Advance Convex Optimization

Gourishanker

Contents

1 Introduction

Definition 1.1 (Optimization Problem and Optimal Value)
Consider the optimization problem

$$\min_{x \in X} f(x),$$

where $f : X \to \mathbb{R}$.
The associated *optimal value* is defined as

$$p^\star := \inf_{x \in X} f(x).$$

The infimum is used instead of the minimum since, in general, the function f may fail to attain its minimum over the set X. That is, there may exist no point $x^\star \in X$ such that

$$f(x^\star) = \min_{x \in X} f(x).$$

Nevertheless, the set $\{f(x) : x \in X\} \subset \mathbb{R}$ always admits a greatest lower bound in $\mathbb{R} \cup \{-\infty\}$, which is precisely the infimum.
If there exists a point $x^\star \in X$ satisfying

$$f(x^\star) = \inf_{x \in X} f(x),$$

then the infimum is said to be *attained*, and we may equivalently write

$$p^\star = \min_{x \in X} f(x).$$

In this case, $x^\star$ is called an *optimal solution*.

Definition 1.2 (Constrained optimization problem)
Consider the constrained optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & g(x) = 0, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$.
The associated *feasible set* is defined by

$$X := \{x \in \mathbb{R}^n : g(x) = 0\}.$$

Definition 1.3 (Local minimum (constrained))
A point $x^\star \in X$ is called a *local minimum* of the constrained optimization problem if there exists $\delta > 0$ such that

$$f(x^\star) \leq f(x) \quad \text{for all } x \in X \cap B(x^\star, \delta),$$

where

$$B(x^\star, \delta) := \{x \in \mathbb{R}^n : \|x - x^\star\| < \delta\}$$

denotes the open ball of radius δ centered at $x^\star$.
Equivalently, $x^\star$ is a local minimum if no feasible point sufficiently close to $x^\star$ achieves a strictly smaller objective value.

	Smooth	**Non-smooth**
Convex	Convex Optimization ☺	Advanced Convex Optimization (Proximal methods, subgradients)
Non-convex	Deep Learning	:(

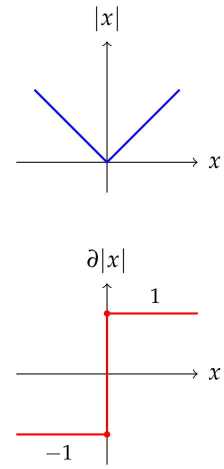

Figure 1: The function $|x|$ and its subdifferential.

Example (Soft-thresholding via subdifferential calculus)
Consider the optimization problem

$$\min_{x\in\mathbb{R}} \left\{ \frac{1}{2}(x-a)^2 + \tau|x| \right\}, \qquad \tau > 0.$$

The objective function is convex but non-smooth due to the absolute value term. The subdifferential of $|x|$ is the set-valued mapping

$$\partial|x| = \begin{cases} \{1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{-1\}, & x < 0. \end{cases}$$

Define

$$\phi(x) := \frac{1}{2}(x-a)^2 + \tau|x|.$$

Since ϕ is convex, a point $x^\star \in \mathbb{R}$ is optimal if and only if

$$0 \in \partial\phi(x^\star) = (x^\star - a) + \tau\partial|x^\star|.$$

Equivalently,

$$a - x^\star \in \tau\partial|x^\star|.$$

A case-by-case analysis yields

$$x^\star = \begin{cases} a - \tau, & a > \tau, \\ 0, & |a| \le \tau, \\ a + \tau, & a < -\tau. \end{cases}$$

This solution can be written compactly as the *soft-thresholding operator*

$$x^\star = \operatorname{sign}(a)\, \max\{|a| - \tau, 0\}.$$

2 Inclusion Problems, Convexity, and Smoothness

Definition 2.1 (Inclusion problem)
Let X be a real Euclidean space. An *inclusion problem* consists in finding

$$x \in X \quad \text{such that} \quad 0 \in T(x),$$

where $T : X \rightrightarrows X$ is a set-valued operator.

Remark
Many convex optimization problems can be written as inclusion problems by taking $T = \partial f$, where ∂f denotes the subdifferential of a convex function.

Definition 2.2 (Convex optimization problem)
Let $C \subset X$ be a convex set and let $f : X \to \mathbb{R}$ be a convex function. The problem

$$\min_{x \in C} f(x)$$

is called a *convex optimization problem*.

Proposition 2.3
If both the objective function f and the constraint set C are convex, then every local minimizer is a global minimizer.

Proposition 2.4
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. If

$$\nabla^2 f(x) \succeq 0 \quad \text{for all } x \in \mathbb{R}^n,$$

then f is convex. If

$$\nabla^2 f(x) \succ 0 \quad \text{for all } x \in \mathbb{R}^n,$$

then f is strictly convex.

2.1 Euclidean spaces and induced norms

Definition 2.5 (Euclidean space)
A real Euclidean space X is a real vector space equipped with an inner product

$$\langle x, y \rangle.$$

The induced norm is defined by

$$\|x\| := \sqrt{\langle x, x \rangle}.$$

Remark
The ℓ^2-norm on $\mathbb{R}^n$ is induced by the standard inner product. The ℓ^1-norm is *not* induced by any inner product.

Definition 2.6
The norm induces a metric

$$d(x,y) := \|x-y\|,$$

which defines a topology on X.

Definition 2.7 (Convergence)
A sequence $(x_k) \subset X$ converges to $x \in X$ if

$$\|x_k - x\| \to 0.$$

Definition 2.8 (Continuity)
A function $f : X \to \mathbb{R}$ is continuous at $x \in X$ if

$$x_k \to x \quad \Rightarrow \quad f(x_k) \to f(x).$$

Definition 2.9 (Topological space)
Let X be a set and let τ be a collection of subsets of X. The pair (X,τ) is called a *topological space* if

- $\varnothing \in \tau$ and $X \in \tau$,
- the union of any collection of sets in τ belongs to τ,
- the intersection of any finite collection of sets in τ belongs to τ.

The elements of τ are called *open sets*.

Definition 2.10 (Convergence in a topological space)
Let (X,τ) be a topological space and let $(x_k) \subset X$. We say that x_k converges to $x \in X$ if for every open set $U \in \tau$ such that $x \in U$, there exists $N \in \mathbb{N}$ for which

$$x_k \in U \quad \text{for all } k \geq N.$$

Remark
This definition of convergence depends only on the topology τ and does not require any notion of distance or norm. In particular, convergence can be defined in spaces that are not metric spaces.

Proposition 2.11
Every convex function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous.

Proposition 2.12

For $0 < p < 1$, the set

$$B_p := \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$$

is not convex.

Definition 2.13 (Differentiability)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be differentiable at $x \in \mathbb{R}^n$ if there exists $z \in \mathbb{R}^n$ such that

$$f(x+h) = f(x) + \langle z, h \rangle + o(\|h\|) \quad \text{as } h \to 0.$$

In this case, $z = \nabla f(x)$. where The term $o(\|h\|)$ satisfies

$$\lim_{\|h\| \to 0} \frac{o(\|h\|)}{\|h\|} = 0.$$

Exercise (Differentiability and inner products)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $x \in \mathbb{R}^n$ in the sense that there exists $z \in \mathbb{R}^n$ such that

$$f(x+h) = f(x) + \langle z, h \rangle + o(\|h\|) \quad \text{as } h \to 0,$$

where

$$\lim_{\|h\| \to 0} \frac{o(\|h\|)}{\|h\|} = 0.$$

1. Show that for a fixed inner product $\langle \cdot, \cdot \rangle$, the vector z is unique.

2. Show that the vector z depends on the choice of inner product.

3. Show that the notion of differentiability does not depend on the choice of inner product on $\mathbb{R}^n$.

4. Show that all norms on $\mathbb{R}^n$ are equivalent.

5. Prove the Riesz representation theorem: every linear functional on a Euclidean space can be represented as an inner product.

Solution.

1. *Uniqueness:* Assume that there exist $z_1, z_2 \in \mathbb{R}^n$ such that

   $$f(x+h) = f(x) + \langle z_1, h \rangle + o(\|h\|)$$

   and

   $$f(x+h) = f(x) + \langle z_2, h \rangle + o(\|h\|).$$

   Subtracting the two expressions gives

   $$\langle z_1 - z_2, h \rangle = o(\|h\|).$$

   Dividing by $\|h\|$ and letting $h \to 0$,

   $$\frac{\langle z_1 - z_2, h \rangle}{\|h\|} \to 0.$$

Choosing $h = z_1 - z_2$ yields

$$\|z_1 - z_2\|^2 = 0,$$

hence $z_1 = z_2$.

2. *Dependence on inner product:* Let $f(x) = a^\mathsf{T} x$, where $a \in \mathbb{R}^n$. For the standard inner product,

$$f(x+h) = f(x) + \langle a, h \rangle,$$

so $z = a$.

Now consider a different inner product

$$\langle x, y \rangle_M := \langle Mx, y \rangle,$$

where M is symmetric positive definite. Then

$$f(x+h) = f(x) + \langle M^{-1}a, h \rangle_M.$$

Hence the representing vector z changes with the inner product.

3. *Independence of differentiability:* Differentiability is equivalent to the existence of a linear map $L : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x+h) = f(x) + L(h) + o(\|h\|).$$

This definition does not involve any inner product.

By the Riesz representation theorem, for any chosen inner product, there exists a unique vector z such that

$$L(h) = \langle z, h \rangle.$$

Thus, while the vector z depends on the inner product, the existence of the derivative does not.

4. *Norm equivalence:* Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms on $\mathbb{R}^n$. To show they are equivalent, we need to find $c, C > 0$ such that $c\|x\|_a \leq \|x\|_b \leq C\|x\|_a$ for all x.

The unit sphere $S = \{x : \|x\|_a = 1\}$ is compact. The function $x \mapsto \|x\|_b$ is continuous on S, so it attains a minimum $c > 0$ and a maximum $C > 0$. By homogeneity of norms, the result follows for all x.

5. *Riesz representation theorem:* Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of X. Define

$$z := \sum_{i=1}^{n} L(e_i)\, e_i.$$

Then for any $h = \sum_i h_i e_i$,

$$\langle z, h \rangle = \sum_{i=1}^{n} L(e_i) h_i = L(h).$$

Uniqueness follows from non-degeneracy of the inner product.

$\square$

Theorem 2.14
Let $X \subset \mathbb{R}^n$ be a convex set and let $f : X \to \mathbb{R}$ be convex and differentiable. If $x^\star \in X$ satisfies

$$\nabla f(x^\star) = 0,$$

then $x^\star$ is a global minimizer of f on X.

Proof. Since f is convex, for any $x, y \in X$ and any $t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Fix $x \in X$ and define the function

$$\varphi(t) := f(x^\star + t(x - x^\star)), \qquad t \in [0, 1].$$

By convexity of f, the function φ is convex on $[0, 1]$.

Since f is differentiable, φ is differentiable and

$$\varphi'(0) = \langle \nabla f(x^\star), x - x^\star \rangle.$$

By assumption, $\nabla f(x^\star) = 0$, hence

$$\varphi'(0) = 0.$$

For a convex function on an interval, the derivative is monotone nondecreasing. Therefore, for all $t \in [0, 1]$,

$$\varphi(t) \geq \varphi(0).$$

In particular, at $t = 1$,

$$f(x) = \varphi(1) \geq \varphi(0) = f(x^\star).$$

Since $x \in X$ was arbitrary, $x^\star$ is a global minimizer of f. □

2.2 Unconstrained and Constrained Optimization

An *unconstrained optimization problem* is of the form

$$\min_{x \in X} f(x),$$

where typically $X = \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$.

In this case, every point in X is feasible, and optimality is characterized by stationarity conditions such as

$$\nabla f(x^\star) = 0.$$

A *constrained optimization problem* is given by

$$\min_{x \in C} f(x),$$

where $C \subset \mathbb{R}^n$ is a constraint set and $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 2.15 (Feasible set)
The set
$$C := \{x \in \mathbb{R}^n : \text{all constraints are satisfied}\}$$
is called the *feasible set*. Points in C are called *feasible points*.

Definition 2.16 (Indicator function)
Let $C \subset \mathbb{R}^n$. The indicator function $\delta_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined as
$$\delta_C(x) := \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$

Then the constrained problem
$$\min_{x \in C} f(x)$$
is equivalent to the unconstrained problem
$$\min_{x \in \mathbb{R}^n} \big(f(x) + \delta_C(x)\big).$$

Definition 2.17 (Lagrangian)
Consider the constrained problem
$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & g(x) \le 0, \end{aligned}$$
where $g : \mathbb{R}^n \to \mathbb{R}^m$. The *Lagrangian* associated with the problem is
$$\mathcal{L}(x, \lambda) := f(x) + \langle \lambda, g(x) \rangle, \qquad \lambda \in \mathbb{R}^m_+.$$

Proposition 2.18 (Optimality and saddle-point property)
If $x^\star$ is a solution of the constrained problem and $\lambda^\star \ge 0$ is a corresponding Lagrange multiplier, then
$$\mathcal{L}(x^\star, \lambda) \le \mathcal{L}(x^\star, \lambda^\star) \le \mathcal{L}(x, \lambda^\star)$$
for all feasible x and all $\lambda \ge 0$.

Thus, optimal solutions correspond to saddle points of the Lagrangian.

Remark (Constraint qualification)
Interior-point conditions (such as Slater's condition) ensure that Lagrange multipliers exist and that optimality conditions are valid.

Example
Consider the problem
$$\min_{x \in \mathbb{R}} x \quad \text{subject to} \quad x \ge 0.$$

The optimal solution is $x^\star = 0$, which lies on the boundary of the feasible set.
If we perturb the constraint slightly, the gradient of the objective cannot be balanced by any Lagrange multiplier unless the feasible set has nonempty interior. This failure prevents the derivation of meaningful optimality conditions.

Remark
Interior feasibility guarantees that constraints are not degenerate and ensures the validity of duality and KKT conditions. Without interior points, multipliers may fail to exist or be unbounded.

3 Lecture 3

Definition 3.1
Let $\mathbb{X}$ be a normed vector space.

- $\Gamma(\mathbb{X}) := \{f : \mathbb{X} \to \mathbb{R} \mid f \text{ is convex}\}$.
- $C^0(\mathbb{X})$ denotes the space of continuous functions on $\mathbb{X}$.
- $C^1(\mathbb{X})$ denotes the space of continuously differentiable functions on $\mathbb{X}$.
- A function $f : \mathbb{X} \to \mathbb{R}$ is called β-Lipschitz if
$$|f(\mathbf{x}) - f(\mathbf{y})| \leq \beta\|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{X}.$$
- A function $f : \mathbb{X} \to \mathbb{R}$ is called β-smooth if $f \in C^1(\mathbb{X})$ and
$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta\|\mathbf{x} - \mathbf{y}\| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{X}.$$
-
$$C^1_\beta(\mathbb{X}) := \{f : \mathbb{X} \to \mathbb{R} \mid f \in C^1(\mathbb{X}) \text{ and } \nabla f \text{ is } \beta\text{-Lipschitz}\}.$$

Examples

- Let $\mathbb{X} = \mathbb{R}^2$. The function
$$f(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2$$
is 1-smooth.
- Let $Q \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite. The quadratic function
$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^\top Q\mathbf{x}$$
is $\lambda_{\max}(Q)$-smooth.
- The function
$$f(\mathbf{x}) = \|\mathbf{x}\|_2^2$$
is 4-smooth on the unit ball
$$\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq 1\}.$$

Theorem 3.2
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then f is β-Lipschitz (with respect to the Euclidean norm) if and only if
$$\|\nabla f(\mathbf{x})\|_2 \leq \beta \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

Proof. ($\Rightarrow$) Assume f is β-Lipschitz. Fix $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{v} \in \mathbb{R}^n$ be any unit vector. For $t \in \mathbb{R}$,
$$|f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})| \leq \beta|t|.$$
Dividing by $|t|$ and letting $t \to 0$, we obtain
$$|\nabla f(\mathbf{x})^\top \mathbf{v}| \leq \beta.$$

Taking the supremum over all unit vectors $\mathbf{v}$ yields

$$\|\nabla f(\mathbf{x})\|_2 \le \beta.$$

($\Leftarrow$) Assume $\|\nabla f(\mathbf{x})\|_2 \le \beta$ for all $\mathbf{x} \in \mathbb{R}^n$. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define

$$\gamma(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x}), \quad t \in [0, 1].$$

By the fundamental theorem of calculus,

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\gamma(t))^\top (\mathbf{y} - \mathbf{x})\, dt.$$

Applying the Cauchy–Schwarz inequality,

$$|f(\mathbf{y}) - f(\mathbf{x})| \le \int_0^1 \|\nabla f(\gamma(t))\|_2 \, \|\mathbf{y} - \mathbf{x}\|_2 \, dt \le \beta \|\mathbf{y} - \mathbf{x}\|_2.$$

Thus, f is β-Lipschitz. □

Exercise (First-order characterizations of convexity)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. The following statements are equivalent:

1. f is convex.
2. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,
$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$
3. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,
$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge 0.$$

Proof. We prove the equivalence by showing $(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)$.

$(1) \Rightarrow (2)$. Assume that f is convex. Fix $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and define

$$\phi(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \quad t \in [0, 1].$$

Since f is convex and differentiable, ϕ is convex and differentiable. By convexity of ϕ,

$$\phi(1) \ge \phi(0) + \phi'(0).$$

Using the chain rule,

$$\phi'(0) = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

Substituting $\phi(1) = f(\mathbf{y})$ and $\phi(0) = f(\mathbf{x})$ yields

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

$(2) \Rightarrow (3)$. Assume that (2) holds. Applying (2) with $(\mathbf{x}, \mathbf{y})$ gives

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

Applying (2) with $(\mathbf{y},\mathbf{x})$ gives

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.$$

Adding the two inequalities and simplifying yields

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq 0.$$

$(3) \Rightarrow (1)$. Assume that (3) holds. Fix $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ and define

$$\gamma(t) := (1-t)\mathbf{x} + t\mathbf{y}, \quad t \in [0,1],$$

and

$$\phi(t) := f(\gamma(t)).$$

By the chain rule,

$$\phi'(t) = \langle \nabla f(\gamma(t)), \mathbf{y}-\mathbf{x}\rangle.$$

Applying (3) with $\mathbf{x} = \gamma(t)$ and $\mathbf{y} = \mathbf{x}$ gives

$$\langle \nabla f(\gamma(t)) - \nabla f(\mathbf{x}), \gamma(t)-\mathbf{x}\rangle \geq 0.$$

Since $\gamma(t) - \mathbf{x} = t(\mathbf{y}-\mathbf{x})$, this implies

$$\phi'(t) \geq \phi'(0).$$

Thus ϕ' is nondecreasing on $[0,1]$, and hence ϕ is convex. Therefore,

$$f(\gamma(t)) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}), \quad \text{for all } t \in [0,1].$$

This is precisely the definition of convexity of f. $\square$

Remark
When $n=1$, condition (3) reduces to

$$(f'(x) - f'(y))(x-y) \geq 0,$$

which is equivalent to monotonicity of f'. Hence, a differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if its derivative is nondecreasing.

Lemma 3.3 (Smoothness Upper Bound or Descent Lemma)
Let $f \in C^1_\beta(\mathbb{X})$, where $\mathbb{X} \subset \mathbb{R}^n$. Then, for all $\mathbf{x},\mathbf{y} \in \mathbb{X}$,

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

Proof. Fix $\mathbf{x},\mathbf{y} \in \mathbb{X}$ and define

$$\gamma(t) := \mathbf{x} + t(\mathbf{y}-\mathbf{x}), \quad t \in [0,1],$$

and

$$\phi(t) := f(\gamma(t)).$$

By the chain rule,

$$\phi'(t) = \langle \nabla f(\gamma(t)), \mathbf{y} - \mathbf{x} \rangle.$$

Using the fundamental theorem of calculus,

$$f(\mathbf{y}) - f(\mathbf{x}) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\, dt.$$

Add and subtract $\nabla f(\mathbf{x})$ inside the integrand:

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) = &\int_0^1 \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\, dt \\ &+ \int_0^1 \langle \nabla f(\gamma(t)) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\, dt. \end{aligned}$$

The first term gives

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

For the second term, by Cauchy–Schwarz and β-Lipschitz continuity of ∇f,

$$\begin{aligned} |\langle \nabla f(\gamma(t)) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &\leq \|\nabla f(\gamma(t)) - \nabla f(\mathbf{x})\|_2 \|\mathbf{y} - \mathbf{x}\|_2 \\ &\leq \beta t \|\mathbf{y} - \mathbf{x}\|_2^2. \end{aligned}$$

Integrating over $t \in [0, 1]$ yields

$$\int_0^1 \beta t \|\mathbf{y} - \mathbf{x}\|_2^2\, dt = \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Combining all terms gives

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

□

Lemma 3.4 (Descent property of gradient descent)
Let $f \in C_\beta^1(\mathbb{X})$ and let $\gamma \in (0, 2/\beta]$. Consider the gradient descent iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k), \quad \mathbf{x}_0 \in \mathbb{X}.$$

Then, for all $k \geq 0$,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k),$$

and more precisely,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \gamma \Big(1 - \tfrac{\beta\gamma}{2}\Big) \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Proof. Fix $k \geq 0$. Apply Lemma 3.4 (the smoothness upper bound) with

$$\mathbf{x} = \mathbf{x}_k, \qquad \mathbf{y} = \mathbf{x}_{k+1}.$$

Since

$$\mathbf{x}_{k+1} - \mathbf{x}_k = -\gamma \nabla f(\mathbf{x}_k),$$

Lemma 3.4 yields

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\beta}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\ &= f(\mathbf{x}_k) - \gamma\|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{\beta}{2}\gamma^2\|\nabla f(\mathbf{x}_k)\|_2^2. \end{aligned}$$

Rearranging gives

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \gamma\left(1 - \tfrac{\beta\gamma}{2}\right)\|\nabla f(\mathbf{x}_k)\|_2^2.$$

If $\gamma \in (0, 2/\beta]$, then $1 - \frac{\beta\gamma}{2} \geq 0$, and hence $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$. □

Theorem 3.5
Let $f \in C^1_\beta(\mathbb{R}^n)$ be convex and let $\mathbf{x}^* \in \arg\min f$ with $p^* := f(\mathbf{x}^*)$. Consider gradient descent with step size $\gamma \in (0, 1/\beta]$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma\nabla f(\mathbf{x}_k).$$

Define

$$\delta_k := f(\mathbf{x}_k) - p^*.$$

Then:

1. $\{f(\mathbf{x}_k)\}$ is nonincreasing and bounded below by p^*;
2. $\sum_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|_2^2 < \infty$;
3. there exists $C > 0$ such that

$$\delta_k \leq \frac{C}{k} \quad \text{for all } k \geq 1.$$

Proof. By the descent lemma, for $\gamma \leq 1/\beta$,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Subtracting p^* from both sides yields

$$\delta_{k+1} \leq \delta_k - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{1}$$

Hence $\{\delta_k\}$ is nonincreasing and bounded below by 0, and therefore

$$\delta_k \to v \geq 0.$$

Summing inequality *(1)* from $k = 0$ to $N - 1$ gives

$$\delta_N \leq \delta_0 - \frac{\gamma}{2}\sum_{k=0}^{N-1} \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Since $\delta_N \geq 0$, we obtain

$$\sum_{k=0}^{\infty} \|\nabla f(\mathbf{x}_k)\|_2^2 \leq \frac{2}{\gamma}\delta_0 < \infty. \tag{2}$$

By convexity of f,

$$f(\mathbf{x}_k) - p^* \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle.$$

Using the Cauchy–Schwarz inequality,

$$\delta_k \leq \|\nabla f(\mathbf{x}_k)\|_2 \|\mathbf{x}_k - \mathbf{x}^*\|_2.$$

Since $\{f(\mathbf{x}_k)\}$ is decreasing and f is coercive on level sets, there exists $R > 0$ such that

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq R \quad \text{for all } k.$$

The boundedness of $\{\mathbf{x}_k\}$ is established later using Fejér monotonicity; see Remark 3.

Thus,

$$\delta_k \leq R \|\nabla f(\mathbf{x}_k)\|_2. \tag{3}$$

Combining *(1)* and *(3)*, we obtain

$$\delta_{k+1} \leq \delta_k - \frac{\gamma}{2R^2} \delta_k^2.$$

This implies

$$\frac{1}{\delta_{k+1}} \geq \frac{1}{\delta_k} + \frac{\gamma}{2R^2}.$$

Iterating,

$$\frac{1}{\delta_k} \geq \frac{1}{\delta_0} + k \frac{\gamma}{2R^2}.$$

Taking reciprocals yields

$$\delta_k \leq \frac{2R^2}{\gamma k} = \frac{C}{k}, \quad C := \frac{2R^2}{\gamma}.$$

□

Remark
The boundedness assumption used in the proof above is not restrictive. In fact, for $\gamma \in (0, 1/\beta]$, the gradient descent iterates $\{\mathbf{x}_k\}$ are Fejér monotone with respect to $\arg\min f$, which implies boundedness and convergence.

Lemma 3.6 (Baillon-Haddad)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and continuously differentiable. If ∇f is β-Lipschitz, then ∇f is $\frac{1}{\beta}$-cocoercive, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

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

Proof. Since $f \in C^1_\beta(\mathbb{R}^n)$, the descent lemma gives, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

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

Exchanging the roles of $\mathbf{x}$ and $\mathbf{y}$ yields

$$f(\mathbf{x}) \le f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2. \tag{2}$$

Adding *(1)* and *(2)* gives

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \le \beta \|\mathbf{x} - \mathbf{y}\|_2^2. \tag{3}$$

On the other hand, since ∇f is β-Lipschitz,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \le \beta \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Rearranging yields the desired inequality:

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge \frac{1}{\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2.$$

□

Definition 3.7 (Fejér monotonicity)
Let $\mathcal{C} \subset \mathbb{R}^n$ be nonempty. A sequence $\{\mathbf{x}_k\} \subset \mathbb{R}^n$ is said to be *Fejér monotone* with respect to $\mathcal{C}$ if

$$\|\mathbf{x}_{k+1} - \mathbf{z}\|_2 \le \|\mathbf{x}_k - \mathbf{z}\|_2 \quad \text{for all } \mathbf{z} \in \mathcal{C} \text{ and all } k \ge 0.$$

Lemma 3.8
Let $f \in C_\beta^1(\mathbb{R}^n)$ be convex and let $\mathcal{X}^* := \arg\min f \neq \varnothing$. Assume $\gamma \in (0, 1/\beta]$ and consider

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k).$$

Then $\{\mathbf{x}_k\}$ is Fejér monotone with respect to $\mathcal{X}^*$.

Proof. Fix $\mathbf{x}^* \in \mathcal{X}^*$. We compute

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \gamma \nabla f(\mathbf{x}_k).$$

Taking squared norms,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2\gamma \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle + \gamma^2 \|\nabla f(\mathbf{x}_k)\|_2^2.$$

By convexity of f,

$$\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \ge f(\mathbf{x}_k) - f(\mathbf{x}^*). \tag{1}$$

Since f is convex and β-smooth, the Baillon–Haddad 3.6 implies that ∇f is $1/\beta$-cocoercive. In particular, using $\nabla f(\mathbf{x}^*) = 0$, we obtain

$$\|\nabla f(\mathbf{x}_k)\|_2^2 \le \beta \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle. \tag{2}$$

Substituting *(1)* and *(2)* into the norm identity yields

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 &\leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2\gamma\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + \beta\gamma^2\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \gamma(2 - \beta\gamma)\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle.\end{aligned}$$

Since $\gamma \leq 1/\beta$, the coefficient $(2 - \beta\gamma)$ is nonnegative, and using *(1)* once more, we conclude

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2.$$

Hence,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2,$$

which proves Fejér monotonicity. □

Lemma 3.9
Let $\{\mathbf{x}_k\}$ be Fejér monotone with respect to a closed set $\mathcal{C} \subset \mathbb{R}^n$. If there exists a subsequence $\mathbf{x}_{k_j} \to \mathbf{x}^* \in \mathcal{C}$, then $\mathbf{x}_k \to \mathbf{x}^*$.

Proof. Since $\{\mathbf{x}_k\}$ is Fejér monotone with respect to $\mathcal{C}$, it is bounded. Hence, by the Bolzano-Weierstrass theorem, it admits a convergent subsequence $\mathbf{x}_{k_j} \to \mathbf{x}^* \in \mathcal{C}$.

Fix $\mathbf{z} = \mathbf{x}^*$. By Fejér monotonicity,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2,$$

so the sequence $\|\mathbf{x}_k - \mathbf{x}^*\|_2$ is nonincreasing and converges to some $\ell \geq 0$.

Along the subsequence,

$$\|\mathbf{x}_{k_j} - \mathbf{x}^*\|_2 \to 0,$$

hence $\ell = 0$. Therefore,

$$\mathbf{x}_k \to \mathbf{x}^*.$$

□

Lemma 3.10 (Fejér monotonicity of gradient descent iterates)
Let $f \in C^1_\beta(\mathbb{R}^n)$ be convex and let $\mathbf{x}^* \in \arg\min f$. Assume $\gamma \in (0, 1/\beta]$ and consider the gradient descent iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma\nabla f(\mathbf{x}_k).$$

Then the sequence $\{\mathbf{x}_k\}$ is Fejér monotone with respect to the solution set $\mathcal{X}^* := \arg\min f$, i.e.,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2 \quad \text{for all } \mathbf{x}^* \in \mathcal{X}^*,$$

and in particular $\{\mathbf{x}_k\}$ is bounded.

Proof. Fix $\mathbf{x}^* \in \mathcal{X}^*$. Using the update rule,

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \gamma \nabla f(\mathbf{x}_k).$$

Taking squared norms,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2\gamma \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle + \gamma^2 \|\nabla f(\mathbf{x}_k)\|_2^2.$$

By convexity of f and optimality of $\mathbf{x}^*$,

$$\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \geq f(\mathbf{x}_k) - f(\mathbf{x}^*). \tag{1}$$

By β-smoothness,

$$\|\nabla f(\mathbf{x}_k)\|_2^2 \leq 2\beta\big(f(\mathbf{x}_k) - f(\mathbf{x}^*)\big). \tag{2}$$

Substituting *(1)* and *(2)* into the squared norm expression yields

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 &\leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2\gamma\big(f(\mathbf{x}_k) - f(\mathbf{x}^*)\big) + 2\beta\gamma^2\big(f(\mathbf{x}_k) - f(\mathbf{x}^*)\big) \\ &= \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2\gamma(1 - \beta\gamma)\big(f(\mathbf{x}_k) - f(\mathbf{x}^*)\big). \end{aligned}$$

Since $\gamma \leq 1/\beta$, the coefficient $2\gamma(1 - \beta\gamma)$ is nonnegative, and hence

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2.$$

Therefore, $\{\mathbf{x}_k\}$ is Fejér monotone with respect to $\mathcal{X}^*$. In particular, the sequence $\{\mathbf{x}_k\}$ is bounded. □

Lemma 3.11 (Convergence of Fejér monotone sequences)
Let $\mathcal{C} \subset \mathbb{R}^n$ be nonempty and closed, and let $\{\mathbf{x}_k\} \subset \mathbb{R}^n$ be Fejér monotone with respect to $\mathcal{C}$, i.e.,

$$\|\mathbf{x}_{k+1} - \mathbf{z}\|_2 \leq \|\mathbf{x}_k - \mathbf{z}\|_2 \quad \text{for all } \mathbf{z} \in \mathcal{C}, \ \forall k \geq 0.$$

If there exists a subsequence $\{\mathbf{x}_{k_j}\}$ and a point $\mathbf{x}^* \in \mathcal{C}$ such that

$$\mathbf{x}_{k_j} \to \mathbf{x}^*,$$

then the whole sequence converges:

$$\mathbf{x}_k \to \mathbf{x}^*.$$

Proof. Fix $\mathbf{z} = \mathbf{x}^* \in \mathcal{C}$. By Fejér monotonicity, the sequence

$$d_k := \|\mathbf{x}_k - \mathbf{x}^*\|_2$$

is nonincreasing and bounded below by 0. Hence, there exists $\ell \geq 0$ such that

$$d_k \to \ell.$$

On the other hand, along the convergent subsequence,

$$d_{k_j} = \|\mathbf{x}_{k_j} - \mathbf{x}^*\|_2 \to 0.$$

Therefore, $\ell = 0$. It follows that

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \to 0,$$

and hence $\mathbf{x}_k \to \mathbf{x}^*$. □

Lemma 3.12 (Convergence via boundedness and Fejér monotonicity)
Let $f \in C^1_\beta(\mathbb{R}^n)$ be convex and let $\mathcal{X}^* := \arg\min f \neq \varnothing$. Assume $\gamma \in (0, 1/\beta]$ and consider the gradient descent iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k).$$

Then the sequence $\{\mathbf{x}_k\}$ converges to a point $\mathbf{x}^* \in \mathcal{X}^*$.

Proof. By Fejér monotonicity with respect to $\mathcal{X}^*$, the sequence $\{\mathbf{x}_k\}$ is bounded. Hence, by the Bolzano–Weierstrass theorem, there exists a subsequence $\{\mathbf{x}_{k_j}\}$ and a point $\mathbf{x}^* \in \mathbb{R}^n$ such that

$$\mathbf{x}_{k_j} \to \mathbf{x}^*.$$

By the descent lemma,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}_k)\|_2^2,$$

which implies

$$\|\nabla f(\mathbf{x}_k)\|_2 \to 0.$$

By continuity of ∇f,

$$\nabla f(\mathbf{x}^*) = 0.$$

Since f is convex, this implies $\mathbf{x}^* \in \mathcal{X}^*$.

By Fejér monotonicity and Lemma 3.11, the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^* \in \mathcal{X}^*$.

□

Lemma 3.13
Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. The following statements are equivalent:

1. f is convex.
2. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

3. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0.$$

4 Projected Gradient Descent

Definition 4.1 (Projection onto a convex set)
Let $\mathbb{X} \subset \mathbb{R}^n$ be nonempty, closed, and convex. The projection operator $\Pi_{\mathbb{X}} : \mathbb{R}^n \to \mathbb{X}$ is defined by

$$\Pi_{\mathbb{X}}(\mathbf{z}) := \arg\min_{\mathbf{x} \in \mathbb{X}} \|\mathbf{x} - \mathbf{z}\|_2.$$

Proposition 4.2 (Properties of the projection)
Let $\mathbb{X} \subset \mathbb{R}^n$ be closed and convex. Then:

1. $\Pi_{\mathbb{X}}$ is well-defined and single-valued.
2. $\Pi_{\mathbb{X}}$ is nonexpansive:

$$\|\Pi_{\mathbb{X}}(\mathbf{x}) - \Pi_{\mathbb{X}}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2.$$

3. (Projection inequality)

$$\langle \mathbf{z} - \Pi_{\mathbb{X}}(\mathbf{z}), \mathbf{x} - \Pi_{\mathbb{X}}(\mathbf{z}) \rangle \leq 0, \quad \forall \mathbf{x} \in \mathbb{X}.$$

Proof of (1): Existence. Let $\mathcal{C} \subset \mathbb{R}^n$ be nonempty, closed, and convex, and fix $\mathbf{x} \in \mathbb{R}^n$. Choose any point $\mathbf{x}_0 \in \mathcal{C}$ and consider the problem:

$$\min_{\mathbf{z} \in \mathcal{C}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2.$$

Set

$$R := \|\mathbf{x} - \mathbf{x}_0\|_2 \quad \text{and} \quad \mathcal{K} := \mathcal{C} \cap \overline{B}(\mathbf{x}, R),$$

where $\overline{B}(\mathbf{x}, R)$ denotes the closed ball centered at $\mathbf{x}$ with radius R.

For any $\mathbf{z} \in \mathcal{C}$ such that $\mathbf{z} \notin \mathcal{K}$, we must have $\|\mathbf{z} - \mathbf{x}\|_2 > R$. Substituting R, this implies:

$$\frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 > \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|_2^2.$$

Since $\mathbf{x}_0 \in \mathcal{K}$ achieves a strictly lower objective value, no point outside $\mathcal{K}$ can be a minimizer. Therefore,

$$\arg\min_{\mathbf{z} \in \mathcal{C}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2 = \arg\min_{\mathbf{z} \in \mathcal{K}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2.$$

The set $\mathcal{K}$ is the intersection of a closed set $\mathcal{C}$ and a closed, bounded ball, so $\mathcal{K}$ is closed and bounded (hence compact). Since the function

$$\mathbf{z} \mapsto \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2$$

is continuous, it attains its minimum on the compact set $\mathcal{K}$ by the *Weierstrass extreme value theorem*. Thus, a minimizer exists.
Proof of (2): Uniqueness TODO □

Projected Gradient Descent

Input: f, C, γ, $\mathbf{x}_0$
for $k = 0, 1, 2, \ldots$ **do**
    $\mathbf{x}_{k+1} = \Pi_C\big(\mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k)\big)$
end for

Remark
- If $\mathbb{X} = \mathbb{R}^n$, Algorithm 4 reduces to standard gradient descent.
- Constraints are enforced via projection rather than penalties.
- The analysis parallels unconstrained gradient descent using Fejér monotonicity.

Theorem 4.3 (Theorem A)
Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex, and let f be differentiable on an open set containing C. Assume that the problem

$$\min_{\mathbf{x}\in C} f(\mathbf{x})$$

is solvable.
Let $\{\mathbf{x}_k\}$ be the sequence generated by projected gradient descent. Then

$$f(\mathbf{x}_k) - \min_{\mathbf{x}\in C} f(\mathbf{x}) \le \mathcal{O}\left(\frac{1}{k}\right),$$

and

$$\mathbf{x}_k \to \mathbf{x}^*, \qquad \mathbf{x}^* \in \arg\min_{\mathbf{x}\in C} f(\mathbf{x}).$$

Proposition 4.4 (Optimality condition)
Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex, and let f be differentiable on an open set containing C. For a point $\mathbf{x}^* \in C$, the following statements are equivalent:

1. $$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in C} f(\mathbf{x})$$
2. $$\mathbf{x}^* = \Pi_C\big(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*)\big), \quad \forall \gamma > 0.$$
3. $$\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}^*)\rangle \ge 0, \quad \forall \mathbf{x} \in C.$$

Proposition 4.5 (Acuteness property of projection)
Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex. For $\mathbf{x}_0 \in \mathbb{R}^n$ and $\mathbf{p} \in C$, the following statements are equivalent:

1. $$\mathbf{p} = \Pi_C(\mathbf{x}_0)$$
2. $$\langle \mathbf{x}_0 - \mathbf{p}, \mathbf{x} - \mathbf{p}\rangle \le 0, \quad \forall \mathbf{x} \in C.$$

Remark

The inequality states that the angle between the vectors $\mathbf{x}_0 - \mathbf{p}$ and $\mathbf{x} - \mathbf{p}$ is obtuse or right for every $\mathbf{x} \in C$. Geometrically, this means that the segment joining $\mathbf{x}_0$ to its projection $\mathbf{p}$ forms a supporting hyperplane to the set C at $\mathbf{p}$.