Advance Convex Optimization

Gourishanker

Contents

1 Lecture 3

Definition 1.1

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 1.2

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

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

($\Leftarrow$) Assume $\|\nabla f(\mathbf{x})\|_2 \leq \beta$ 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,

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

Thus, f is β-Lipschitz. $\qquad\square$

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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

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

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

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

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,

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

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

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

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

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

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. $\qquad\square$

> **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 1.3 (Smoothness Upper Bound or Descent Lemma)
Let $f \in C_\beta^1(\mathbb{X})$, 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:
$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\, dt$$
$$+ \int_0^1 \langle \nabla f(\gamma(t)) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\, dt.$$

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

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

$\square$

Lemma 1.4 (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,
$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \gamma\left(1 - \frac{\beta\gamma}{2}\right)\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Proof. Fix $k \geq 0$. Apply Lemma 1.4 (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 1.4 yields

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\beta}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$$

$$= f(\mathbf{x}_k) - \gamma \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{\beta}{2} \gamma^2 \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Rearranging gives

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

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

Theorem 1.5

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

$$\delta_k \le \frac{C}{k} \quad \text{for all } k \ge 1.$$

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

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

Subtracting p^* from both sides yields

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

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

$$\delta_k \to v \ge 0.$$

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

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

Since $\delta_N \ge 0$, we obtain

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

5

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 1.

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

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

$\square$

> **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 1.6 (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 $f \in C_\beta^1(\mathbb{R}^n)$, 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

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

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

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

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

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

Rearranging yields the desired inequality:

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

$\square$

Definition 1.7 (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

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

Lemma 1.8

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,

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

Since f is convex and β-smooth, the Baillon–Haddad 1.6 implies that ∇f is $1/\beta$-cocoercive. In particular, using $\nabla f(\mathbf{x}^*) = 0$, we obtain

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

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

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

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 1.9
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 1.10 (Fejér monotonicity of gradient descent iterates)
Let $f \in C_\beta^1(\mathbb{R}^n)$ 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

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

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. $\qquad\square$

Lemma 1.11 (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 $\mathbf{x}_k \to \mathbf{x}^*$. $\qquad\square$

> **Lemma 1.12** (Convergence via boundedness and Fejér monotonicity)
> 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 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 1.11, the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^* \in \mathcal{X}^*$.

$\square$

> **Lemma 1.13**
> 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,
>
> $$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0.$$

2 Projected Gradient Descent

Definition 2.1 (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 2.2 (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 2 reduces to standard gradient descent.
- Constraints are enforced via projection rather than penalties.
- The analysis parallels unconstrained gradient descent using Fejér monotonicity.

Theorem 2.3 (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

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

and

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

Proposition 2.4 (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.
$$\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}^*) \rangle \geq 0, \quad \forall \mathbf{x} \in C.$$

Proposition 2.5 (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.
$$\langle \mathbf{x}_0 - \mathbf{p}, \mathbf{x} - \mathbf{p} \rangle \leq 0, \quad \forall \mathbf{x} \in C.$$

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 $\mathbf{p}$.