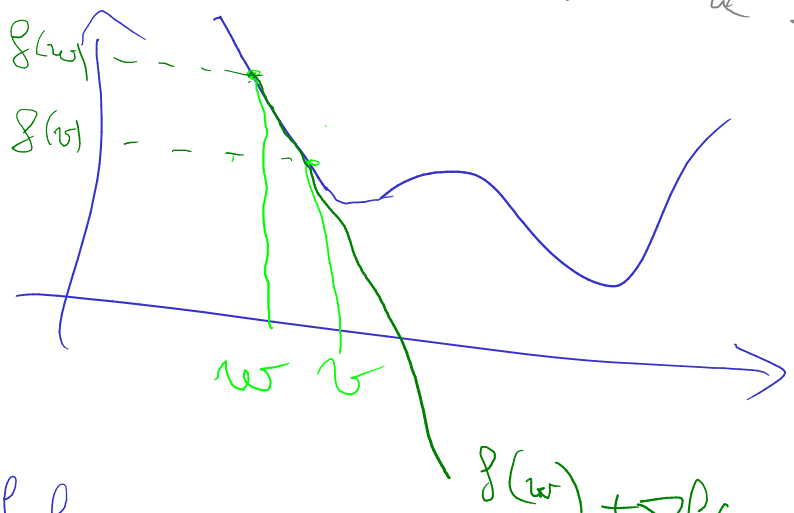
scalar product

$\nabla f(w) \in \mathbb{R}^d$ is called the gradient of f at w and it is assumed to be continuous as well.

$(\nabla f: \mathbb{R}^d \to \mathbb{R}^d)$

$\forall (a, b) \in (\mathbb{R}^d)^2$,

$$a^T b = \sum_{i=1}^{d} a_i b_i$$



$f(w) + \nabla f(w)^T (v - w)$

If f is a C^1 function, then at every point $w \in \mathbb{R}^d$ the function can be approximated locally by a linear function involving $\nabla f(w)$.

Theorem: First-order optimality conditions

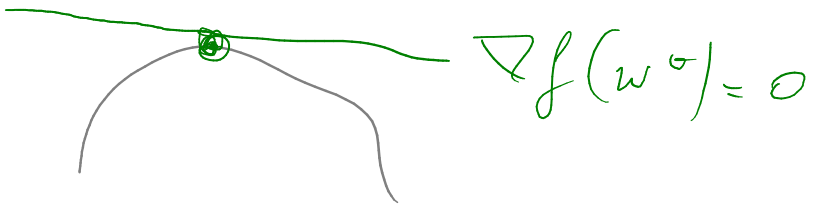
Let $f \in C^1$ and $w^* \in \mathbb{R}^d$

If w^* is a local minimum of f,

then $\|\nabla f(w^*)\| = 0$

⚠ Necessary condition (not sufficient):

$\|\nabla f(w^*)\| = 0$ does not imply that w^* is a local minimum.



⇒ We need more assumptions on f in order to define sufficient conditions for computing local minima

Def: f is C^2 if it is C^1 and $\forall w \in \mathbb{R}^d$, $\forall v \in \mathbb{R}^d$

$$f(v) \approx f(w) + \nabla f(w)^T (v - w)$$

$$+ \frac{1}{2} \underbrace{(v-w)^T}_{\in \mathbb{R}^d} \underbrace{\nabla^2 f(w)}_{\in \mathbb{R}^{d \times d}} \underbrace{(v-w)}_{\in \mathbb{R}^d}$$

where $\nabla^2 f(w) \in \mathbb{R}^{d \times d}$ (matrix with d rows and d columns) is called the Hessian matrix of f at w

Theorem: Let f be C^2 and $w^* \in \mathbb{R}^d$.

- Second-order necessary conditions

[w^* is a local minimum of f]

$\Rightarrow$ [$\|\nabla f(w^*)\| = 0$ and $\nabla^2 f(w^*) \geq 0$]

"$\nabla^2 f(w^*)$ is positive semidefinite"

$\mathbb{R}^{d \times d} \ni A \geq 0$ if $\forall v \in \mathbb{R}^d$, $v^T A v \geq 0$

- Second-order sufficient conditions

[$\|\nabla f(w^*)\| = 0$ and $\nabla^2 f(w^*) > 0$]

$\Rightarrow$ [w^* is a local minimum of f]

$A \in \mathbb{R}^{d \times d}$

$A \succ 0$ "positive definite"

$\forall v \in \mathbb{R}^d, \; v^T A v > 0$ if $v \neq 0$

- These necessary/sufficient conditions can be checked in finite time
- they involve finite representations of the changes in f ($\nabla f(w), \nabla^2 f(w)$)
- The conditions + gradients/Hessians are what we use to develop optimization algorithms, that solve optimization problems in an iterative fashion.

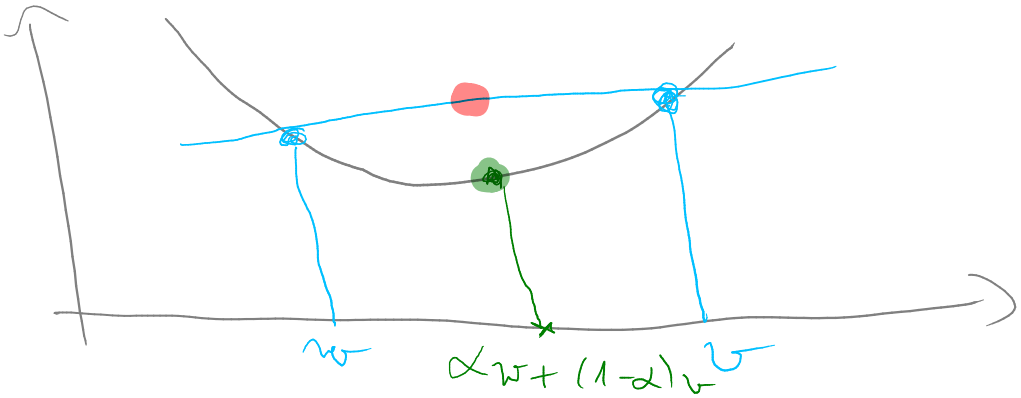
(4) Convexity

Let f function $f : \mathbb{R}^d \to \mathbb{R}$ is

Convex if

$\forall (w, v) \in (\mathbb{R}^d)^2, \; \forall \alpha \in [0, 1]$

$$f(\alpha w + (1-\alpha) v) \leq \alpha f(w) + (1-\alpha) f(v),$$



Example of convex functions:

Linear functions: $f(w) = a^T w + b$ with $a \in \mathbb{R}^d$, $b \in \mathbb{R}$.

Quadratic function:

$$f(w) = c + b^T w + \frac{1}{2} w^T A w$$

with $A \succeq 0$

Properties of convex functions

Let f be convex. Then

(i) Every local minimum of f is a global minimum.

(ii) If f is C^1, then

[$w^* \in \mathbb{R}^d$ is a global minimum]

$\Leftrightarrow$ [$\|\nabla f(w^*)\| = 0$]

(iii) If f is C^2, then

$\nabla^2 f(w) \geq 0 \quad \forall w \in \mathbb{R}^d$

and the converse is true.

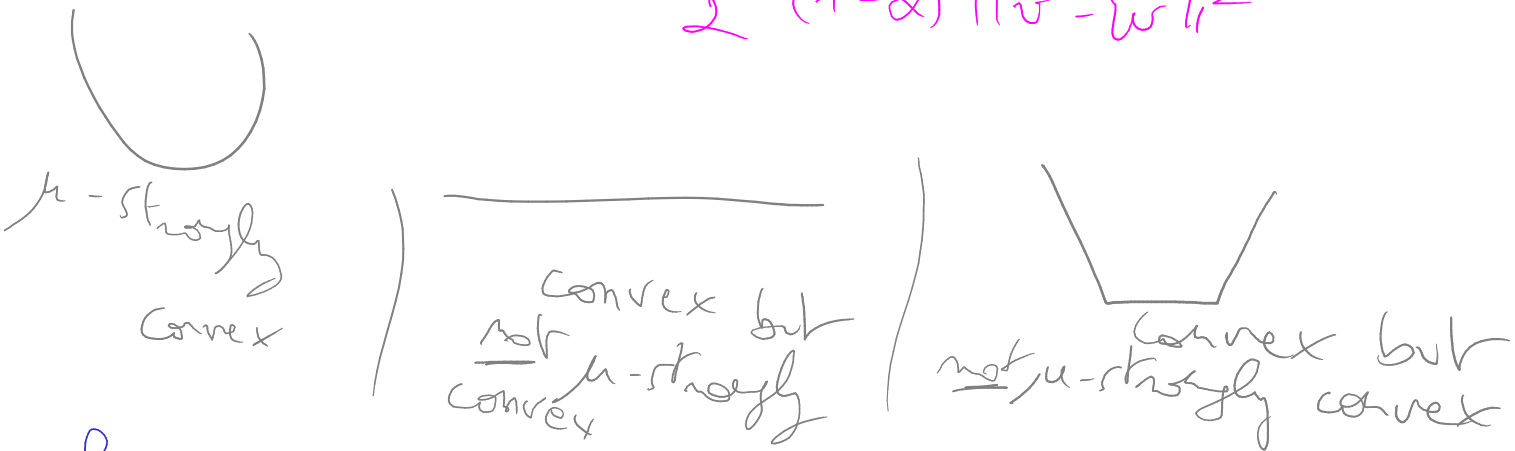
(f C^2 and $\nabla^2 f(w) \geq 0 \ \forall w \Rightarrow f$ is convex)

A special class of convex functions

Def: $f: \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex ($\mu > 0$) if

$\forall (w, v) \in (\mathbb{R}^d)^2$, $\forall \alpha \in [0, 1]$,

$$f(\alpha w + (1-\alpha)v) \leq \alpha f(w) + (1-\alpha) f(v) - \frac{\mu}{2}\alpha(1-\alpha)\|v - w\|^2$$



Th A μ-strongly convex function has a unique global minimum.

Other properties of μ-strongly convex functions:

- If $f \in C^1$ and μ-strongly convex, then there exists a unique $w^* \in \mathbb{R}^d$ such that $\|\nabla f(w^*)\| = 0$
- For $f \in C^2$,

[f μ-strongly convex]

$\Leftrightarrow$ [$\forall w \in \mathbb{R}^d$, $\nabla^2 f(w) \geq \mu I_d$]

$$I_d = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}$$

$$\nabla^2 f(w) - \mu I_d \geq 0$$

Conclusion: Convex and strongly convex functions are great for optimization!

$\Rightarrow$ Can compute global minima by finding points with zero gradients

$\Rightarrow$ Even get uniqueness of the global minimum in the strongly convex case

(5) Examples of convex optimization problems

Motivation: Linear models in data analysis

Data: $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times d}$

$\begin{pmatrix} (x_i)_1 \\ \vdots \\ (x_i)_d \end{pmatrix} = x_i \in \mathbb{R}^d$

$x_i^T = \begin{bmatrix} (x_i)_1 & \cdots & (x_i)_d \end{bmatrix}$ "matrix of feature vectors"

$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$ "vector of labels"

Goal: Find a linear model that explains the data.

Mathematically, we want $w \in \mathbb{R}^d$ such that

$$Xw = y$$

$$\Leftrightarrow \quad x_i^T w = y_i$$

⚠ $Xw = y$ does not always have a solution

⇒ In that case, we seek a w such that Xw is the closest possible to y

Can formulate that as an optimization problem!

Most classical way: Linear least squares

Given (X, y), solve

(LS) $\underset{w \in \mathbb{R}^d}{\text{minimize}} \; f(w) := \frac{1}{2} \|Xw - y\|^2$

$\|\cdot\|$: quantifies how far $Xw - y$ is from 0 (ideal goal)

$\|\cdot\|_2$: makes it C^2

$\frac{1}{2}$: makes it convenient for computing the gradient

Properties of (LS)

- The objective f is convex and C^2
- $\nabla f(w) = X^T(Xw - y)$ $\forall w \in \mathbb{R}^d$

and can compute a solution to $\|\nabla f(w)\| = 0$ explicitly (in closed form)

always has a solution! if $X \neq 0$

- If $Xw = y$ has a solution, then $\operatorname{argmin}_{w \in \mathbb{R}^d} f(w)$ is the set of solutions of $Xw = y$

and $\min_{w \in \mathbb{R}^d} f(w) = 0$