Chapter 8

Optimization

In this chapter, we focus on optimization problems of the following form:

$$\text{Find} \quad \boldsymbol{x}_* \in \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{K}} J(\boldsymbol{x}), \tag{8.1}$$

where $\mathcal{K}$ is a given subset of $\mathbf{R}^n$ and $J \colon \mathcal{K} \to \mathbf{R}$ is a given *objective function*. We came across several examples of such problems earlier in these notes:

- In Chapter 2, in the context of least-squares approximation, we considered the problem of minimizing
$$J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b}\|^2.$$

- In Chapter 4, we observed that, if A is a symmetric and positive definite matrix, then solving the linear system $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ amounts to finding the minimizer of the functional
$$J(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^T \mathsf{A} \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}.$$

When $\mathcal{K} = \mathbf{R}^n$, equation (8.1) is an *unconstrained* optimization problem, and when $\mathcal{K} \subsetneq \mathbf{R}^n$, equation (8.1) is a *constrained* optimization problem. In practice, the set $\mathcal{K}$ is often an intersection of sets of the form

$$\big\{\boldsymbol{x} \in \mathbf{R}^n : \phi(\boldsymbol{x}) \leqslant 0\big\}, \qquad \text{or} \qquad \big\{\boldsymbol{x} \in \mathbf{R}^n : \phi(\boldsymbol{x}) = 0\big\},$$

for appropriate $\phi \colon \mathbf{R}^n \to \mathbf{R}$. Constraints of the former form are called *inequality constraints*, while constraints of the latter form are called *equality constraints*. Our aim in this chapter is to give a brief introduction to numerical optimization. We focus on the simplest method, namely the *steepest descent method* with fixed step. The rest of this chapter is organized as follows:

- We begin in Section 8.1 by defining the notions of *convexity*, *strict convexity* and *strong convexity*, which play an important role in optimization.

- Then, in Section 8.2, we analyze the steepest descent method with fixed step in the setting of unconstrained optimization. To this end, we first establish conditions under which (8.1) is well posed.

- Finally, in Section 8.3, we extend the steepest descent method to the case of optimization with constraints.

Remark 8.1. For generality, we could consider the setting where the set $\mathcal{K}$ in (8.1) is a subset of some finite dimensional or infinite dimensional vector space V. An optimization problem over (a subset of) a finite dimensional vector space of dimension n can always be recast as an optimization problem over (a subset of) $\mathbf{R}^n$ – the type we study in this chapter – by fixing a basis. The case of an infinite dimensional vector space, however, is more delicate, and we do not address it here.

8.1 Definition and characterization of convexity

Definition 8.1 (Convexity)**.** Assume that $J \colon \mathcal{K} \to \mathbf{R}$.

- The function J is said to be *convex* if

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K} \times \mathcal{K}, \qquad \forall \theta \in [0,1], \qquad J\bigl(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}\bigr) \leqslant \theta J(\boldsymbol{x}) + (1-\theta) J(\boldsymbol{y}). \tag{8.2}$$

- The function J is called *strictly convex* if (8.2) holds with strict inequality if $\boldsymbol{x} \neq \boldsymbol{y}$ and $\theta \in (0,1)$.

- The function J is called *strongly convex* with parameter $\alpha > 0$ if for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K} \times \mathcal{K}$ and for all $\theta \in [0,1]$,

$$J\bigl(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}\bigr) \leqslant \theta J(\boldsymbol{x}) + (1-\theta) J(\boldsymbol{y}) - \frac{\alpha}{2}\theta(1-\theta)\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{8.3}$$

If the function J is differentiable, then convexity, strict convexity and strong convexity can be characterized in terms of the gradient ∇J. We illustrate this for strong convexity, noting that a characterization of convexity is obtained by substituting $\alpha = 0$ in the following result.

Proposition 8.1. *A differentiable function* $J \colon \mathbf{R}^n \to \mathbf{R}$ *is strongly convex with parameter* α *if and only if*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \qquad J(\boldsymbol{x}) \geqslant J(\boldsymbol{y}) + \bigl\langle \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \bigr\rangle + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \tag{8.4}$$

or, equivalently,

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \qquad \bigl\langle \nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \bigr\rangle \geqslant \alpha \|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{8.5}$$

Proof. For clarity, we divide the proof into items and prove one implication per item.

- (8.3) $\Rightarrow$ (8.4). Rearranging (8.3), we have

$$\frac{J\bigl(\boldsymbol{y} + \theta(\boldsymbol{x} - \boldsymbol{y})\bigr) - J(\boldsymbol{y})}{\theta} \leqslant J(\boldsymbol{x}) - J(\boldsymbol{y}) - \frac{\alpha}{2}(1-\theta)\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Taking the limit $\theta \to 0$, we deduce that

$$\langle \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leqslant J(\boldsymbol{x}) - J(\boldsymbol{y}) - \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

This gives (8.4) after rearranging.

- (8.4) $\Rightarrow$ (8.3). To prove this implication, suppose that (8.4) holds, take $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n$ and let $\boldsymbol{z} = \theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y}$. Using (8.4) successively with $(\boldsymbol{x}, \boldsymbol{z})$ and $(\boldsymbol{y}, \boldsymbol{z})$, we deduce

$$\begin{aligned} J(\boldsymbol{x}) &\geqslant J(\boldsymbol{z}) + \langle \nabla J(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{z}\|^2, \\ J(\boldsymbol{y}) &\geqslant J(\boldsymbol{z}) + \langle \nabla J(\boldsymbol{z}), \boldsymbol{y} - \boldsymbol{z} \rangle + \frac{\alpha}{2} \|\boldsymbol{y} - \boldsymbol{z}\|^2. \end{aligned}$$

  Combining these inequalities, we deduce that

$$\begin{aligned} \theta J(\boldsymbol{x}) + (1 - \theta) J(\boldsymbol{y}) \geqslant J(\boldsymbol{z}) &+ \Big\langle \nabla J(\boldsymbol{z}), \theta \boldsymbol{x} + (1 - \theta)\boldsymbol{y} - \boldsymbol{z} \Big\rangle \\ &+ \frac{\alpha \theta}{2} \|\boldsymbol{x} - \boldsymbol{z}\|^2 + \frac{\alpha(1 - \theta)}{2} \|\boldsymbol{y} - \boldsymbol{z}\|^2 \\ &= J(\boldsymbol{z}) + 0 + \frac{\alpha}{2} \theta (1 - \theta) \|\boldsymbol{x} - \boldsymbol{y}\|^2. \end{aligned}$$

  Rearranging gives (8.3).

- (8.4) $\Rightarrow$ (8.5). Assuming that (8.4) holds and applying this inequality first to $(\boldsymbol{x}, \boldsymbol{y})$ and then to $(\boldsymbol{y}, \boldsymbol{x})$, we obtain

$$\begin{aligned} J(\boldsymbol{x}) &\geqslant J(\boldsymbol{y}) + \langle \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ J(\boldsymbol{y}) &\geqslant J(\boldsymbol{x}) + \langle \nabla J(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2. \end{aligned}$$

  Adding these equations and rearranging, we deduce (8.5).

- (8.5) $\Rightarrow$ (8.4). Suppose that (8.5) holds and take $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n$. Using the fundamental theorem of analysis and (8.5), we have

$$\begin{aligned} J(\boldsymbol{x}) &= J(\boldsymbol{y}) + \int_0^1 \Big\langle \nabla J\big(\boldsymbol{y} + \theta(\boldsymbol{x} - \boldsymbol{y})\big), \boldsymbol{x} - \boldsymbol{y} \Big\rangle \, \mathrm{d}\theta \\ &\geqslant J(\boldsymbol{y}) + \int_0^1 \Big\langle \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \Big\rangle + \alpha \theta \|\boldsymbol{x} - \boldsymbol{y}\|^2 \, \mathrm{d}\theta \\ &= J(\boldsymbol{y}) + \Big\langle \nabla J(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \Big\rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \end{aligned}$$

  which gives (8.4).

We have proved all the implications required to conclude the proof. □

8.2 Unconstrained optimization

Throughout this section $\mathcal{K} = \mathbf{R}^n$. We begin by establishing conditions under which the optimization problem (8.1) admits a unique solution in this setting. We first prove existence of a global minimizer under appropriate conditions.

Proposition 8.2 (Existence of a global minimizer)**.** *Suppose that* $J\colon \mathbf{R}^n \to \mathbf{R}$ *is continuous and coercive, the latter meaning that* $J(\boldsymbol{x}) \to \infty$ *when* $\|\boldsymbol{x}\| \to \infty$*. Then there exists a global minimizer of* J *in* $\mathbf{R}^n$*.*

Proof. Let $(\boldsymbol{x}_n)_{n\in\mathbf{N}}$ be a minimizing sequence of J, i.e. a sequence in $\mathbf{R}^n$ such that

$$J(\boldsymbol{x}_n) \to \inf_{\boldsymbol{x}\in\mathbf{R}^n} J(\boldsymbol{x}) \qquad \text{as } n \to \infty.$$

The sequence $(\boldsymbol{x}_n)$ is bounded, because otherwise it would hold that $J(\boldsymbol{x}_n) \to \infty$ by coercivity. Therefore, since closed bounded sets in $\mathbf{R}^n$ are compact, there is a subsequence $(\boldsymbol{x}_{n_k})_{k\in\mathbf{N}}$ converging to some $\boldsymbol{x}_* \in \mathbf{R}^n$. Since J is continuous, we have that

$$J(\boldsymbol{x}_*) = \lim_{k\to\infty} J(\boldsymbol{x}_{n_k}) = \inf_{\boldsymbol{x}\in\mathbf{R}^n} J(\boldsymbol{x}).$$

We conclude that $\boldsymbol{x}_*$ is a minimizer of J. □

Remark 8.2. We relied crucially in the proof of Proposition 8.2 on the fact that closed bounded sets in $\mathbf{R}^n$ are compact. In the infinite-dimensional setting, coercivity and continuity alone are not sufficient to guarantee the existence of a minimizer.

Uniqueness of the minimizer can be established under a strict convexity assumption.

Proposition 8.3 (Uniqueness of the minimizer)**.** *If* J *is strictly convex, then there exists at most one global minimizer.*

Proof. Suppose for contradiction that there were two minimizers $\boldsymbol{x}_*$ and $\boldsymbol{y}_*$. Then by strict convexity we have

$$J\left(\frac{\boldsymbol{x}_* + \boldsymbol{y}_*}{2}\right) < \frac{1}{2}\big(J(\boldsymbol{x}_*) + J(\boldsymbol{y}_*)\big) = J(\boldsymbol{x}_*),$$

which contradicts the minimality of $J(\boldsymbol{x}_*)$. □

Finally, before introducing the steepest descent algorithm, we recall the following standard result from analysis, the proof of which is left as an exercise.

Theorem 8.4 (Euler condition)**.** *Suppose that* $J\colon \mathbf{R}^n \to \mathbf{R}$ *is differentiable.*

- *If* $\boldsymbol{x}_*$ *is a local minimizer of* J*, then* $\nabla J(\boldsymbol{x}_*) = 0$*.*
- *If* J *is convex, then* $\nabla J(\boldsymbol{x}_*) = 0$ *if and only if* $\boldsymbol{x}_*$ *is a global minimizer.*

Steepest descent method. In this section, we study the more general version of the steepest descent with *fixed step* given in Algorithm 17.

Algorithm 17 Steepest descent method

1: Pick λ, and initial $\boldsymbol{x}_0$.
2: **for** $k \in \{0, 1, \dots\}$ **do**
3: $\quad \boldsymbol{d}_k \leftarrow \nabla J(\boldsymbol{x}_k)$
4: $\quad \boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \lambda \boldsymbol{d}_k$
5: **end for**

Remark 8.3. We encountered the steepest descent with fixed step for a quadratic objective function when we analyzed Richardson's method for solving linear equations in Chapter 4.

In practice, Algorithm 17 must be supplemented with an appropriate stopping criterion. This could be, for example, a criterion of the form $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leqslant \varepsilon$, or $\left|J(\boldsymbol{x}_{k+1}) - J(\boldsymbol{x}_k)\right| \leqslant \varepsilon$. It is sometimes also useful to use a normalized criterion of the form $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leqslant \varepsilon \|\boldsymbol{x}_0\|$. The steepest descent method may be viewed as a fixed point iteration for the function

$$\boldsymbol{F}_\lambda(\boldsymbol{x}) = \boldsymbol{x} - \lambda \nabla J(\boldsymbol{x}). \tag{8.6}$$

A point $\boldsymbol{x}_* \in \mathbf{R}^n$ is a fixed point of this function if and only if $\boldsymbol{x}_*$ is a solution to the nonlinear equation $\nabla J(\boldsymbol{x}_*) = 0$. We shall now prove the convergence of the steepest descent under appropriate assumptions on the function J.

Theorem 8.5 (Convergence of the steepest descent method)**.** *Suppose that J is differentiable, strongly convex with parameter α, and that its gradient $\nabla J \colon \mathbf{R}^n \to \mathbf{R}^n$ is Lipschitz continuous with parameter L:*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \qquad \|\nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{y})\| \leqslant L \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{8.7}$$

Then provided that

$$0 < \lambda < \frac{2\alpha}{L}, \tag{8.8}$$

the steepest descent method with fixed step is convergent. More precisely, there exists $\rho \in (0, 1)$ such that for all $k \geqslant 0$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leqslant \rho^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|. \tag{8.9}$$

Proof. Under the assumptions of the theorem, there exists a unique global minimizer of J, which is the unique fixed point of $\boldsymbol{F}_\lambda$. We begin by proving that $\boldsymbol{F}_\lambda$ defined in (8.6) is globally Lipschitz continuous. We have

$$\begin{aligned}
\|\boldsymbol{F}_\lambda(\boldsymbol{x}) - \boldsymbol{F}_\lambda(\boldsymbol{y})\|^2 &= \left\|\boldsymbol{x} - \boldsymbol{y} - \lambda\big(\nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{y})\big)\right\|^2 \\
&= \|\boldsymbol{x} - \boldsymbol{y}\|^2 - 2\lambda \big\langle \boldsymbol{x} - \boldsymbol{y}, \nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{y}) \big\rangle + \lambda^2 \|\nabla J(\boldsymbol{x}) - \nabla J(\boldsymbol{y})\|^2 \\
&\leqslant (1 - 2\alpha\lambda + \lambda^2 L) \|\boldsymbol{x} - \boldsymbol{y}\|^2,
\end{aligned}$$

where we employed (8.5) for the second term and (8.7) for the third term. Thus, $\boldsymbol{F}_\lambda$ is globally Lipschitz continuous with constant $\rho = \sqrt{1 - 2\alpha\lambda + \lambda^2 L}$, which is less than 1 if and only (8.8) is satisfied. The bound (8.9) then follows by noting that

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\| = \|\boldsymbol{F}_\lambda(\boldsymbol{x}_{k-1}) - \boldsymbol{F}_\lambda(\boldsymbol{x}_*)\| \leqslant \rho\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_*\| \leqslant \ldots \leqslant \rho^k\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|,$$

which concludes the proof. (Note that (8.9) also follows from Theorem 5.2.) □

Remark 8.4 (Convergence speed). The choice of λ minimizing the Lipschitz constant ρ is given by $\lambda_* = \frac{\alpha}{L^2}$, which corresponds to $\rho_* = 1 - \left(\frac{\alpha}{L}\right)^2$. Often, in practice, it holds that $\alpha \ll L$, in which case the convergence of the steepest descent with fixed step is slow.

8.3 Constrained optimization

In this section, we assume that $\mathcal{K} \subset \mathbf{R}^n$. We begin by establishing well-posedness of the optimization problem (8.1) in this setting.

Proposition 8.6 (Well posedness of (8.1) in the constrained setting). *The two items below concern existence and uniqueness, respectively.*

- *Suppose that $\mathcal{K} \subset \mathbf{R}^n$ is* closed *and that $J\colon \mathcal{K} \to \mathbf{R}$ is continuous and coercive. Then there exists a global minimizer of J in $\mathcal{K}$.*
- *Suppose that $\mathcal{K} \subset \mathbf{R}^n$ is* convex *and that $J\colon \mathcal{K} \to \mathbf{R}$ is strictly convex. Then there exists at most one global minimizer.*

Proof. The proof is very similar to those of Proposition 8.2 and Proposition 8.3, and so we leave it to the reader. Note that the set $\mathcal{K}$ must be closed to ensure existence, and convex to guarantee uniqueness. These assumptions are clearly satisfied when $\mathcal{K} = \mathbf{R}^n$, so Proposition 8.6 indeed generalizes Propositions 8.2 and 8.3. □

The following theorem, which generalizes (8.4), establishes a characterization of the minimizer when J is differentiable.

Theorem 8.7 (Euler–Lagrange conditions). *Suppose that $J\colon \mathcal{K} \to \mathbf{R}$ is differentiable and that $\mathcal{K} \subset \mathbf{R}^n$ is closed and convex. Then the following statements hold.*

- *If $\boldsymbol{x}_*$ is a local minimizer of J, then*

$$\forall \boldsymbol{x} \in \mathcal{K}, \qquad \langle \nabla J(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle \geqslant 0. \tag{8.10}$$

- *Conversely, if (8.10) is satisfied and J is convex, then $\boldsymbol{x}_*$ is a global minimizer of J.*

Proof. Suppose that $\boldsymbol{x}_*$ is a local minimizer of J. This means that there exists $\delta > 0$ such that

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*) \cap \mathcal{K}, \qquad J(\boldsymbol{x}_*) \leqslant J(\boldsymbol{x}).$$

Therefore $J(\boldsymbol{x}_*) \leqslant J\big((1-t)\boldsymbol{x}_* + t\boldsymbol{x}\big)$ for all $t \in [0,1]$ sufficiently small. But then

$$\langle \nabla J(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle = \lim_{t \to 0} \frac{J\big((1-t)\boldsymbol{x}_* + t\boldsymbol{x}\big) - J(\boldsymbol{x}_*)}{t} \geqslant 0.$$

Conversely, suppose that (8.10) is satisfied and that J is convex. Since J is convex, equation (8.4) holds with $\alpha = 0$, and applying this equation with $\boldsymbol{y} = \boldsymbol{x}_*$, we deduce that $\boldsymbol{x}_*$ is a global minimizer. □

The steepest descent Algorithm 17 can be extended to optimization problems with constraints by introducing an additional projection step. In order to precisely formulate the algorithm, we begin by introducing the projection operator $\Pi_{\mathcal{K}}$.

Proposition 8.8 (Projection on a closed convex set)**.** *Suppose that $\mathcal{K}$ is a closed convex subset of $\mathbf{R}^n$. Then for all $\boldsymbol{x} \in \mathbf{R}^n$ there a unique $\Pi_{\mathcal{K}}\boldsymbol{x} \in \mathcal{K}$, called the orthogonal projection of $\boldsymbol{x}$ onto $\mathcal{K}$, such that*

$$\|\Pi_{\mathcal{K}}\boldsymbol{x} - \boldsymbol{x}\| = \inf_{\boldsymbol{y} \in \mathcal{K}} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

Proof. The functional $J_{\boldsymbol{x}}(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{x}\|^2$ is strongly convex, and so Proposition 8.6 immediately implies the existence and uniqueness of $\Pi_{\mathcal{K}}\boldsymbol{x}$. □

Remark 8.5. In view of Theorem 8.7, the projection $\Pi_{\mathcal{K}}\boldsymbol{x}$ is the unique element of $\mathcal{K}$ which satisfies

$$\forall \boldsymbol{y} \in \mathcal{K}, \qquad \langle \Pi_{\mathcal{K}}\boldsymbol{x} - \boldsymbol{x}, \boldsymbol{y} - \Pi_{\mathcal{K}}\boldsymbol{x} \rangle \geqslant 0. \tag{8.11}$$

We are now ready to present the steepest descent method with projection: see Algorithm 18. Like Algorithm 17, the steepest descent with projection may be viewed as a fixed point iteration, this time for the function

$$\boldsymbol{F}_\lambda(\boldsymbol{x}) := \Pi_{\mathcal{K}}\big(\boldsymbol{x} - \lambda \nabla J(\boldsymbol{x})\big). \tag{8.12}$$

We now prove the convergence of the method.

Algorithm 18 Steepest descent with projection

1: Pick λ, and initial $\boldsymbol{x}_0$.
2: **for** $k \in \{0, 1, \dots\}$ **do**
3: $\quad \boldsymbol{d}_k \leftarrow \nabla J(\boldsymbol{x}_k)$
4: $\quad \boldsymbol{x}_{k+1} \leftarrow \Pi_{\mathcal{K}}(\boldsymbol{x}_k - \lambda \boldsymbol{d}_k)$
5: **end for**

Theorem 8.9 (Convergence of steepest descent with projection)**.** *Suppose that J is differentiable, strongly convex with parameter α, and that its gradient $\nabla J \colon \mathbf{R}^n \to \mathbf{R}^n$ is Lipschitz continuous with parameter L. Assume also that $\mathcal{K} \subset \mathbf{R}^n$ is closed and convex. Then provided that*

$$0 < \lambda < \frac{2\alpha}{L}, \tag{8.13}$$

the steepest descent method with fixed step is convergent. More precisely, there exists $\rho \in (0, 1)$ *such that for all* $k \geqslant 0$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leqslant \rho^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|.$$

Proof. Under the assumptions, there exists a unique global minimizer $\boldsymbol{x}_* \in \mathcal{K}$. We already showed in the proof of Theorem 8.5 that the mapping $\boldsymbol{x} \mapsto \boldsymbol{x} - \lambda \nabla J(\boldsymbol{x})$ is a contraction if and only if λ satisfies (8.13). In order to prove that $\boldsymbol{F}_\lambda$ given in (8.12) is a contraction under the same condition, it is sufficient to prove that $\Pi_{\mathcal{K}} \colon \mathbf{R}^n \to \mathcal{K}$ satisfies the following estimate:

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n, \qquad \|\Pi_{\mathcal{K}} \boldsymbol{x} - \Pi_{\mathcal{K}} \boldsymbol{y}\| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|.$$

To this end, take $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbf{R}^n \times \mathbf{R}^n$ and let $\boldsymbol{\delta} = \Pi_{\mathcal{K}} \boldsymbol{x} - \Pi_{\mathcal{K}} \boldsymbol{y}$. By (8.11), it holds that

$$\begin{aligned}
\|\boldsymbol{\delta}\|^2 &= \langle \boldsymbol{\delta}, \Pi_{\mathcal{K}} \boldsymbol{x} - \boldsymbol{x} \rangle + \langle \boldsymbol{\delta}, \boldsymbol{x} - \boldsymbol{y} \rangle + \langle \boldsymbol{\delta}, \boldsymbol{y} - \Pi_{\mathcal{K}} \boldsymbol{y} \rangle \\
&\leqslant 0 + \langle \boldsymbol{\delta}, \boldsymbol{x} - \boldsymbol{y} \rangle + 0 \leqslant \|\boldsymbol{\delta}\| \|\boldsymbol{x} - \boldsymbol{y}\|,
\end{aligned}$$

which yields the required inequality. Therefore $\boldsymbol{F}_\lambda$ in (8.12) is a contraction and so, by the Banach fixed point theorem, it admits a unique fixed point $\boldsymbol{y}_* \in \mathcal{K}$. To show that $\boldsymbol{y}_* = \boldsymbol{x}_*$, note that if $\boldsymbol{F}_\lambda(\boldsymbol{y}_*) = \boldsymbol{y}_*$, then by (8.11) it holds that

$$\forall \boldsymbol{y} \in \mathcal{K}, \qquad \langle \lambda \nabla J(\boldsymbol{y}_*), \boldsymbol{y} - \boldsymbol{y}_* \rangle \geqslant 0.$$

Therefore, using Theorem 8.7, we obtain that $\boldsymbol{y}_*$ is a global minimizer of J, so $\boldsymbol{y}_* = \boldsymbol{x}_*$. □

Remark 8.6. The applicability of Algorithm 18 is limited in practice, as computing $\Pi_{\mathcal{K}}(\boldsymbol{x})$ analytically is possible only in simple settings.