Optimization Methods

with Applications in Statistics

Lutz Dümbgen
University of Bern

July 11, 2025

Bibliography

[1] A. BEN-TAL and A. NEMIROVSKI (2001). *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications.* SIAM, Philadelphia

[2] R.J. BERAN and L. DÜMBGEN (2010). Least Squares and Shrinkage Estimation under Bimonotonicity Constraints. *Statistics and Computing* ***20****(2)*, 177–189.

[3] CORMEN, T.H., C.E. LEISERSON and R.L. RIVEST (1990). *Introduction to Algorithms.* M.I.T. Press

[4] K. DEIMLING (1985). *Nonlinear Functional Analysis.* Springer, Berlin Heidelberg

[5] R. FLETCHER (1987). *Practical Methods of Optimization (2nd edition).* Wiley, New York

[6] P.E. GILL, W. MURRAY and M.H. WRIGHT (1981). *Practical Optimization.* Academic Press

[7] U. GRENANDER (1956), On the Theory of Mortality Measurement, Part II. *Skandinavisk Aktuarietidskrift* ***39***, 125–153.

[8] F. JARRE und J. STOER (2004). *Optimierung.* Springer, Berlin Heidelberg

[9] K. LANGE (1999). *Numerical Analysis for Statisticians.* Springer, New York

[10] G. OPFER (1994). *Numerische Mathematik für Anfänger.* Vieweg, Braunschweig Wiesbaden.

[11] T. ROBERTSON and P. WALTMAN (1968). On Estimating Monotone Parameters. *Annals of Mathematical Statistics* ***39***, 1030–1039

[12] T. ROBERTSON, F.T. WRIGHT and R.L. DYKSTRA (1988). *Order Restricted Statistical Inference.* Wiley, New York

Acknowledgements. Sebastian Arnold, David Ginsbourger, Micha Grütter, Dirk Klingbiel, Jürg Krähenbühl, Christof Mahnig, Alexandre Mösching, Kaspar Rufibach, Adrian Schmucker, Blendi Shala and Christof Strähl gave numerous hints to different versions of these lecture notes – thanks a lot to all of them!

Contents

Introduction

In this course we are dealing with various optimization methods which are used frequently in statistics and other fields. The optimization problems we'll encounter may be divided roughly into three categories:

(a) Minimization problems. A function $f : \mathcal{X} \to \overline{\mathbb{R}} = [-\infty, \infty]$ has to be minimized. This concerns both the minimal value,

$$\inf_{\mathcal{X}} f \;=\; \inf_{x \in \mathcal{X}} f(x),$$

as well as the set of minimizers,

$$\arg\min_{\mathcal{X}} f \;=\; \arg\min_{x \in \mathcal{X}} f(x) \;:=\; \Big\{ x \in \mathcal{X} : f(x) = \inf_{\mathcal{X}} f \Big\}.$$

(b) Finding roots. For a mapping $f : \mathcal{X} \to \mathcal{Y}$ and a point $b \in \mathcal{Y}$ we want to determine the set

$$f^{-1}(b) \;=\; \{x \in \mathcal{X} : f(x) = b\}$$

or at least one of its elements.

(c) Fixed point problems. For a mapping $f : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{X} \subset \mathcal{Y}$ we want to determine its set of fixed points,

$$\mathrm{FP}(f) \;:=\; \{x \in \mathcal{X} : f(x) = x\}.$$

Quite often one translates minimization problems into root finding or fixed point problems.

Chapter 1

Univariate Procedures

In this chapter we consider functions $f : \mathcal{X} \to \mathbb{R}$ on a real interval $\mathcal{X} \subset \mathbb{R}$.

1.1 Bisection Methods

Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function such that $f(a) \leq 0 \leq f(b)$. According to the intermediate value theorem for continuous functions, there exists at least one point $x_* \in [a, b]$ with $f(x_*) = 0$. This fact is used in the algorithm in Table 1.1. It returns a pair (x_a, x_b) of points in $[a, b]$ such that $0 \leq x_b - x_a \leq \delta$ and $f(x_*) = 0$ for some $x_* \in [x_a, x_b]$. It terminates after

$$\left\lceil \log_2\Big(\frac{b-a}{\delta}\Big) \right\rceil$$

executions of the while-loop, because after $n \geq 0$ executions,

$$x_b - x_a \;=\; (b-a)2^{-n}.$$

Sometimes one would also like to ensure that $|f(x)| \leq \delta$ for all $x \in [x_a, x_b]$. To do so, we now assume that f is also isotonic (non-decreasing) and modify the algorithm from Table 1.1 as described in Table 1.2. In case of a differentiable function f one can verify easily that the

Algorithm $(x_a, x_b) \leftarrow$ **Bisection**(f, a, b, δ)
$(x_a, x_b) \leftarrow (a, b)$
while $x_b - x_a > \delta$ **do**
    $x_o \leftarrow (x_a + x_b)/2$
    if $f(x_o) \leq 0$ **then**
        $x_a \leftarrow x_o$
    else
        $x_b \leftarrow x_o$
    end if
end while.

Table 1.1: Bisection method I.

Algorithm $(x_a, x_b) \leftarrow$ **Bisection2**(f, a, b, δ)
$(x_a, x_b) \leftarrow (a, b)$
$(f_a, f_b) \leftarrow (f(a), f(b))$
while $x_b - x_a > \delta$ **or** $f_b - f_a > \delta$ **do**
    $x_o \leftarrow (x_a + x_b)/2$
    $f_o \leftarrow f(x_o)$
    if $f_o \leq 0$ **then**
        $x_a \leftarrow x_o$
        $f_a \leftarrow f_o$
    else
        $x_b \leftarrow x_o$
        $f_b \leftarrow f_o$
    end if
end while.

Table 1.2: Bisection method II.

while-loop is executed no more than

$$\left\lceil \log_2\Big(\frac{b-a}{\delta}\ \max\Big\{1, \sup_{[a,b]} f'\Big\}\Big)\right\rceil$$

times.

Application to exact confidence bounds

Let Y be a random variable with values in $\mathbb{Z}$ and distribution function F_{θ_*}, where $(F_\theta)_{\theta\in\Theta}$ is a given family of distribution functions, and θ_* is an unknown parameter in Θ. Thus

$$\mathbb{P}(Y \le c) \ = \ F_{\theta_*}(c).$$

One can easily show that

$$\mathbb{P}(F_{\theta_*}(Y) > \alpha) \ \ge \ 1-\alpha \quad \text{and} \quad \mathbb{P}(F_{\theta_*}(Y-1) < 1-\alpha) \ \ge \ 1-\alpha$$

for any fixed $\alpha \in (0,1)$. Thus both

$$C_\alpha^{\text{left}}(Y) \ := \ \big\{\theta \in \Theta : F_\theta(Y) > \alpha\big\}$$

and

$$C_\alpha^{\text{right}}(Y) \ := \ \big\{\theta \in \Theta : F_\theta(Y-1) < 1-\alpha\big\}$$

are $(1-\alpha)$-confidence regions for θ_*. The set $C_\alpha^{\text{left}}(Y)$ consists of all parameters θ such that Y is not "suspiciously small" in the sense that the left-sided p-value $F_\theta(Y)$ is larger than α. Analogously, $C_\alpha^{\text{right}}(Y)$ consists of all parameters θ such that Y is not "suspiciously large" in the sense that the right-sided p-value $1 - F_\theta(Y-1)$ is larger than α.

Suppose that Θ is an interval of real numbers. For any integer c let $\Theta \ni \theta \mapsto F_\theta(c)$ be continuous and antitonic (non-increasing). In this case $C_\alpha^{\text{left}}(Y)$ yields an upper and $C_\alpha^{\text{right}}(Y)$ yields a lower confidence bound for θ_*.

More precisely, let $F_\theta(Y)$ be continuous and strictly antitonic in $\theta \in \Theta$ with limits

$$\lim_{\theta \to \inf(\Theta)} F_\theta(Y) \;=\; 1 \quad \text{and} \quad \lim_{\theta \to \sup(\Theta)} F_\theta(Y) \;=\; 0.$$

Then

$$C_\alpha^{\text{left}}(Y) \;=\; \big\{\theta \in \Theta : \theta < b_\alpha(Y)\big\}$$

with the unique number $b_\alpha(Y) \in \Theta$ such that

$$F_{b_\alpha(Y)}(Y) \;=\; \alpha.$$

Thus we obtain an exact upper confidence bound which may be computed via a bisection procedure.

Analogously let $F_\theta(Y-1)$ be continuous and strictly antitonic in $\theta \in \Theta$ with limits

$$\lim_{\theta \to \inf(\Theta)} F_\theta(Y-1) \;=\; 1 \quad \text{and} \quad \lim_{\theta \to \sup(\Theta)} F_\theta(Y-1) \;=\; 0.$$

Then

$$C_\alpha^{\text{right}}(Y) \;=\; \big\{\theta \in \Theta : \theta > a_\alpha(Y)\big\}$$

with the unique number $a_\alpha(Y) \in \Theta$ such that

$$F_{a_\alpha(Y)}(Y-1) \;=\; 1-\alpha.$$

Example 1.1 (Binomial parameters). Let Y follow a binomial distribution $\mathrm{Bin}(n, \theta_*)$ with given $n \in \mathbb{N}$ and unknown parameter $\theta_* \in \Theta = [0,1]$. Then

$$F_\theta(c) \;=\; \sum_{k=0}^{c} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

for arbitrary $c \in \{0, 1, \ldots, n\}$. As a function of θ, this is a polynomial of order n and thus infinitely often differentiable. In case of $0 \le c < n$ one can even show that $F_\theta(c)$ is strictly antitonic in $\theta \in [0,1]$ with limits $F_0(c) = 1$ and $F_1(c) = 0$. Consequently,

$$C_\alpha^{\text{left}}(Y) \;=\; \begin{cases} [0,1] & \text{if } Y = n, \\ [0, b_\alpha(Y)) & \text{if } Y < n, \text{ where } F_{b_\alpha(Y)}(Y) = \alpha, \end{cases}$$

and

$$C_\alpha^{\text{right}}(Y) \;=\; \begin{cases} [0,1] & \text{if } Y = 0, \\ (a_\alpha(Y), 1] & \text{if } Y > 0, \text{ where } F_{a_\alpha(Y)}(Y-1) = 1-\alpha. \end{cases}$$

For the computation of both confidence bounds $a_\alpha(Y)$ and $b_\alpha(Y)$ one may employ the algorithm in Table 1.3 which returns for given numbers $c \in \{0, 1, \ldots, n-1\}$, $\gamma \in (0,1)$ and $\delta > 0$ two numbers x_a, x_b with the following properties:

$$0 \le x_a < x_b \le 1,\; x_b - x_a \le \delta,$$

and

$$F_{x_a}(c) \ge \gamma \ge F_{x_b}(c),\; F_{x_a}(c) - F_{x_b}(c) \le \delta.$$

It requires an auxiliary function $\texttt{FBino}(\cdot, n, \theta)$ for the computation of the distribution function of $\mathrm{Bin}(n, \theta)$.

Algorithm $(x_a, x_b) \leftarrow \mathbf{BinoCB}(n, c, \gamma, \delta)$
$(x_a, x_b) \leftarrow (0, 1)$
$(F_a, F_b) \leftarrow (1, 0)$
while $x_b - x_a > \delta$ **or** $F_a - F_b > \delta$ **do**
    $x_o \leftarrow (x_a + x_b)/2$
    $F_o \leftarrow \texttt{FBino}(c, n, x_o)$
    if $F_o \geq \gamma$ **then**
        $x_a \leftarrow x_o$
        $F_a \leftarrow F_o$
    else
        $x_b \leftarrow x_o$
        $F_b \leftarrow F_o$
    end if
end while.

Table 1.3: Computation of confidence bounds for a binomial parameter.

Example 1.2 (Poisson parameters). Let Y follow a Poisson distribution $\mathrm{Poiss}(\theta_*)$ with unknown parameter $\theta_* \in \Theta = [0, \infty)$. Then

$$F_\theta(c) \;=\; \exp(-\theta) \sum_{k=0}^{c} \frac{\theta^k}{k!}$$

for arbitrary $c \in \mathbb{N}_0$. As a function of θ, this is continuous and strictly antitonic with limits $F_0(c) = 1$ and $F_\infty(c) = 0$. Thus

$$C_\alpha^{\mathrm{left}}(Y) \;=\; [0, b_\alpha(Y)) \quad \text{with } F_{b_\alpha(Y)}(Y) = \alpha,$$

and

$$C_\alpha^{\mathrm{right}}(Y) \;=\; \begin{cases} [0, \infty) & \text{if } Y = 0, \\ (a_\alpha(Y), \infty) & \text{if } Y > 0, \text{ where } F_{a_\alpha(Y)}(Y-1) = 1 - \alpha. \end{cases}$$

For the computation of both confidence bounds $a_\alpha(Y)$ and $b_\alpha(Y)$ one has to modify algorithm **BinoCB** in two respects. On the one hand we replace the function $\texttt{FBino}(\cdot, n, \theta)$ with the distribution function $\texttt{FPoiss}(\cdot, \theta)$ of $\mathrm{Poiss}(\theta)$. Moreover, at first one has to find a starting interval for the bisection search, see Table 1.4.

Exercise 1.3. Let $w_0, w_1, w_2, \ldots$ be nonnegative weights such that $w_0 > 0$ and

$$C(\theta) := \sum_{k=0}^{\infty} w_k \theta^k \;<\; \infty \quad \text{for all } \theta \geq 0.$$

With these weights we define a family of distributions on $\mathbb{N}_0$: For $\theta \geq 0$ and $c \in \mathbb{N}_0$ let

$$F_\theta(c) \;:=\; C(\theta)^{-1} \sum_{k=0}^{c} w_k \theta^k.$$

(a) Show that $\theta \mapsto F_\theta(c)$ is continuous and antitonic on $[0, \infty)$ with $F_0(c) = 1$. Show that it is even strictly antitonic with limit $F_\infty(c) = 0$, provided that $w_k > 0$ for at least one index $k > c$.

Algorithm $(x_a, x_b) \leftarrow \mathbf{PoissCB}(c, \gamma, \delta)$
$(x_a, x_b) \leftarrow (0, 1)$
$(F_a, F_b) \leftarrow (1, \texttt{FPoiss}(c, 1))$
while $F_b > \gamma$ **do**
    $(x_a, x_b) \leftarrow (x_b, 2x_b)$
    $(F_a, F_b) \leftarrow (F_b, \texttt{FPoiss}(c, x_b))$
end while
while $x_b - x_a > \delta$ **or** $F_a - F_b > \delta$ **do**
    $x_o \leftarrow (x_a + x_b)/2$
    $F_o \leftarrow \texttt{FPoiss}(c, x_o)$
    if $F_o \geq \gamma$ **then**
        $x_a \leftarrow x_o$
        $F_a \leftarrow F_o$
    else
        $x_b \leftarrow x_o$
        $F_b \leftarrow F_o$
    end if
end while.

Table 1.4: Computation of confidence bounds for a Poisson parameter.

(b) This result applies immediately to Poisson distributions, where $w_k = 1/k!$. How could you apply it to binomial distributions? How would you modify the result to be applicable to negative binomial distributions?

1.2 Newton's Method (Univariate Case)

Now we consider a differentiable function $f : \mathcal{X} \to \mathbb{R}$ on an open interval $\mathcal{X} \subset \mathbb{R}$ with strictly positive derivative f'. Moreover we assume that $\inf_{\mathcal{X}} f < 0 < \sup_{\mathcal{X}} f$, hence there exists a unique $x_* \in f^{-1}(0)$.

To find or approximate this point x_* we consider an arbitrary first candidate $x_o \in \mathcal{X}$. Now we approximate f by an affine function:

$$x \;\mapsto\; f(x_o) + f'(x_o)(x - x_o).$$

This auxiliary function attains the value 0 at

$$\psi(x_o) \;:=\; x_o - \frac{f(x_o)}{f'(x_o)}.$$

Note that $\psi(x_o) = x_o$ if and only if $f(x_o) = 0$, whence $x_o = x_*$. Otherwise we hope that $\psi(x_o)$ is closer to x_* than x_o.

Now we would like to iterate this mapping ψ. That means, for a starting value $x_0 \in \mathcal{X}$ one computes inductively

$$x_n \;:=\; \psi(x_{n-1}) \quad \text{for } n = 1, 2, 3, \ldots.$$

When doing so we hope that

(i) this sequence is well-defined in the sense that all points x_n are contained in $\mathcal{X}$,
(ii) it converges to x_*.

Local convergence. The next theorem implies that the sequence $(x_n)_{n=0}^{\infty}$ does have these properties, provided that the starting value x_0 is sufficiently close to x_*.

Theorem 1.4 (Local convergence of Newton's method).

(a) *Let f' be continuous at x_*. Then*

$$\lim_{x \to x_*} \frac{\psi(x) - x_*}{x - x_*} \;=\; 0.$$

(If $x = x_*$, then $\psi(x) = x_*$, and we interpret $0/0$ as 0.)

(b) *Let f be twice differentiable, and let f'' be continuous at x_*. Then*

$$\lim_{x \to x_*} \frac{\psi(x) - x_*}{(x - x_*)^2} \;=\; \frac{f''(x_*)}{2f'(x_*)}.$$

For the sequence $(x_n)_{n=0}^{\infty}$ defined earlier, Theorem 1.4 (a) has the following consequences: For any $\epsilon \in (0,1)$ there exists a $\delta(\epsilon) > 0$ such that $[x_* \pm \delta(\epsilon)] \subset \mathcal{X}$ and

$$|\psi(x) - x_*| \;\le\; \epsilon |x - x_*| \quad \text{for } x \in [x_* \pm \delta(\epsilon)].$$

In particular, if $|x_0 - x_*| \le \delta(\epsilon)$, then $|x_n - x_*| \le |x_0 - x_*|\epsilon^n$ for all $n \ge 0$. More generally, if $|x_{n(\epsilon)} - x_*| \le \delta(\epsilon)$ for some $n(\epsilon) \in \mathbb{N}_0$, then $|x_n - x_*| \le C(\epsilon)\epsilon^n$ for all $n \ge 0$, where $C(\epsilon) := \max_{m \le n(\epsilon)} \epsilon^{-m} |x_m - x_*|$.

This consideration shows that

$$\lim_{n \to \infty} |x_n - x_*| \;=\; 0 \quad \text{and} \quad \lim_{n \to \infty} \frac{|x_{n+1} - x_*|}{|x_n - x_*|} \;=\; 0, \tag{1.1}$$

provided that x_0 is sufficiently close to x_*. We'll show later how to avoid the latter restriction. Property (1.1) of $(x_n)_{n \ge 0}$ is called "super-linear convergence (to x_*)".

For starting values x_0 sufficiently close to x_*, Theorem 1.4 (b) yields even the more precise statement that

$$\lim_{n \to \infty} |x_n - x_*| \;=\; 0 \quad \text{und} \quad \lim_{n \to \infty} \frac{|x_{n+1} - x_*|}{|x_n - x_*|^2} \;=\; \frac{|f''(x_*)|}{2f'(x_*)}. \tag{1.2}$$

This property of $(x_n)_{n=0}^{\infty}$ is called "quadratic convergence (to x_*)".

Proof of Theorem 1.4. For real numbers a, b let $a \wedge b$ and $a \vee b$ denote their minimum and maximum, respectively. According to the mean value theorem, for $x \in \mathcal{X} \setminus \{x_*\}$, there exists a point $\xi(x) \in (x \wedge x_*, x \vee x_*)$ such that

$$f(x) \;=\; f(x) - f(x_*) \;=\; f'(\xi(x))(x - x_*),$$

whence

$$\frac{\psi(x) - x_*}{x - x_*} \;=\; 1 - \frac{f(x)}{f'(x)(x - x_*)} \;=\; 1 - \frac{f'(\xi(x))(x - x_*)}{f'(x)(x - x_*)} \;=\; 1 - \frac{f'(\xi(x))}{f'(x)}.$$

By continuity of f' at x_* with $f'(x_*) > 0$, this implies that

$$\lim_{x\to x_*} \frac{\psi(x) - x_*}{x - x_*} \;=\; 1 - \lim_{x\to x_*} \frac{f'(\xi(x))}{f'(x)} \;=\; 1 - \frac{f'(x_*)}{f'(x_*)} \;=\; 0.$$

This proves part (a).

For part (b) we use Taylor's formula. For $x \in \mathcal{X} \setminus \{x_*\}$ and $h := x - x_*$,

$$f(x) \;=\; f(x_*) + f'(x_*)h + f''(\xi(x))h^2/2 \;=\; f'(x_*)h + f''(\xi(x))h^2/2$$

and

$$f'(x) \;=\; f'(x_*) + f''(\eta(x))h$$

with suitable points $\xi(x), \eta(x)$ in $(x \wedge x_*, x \vee x_*)$. Thus,

$$\begin{aligned}
\frac{\psi(x) - x_*}{(x - x_*)^2} \;&=\; \frac{h - f(x)/f'(x)}{h^2} \\
&=\; \frac{f'(x)h - f(x)}{f'(x)h^2} \\
&=\; \frac{f'(x_*)h + f''(\eta(x))h^2 - f'(x_*)h - f''(\xi(x))h^2/2}{(f'(x_*) + f''(\eta(x))h)h^2} \\
&=\; \frac{f''(\eta(x)) - f''(\xi(x))/2}{f'(x_*) + f''(\eta(x))h} \\
&\to\; \frac{f''(x_*) - f''(x_*)/2}{f'(x_*)} \;=\; \frac{f''(x_*)}{2f'(x_*)}
\end{aligned}$$

as $x \to x_*$, i.e. $h \to 0$. □

Example 1.5. Let us illustrate the potential problems with starting values x_0 which are too far from x_* with the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x(1 + x^2)^{-1/2}$. This function is strictly increasing with $f'(x) = (1 + x^2)^{-3/2}$, and the limits are $f(\pm\infty) = \pm 1$. Here $x_* = 0$, and

$$\psi(x) \;=\; x - x(1 + x^2) \;=\; -x^3.$$

Hence the Newton sequence $(x_n)_{n\geq 0}$ converges if and only if $|x_0| < 1$. Precisely, one can show by induction that

$$x_n \;=\; (-1)^n x_0^{3^n}$$

for all $n \geq 0$.

Global convergence. Under the additional assumption that f is convex and $x_0 \geq x_*$, Newton's method yields always a sequence with limit x_*. This is a consequence of the following result:

Theorem 1.6 (Global convergence of Newton's method). *Suppose that f is convex, that means, f' is isotonic. Then $\psi(x) \geq x_*$ whenever $x \leq x_*$. For $x > x_*$,*

$$0 \;\leq\; \frac{\psi(x) - x_*}{x - x_*} \;\leq\; 1 - \frac{f'(x_* +)}{f'(x)},$$

where $f'(x_* +)$ denotes the rightsided limit of f' at x_*.

Under the assumptions of Theorem 1.6, the sequence $(x_n)_{n\geq 0}$ is monotone decreasing with limit x_*, provided that $x_0 \geq x_*$. The monotonicity is even strikt, unless $x_n = x_*$ for some $n \geq 0$. But $\psi(x) = x_*$ for some $x > x_*$ would imply that $f' \equiv f'(x_* +)$ on (x_*, x).

The restriction $x_0 \geq x_*$ is unnecessary if $\mathcal{X} = (a, \infty)$ for some $a \in [-\infty, \infty)$. For then $x_1 = \psi(x_0) \geq x_*$, and $(x_n)_{n=1}^{\infty}$ is monotone decreasing with limit x_*.

Example 1.7. A classical application of Newton's procedure is the computation of square roots. For any number $\gamma > 0$ we consider the function

$$(0, \infty) \ni x \;\mapsto\; f(x) := x^2 - \gamma.$$

This function is infinitely often differentiable, strictly isotonic and convex, and $f^{-1}(0) = \{\sqrt{\gamma}\}$. Here

$$\psi(x) \;=\; x - \frac{x^2 - \gamma}{2x} \;=\; \frac{x + \gamma/x}{2},$$

and iterating this mapping yields always a sequence converging quadratically to $\sqrt{\gamma}$. Indeed,

$$\frac{\psi(x) - \sqrt{\gamma}}{x - \sqrt{\gamma}} \;=\; \frac{x - \sqrt{\gamma}}{2x} \begin{cases} < \; 1/2 \\ > \; 0 \qquad \text{if } x > \sqrt{\gamma} \end{cases}$$

and

$$\frac{\psi(x) - \sqrt{\gamma}}{(x - \sqrt{\gamma})^2} \;=\; \frac{1}{2x}$$

for any $x > 0$.

Proof of Theorem 1.6. For $x < x_*$ there exists a point $\xi(x) \in (x, x_*)$ such that

$$\psi(x) \;=\; x - \frac{f'(\xi(x))(x - x_*)}{f'(x)} \;=\; x_* + (x - x_*)\Big(1 - \frac{f'(\xi(x))}{f'(x)}\Big) \;\geq\; x_*,$$

because $0 < f'(x) \leq f'(\xi(x))$.

For $x > x_*$ there exists a point $\xi(x) \in (x_*, x)$ with

$$\frac{\psi(x) - x_*}{x - x_*} \;=\; 1 - \frac{f'(\xi(x))}{f'(x)} \;\in\; \Big[0, 1 - \frac{f'(x_* +)}{f'(x)}\Big],$$

because $f'(x_* +) \leq f'(\xi(x)) \leq f'(x)$. □

Exercise 1.8. Consider the function $f : (0, \infty) \to \mathbb{R}$, $f(x) := x^k - \gamma$, with given constants $k > 1$ and $\gamma > 0$.

(a) Show that f is strictly isotonic and convex with unique zero at $x_* = \gamma^{1/k}$.

(b) Show that the corresponding algorithmic mapping $\psi(x) = x - f(x)/f'(x)$ for Newton's method has the following properties:

$$h(x) \;:=\; \frac{\psi(x) - x_*}{x - x_*}$$

is isotonic on (x_*, ∞) with limits $h(x_*) = 0$ and $h(\infty) = 1 - 1/k$.

Exercise 1.9. For $k \in \mathbb{N}$ let $F : [0, \infty) \to \mathbb{R}$ be given by

$$F(x) \;:=\; 1 - \exp(-x) \sum_{j=0}^{k} x^j / j! \,.$$

(a) Show that F is continuously differentiable and strictly isotonic on $[0, \infty)$ with $F(0) = 0$ and limit $F(\infty) = 1$.

(b) Describe Newton's procedure for the (approximate) solution of the equation $F(x) = \gamma$, where γ is any given number in $(0, 1)$.

(c) Specify a starting value $x_0 > 0$ such that Newton's method yields for arbitrary $\gamma \in (0, 1)$ a convergent sequence.

Stopping criteria. So far we analyzed the algorithmic mapping ψ underlying Newton's method. For explicit implementations we also need a stopping criterion involving some upper bound δ for the approximation error. Since we are looking for a zero of f, we could run the algorithm until $|f(x_n)| \le \delta$. Sometimes one combines this criterium with the requirement that $|x_n - x_{n-1}| \le \delta$.

Suppose we want to guarantee that $|x_n - x_*| \le \delta$. If we can show that $f'(x) \ge \kappa > 0$ for all x in an interval containing x_* and all points x_n, then it follows from

$$\frac{f(x_n)}{x_n - x_*} \;=\; \frac{f(x_n) - f(x_*)}{x_n - x_*} \;=\; f'(\xi_n)$$

with suitable $\xi_n \in (x_n \wedge x_*, x_n \vee x_*)$ that

$$|x_n - x_*| \;\le\; \frac{|f(x_n)|}{\kappa}.$$

Thus the stopping criterion $|f(x_n)| \le \delta\kappa$ guarantees that $|x_n - x_*| \le \delta$.

A simpler rule of thumb which is often implemented relies on the fact that in case of (1.1) also

$$\lim_{n\to\infty} \frac{x_n - x_*}{x_n - x_{n-1}} \;=\; 0,$$

see the next exercise. Consequently, if we iterate the mapping ψ until $|x_n - x_{n-1}|$ is smaller than a (very) small number δ, chances are high that $|x_n - x_*| \le \delta$, too.

Exercise 1.10 (Rates of convergence). Let $(\mathcal{X}, d)$ be a metric space and $(x_n)_{n=0}^{\infty}$ a sequence in $\mathcal{X}$ with limit $x_* \in \mathcal{X}$, that is, $\lim_{n\to\infty} d(x_n, x_*) = 0$.

(a) Suppose that

$$\lim_{n\to\infty} \frac{d(x_{n+1}, x_*)}{d(x_n, x_*)} \;=\; 0$$

(with the conventions that $0/0 := 0$ and $a/0 := \infty$ for $a > 0$). Show that

$$\lim_{n\to\infty} \frac{d(x_n, x_*)}{d(x_n, x_{n-1})} \;=\; 0$$

Show also that for any $\epsilon \in (0, 1)$ there exists a $K = K(\epsilon) \in (0, \infty)$ such that

$$d(x_n, x_*) \;\le\; K\epsilon^n \quad \text{for all } n \ge 0.$$

(b) Suppose that

$$\frac{d(x_{n+1}, x_*)}{d(x_n, x_*)^2} \;\le\; C \;<\; \infty \quad \text{for all } n \ge 0.$$

Show that for each $\epsilon \in (0,1)$ and $n, k \in \mathbb{N}_0$,

$$d(x_{n+k}, x_*) \;\le\; \epsilon^{(2^k)}/C \quad \text{if } d(x_n, x_*) \le \epsilon/C.$$

Exercise 1.11. For a given constant $c > 1$, let $f(x) := -\log(1-x) - cx$.

(a) Show that there is a unique point $x_* \in (0,1)$ such that $f(x_*) = 0$.

(b) Describe explicit programs to approximate x_* up to a given error $\delta \in (0,1)$ by means of a

(b.1) bisection method,
(b.2) Newton method.

Your descriptions should specify how to determine a starting interval or starting point, respectively, and provide a stopping criterion such that the approximation x_n of x_* satisfies $|x_n - x_*| \le \delta$.

Exercise 1.12. For a given parameter $c > 1$, let $f(x) := e^x - 1 - cx$.

(a) Show that there is a unique point $x_* > 0$ such that $f(x_*) = 0$.

(b) Show that $\log c < x_* < 2 \log c$.

(c) Show that $f'(x_*) = h(x_*) := e^{x_*} - (e^{x_*} - 1)/x_*$. Then show that h is strictly positive and strictly isotonic on $(0, \infty)$, and deduce from that the inequality $f'(x_*) > c - (c-1)/\log c > 0$.

(d) Describe an explicit Newton procedure to approximate x_*. For a given threshold $\delta > 0$, your procedure should return a number $x_{**} > 0$ such that $|x_{**} - x_*| \le \delta$ and $|f(x_{**})| \le \delta$.

Exercise 1.13. Suppose that f is convex on $\mathcal{X}$ but $f'(x)$ is concave in $x \ge x_*$. Show that

$$0 \;\le\; \frac{\psi(x) - x_*}{x - \psi(x)} \;<\; 1 \quad \text{for any } x > x_*.$$

For a Newton sequence $(x_n)_{n \ge 0}$ with starting value $x_0 \ge x_*$ this implies that

$$0 \;\le\; x_n - x_* \;\le\; x_{n-1} - x_n \quad \text{for arbitrary } n \ge 1.$$

Exercise 1.14. Consider independent random variables $X_1, X_2, \ldots, X_n \in [0,1]$ with density function

$$f_\theta(x) \;:=\; c(\theta)^{-1} \exp(\theta x), \quad x \in [0,1],$$

where $\theta \in \mathbb{R}$ is some unknown parameter and $c(\theta) > 0$ some norming constant.

(a) Compute $c(\theta)$ and verify that

$$\mu(\theta) := \mathbb{E}_\theta(X_1) \;=\; \frac{d}{d\theta} \log c(\theta).$$

(b) Show that μ is differentiable with

$$\mu'(\theta) \;=\; \mathrm{Var}_\theta(X_1) \;>\; 0$$

and verify that its limits are $\mu(-\infty) = 0$ and $\mu(\infty) = 1$.

(c) Show that $\mu(-\theta) = 1 - \mu(\theta)$ for all $\theta \in \mathbb{R}$.

(d) Show that $\mu(\theta)$ is convex in $\theta \in (-\infty, 0]$ and concave in $\theta \in [0, \infty)$.

(e) Show that the log-likelihood function

$$\theta \;\mapsto\; \sum_{i=1}^{n} \log f_\theta(X_i)$$

has a unique maximizer $\widehat{\theta} \in \mathbb{R}$ which is given by

$$\bar{X} \;=\; \mu(\widehat{\theta})$$

with the sample mean $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.

(f) Describe and implement an explicit procedure to compute $\widehat{\theta}$.

Hint: For small values of $|\theta|$ (e.g. $|\theta| < 0.01$) the function $\mu(\theta)$ may be numerically instable. (Why?) Use a suitable Taylor approximation of μ to avoid such problems.

1.3 Golden Section Search

We end this chapter with a simple procedure to minimize a "bathtub-shaped" function $f : [a, b] \to (-\infty, \infty]$. Precisely, suppose that there exists a unique minimizer x_* of f. Further let f be strictly antitonic on $[a, x_*]$ and strictly isotonic on $[x_*, b]$. Our goal is to find (an approximation of) x_* without referring to any derivative of f.

To localize the minimizer x_* we rely on the following simple consideration: Let $a \le r < s \le b$. Then it follows from $f(r) \le f(s)$ that $x_* < s$. For if $x_* \ge s$, then $f(r) > f(s)$, because f is strictly antitonic on $[a, x_*] \supset [r, s]$. Analogously it follows from $f(r) \ge f(s)$ that $x_* > r$.

This consideration gives rise to an iterative procedure: Let $a \le q < r < s < t \le b$, where it is known that $x_* \in [q, t]$. To fulfill the latter requirement, we start with $q = a$ and $t = b$. In case of $f(r) \le f(s)$ we know that $x_* \in [q, s]$, so we may replace the quadruple (q, r, s, t) with $(q, \dot{r}, r, s)$, where $q < \dot{r} < r$. In case of $f(r) > f(s)$ it is clear that $x_* \in [r, t]$, so we replace (q, r, s, t) with $(r, s, \dot{s}, t)$, where $s < \dot{s} < t$, see figure 1.1.

For the explicit choice of the new points $\dot{r}$ and $\dot{s}$ there are various possibilities. For instance one could take $\dot{r} = (q + r)/2$ and $\dot{s} = (s + t)/2$. Here is another proposal: We want to ensure that the three quadrupels (q, r, s, t), $(q, \dot{r}, r, s)$ and $(r, s, \dot{s}, t)$ are identical up to affine transformations. To this end we consider the ratios

$$B \;:=\; \frac{r - q}{t - q} \quad \text{and} \quad C \;:=\; \frac{s - q}{t - q},$$

so $0 < B < C < 1$ and

$$r = (1 - B)q + Bt, \quad s = (1 - C)q + Ct.$$

The quadrupels $(q, \dot{r}, r, s)$ and (q, r, s, t) have the same 'shape' if and only if

$$\frac{\dot{r} - q}{r - q} \;=\; \frac{r - q}{s - q} \;=\; \frac{B}{C} \quad \text{and} \quad \frac{r - q}{s - q} \;=\; \frac{s - q}{t - q} \;=\; C,$$

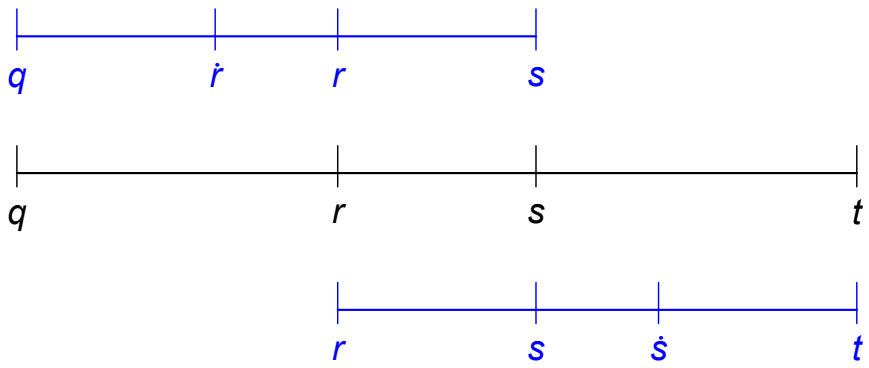

Figure 1.1: Partitions for Golden Section Search.

so

$$B \;=\; C^2 \quad \text{and} \quad \dot r \;=\; q + C(r-q) \;=\; (1-C)q + Cr.$$

The quadrupels $(r, s, \dot s, t)$ and (q, r, s, t) have the same shape if and only if

$$\frac{t-\dot s}{t-s} \;=\; \frac{t-s}{t-r} \;=\; \frac{1-C}{1-B} \quad \text{and} \quad \frac{t-s}{t-r} \;=\; \frac{t-r}{t-q} \;=\; 1-B,$$

so

$$1-C \;=\; (1-B)^2 \quad \text{and} \quad \dot s \;=\; t-(1-B)(t-s) \;=\; (1-B)s + Bt.$$

Plugging in $B = C^2$ into $1 - C = (1 - B)^2$ yields the equation

$$C^4 - 2C^2 + C \;=\; 0.$$

Obviously, $C = 0$ and $C = 1$ are two irrelevant solutions of this equation. Thus we write

$$C^4 - 2C^2 + C \;=\; C(C^3 - 2C + 1) \;=\; C(C-1)(C^2 + C - 1)$$

and determine a solution $C \in (0, 1)$ of the equation

$$0 \;=\; C^2 + C - 1 \;=\; (C + 1/2)^2 - 5/4.$$

This leads to

$$C \;=\; \frac{\sqrt{5}-1}{2} \;=\; \frac{2}{\sqrt{5}+1} \;\approx\; 0.618.$$

Then the corresponding value for $B = C^2 = 1 - C$ equals

$$B \;=\; \frac{3-\sqrt{5}}{2} \;=\; \frac{2}{3+\sqrt{5}} \;\approx\; 0.382.$$

In figure 1.1 we used these particular ratios B and C. Since $B + C = 1$, we may rewrite

$$r \;=\; Cq + Bt, \quad s = Bq + Ct, \quad \dot r \;=\; Bq + Cr, \quad \dot s \;=\; Cs + Bt.$$

These considerations lead to the algorithm in Table 1.5. It yields an interval $[q, t]$ of length at most $C\delta$ containing the minimizer x_*. Note that it terminates after at most

$$\left\lceil \log_{1/C}\Big(\frac{b-a}{\delta}\Big) \right\rceil$$

repetitions of the while-loop.

Algorithm $(q,t) \leftarrow \mathbf{GoldenSection}(f,a,b,\delta)$
$(B,C) \leftarrow (0.382, 0.618)$
$(q,r,s,t) \leftarrow (a, Ca+Bb, Ba+Cb, b)$
$(f_r, f_s) \leftarrow (f(r), f(s))$
while $t-q > \delta$ **do**
    if $f_r \leq f_s$ **then**
        $(r,s,t) \leftarrow (Bq+Cr, r, s)$
        $(f_r, f_s) \leftarrow (f(r), f_r)$
    else
        $(q,r,s) \leftarrow (r, s, Cs+Bt)$
        $(f_r, f_s) \leftarrow (f_s, f(s))$
    end if
end while
if $f_r \leq f_s$ **then**
    $t \leftarrow s$
else
    $q \leftarrow r$
end if.

Table 1.5: Golden section search.

Exercise 1.15. Implement the golden section search to approximate the minimizer of the function

$$x \;\mapsto\; f(x) \;:=\; (1-x)\log(1-x) + x - x^2$$

on $[0,1]$ with given precision $\delta > 0$. (Here $0\log(0) := 0$.) Why is golden section search justified here? What is your result in case of $\delta = 0.0001$?

Exercise 1.16. What can be said about the algorithm for golden section search if we assume only that f is

(a) continuous,

(b) differentiable,

(c) continuous and convex

on $[a,b]$?

Exercise 1.17. Suppose one replaces algorithm $\mathbf{GoldenSection}(f,a,b,\delta)$ with the algorithm in Table 1.6.

(a) Which ratios

$$(B,C) \;:=\; \Big(\frac{r-q}{t-q}, \frac{s-q}{t-q}\Big)$$

can occur while running this algorithm?

(b) In what sense is the new algorithm worse than golden section search?

Algorithm $(q,t) \leftarrow \mathbf{SilverSection}(f,a,b,\delta)$
$(q,r,s,t) \leftarrow (a, 2a/3+b/3, a/3+2b/3, b)$
$(f_r, f_s) \leftarrow (f(r), f(s))$
while $t-q > \delta$ **do**
    if $f_r \le f_s$ **then**
        $(r,s,t) \leftarrow ((q+r)/2, r, s)$
        $(f_r, f_s) \leftarrow (f(r), f_r)$
    else
        $(q,r,s) \leftarrow (r, s, (s+t)/2)$
        $(f_r, f_s) \leftarrow (f_s, f(s))$
    end if
end while
if $f_r \le f_s$ **then**
    $t \leftarrow s$
else
    $q \leftarrow r$
end if.

Table 1.6: 'Silver section' search.

Chapter 2

Convex Sets

In the current and next chapter we cover some basic concepts and results of convex analysis. In what follows we define and analyze properties of subsets of a real vector space $\boldsymbol{V}$.

2.1 Convex Sets and Cones

Definition 2.1 (Convex sets). A subset $\boldsymbol{C}$ of $\boldsymbol{V}$ is called *convex*, if for arbitrary points $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}$ their connecting line segment is contained in $\boldsymbol{C}$ as well:

$$\big\{(1-t)\boldsymbol{x} + t\boldsymbol{y} : t \in [0,1]\big\} \ \subset \ \boldsymbol{C}.$$

Via induction this implies that

$$\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \ \in \ \boldsymbol{C}$$

for arbitrary $n \in \mathbb{N}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{C}$ and $\lambda_1, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^n \lambda_i = 1$. Such a point $\sum_{i=1}^n \lambda_i \boldsymbol{x}_i$ is called a "convex combination of the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (with weights $\lambda_1, \ldots, \lambda_n$)".

Example 2.2. Suppose that $(\boldsymbol{V}, \|\cdot\|)$ is a normed vector space. Then any closed ball $\boldsymbol{B}(\boldsymbol{x}_o, \delta) := \{\boldsymbol{x} \in \boldsymbol{V} : \|\boldsymbol{x} - \boldsymbol{x}_o\| \leq \delta\}$ with center $\boldsymbol{x}_o \in \boldsymbol{V}$ and radius $\delta \geq 0$ is a convex set. Indeed, if $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{B}(\boldsymbol{x}_o, \delta)$ and $\lambda \in [0,1]$, then

$$\begin{aligned}
\big\|(1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y} \ - \ \boldsymbol{x}_o\big\| \ &= \ \big\|(1-\lambda)(\boldsymbol{x} - \boldsymbol{x}_o) + \lambda(\boldsymbol{y} - \boldsymbol{x}_o)\big\| \\
&\leq \ \big\|(1-\lambda)(\boldsymbol{x} - \boldsymbol{x}_o)\big\| + \big\|\lambda(\boldsymbol{y} - \boldsymbol{x}_o)\big\| \\
&= \ (1-\lambda)\|\boldsymbol{x} - \boldsymbol{x}_o\| + \lambda\|\boldsymbol{y} - \boldsymbol{x}_o\| \\
&\leq \ (1-\lambda)\delta + \lambda\delta \ = \ \delta.
\end{aligned}$$

With the same arguments one can show that any open ball is a convex set too.

Definition 2.3 (Convex cone). A subset $\boldsymbol{C}$ of $\boldsymbol{V}$ is called a *convex cone*, if for arbitrary points $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}$ the set of their linear combinations with nonnegative coefficients is contained in $\boldsymbol{C}$ as well:

$$\big\{\lambda\boldsymbol{x} + \mu\boldsymbol{y} : \lambda, \mu \geq 0\big\} \ \subset \ \boldsymbol{C}.$$

In other words, $\boldsymbol{C}$ is convex, and for arbitrary $\boldsymbol{x} \in \boldsymbol{C}$ the set of its nonnegative multiples is contained in $\boldsymbol{C}$:

$$\{\lambda \boldsymbol{x} : \lambda \geq 0\} \subset \boldsymbol{C}.$$

Again one can show via induction that

$$\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \in \boldsymbol{C}$$

for arbitrary $n \in \mathbb{N}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{C}$ and $\lambda_1, \ldots, \lambda_n \geq 0$.

Convex hulls etc. Now let $\boldsymbol{M}$ be an arbitrary nonvoid subset of $\boldsymbol{V}$. Then one can show easily that

$$\operatorname{conv}(\boldsymbol{M}) := \left\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \,:\, n \in \mathbb{N};\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{M};\, \lambda_1, \ldots, \lambda_n \geq 0;\, \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

is the smallest convex set containing $\boldsymbol{M}$. We call this set the *convex hull of* $\boldsymbol{M}$. Moreover,

$$\operatorname{cone}(\boldsymbol{M}) := \left\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \,:\, n \in \mathbb{N};\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{M};\, \lambda_1, \ldots, \lambda_n \geq 0 \right\}$$

is the smallest convex cone containing $\boldsymbol{M}$. We call it the *convex cone spanned by* $\boldsymbol{M}$. In case of a finite-dimensional space $\boldsymbol{V}$, these representations may be simplified as follows:

Lemma 2.4 (Carathéodory[1], 1907). *Let $d := \dim(\boldsymbol{V})$ be finite. Then for arbitrary subsets $\boldsymbol{M}$ of $\boldsymbol{V}$,*

$$\begin{aligned}
\operatorname{conv}(\boldsymbol{M}) &= \left\{ \sum_{i=1}^{d+1} \lambda_i \boldsymbol{x}_i \,:\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+1} \in \boldsymbol{M};\, \lambda_1, \ldots, \lambda_{d+1} \geq 0;\, \sum_{i=1}^{d+1} \lambda_i = 1 \right\}, \\
\operatorname{cone}(\boldsymbol{M}) &= \left\{ \sum_{i=1}^{d} \lambda_i \boldsymbol{x}_i \,:\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \boldsymbol{M};\, \lambda_1, \ldots, \lambda_d \geq 0 \right\}.
\end{aligned}$$

Proof of Lemma 2.4. We start with the representation of $\operatorname{cone}(\boldsymbol{M})$. Let $\boldsymbol{x} = \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i$ be a point in $\operatorname{cone}(\boldsymbol{M})$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{M}$, $\lambda_1, \ldots, \lambda_n > 0$ and $n > d$. Then the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are linearly dependent. That means,

$$\sum_{i=1}^{n} \mu_i \boldsymbol{x}_i = \boldsymbol{0}$$

for suitable numbers $\mu_1, \ldots, \mu_n$ such that $\mu_i \neq 0$ for at least one index i. Without loss of generatily let $\mu_i < 0$ for at least one index i. Then for arbitrary $t \geq 0$,

$$\boldsymbol{x} = \sum_{i=1}^{n} (\lambda_i + t\mu_i) \boldsymbol{x}_i.$$

[1] Constantin Carathéodory (1873–1950): Greek mathematician who spent most of his academic life in Germany with significant contributions to measure theory and other fields.

If we choose

$$t \;:=\; \min_{i\,:\,\mu_i<0} \frac{\lambda_i}{|\mu_i|},$$

then all coefficients $\widetilde{\lambda}_i := \lambda_i + t\mu_i$ are nonnegative, and at least one of them equals zero. Thus we may represent $\boldsymbol{x}$ as a linear combination of $n-1$ vectors in $\mathrm{cone}(\boldsymbol{M})$ with nonnegative coefficients. We may iterate this reduction until $\boldsymbol{x}$ is a linear combination of d vectors in $\mathrm{cone}(\boldsymbol{M})$ with nonnegative coefficients.

As to $\mathrm{conv}(\boldsymbol{M})$, note that the equations $\boldsymbol{x} = \sum_{i=1}^n \lambda_i \boldsymbol{x}_i$ and $\sum_{i=1}^n \lambda_i = 1$ with $\boldsymbol{x}_i \in \boldsymbol{M}$ and $\lambda_i > 0$ may be combined to one equation

$$(\boldsymbol{x}, 1) \;=\; \sum_{i=1}^n \lambda_i (\boldsymbol{x}_i, 1)$$

involving points in the $(d+1)$-dimensional linear space $\boldsymbol{V} \times \mathbb{R}$. If $n > d+1$, we may imitate the arguments for $\mathrm{cone}(\boldsymbol{M})$ to show that $(\boldsymbol{x}, 1)$ is a linear combination of $n-1$ points in $\boldsymbol{M} \times \{1\}$ with non-negative weights. Again iteration of this argument shows that $(\boldsymbol{x}, 1)$ is a linear combination of $d+1$ points in $\boldsymbol{M} \times \{1\}$ with non-negative weights. In other words, $\boldsymbol{x}$ is a convex combination of $d+1$ points in $\boldsymbol{M}$. □

Exercise 2.5. Determine the convex hull of $\boldsymbol{M} \subset \mathbb{R} \times \mathbb{R}$ in the following cases:

(a) $\boldsymbol{M} = \{(x, 1/|x|) : x \in \mathbb{R}, x \neq 0\}$.

(b) $\boldsymbol{M} = \{(x, 1/x) : x \in \mathbb{R}, x \neq 0\}$.

Exercise 2.6 (Linear mappings and convexity). Let $L : \boldsymbol{V} \to \boldsymbol{W}$ be a linear mapping from $\boldsymbol{V}$ into another real vector space $\boldsymbol{W}$.

(a) Show that $L(\boldsymbol{C})$ is convex for any convex set $\boldsymbol{C} \subset \boldsymbol{V}$.

(b) Show that $L^{-1}(\boldsymbol{D})$ is convex for any convex set $\boldsymbol{D} \subset \boldsymbol{W}$.

Exercise 2.7. Let $(\boldsymbol{V}, \|\cdot\|)$ be a real normed vector space and $\boldsymbol{M}$ be a nonvoid subset of $\boldsymbol{V}$. Prove or falsify the following claims:

(a) If $\boldsymbol{M}$ is open, then $\mathrm{conv}(\boldsymbol{M})$ is open, too.

(b) If $\boldsymbol{M}$ is closed, then $\mathrm{conv}(\boldsymbol{M})$ is closed, too.

(c) If $\boldsymbol{M}$ is compact and $\dim(\boldsymbol{V}) < \infty$, then $\mathrm{conv}(\boldsymbol{M})$ is compact, too.

Exercise 2.8. Let $\boldsymbol{C}$ be a convex subset of a real normed vector space $(\boldsymbol{V}, \|\cdot\|)$. Show that its closure $\overline{\boldsymbol{C}}$ is convex, too.

Exercise 2.9 (Minkowski operations). For $n \in \mathbb{N}$, nonvoid sets $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_n \subset \boldsymbol{V}$ and real numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ let

$$\sum_{i=1}^n \lambda_i \boldsymbol{C}_i \;:=\; \Big\{ \sum_{i=1}^n \lambda_i \boldsymbol{x}_i : \boldsymbol{x_i} \in \boldsymbol{C}_i \text{ for } 1 \le i \le n \Big\}.$$

Especially for $\boldsymbol{x} \in \boldsymbol{V}$ and $\boldsymbol{C} \subset \boldsymbol{V}$ we set $\boldsymbol{x} + \boldsymbol{C} := \{\boldsymbol{x}\} + \boldsymbol{C}$.

(a) Show that $\sum_{i=1}^n \lambda_i \boldsymbol{C}_i$ is convex if all sets $\boldsymbol{C}_i$ are convex.

(b) Show that for a nonvoid convex set $\boldsymbol{C} \subset \boldsymbol{V}$ and nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\sum_{i=1}^{n}(\lambda_i \boldsymbol{C}) \;=\; \Big(\sum_{i=1}^{n}\lambda_i\Big)\boldsymbol{C}.$$

Exercise 2.10. Let $\boldsymbol{M}$ be a subset of $\mathbb{R}^d$.

(a) Suppose that $\boldsymbol{M}$ is finite. Show that $\mathrm{cone}(\boldsymbol{M})$ is closed. (Hint: Consider first the situation that the elements of $\boldsymbol{M}$ are linearly independent. Then argue as in the proof of Lemma 2.4.)

(b) Find an example of a compact set $\boldsymbol{M} \subset \mathbb{R}^2$ such that $\mathrm{cone}(\boldsymbol{M})$ is not closed.

2.2 Metric Projections and Separating Hyperplanes

In this section let $(\boldsymbol{V}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be a real Hilbert space, for instance, $\boldsymbol{V} = \mathbb{R}^d$ with the standard inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \boldsymbol{x}^\top \boldsymbol{y}$. The closed unit ball is denoted by $\boldsymbol{B}$,

$$\boldsymbol{B} \;=\; \{\boldsymbol{x} \in \boldsymbol{V} : \|\boldsymbol{x}\| \le 1\}.$$

Throughout this section let $\boldsymbol{C}$ be a nonvoid convex and closed subset of $\boldsymbol{V}$.

2.2.1 Metric projections

The next theorem shows that for any point $\boldsymbol{x} \in \boldsymbol{V}$ there exists a unique closest point $\Pi\boldsymbol{x} = \Pi_{\boldsymbol{C}}\boldsymbol{x}$ in $\boldsymbol{C}$, that means,

$$\|\boldsymbol{x} - \Pi\boldsymbol{x}\| \;=\; d(\boldsymbol{x}, \boldsymbol{C}) := \inf_{\boldsymbol{y}\in\boldsymbol{C}} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

The corresponding mapping $\Pi : \boldsymbol{V} \to \boldsymbol{C}$ is called the *metric projection of $\boldsymbol{V}$ onto $\boldsymbol{C}$*. It is Lipschitz continuous with constant one:

Theorem 2.11 (Metric projection onto a closed convex set).

(a) *To each point $\boldsymbol{x} \in \boldsymbol{V}$ there exists a unique point $\Pi\boldsymbol{x} \in \boldsymbol{C}$ such that*

$$\|\boldsymbol{x} - \Pi\boldsymbol{x}\| \;=\; \min_{\boldsymbol{y}\in\boldsymbol{C}} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

(b) *A point $\boldsymbol{y}_o \in \boldsymbol{C}$ equals $\Pi\boldsymbol{x}$ if and only if*

$$\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} - \boldsymbol{y}_o\rangle \;\le\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C}.$$

(c) *For arbitrary $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{V}$,*

$$\|\Pi\boldsymbol{x}_1 - \Pi\boldsymbol{x}_2\| \;\le\; \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

In case of closed convex cones or closed linear subspaces, the characterization in part (b) of Theorem 2.11 may be simplified somewhat, and the metric projection Π has additional properties:

Corollary 2.12 (Metric projection onto a closed convex cone). *Let $\boldsymbol{C}$ be a closed convex cone. Then a point $\boldsymbol{y}_o \in \boldsymbol{C}$ equals $\Pi\boldsymbol{x}$ if and only if*

$$\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y}_o \rangle \;=\; 0 \quad \text{and} \quad \langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} \rangle \;\le\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C}.$$

Moreover, $\Pi(\lambda\boldsymbol{x}) = \lambda\Pi\boldsymbol{x}$ for arbitrary $\boldsymbol{x} \in \boldsymbol{V}$ and $\lambda \ge 0$.

Corollary 2.13 (Metric projection onto a closed linear space). *Let $\boldsymbol{C}$ be a closed linear subspace of $\boldsymbol{V}$. Then a point $\boldsymbol{y}_o \in \boldsymbol{C}$ equals $\Pi\boldsymbol{x}$ if and only if*

$$\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} \rangle \;=\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C}.$$

Moreover, $\Pi : \boldsymbol{V} \to \boldsymbol{C}$ is a linear mapping, the so-called orthogonal projection from $\boldsymbol{V}$ onto $\boldsymbol{C}$.

Proof of Theorem 2.11. *Proof of part (a).* We first show that the distance between two points $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \boldsymbol{C}$ has to be rather small if both $\|\boldsymbol{x} - \boldsymbol{y}_1\|$ and $\|\boldsymbol{x} - \boldsymbol{y}_2\|$ are close to $d(\boldsymbol{x}, \boldsymbol{C})$. To this end we use the parallelogram identity,

$$\|\boldsymbol{a} + \boldsymbol{b}\|^2 + \|\boldsymbol{a} - \boldsymbol{b}\|^2 \;=\; 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2 \quad \text{für } \boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{V},$$

which is a consequence of the equations $\|\boldsymbol{a} \pm \boldsymbol{b}\|^2 \;=\; \|\boldsymbol{a}\|^2 \pm 2\langle \boldsymbol{a}, \boldsymbol{b} \rangle + \|\boldsymbol{b}\|^2$. This identity is applied to $\boldsymbol{a} = \boldsymbol{x} - \boldsymbol{y}_1$ and $\boldsymbol{b} = \boldsymbol{x} - \boldsymbol{y}_2$, and we utilize the fact that due to convexity of $\boldsymbol{C}$,

$$\boldsymbol{y}_o \;:=\; 2^{-1}(\boldsymbol{y}_1 + \boldsymbol{y}_2)$$

belongs to $\boldsymbol{C}$, too. Thus

$$\begin{aligned}
\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|^2 \;&=\; \big\|(\boldsymbol{x} - \boldsymbol{y}_1) - (\boldsymbol{x} - \boldsymbol{y}_2)\big\|^2 \\
&=\; 2\|\boldsymbol{x} - \boldsymbol{y}_1\|^2 + 2\|\boldsymbol{x} - \boldsymbol{y}_2\|^2 - \big\|(\boldsymbol{x} - \boldsymbol{y}_1) + (\boldsymbol{x} - \boldsymbol{y}_2)\big\|^2 \\
&=\; 2\|\boldsymbol{x} - \boldsymbol{y}_1\|^2 + 2\|\boldsymbol{x} - \boldsymbol{y}_2\|^2 - 4\|\boldsymbol{x} - \boldsymbol{y}_o\|^2 \\
&\le\; 2\|\boldsymbol{x} - \boldsymbol{y}_1\|^2 + 2\|\boldsymbol{x} - \boldsymbol{y}_2\|^2 - 4d(\boldsymbol{x}, \boldsymbol{C})^2,
\end{aligned}$$

whence

$$\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|^2 \;\le\; 2\big(\|\boldsymbol{x} - \boldsymbol{y}_1\|^2 - d(\boldsymbol{x}, \boldsymbol{C})^2\big) + 2\big(\|\boldsymbol{x} - \boldsymbol{y}_2\|^2 - d(\boldsymbol{x}, \boldsymbol{C})^2\big).$$

This inequality shows already that there is at most one point $\boldsymbol{y} \in \boldsymbol{C}$ such that $\|\boldsymbol{x} - \boldsymbol{y}\| = d(\boldsymbol{x}, \boldsymbol{C})$. It also implies existence of such a point. To see this, let $(\boldsymbol{y}_n)_{n=1}^{\infty}$ be a sequence in $\boldsymbol{C}$ such that

$$\lim_{n\to\infty} \|\boldsymbol{x} - \boldsymbol{y}_n\| \;=\; d(\boldsymbol{x}, \boldsymbol{C}).$$

For $N \in \mathbb{N}$,

$$\begin{aligned}
\sup_{m,n\ge N} \|\boldsymbol{y}_m - \boldsymbol{y}_n\|^2 \;&\le\; 4\Big(\sup_{n\ge N} \|\boldsymbol{x} - \boldsymbol{y}_n\|^2 - d(\boldsymbol{x}, \boldsymbol{C})^2\Big) \\
&\to\; 0 \quad \text{as } N \to \infty.
\end{aligned}$$

Hence $(\boldsymbol{y}_n)_{n=1}^{\infty}$ is a Cauchy sequence in $\boldsymbol{V}$. Since $(\boldsymbol{V}, \|\cdot\|)$ is complete and $\boldsymbol{C}$ is closed, this sequence has a limit $\boldsymbol{y} \in \boldsymbol{C}$, and $\|\boldsymbol{x} - \boldsymbol{y}\| = d(\boldsymbol{x}, \boldsymbol{C})$.

Proof of part (b). For any two points $\boldsymbol{y}_o, \boldsymbol{y} \in \boldsymbol{C}$ and $t \in [0,1]$, the point $\boldsymbol{y}_t := (1-t)\boldsymbol{y}_o + t\boldsymbol{y}$ belongs to $\boldsymbol{C}$, too, and

$$\|\boldsymbol{x} - \boldsymbol{y}_t\|^2 \;=\; \|\boldsymbol{x} - \boldsymbol{y}_o - t(\boldsymbol{y} - \boldsymbol{y}_o)\|^2 \;=\; \|\boldsymbol{x} - \boldsymbol{y}_o\|^2 - 2t\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} - \boldsymbol{y}_o\rangle + t^2\|\boldsymbol{y} - \boldsymbol{y}_o\|^2.$$

The right hand side is greater than or equal to $\|\boldsymbol{x} - \boldsymbol{y}_o\|^2$ for arbitrary $t \in [0,1]$ if and only if

$$\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} - \boldsymbol{y}_o\rangle \;\le\; 0.$$

Proof of part (c). According to part (b),

$$\begin{aligned}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \;&= \big\|(\Pi\boldsymbol{x}_1 - \Pi\boldsymbol{x}_2) + (\boldsymbol{x}_1 - \Pi\boldsymbol{x}_1 - \boldsymbol{x}_2 + \Pi\boldsymbol{x}_2)\big\|^2 \\ &\ge \|\Pi\boldsymbol{x}_1 - \Pi\boldsymbol{x}_2\|^2 \\ &\quad -\; 2\langle \boldsymbol{x}_1 - \Pi\boldsymbol{x}_1, \Pi\boldsymbol{x}_2 - \Pi\boldsymbol{x}_1\rangle - 2\langle \boldsymbol{x}_2 - \Pi\boldsymbol{x}_2, \Pi\boldsymbol{x}_1 - \Pi\boldsymbol{x}_2\rangle \\ &\ge \|\Pi\boldsymbol{x}_1 - \Pi\boldsymbol{x}_2\|^2.\end{aligned}$$

Hence Π is Lipschitz continuous with constant 1. □

Proof of Corollary 2.12. Suppose that a point $\boldsymbol{y}_o \in \boldsymbol{C}$ satisfies the equation $\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y}_o\rangle = 0$ and the inequality $\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y}\rangle \le 0$ for any $\boldsymbol{y} \in \boldsymbol{C}$. Then

$$\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} - \boldsymbol{y}_o\rangle \;=\; \langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y}\rangle - \langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y}_o\rangle \;\le\; 0$$

for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$. Consequently, according to Theorem 2.11 (b), $\boldsymbol{y}_o = \Pi\boldsymbol{x}$.

On the other hand, since $\boldsymbol{C}$ is a convex cone, for any number $t \ge 0$ the point $t\,\Pi\boldsymbol{x}$ belongs to $\boldsymbol{C}$, too, whence

$$0 \;=\; \frac{d}{dt}\bigg|_{t=1} \|\boldsymbol{x} - t\,\Pi\boldsymbol{x}\|^2 \;=\; -2\langle \boldsymbol{x}, \Pi\boldsymbol{x}\rangle + 2\|\Pi\boldsymbol{x}\|^2 \;=\; -2\langle \boldsymbol{x} - \Pi\boldsymbol{x}, \Pi\boldsymbol{x}\rangle.$$

Furthermore, for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$, the sum $\Pi\boldsymbol{x} + \boldsymbol{y}$ is an element of $\boldsymbol{C}$, too, whence Theorem 2.11 (b) implies that

$$0 \;\ge\; \langle \boldsymbol{x} - \Pi\boldsymbol{x}, (\Pi\boldsymbol{x} + \boldsymbol{y}) - \Pi\boldsymbol{x}\rangle \;=\; \langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{y}\rangle.$$

For any $\boldsymbol{x} \in \boldsymbol{V}$ and $\lambda \ge 0$, the point $\lambda\Pi\boldsymbol{x} \in \boldsymbol{C}$ satisfies the (in)equalities

$$\begin{aligned}\langle \lambda\boldsymbol{x} - \lambda\Pi\boldsymbol{x}, \lambda\Pi\boldsymbol{x}\rangle \;&=\; \lambda^2\langle \boldsymbol{x} - \Pi\boldsymbol{x}, \Pi\boldsymbol{x}\rangle \;=\; 0, \\ \langle \lambda\boldsymbol{x} - \lambda\Pi\boldsymbol{x}, \boldsymbol{y}\rangle \;&=\; \lambda\langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{y}\rangle \;\le\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C}.\end{aligned}$$

Hence $\Pi(\lambda\boldsymbol{x}) = \lambda\Pi\boldsymbol{x}$. □

Proof of Corollary 2.13. Suppose that a point $\boldsymbol{y}_o \in \boldsymbol{C}$ satisfies the equation $\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{z}\rangle = 0$ for arbitrary $\boldsymbol{z} \in \boldsymbol{C}$. Then $\langle \boldsymbol{x} - \boldsymbol{y}_o, \boldsymbol{y} - \boldsymbol{y}_o\rangle = 0$ for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$, and Theorem 2.11 (b) shows that $\boldsymbol{y}_o = \Pi\boldsymbol{x}$.

On the other hand, $\boldsymbol{z} := \Pi\boldsymbol{x} \pm \boldsymbol{y} \in \boldsymbol{C}$ for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$, whence Theorem 2.11 (b) implies that

$$0 \;\geq\; \langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{z} - \Pi\boldsymbol{x} \rangle \;=\; \langle \boldsymbol{x} - \Pi\boldsymbol{x}, \pm\boldsymbol{y} \rangle \;=\; \pm\langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{y} \rangle.$$

Consequently, $\langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{y} \rangle = 0$ for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$.

For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{V}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$\big\langle (\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2) - (\lambda_1\Pi\boldsymbol{x}_1 + \lambda_2\Pi\boldsymbol{x}_2), \boldsymbol{y} \big\rangle \;=\; \lambda_1\langle \boldsymbol{x}_1 - \Pi\boldsymbol{x}_1, \boldsymbol{y} \rangle + \lambda_2\langle \boldsymbol{x}_2 - \Pi\boldsymbol{x}_2, \boldsymbol{y} \rangle \;=\; 0.$$

Hence $\Pi(\lambda_1\boldsymbol{x}_1 + \lambda\boldsymbol{x}_2) = \lambda_1\Pi\boldsymbol{x}_1 + \lambda_2\Pi\boldsymbol{x}_2$. In other words, Π is a linear mapping from $\boldsymbol{V}$ onto $\boldsymbol{C}$. □

Exercise 2.14 (Distance functions). Let A be a nonvoid subset of a metric space $(\mathcal{X}, d)$. For $y \in \mathcal{X}$ let

$$d(y, A) \;:=\; \inf_{x \in A} d(x, y).$$

Show that

(a) $d(y, A) = 0$ if and only if y is in the closure $\overline{A}$ of A,
(b) $d(y, A) = d(y, \overline{A})$,
(c) $\big|d(y, A) - d(z, A)\big| \leq d(y, z)$ for arbitrary $y, z \in \mathcal{X}$.

Exercise 2.15. Let $\boldsymbol{V} = \mathbb{R}^4$ and $\boldsymbol{C} := \{\boldsymbol{x} \in \mathbb{R}^4 : x_1 \leq x_2 \leq x_3 \leq x_4\}$. Show that $\boldsymbol{C}$ is a closed convex cone. Guess the projection $\Pi\boldsymbol{x}$ of the vector $\boldsymbol{x} := (3, 1, 5, 3)^\top$ onto $\boldsymbol{C}$. Verify your guess by checking the conditions in Corollary 2.12.

Exercise 2.16 (Riesz'[2] representation theorem in Hilbert spaces). By means of Corollary 2.13 one can easily prove a key result in functional analysis: Let $L : \boldsymbol{V} \to \mathbb{R}$ be a continuous linear function. Then there exists a unique vector $\boldsymbol{v} \in \boldsymbol{V}$ such that

$$L(\boldsymbol{x}) \;=\; \langle \boldsymbol{x}, \boldsymbol{v} \rangle \quad \text{for all } \boldsymbol{x} \in \boldsymbol{V}.$$

Prove this result along the following lines:

(a) Show that $\boldsymbol{C} := \{\boldsymbol{x} \in \boldsymbol{V} : L(\boldsymbol{x}) = 0\}$ is a closed linear subspace of $\boldsymbol{V}$.

(b) If $L \not\equiv 0$, there exists a $\boldsymbol{x}_1 \in \boldsymbol{V}$ such that $L(\boldsymbol{x}_1) = 1$. Now consider the vector $\boldsymbol{v} := \|\boldsymbol{x}_1 - \Pi\boldsymbol{x}_1\|^{-2}(\boldsymbol{x}_1 - \Pi\boldsymbol{x}_1)$ which satisifes $\boldsymbol{v} \perp \boldsymbol{C}$ by Corollary 2.13.

(b.1) Verify that $\langle \boldsymbol{x}_1, \boldsymbol{v} \rangle = 1$.

(b.2) Show that for any $\boldsymbol{x} \in \boldsymbol{V}$, the vector $\boldsymbol{x} - L(\boldsymbol{x})\boldsymbol{x}_1$ belongs to $\boldsymbol{C}$. Deduce from the latter fact that $L(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{v} \rangle$.

(c) Verify uniqueness of $\boldsymbol{v}$.

[2]Frigyes Riesz (1880–1956): Hungarian mathematician who made fundamental contributions to functional analysis.

2.2.2 Support functions and half spaces

Starting from Theorem 2.11 one can deduce various properties of convex sets. At first we show that a closed concex subset of $\boldsymbol{V}$ is always an intersection of closed halfspaces. A set $\boldsymbol{H} \subset \boldsymbol{V}$ is called a closed halfspace if

$$\boldsymbol{H} \;=\; \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle \le r\big\}$$

for some $\boldsymbol{v} \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}$ and $r \in \mathbb{R}$. One can easily verify that this set $\boldsymbol{H}$ is closed and convex, see also Exercise 2.6. Its boundary is the hyperplane

$$\big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle = r\big\} \;=\; \frac{r}{\|\boldsymbol{v}\|^2}\,\boldsymbol{v} + \boldsymbol{v}^\perp$$

with the orthogonal complement

$$\boldsymbol{v}^\perp \;=\; \big\{\boldsymbol{y} \in \boldsymbol{V} : \langle \boldsymbol{y}, \boldsymbol{v}\rangle = 0\big\}$$

of $\boldsymbol{v}$, a closed linear subspace of $\boldsymbol{V}$.

Definition 2.17 (Support function). The *support function* of a nonvoid set $\boldsymbol{M} \subset \boldsymbol{V}$ is defined as

$$h_{\boldsymbol{M}} : \boldsymbol{V} \to (-\infty, \infty], \quad h_{\boldsymbol{M}}(\boldsymbol{v}) \;:=\; \sup_{\boldsymbol{x}\in\boldsymbol{M}} \langle \boldsymbol{x}, \boldsymbol{v}\rangle.$$

Exercise 2.18 (Support functions of convex hulls). Show that

$$h_{\boldsymbol{M}} \;\equiv\; h_{\overline{\boldsymbol{M}}} \quad \text{and} \quad h_{\boldsymbol{M}} \;\equiv\; h_{\mathrm{conv}(\boldsymbol{M})},$$

where $\overline{\boldsymbol{M}}$ denotes the closure of $\boldsymbol{M}$. In particular, the support function of $\boldsymbol{M}$ coincides with the support function of $\overline{\mathrm{conv}(\boldsymbol{M})}$, which is the smallest closed and convex set containing $\boldsymbol{M}$, see also Exercise 2.8.

Recall that $\boldsymbol{C}$ is a nonvoid closed and convex subset of $\boldsymbol{V}$ and that $\boldsymbol{B}$ is the closed unit ball in $\boldsymbol{V}$, so $\partial\boldsymbol{B}$ is the unit sphere $\{\boldsymbol{u} \in \boldsymbol{V} : \|\boldsymbol{u}\| = 1\}$.

Corollary 2.19 (Support functions and halfspaces). *The support function $h = h_{\boldsymbol{C}}$ of $\boldsymbol{C}$ has the following properties:*

$$\boldsymbol{C} \;=\; \bigcap_{\boldsymbol{v}\in\boldsymbol{V}} \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle \le h(\boldsymbol{v})\big\} \;=\; \bigcap_{\boldsymbol{u}\in\partial\boldsymbol{B}} \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{u}\rangle \le h(\boldsymbol{u})\big\}.$$

For $\boldsymbol{x} \in \boldsymbol{V} \setminus \boldsymbol{C}$,

$$d(\boldsymbol{x}, \boldsymbol{C}) \;=\; \max_{\boldsymbol{u}\in\partial\boldsymbol{B}} \big(\langle \boldsymbol{x}, \boldsymbol{u}\rangle - h(\boldsymbol{u})\big).$$

Note that the set $\big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle \le h(\boldsymbol{v})\big\}$ is a closed halfspace in $\boldsymbol{V}$ if $\boldsymbol{v} \neq \boldsymbol{0}$ and $h(\boldsymbol{v}) < \infty$. Otherwise it is the full space $\boldsymbol{V}$.

Proof of Corollary 2.19. Obviously, the set $\boldsymbol{C}$ is contained in any set $\big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle \le h(\boldsymbol{v})\big\}$, so

$$\boldsymbol{C} \;\subset\; \bigcap_{\boldsymbol{v}\in\boldsymbol{V}} \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{v}\rangle \le h(\boldsymbol{v})\big\} \;\subset\; \bigcap_{\boldsymbol{u}\in\partial\boldsymbol{B}} \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{u}\rangle \le h(\boldsymbol{u})\big\}.$$

To show that the latter set coincides with $\boldsymbol{C}$, it suffices to verify the particular representation of $d(\boldsymbol{x}, \boldsymbol{C})$ for $\boldsymbol{x} \in \boldsymbol{V} \setminus \boldsymbol{C}$. To this end we consider the metric projection $\Pi\boldsymbol{x}$ of $\boldsymbol{x}$ onto $\boldsymbol{C}$. For any unit vector $\boldsymbol{u}$ in $\boldsymbol{V}$, the definition of $h(\boldsymbol{u})$ and the Cauchy–Schwarz inequality imply that

$$\langle \boldsymbol{x}, \boldsymbol{u} \rangle - h(\boldsymbol{u}) \;\leq\; \langle \boldsymbol{x}, \boldsymbol{u} \rangle - \langle \Pi\boldsymbol{x}, \boldsymbol{u} \rangle \;=\; \langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{u} \rangle \;\leq\; \|\boldsymbol{x} - \Pi\boldsymbol{x}\| \;=\; d(\boldsymbol{x}, \boldsymbol{C}).$$

Specifically let $\boldsymbol{u} = \|\boldsymbol{x} - \Pi\boldsymbol{x}\|^{-1}(\boldsymbol{x} - \Pi\boldsymbol{x})$. Then it follows from Theorem 2.11 (b) that

$$\langle \boldsymbol{y} - \Pi\boldsymbol{x}, \boldsymbol{u} \rangle \;=\; \|\boldsymbol{x} - \Pi\boldsymbol{x}\|^{-1} \langle \boldsymbol{y} - \Pi\boldsymbol{x}, \boldsymbol{x} - \Pi\boldsymbol{x} \rangle \;\leq\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C}.$$

In other words, $\langle \boldsymbol{y}, \boldsymbol{u} \rangle \leq \langle \Pi\boldsymbol{x}, \boldsymbol{u} \rangle$ for all $\boldsymbol{y} \in \boldsymbol{C}$, whence

$$h(\boldsymbol{u}) \;=\; \sup_{\boldsymbol{y} \in \boldsymbol{C}} \langle \boldsymbol{y}, \boldsymbol{u} \rangle \;=\; \langle \Pi\boldsymbol{x}, \boldsymbol{u} \rangle,$$

and

$$\langle \boldsymbol{x}, \boldsymbol{u} \rangle \;=\; \langle \Pi\boldsymbol{x}, \boldsymbol{u} \rangle + \langle \boldsymbol{x} - \Pi\boldsymbol{x}, \boldsymbol{u} \rangle \;=\; h(\boldsymbol{u}) + \|\boldsymbol{x} - \Pi\boldsymbol{x}\|. \qquad \square$$

Exercise 2.20 (Support functions and interior points). Show that for any $\boldsymbol{x} \in \boldsymbol{V}$ and $\delta > 0$ the following two statements are equivalent:

(i) The closed ball $\boldsymbol{x} + \delta\boldsymbol{B} = \{\boldsymbol{x} + \delta\boldsymbol{v} : \boldsymbol{v} \in \boldsymbol{B}\}$ with center $\boldsymbol{x}$ and radius δ is contained in $\boldsymbol{C}$;

(ii) $h(\boldsymbol{u}) \geq \langle \boldsymbol{x}, \boldsymbol{u} \rangle + \delta$ for any unit vector $\boldsymbol{u}$ in $\boldsymbol{V}$.

2.2.3 Separating hyperplanes

The next consequence of Theorem 2.11 is that for any point on the boundary of $\boldsymbol{C}$ there exists a tangent hyperplane, at least in case of $\boldsymbol{V}$ being finite-dimensional.

Corollary 2.21 (Tangent hyperplanes). *Let $\dim(\boldsymbol{V}) < \infty$. For each $\boldsymbol{x} \in \partial\boldsymbol{C}$ there exists a unit vector $\boldsymbol{u} \in \boldsymbol{V}$ such that $\langle \boldsymbol{x}, \boldsymbol{u} \rangle \geq \langle \boldsymbol{y}, \boldsymbol{u} \rangle$ for all $\boldsymbol{y} \in \boldsymbol{C}$.*

Proof of Corollary 2.21. Let $(\boldsymbol{x}_n)_{n=1}^{\infty}$ be a sequence in $\boldsymbol{V} \setminus \boldsymbol{C}$ with limit $\boldsymbol{x}$. We define

$$\boldsymbol{u}_n \;:=\; \|\boldsymbol{x}_n - \Pi\boldsymbol{x}_n\|^{-1}(\boldsymbol{x}_n - \Pi\boldsymbol{x}_n).$$

Then $(\boldsymbol{u}_n)_{n=1}^{\infty}$ is a sequence of unit vectors such that

$$\langle \boldsymbol{y} - \boldsymbol{x}_n, \boldsymbol{u}_n \rangle \;=\; \langle \boldsymbol{y} - \Pi\boldsymbol{x}_n, \boldsymbol{u}_n \rangle - \|\boldsymbol{x}_n - \Pi\boldsymbol{x}_n\| \;\leq\; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{C},$$

see Theorem 2.11 (b). Since the unit sphere of a finite-dimensional Hilbert space is compact, we may assume without loss of generality that $(\boldsymbol{u}_n)_{n=1}^{\infty}$ converges to some unit vector $\boldsymbol{u}$. But then

$$\langle \boldsymbol{y}, \boldsymbol{u} \rangle - \langle \boldsymbol{x}, \boldsymbol{u} \rangle \;=\; \langle \boldsymbol{y} - \boldsymbol{x}, \boldsymbol{u} \rangle \;=\; \lim_{n \to \infty} \langle \boldsymbol{y} - \boldsymbol{x}_n, \boldsymbol{u}_n \rangle \;\leq\; 0$$

for any $\boldsymbol{y} \in \boldsymbol{C}$. $\square$

Finally we show that two disjoint convex subsets can often be separated by a hyperplane:

Theorem 2.22 (Separating hyperplanes). *Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_1$ be nonvoid convex subsets of $\boldsymbol{V}$ such that $\boldsymbol{D}_1 \cap \boldsymbol{D}_2 = \emptyset$.*

(a) *If $\boldsymbol{D}_1$ is compact and $\boldsymbol{D}_2$ is closed, then there exists a unit vector $\boldsymbol{u} \in \boldsymbol{V}$ such that*

$$\sup_{\boldsymbol{x} \in \boldsymbol{D}_1} \langle \boldsymbol{x}, \boldsymbol{u} \rangle \; < \; \inf_{\boldsymbol{y} \in \boldsymbol{D}_2} \langle \boldsymbol{y}, \boldsymbol{u} \rangle.$$

(b) *If $\dim(\boldsymbol{V}) < \infty$, then there exists a unit vector $\boldsymbol{u} \in \boldsymbol{V}$ such that*

$$\sup_{\boldsymbol{x} \in \boldsymbol{D}_1} \langle \boldsymbol{x}, \boldsymbol{u} \rangle \; \leq \; \inf_{\boldsymbol{y} \in \boldsymbol{D}_2} \langle \boldsymbol{y}, \boldsymbol{u} \rangle.$$

Proof of Theorem 2.22. We start with part (a). As shown in Exercise 2.14, the mapping $\boldsymbol{x} \mapsto d(\boldsymbol{x}, \boldsymbol{D}_2)$ is continuous. Hence compactness of $\boldsymbol{D}_1$ implies existence of a point $\boldsymbol{a} \in \boldsymbol{D}_1$ such that $d(\boldsymbol{a}, \boldsymbol{D}_2) \leq d(\boldsymbol{x}, \boldsymbol{D}_2)$ for all $\boldsymbol{x} \in \boldsymbol{D}_1$. If we denote the metric projection of $\boldsymbol{a}$ onto $\boldsymbol{D}_2$ with $\boldsymbol{b}$, then

$$0 \; < \; \|\boldsymbol{b} - \boldsymbol{a}\| \; = \; \inf_{\boldsymbol{x} \in \boldsymbol{D}_1, \, \boldsymbol{y} \in \boldsymbol{D}_2} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

In particular, $\boldsymbol{a}$ is the metric projection of $\boldsymbol{b}$ onto $\boldsymbol{D}_1$. Hence it follows from Theorem 2.11 (b) applied to $\boldsymbol{C} = \boldsymbol{D}_1, \boldsymbol{D}_2$ that the unit vector

$$\boldsymbol{u} \; := \; \|\boldsymbol{b} - \boldsymbol{a}\|^{-1} (\boldsymbol{b} - \boldsymbol{a})$$

satisfies the following inequalities:

$$\begin{aligned} \langle \boldsymbol{x} - \boldsymbol{a}, \boldsymbol{u} \rangle \; &\leq \; 0 \quad \text{for all } \boldsymbol{x} \in \boldsymbol{D}_1, \\ \langle \boldsymbol{y} - \boldsymbol{b}, \boldsymbol{u} \rangle \; &\geq \; 0 \quad \text{for all } \boldsymbol{y} \in \boldsymbol{D}_2. \end{aligned}$$

This implies that

$$\begin{aligned} \sup_{\boldsymbol{x} \in \boldsymbol{D}_1} \langle \boldsymbol{x}, \boldsymbol{u} \rangle \; &\leq \; \langle \boldsymbol{a}, \boldsymbol{u} \rangle, \\ \inf_{\boldsymbol{y} \in \boldsymbol{D}_2} \langle \boldsymbol{y}, \boldsymbol{u} \rangle \; &\geq \; \langle \boldsymbol{b}, \boldsymbol{u} \rangle, \end{aligned}$$

and

$$\langle \boldsymbol{b}, \boldsymbol{u} \rangle - \langle \boldsymbol{a}, \boldsymbol{u} \rangle \; = \; \langle \boldsymbol{b} - \boldsymbol{a}, \boldsymbol{u} \rangle \; = \; \|\boldsymbol{b} - \boldsymbol{a}\| \; > \; 0.$$

Now to part (b). Since $\boldsymbol{V}$ is separable, which is a consequence of $\dim(\boldsymbol{V}) < \infty$, for $j = 1, 2$ there exists a countable dense subset $\{\boldsymbol{z}_{j1}, \boldsymbol{z}_{j2}, \boldsymbol{z}_{j3}, \ldots\}$ of $\boldsymbol{D}_j$; see also Remark 2.23. It follows from Exercise 2.7 and convexity of $\boldsymbol{D}_j$ that $\boldsymbol{D}_{jn} := \operatorname{conv}\{\boldsymbol{z}_{j1}, \ldots, \boldsymbol{z}_{jn}\}$ is a compact subset of $\boldsymbol{D}_j$. Moreover, this construction implies that $\boldsymbol{D}_{j1} \subset \boldsymbol{D}_{j2} \subset \boldsymbol{D}_{j3} \subset \cdots$ with

$$\lim_{n \to \infty} d(\boldsymbol{x}, \boldsymbol{D}_{jn}) \; = \; 0 \quad \text{for any } \boldsymbol{x} \in \boldsymbol{D}_j.$$

According to part (a), for any index n there exists a unit vector $\boldsymbol{u}_n \in \boldsymbol{V}$ such that

$$\inf_{\boldsymbol{y}_n \in \boldsymbol{D}_{2n}} \langle \boldsymbol{y}_n, \boldsymbol{u}_n \rangle \; > \; \sup_{\boldsymbol{x}_n \in \boldsymbol{D}_{1n}} \langle \boldsymbol{x}_n, \boldsymbol{u}_n \rangle.$$

Since $\dim(\boldsymbol{V}) < \infty$, the unit sphere of $\boldsymbol{V}$ is compact. Hence we may replace $\big((\boldsymbol{D}_{1n}, \boldsymbol{D}_{2n}, \boldsymbol{u}_n)\big)_n$ with a subsequence, if necessary, such that $(\boldsymbol{u}_n)_n$ converges to a unit vector $\boldsymbol{u}$. But then for arbitrary $\boldsymbol{x} \in \boldsymbol{D}_1$ and $\boldsymbol{y} \in \boldsymbol{D}_2$ the metric projections $\boldsymbol{x}_n := \Pi_{\boldsymbol{D}_{1n}}\boldsymbol{x}$ and $\boldsymbol{y}_n := \Pi_{\boldsymbol{D}_{2n}}\boldsymbol{y}$ satisfy $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}$ and $\lim_{n\to\infty} \boldsymbol{y}_n = \boldsymbol{y}$, so the properties of $(\boldsymbol{u}_n)_n$ imply that

$$\langle \boldsymbol{y}, \boldsymbol{u}\rangle - \langle \boldsymbol{x}, \boldsymbol{u}\rangle \;=\; \lim_{n\to\infty} \big(\langle \boldsymbol{y}_n, \boldsymbol{u}_n\rangle - \langle \boldsymbol{x}_n, \boldsymbol{u}_n\rangle\big) \;\geq\; 0.$$

Consequently, $\inf_{\boldsymbol{y}\in\boldsymbol{D}_2}\langle \boldsymbol{y}, \boldsymbol{u}\rangle$ is greater than or equal to $\sup_{\boldsymbol{x}\in\boldsymbol{D}_1}\langle \boldsymbol{x}, \boldsymbol{u}\rangle$. □

Remark 2.23 (Separable metric spaces). A metric space $(\mathcal{X}, d)$ is called *separable* if there exists a countable set $\mathcal{X}_o \subset \mathcal{X}$ which is dense in $\mathcal{X}$. The latter property means that for arbitrary $x \in \mathcal{X}$ and $\epsilon > 0$ there exists a $x_o \in \mathcal{X}_o$ with $d(x, x_o) \leq \epsilon$. (A standard example is $\mathcal{X} = \mathbb{R}^d$ with the standard Euclidean distance and its countable subset $\mathcal{X}_o = \mathbb{Q}^d$.)

For any nonvoid subset $\mathcal{Y}$ of $\mathcal{X}$, the metric space $(\mathcal{Y}, d)$ is separable, too. To see this, we 'choose' for each $x_o \in \mathcal{X}_o$ and $k \in \mathbb{N}$ a point $y(x_o, k) \in \mathcal{Y}$ such that $d(x_o, y(x_o, k)) \leq d(x_o, \mathcal{Y}) + 1/k$. Then $\mathcal{Y}_o := \{y(x_o, k) : x_o \in \mathcal{X}_o, k \in \mathbb{N}\}$ is a countable dense subset of $\mathcal{Y}$. Indeed, consider an arbitrary point $y \in \mathcal{Y}$. For $\epsilon > 0$, there exists a point $x_o \in \mathcal{X}_o$ such that $d(y, x_o) \leq \epsilon/3$. Then for any integer $k \geq 3/\epsilon$, the point $y(x_o, k) \in \mathcal{Y}_o$ satisfies

$$\begin{aligned} d(y, y(x_o, k)) \;&\leq\; d(y, x_o) + d(x_o, y(x_o, k)) \\ &\leq\; d(y, x_o) + d(x_o, \mathcal{Y}) + 1/k \\ &\leq\; 2d(y, x_o) + 1/k \\ &\leq\; \epsilon. \end{aligned}$$

2.2.4 Polar sets and cones

For a subset $\boldsymbol{M}$ of $\boldsymbol{V}$ its *polar set* is defined as

$$\boldsymbol{M}^* \;:=\; \big\{\boldsymbol{y} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq 1 \;\text{ for all } \boldsymbol{x} \in \boldsymbol{M}\big\}.$$

Obviously, $\boldsymbol{M}^*$ contains the point $\mathbf{0}$, and writing

$$\boldsymbol{M}^* \;=\; \bigcap_{\boldsymbol{x}\in\boldsymbol{M}} \big\{\boldsymbol{y} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq 1\big\}$$

shows that $\boldsymbol{M}^*$ is closed and convex. With the support function of $\boldsymbol{M}$ one may also write

$$\boldsymbol{M}^* \;=\; \big\{\boldsymbol{y} \in \boldsymbol{V} : h_{\boldsymbol{M}}(\boldsymbol{y}) \leq 1\big\}.$$

If the closed and convex set $\boldsymbol{C}$ contains the point $\mathbf{0}$, its polar set $\boldsymbol{C}^*$ has remarkable properties:

Theorem 2.24 (Polar sets). *Suppose that $\boldsymbol{C}$ contains $\mathbf{0}$. Then*

$$(\boldsymbol{C}^*)^* \;=\; \boldsymbol{C}.$$

The set $\boldsymbol{C}$ is bounded if and only if $\mathbf{0}$ is an interior point of $\boldsymbol{C}^*$. The set $\boldsymbol{C}^*$ is bounded if and only if $\mathbf{0}$ is an interior point of $\boldsymbol{C}$.

Proof of Theorem 2.24. Since $\langle x, y\rangle \le 1$ for arbitrary $\boldsymbol{x} \in \boldsymbol{C}$ and $\boldsymbol{y} \in \boldsymbol{C}^*$, we have the inclusion

$$\boldsymbol{C} \;\subset\; \big\{\boldsymbol{x} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \le 1 \;\text{ for all } \boldsymbol{y} \in \boldsymbol{C}^*\big\} \;=\; (\boldsymbol{C}^*)^*.$$

It remains to show that for any $\boldsymbol{x}_1 \in \boldsymbol{V} \setminus \boldsymbol{C}$ there exists a $\boldsymbol{y} \in \boldsymbol{C}^*$ such that $\langle \boldsymbol{x}_1, \boldsymbol{y}\rangle > 1$. To this end let $\boldsymbol{x}_0 := \Pi_{\boldsymbol{C}}\boldsymbol{x}_1$. Then $\boldsymbol{v} := \boldsymbol{x}_1 - \boldsymbol{x}_0 \neq \boldsymbol{0}$, and by Theorem 2.11 (b), $\langle \boldsymbol{x} - \boldsymbol{x}_0, \boldsymbol{v}\rangle \le 0$ for all $\boldsymbol{x} \in \boldsymbol{C}$. Consequently,

$$\sup_{\boldsymbol{x}\in\boldsymbol{C}} \langle \boldsymbol{x}, \boldsymbol{v}\rangle \;=\; \langle \boldsymbol{x}_0, \boldsymbol{v}\rangle \;=\; \langle \boldsymbol{x}_1, \boldsymbol{v}\rangle - \|\boldsymbol{v}\|^2 \;<\; \langle \boldsymbol{x}_1, \boldsymbol{v}\rangle.$$

Since $\boldsymbol{0} \in \boldsymbol{C}$, we know that

$$0 \;\le\; \sup_{\boldsymbol{x}\in\boldsymbol{C}} \langle \boldsymbol{x}, \boldsymbol{v}\rangle \;=\; \langle \boldsymbol{x}_0, \boldsymbol{v}\rangle \;<\; \langle \boldsymbol{x}_1, \boldsymbol{v}\rangle.$$

In case of $\langle \boldsymbol{x}_0, \boldsymbol{v}\rangle = 0$, any point $\boldsymbol{y} = t\boldsymbol{v}$ with $t \ge 0$ belongs to $\boldsymbol{C}^*$, and $\langle \boldsymbol{x}_1, \boldsymbol{y}\rangle = t\langle \boldsymbol{x}_1, \boldsymbol{v}\rangle > 1$ for sufficiently large $t > 0$. In case of $\langle \boldsymbol{x}_0, \boldsymbol{v}\rangle > 0$ the point $\boldsymbol{y} := \langle \boldsymbol{x}_0, \boldsymbol{v}\rangle^{-1}\boldsymbol{v}$ belongs to $\boldsymbol{C}^*$ with $\langle \boldsymbol{x}_1, \boldsymbol{y}\rangle > 1 = \langle \boldsymbol{x}_0, \boldsymbol{y}\rangle = \sup_{\boldsymbol{x}\in\boldsymbol{C}}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$.

Suppose that $\boldsymbol{C}$ is bounded, that means, $\boldsymbol{C} \subset r\boldsymbol{B}$ for some $r > 0$ with the closed unit ball $\boldsymbol{B}$ of $\boldsymbol{V}$. This implies that

$$\boldsymbol{C}^* \;\supset\; (r\boldsymbol{B})^* \;=\; r^{-1}\boldsymbol{B},$$

see Exercise 2.25. Thus $\boldsymbol{0}$ is an interior point of $\boldsymbol{C}^*$. On the other hand, suppose that $\boldsymbol{0}$ is an interior point of $\boldsymbol{C}$, that means, $r\boldsymbol{B} \subset \boldsymbol{C}$ for some $r > 0$. Then

$$\boldsymbol{C}^* \;\subset\; (r\boldsymbol{B})^* \;=\; r^{-1}\boldsymbol{B},$$

whence $\boldsymbol{C}^*$ is bounded.

Since $(\boldsymbol{C}^*)^* = \boldsymbol{C}$, the previous two statements remain valid if we interchange $\boldsymbol{C}$ and $\boldsymbol{C}^*$. □

Exercise 2.25 (Polar sets of balls centered at $\boldsymbol{0}$). Show that for any $r > 0$ and the closed unit ball $\boldsymbol{B}$ of $\boldsymbol{V}$,

$$(r\boldsymbol{B})^* \;=\; r^{-1}\boldsymbol{B}.$$

Remark 2.26 (Polar cones). Suppose that $\boldsymbol{C}$ is a closed convex cone. Then one can show that

$$\boldsymbol{C}^* \;=\; \big\{\boldsymbol{y} \in \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \le 0\big\}.$$

The latter set is a closed convex cone, too, the so-called *polar cone* of $\boldsymbol{C}$.

Figure 2.1 indicates a closed convex cone $\boldsymbol{C} \subset \mathbb{R}^2$ and its polar cone $\boldsymbol{C}^*$.

2.3 Extremal Points

In this section we consider again an arbitrary real linear space $\boldsymbol{V}$, and $\boldsymbol{C}$ is always a nonvoid convex subset of $\boldsymbol{V}$. Recall the notion of convex hulls: For a nonvoid set $\boldsymbol{M} \subset \boldsymbol{V}$ we defined its convex hull $\mathrm{conv}(\boldsymbol{M})$, the smallest convex set containing $\boldsymbol{M}$. Now the question is whether for a given convex set $\boldsymbol{C}$ there is a *smallest* set $\boldsymbol{M} \subset \boldsymbol{V}$ such that $\boldsymbol{C} = \mathrm{conv}(\boldsymbol{M})$. If such a set $\boldsymbol{M}$ exists, it will certainly contain all "extremal points" of $\boldsymbol{C}$:

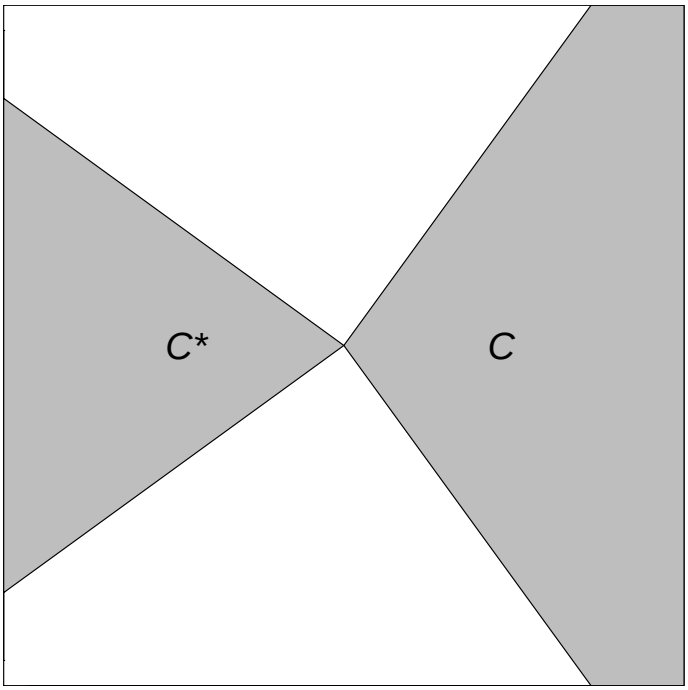

Figure 2.1: A closed convex cone $\boldsymbol{C}$ and its polar cone $\boldsymbol{C}^*$.

Definition 2.27 (Extremal points). A point $\boldsymbol{x} \in \boldsymbol{C}$ is called *extremal point of* $\boldsymbol{C}$ (or *extremal in* $\boldsymbol{C}$) if it satisfies the following condition: If $\boldsymbol{x} = (1-\lambda)\boldsymbol{y} + \lambda\boldsymbol{z}$ with $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{C}$ and $0 < \lambda < 1$, then $\boldsymbol{y} = \boldsymbol{z}$.

The set of all extremal points of $\boldsymbol{C}$ is denoted with $\mathrm{extr}(\boldsymbol{C})$. Any point $\boldsymbol{x} \in \boldsymbol{C} \setminus \mathrm{extr}(\boldsymbol{C})$ can be written as $\boldsymbol{x} = (1-\lambda)\boldsymbol{y} + \lambda\boldsymbol{z}$ with two *different* points $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{C}$ and some $\lambda \in (0,1)$.

Remark 2.28 (Simplified criterion for extremal points). A point $\boldsymbol{x} \in \boldsymbol{C}$ is extremal in $\boldsymbol{C}$ if and only if it satisfies the following condition: If $\boldsymbol{x} = 2^{-1}(\boldsymbol{y} + \boldsymbol{z})$ with $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{C}$, then $\boldsymbol{y} = \boldsymbol{z}$.

Proof: It is clear that any point $\boldsymbol{x} \in \mathrm{extr}(\boldsymbol{C})$ satisfies the latter condition. On the other hand, suppose that $\boldsymbol{x} \in \boldsymbol{C}$ satisfies the latter condition and equals $(1-\lambda)\boldsymbol{y} + \lambda\boldsymbol{z}$ with $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{C}$ and $\lambda \in (0,1)$. Without loss of generality let $\lambda \leq 1/2$, because otherwise we could interchange $\boldsymbol{y}$ and $\boldsymbol{z}$. Then $\boldsymbol{x} = 2^{-1}(\boldsymbol{y} + \boldsymbol{z}')$ with $\boldsymbol{z}' = (1-2\lambda)\boldsymbol{y} + 2\lambda\boldsymbol{z} \in \boldsymbol{C}$. But then $\boldsymbol{y} = \boldsymbol{z}'$, which is equivalent to $\boldsymbol{y} = \boldsymbol{z}$.

Exercise 2.29. Let $\boldsymbol{x}$ be an element of the convex set $\boldsymbol{C} \subset \boldsymbol{V}$. Show that $\boldsymbol{x} \in \mathrm{extr}(\boldsymbol{C})$ if and only if $\boldsymbol{C} \setminus \{\boldsymbol{x}\}$ is convex.

Exercise 2.30. Let $\boldsymbol{C} = \mathrm{conv}(\boldsymbol{M})$ for some set $\boldsymbol{M} \subset \boldsymbol{V}$. Show that

$$\mathrm{extr}(\boldsymbol{C}) \;\subset\; \boldsymbol{M}.$$

More precisely, show that

$$\mathrm{extr}(\boldsymbol{C}) \;=\; \big\{\boldsymbol{x} \in \boldsymbol{M} : \boldsymbol{x} \notin \mathrm{conv}(\boldsymbol{M} \setminus \{\boldsymbol{x}\})\big\}.$$

Remark 2.31. Suppose that $(\boldsymbol{V}, \|\cdot\|)$ is a normed space. Then

$$\mathrm{extr}(\boldsymbol{C}) \;\subset\; \partial\boldsymbol{C}.$$

Proof: If $\boldsymbol{x}$ is an interior point of $\boldsymbol{C}$, then for some $\epsilon > 0$, the closed ball with center $\boldsymbol{x}$ and radius ϵ is contained in $\boldsymbol{C}$. Hence, for any $\boldsymbol{w} \in \boldsymbol{V}$ with $\|\boldsymbol{w}\| = \epsilon$, the points $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{w}$ and $\boldsymbol{z} := \boldsymbol{x} + \boldsymbol{w}$ are different points in $\boldsymbol{C}$ such that $\boldsymbol{x} = 2^{-1}(\boldsymbol{y} + \boldsymbol{z})$.

Exercise 2.32. Let $(\boldsymbol{V}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be a Hilbert space. Show that any unit vector is an extremal point of the closed unit ball $\boldsymbol{B} = \{\boldsymbol{x} \in \boldsymbol{V} : \|\boldsymbol{x}\| \leq 1\}$. Precisely,

$$\mathrm{extr}(\boldsymbol{B}) \;=\; \partial\boldsymbol{B}.$$

For compact and convex subsets of $\mathbb{R}^d$ there is a simple representation theorem which answers our initial question:

Theorem 2.33 (Minkowski). *Let $\boldsymbol{K}$ be a compact and convex subset of $\mathbb{R}^d$. Then $\mathrm{extr}(\boldsymbol{K})$ is nonempty, and*

$$\boldsymbol{K} \;=\; \mathrm{conv}(\mathrm{extr}(\boldsymbol{K})).$$

In Functional Analysis there exist extensions of this theorem to normed and topological vector spaces, e.g. the Krein–Milman theorem or Choquet's theorem. But for our purposes the finite-dimensional variant will be sufficient.

Proof of Theorem 2.33. In case of $d = 1$ the statement is easily verified: Here $\boldsymbol{K} = [a, b]$ with real numbers $a \leq b$, and one can show easily that $\mathrm{extr}(\boldsymbol{K}) = \{a, b\}$ and $\boldsymbol{K} = \mathrm{conv}(\{a, b\})$.

Suppose that for some $d > 1$ the assertion is correct for compact, convex subsets of $\mathbb{R}^{d-1}$. Now we have to show that an arbitrary point $\boldsymbol{x} \in \boldsymbol{K}$ may be represented as a convex combination of points in $\mathrm{extr}(\boldsymbol{K})$.

Suppose first that $\boldsymbol{x}$ is in the interior of $\boldsymbol{K}$. Fixing an arbitrary unit vector $\boldsymbol{u}$, we may move from $\boldsymbol{x}$ in direction $\pm\boldsymbol{u}$ until we hit boundary points $\boldsymbol{y} = \boldsymbol{x} - s\boldsymbol{u}$ and $\boldsymbol{z} = \boldsymbol{x} + t\boldsymbol{u}$ of $\boldsymbol{K}$, where $s, t > 0$. But then $\boldsymbol{x} = (1 - \lambda)\boldsymbol{y} + \lambda\boldsymbol{z}$ with $\lambda := s/(s + t) \in (0, 1)$. Thus it suffices to consider boundary points $\boldsymbol{x}$ of $\boldsymbol{K}$.

Let $\boldsymbol{x}$ be on the boundary of $\boldsymbol{K}$. As shown in Corollary 2.21, there exists a unit vector $\boldsymbol{u}$ such that

$$r_o := \langle\boldsymbol{x}, \boldsymbol{u}\rangle \;=\; \max_{\boldsymbol{y}\in\boldsymbol{K}} \langle\boldsymbol{y}, \boldsymbol{u}\rangle.$$

In particular, $\boldsymbol{x}$ belongs to the compact, convex set

$$\boldsymbol{K}_o \;:=\; \big\{\boldsymbol{y} \in \boldsymbol{K} : \langle\boldsymbol{y}, \boldsymbol{u}\rangle = r_o\big\}.$$

But this set is contained in the hyperplane

$$\boldsymbol{H}_o \;:=\; \big\{\boldsymbol{y} \in \mathbb{R}^d : \langle\boldsymbol{y}, \boldsymbol{u}\rangle = r_o\big\} \;=\; \big\{r_o\boldsymbol{u} + \boldsymbol{z} : \boldsymbol{z} \in \boldsymbol{u}^\perp\big\}$$

which is geometrically equivalent to $\mathbb{R}^{d-1}$. Consequently, by our induction hypothesis, $\boldsymbol{x}$ is a convex combination of extremal points of $\boldsymbol{K}_o$. Now the assertion follows from the fact that any extremal point $\boldsymbol{x}_o$ of $\boldsymbol{K}_o$ is automatically an extremal point of $\boldsymbol{K}$. For if $\boldsymbol{x}_o = 2^{-1}(\boldsymbol{y} + \boldsymbol{z})$ with $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{K}$, then

$$r_o \;=\; \langle\boldsymbol{x}_o, \boldsymbol{u}\rangle \;=\; 2^{-1}\big(\underbrace{\langle\boldsymbol{y}, \boldsymbol{u}\rangle}_{\leq r_o} + \underbrace{\langle\boldsymbol{z}, \boldsymbol{u}\rangle}_{\leq r_o}\big),$$

whence $\langle \boldsymbol{y}, \boldsymbol{u} \rangle = \langle \boldsymbol{z}, \boldsymbol{u} \rangle = r_o$. Consequently, $\boldsymbol{y}$ and $\boldsymbol{z}$ belong to $\boldsymbol{K}_o$, so $\boldsymbol{x}_o \in \operatorname{extr}(\boldsymbol{K}_o)$ implies that $\boldsymbol{y} = \boldsymbol{z} = \boldsymbol{x}_o$. □

Exercise 2.34. Prove or falsify the following claim for $d = 1, 2, 3$: If $\boldsymbol{K} \subset \mathbb{R}^d$ is compact and convex, then $\operatorname{extr}(\boldsymbol{K})$ is compact, too.

Exercise 2.35. Let $\boldsymbol{C} \subset [a, b]^d$ for real numbers $a < b$. Show that

$$\boldsymbol{C} \cap \{a, b\}^d \ \subset \ \operatorname{extr}(\boldsymbol{C}).$$

Then deduce from Exercise 2.30 that

$$\operatorname{extr}(\boldsymbol{C}) \ = \ \boldsymbol{C} \cap \{a, b\}^d \quad \text{if } \boldsymbol{C} = \operatorname{conv}\big(\boldsymbol{C} \cap \{a, b\}^d\big).$$

Example 2.36. Consider the following subsets of $\mathbb{R}^d$:

$$\begin{aligned} \boldsymbol{C}_1 \ &:= \ \Big\{\boldsymbol{x} \in [0, \infty)^d : \sum_{i=1}^{d} x_i \leq 1\Big\}, \\ \boldsymbol{C}_2 \ &:= \ \big\{\boldsymbol{x} \in [0, 1]^d : x_1 \leq x_2 \leq \cdots \leq x_d\big\}. \end{aligned}$$

Both sets are the intersection of finitely many closed halfspaces in $\mathbb{R}^d$, so they are closed and convex. (Precisely, both sets are the intersection of $d + 1$ closed halfspaces.) Moreover, both are contained in $[0, 1]^d$, so they are bounded and thus compact.

Now we show that for $j = 1, 2$,

$$\operatorname{extr}(\boldsymbol{C}_j) \ = \ \boldsymbol{C}_j \cap \{0, 1\}^d.$$

By Exercise 2.35 this is true, provided that $\boldsymbol{C}_j$ is the convex hull of $\boldsymbol{C}_j \cap \{0, 1\}^d$.

The set $\boldsymbol{C}_1 \cap \{0, 1\}^d$ consists of the standard basis vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ and $\boldsymbol{0}$. Any $\boldsymbol{x} \in \boldsymbol{C}_1$ can be written as

$$\boldsymbol{x} \ = \ \sum_{i=1}^{n} x_i \boldsymbol{e}_i \ = \ \sum_{i=1}^{d} x_i \boldsymbol{e}_i + x_{d+1} \boldsymbol{0}$$

with the nonnegative weights $x_1, \ldots, x_d$ and $x_{d+1} := 1 - \sum_{i=1}^{d} x_i$ summing to one. Thus $\boldsymbol{C}_1$ is indeed the convex hull of $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d, \boldsymbol{0}\}$.

Similarly, the set $\boldsymbol{C}_2 \cap \{0, 1\}^d$ consists of $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{d+1}$ with $\boldsymbol{b}_j := (1_{[i \geq j]})_{i=1}^{d}$, i.e. $\boldsymbol{b}_{d+1} = \boldsymbol{0}$. Then any $\boldsymbol{x} \in \boldsymbol{C}_2$ may be written as

$$\boldsymbol{x} \ = \ \sum_{j=1}^{d+1} \lambda_j \boldsymbol{b}_j$$

with the nonnegative weights $\lambda_1 := x_1$, $\lambda_j := x_j - x_{j-1}$ for $2 \leq j \leq d$ and $\lambda_{d+1} := 1 - x_d$ summing to one.

The set $\boldsymbol{C}_2$ in the previous example may be seen as a special case of the set $\boldsymbol{C} \cap [0, 1]^d$ in the next exercise.

Exercise 2.37. Let $\mathcal{O}$ be a nonempty set of index pairs (i,j), where $i,j \in \{1,\dots,d\}$ and $i \neq j$. Then define

$$\boldsymbol{C} \;:=\; \big\{\boldsymbol{x} \in \mathbb{R}^d : x_i \leq x_j \text{ whenever } (i,j) \in \mathcal{O}\big\}.$$

(a) Show that $\boldsymbol{C}$ is a closed convex cone.

(b) Show that

$$\boldsymbol{C} \cap [0,1]^d \;=\; \operatorname{conv}(\boldsymbol{E}_+) \quad \text{with} \quad \boldsymbol{E}_+ \;:=\; \boldsymbol{C} \cap \{0,1\}^d.$$

(Hence, by Exercise 2.35, the elements of $\boldsymbol{E}_+$ are the extremal points of $\boldsymbol{C} \cap [0,1]^d$.)

(c) Show that

$$\boldsymbol{C} \;=\; \operatorname{span}(\mathbf{1}_d) + \boldsymbol{C}_+ \quad \text{with} \quad \boldsymbol{C}_+ \;:=\; \boldsymbol{C} \cap [0,\infty)^d.$$

(d) Show that

$$\boldsymbol{C}_+ \;=\; \operatorname{cone}(\boldsymbol{E}_+) \quad \text{and} \quad \boldsymbol{C} \;=\; \operatorname{cone}(\boldsymbol{E}_+ \cup \{-\mathbf{1}_d\}).$$

2.4 Convex Polyhedra

In this section we consider the space $\boldsymbol{V} = \mathbb{R}^d$ with its standard inner product. In the last section we encountered already various examples of a special type of subsets of $\mathbb{R}^d$:

Definition 2.38 (Convex polyhedron). A nonvoid subset of $\mathbb{R}^d$ is called a *convex polyhedron* if it is the intersection of finitely many closed halfspaces.

Since any closed halfspace is convex, a convex polyhedron is a closed and convex set. The next theorem provides some further facts which are important in various settings.

Theorem 2.39 (Properties of convex polyhedra). *Consider a convex polyhedron*

$$\boldsymbol{C} \;=\; \bigcap_{i=1}^{m} \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \leq r_i\big\}$$

with $m \in \mathbb{N}$, $\boldsymbol{v}_i \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $r_i \in \mathbb{R}$.

***(a)** The following three properties are equivalent:*

(2.1) $\boldsymbol{C}$ *is bounded*;

(2.2) $\max\limits_{1 \leq i \leq m} \langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \;>\; 0$ *for all* $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$;

(2.3) $\mathbf{0}$ *is an interior point of* $\operatorname{conv}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_m)$.

***(b)** A point $\boldsymbol{x} \in \boldsymbol{C}$ is extremal in $\boldsymbol{C}$ if and only if*

(2.4) $$\operatorname{span}\big\{\boldsymbol{v}_i : 1 \leq i \leq m, \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle = r_i\big\} \;=\; \mathbb{R}^d\,.$$

In particular, $\operatorname{extr}(\boldsymbol{C})$ comprises at most $\binom{m}{d}$ different points.

Proof of Theorem 2.39. We start with the equivalence of (2.1) and (2.2). Suppose that property (2.2) is violated. That means, for some $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$,

$$\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \ \leq \ 0 \quad \text{for } i \in \{1, \ldots, m\}.$$

For a fixed point $\boldsymbol{x}_o \in \boldsymbol{C}$ and arbitrary $t \geq 0$, the point $\boldsymbol{x}_o + t\boldsymbol{w}$ lies in $\boldsymbol{C}$ as well, because $\langle \boldsymbol{x}_o + t\boldsymbol{w}, \boldsymbol{v}_i \rangle \leq r_i + t\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \leq r_i$ for $i = 1, \ldots, m$. Consequently, $\boldsymbol{C}$ is unbounded.

On the other hand, suppose that $\boldsymbol{C}$ is unbounded. Then there exists a sequence $(\boldsymbol{x}_n)_n$ in $\boldsymbol{C}$ with $0 < \|\boldsymbol{x}_n\| \to \infty$ as $n \to \infty$. We may replace this sequence with a subsequence, if necessary, such that $\lim_{n\to\infty} \|\boldsymbol{x}_n\|^{-1}\boldsymbol{x}_n = \boldsymbol{w}$ for some unit vector $\boldsymbol{w} \in \mathbb{R}^d$. But this vector satisfies the condition

$$\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \ = \ \lim_{n\to\infty} \|\boldsymbol{x}_n\|^{-1} \underbrace{\langle \boldsymbol{x}_n, \boldsymbol{v}_i \rangle}_{\leq r_i} \ \leq \ \lim_{n\to\infty} \|\boldsymbol{x}_n\|^{-1} r_i \ = \ 0$$

for arbitrary $i \leq \{1, \ldots, m\}$. Hence property (2.2) is violated as well.

Now we prove equivalence of (2.2) and (2.3). Let $\boldsymbol{D} = \operatorname{conv}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$. As shown in Exercise 2.18,

$$h_{\boldsymbol{D}}(\boldsymbol{w}) := \sup_{\boldsymbol{x}\in\boldsymbol{D}} \langle \boldsymbol{x}, \boldsymbol{w} \rangle \ = \ \max_{1\leq i\leq m} \langle \boldsymbol{v}_i, \boldsymbol{w} \rangle.$$

Consequently, (2.2) is equivalent to

$$h_{\boldsymbol{D}}(\boldsymbol{u}) \ > \ 0 \quad \text{for any } \boldsymbol{u} \in \partial\boldsymbol{B}.$$

Since $h_{\boldsymbol{D}}$ is continuous and $\partial\boldsymbol{B}$ is compact, this is even equivalent to

$$\min_{\boldsymbol{u}\in\partial\boldsymbol{B}} h_{\boldsymbol{D}}(\boldsymbol{u}) \ > \ 0.$$

But it follows from Exercise 2.20, that for any $\delta > 0$, the set $\mathbf{0} + \delta\boldsymbol{B} = \delta\boldsymbol{B}$ is contained in $\boldsymbol{D}$ if and only if $h_{\boldsymbol{D}} \geq \delta$ on $\partial\boldsymbol{B}$. Thus (2.2) is equivalent to $\mathbf{0}$ being an interior point of $\boldsymbol{D}$, which is (2.3).

Now we verify part (b). Suppose first that $\boldsymbol{x} \in \boldsymbol{C}$ satisfies (2.4). To show that $\boldsymbol{x} \in \operatorname{extr}(\boldsymbol{C})$, let $\boldsymbol{x} = 2^{-1}(\boldsymbol{y} + \boldsymbol{z})$ with $\boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{C}$. Now we have to show that $\boldsymbol{y} = \boldsymbol{z}$. Indeed,

$$\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \ = \ 2^{-1} \underbrace{\langle \boldsymbol{y}, \boldsymbol{v}_i \rangle}_{\leq r_i} + 2^{-1} \underbrace{\langle \boldsymbol{z}, \boldsymbol{v}_i \rangle}_{\leq r_i} \ < \ r_i \quad \text{if } \min\big(\langle \boldsymbol{y}, \boldsymbol{v}_i \rangle, \langle \boldsymbol{z}, \boldsymbol{v}_i \rangle\big) \ < \ r_i,$$

whence $\langle \boldsymbol{y} - \boldsymbol{z}, \boldsymbol{v}_i \rangle = 0$ for all indices i such that $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle = r_i$. By assumption (2.4), these vectors $\boldsymbol{v}_i$ span the full space, so $\boldsymbol{y} - \boldsymbol{z} = \mathbf{0}$.

Secondly, suppose that $\boldsymbol{x} \in \boldsymbol{C}$ violates condition (2.4). Then there exists a unit vector $\boldsymbol{w} \in \mathbb{R}^d$ such that $\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle = 0$ whenever $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle = r_i$. For $\epsilon > 0$,

$$\langle \boldsymbol{x} \pm \epsilon\boldsymbol{w}, \boldsymbol{v}_i \rangle \ = \ \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \pm \epsilon\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle \ \leq \ r_i \quad \text{for all } i \in \{1, 2, \ldots, m\},$$

provided that

$$\epsilon \ \leq \ \min\Big\{ \frac{r_i - \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle}{\big|\langle \boldsymbol{w}, \boldsymbol{v}_i \rangle\big|} : 1 \leq i \leq m, \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle < r_i \Big\}.$$

In the latter case, $\boldsymbol{x} = 2^{-1}\boldsymbol{y} + 2^{-1}\boldsymbol{z}$ with $\boldsymbol{y} := \boldsymbol{x} - \epsilon\boldsymbol{w} \in \boldsymbol{C}$ and $\boldsymbol{z} := \boldsymbol{x} + \epsilon\boldsymbol{w} \in \boldsymbol{C}$. Thus $\boldsymbol{x}$ is not an extremal point of $\boldsymbol{C}$.

The upper bound for the cardinality of $\mathrm{extr}(\boldsymbol{C})$ may be verified as follows: If $m < d$, the set $\boldsymbol{C}$ has no extremal points at all. Otherwise one could determine all extremal points of $\boldsymbol{C}$ as follows (in principle): For arbitrary indices $1 \le i(1) < \cdots < i(d) \le m$ check whether the vectors $\boldsymbol{v}_{i(1)}$, $\ldots$, $\boldsymbol{v}_{i(d)}$ are linearly independent. If yes, let $\boldsymbol{x}$ be the unique vector satisfying $\langle \boldsymbol{x}, \boldsymbol{v}_{i(j)} \rangle = r_{i(j)}$ for $1 \le j \le d$. Then check whether $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \le r_i$ for all remaining indices i. If yes, $\boldsymbol{x}$ is an extremal point of $\boldsymbol{C}$. Since there are $\binom{m}{d}$ possibilities for chosing $i(1), \ldots, i(d)$, there are at most $\binom{m}{d}$ extremal points in $\boldsymbol{C}$. □

For compact convex polyhedra, there is an alternative simple representation:

Theorem 2.40 (Compact convex polyhedra). *A set $\boldsymbol{C} \subset \mathbb{R}^d$ is a compact convex polyhedron if and only if it is the convex hull of a finite subset of $\mathbb{R}^d$.*

Before proving this theorem, we need a simple auxiliary result about convex sets.

Lemma 2.41. *Let $\boldsymbol{C} = \mathrm{conv}(\boldsymbol{M})$ for some set $\boldsymbol{M} \subset \mathbb{R}^d$ with at least two elements. For any fixed $\boldsymbol{x}_o \in \boldsymbol{C}$, the following two properties of $\boldsymbol{M}$ are equivalent:*

(i) *The set $\boldsymbol{C}$ has nonempty interior.*

(ii) *The vectors $\boldsymbol{x} - \boldsymbol{x}_o$, $\boldsymbol{x} \in \boldsymbol{M}$, span the whole space $\mathbb{R}^d$.*

Proof of Lemma 2.41. Note first that $\boldsymbol{C} \subset \boldsymbol{x}_o + \mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o)$. Indeed, any point $\boldsymbol{x} \in \boldsymbol{C}$ can be written as $\boldsymbol{x} = \sum_{i=1}^n \lambda_i \boldsymbol{x}_i$ with $n \in \mathbb{N}$ positive numbers $\lambda_1, \ldots, \lambda_n$ summing to one and points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{M}$. But this may be rewritten as $\boldsymbol{x} = \boldsymbol{x}_o + \sum_{i=1}^n \lambda_i(\boldsymbol{x}_i - \boldsymbol{x}_o)$, a point in the affine space $\boldsymbol{x}_o + \mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o)$.

Suppose that (ii) is violated, that means, $\mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o)$ has dimension $d' < d$. Then there exists a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ which is perpendicular to $\mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o)$, in other words, $\mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o) \subset \boldsymbol{u}^\perp$. In particular, $\boldsymbol{C} \subset \boldsymbol{x}_o + \mathrm{span}(\boldsymbol{M} - \boldsymbol{x}_o)$ is contained in the hyperplane $\boldsymbol{x}_o + \boldsymbol{u}^\perp$. Thus, $\boldsymbol{C}$ has no interior points, whence (i) is violated, too.

Suppose that (ii) is satisfied. Then there exist points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \boldsymbol{M}$ such that $\boldsymbol{x}_1 - \boldsymbol{x}_o, \ldots, \boldsymbol{x}_d - \boldsymbol{x}_o$ are linearly independent. In other words, the matrix $\boldsymbol{A} := [\boldsymbol{x}_1 - \boldsymbol{x}_o, \ldots, \boldsymbol{x}_d - \boldsymbol{x}_o] \in \mathbb{R}^{d\times d}$ is nonsingular, and $\phi(\boldsymbol{\lambda}) := \boldsymbol{x}_o + \boldsymbol{A}\boldsymbol{\lambda}$ defines a homeomorphism $\phi : \mathbb{R}^d \to \mathbb{R}^d$. In particular, $\boldsymbol{C}$ contains the nonempty open set $\phi(U)$, where $U := \{\boldsymbol{\lambda} \in (0,\infty)^d : \sum_{i=1}^d \lambda_i < 1\}$. Hence, (i) is satisfied as well. □

Proof of Theorem 2.40. Theorem 2.33 implies that a compact convex polyhedron $\boldsymbol{C}$ is equal to $\mathrm{conv}(\mathrm{extr}(\boldsymbol{C}))$. According to Theorem 2.39 (b), the set $\mathrm{extr}(\boldsymbol{C})$ is finite. Thus $\boldsymbol{C}$ is the convex hull of a finite subset of $\mathbb{R}^d$.

On the other hand, suppose that $\boldsymbol{C} = \mathrm{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ with $m \in \mathbb{N}$ and pairwise different points $\boldsymbol{x}_i \in \mathbb{R}^d$. In case of $m = 1$, $\boldsymbol{C} = \{\boldsymbol{x}_1\}$ is easily seen to be the intersection of $2d$ closed halfspaces, so we only consider the case $m \ge 2$.

Suppose first that $\boldsymbol{C}$ has nonvoid interior. Without loss of generality let $\mathbf{0}$ be an interior point of $\boldsymbol{C}$. Then it follows from Theorem 2.24 that the polar set $\boldsymbol{C}^* = \{\boldsymbol{y} \in \mathbb{R}^d : h_{\boldsymbol{C}}(\boldsymbol{y}) \leq 1\}$ is also a compact and convex set with interior point $\mathbf{0}$. But since $\boldsymbol{C}$ is the convex hull of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, it follows from Exercise 2.18 that

$$\boldsymbol{C}^* \;=\; \Big\{\boldsymbol{y} \in \mathbb{R}^d : \max_{1\leq i\leq m} \langle \boldsymbol{x}_i, \boldsymbol{y}\rangle \leq 1\Big\} \;=\; \bigcap_{i=1}^{m} \big\{\boldsymbol{y} \in \mathbb{R}^d : \langle \boldsymbol{x}_i, \boldsymbol{y}\rangle \leq 1\big\},$$

i.e. $\boldsymbol{C}^*$ is a compact convex polyhedron. According to the first part, $\boldsymbol{C}^*$ is the convex hull of a finite subset of $\mathbb{R}^d$. But then we may apply the previous considerations with $\boldsymbol{C}^*$ in place of $\boldsymbol{C}$ and $(\boldsymbol{C}^*)^* = \boldsymbol{C}$ in place of $\boldsymbol{C}^*$ to show that $\boldsymbol{C}$ is a convex polyhedron, too.

In general, without loss of generality let $\mathbf{0} \in \boldsymbol{C}$, and define

$$\boldsymbol{W} \;:=\; \operatorname{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}.$$

Considering $\boldsymbol{C}$ as a subset of $\boldsymbol{W}$, it has nonempty interior, see Lemma 2.41. Thus $\boldsymbol{C}$ may be represented as

$$\boldsymbol{C} \;=\; \bigcap_{i=1}^{n} \big\{\boldsymbol{x} \in \boldsymbol{W} : \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \leq r_i\big\}$$

with $n \in \mathbb{N}$ nonzero vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \boldsymbol{W}$ and real numbers $r_1, \ldots, r_n$. But if $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ is an orthonormal basis of $\boldsymbol{W}^{\perp}$, then we may rewrite $\boldsymbol{C}$ as

$$\begin{aligned}\boldsymbol{C} \;&=\; \bigcap_{i=1}^{n} \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \leq r_i\big\} \\ &\quad\cap \bigcap_{j=1}^{k} \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{b}_j\rangle \leq 0\big\} \cap \bigcap_{j=1}^{k} \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, -\boldsymbol{b}_j\rangle \leq 0\big\},\end{aligned}$$

so $\boldsymbol{C}$ is a convex polygon. □

Exercise 2.42. Let $\boldsymbol{M}$ be an arbitrary nonvoid subset of a real vector space $\boldsymbol{V}$. Show that for any fixed $\boldsymbol{x}_o \in \operatorname{conv}(\boldsymbol{M})$ the following three sets are identical:

$$\begin{aligned}\boldsymbol{W}_1 \;&:=\; \operatorname{span}\{\boldsymbol{x} - \boldsymbol{x}_o : \boldsymbol{x} \in \boldsymbol{M}\},\\ \boldsymbol{W}_2 \;&:=\; \operatorname{span}\{\boldsymbol{x} - \boldsymbol{y} : \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{M}\},\\ \boldsymbol{W}_3 \;&:=\; \operatorname{span}\Big\{\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i : n \in \mathbb{N}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \boldsymbol{M}; \lambda_1, \ldots, \lambda_n \in \mathbb{R} \text{ with } \sum_{i=1}^{n} \lambda_i = 0\Big\}.\end{aligned}$$

2.4.1 Linear programs

In numerous applications one has to minimize a linear function $f : \mathbb{R}^d \to \mathbb{R}$,

$$f(\boldsymbol{x}) \;=\; \langle \boldsymbol{x}, \boldsymbol{c}\rangle \tag{2.5}$$

with a given nonzero vector $\boldsymbol{c} \in \mathbb{R}^d$, under certain constraints on $\boldsymbol{x}$. Precisely, let $\boldsymbol{x}$ be an element of

$$\boldsymbol{C} \;:=\; \big\{\boldsymbol{x} \in [0, \infty)^d : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\big\} \tag{2.6}$$

with a given matrix $\boldsymbol{A} = (a_{i,j})_{i\le k, j\le d} \in \mathbb{R}^{k\times d}$ of rank $k < d$ and a given vector $\boldsymbol{b} \in \mathbb{R}^k$. This set $\boldsymbol{C}$ is a convex polyhedron determined by $d + 2k$ linear inequalities:

$$\begin{aligned} \langle \boldsymbol{x}, \pm\boldsymbol{a}_i\rangle &\le \pm b_i \quad \text{for } 1 \le i \le k, \\ \langle \boldsymbol{x}, -\boldsymbol{e}_j\rangle &\le 0 \quad \text{for } 1 \le j \le d, \end{aligned}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_k]^\top$ with linearly independent vectors $\boldsymbol{a}_1, \dots, \boldsymbol{a}_k \in \mathbb{R}^d$, and $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d$ denotes the standard basis of $\mathbb{R}^d$. The following corollary to Theorem 2.39 provides additional information about $\boldsymbol{C}$ and its extremal points.

Corollary 2.43. *Let $\boldsymbol{C}$ be the polyhedron defined in* (2.6)*. This set is bounded if and only if*

$$\big\{\boldsymbol{w} \in [0,\infty)^d : \boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}\big\} = \{\boldsymbol{0}\}.$$

A point $\boldsymbol{x} \in \boldsymbol{C}$ is extremal in $\boldsymbol{C}$ if and only if

$$\operatorname{span}\big(\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_k\} \cup \{\boldsymbol{e}_i : x_i = 0\}\big) = \mathbb{R}^d.$$

Proof of Corollary 2.43. Recall that $\boldsymbol{C}$ is a convex polyhedron determined by $m = 2k + d$ linear inequalities with directions $\pm\boldsymbol{a}_1, \dots, \pm\boldsymbol{a}_k$ and $-\boldsymbol{e}_1, \dots, -\boldsymbol{e}_d$. Suppose that $\boldsymbol{w} \in \mathbb{R}^d$ satisfies $\langle \boldsymbol{w}, \pm\boldsymbol{a}_j\rangle \le 0$ for $1 \le j \le k$ and $\langle \boldsymbol{w}, -\boldsymbol{e}_i\rangle \le 0$ for $1 \le i \le d$. This is equivalent to saying that $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}$ and $\boldsymbol{w} \in [0,\infty)^d$. Hence, by Theorem 2.39 (a), $\boldsymbol{C}$ is bounded if and only if $\boldsymbol{w} = \boldsymbol{0}$ is the only vector with these properties.

The characterization of $\boldsymbol{x} \in \operatorname{extr}(\boldsymbol{C})$ is an immediate consequence of Theorem 2.39 (b). □

In case of $\boldsymbol{C}$ being compact, we know that $\boldsymbol{C} = \operatorname{conv}(\operatorname{extr}(\boldsymbol{C}))$, and

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\operatorname{extr}(\boldsymbol{C})} f(\boldsymbol{x}).$$

This equation can be verified easily. It also follows from Exercise 2.18 and the fact that

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) = -h_{\boldsymbol{C}}(-\boldsymbol{c}).$$

Even if $\boldsymbol{C}$ is unbounded, either the minimum is attained at an extremal point of $\boldsymbol{C}$ or f is unbounded from below:

Theorem 2.44 (Solutions of a linear program). *Let f and $\boldsymbol{C}$ be given by* (2.5) *and* (2.6)*, respectively. If there exists a vector $\boldsymbol{w} \in [0,\infty)^d$ with $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}$ and $f(\boldsymbol{w}) < 0$, then*

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) = -\infty.$$

Otherwise,

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\operatorname{extr}(\boldsymbol{C})} f(\boldsymbol{x}) \in \mathbb{R}.$$

Proof of Theorem 2.44. Suppose first that $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}$ and $f(\boldsymbol{w}) < 0$ for some $\boldsymbol{w} \in [0,\infty)^d$. For any $\boldsymbol{x} \in \boldsymbol{C}$ and arbitrary $t \ge 0$, the point $\boldsymbol{x} + t\boldsymbol{w}$ belongs to $\boldsymbol{C}$ as well, and

$$f(\boldsymbol{x} + t\boldsymbol{w}) = f(\boldsymbol{x}) + t f(\boldsymbol{w}) \to -\infty \quad \text{as } t \to \infty.$$

Thus

$$\inf_{\boldsymbol{y}\in\boldsymbol{C}} f(\boldsymbol{y}) \;=\; -\infty.$$

Now suppose that $f(\boldsymbol{w}) \geq 0$ whenever $\boldsymbol{w} \in [0,\infty)^d$ and $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{0}$. Suppose that $\boldsymbol{x} \in \boldsymbol{C}$ is not an extremal point of $\boldsymbol{C}$. That means,

$$\operatorname{span}\big(\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k\} \cup \{\boldsymbol{e}_j : x_j = 0\}\big) \;\neq\; \mathbb{R}^d.$$

Hence there exists a unit vector $\boldsymbol{w}$ which is perpendicular to the latter space. In other words, $\boldsymbol{w}$ is a unit vector such that

$$\boldsymbol{A}\boldsymbol{w} \;=\; \boldsymbol{0} \quad\text{and}\quad w_j \;=\; 0 \;\text{ whenever } x_j = 0.$$

Replacing $\boldsymbol{w}$ with $-\boldsymbol{w}$, if necessary, we may assume without loss of generality that $f(\boldsymbol{w}) \geq 0$ and $w_j > 0$ for at least one index j. (Indeed, if $f(\boldsymbol{w}) > 0$ but $\boldsymbol{w} \in (-\infty,0]^d$, then $\tilde{\boldsymbol{w}} := -\boldsymbol{w}$ would satisfy $\boldsymbol{A}\tilde{\boldsymbol{w}} = \boldsymbol{0}$, $\tilde{\boldsymbol{w}} \in [0,\infty)^d$ and $f(\tilde{\boldsymbol{w}}) < 0$.) But then

$$\boldsymbol{x}^{\mathrm{new}} \;:=\; \boldsymbol{x} - t\boldsymbol{w} \quad\text{with}\quad t \;:=\; \min\Big\{\frac{x_j}{w_j} : w_j > 0\Big\} \;>\; 0.$$

would be a new vector in $\boldsymbol{C}$ with $f(\boldsymbol{x}^{\mathrm{new}}) \leq f(\boldsymbol{x})$ and

$$\operatorname{span}\big(\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k\} \cup \{\boldsymbol{e}_j : x_j = 0\}\big) \;\subsetneq\; \operatorname{span}\big(\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k\} \cup \{\boldsymbol{e}_j : x_j^{\mathrm{new}} = 0\}\big).$$

Replacing $\boldsymbol{x}$ with $\boldsymbol{x}^{\mathrm{new}}$ and repeating this step, if necessary, we reach eventually a point $\boldsymbol{x} \in \operatorname{extr}(\boldsymbol{C})$ with the same or a smaller value of $f(\boldsymbol{x})$ than the starting point. Since $\operatorname{extr}(\boldsymbol{C})$ is finite, this shows that

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) \;=\; \min_{\boldsymbol{x}\in\operatorname{extr}(\boldsymbol{C})} f(\boldsymbol{x}) \;\in\; \mathbb{R}.$$

□

Explicit formulae for extremal points of $\boldsymbol{C}$. As shown in Corollary 2.43, a point $\boldsymbol{x} \in \boldsymbol{C}$ is extremal in $\boldsymbol{C}$ if and only if

$$\operatorname{span}\big(\{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k\} \cup \{\boldsymbol{e}_j : x_j = 0\}\big) \;=\; \mathbb{R}^d.$$

By means of basis completion one can show that this condition is equivalent to the following one: There exist indices $1 \leq j(1) < \cdots < j(d-k) \leq d$ such that

$$\det\big(\big[\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k,\, \boldsymbol{e}_{j(1)},\ldots,\boldsymbol{e}_{j(d-k)}\big]\big) \;\neq\; 0 \quad\text{and}\quad x_{j(1)} = \cdots = x_{j(d-k)} \;=\; 0.$$

If $h(1) < \cdots < h(k)$ are the elements of $\{1,\ldots,d\} \setminus \{j(1),\ldots,j(d-k)\}$, then the former determinant equals $\pm\det\big((a_{i,h(\ell)})_{i,\ell=1}^k\big)$. Hence we arrive at the following conclusion:

A point $\boldsymbol{x} \in \boldsymbol{C}$ is extremal in $\boldsymbol{C}$ if and only if there exist indizes $h(1) < \ldots < h(k)$ in $\{1,\ldots,d\}$ such that

$$\det\big((a_{i,h(\ell)})_{i,\ell=1}^k\big) \;\neq\; 0 \quad\text{and}\quad x_j \;=\; 0 \;\text{ if } j \notin \{h(1),\ldots,h(k)\}.$$

In this case,

$$(x_{h(\ell)})_{\ell=1}^k \;=\; \big((a_{i,h(\ell)})_{i,\ell=1}^k\big)^{-1}\boldsymbol{b}.$$

The latter representation is at the heart of the so-called *simplex method*, invented by G.B. Dantzig in 1947. Roughly speaking, one starts with one extremal point $\boldsymbol{x}^{\mathcal{H}}$ of $\boldsymbol{C}$, where $\mathcal{H}$ stands for the index set $\{h(1), \ldots, h(k)\}$, and then one modifies $\mathcal{H}$ step by step by exchanging one of its elements such that the value $f(\boldsymbol{x}^{\mathcal{H}})$ decreases.

Example 2.45 (Nutrients of meals). Let $F_1, F_2, \ldots, F_d$ be given food ingredients, $d \geq 4$. For $1 \leq i \leq d$ let $a_{1i}, a_{2i}, a_{3i} \geq 0$ be the weight fractions of proteins, carbohydrates and fat, respectively, contained in one weight unit of F_i. Suppose one is planning to compose a meal, given by a vector $\boldsymbol{x} \in [0, \infty)^d$, where x_i specifies the amount (weight) of F_i. The goal is to have a meal containing given amounts $b_1, b_2, b_3 > 0$ of proteins, carbohydrates and fat, respectively. The set of vectors $\boldsymbol{x}$ with these properties corresponds to the set $\boldsymbol{C}$ above, where $k = 3$. It follows from Corollary 2.43 that this set is bounded, unless $a_{1i} = a_{2i} = a_{3i} = 0$ for some ingredient F_j.

Now one could think about finding a 'meal' $\boldsymbol{x} \in \boldsymbol{C}$ with a maximal amount of one's favourite ingredient F_{j_*}, i.e. one would like to minimize $f(\boldsymbol{x}) := \langle \boldsymbol{x}, -\boldsymbol{e}_{j_*} \rangle$. Alternatively, let c_j be the price of F_j per unit weight. Then $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{c} \rangle$ would be the overall price of the 'meal' $\boldsymbol{x}$. One could also consider the volume c_j of one weight unit of F_j and thus try to minimize the overall volume of a meal. With $\boldsymbol{c} = \mathbf{1}_d := (1, \ldots, 1)^\top$ one would consider the overall weight of a meal.

Obviously, the extremal points of $\boldsymbol{C}$ may fail to be 'delicious' themselves; their main purpose is to determine the minimum of f over $\boldsymbol{C}$. A data example will be presented in the lecture; see also the supplementary files.

Exercise 2.46. For a given program of study and a certain time interval, let p_i be the relative frequency of degrees with final grade $x_i = 4 + (i-1)/2$, $1 \leq i \leq 5$, among all degrees awarded. Thus $\boldsymbol{p} \in [0,1]^5$ with $\sum_{i=1}^5 p_i = 1$. Suppose we only know that

$$\sum_{i=1}^5 p_i x_i \;=\; \mu \in (4,6) \quad \text{and} \quad \Big(\sum_{i=1}^5 p_i (x_i - \mu)^2\Big)^{1/2} \;=\; \sigma > 0.$$

(a) Show that the set $\boldsymbol{P}(\mu, \sigma)$ of all vectors $\boldsymbol{p}$ with these properties is a compact polyhedron with at most 10 extremal points.

(b) Write a program which generates for given values μ and σ a list of the extremal points of $\boldsymbol{P}(\mu, \sigma)$.

(c) Determine with your program sharp bounds for p_j, $1 \leq j \leq 5$, and $\sum_{i=1}^k p_i$, $1 \leq k \leq 4$, in case of $\mu = 5.5$ and $\sigma = 0.5$.

Inequality constraints and slack variables. Suppose we want to replace some of the constraints $\boldsymbol{a}_i^\top \boldsymbol{x} = b_i$ with the weaker constraint $\boldsymbol{a}_i^\top \boldsymbol{x} \leq b_i$. After rearranging the rows of $\boldsymbol{A}$, if necessary, let the constraint set $\boldsymbol{C}$ consist of all vectors $\boldsymbol{x} \in [0, \infty)^d$ such that

$$\begin{aligned} \boldsymbol{a}_i^\top \boldsymbol{x} \;&\leq\; b_i \quad \text{if } 1 \leq i \leq k', \\ \boldsymbol{a}_i^\top \boldsymbol{x} \;&=\; b_i \quad \text{if } k' < i \leq k. \end{aligned}$$

Then we may introduce a vector $\boldsymbol{y} \in [0, \infty)^{k'}$ of additional "slack variables" and write

$$b_i \;=\; \boldsymbol{a}_i^\top \boldsymbol{x} + y_i \quad \text{for } i = 1, \ldots, k'.$$

Thus we consider the augmented vector $\tilde{\boldsymbol{x}} = [\boldsymbol{x}^\top, \boldsymbol{y}]^\top \in [0,\infty)^{d+k'}$ and want to minimize $\tilde{f}(\tilde{\boldsymbol{x}}) := f(\boldsymbol{x})$ under the constraint that $\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \boldsymbol{b}$, where

$$\tilde{\boldsymbol{A}} := [\boldsymbol{A}, \boldsymbol{D}] \quad \text{with} \quad \boldsymbol{D} := (1_{[i=s]})_{i\le k, s\le k'}.$$

Hence by increasing the dimension d we may translate the weaker constraint set in a polyhedron of type (2.6).

2.4.2 Doubly stochastic matrices and the Hoffmann–Wielandt inequalities

Doubly stochastic matrices. We consider a special example of a convex polyhedron. The matrix space $\mathbb{R}^{d\times d}$ becomes a d^2-dimensional Euclidean space, if we define $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \sum_{i,j=1}^d A_{ij}B_{ij} = \operatorname{trace}(\boldsymbol{A}^\top \boldsymbol{B})$. The corresponding norm is the so-called Frobenius norm $\|\cdot\|_F$.

Now we consider the set Θ of all matrices $\boldsymbol{\theta} \in [0,\infty)^{d\times d}$ whose row and column sums are all equal to one,

$$\sum_{j=1}^d \theta_{ij} = 1 \text{ for } 1\le i\le d \quad \text{and} \quad \sum_{i=1}^d \theta_{ij} = 1 \text{ for } 1\le j\le d.$$

This set is characterized by $d^2 + 4d$ linear inequalities, and since it is contained in $[0,1]^d$, it is compact. Hence our general results imply that $\operatorname{extr}(\Theta)$ is a finite set, and $\Theta = \operatorname{conv}(\operatorname{extr}(\Theta))$. The set of extremal points in Θ may be characterized as follows:

Theorem 2.47 (Birkhoff–von Neumann). *The set* $\operatorname{extr}(\Theta)$ *is the set of all permutation matrices* $\boldsymbol{\theta}$*, that means,*

$$\boldsymbol{\theta} = \big(1_{[j=\sigma(i)]}\big)_{i,j=1}^d$$

for some permutation σ *of* $\{1,\dots,d\}$*.*

In other words, $\boldsymbol{\theta} \in \mathbb{R}^{d\times d}$ is an extremal point of Θ if and only if each row and each column contains precisely one entry 1 and $d-1$ entries 0.

Proof of Theorem 2.47. Instead of using Theorem 2.39 (b) we provide an elementary proof. Note first that $\Theta \cap \{0,1\}^{d\times d}$ is the set of all permutation matrices: The definition of Θ entails that any matrix in $\{0,1\}^{d\times d}$ belongs to Θ if and only if it has precisely on entry 1 and $d-1$ entries 0 in each row and in each column, that means, it is a permutation matrix.

As shown in Exercise 2.35, any matrix in $\boldsymbol{\theta} \in \Theta\cap\{0,1\}^{d\times d}$ is extremal in Θ. Hence, it suffices to show that any matrix $\boldsymbol{\theta}$ in $\Theta\setminus\{0,1\}^{d\times d}$ is not extremal. Indeed, the set $\mathcal{K} := \big\{(i,j) : 0<\theta_{ij}<1\big\}$ is nonempty. The definition of Θ implies that with $(i,j)\in\mathcal{K}$ also $(i',j),(i,j')\in\mathcal{K}$ for some indices $i'\ne i$ and $j'\ne j$. From this one can deduce that there is an even number $k\ge 4$ of different pairs $(i_1,j_1),\dots,(i_k,j_k)\in\mathcal{K}$ such that

$$\begin{aligned} &i_t = i_{t+1} \text{ for odd } t\in\{1,\dots,k-1\},\\ &j_t = j_{t+1} \text{ for even } t\in\{2,3,\dots,k\} \quad \text{with } j_{k+1} := j_1. \end{aligned}$$

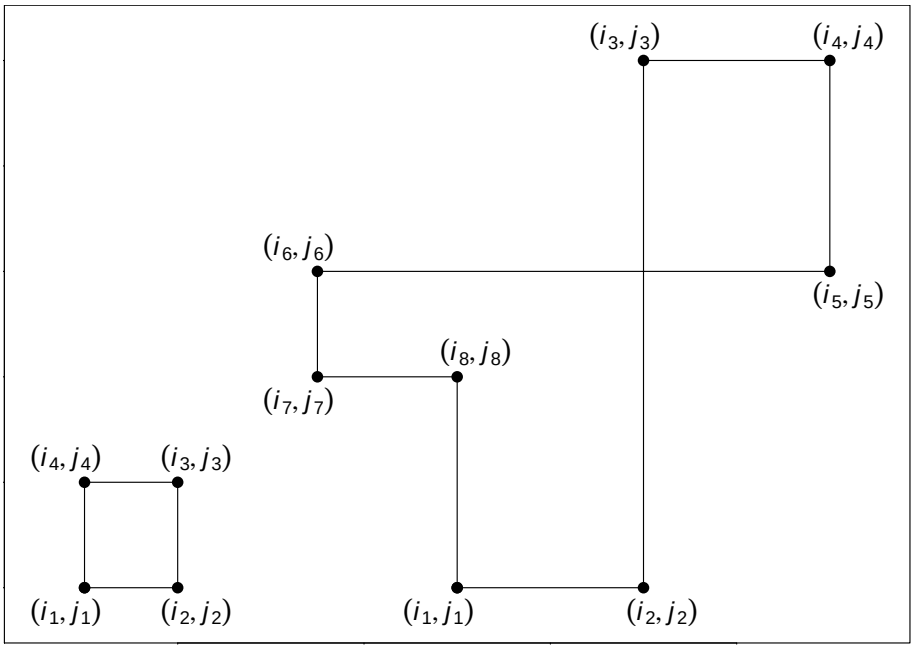

Figure 2.2: Two possible configurations of $(i_1, j_1), \ldots, (i_k, j_k)$.

Figure 2.2 illustrates the definition of this sequence $(i_1, j_1), \ldots, (i_k, j_k)$. Now we define the non-zero matrix $\boldsymbol{\delta} \in \mathbb{R}^{d\times d}$ with components

$$\delta_{ij} \;:=\; \begin{cases} (-1)^t & \text{if } (i,j) = (i_t, j_t), 1 \le t \le k, \\ 0 & \text{else.} \end{cases}$$

All row and column sums of this matrix $\boldsymbol{\delta}$ are equal to 0, and for sufficiently small $\epsilon > 0$,

$$\boldsymbol{\theta} \;=\; 2^{-1}\big((\boldsymbol{\theta} + \epsilon\boldsymbol{\delta}) + (\boldsymbol{\theta} - \epsilon\boldsymbol{\delta})\big) \quad \text{and} \quad \boldsymbol{\theta} \pm \epsilon\boldsymbol{\delta} \;\in\; \Theta.$$

Thus $\boldsymbol{\theta}$ is not an extremal point in Θ. □

The Hoffmann–Wielandt inequalities. The preceding considerations enable an elegant proof of two famous inequalities about eigenvalues of symmetric matrices.

Theorem 2.48 (Hoffmann–Wielandt). *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be symmetric matrices in $\mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ and $\mu_1 \le \mu_2 \le \cdots \le \mu_d$, respectively. Then*

$$\sum_{i=1}^{d} (\lambda_i - \mu_i)^2 \;\le\; \sum_{i,j=1}^{d} (A_{ij} - B_{ij})^2 \;\le\; \sum_{i=1}^{d} (\lambda_i - \mu_{d+1-i})^2.$$

Both inequalities are sharp. Just consider the diagonal matrices $\boldsymbol{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and $\boldsymbol{B} = \operatorname{diag}(\mu_1, \ldots, \mu_d), \operatorname{diag}(\mu_d, \ldots, \mu_1)$. The first inequality may be interpreted as a Lipschitz property of eigenvalues: The standard Euclidean distance between the two vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of ordered eigenvalues is bounded from above by the Frobenius norm of $\boldsymbol{A} - \boldsymbol{B}$,

$$\|\boldsymbol{\lambda} - \boldsymbol{\mu}\| \;\le\; \|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

Proof of Theorem 2.48. The spectral decomposition of symmetric matrices shows that there exist orthogonal matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$ such that

$$\boldsymbol{A} \;=\; \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top \quad \text{and} \quad \boldsymbol{B} \;=\; \boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\mu})\boldsymbol{V}^\top.$$

The general rule $\mathrm{trace}(\boldsymbol{DE}) = \mathrm{trace}(\boldsymbol{ED})$ for arbitrary matrices $\boldsymbol{D}, \boldsymbol{E} \in \mathbb{R}^{d\times d}$ implies that $\|\boldsymbol{U}^\top \boldsymbol{D}\|_F = \|\boldsymbol{D}\|_F = \|\boldsymbol{D}\boldsymbol{V}\|_F$. In particular,

$$\begin{aligned}
\|\boldsymbol{A}-\boldsymbol{B}\|_F^2 \;&=\; \|\boldsymbol{U}^\top\boldsymbol{A}\boldsymbol{V} - \boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{V}\|_F^2 \\
&=\; \|\,\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top\boldsymbol{V} - \boldsymbol{U}^\top\boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\mu})\|_F^2 \\
&=\; \sum_{i,j=1}^d (\lambda_i-\mu_j)^2 (\boldsymbol{U}^\top\boldsymbol{V})_{ij}^2 \\
&=\; \sum_{i,j=1}^d (\lambda_i-\mu_j)^2 \theta_{ij},
\end{aligned}$$

with $\boldsymbol{\theta} := (\theta_{ij})_{i,j=1}^d$ given by

$$\theta_{ij} \;:=\; (\boldsymbol{U}^\top\boldsymbol{V})_{ij}^2.$$

Hence for fixed vectors $\boldsymbol{\lambda}, \boldsymbol{\mu}$, the squared norm $\|\boldsymbol{A}-\boldsymbol{B}\|_F^2$ is a linear function of the matrix $\boldsymbol{\theta}$. Note that $\boldsymbol{W} := \boldsymbol{U}^\top\boldsymbol{V}$ is orthogonal, too, so

$$\begin{aligned}
\sum_{j=1}^d \theta_{ij} \;&=\; (\boldsymbol{W}\boldsymbol{W}^\top)_{ii} \;=\; 1 \quad \text{for } 1 \le i \le d, \\
\sum_{i=1}^d \theta_{ij} \;&=\; (\boldsymbol{W}^\top\boldsymbol{W})_{jj} \;=\; 1 \quad \text{for } 1 \le j \le d.
\end{aligned}$$

That means, $\boldsymbol{\theta}$ belongs to the set $\boldsymbol{\Theta}$ of all doubly stochastic matrices in $\mathbb{R}^{d\times d}$. As stated in Theorem 2.47, $\Theta = \mathrm{conv}(\boldsymbol{P})$ with $\boldsymbol{P}$ the set of all permutation matrices in $\mathbb{R}^{d\times d}$. Thus

$$\min_{\sigma\in\mathcal{S}_d} \sum_{i=1}^d (\lambda_i - \mu_{\sigma(i)})^2 \;\le\; \|\boldsymbol{A}-\boldsymbol{B}\|_F^2 \;\le\; \max_{\sigma\in\mathcal{S}_d} \sum_{i=1}^d (\lambda_i - \mu_{\sigma(i)})^2$$

with the set $\mathcal{S}_d$ of all permutations of $\{1, 2, \ldots, d\}$. Now the assertion follows from Lemma 2.49 below. □

Lemma 2.49 (Hardy–Littlewood–Polya). *For arbitrary real numbers $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ and $\mu_1 \le \mu_2 \le \cdots \le \mu_d$ and any permutation σ of $\{1, 2, \ldots, d\}$,*

$$\sum_{i=1}^d (\lambda_i - \mu_i)^2 \;\le\; \sum_{i=1}^d (\lambda_i - \mu_{\sigma(i)})^2 \;\le\; \sum_{i=1}^d (\lambda_i - \mu_{d+1-i})^2.$$

Proof of Lemma 2.49. This lemma is essentially a consequence of the fact that any permutation may be represented as a combination of finitely many transpositions (i.e. exchanges of pairs). For fixed indices $1 \le j < k \le d$ we define a new permutation $\tilde\sigma$ via

$$\tilde\sigma(i) \;:=\; \begin{cases} \sigma(i) & \text{for } i \notin \{j,k\}, \\ \sigma(k) & \text{for } i = j, \\ \sigma(j) & \text{for } i = k. \end{cases}$$

Then

$$\begin{aligned}
&\sum_{i=1}^{d}(\lambda_i-\mu_{\tilde\sigma(i)})^2-\sum_{i=1}^{d}(\lambda_i-\mu_{\sigma(i)})^2\\
&\quad=\ (\lambda_j-\mu_{\sigma(k)})^2+(\lambda_k-\mu_{\sigma(j)})^2-(\lambda_j-\mu_{\sigma(j)})^2-(\lambda_k-\mu_{\sigma(k)})^2\\
&\quad=\ 2(\lambda_k-\lambda_j)(\mu_{\sigma(k)}-\mu_{\sigma(j)})\ \begin{cases}\le 0 & \text{if } \sigma(k)<\sigma(j),\\ \ge 0 & \text{if } \sigma(k)>\sigma(j).\end{cases}
\end{aligned}\tag{2.7}$$

On the one hand, let j be the smallest index such that $\sigma(j)\neq j$. Then $\sigma(j)>j$ and $k:=\sigma^{-1}(j)>j$ while $\sigma(k)<\sigma(j)$. Then we replace σ with $\tilde\sigma$, so $\sigma(i)=i$ for $i\le j$. Repeating this step as often as necessary, we end up with $\sigma(i)=i$ for all i. In each step the sum of all squares $(\lambda_i-\mu_{\sigma(i)})^2$ stays constant or decreases by (2.7). This yields the first asserted inequality.

On the other hand, let j be the smallest index such that $\sigma(j)\neq d+1-j$. Then $\sigma(j)<d+1-j$ and $k:=\sigma^{-1}(d+1-j)>j$ while $\sigma(k)>\sigma(j)$. Then we replace σ with $\tilde\sigma$, so $\sigma(i)=d+1-i$ for $i\le j$. This step is repeated until finally $\sigma(i)=d+1-i$ for all i. Since in each step the sum of all squares $(\lambda_i-\mu_{\sigma(i)})^2$ stays constant or increases by (2.7), this yields the second asserted inequality. □

2.4.3 Polyhedral cones

So far, we analyzed bounded convex polyhedra in quite some detail. Now we consider another special subfamily of convex polyhedra.

Definition 2.50 (Polyhedral cone). A nonvoid subset $\mathbb{R}^d$ is called a *polyhedral cone* if it is the intersection of finitely many closed halfspaces of the form $\{\boldsymbol{x}\in\mathbb{R}^d:\langle\boldsymbol{x},\boldsymbol{v}\rangle\le 0\}$ with $\boldsymbol{v}\in\mathbb{R}^d\setminus\{\boldsymbol{0}\}$.

Here is an analogue to Theorem 2.40 for polyhedral cones.

Theorem 2.51 (Polyhedral cones). *A set $\boldsymbol{C}\subset\mathbb{R}^d$ is a polyhedral cone if and only if it is the cone generated by a finite subset of $\mathbb{R}^d$.*

This theorem can be deduced from Theorem 2.40 by means of a special decomposition of $\mathbb{R}^d$ which is explained first. Let

$$\boldsymbol{C}\ =\ \bigcap_{i=1}^{m}\big\{\boldsymbol{x}\in\mathbb{R}^d:\langle\boldsymbol{x},\boldsymbol{v}_i\rangle\le 0\big\}\tag{2.8}$$

with $m\in\mathbb{N}$ and nonzero vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\in\mathbb{R}^d$. Now let

$$\boldsymbol{V}\ :=\ \operatorname{span}(\boldsymbol{v}_1,\dots,\boldsymbol{v}_m)\quad\text{and}\quad\boldsymbol{W}\ :=\ \boldsymbol{V}^\perp\ =\ \{\boldsymbol{v}_1,\dots,\boldsymbol{v}_m\}^\perp.$$

Obviously, $\boldsymbol{W}$ is a subset of $\boldsymbol{C}$. If $\boldsymbol{x}\in\mathbb{R}^d$ is written as $\boldsymbol{x}=\boldsymbol{x}_V+\boldsymbol{x}_W$ with $\boldsymbol{x}_V\in\boldsymbol{V}$ and $\boldsymbol{x}_W\in\boldsymbol{W}$, then $\langle\boldsymbol{x},\boldsymbol{v}_i\rangle=\langle\boldsymbol{x}_V,\boldsymbol{v}_i\rangle$ for $1\le i\le m$, whence $\boldsymbol{x}\in\boldsymbol{C}$ if and only if $\boldsymbol{x}_V\in\boldsymbol{C}$. This shows that

$$\boldsymbol{C}\ =\ \boldsymbol{C}\cap\boldsymbol{V}+\boldsymbol{W}.$$

Note that $\boldsymbol{C} \cap \boldsymbol{V}$ is a "pointed cone" in the sense that for $\boldsymbol{x} \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}$, $\boldsymbol{x} \in \boldsymbol{C}$ implies that $-\boldsymbol{x} \notin \boldsymbol{C}$.

Proof of Theorem 2.51. Suppose first that $\boldsymbol{C}$ is given by (2.8). With the decomposition $\boldsymbol{C} = \boldsymbol{C} \cap \boldsymbol{V} + \boldsymbol{W}$ just mentioned, it suffices to show that $\boldsymbol{C} \cap \boldsymbol{V}$ and $\boldsymbol{W}$ are cones generated by finite sets. Since $\{\boldsymbol{0}\} = \mathrm{cone}(\boldsymbol{0})$, it suffices to consider the cases $\boldsymbol{C} \cap \boldsymbol{V} \neq \{\boldsymbol{0}\}$ and $\boldsymbol{W} \neq \{\boldsymbol{0}\}$.

Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$ be a basis of $\boldsymbol{W}$, and let $\boldsymbol{b}_0 := -\sum_{i=1}^q \boldsymbol{b}_i$. Any point $\boldsymbol{x} \in \boldsymbol{W}$ may be written as $\boldsymbol{x} = \sum_{i=1}^q \mu_i \boldsymbol{b}_i$ with suitable numbers $\mu_1, \ldots, \mu_q \in \mathbb{R}$. But for any $\lambda_0 \geq 0$,

$$\boldsymbol{x} \;=\; \lambda_0 \boldsymbol{b}_0 + \sum_{i=1}^{q} (\mu_i + \lambda_0) \boldsymbol{b}_i,$$

and for sufficienlty large λ_0, all factors $\mu_i + \lambda_0$, $1 \leq i \leq q$, are positive. Hence, $\boldsymbol{W} = \mathrm{cone}(\boldsymbol{b}_0, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)$.

Suppose that $\boldsymbol{C} \cap \boldsymbol{V} \neq \{\boldsymbol{0}\}$. With $\boldsymbol{b} := -\sum_{i=1}^m \boldsymbol{v}_i$, every nonzero vector $\boldsymbol{x} \in \boldsymbol{C} \cap \boldsymbol{V}$ satisfies $\langle \boldsymbol{x}, \boldsymbol{b} \rangle = 1$. Indeed, if $\boldsymbol{x} \in \boldsymbol{C}$, then $\langle \boldsymbol{x}, \boldsymbol{b} \rangle = -\sum_{i=1}^m \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \geq 0$ with equality if and only if $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle = 0$ for $1 \leq i \leq m$, whence $\boldsymbol{x} \in \boldsymbol{V}^\perp = \boldsymbol{W}$. Consequently, $\boldsymbol{C} \cap \boldsymbol{V}$ may be written as

$$\boldsymbol{C} \cap \boldsymbol{V} \;=\; \{\lambda \boldsymbol{x} : \lambda \geq 0, \boldsymbol{x} \in \boldsymbol{K}\}$$

with

$$\boldsymbol{K} \;:=\; \big\{\boldsymbol{x} \in \boldsymbol{C} \cap \boldsymbol{V} : \langle \boldsymbol{x}, \boldsymbol{b} \rangle = 1\big\}.$$

Note that $\boldsymbol{K}$ is a convex polyhedron, determined by $m + 2\dim(\boldsymbol{W}) + 2$ linear inequalities. If $\boldsymbol{K}$ is bounded, it follows from Theorem 2.40 that $\boldsymbol{K} = \mathrm{conv}(\boldsymbol{M})$ for some finite subset of $\boldsymbol{C} \cap \boldsymbol{V}$, whence $\boldsymbol{C} \cap \boldsymbol{V} = \mathrm{cone}(\boldsymbol{M})$. To verify boundedness of $\boldsymbol{K}$, assume the contrary. Then there exists a sequence $(\boldsymbol{x}_n)_n$ in $\boldsymbol{K}$ such that $\|\boldsymbol{x}_n\| \to \infty$ as $n \to \infty$. Without loss of generality we may assume that $\boldsymbol{u}_n := \|\boldsymbol{x}_n\|^{-1} \boldsymbol{x}_n$ converges to a unit vector $\boldsymbol{u}$ as $n \to \infty$. Since $\boldsymbol{C} \cap \boldsymbol{V}$ is a closed convex cone, all points $\boldsymbol{u}_n$ and the limit $\boldsymbol{u}$ belong to $\boldsymbol{C} \cap \boldsymbol{V}$. But this would imply the contradiction that

$$\langle \boldsymbol{u}, \boldsymbol{b} \rangle \;=\; \lim_{n \to \infty} \langle \boldsymbol{u}_n, \boldsymbol{b} \rangle \;=\; \lim_{n \to \infty} \|\boldsymbol{x}_n\|^{-1} \;=\; 0.$$

Now suppose that $\boldsymbol{C} = \mathrm{cone}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ for some $m \in \mathbb{N}$ and nonzero vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{R}^d$. As shown in Exercise 2.10, $\boldsymbol{C}$ is closed, whence $\boldsymbol{C} = (\boldsymbol{C}^*)^*$ with $\boldsymbol{M}^*$ denoting the polar set of $\boldsymbol{M} \subset \mathbb{R}^d$. But

$$\boldsymbol{C}^* \;=\; \big\{\boldsymbol{y} \in \mathbb{R}^d : \langle \boldsymbol{y}, \boldsymbol{x} \rangle \leq 0 \text{ for all } \boldsymbol{x} \in \boldsymbol{C}\big\} \;=\; \big\{\boldsymbol{y} \in \mathbb{R}^d : \langle \boldsymbol{y}, \boldsymbol{v}_i \rangle \leq 0 \text{ for } 1 \leq i \leq m\big\},$$

which is a polyhedral cone. As shown before, there exists a finite subset $\boldsymbol{M}$ of $\mathbb{R}^d$ such that $\boldsymbol{C}^* = \mathrm{cone}(\boldsymbol{M})$. But then,

$$\boldsymbol{C} \;=\; \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 0 \text{ for all } \boldsymbol{y} \in \mathrm{cone}(\boldsymbol{M})\big\} \;=\; \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 0 \text{ for all } \boldsymbol{y} \in \boldsymbol{M}\big\},$$

which is a polyhedral cone as well. □

The polar cone of a polyhedral cone. It follows from the general theory that for any closed convex cone $\boldsymbol{C} \subset \mathbb{R}^d$ and its polar cone $\boldsymbol{C}^*$, every point $\boldsymbol{x} \in \mathbb{R}^d$ can be written as $\boldsymbol{x} = \Pi\boldsymbol{x} + \Pi^*\boldsymbol{x}$, where Π and Π^* denote the metric projections onto $\boldsymbol{C}$ and $\boldsymbol{C}^*$, respectively. Moreover, $\Pi\boldsymbol{x} \perp \Pi^*\boldsymbol{x}$.

Now suppose that $\boldsymbol{C}$ is a polyhedral cone as in (2.8). Then,

$$\boldsymbol{C}^* \;=\; \mathrm{cone}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m).$$

For $\boldsymbol{x} \in \mathbb{R}^d$ let

$$J(\boldsymbol{x}) \;:=\; \big\{i \in \{1, \ldots, m\} : \langle \Pi\boldsymbol{x}, \boldsymbol{v}_i\rangle = 0\big\}.$$

Then

$$\Pi\boldsymbol{x} \;\in\; \{\boldsymbol{v}_j : j \in J(\boldsymbol{x})\}^{\perp} \quad \text{and} \quad \Pi^*\boldsymbol{x} \;\in\; \mathrm{cone}(\boldsymbol{v}_j : j \in J(\boldsymbol{x})), \tag{2.9}$$

where $\mathrm{cone}(\emptyset) := \{\boldsymbol{0}\}$. This implies that $\Pi\boldsymbol{x}$ is the orthogonal projection onto the linear space $\{\boldsymbol{v}_j : j \in J(\boldsymbol{x})\}^{\perp}$ and $\Pi^*\boldsymbol{x}$ is the orthogonal projection onto the linear space $\mathrm{span}(\boldsymbol{v}_j : j \in J(\boldsymbol{x}))$. The first part of (2.9) is just a consequence of the definition of $J(\boldsymbol{x})$. As to the second part of (2.9), let $\Pi^*\boldsymbol{x} = \sum_{i=1}^m \lambda_i \boldsymbol{v}_i$ with $\lambda_1, \ldots, \lambda_m \geq 0$. For any index j with $\lambda_j > 0$, the vector $\Pi^*\boldsymbol{x} - \lambda_j \boldsymbol{v}_j$ belongs to $\boldsymbol{C}^*$ too, whence

$$0 \;\leq\; -\langle \Pi\boldsymbol{x}, \boldsymbol{v}_j\rangle \;=\; \lambda_j^{-1}\langle \Pi\boldsymbol{x}, \Pi^*\boldsymbol{x} - \lambda_j\boldsymbol{v}_j\rangle \;\leq\; 0,$$

whence $j \in J(\boldsymbol{x})$. Consequently, $\Pi^*\boldsymbol{x} \in \mathrm{cone}(\boldsymbol{v}_j : j \in J(\boldsymbol{x}))$.

Chapter 3

Convex Functions

3.1 Convex Functions on the Real Line

In Section 1.2 we encountered already convex functions. There, the functions f under consideration were at least differentiable, and convexity was just an additional feature. Now we review the definition of convexity of univariate functions and its implications in full generality and detail. Throughout this section let C be a nondegenerate interval in $\mathbb{R}$.

Definition 3.1 (Convex and concave functions, I). **(a)** A function $f : C \to (-\infty, \infty]$ is called *convex*, if for arbitrary points $x, y \in C$ and $\lambda \in (0, 1)$,

$$f((1-\lambda)x + \lambda y) \;\le\; (1-\lambda)f(x) + \lambda f(y).$$

A function $f : C \to [-\infty, \infty)$ is called *concave*, if $-f$ is convex.

(b) A function $f : C \to \mathbb{R}$ is called *strictly convex*, if for arbitrary different points $x, y \in C$ and $\lambda \in (0, 1)$,

$$f((1-\lambda)x + \lambda y) \;<\; (1-\lambda)f(x) + \lambda f(y).$$

A function $f : C \to \mathbb{R}$ is called *strictly concave*, if $-f$ is strictly convex.

Note that for points $x < y$ in C, the set $\{(1-\lambda)x + \lambda y : \lambda \in (0, 1)\}$ is the interval (x, y). Hence, if we define the *domain* of $f : C \to (-\infty, \infty]$ to be the set

$$\mathrm{dom}(f) \;:=\; \{x \in C : f(x) < \infty\},$$

then convexity of f implies that $\mathrm{dom}(f)$ is a subinterval of C. If we set $f(x) := \infty$ for $x \in \mathbb{R} \setminus C$, then $\mathrm{dom}(f)$ remains unaltered, and convexity of f on C implies convexity of f on $\mathbb{R}$.

Exercise 3.2 (Simplified criterion for convexity). Suppose that $f : C \to \mathbb{R}$ is continuous. Show that f is convex (strictly convex) if and only if

$$f(2^{-1}(x+y)) \;\le (<)\; 2^{-1}\big(f(x) + f(y)\big)$$

for arbitrary different points $x, y \in C$.

Exercise 3.3 (A contraction principle). Let $f : [-1, 1] \to \mathbb{R}$ be convex. Show that for any $x \in (-1, 1)$,

$$f(x) + f(-x) \;\le\; f(1) + f(-1).$$

Alternative characterizations. For three numbers $x < y < z$ in C one can always write

$$y \;=\; (1-\lambda)x + \lambda z \quad \text{with} \quad \lambda \;:=\; \frac{y-x}{z-x} \;\in\; (0,1) \;\text{ and }\; 1-\lambda \;=\; \frac{z-y}{z-x}.$$

Hence, f is (strictly) convex on C if and only if for any choice of the points $x < y < z$ in C,

$$f(y) \;\leq (<)\; (1-\lambda)f(x) + \lambda\, f(z).$$

Now we want to reinterpret such inequalities in terms of the slopes (difference ratios)

$$f'(r,s) \;:=\; \frac{f(s)-f(r)}{s-r} \;\in\; [-\infty,\infty]$$

for arbitrary points $r, s \in C$ such that $r \neq s$ and at least one of them is in $\mathrm{dom}(f)$. Note that $f'(r,s) = f'(s,r)$.

In case of $f(y) < \infty$, the inequality

$$f(y) \;\leq\; (1-\lambda)f(x) + \lambda f(z) \tag{3.1}$$

is equivalent to

$$f'(x,y) \;\leq\; f'(y,z). \tag{3.2}$$

Indeed,

$$\begin{aligned}(1-\lambda)f(x) + \lambda f(z) \;-\; f(y) \;&=\; (1-\lambda)(f(x)-f(y)) + \lambda(f(z)-f(y)) \\ &=\; \frac{(z-y)(f(x)-f(y)) + (y-x)(f(z)-f(y))}{z-x} \\ &=\; \frac{(y-x)(z-y)}{z-x}\big(f'(y,z) - f'(x,y)\big).\end{aligned}$$

In case of $f(x) < \infty$, condition (3.1) is equivalent to

$$f'(x,y) \;\leq\; f'(x,z). \tag{3.3}$$

Indeed, if $f(z) = \infty$, then both (3.1) and (3.3) are trivial. If $f(z) < \infty$, then both (3.1) and (3.3) imply that $f(y) < \infty$, and

$$\begin{aligned}f'(x,z) - f'(x,y) \;&=\; \frac{1}{y-x}\big(\lambda(f(z)-f(x)) - f(y) + f(x)\big) \\ &=\; \frac{1}{y-x}\big((1-\lambda)f(x) + \lambda f(z) \;-\; f(y)\big).\end{aligned}$$

Analogously one can show that in case of $f(z) < \infty$, condition (3.1) is equivalent to

$$f'(x,z) \;\leq\; f'(y,z). \tag{3.4}$$

These considerations lead to the following characterizations of convexity.

Lemma 3.4 (Characterizing convexity). *For a function $f : \mathbb{R} \to (-\infty, \infty]$ the following conditions are equivalent:*

(a) *f is convex;*

(b) *$\mathrm{dom}(f)$ is an interval, and for arbitrary points $x < y < z$ in $\mathrm{dom}(f)$, inequality* (3.2) *holds true;*

(c) *for arbitrary points $x \in \mathrm{dom}(f)$, the function*

$$\mathbb{R} \setminus \{x\} \ni y \;\mapsto\; f'(x, y) \in [-\infty, \infty]$$

is isotonic.

With obvious modifications one can also verify the following characterization of strict convexity:

Lemma 3.5 (Characterizing strict convexity). *For a function $f : C \to \mathbb{R}$, the following two conditions are equivalent:*

(a) *f is strictly convex;*

(b) *for arbitrary points $x < y < z$ in C, inequality* (3.2) *holds strictly;*

(c) *for arbitrary points $x \in C$, the function*

$$C \setminus \{x\} \ni y \;\mapsto\; f'(x, y) \in \mathbb{R}$$

is strictly isotonic.

Smoothness and integral representations. It follows from characterization (c) in Lemma 3.4 that any convex function f admits finite one-sided derivatives on the interior of $\mathrm{dom}(f)$, and the latter satisfy certain inequalities.

Corollary 3.6 (One-sided derivatives). ***(a)*** *Let $f : \mathbb{R} \to (-\infty, \infty]$ be convex. Then for each $x \in \mathrm{dom}(f)$ the one-sided derivatives*

$$\begin{aligned} f'(x-) &:= \lim_{y \to x, y < x} f'(x, y), \\ f'(x+) &:= \lim_{y \to x, y > x} f'(x, y) \end{aligned}$$

exist in $[-\infty, \infty]$, and

$$f'(x-) \;\le\; f'(x+).$$

Moreover, for arbitrary $x, y \in \mathrm{dom}(f)$ with $x < y$:

$$f'(x+) \;\le\; f'(x, y) \;\le\; f'(y-).$$

In particular, $f'(x \pm) \in \mathbb{R}$ for any interior point x of $\mathrm{dom}(f)$.

(b) *Let $f : C \to \mathbb{R}$ be strictly convex. Then for arbitrary $x, y \in C$ with $x < y$,*

$$f'(x+) \;<\; f'(x, y) \;<\; f'(y-).$$

Remark 3.7. For a convex function $f : \mathbb{R} \to (-\infty, \infty]$, Corollary 3.6 (a) has the following implications:

(i) The functions $x \mapsto f'(x\pm)$ are monotonically increasing on $\mathrm{dom}(f)$ and real-valued on the interior of $\mathrm{dom}(f)$.

(ii) There are at most countably many points $x \in \mathrm{dom}(f)$ such that $f'(x-) < f'(x+)$.

For if C_o is the set of all such points $x \in \mathrm{dom}(f)$, then the intervals $\big(f'(x-), f'(x+)\big)$, $x \in C_o$, are pairwise disjoint. Consequently, if we choose a rational number $q(x) \in \big(f'(x-), f'(x+)\big)$, then $C_o \ni x \mapsto q(x) \in \mathbb{Q}$ is injective.

(iii) For any nondegenerate interval $[a, b] \subset \mathrm{dom}(f)$,

$$\big|f(y) - f(x)\big| \;\le\; L|y - x| \quad \text{for } x, y \in [a, b],$$

where $L := \max\{f'(b-), -f'(a+)\}$. In particular, f is Lipschitz-continuous on $[a, b]$ whenever a and b are interior points of $\mathrm{dom}(f)$.

To see this, note that for $a \le x < y \le b$,

$$f'(x, y) \begin{cases} \le\; f'(x, b) \;\le\; f'(b-), \\ \ge\; f'(a, y) \;\ge\; f'(a+). \end{cases}$$

Together with the mean value theorem, Lemmas 3.4, 3.5 and Corollary 3.6 imply the following facts for smooth functions:

Lemma 3.8 (Convexity of differentiable functions). *Let $f : C \to \mathbb{R}$ be continuous, and let C^o be the interior of C.*

(a) *If f is differentiable on C^o, it is (strictly) convex if and only if f' is (strictly) isotonic on C^o.*

(b) *If f is twice differentiable on C^o, it is convex if and only if $f'' \ge 0$ on C^o.*

(c) *If f is twice differentiable on C^o, it is strictly convex if and only if $f'' \ge 0$ on C^o and $\{x \in C^o : f''(x) > 0\}$ is dense in C^o.*

Proof of Lemma 3.8. . Suppose that $f : C \to \mathbb{R}$ is (strictly) convex. Then by Corollary 3.6, f' is (strictly) isotonic on C^o. On the other hand, suppose that f' is (strictly) isotonic. According to the mean value theorem, for arbitrary points $x < y < z$ in C, there exist points $\xi_1 \in (x, y)$ and $\xi_2 \in (y, z)$ such that

$$f'(x, y) = f'(\xi_1) \;\le (<)\; f'(\xi_2) = f'(y, z).$$

Thus it follows from Lemma 3.4 (b) (Lemma 3.5 (b)) that f is (strictly) convex. This proves part (a).

As to part (b), if f is twice differentiable, then f' is isotonic if and only if $f'' \ge 0$, so by part (a), f is convex if and only if $f'' \ge 0$.

As to part (c), by parts (a) and (b), we may assume that f' is isotonic and $f'' \ge 0$. Now we have to show that f' is strictly isotonic if and only if $\{x \in C^o : f''(x) > 0\}$ is dense in C^o. If the latter condition is violated, then $f'' = 0$ on an interval $[a, b] \subset C^o$ with $a < b$. But then, f' is constant

on $[a,b]$, so f' is not strictly increasing. If $\{x \in C^o : f''(x) > 0\}$ is dense in C^o, then for arbitrary $a, b \in C^o$ with $a < b$, there exists a point $x \in (a,b)$ such that $f''(x) > 0$. But then for sufficiently small $\epsilon > 0$, $x \pm \epsilon \in [a,b]$ and

$$f'(a) \;\le\; f'(x-\epsilon) \;<\; f'(x+\epsilon) \;\le\; f'(b).$$

This shows that f' is strictly increasing on C^o. □

Exercise 3.9. Determine for each of the following functions $f : \mathbb{R} \to \mathbb{R}$ maximal intervals C on which it is convex or concave[1].

$$f(x) := \sqrt{1+x^2}; \qquad f(x) := \frac{x^2}{1+x^2};$$
$$f(x) := x^2/\sqrt{1+x^2}; \qquad f(x) := \log(1+x^2);$$
$$f(x) := \log(1+e^x).$$

Example 3.10. For $x \in \mathbb{R}$, let $f(x) := |x|$. With the triangle inequality one can verify rigorously that f is convex with $\mathrm{dom}(f) = \mathbb{R}$. Moreover,

$$f'(x\pm) \;=\; \begin{cases} \mathrm{sign}(x) & \text{if } x \neq 0, \\ \pm 1 & \text{if } x = 0. \end{cases}$$

As a second example consider

$$f(x) \;:=\; \begin{cases} -\sqrt{1-x^2} & \text{if } |x| \le 1, \\ \infty & \text{if } |x| > 1. \end{cases}$$

Here $\mathrm{dom}(f) = [-1,1]$. Moreover, f is convex on $\mathbb{R}$ and strictly convex on $[-1,1]$, because $f < 0 = f(\pm 1)$ on $(-1,1)$, and for $x \in (-1,1)$,

$$f'(x) \;=\; x(1-x^2)^{-1/2}, \quad f''(x) \;=\; (1-x^2)^{-3/2}.$$

Moreover,

$$f'(x\pm) \;=\; \begin{cases} -\infty & \text{if } x = -1, \\ +\infty & \text{if } x = +1. \end{cases}$$

Example 3.11 (An exotic convex function). Let $Q = \{q_1, q_2, q_3, \ldots\}$ be a countable subset of $\mathbb{R}$, and let $a_1, a_2, a_3, \ldots > 0$ such that $\sum_{k=1}^{\infty} a_k(1+|q_k|) < \infty$. Then,

$$f(x) \;:=\; \sum_{k=1}^{\infty} a_k |x - q_k|$$

defines a convex function $f : \mathbb{R} \to \mathbb{R}$ such that f is differentiable at $x \in \mathbb{R}$ if and only if $x \in Q$. Moreover, if $Q \cap C$ is dense in some nondegenerate interval $C \subset \mathbb{R}$, then f is strictly convex on C. The verification of these properties is left to the reader as an exercise.

Our last finding about convex functions on the real line ist the fact that they are absolutely continuous on the interior of their domain:

[1] i.e. $-f$ is convex

Lemma 3.12 (Absolute continuity). *Let $f : \mathbb{R} \to (-\infty, \infty]$ be convex. For $x \in \operatorname{dom}(f)$ let $f'(x)$ be an arbitrary number in $[f'(x-), f'(x+)]$. Then f' is isotonic on $\operatorname{dom}(f)$, and for arbitrary interior points a, b of $\operatorname{dom}(f)$,*

$$f(b) - f(a) \;=\; \int_a^b f'(x)\,dx.$$

Proof of Lemma 3.12. Isotonicity of f' follows from Corollary 3.6. Without loss of generality let $a < b$. For arbitrary $k \in \mathbb{N}$ and $i \in \{0, \dots, k\}$ we define $\delta_k := (b-a)/k$ and $t_{k,i} := a + i\delta_k$. Then it follows from Corollary 3.6 and the construction of f' that

$$f(b) - f(a) = \sum_{i=0}^{k-1} \big(f(t_{k,i+1}) - f(t_{k,i})\big) = \sum_{i=0}^{k-1} f'(t_{k,i}, t_{k,i+1})\delta_k \quad \begin{cases} \geq \sum_{i=0}^{k-1} f'(t_{k,i})\delta_k, \\ \leq \sum_{j=1}^{k} f'(t_{k,j})\delta_k. \end{cases}$$

Now, isotonicity of f' implies that

$$\begin{aligned} \sum_{i=0}^{k-1} f'(t_{k,i})\delta_k \;&=\; \sum_{i=1}^{k} f'(t_{k,i})\delta_k - \big(f'(b) - f'(a)\big)\delta_k \\ &\geq\; \sum_{i=1}^{k} \int_{t_{k,i-1}}^{t_{k,i}} f'(x)\,dx - \big(f'(b) - f'(a)\big)\delta_k \\ &=\; \int_a^b f'(x)\,dx - \big(f'(b) - f'(a)\big)\delta_k \end{aligned}$$

and

$$\begin{aligned} \sum_{j=1}^{k} f'(t_{k,j})\delta_k \;&=\; \sum_{j=0}^{k-1} f'(t_{k,j})\delta_k + \big(f'(b) - f'(a)\big)\delta_k \\ &\leq\; \int_a^b f'(x)\,dx + \big(f'(b) - f'(a)\big)\delta_k \end{aligned}$$

As $k \to \infty$, it follows that $f(b) - f(a) = \int_a^b f'(x)\,dx$. □

Exercise 3.13. Prove or disprove the following claim: A function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 0$ is convex if the function $\mathbb{R} \setminus \{0\} \ni x \mapsto f(x)/x$ is isotonic.

Exercise 3.14. Let $h : C \to \mathbb{R}$ be an isotonic (a strictly isotonic) function. For any fixed $x_o \in C$ define $f : C \to \mathbb{R}$ via

$$f(x) \;:=\; \int_{x_o}^{x} h(t)\,dt.$$

Show that f is (strictly) convex with $f'(x\pm) = h(x\,pm)$.

Exercise 3.15. Consider functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to (-\infty, \infty]$.

(a) Show that $h \circ g$ is convex, provided that g is convex and h is convex and isotonic.

(b) Why is it essential in part (a) that h is isotonic?

(c) Suppose that h is convex and antitonic. Which property of g would guarantee convexity of $h \circ g$?

(d) Find conditions on g and h which guarantee that $h \circ g$ is strictly convex.

3.2 Convex Functions on Linear Spaces

Throughout this section let $\boldsymbol{C}$ be a nondegenerate convex subset of a real vector space $\boldsymbol{V}$.

Definition 3.16 (Convex functions, II). **(a)** A function $f : \boldsymbol{C} \to (-\infty, \infty]$ is called *convex* if for arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}$ and $\lambda \in (0, 1)$,

$$f((1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y}) \;\leq\; (1-\lambda)f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}).$$

A function $f : \boldsymbol{C} \to [-\infty, \infty)$ is called *concave* is $-f$ is convex.

(b) A function $f : \boldsymbol{C} \to \mathbb{R}$ is called strictly convex if for arbitrary different points $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{C}$ and $\lambda \in (0, 1)$,

$$f((1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y}) \;<\; (1-\lambda)f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}).$$

A function $f : \boldsymbol{C} \to \mathbb{R}$ is called *strictly concave* is $-f$ is strictly convex.

Exercise 3.17 (Norms). Let $(\boldsymbol{V}, \|\cdot\|)$ a real normed vector space. Show that for any $\boldsymbol{x}_o \in \boldsymbol{V}$, $f(\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{x}_o\|$ defines a convex function f on $\mathbb{R}$.

Exercise 3.18 (Sublevel sets). Suppose that $f : \boldsymbol{C} \to (-\infty, \infty]$ is convex. Show that all sublevel sets $\{\boldsymbol{x} \in \boldsymbol{C} : f(\boldsymbol{x}) \leq r\}$ and $\{\boldsymbol{x} \in \boldsymbol{C} : f(\boldsymbol{x}) < r\}$ with $r \in \mathbb{R}$ are convex. Does the latter property imply convexity?

Again we define

$$\operatorname{dom}(f) \;:=\; \{\boldsymbol{x} \in \boldsymbol{C} : f(\boldsymbol{x}) < \infty\}$$

for a function $f : \boldsymbol{C} \to (-\infty, \infty]$. If f is convex, then $\operatorname{dom}(f)$ is a convex subset of $\boldsymbol{C}$. If we define $f(\boldsymbol{x}) := \infty$ for $\boldsymbol{x} \in \boldsymbol{V} \setminus \boldsymbol{C}$, then convexity of f on $\boldsymbol{C}$ is equivalent to convexity of f on $\boldsymbol{V}$ while $\operatorname{dom}(f)$ remains unaltered.

Remark 3.19 (Epigraphs). For a function $f : \boldsymbol{V} \to (-\infty, \infty]$, its *epigraph* is defined as

$$\operatorname{epi}(f) \;:=\; \big\{(\boldsymbol{x}, t) \in \boldsymbol{V} \times \mathbb{R} : t \geq f(\boldsymbol{x})\big\}$$

Note that

$$\big\{(\boldsymbol{x}, f(\boldsymbol{x})) : \boldsymbol{x} \in \operatorname{dom}(f)\big\} \;\subset\; \operatorname{epi}(f) \;\subset\; \operatorname{dom}(f) \times \mathbb{R}.$$

A useful fact is that $f : \boldsymbol{V} \to (-\infty, \infty]$ is convex if and only if $\operatorname{epi}(f)$ is a convex subset of $\boldsymbol{V} \times \mathbb{R}$.

Indeed, suppose first that f is a convex function. Let $(\boldsymbol{x}, s), (\boldsymbol{y}, t) \in \operatorname{epi}(f)$. Then for $\lambda \in (0, 1)$,

$$(1-\lambda)s + \lambda t \;\geq\; (1-\lambda)f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \;\geq\; f\big((1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y}\big)$$

and therefore

$$(1-\lambda)(\boldsymbol{x},s)+\lambda(\boldsymbol{y},t) \;=\; \big((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y},\,(1-\lambda)s+\lambda t\big) \;\in\; \mathrm{epi}(f).$$

This shows convexity of $\mathrm{epi}(f)$.

Assume now that $\mathrm{epi}(f)$ is a convex set. For $\boldsymbol{x},\boldsymbol{y}\in\mathrm{dom}(f)$ and $\lambda\in(0,1)$,

$$\big((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y},\,(1-\lambda)f(\boldsymbol{x})+\lambda f(\boldsymbol{y})\big) \;=\; (1-\lambda)(\boldsymbol{x},f(\boldsymbol{x}))+\lambda(\boldsymbol{y},f(\boldsymbol{y})) \;\in\; \mathrm{epi}(f),$$

because $(\boldsymbol{x},f(\boldsymbol{x})),(\boldsymbol{y},f(\boldsymbol{y}))\in\mathrm{epi}(f)$ and $\mathrm{epi}(f)$ is convex. This shows that

$$(1-\lambda)f(\boldsymbol{x})+\lambda f(\boldsymbol{y}) \;\geq\; f\big((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y}\big).$$

Hence f is convex.

Convexity of a function can often be established by means of the following lemma the proof of which is left to the reader as an exercise:

Lemma 3.20 (Verifying convexity).

(a) *Let $m\in\mathbb{N}$, and let $f:=\sum_{i=1}^m a_i f_i$ with numbers $a_i>0$ and convex functions $f_i:\boldsymbol{V}\to(-\infty,\infty]$. Show that f is convex, too.*

(b) *Let L be an affine function from $\boldsymbol{V}$ into another real vector space $\boldsymbol{W}$, and let $g:\boldsymbol{W}\to(-\infty,\infty]$ be a convex function. Then $g\circ L$ is a convex function on $\boldsymbol{V}$.*

(c) *Let $(f_\lambda)_{\lambda\in\Lambda}$ be a family of convex functions $f_\lambda:\boldsymbol{V}\to(-\infty,\infty]$. Then the pointwise supremum $\sup_{\lambda\in\Lambda}f_\lambda$ is convex, too.*

Here is a simple but very useful observation: A function $f:\boldsymbol{C}\to(-\infty,\infty]$ is convex if and only if for arbitrary points $\boldsymbol{x}\in\boldsymbol{C}$ and vectors $\boldsymbol{v}\in\boldsymbol{V}\setminus\{\boldsymbol{0}\}$ the function

$$t \;\mapsto\; f(\boldsymbol{x}+t\boldsymbol{v})$$

is convex on the interval $\{t\in\mathbb{R}:\boldsymbol{x}+t\boldsymbol{v}\in\boldsymbol{C}\}$. Analogously, a function $f:\boldsymbol{C}\to\mathbb{R}$ is strictly concave, if the previous univariate funtion is always strictly convex.

Exercise 3.21 (Affine functions). Let $f:\boldsymbol{V}\to\mathbb{R}$ be both convex and concave. Show that f is an *affine* function, that means,

$$f(\boldsymbol{x}) \;=\; c+L(\boldsymbol{x})$$

for some constant $c\in\mathbb{R}$ and a linear function $L:\boldsymbol{V}\to\mathbb{R}$.

Lemma 3.22 (Monotone gradients). *Let $\boldsymbol{C}$ be an open, convex subset of $\mathbb{R}^d$, and let $f:\boldsymbol{C}\to\mathbb{R}$ be differentiable. Then f is convex (strictly convex) if and only if*

$$\big(\nabla f(\boldsymbol{y})-\nabla f(\boldsymbol{x})\big)^\top(\boldsymbol{y}-\boldsymbol{x}) \;\geq(>)\; 0$$

for arbitrary different points $\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{C}$.

Lemma 3.23 (Hessian matrices). *Let $\boldsymbol{C}$ be an open, convex subset of $\mathbb{R}^d$, and let $f : \boldsymbol{C} \to \mathbb{R}$ be twice differentiable.*

(a) *f is convex if and only if the Hessian matrix $D^2 f(\boldsymbol{x})$ is positive semidefinite for any point $\boldsymbol{x} \in \boldsymbol{C}$.*

(b) *f is strictly convex if the Hessian matrix $D^2 f(\boldsymbol{x})$ is positive definite for any point $\boldsymbol{x} \in \boldsymbol{C}$.*

The previous two lemmata follow essentially from our results about univariate functions (Section 3.1) and the observation that for any $\boldsymbol{x} \in \boldsymbol{C}$ and $\boldsymbol{v} \in \mathbb{R}^d$, $g(t) := f(\boldsymbol{x} + t\boldsymbol{v})$ satisfies

$$g'(t) \;=\; \nabla f(\boldsymbol{x} + t\boldsymbol{v})^\top \boldsymbol{v} \quad \text{and} \quad g''(t) \;=\; \boldsymbol{v}^\top D^2 f(\boldsymbol{x} + t\boldsymbol{v}) \boldsymbol{v}.$$

Example 3.24 (Quadratic functions). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a quadratic function, i.e. $f(\boldsymbol{x}) = c + \boldsymbol{b}^\top \boldsymbol{x} + 2^{-1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$ with a constant $c \in \mathbb{R}$, a vector $\boldsymbol{b} \in \mathbb{R}^d$ and a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. Here

$$\nabla f(\boldsymbol{x}) \;=\; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \quad \text{and} \quad D^2 f(\boldsymbol{x}) \;=\; \boldsymbol{A}.$$

Note also that

$$\big\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \big\rangle \;=\; (\boldsymbol{y} - \boldsymbol{x})^\top \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{x}).$$

Hence f is convex if, and only, if $\boldsymbol{A}$ is positiv semidefinite. Strict convexity of f is equivalent to $\boldsymbol{A}$ being positive definite.

A simple version of Jensen's inequality. Inductively one can show that a convex function f on $\boldsymbol{C}$ has the following property: For any $n \in \mathbb{N}$,

$$f\Big(\sum_{i=1}^n \lambda_i \boldsymbol{x}_i\Big) \;\le\; \sum_{i=1}^n \lambda_i f(\boldsymbol{x}_i)$$

whenever $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \boldsymbol{C}$ and $\lambda_1, \dots, \lambda_n > 0$ such that $\sum_{i=1}^n \lambda_i = 1$. This inequality is sometimes called Jensen's inequality. It implies a simple fact which we shall use several times:

Lemma 3.25. *Let $\boldsymbol{C} = \operatorname{conv}(\boldsymbol{S})$ for a nonvoid set $\boldsymbol{S} \subset \boldsymbol{V}$. For any convex function $f : \boldsymbol{C} \to (-\infty, \infty]$,*

$$\sup_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x}).$$

For any affine function $f : \boldsymbol{C} \to \mathbb{R}$,

$$\inf_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x}) \;=\; \inf_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x}) \quad \textit{and} \quad \sup_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x}).$$

Proof of Lemma 3.25. Since $\boldsymbol{C} = \operatorname{conv}(\boldsymbol{S})$, $\sup_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x})$ is the supremum of

$$f\Big(\sum_{i=1}^n \lambda_i \boldsymbol{x}_i\Big)$$

over all $n \in \mathbb{N}$, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \boldsymbol{S}$ and $\lambda_1, \dots, \lambda_n > 0$ with $\sum_{i=1}^n \lambda_i = 1$. Taking $n = 1$ shows that $\sup_{\boldsymbol{x} \in \boldsymbol{C}} f(\boldsymbol{x}) \ge \sup_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x})$. But by Jensen's inequality,

$$f\Big(\sum_{i=1}^n \lambda_i \boldsymbol{x}_i\Big) \;\le\; \sum_{i=1}^n \lambda_i f(\boldsymbol{x}_i) \;\le\; \sup_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x}).$$

If f is an affine function, then f as well as $-f$ are convex, so the second part is an immediate consequence of the first part. □

Exercise 3.26 (Minkowski inequality). Prove the generalized Minkowski inequality: For arbitrary positive numbers $x_1, \dots, x_n$ and $\lambda_1, \dots, \lambda_n$ with $\sum_{i=1}^n \lambda_i = 1$,

$$\prod_{i=1}^n x_i \;\le\; \sum_{i=1}^n \lambda_i x_i^{1/\lambda_i}.$$

Exercise 3.27 (Investment plans). Mr. Müller and Mr. Lüdenscheid, two wealthy gentlemen, invest money at time points $t_1 < t_2 < \cdots < t_n$ in shares of the promising company *IMSV*. The strategy of Mr. Müller is to invest each time a fixed amount $c > 0$, regardless of the current price. (For simplicity we assume that one can buy fractions of shares.) Mr. Lüdenscheid, however, buys each time a fixed number $d > 0$ of shares. Who of the two will end up with the higher ratio

$$\frac{\text{amount of shares}}{\text{amount of invested money}}\,?$$

3.3 Directional Derivatives and Minimizers

One of the most important features of convex functions is the fact that any local minimizer is also a global minimizer. This will be derived subsequently.

Let $f : \boldsymbol{V} \to (-\infty, \infty]$ be a convex function. For any point $\boldsymbol{x} \in \operatorname{dom}(f)$ and arbitrary vectors $\boldsymbol{v} \in \boldsymbol{V}$, the one-sided directional derivative

$$Df(\boldsymbol{x}, \boldsymbol{v}) \;:=\; \lim_{t \to 0+} \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} \;\in\; [-\infty, \infty]$$

is well-defined. This follows from Corollary 3.6, applied to the convex function $\mathbb{R} \ni t \mapsto f(\boldsymbol{x} + t\boldsymbol{v})$. As a function of $\boldsymbol{v}$, $Df(\boldsymbol{x}, \boldsymbol{v})$ is *sublinear*, that means,

$$Df(\boldsymbol{x}, \lambda\boldsymbol{v}) \;=\; \lambda Df(\boldsymbol{x}, \boldsymbol{v}) \qquad \text{for } \boldsymbol{v} \in \boldsymbol{V} \text{ and } \lambda \ge 0, \tag{3.5}$$

$$Df(\boldsymbol{x}, \boldsymbol{v} + \boldsymbol{w}) \;\le\; Df(\boldsymbol{x}, \boldsymbol{v}) + Df(\boldsymbol{x}, \boldsymbol{w}) \qquad \text{for } \boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{V} \text{ such that } \big\{Df(\boldsymbol{x}, \boldsymbol{v}), Df(\boldsymbol{x}, \boldsymbol{w})\big\} \neq \{-\infty, \infty\}. \tag{3.6}$$

In particular, if $Df(\boldsymbol{x}, \boldsymbol{v}) > -\infty$ for all $\boldsymbol{v} \in \boldsymbol{V}$, then $Df(\boldsymbol{x}, \cdot)$ is a convex function on $\boldsymbol{V}$.

Exercise 3.28. Prove the statements (3.5) and (3.6). Show also that $Df(\boldsymbol{x}, \cdot)$ is a convex function on $\boldsymbol{V}$, provided that $Df(\boldsymbol{x}, \cdot) > -\infty$.

Example 3.29. To see that for fixed $\boldsymbol{x} \in \operatorname{dom}(f)$, $Df(\boldsymbol{x}, \cdot)$ may take the values $\pm\infty$, let $\boldsymbol{V} = \mathbb{R}$ and

$$f(x) \;:=\; \begin{cases} \infty & \text{if } |x| > 1, \\ -\sqrt{1 - x^2} & \text{if } |x| \le 1. \end{cases}$$

Then $Df(x, v) = vf'(x) = vx/\sqrt{1 - x^2}$ for $x \in (-1, 1)$ and $v \in \mathbb{R}$, but $Df(1, \pm 1) = \pm\infty$ and $Df(-1, \pm 1) = \mp\infty$.

It follows from Corollary 3.6 that

$$f(\boldsymbol{x}) + Df(\boldsymbol{x}, \boldsymbol{v}) \;\leq\; f(\boldsymbol{x} + \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}. \tag{3.7}$$

This statement and the definition of directional derivatives imply the aforementioned fact that every local minimizer of a convex function is automatically a global minimizer:

Theorem 3.30. *Let $f : \boldsymbol{V} \to (-\infty, \infty]$ be a convex function. The following two statements about $\boldsymbol{x} \in \operatorname{dom}(f)$ are equivalent:*

(a) *$\boldsymbol{x}$ is a global minimizer of f, that means, $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \boldsymbol{V}$;*

(b) *$Df(\boldsymbol{x}, \boldsymbol{v}) \geq 0$ for arbitrary $\boldsymbol{v} \in \boldsymbol{V}$.*

Here is a version of the same result for convex functions on convex subsets of $\boldsymbol{V}$:

Theorem 3.31. *Let f be a real-valued function on a convex subset $\boldsymbol{C}$ of $\boldsymbol{V}$. The following two statements about $\boldsymbol{x} \in \boldsymbol{C}$ are equivalent:*

(a) *$\boldsymbol{x}$ is a global minimizer of f on $\boldsymbol{C}$, that means, $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \boldsymbol{C}$;*

(b) *$Df(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x}) \geq 0$ for arbitrary $\boldsymbol{y} \in \boldsymbol{C}$.*

Exercise 3.32 (ℓ_p-norms). For $p \geq 1$ and $\boldsymbol{x} \in \mathbb{R}^d$ let $\|\boldsymbol{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. Determine for $p, q \geq 1$ the number

$$C_{p,q} \;:=\; \max_{\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}} \frac{\|\boldsymbol{x}\|_q}{\|\boldsymbol{x}\|_p}.$$

Hint: One can write $\big(\|\boldsymbol{x}\|_q / \|\boldsymbol{x}\|_p\big)^q$ as a symmetric convex or concave function of the vector $\boldsymbol{w} := \|\boldsymbol{x}\|_p^{-p} (|x_i|^p)_{i=1}^d$ which lies in the convex polyhedron $\boldsymbol{C} = \big\{\boldsymbol{w} \in [0, \infty)^d : \sum_{i=1}^d w_i = 1\big\}$.

Example 3.33 (Differentiable and quadratic functions). Let $\boldsymbol{C}$ be an open subset of $\mathbb{R}^d$, and let $f : \boldsymbol{C} \to \mathbb{R}$ be differentiable and convex. Since $Df(\boldsymbol{x}, \boldsymbol{v}) = \nabla f(\boldsymbol{x})^\top \boldsymbol{v}$ for arbitrary $\boldsymbol{x} \in \boldsymbol{C}$ and $\boldsymbol{v} \in \mathbb{R}^d$, a point $\boldsymbol{x} \in \boldsymbol{C}$ minimizes f if and only if

$$\nabla f(\boldsymbol{x}) \;=\; \boldsymbol{0}.$$

In particular, let $\boldsymbol{C} = \mathbb{R}^d$, and let f be a quadratic function, i.e. $f(\boldsymbol{x}) = c + \boldsymbol{b}^\top \boldsymbol{x} + 2^{-1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$ for a constant $c \in \mathbb{R}$, a vector $\boldsymbol{b} \in \mathbb{R}^d$ and a symmetric, positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$. Since $\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, a point $\boldsymbol{x}$ minimizes f if and only if

$$\boldsymbol{A}\boldsymbol{x} \;=\; -\boldsymbol{b}.$$

In particular, a minimizer exists if and only if $\boldsymbol{b}$ lies in the column space of $\boldsymbol{A}$. In case of $\boldsymbol{A}$ being positive definite, the function f is strictly convex with unique minimizer

$$\boldsymbol{x} \;=\; -\boldsymbol{A}^{-1} \boldsymbol{b}.$$

Exercise 3.34. For $\boldsymbol{x} \in (0, \infty)^2$ let

$$f(\boldsymbol{x}) \;:=\; \frac{x_1^2}{x_2} + \frac{x_2^2}{x_1} - x_1 - x_2.$$

Show that $f \geq 0$, and determine the set of minimizers of f.

Remark 3.35. Let f be a convex function on an open convex subset $\boldsymbol{C}$ of $\mathbb{R}^d$, and let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ be the standard basis of $\mathbb{R}^d$. In case of $d \geq 2$, it may happen that a point $\boldsymbol{x} \in \boldsymbol{C}$ satisfies the $2d$ constraints $Df(\boldsymbol{x}, \pm\boldsymbol{e}_i) \geq 0$, $1 \leq i \leq d$, but $\boldsymbol{x}$ is *not* a minimizer of f. As a counterexample, consider

$$f(\boldsymbol{x}) \;:=\; \max_{1\leq i\leq d} x_i.$$

Since f is a maximum of d linear functions, it is convex. Moreover, one can show that

$$Df(\boldsymbol{x}, \boldsymbol{v}) \;=\; \max_{i\in I(\boldsymbol{x})} v_i \quad \text{with} \quad I(\boldsymbol{x}) \;:=\; \operatorname*{arg\,max}_{i\in\{1,\ldots,d\}} x_i.$$

In particular, if $\boldsymbol{x} = r\mathbf{1}_d$, then for $1 \leq i \leq d$,

$$Df(\boldsymbol{x}, \boldsymbol{e}_i) \;=\; 1 \quad \text{and} \quad Df(\boldsymbol{x}, -\boldsymbol{e}_i) \;=\; 0.$$

But $Df(\boldsymbol{y}, -\mathbf{1}_d) = -1$ for any $\boldsymbol{y} \in \mathbb{R}^d$, whence f has no minimizer on any open set $\boldsymbol{C} \subset \mathbb{R}^d$.

Exercise 3.36. Find a convex function $f : \mathbb{R}^d \to \mathbb{R}$ such that for a given basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d$ of $\mathbb{R}^d$ and some $\boldsymbol{x} \in \mathbb{R}^d$, the $2d$ inequalites $Df(\boldsymbol{x}, \pm\boldsymbol{b}_i) > 0$, $1 \leq i \leq d$, are satisfied but $\boldsymbol{x} \notin \arg\min_{\mathbb{R}^d}(f)$.

Exercise 3.37 (Spatial median, I). For given points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ we consider the function $f : \mathbb{R}^d \to \mathbb{R}$ with

$$f(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}\|.$$

(a) Show that f is a convex function satisfying $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$.

(b) Derive an explicit expression for the directional derivatives $Df(\boldsymbol{x}, \boldsymbol{v})$. (You should distinguish the index sets $\{i : \boldsymbol{x}_i = \boldsymbol{x}\}$ and $\{i : \boldsymbol{x}_i \neq \boldsymbol{x}\}$.)

(c) Show that

$$\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \;\subset\; \operatorname{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).$$

Two examples of regularized estimators

Example 3.38 (LASSO penalization). In statistics one is often interested in minimizing a differentiable convex function $g : \mathbb{R}^d \to \mathbb{R}$, for instance a sum of squared residuals

$$g(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} (y_i - \boldsymbol{z}_i^\top \boldsymbol{x})^2/2$$

with a given data vector $\boldsymbol{y} \in \mathbb{R}^n$ and given design vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}^d$. In this particular example for g, if the vectors $\boldsymbol{z}_i$ span $\mathbb{R}^d$, there exists a unique minimizer $\widehat{\boldsymbol{x}}$ of g: With $\boldsymbol{Z} := [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^\top \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\Gamma} := \boldsymbol{Z}^\top\boldsymbol{Z} = \sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top$,

$$\begin{aligned}
g(\boldsymbol{x}) \;&=\; \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{x}\|^2/2 \\
&=\; \|\boldsymbol{y}\|^2/2 - \boldsymbol{x}^\top\boldsymbol{Z}^\top\boldsymbol{y} + \boldsymbol{x}^\top\boldsymbol{\Gamma}\boldsymbol{x}/2 \\
&=\; \|\boldsymbol{y}\|^2/2 - \boldsymbol{y}^\top\boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^\top\boldsymbol{y}/2 + (\boldsymbol{x} - \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^\top\boldsymbol{y})^\top\boldsymbol{\Gamma}(\boldsymbol{x} - \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^\top\boldsymbol{y})/2 \\
&\geq\; \|\boldsymbol{y}\|^2/2 - \boldsymbol{y}^\top\boldsymbol{Z}\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^\top\boldsymbol{y}/2
\end{aligned}$$

with equality if and only if

$$\boldsymbol{x} \;=\; \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}^{\top}\boldsymbol{y}.$$

But often one is interested in computing parameter vectors $\boldsymbol{x}$ having only a few nonzero components while $g(\boldsymbol{x})$ is still relatively small. In the least squares example it may even happen that $d > n$, so the minimization problem has no unique solution. One way to deal with these problems is to minimize the penalized function

$$f_\lambda(\boldsymbol{x}) \;:=\; g(\boldsymbol{x}) + \lambda \sum_{j=1}^{d} |x_j|.$$

for some tuning parameter $\lambda > 0$. (This method is called **l**east **a**bsolute **s**hrinkage and **s**election **o**perator.) Here one can easily show that

$$\begin{aligned} Df_\lambda(\boldsymbol{x},\boldsymbol{v}) \;&=\; \sum_{j=1}^{d}\Big(\frac{\partial g}{\partial x_j}(\boldsymbol{x})v_j + \lambda\,\mathrm{sign}(x_j)v_j + \lambda 1_{[x_j=0]}|v_j|\Big) \\ &=\; \sum_{j\,:\,x_j\neq 0}\Big(\frac{\partial g}{\partial x_j}(\boldsymbol{x}) + \lambda\,\mathrm{sign}(x_j)\Big)v_j + \sum_{j\,:\,x_j=0}\Big(\frac{\partial g}{\partial x_j}(\boldsymbol{x})v_j + \lambda|v_j|\Big). \end{aligned}$$

From this representation one can easily deduce that a vector $\widehat{\boldsymbol{x}} \in \mathbb{R}^d$ minimizes f_λ if and only if

$$\frac{\partial g}{\partial x_j}(\widehat{\boldsymbol{x}}) \;\begin{cases} \in\; [-\lambda,\lambda] & \text{if } \widehat{x}_j = 0, \\ =\; -\lambda\,\mathrm{sign}(\widehat{x}_j) & \text{if } \widehat{x}_j \neq 0. \end{cases}$$

In particular, $\widehat{x}_j = 0$ whenever the modulus of $\partial g(\widehat{\boldsymbol{x}})/\partial x_j$ is less than λ. This explains why the LASSO approach tends to produce vectors $\widehat{\boldsymbol{x}}$ with several components equal to zero.

Exercise 3.39. Let $g : \mathbb{R}^d \to \mathbb{R}$ be given by $g(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{x}\|^2/2$ with given numbers $y_1,\ldots,y_n \in \mathbb{R}$ and a matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ such that $\sum_{i=1}^n Z_{ij}^2 = 1$ for $1 \le j \le d$. For $\lambda > 0$ we consider $f_\lambda(\boldsymbol{x}) := g(\boldsymbol{x}) + \lambda\sum_{j=1}^d |x_j|$.

(a) Determine for $\boldsymbol{x} \in \mathbb{R}^d$ and any $k \in \{1,\ldots,d\}$ the (unique!) vector

$$\boldsymbol{\psi}^{(k)}(\boldsymbol{x}) \;:=\; \arg\min\Big\{f_\lambda(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \in \mathbb{R}^d, \tilde{x}_j = x_j \text{ whenever } j \neq k\Big\}.$$

Show that this defines a continuous mapping $\boldsymbol{\psi}^{(k)} : \mathbb{R}^d \to \mathbb{R}^d$.

These mappings $\boldsymbol{\psi}^{(k)}$ are the basis of a simple iterative algorithm to minimize f_λ: For a given candidate $\boldsymbol{x}$ we compute $\boldsymbol{x}_{\mathrm{new}}$ as follows: After initializing $\boldsymbol{x}_{\mathrm{new}} \leftarrow \boldsymbol{x}$, we execute $\boldsymbol{x}_{\mathrm{new}} \leftarrow \boldsymbol{\psi}^{(k)}(\boldsymbol{x}_{\mathrm{new}})$ for $k = 1,\ldots,d$. As long as the difference $f_\lambda(\boldsymbol{x}) - f_\lambda(\boldsymbol{x}_{\mathrm{new}})$ is larger than a given small constant $\delta > 0$, we set $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\mathrm{new}}$ and repeat this loop. A justification of this algorithm will be given in the next chapter.

(b) Suppose that $\boldsymbol{x} = \boldsymbol{\psi}^{(k)}(\boldsymbol{x})$ for $1 \le k \le d$. Show that $\boldsymbol{x}$ is a minimizer of f_λ.

Example 3.40 (Taut Strings). Consider a vector $\boldsymbol{y} \in \mathbb{R}^n$ of observations $y_1, y_2, \ldots, y_n$ corresponding, for instance, to measurements at n consecutive time points. If these measurements are noisy, one would like to approximate $\boldsymbol{y}$ by a 'smoother' vector $\boldsymbol{x}$.

A classical approach is to minimize for a given tuning parameter $\lambda > 0$ the target function

$$\begin{aligned} f_\lambda(\boldsymbol{x}) \;&:=\; \sum_{i=1}^{n}(y_i - x_i)^2 + \lambda \sum_{j=1}^{n-1}(x_{j+1} - x_j)^2 \\ &= \|\boldsymbol{y} - \boldsymbol{x}\|^2 + \lambda \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} \end{aligned}$$

with the symmetric and positive semidefinite matrix $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ given by

$$S_{ij} \;:=\; \begin{cases} 1 & \text{if } i = j \in \{1, n\}, \\ 2 & \text{if } 1 < i = j < n, \\ 1 & \text{if } |i - j| = 1, \\ 0 & \text{else.} \end{cases}$$

The first summand $\|\boldsymbol{y} - \boldsymbol{x}\|^2$ of $f_\lambda(\boldsymbol{x})$ measures the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ while the second summand $\lambda \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$ quantifies how strongly consecutive components of $\boldsymbol{x}$ differ. This function f_λ has a unique minimizer $\widehat{\boldsymbol{x}}_\lambda$:

$$\begin{aligned} f_\lambda(\boldsymbol{x}) \;&= \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^\top \boldsymbol{y} + \boldsymbol{x}^\top (\boldsymbol{I}_n + \lambda \boldsymbol{S})\boldsymbol{x} \\ &= \|\boldsymbol{y}\|^2 - \boldsymbol{y}^\top (\boldsymbol{I}_n + \lambda \boldsymbol{S})^{-1}\boldsymbol{y} \\ &\quad + \big(\boldsymbol{x} - (\boldsymbol{I}_n + \lambda \boldsymbol{S})^{-1}\boldsymbol{y}\big)^\top (\boldsymbol{I}_n + \lambda \boldsymbol{S})\big(\boldsymbol{x} - (\boldsymbol{I}_n + \lambda \boldsymbol{S})^{-1}\boldsymbol{y}\big) \\ &\geq \|\boldsymbol{y}\|^2 - \boldsymbol{y}^\top (\boldsymbol{I}_n + \lambda \boldsymbol{S})^{-1}\boldsymbol{y} \end{aligned}$$

with equality if and only if

$$\boldsymbol{x} \;=\; (\boldsymbol{I}_n + \lambda \boldsymbol{S})^{-1}\boldsymbol{y}.$$

Let us now consider an alternative approach: We want to minimize

$$f_\lambda(\boldsymbol{x}) \;:=\; \|\boldsymbol{y} - \boldsymbol{x}\|^2/2 + \lambda \sum_{j=1}^{n-1} |x_{j+1} - x_j|.$$

The latter penalty term favours vectors $\boldsymbol{x}$ such that $x_j = x_{j+1}$ for several indices $j < n$. Note first that f_λ is a strictly convex function such that $f_\lambda(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$. This implies that f_λ has a unique minimizer $\widehat{\boldsymbol{x}}_\lambda$, see also Section 3.5.3. Now we characterize $\widehat{\boldsymbol{x}}_\lambda$ by means of one-sided directional derivatives. One can easily show that for arbitrary vectors $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^n$,

$$Df_\lambda(\boldsymbol{x}, \boldsymbol{v}) \;=\; (\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{v} + \lambda \sum_{j=1}^{n-1} \big(\operatorname{sign}(x_{j+1} - x_j)(v_{j+1} - v_j) + 1_{[x_j = x_{j+1}]}|v_{j+1} - v_j|\big).$$

This does not look too illuminating at first glance, but plugging in some special vectors $\boldsymbol{v}$ reveals interesting information about $\widehat{\boldsymbol{x}}_\lambda$. Specifically, let $\boldsymbol{b}_k := (1_{[i \leq k]})_{i=1}^n$ for $k = 1, 2, \ldots, n$.

If $\boldsymbol{v} = \pm \boldsymbol{b}_n$, then

$$Df_\lambda(\boldsymbol{x}, \boldsymbol{v}) \;=\; \pm(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{b}_n \;=\; \pm\Big(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\Big).$$

Hence a first necessary condition for $\boldsymbol{x} = \widehat{\boldsymbol{x}}_\lambda$ is

$$\sum_{i=1}^{n} x_i \;=\; \sum_{i=1}^{n} y_i. \tag{3.8}$$

Next consider $\boldsymbol{v} = \pm\boldsymbol{b}_k$ for some index $1 \le k < n$. Then

$$\begin{aligned} Df_\lambda(\boldsymbol{x},\boldsymbol{v}) &= \pm(\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{b}_k \mp \lambda \operatorname{sign}(x_{k+1}-x_k) + \lambda 1_{[x_k=x_{k+1}]} \\ &= \pm\Big(\sum_{i=1}^k x_i - \sum_{i=1}^k y_i\Big) \mp \lambda \operatorname{sign}(x_{k+1}-x_k) + \lambda 1_{[x_k=x_{k+1}]}. \end{aligned}$$

This leads to further necessary conditions for $\boldsymbol{x} = \widehat{\boldsymbol{x}}_\lambda$: For $1 \le k < n$,

$$\Big|\sum_{i=1}^k x_i - \sum_{i=1}^k y_i\Big| \;\le\; \lambda, \tag{3.9}$$

$$\sum_{i=1}^k x_i \;=\; \sum_{i=1}^k y_i + \lambda \operatorname{sign}(x_{k+1}-x_k) \quad \text{if } x_k \ne x_{k+1}. \tag{3.10}$$

In fact, Conditions (3.8), (3.9) and (3.10) are even sufficient for $\boldsymbol{x} = \widehat{\boldsymbol{x}}_\lambda$. This can be verified by writing an arbitrary vector $\boldsymbol{v}$ as

$$\boldsymbol{v} \;=\; \sum_{k=1}^{n-1}(v_k - v_{k+1})\boldsymbol{b}_k + v_n\boldsymbol{b}_n$$

and noting that

$$\begin{aligned} Df_\lambda(\boldsymbol{x},\boldsymbol{v}) &= \sum_{k=1}^{n-1}(v_k - v_{k+1})(\boldsymbol{x}-\boldsymbol{y})^\top\boldsymbol{b}_k + v_n \underbrace{(\boldsymbol{x}-\boldsymbol{y})^\top\boldsymbol{b}_n}_{=0} \\ &\quad + \lambda\sum_{j=1}^{n-1}\big(\operatorname{sign}(x_{j+1}-x_j)(v_{j+1}-v_j) + 1_{[x_j=x_{j+1}]}|v_{j+1}-v_j|\big) \\ &= \sum_{k=1}^{n-1}(v_k - v_{k+1})(\boldsymbol{x}-\boldsymbol{y})^\top\boldsymbol{b}_k \\ &\quad + \lambda\sum_{k=1}^{n-1}\big(-(v_k - v_{k+1})\operatorname{sign}(x_{k+1}-x_k) + 1_{[x_j=x_{j+1}]}|v_k - v_{k+1}|\big) \\ &= \sum_{k=1}^{n-1}|v_k - v_{k+1}| Df_\lambda\big(\boldsymbol{x}, \operatorname{sign}(v_k - v_{k+1})\boldsymbol{b}_k\big) \end{aligned}$$

is nonnegative.

There is a nice geometrical interpretation of Conditions (3.8), (3.9) and (3.10): Any vector $\boldsymbol{w} \in \mathbb{R}^n$ can be identified with its partial sum function,

$$\{0,1,\dots,n\} \ni k \;\mapsto\; S_{\boldsymbol{w}}(k) \;:=\; \begin{cases} 0 & \text{if } k = 0, \\ \sum_{i=1}^k w_i & \text{if } k > 0. \end{cases}$$

Condition (3.8) is equivalent to saying that $S_{\boldsymbol{x}}$ and $S_{\boldsymbol{y}}$ coincide at 0 and n. Condition (3.9) means that $S_{\boldsymbol{y}} - \lambda \le S_{\boldsymbol{x}} \le S_{\boldsymbol{y}} + \lambda$. Condition (3.10) means the following: If $S_{\boldsymbol{x}}$ bends upwards at an index $k \in \{1,\dots,n-1\}$ (i.e. $x_k < x_{k+1}$), then it hits the upper boundary $S_{\boldsymbol{y}} + \lambda$ there; if it bends downwards, then it hits the lower boundary $S_{\boldsymbol{y}} - \lambda$.

This is illustrated in Figures 3.1, 3.2 and 3.3. For $n = 70$ and $\lambda = 2, 1, 0.5$, respectively, the upper panel shows a scatter plot of the pairs (i, y_i) (black bullets) and $(i, \widehat{x}_{\lambda,i})$ (blue circles and line). In the lower panel one sees the partial sum function $S_{\boldsymbol{y}}$ (thin black line) and the same function $\pm\lambda$ on $\{1, \ldots, n-1\}$ (black lines and bullets). In between one sees the partial sum function $S_{\widehat{\boldsymbol{x}}_\lambda}$ (thick blue line). All these functions are interpolated linearly between consecutive integers. Obviously the fitted vector $\widehat{\boldsymbol{x}}_\lambda$ gets closer to $\boldsymbol{y}$ if λ is lowered.

When looking at the lower panels of these figures, one can imagine the following procedure: Suppose the coordinate system is a wooden board, and each bullet represents a nail. Now we connect the leftmost and rightmost nail with a string lying loosely between the lower and upper row of nails. Now we pull the string tight and obtain the graph of the partial sum function of $\widehat{\boldsymbol{x}}_\lambda$. This explains the name 'taut string' method. The actual computation is possible in $O(n)$ steps, but the precise algorithm is beyond the scope of this lecture.

3.4 Smoothness Properties

In this section we show that convex functions have almost automatically certain smoothness properties. Throughout this section let $(\boldsymbol{V}, \|\cdot\|)$ be a real normed vector space with closed unit ball $\boldsymbol{B}$, and let $f : \boldsymbol{V} \to (-\infty, \infty]$ be a convex function.

3.4.1 Local Lipschitz continuity

The next two results are about local Lipschitz continuity of convex functions.

Theorem 3.41 (Local Lipschitz continuity, I). *Suppose that for some $\boldsymbol{x} \in \boldsymbol{V}$ and $\epsilon > 0$ the function f is bounded from above on the set $\boldsymbol{x} + \epsilon\boldsymbol{B}$. Then f is Lipschitz continuous on $\boldsymbol{x} + \delta\boldsymbol{B}$ whenever $0 < \delta < \epsilon$.*

Theorem 3.42 (Local Lipschitz continuity, II). *Let $\boldsymbol{V} = \mathbb{R}^d$. Then f is continuous on the interior of $\operatorname{dom}(f)$. It is even Lipschitz continuous on any compact subset of the interior of $\operatorname{dom}(f)$.*

Proof of Theorem 3.41. For two different points $\boldsymbol{y}, \boldsymbol{z} \in \operatorname{dom}(f)$ we consider the mapping $\tilde{f} : \mathbb{R} \to (-\infty, \infty]$ with $\tilde{f}(t) := f(\boldsymbol{y} + t(\boldsymbol{z} - \boldsymbol{y}))$ and the unit vector $\boldsymbol{v} := \|\boldsymbol{z} - \boldsymbol{y}\|^{-1}(\boldsymbol{z} - \boldsymbol{y})$. Then

$$f(\boldsymbol{z}) - f(\boldsymbol{y}) \begin{cases} \leq \tilde{f}'(1+) \;=\; +\|\boldsymbol{z} - \boldsymbol{y}\| Df(\boldsymbol{z}, +\boldsymbol{v}), \\ \geq \tilde{f}'(0-) \;=\; -\|\boldsymbol{z} - \boldsymbol{y}\| Df(\boldsymbol{y}, -\boldsymbol{v}). \end{cases}$$

Hence it suffices to show that

$$\sup_{\boldsymbol{y} \in \boldsymbol{x} + \delta\boldsymbol{B}, \boldsymbol{v} \in \boldsymbol{B}} Df(\boldsymbol{y}, \boldsymbol{v}) \;<\; \infty$$

for $0 < \delta < \epsilon$.

To this end let $M := \sup_{\boldsymbol{x} + \epsilon\boldsymbol{B}}(f) < \infty$. For any $\boldsymbol{y} \in \boldsymbol{x} + \epsilon\boldsymbol{B}$, the vector $\widetilde{\boldsymbol{y}} = \boldsymbol{x} - (\boldsymbol{y} - \boldsymbol{x})$ belongs to $\boldsymbol{x} + \epsilon\boldsymbol{B}$, too, so it follows from $\boldsymbol{x} = (\boldsymbol{y} + \widetilde{\boldsymbol{y}})/2$ that $(f(\boldsymbol{y}) + f(\widetilde{\boldsymbol{y}}))/2 \geq f(\boldsymbol{x})$, whence

$$f(\boldsymbol{y}) \;\geq\; 2f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{y}}) \;\geq\; 2f(\boldsymbol{x}) - M.$$

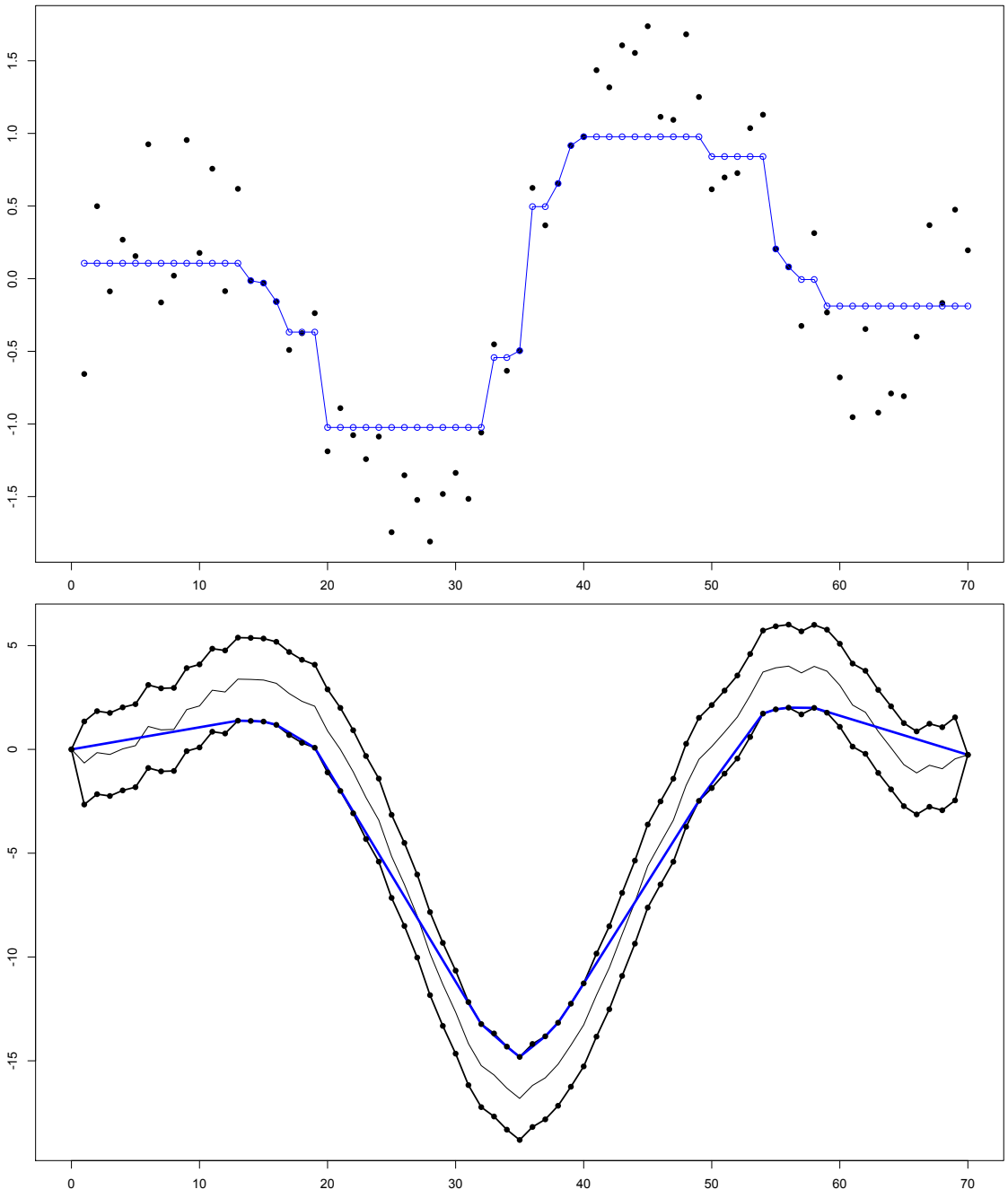

Figure 3.1: Illustration of the Taut String method, $\lambda = 2$.

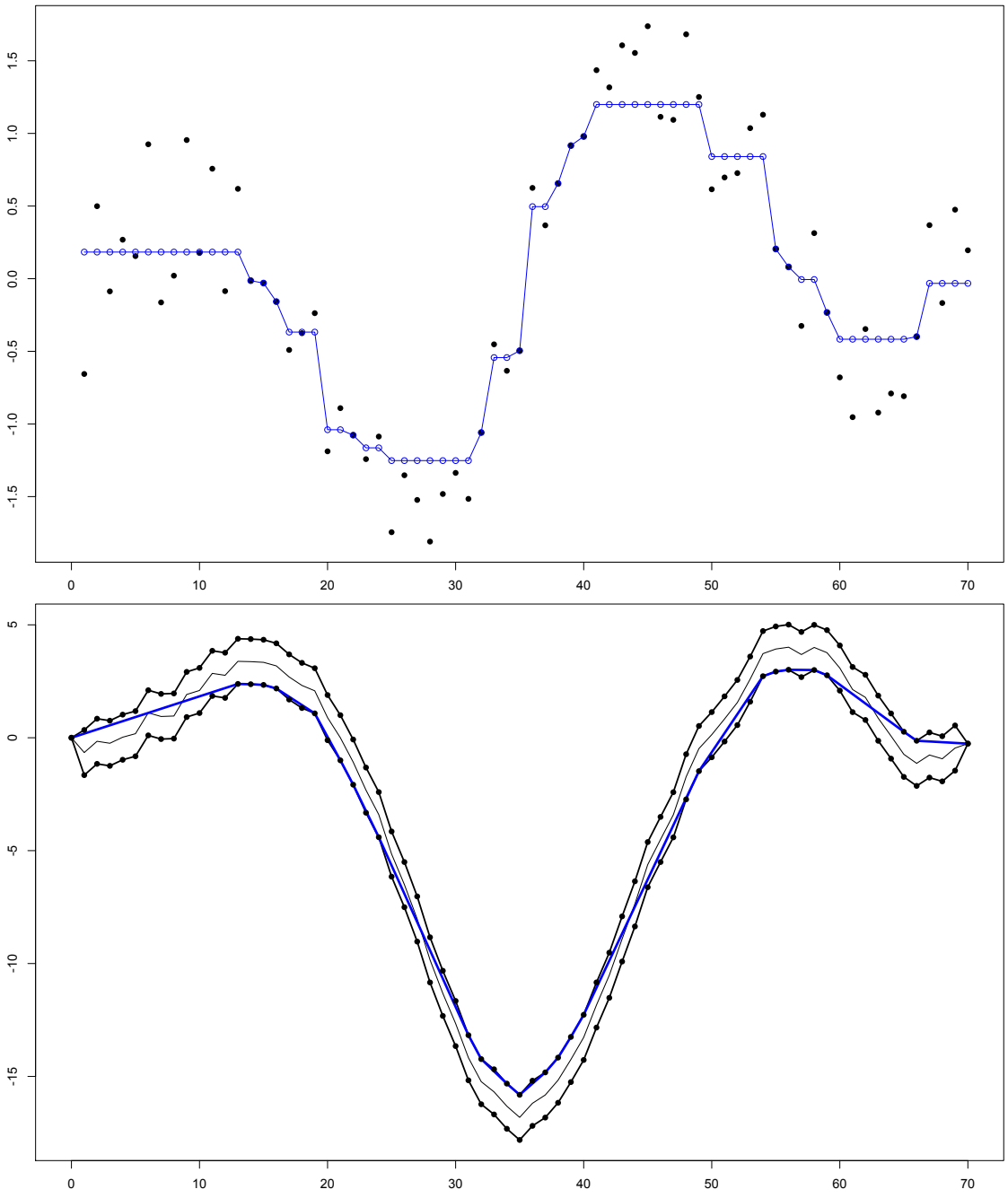

Figure 3.2: Illustration of the Taut String method, $\lambda = 1$.

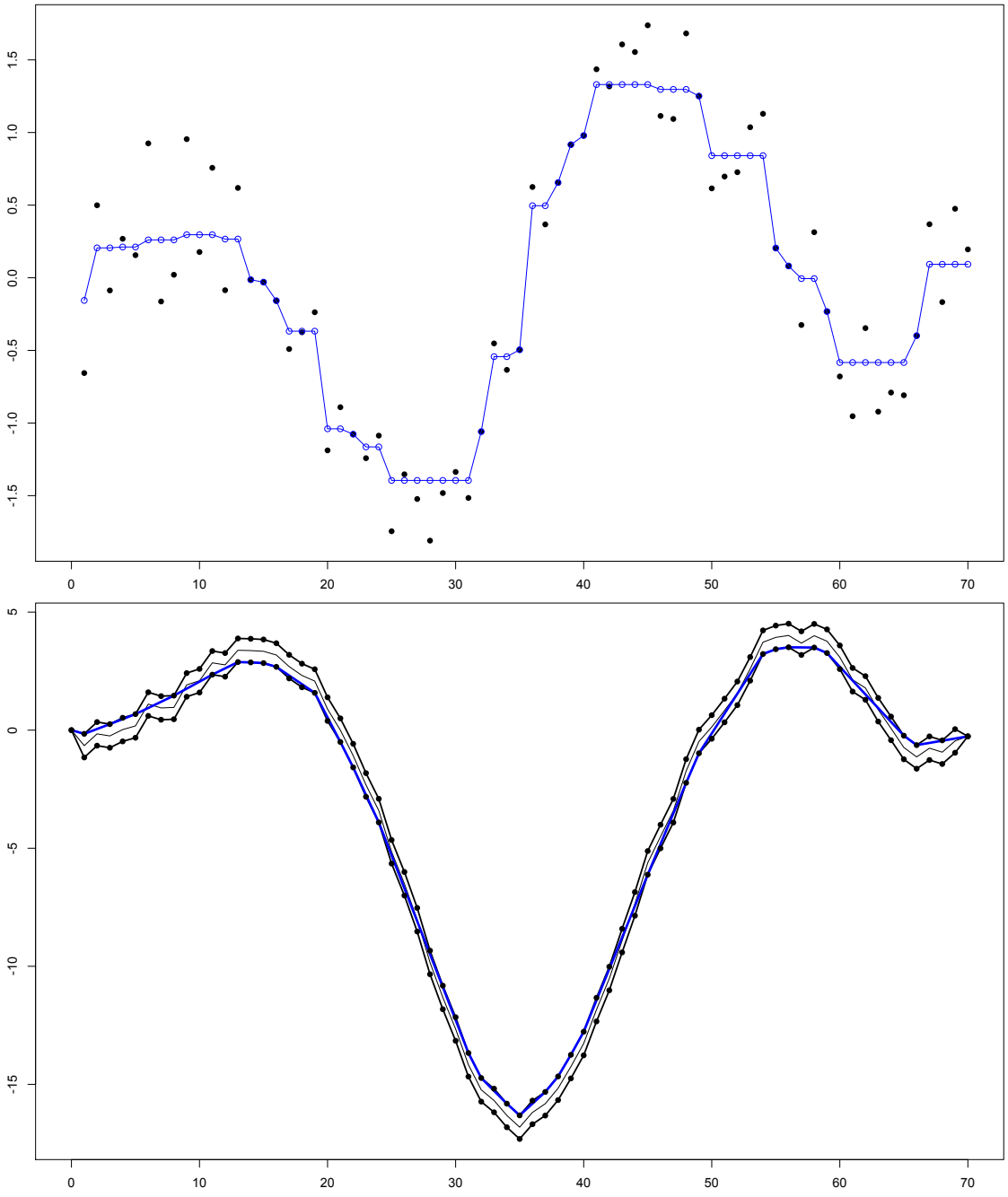

Figure 3.3: Illustration of the Taut String method, $\lambda = 0.5$.

This shows that

$$\inf_{\boldsymbol{x}+\epsilon\boldsymbol{B}} f \;\geq\; 2f(\boldsymbol{x}) - M.$$

Hence for $\boldsymbol{y} \in \boldsymbol{x} + \delta\boldsymbol{B}$ and $\boldsymbol{v} \in \boldsymbol{B}$,

$$Df(\boldsymbol{y}, \boldsymbol{v}) \;\leq\; \frac{f(\boldsymbol{y} + (\epsilon - \delta)\boldsymbol{v}) - f(\boldsymbol{y})}{\epsilon - \delta} \;\leq\; \frac{M - (2f(\boldsymbol{x}) - M)}{\epsilon - \delta} \;=\; \frac{2(M - f(\boldsymbol{x}))}{\epsilon - \delta}. \qquad \square$$

Proof of Theorem 3.42. Since all norms on $\mathbb{R}^d$ are equivalent, we consider temporarily the norm $\|\boldsymbol{v}\| := \max_i |v_i|$ with the corresponding unit ball $\boldsymbol{B} = [-1, 1]^d$. Now let $\boldsymbol{x}$ be an interior point of $\mathrm{dom}(f)$. Then there exists an $\epsilon = \epsilon(\boldsymbol{x}) > 0$ such that $\boldsymbol{x} + \epsilon\boldsymbol{B} \subset \mathrm{dom}(f)$. Since the cube $\boldsymbol{x} + \epsilon\boldsymbol{B}$ is the convex hull of the finite set

$$\big\{\boldsymbol{x} + \epsilon\boldsymbol{\xi} : \boldsymbol{\xi} \in \{-1, 1\}^d\big\},$$

it follows from Lemma 3.25 that f is bounded from above on $\boldsymbol{x} + \epsilon\boldsymbol{B}$. In particular, Theorem 3.41 implies that f is Lipschitz continuous on $\boldsymbol{x} + \delta\boldsymbol{B}$, provided that $0 < \delta < \epsilon$.

Now let $\boldsymbol{K}$ be a compact subset of $\mathrm{interior}(\mathrm{dom}(f))$. Then there exists an $\epsilon > 0$ such that $\boldsymbol{K} + \epsilon\boldsymbol{B}$ is contained in $\mathrm{interior}(\mathrm{dom}(f))$ as well. Since the latter set $\boldsymbol{K} + \epsilon\boldsymbol{B}$ is compact, too, it follows from continuity of f on $\mathrm{interior}(\mathrm{dom}(f))$ that

$$\tilde{L} \;:=\; \sup_{\boldsymbol{y},\boldsymbol{z} \in \boldsymbol{K}+\epsilon\boldsymbol{B}} \big|f(\boldsymbol{y}) - f(\boldsymbol{z})\big| \;<\; \infty.$$

Consequently, for arbitrary $\boldsymbol{x} \in \boldsymbol{K}$ and unit vectors $\boldsymbol{v} \in \boldsymbol{B}$,

$$Df(\boldsymbol{x}, \boldsymbol{v}) \;\leq\; \epsilon^{-1}\big(f(\boldsymbol{x} + \epsilon\boldsymbol{v}) - f(\boldsymbol{x})\big) \;\leq\; \epsilon^{-1}\tilde{L}.$$

As in the proof of Theorem 3.41 one can deduce from this inequality that f is Lipschitz continuous on $\boldsymbol{K}$ with Lipschitz constant $\epsilon^{-1}\tilde{L}$. $\square$

3.4.2 Subdifferentials and smoothness

A linear functional $L : \boldsymbol{V} \to \mathbb{R}$ is called *subdifferential of f at the point $\boldsymbol{x} \in \mathrm{dom}(f)$* if

$$f(\boldsymbol{x}) + L(\boldsymbol{v}) \;\leq\; f(\boldsymbol{x} + \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}. \tag{3.11}$$

It follows from our definition of directional derivatives and inequality (3.7) that condition (3.11) is equivalent to

$$L(\boldsymbol{v}) \;\leq\; Df(\boldsymbol{x}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}. \tag{3.12}$$

In case of $Df(\boldsymbol{x}, -\boldsymbol{v}) = -Df(\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}$ for arbitrary $\boldsymbol{v} \in \boldsymbol{V}$, the function $Df(\boldsymbol{x}, \cdot)$ itself is the unique subdifferential of f at $\boldsymbol{x}$.

Exercise 3.43 (Sublinearity and linearity). Let $L : \boldsymbol{V} \to \mathbb{R}$ be sublinear, i.e. for arbitrary $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{V}$ and $\lambda \geq 0$,

$$L(\lambda\boldsymbol{v}) \;=\; \lambda L(\boldsymbol{v}) \quad \text{and} \quad L(\boldsymbol{v} + \boldsymbol{w}) \;\leq\; L(\boldsymbol{v}) + L(\boldsymbol{w}).$$

(a) Show that L is linear if and only if

$$L(-\boldsymbol{v}) \;=\; -L(\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{V}.$$

(b) Let $\boldsymbol{V} = \operatorname{span}(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$. Show that L is linear if and only if

$$L(-\boldsymbol{b}_i) \;=\; -L(\boldsymbol{b}_i) \quad \text{for } 1 \le i \le d.$$

Exercise 3.44 (Sublinearity and Lipschitz-continuity). Let $L : \mathbb{R}^d \to \mathbb{R}$ be sublinear. Show that for arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\big|L(\boldsymbol{x}) - L(\boldsymbol{y})\big| \;\le\; \begin{cases} \max\big\{L(\boldsymbol{v}) : \boldsymbol{v} \in \{-1,1\}^d\big\}\, \|\boldsymbol{x} - \boldsymbol{y}\|_\infty, \\ \max\big\{L(\boldsymbol{v}) : \boldsymbol{v} \in \{\pm\boldsymbol{e}_1, \ldots, \pm\boldsymbol{e}_d\}\big\}\, \|\boldsymbol{x} - \boldsymbol{y}\|_1, \end{cases}$$

where $\|\boldsymbol{x}\|_\infty := \max_{i=1,\ldots,d} |x_i|$, $\|\boldsymbol{x}\|_1 := \sum_{i=1}^d |x_i|$, and $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ is the standard basis of $\mathbb{R}^d$.

The next theorem shows that a convex function f on $\mathbb{R}^d$ is differentiable almost everywhere in the interior of $\operatorname{dom}(f)$.

Theorem 3.45. *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be convex. For any interior point of $\operatorname{dom}(f)$ there exists a subdifferential of f. The set of all interior points of $\operatorname{dom}(f)$ at which f is not differentiable is a Borel set with Lebesgue measure 0.*

Proof of Theorem 3.45. Let $\boldsymbol{x} \in \Omega := \operatorname{interior}(\operatorname{dom}(f))$. We have to show that there exists a linear function L on $\mathbb{R}^d$ such that $f(\boldsymbol{x}) + L \le f(\boldsymbol{x} + \cdot)$. For this we apply Theorem 2.22 (b) to the disjoint convex subsets $\boldsymbol{D}_1 := \{\boldsymbol{x}\} \times (-\infty, f(\boldsymbol{x}))$ and $\boldsymbol{D}_2 := \operatorname{epi}(f)$ of $\mathbb{R}^d \times \mathbb{R}$. There exists a nonzero vector $(\boldsymbol{w}, t) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\big\langle (\boldsymbol{x}, r), (\boldsymbol{w}, t) \big\rangle \;\le\; \big\langle (\boldsymbol{x} + \boldsymbol{v}, s), (\boldsymbol{w}, t) \big\rangle$$

whenever $r < f(\boldsymbol{x})$, $\boldsymbol{v} \in \mathbb{R}^d$ and $s \ge f(\boldsymbol{x} + \boldsymbol{v})$. Writing $\big\langle (\boldsymbol{x}, r), (\boldsymbol{w}, t) \big\rangle = \boldsymbol{x}^\top \boldsymbol{w} + rt$ and $\big\langle (\boldsymbol{x} + \boldsymbol{v}, s), (\boldsymbol{w}, t) \big\rangle = (\boldsymbol{x} + \boldsymbol{v})^\top \boldsymbol{w} + st$, we get the inequalities

$$rt \;\le\; \boldsymbol{v}^\top \boldsymbol{w} + st \quad \text{whenever } r < f(\boldsymbol{x}), \boldsymbol{v} \in \mathbb{R}^d, s \ge f(\boldsymbol{x} + \boldsymbol{v}). \tag{3.13}$$

For $\boldsymbol{v} = \boldsymbol{0}$ we see that $rt \le st$ for arbitrary $r < f(\boldsymbol{x}) \le s$, so $t \ge 0$. If $t = 0$, then $\boldsymbol{w} \ne \boldsymbol{0}$, and (3.13) could be rephrased as $\boldsymbol{v}^\top \boldsymbol{w} \ge 0$ for arbitrary $\boldsymbol{v} \in \mathbb{R}^d$ such that $\boldsymbol{x} + \boldsymbol{v} \in \operatorname{dom}(f)$. But this would contradict our assumption that $\boldsymbol{x}$ is an interior point of $\operatorname{dom}(f)$. Thus $t > 0$, and we may assume without loss of generality that $t = 1$. Then (3.13) reads

$$f(\boldsymbol{x}) \;\le\; \boldsymbol{v}^\top \boldsymbol{w} + f(\boldsymbol{x} + \boldsymbol{v}) \quad \text{for arbitrary } \boldsymbol{v} \in \mathbb{R}^d,$$

so $L(\boldsymbol{v}) := -\boldsymbol{v}^\top \boldsymbol{w}$ defines a subdifferential of f at $\boldsymbol{x}$.

Suppose that

$$Df(\boldsymbol{x}, -\boldsymbol{e}_i) \;=\; -Df(\boldsymbol{x}, \boldsymbol{e}_i) \quad \text{for } 1 \le i \le d, \tag{3.14}$$

where $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ is the standard basis of $\mathbb{R}^d$. Then $Df(\boldsymbol{x}, \cdot)$ is linear, see Exercise 3.43. But this implies that f is differentiable at $\boldsymbol{x}$, see Exercise 3.46. On the other hand, differentiability of f at $\boldsymbol{x}$ implies (3.14).

Consequently, the set of all $\boldsymbol{x} \in \Omega$ at which f is not differentiable coincides with $\bigcup_{i=1}^d \Omega_i$ with

$$\Omega_i \;:=\; \big\{\boldsymbol{x} \in \Omega : Df(\boldsymbol{x}, -\boldsymbol{e}_i) > -Df(\boldsymbol{x}, \boldsymbol{e}_i)\big\}.$$

One can easily verify that Ω_i is a Borel set, so it remains to be shown that Ω_i has Lebesgue measure 0. By Fubini's theorem,

$$\mathrm{Leb}_d(\Omega_i) \;=\; \int_{\boldsymbol{e}_i^\perp} \mathrm{Leb}_1(N_i(\boldsymbol{z}))\; \mathrm{Leb}_{d-1}(d\boldsymbol{z})$$

with

$$N_i(\boldsymbol{z}) \;:=\; \big\{t \in \mathbb{R} : \boldsymbol{z} + t\boldsymbol{e}_i \in \Omega, Df(\boldsymbol{z} + t\boldsymbol{e}_i, -\boldsymbol{e}_i) > -Df(\boldsymbol{z} + t\boldsymbol{e}_i, \boldsymbol{e}_i)\big\}.$$

But $g_{\boldsymbol{z}}(t) := f(\boldsymbol{z} + t\boldsymbol{e}_i)$ defines a convex function $g_{\boldsymbol{z}}$ on $\mathbb{R}$, and for $t \in \mathrm{dom}(g_{\boldsymbol{z}})$,

$$g'_{\boldsymbol{z}}(t-) \;=\; -Df(\boldsymbol{z} + t\boldsymbol{e}_i, -\boldsymbol{e}_i), \quad g'_{\boldsymbol{z}}(t+) \;=\; Df(\boldsymbol{z} + t\boldsymbol{e}_i, \boldsymbol{e}_i).$$

As shown in Section 3.1, the set $N_i(\boldsymbol{z}) = \{t \in \mathrm{dom}(g_{\boldsymbol{z}}) : g'_{\boldsymbol{z}}(t-) < g\boldsymbol{z}'(t+)\}$ is at most countable and thus has Lebesgue measure 0. □

Exercise 3.46 ("Subdifferentiability"). Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be convex, and let $\boldsymbol{x}$ be an interior point of $\mathrm{dom}(f)$. Show that

$$\sup_{\boldsymbol{v} \in \boldsymbol{B}} \Big| \frac{f(\boldsymbol{x} + t\boldsymbol{v}) - f(\boldsymbol{x})}{t} - Df(\boldsymbol{x}, \boldsymbol{v}) \Big| \;\to\; 0 \quad \text{as } t \downarrow 0.$$

3.4.3 Jensen's inequality and Bregman divergences

In what follows let $\boldsymbol{Y} = (Y_i)_{i=1}^d$ be a random vector such that $\mathbb{E}\,\|\boldsymbol{Y}\| < \infty$, and let $\boldsymbol{\mu} := (\mathbb{E}\,Y_i)_{i=1}^d$ be its mean.

Theorem 3.47 (Jensen). *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a convex and measurable function such that $\mathbb{P}(\boldsymbol{Y} \in \mathrm{dom}(f)) = 1$. Then $\boldsymbol{\mu} \in \mathrm{dom}(f)$ as well, and*

$$\mathbb{E}\, f(\boldsymbol{Y}) \;\geq\; f(\boldsymbol{\mu}).$$

If f is strictly convex on $\mathrm{dom}(f)$ and $\mathbb{P}(\boldsymbol{Y} \neq \boldsymbol{\mu}) > 0$, then even

$$\mathbb{E}\, f(\boldsymbol{Y}) \;>\; f(\boldsymbol{\mu}).$$

Before proving this theorem, let us verify an auxiliary result.

Lemma 3.48. *Let $\boldsymbol{C} \subset \mathbb{R}^d$ be a convex set.*

(a) *Either $\boldsymbol{C}$ is contained in a hyperplane or it has nonvoid interior. In the latter case,*

$$\boldsymbol{C} \;\subset\; \overline{\mathrm{interior}(\boldsymbol{C})} \;=\; \overline{\boldsymbol{C}}.$$

(b) *Suppose that $\boldsymbol{C}$ is a Borel set with $\mathbb{P}(\boldsymbol{Y} \in \boldsymbol{C}) = 1$. Then either*

$$\boldsymbol{\mu} \ \in \ \mathrm{interior}(\boldsymbol{C}),$$

or there exists a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that $\boldsymbol{u}^\top \boldsymbol{Y} = \boldsymbol{u}^\top \boldsymbol{\mu}$ almost surely.

Proof of Lemma 3.48. Concerning part (a), let $\boldsymbol{x}_0$ be an arbitrary point in $\boldsymbol{C}$, and let

$$\boldsymbol{W} \ := \ \mathrm{span}\{\boldsymbol{x} - \boldsymbol{x}_0 : \boldsymbol{x} \in \boldsymbol{C}\}.$$

If $\dim(\boldsymbol{W}) < d$, there exists a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ which is perpendicular to $\boldsymbol{W}$, and $\boldsymbol{C} \subset \boldsymbol{x}_0 + \boldsymbol{W}$ is contained in the hyperplane $\boldsymbol{H} := \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}^\top \boldsymbol{x} = \boldsymbol{u}^\top \boldsymbol{x}_0\}$. If $\dim(\boldsymbol{W}) = d$, the set $\boldsymbol{C}$ has nonvoid interior as shown in Lemma 2.41.

Now we change $\boldsymbol{x}_0$, if necessary, such that it is an interior point of $\boldsymbol{C}$. That means, for some $\epsilon > 0$,

$$\boldsymbol{x}_0 + \epsilon \boldsymbol{B} \ \subset \ \boldsymbol{C}.$$

For any point $\boldsymbol{x}_1 \in \boldsymbol{C}$ let $\boldsymbol{x}_\lambda := (1-\lambda)\boldsymbol{x}_0 + \lambda \boldsymbol{x}_1$ for $\lambda \in [0, 1)$. Then $\boldsymbol{x}_\lambda \to \boldsymbol{x}_1$ as $\lambda \uparrow 1$. But $\boldsymbol{x}_\lambda$ is an interior point of $\boldsymbol{C}$ for any $\lambda \in [0, 1)$, because

$$\boldsymbol{x}_\lambda + (1-\lambda)\epsilon \boldsymbol{B} \ = \ (1-\lambda)(\boldsymbol{x}_0 + \epsilon \boldsymbol{B}) + \lambda \boldsymbol{x}_1 \ \subset \ \boldsymbol{C}.$$

Thus $\boldsymbol{x}_1 \in \overline{\mathrm{interior}(\boldsymbol{C})}$. This shows that $\boldsymbol{C} \subset \overline{\mathrm{interior}(\boldsymbol{C})}$. The latter inclusion implies that even $\overline{\boldsymbol{C}} \subset \overline{\mathrm{interior}(\boldsymbol{C})}$. Since obviously, $\overline{\mathrm{interior}(\boldsymbol{C})} \subset \overline{\boldsymbol{C}}$, we end up with the equation $\overline{\boldsymbol{C}} = \overline{\mathrm{interior}(\boldsymbol{C})}$.

Now we prove part (b). If $\boldsymbol{C}$ has no interior points, part (a) shows that it is contained in a hyperplane. That means, for some unit vector $\boldsymbol{u} \in \mathbb{R}^d$ and $r \in \mathbb{R}$, $C \subset \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}^\top \boldsymbol{x} = r\}$. This implies that $\boldsymbol{u}^\top \boldsymbol{Y} = r$ almost surely, and taking the expectation of this equation yields $r = \boldsymbol{u}^\top \boldsymbol{\mu}$. Now let $\boldsymbol{C}$ have interior points. If $\boldsymbol{\mu}$ is not an interior point of $\boldsymbol{C}$, it lies on the boundary or in the complement of $\overline{\boldsymbol{C}}$. Since $\overline{\boldsymbol{C}}$ is a closed and convex set, there exists a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that

$$\boldsymbol{u}^\top \boldsymbol{\mu} \ \geq \ \boldsymbol{u}^\top \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \overline{\boldsymbol{C}},$$

see Corollaries 2.19 and 2.21. In particular, $\boldsymbol{u}^\top \boldsymbol{Y} \leq \boldsymbol{u}^\top \boldsymbol{\mu} = \mathbb{E}(\boldsymbol{u}^\top \boldsymbol{Y})$ almost surely, and this implies that even $\boldsymbol{u}^\top \boldsymbol{Y} = \boldsymbol{u}^\top \boldsymbol{\mu}$ almost surely. □

Proof of Theorem 3.47. We may assume without loss of generality that $\boldsymbol{\mu}$ is an interior point of $\mathrm{dom}(f)$. Otherwise we could deduce from Lemma 3.48 that $\mathbb{P}(\boldsymbol{Y} \in \boldsymbol{H}) = 1$ for some hyperplane $\boldsymbol{H} \subset \mathbb{R}^d$, so we could replace $\mathbb{R}^d$ with $\boldsymbol{H}$, and the latter set could be identified with $\mathbb{R}^{d-1}$. This reduction could be repeated until $\boldsymbol{\mu}$ is an interior point of $\mathrm{dom}(f)$ and still $d \geq 1$, or $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}_o) = 1$ for a fixed point $\boldsymbol{y}_o \in \mathbb{R}^d$. The latter case is obviously trivial.

If $\boldsymbol{\mu}$ is an interior point of $\mathrm{dom}(f)$, by Theorem 3.45 there exists a linear function $L : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(\boldsymbol{y}) \ \geq \ f(\boldsymbol{\mu}) + L(\boldsymbol{y} - \boldsymbol{\mu}) \quad \text{for arbitrary } \boldsymbol{y} \in \mathbb{R}^d.$$

Setting $\boldsymbol{y} = \boldsymbol{Y}$ and taking the expectation of both sides yields that

$$\mathbb{E}\, f(\boldsymbol{Y}) \;\geq\; f(\boldsymbol{\mu}) + \mathbb{E}\, L(\boldsymbol{Y} - \boldsymbol{\mu}) \;=\; f(\boldsymbol{\mu}) + L\big(\underbrace{\mathbb{E}(\boldsymbol{Y} - \boldsymbol{\mu})}_{=\mathbf{0}}\big) \;=\; f(\boldsymbol{\mu}).$$

If f is strictly convex on $\mathrm{dom}(f)$, then even $f(\boldsymbol{y}) > f(\boldsymbol{\mu}) + L(\boldsymbol{y} - \boldsymbol{\mu})$ for all $\boldsymbol{y} \in \mathbb{R}^d \setminus \{\boldsymbol{\mu}\}$, and we obtain the strict inequality $\mathbb{E}\, f(\boldsymbol{Y}) > f(\boldsymbol{\mu})$, unless $\mathbb{P}(\boldsymbol{Y} \neq \boldsymbol{\mu}) = 0$. □

The difference $\mathbb{E}\, f(\boldsymbol{Y}) - f(\boldsymbol{\mu})$ can be quantified in the special setting when $\mathrm{dom}(f)$ has nonvoid interior and f is continuously differentiable and strictly convex on $\mathrm{dom}(f)$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ we define

$$D_f(\boldsymbol{y}, \boldsymbol{x}) \;=\; f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}),$$

the so-called *Bregman divergence* of $\boldsymbol{y}$ from $\boldsymbol{x}$ (with respect to f). Note that $h(t) := f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$ is strictly convex in $t \in [0, 1]$ with derivative $h'(t) = \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})$, so

$$D_f(\boldsymbol{y}, \boldsymbol{x}) \;=\; h(1) - h(0) - h'(0) \;=\; \int_0^1 \big(h'(t) - h'(0)\big)\, dt \;\geq\; 0,$$

with equality if and only if $\boldsymbol{y} = \boldsymbol{x}$.

Theorem 3.49 (Minimizing expected Bregman divergence). *Suppose that $\mathbb{E}\, f(\boldsymbol{Y}) < \infty$. Then $\mathbb{E}\, D_f(\boldsymbol{Y}, \boldsymbol{x}) < \infty$ for arbitrary $\boldsymbol{x} \in \mathrm{dom}(f)$, and*

$$\mathbb{E}\, D_f(\boldsymbol{Y}, \boldsymbol{x}) \;\geq\; \mathbb{E}\, D_f(\boldsymbol{Y}, \boldsymbol{\mu})$$

with equality if and only if $\boldsymbol{x} = \boldsymbol{\mu}$.

More generally, for arbitrary points $\boldsymbol{x}, \boldsymbol{x}_o \in \mathrm{dom}(f)$,

$$\mathbb{E}\big(D_f(\boldsymbol{Y}, \boldsymbol{x}) - D_f(\boldsymbol{Y}, \boldsymbol{x}_o)\big)$$

exists in $\mathbb{R}$ and is minimal with respect to $\boldsymbol{x}$ if and only if $\boldsymbol{x} = \boldsymbol{\mu}$.

Proof of Theorem 3.49. Recall from Theorem 3.47 that $\boldsymbol{\mu} \in \mathrm{dom}(f)$. Note that

$$D_f(\boldsymbol{Y}, \boldsymbol{x}) - D_f(\boldsymbol{Y}, \boldsymbol{x}_o) \;=\; f(\boldsymbol{x}_o) + \nabla f(\boldsymbol{x}_o)^\top (\boldsymbol{Y} - \boldsymbol{x}_o) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{Y} - \boldsymbol{x})$$

is integrable, because $\mathbb{E}\, \|\boldsymbol{Y}\|$ is finite. The function $h(\boldsymbol{x}, \boldsymbol{x}_o) := \mathbb{E}\big(D_f(\boldsymbol{Y}, \boldsymbol{x}) - D_f(\boldsymbol{Y}, \boldsymbol{x}_o)\big)$ satisfies

$$\begin{aligned} h(\boldsymbol{x}, \boldsymbol{x}_o) - h(\boldsymbol{\mu}, \boldsymbol{x}_o) \;&=\; h(\boldsymbol{x}, \boldsymbol{\mu}) \\ &=\; \mathbb{E}\big[f(\boldsymbol{\mu}) + \nabla f(\boldsymbol{\mu})^\top (\boldsymbol{Y} - \boldsymbol{\mu}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{Y} - \boldsymbol{x})\big] \\ &=\; f(\boldsymbol{\mu}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{\mu} - \boldsymbol{x}) \\ &=\; D_f(\boldsymbol{\mu}, \boldsymbol{x}). \end{aligned}$$

This shows that $\boldsymbol{\mu}$ is the unique minimizer of $h(\cdot, \boldsymbol{x}_o)$ on $\mathrm{dom}(f)$. □

Remark 3.50. Note that under the assumptions of the previous theorem,

$$\mathbb{E}\, f(\boldsymbol{Y}) - f(\boldsymbol{\mu}) \;=\; \mathbb{E}\big[f(\boldsymbol{Y}) - f(\boldsymbol{\mu}) - \nabla f(\boldsymbol{\mu})^\top (\boldsymbol{Y} - \boldsymbol{\mu})\big] \;=\; \mathbb{E}\, D_f(\boldsymbol{Y}, \boldsymbol{\mu}).$$

Exercise 3.51. For $\boldsymbol{x} \in \mathbb{R}^d$ let

$$f(\boldsymbol{x}) \;:=\; \frac{\|\boldsymbol{x}\|^2}{\sqrt{1+\|\boldsymbol{x}\|^2}}.$$

Show that $f : \mathbb{R}^d \to \mathbb{R}$ is strictly convex and continuously differentiable, and determine the Bregman divergence $D_f(\boldsymbol{y}, \boldsymbol{x})$.

3.5 Lower Semicontinuity, Minimizers and Convex Conjugates

So far we know that convex functions on $\mathbb{R}^d$ behave quite nicely on the interior of their domain. From that perspective it looks a bit strange that we talk about a convex and measurable function in Jensen's inequality. To see that strange things may happen on the boundary, note that *any* function $f : \mathbb{R}^d \to (-\infty, \infty]$ with

$$f(\boldsymbol{x}) \begin{cases} = 0 & \text{if } \|\boldsymbol{x}\| < 1 \\ \geq 0 & \text{if } \|\boldsymbol{x}\| = 1 \\ = \infty & \text{if } \|\boldsymbol{x}\| > 1 \end{cases}$$

is convex. This follows from the fact that for the standard Euclidean norm $\|\cdot\|$, any point on the unit sphere $\partial \boldsymbol{B}$ is an extremal point of the closed unit ball $\boldsymbol{B}$. In the next subsection we introduce an additional regularity condition which has some useful consequences if applied to convex functions.

3.5.1 Lower semicontinuity

In this subsection we consider an arbitrary metric space $(\mathcal{X}, d)$ and study a special property of functions on $\mathcal{X}$.

Definition 3.52 (Lower semicontinuity). A function $f : \mathcal{X} \to [-\infty, \infty]$ is called *lower semicontinuous* if

$$f(x) \;\leq\; \liminf_{y \to x} f(y)$$

for arbitrary $x \in \mathcal{X}$.

The previous limes inferior is defined as follows:

$$\liminf_{y \to x} f(y) \;=\; \lim_{\delta \to 0+} \; \inf_{y \in \mathcal{X} \,:\, d(x,y) \leq \delta} f(y).$$

Note that the limit on the right-hand side exists in $[-\infty, \infty]$, because $\inf_{y \,:\, d(x,y) \leq \delta} f(y)$ is antitonic in $\delta > 0$. Note that for any sequence $(x_n)_n$ in $\mathcal{X}$ with limit x,

$$\liminf_{n \to \infty} \; f(x_n) \;\geq\; \liminf_{y \to x} \; f(y).$$

If we choose $x_n \in \mathcal{X}$ such that $d(x, x_n) \leq 1/n$ and $f(x_n) \leq \inf_{y \,:\, d(x,y) \leq 1/n} f(y) + 1/n$, then

$$\lim_{n \to \infty} \; f(x_n) \;=\; \liminf_{y \to x} \; f(y).$$

Lemma 3.53 (Characterizing lower semicontinuity). *For a function $f : \mathcal{X} \to [-\infty, \infty]$ on a metric space $(\mathcal{X}, d)$ the following three statements are equivalent:*

(i) *f is lower semicontinuous.*

(ii) *For any $r \in \mathbb{R}$, the set $\{x \in \mathcal{X} : f(x) \le r\}$ is a closed subset of $\mathcal{X}$.*

(iii) *The epigraph $\operatorname{epi}(f) = \big\{(x, r) \in \mathcal{X} \times \mathbb{R} : f(x) \le r\big\}$ is a closed subset of $\mathcal{X} \times \mathbb{R}$.*

In part (iii), we equip the cartesian product $\mathcal{X} \times \mathbb{R}$ with the metric

$$d\big((x, r), (y, s)\big) \;:=\; \sqrt{d(x, y)^2 + |r - s|^2}.$$

Alternatively, one could define $d\big((x, r), (y, s)\big)$ to be $d(x, y) + |r - s|$ or $\max\big(d(x, y), |r - s|\big)$.

Proof of Lemma 3.53. Suppose that f is lower semicontinuous. If $(x_n)_n$ is a sequence in $\{f \le r\}$ with limit $x \in \mathcal{X}$, then $f(x) \le \liminf_{n\to\infty} f(x_n) \le r$, so $x \in \{f \le r\}$, too. This shows that condition (ii) holds true.

Suppose that f satisfies (ii). Let $\big((x_n, r_n)\big)_n$ be a sequence in $\operatorname{epi}(f)$ with limit $(x, r) \in \mathcal{X} \times \mathbb{R}$. For arbitrary fixed $r' > r$ and sufficiently large n, $f(x_n) \le r_n \le r'$, so x is a limit point of a sequence in $\{f \le r'\}$. Since the latter set is closed by assumption, $x \in \{f \le r'\}$, so $f(x) \le r'$. Since r' can be chosen arbitrarily close to r, we may conclude that $f(x) \le r$, so $(x, r) \in \operatorname{epi}(f)$. This proves (iii).

Suppose that f satisfies (iii). For any $x \in \mathcal{X}$ let $(x_n)_n$ be a sequence in $\mathcal{X}$ with limit x such that $r := \lim_{n\to\infty} f(x_n) = \liminf_{y\to x} f(y)$. We have to show that $r \ge f(x)$. This is trivial in case of $r = \infty$. Otherwise let $r' > r$. Then for a suitable n_o, $f(x_n) \le r'$ for all $n \ge n_o$. Hence $(x, r') = \lim_{n\to\infty}(x_n, r')$, and $(x_n, r') \in \operatorname{epi}(f)$, for all $n \ge n_o$. By assumption (iii), the point (x, r') belongs to $\operatorname{epi}(f)$, too, whence $f(x) \le r'$. Since r' can be chosen arbitrarily close to r, this shows that $f(x) \le r$, i.e. f is lower semicontinuous. □

Exercise 3.54. Let $(f_\lambda)_{\lambda\in\Lambda}$ be a family of functions $f_\lambda : \mathcal{X} \to [-\infty, \infty]$ on some set $\mathcal{X}$, and let $f := \sup_{\lambda\in\Lambda} f_\lambda$ (pointwise).

(a) Show that

$$\operatorname{epi}(f) \;=\; \bigcap_{\lambda\in\Lambda} \operatorname{epi}(f_\lambda).$$

(b) Now suppose that $(\mathcal{X}, d)$ is a metric space and that all f_λ, $\lambda \in \Lambda$, are lower semicontinuous. Show that f is lower semicontinuous, too.

Exercise 3.55. Let $(\mathcal{X}, d)$ be a metric space, and let $f : \mathcal{X} \to [-\infty, \infty]$ be lower semicontinuous. Show that for any compact set $\mathcal{X}_o \subset \mathcal{X}$, $\arg\min_{\mathcal{X}_o} f$ is a compact (and thus nonvoid) set.

3.5.2 Convexity, lower semicontinuity and affine functions

Now we return to functions $f : \mathbb{R}^d \to (-\infty, \infty]$. As we shall see here and in other parts, combining convexity and lower semicontinuity is particularly fruitful. Recall that f is convex

and lower semicontinuous if and only if its epigraph

$$\operatorname{epi}(f) \;=\; \big\{(\boldsymbol{x}, r) \in \mathbb{R}^d \times \mathbb{R} : f(\boldsymbol{x}) \le r\big\}$$

is a closed and convex subset of $\mathbb{R}^d \times \mathbb{R}$. Then $\operatorname{epi}(f)$ is an intersection of closed halfspaces in $\mathbb{R}^d \times \mathbb{R}$. As shown in Exercise 3.57, the epigraph of an affine function $A : \mathbb{R}^d \to \mathbb{R}$ is a closed halfspace in $\mathbb{R}^d \times \mathbb{R}$. Thus the next result is rather plausible in view of Corollary 2.19:

Theorem 3.56 (Convex functions as pointwise suprema of affine functions). *For a function $f : \mathbb{R}^d \to (-\infty, \infty]$ the following two statements are equivalent:*

(i) *f is convex and lower semicontinuous.*

(ii) *f is the pointwise supremum of all affine functions $A : \mathbb{R}^d \to \mathbb{R}$ such that $A \le f$ on $\mathbb{R}^d$.*

Note that $\sup(\emptyset) := -\infty$, so part (ii) implies that there is some affine function A with $A \le f$.

Proof of Theorem 3.56. Let $\mathcal{A}$ be the set of all affine functions $A : \mathbb{R}^d \to \mathbb{R}$ such that $A \le f$, and suppose that $\mathcal{A}$ is nonvoid. Since every function $A \in \mathcal{A}$ is convex, real-valued and continuous, it follows from Lemma 3.20 (d) and Exercise 3.54 (b) that the pointwise supremum $\underline{f} := \sup_{A \in \mathcal{A}} A$ is a convex and lower semicontinuous function $\underline{f} : \mathbb{R}^d \to (-\infty, \infty]$ such that $\underline{f} \le f$. This shows already that condition (ii), which states that $\underline{f} \equiv f$, implies condition (i).

It remains to show that condition (i) implies condition (ii). For this it suffices to show that for any point $(\boldsymbol{x}_o, r_o) \in \mathbb{R}^d \times \mathbb{R}$ with $r_o < f(\boldsymbol{x}_o)$ there exists an $A \in \mathcal{A}$ such that $A(\boldsymbol{x}_o) \ge r_o$. In case of $f \equiv \infty$ this is trivial; just take the constant function $A \equiv r_o$. Hence let $\operatorname{dom}(f)$ be nonempty. In case of $\boldsymbol{x}_o \notin \overline{\operatorname{dom}(f)}$ let $\boldsymbol{x}_*$ be the metric projection of $\boldsymbol{x}_o$ onto $\overline{\operatorname{dom}(f)}$, and pick some number $r_* < f(\boldsymbol{x}_*)$. Otherwise let $(\boldsymbol{x}_*, r_*) := (\boldsymbol{x}_o, r_o)$. In both cases,

$$\boldsymbol{D} \;:=\; \operatorname{conv}\big\{(\boldsymbol{x}_o, r_o), (\boldsymbol{x}_*, r_*)\big\}$$

is a compact, convex subset of $\mathbb{R}^d \times \mathbb{R}$ with

$$\boldsymbol{D} \cap \operatorname{epi}(f) \;=\; \emptyset.$$

Hence by Theorem 2.22 (a) there exists a nonzero vector $(\boldsymbol{b}, c) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\max_{(\boldsymbol{x}, r) \in \boldsymbol{D}} (\boldsymbol{b}^\top \boldsymbol{x} + cr) \;<\; \inf_{(\boldsymbol{y}, s) \in \operatorname{epi}(f)} (\boldsymbol{b}^\top \boldsymbol{y} + cs). \tag{3.15}$$

Suppose that $c \le 0$. Then consider a sequence $(\boldsymbol{y}_n)_n$ in $\operatorname{dom}(f)$ with limit $\boldsymbol{x}_*$. By lower semicontinuity of f,

$$\limsup_{n \to \infty} (\boldsymbol{b}^\top \boldsymbol{y}_n + c f(\boldsymbol{y}_n)) \;\le\; \boldsymbol{b}^\top \boldsymbol{x}_* + c f(\boldsymbol{x}_*) \;\le\; \boldsymbol{b}^\top \boldsymbol{x}_* + c r_*,$$

a contradiction to (3.15). Hence $c > 0$ and we may assume without loss of generality that $c = 1$. Then (3.15) implies that

$$\boldsymbol{b}^\top \boldsymbol{x}_o + r_o \;<\; \boldsymbol{b}^\top \boldsymbol{y} + f(\boldsymbol{y}) \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^d.$$

In other words, $A(\boldsymbol{y}) := \boldsymbol{b}^\top \boldsymbol{x}_o + r_o - \boldsymbol{b}^\top \boldsymbol{y}$ defines an affine function such that $A < f$ and $A(\boldsymbol{x}_o) = r_o$. □

Exercise 3.57 (Epigraphs of affine functions). Show that a function $f : \mathbb{R}^d \to \mathbb{R}$ is affine if and only if its epigraph is a closed halfspace in $\mathbb{R}^d \times \mathbb{R}$.

3.5.3 Existence of minimizers and coercivity

Let us define an important property in connection with minimization of functions:

Definition 3.58 (Coercivity). A function $f : \mathbb{R}^d \to (-\infty, \infty]$ is called *coercive* if

$$f(\boldsymbol{x}) \;\to\; \infty \quad \text{as } \|\boldsymbol{x}\| \to \infty.$$

The next lemma shows that lower semicontinuity and coercivity of a function imply the existence of a compact set of minimizers:

Lemma 3.59 (Existence of minimizers). *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be lower semicontinuous and coercive with $\operatorname{dom}(f) \neq \emptyset$. Then*

$$\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \;=\; \big\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) = \inf_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{y})\big\}$$

is compact (and thus nonvoid).

Proof of Lemma 3.59. Let $(\boldsymbol{x}_n)_n$ be a sequence in $\mathbb{R}^d$ such that $f(\boldsymbol{x}_n)$ converges to $\inf(f) := \inf_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{y})$ as $n \to \infty$. It follows from f being coercive that $(\boldsymbol{x}_n)_n$ is bounded. Thus we may replace $(\boldsymbol{x}_n)_n$ with a subsequence, if necessary, such that $(\boldsymbol{x}_n)_n$ has a limit $\boldsymbol{x} \in \mathbb{R}^d$. By lower semicontinuity of f,

$$f(\boldsymbol{x}) \;\le\; \liminf_{n\to\infty} f(\boldsymbol{x}_n) \;=\; \inf(f),$$

whence $f(\boldsymbol{x}) = \inf(f)$. This shows that $\arg\min(f) := \{f = \inf(f)\}$ is nonempty. Since $\arg\min(f)$ equals $\{f \le \inf(f)\}$, lower semicontinuity of f implies that it is closed, while coercivity of f implies that it is bounded. Consequently, $\arg\min(f)$ is compact. □

For convex functions one can give a rather complete picture. First of all, for a convex and lower semicontinuous function, coercivity and compactness of its set of minimizers are equivalent.

Theorem 3.60 (Convexity and coercivity). *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be convex and lower semicontinuous with $\operatorname{dom}(f) \neq \emptyset$. Then f is coercive if and only if the set*

$$\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \;=\; \Big\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) = \inf_{\boldsymbol{y}\in\mathbb{R}^d} f(\boldsymbol{y})\Big\}$$

is compact (and thus nonvoid).

For nondegenerate convex functions which are strictly convex on their domain, it is clear that they have at most one minimizer. Hence Theorem 3.60 yields the following corollaries:

Corollary 3.61. *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be lower semicontinuous and convex with $\operatorname{dom}(f) \neq \emptyset$. If f is strictly convex on $\operatorname{dom}(f)$, then there exists a unique minimizer of f if and only if f is coercive.*

Since a real-valued convex function on $\mathbb{R}^d$ is automatically continuous, there is a simplified version of the previous corollary for this situation:

Corollary 3.62. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strictly convex. There exists a unique minimizer of f if and only if f is coercive.*

Proof of Theorem 3.60. We know already from Lemma 3.59 that the set $\arg\min(f) = \{f \leq \inf(f)\}$ is compact if f is coercive. It remains to show that compactness of $\arg\min(f)$ implies coercivity. To this end let $\boldsymbol{x}_o \in \arg\min(f)$ and let $r > 0$ such that $\arg\min(f)$ is contained in $\boldsymbol{x}_o + r\boldsymbol{B}$. Since the sphere with center $\boldsymbol{x}_o$ and radius $r+1$ is compact, lower semicontinuity of f implies the existence of

$$c := \min_{\boldsymbol{x}\in\mathbb{R}^d\,:\,\|\boldsymbol{x}-\boldsymbol{x}_o\|=r+1} f(\boldsymbol{x}) \;>\; f(\boldsymbol{x}_o);$$

see also Exercise 3.55 or the proof of Lemma 3.59. For any $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x} - \boldsymbol{x}_o\| \geq r+1$ we may write $\boldsymbol{x} = \boldsymbol{x}_o + \|\boldsymbol{x} - \boldsymbol{x}_o\|\boldsymbol{u}$ with a unit vector $\boldsymbol{u}$, and by convexity of $f(\boldsymbol{x}_o + t\boldsymbol{u})$ in $t \in \mathbb{R}$,

$$\begin{aligned} f(\boldsymbol{x}) \;&=\; f(\boldsymbol{x}_o) + \big(f(\boldsymbol{x}_o + \|\boldsymbol{x} - \boldsymbol{x}_o\|\boldsymbol{u}) - f(\boldsymbol{x}_o)\big) \\ &\geq\; f(\boldsymbol{x}_o) + \|\boldsymbol{x} - \boldsymbol{x}_o\|\frac{f(\boldsymbol{x}_o + (r+1)\boldsymbol{u}) - f(\boldsymbol{x}_o)}{r+1} \\ &\geq\; f(\boldsymbol{x}_o) + \|\boldsymbol{x} - \boldsymbol{x}_o\|\frac{c - f(\boldsymbol{x}_o)}{r+1} \\ &\geq\; f(\boldsymbol{x}_o) + \big(\|\boldsymbol{x}\| - \|\boldsymbol{x}_o\|\big)\frac{c - f(\boldsymbol{x}_o)}{r+1}. \end{aligned}$$

The right-hand side converges to ∞ as $\|\boldsymbol{x}\| \to \infty$, whence f is coercive. □

We end this subsection with a specific characterization of coercivity in terms of directional derivatives which is often rather useful:

Lemma 3.63 (Criterion for coercivity). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. Then f is coercive if and only if for any fixed unit vector $\boldsymbol{u} \in \mathbb{R}^d$,*

$$\lim_{t\to\infty} Df(t\boldsymbol{u}, \boldsymbol{u}) \;>\; 0. \tag{3.16}$$

Note that convexity of f implies that for a fixed unit vector $\boldsymbol{u} \in \mathbb{R}^d$, the function $h_{\boldsymbol{u}} : \mathbb{R} \to \mathbb{R}$, $h_{\boldsymbol{u}}(t) := f(t\boldsymbol{u})$, is convex with $h'_{\boldsymbol{u}}(t+) = Df(t\boldsymbol{u}, \boldsymbol{u})$. Thus $Df(t\boldsymbol{u}, \boldsymbol{u})$ is isotonic in t, and the limit $\lim_{t\to\infty} Df(t\boldsymbol{u}, \boldsymbol{u})$ exists in $(-\infty, \infty]$.

Proof of Lemma 3.63. Suppose first that f is coercive. Then for sufficiently large $r > 0$,

$$c := \min_{\boldsymbol{x}\in r\partial\boldsymbol{B}} f(\boldsymbol{x}) \;>\; f(\boldsymbol{0}).$$

But by convexity of f, for any fixed unit vector $\boldsymbol{u} \in \partial\boldsymbol{B}$, the function $h_{\boldsymbol{u}}(t) := f(t\boldsymbol{u})$ is convex in $t \in \mathbb{R}$ with $h'_{\boldsymbol{u}}(t+) = Df(t\boldsymbol{u}, \boldsymbol{u})$ being isotonic in t. Thus

$$\lim_{t\to\infty} Df(t\boldsymbol{u}, \boldsymbol{u}) \;\geq\; h'_{\boldsymbol{u}}(r+) \;\geq\; r^{-1}(h_{\boldsymbol{u}}(r) - h_{\boldsymbol{u}}(0)) \;\geq\; r^{-1}(c - f(\boldsymbol{0})) \;>\; 0.$$

Hence (3.16) is satisfied.

Now suppose that f fails to be coercive. That means, for some sequence $(\boldsymbol{x}_n)_n$ in $\mathbb{R}^d \setminus \{\mathbf{0}\}$ and some real constant C,

$$\lim_{n\to\infty} \|\boldsymbol{x}_n\| \;=\; \infty \quad \text{but} \quad f(\boldsymbol{x}_n) \;\le\; C \;\text{ for all } n.$$

We may replace $(\boldsymbol{x}_n)_n$ with a subsequence, if necessary, such that in addition

$$\boldsymbol{u}_n := \|\boldsymbol{x}_n\|^{-1}\boldsymbol{x}_n \;\to\; \boldsymbol{u} \in \partial \boldsymbol{B} \quad \text{as } n \to \infty.$$

But then for any fixed $t > 0$, convexity and continuity of f imply that

$$\begin{aligned} Df(t\boldsymbol{u}, \boldsymbol{u}) \;&\le\; f((t+1)\boldsymbol{u}) - f(t\boldsymbol{u}) \\ &=\; \lim_{n\to\infty} \big(f((t+1)\boldsymbol{u}_n) - f(t\boldsymbol{u}_n)\big) \\ &\le\; \lim_{n\to\infty} \frac{f(\|\boldsymbol{x}_n\|\boldsymbol{u}_n) - f(t\boldsymbol{u}_n)}{\|\boldsymbol{x}_n\| - t} \\ &=\; \lim_{n\to\infty} \frac{f(\boldsymbol{x}_n) - f(t\boldsymbol{u}_n)}{\|\boldsymbol{x}_n\| - t} \\ &\le\; \lim_{n\to\infty} \frac{C - \min_{\boldsymbol{y}\in t\boldsymbol{B}} f(\boldsymbol{y})}{\|\boldsymbol{x}_n\| - t} \\ &=\; 0. \end{aligned}$$

Thus (3.16) is violated as well. □

Example 3.64 (Logistic regression). In logistic regression one considers observations $(\boldsymbol{z}_1, y_1)$, …, $(\boldsymbol{z}_n, y_n)$ in $\mathbb{R}^d \times \{0, 1\}$ and wants to minimize the negative log-likelihood function $f : \mathbb{R}^p \to \mathbb{R}$ given by

$$f(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} \big(-y_i \boldsymbol{z}_i^\top \boldsymbol{x} + \log(1 + e^{\boldsymbol{z}_i^\top \boldsymbol{x}})\big).$$

An obvious question is whether there exists a unique minimizer $\boldsymbol{x}_*$ of f.

Fist of all, elementary calculations reveal that the gradient and Hessian matrix of f are given by

$$\nabla f(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} (\ell(\boldsymbol{z}_i^\top \boldsymbol{x}) - y_i)\, \boldsymbol{z}_i, \tag{3.17}$$

$$D^2 f(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} \ell'(\boldsymbol{z}_i^\top \boldsymbol{x})\, \boldsymbol{z}_i \boldsymbol{z}_i^\top, \tag{3.18}$$

where $\ell : \mathbb{R} \to (0, 1)$ is the logistic distribution function,

$$\ell(t) \;:=\; \frac{e^t}{1 + e^t} \;=\; (1 + e^{-t})^{-1},$$

with derivative $\ell' = \ell(1 - \ell) > 0$. Hence for any $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top D^2 f(\boldsymbol{x}) \boldsymbol{v} \;=\; \sum_{i=1}^{n} \ell'(\boldsymbol{z}_i^\top \boldsymbol{x})(\boldsymbol{z}_i^\top \boldsymbol{v})^2 \;\begin{cases} \ge\; 0, \\ >\; 0 \quad \text{if } \boldsymbol{v} \notin \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}^\perp. \end{cases}$$

This shows that f is convex, and it is even strictly convex if

$$\operatorname{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \;=\; \mathbb{R}^d.$$

As to coercivity, for any fixed unit vector $\boldsymbol{u} \in \mathbb{R}^d$ and $t \in \mathbb{R}$,

$$\begin{aligned} Df(t\boldsymbol{u}, \boldsymbol{u}) &= \nabla f(t\boldsymbol{u})^\top \boldsymbol{u} \\ &= \sum_{i=1}^n (\ell(t\boldsymbol{z}_i^\top \boldsymbol{u}) - y_i)\boldsymbol{z}_i^\top \boldsymbol{u} \\ &\to \sum_{i=1}^n \big((1-y_i)(\boldsymbol{z}_i^\top \boldsymbol{u})^+ + y_i(\boldsymbol{z}_i^\top \boldsymbol{u})^-\big) \quad \text{as } t \to \infty. \end{aligned}$$

The latter limit is nonnegative, and it is strictly positive unless

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{z}_i^\top \boldsymbol{u} > 0, \\ 0 & \text{if } \boldsymbol{z}_i^\top \boldsymbol{u} < 0. \end{cases}$$

Hence the function f is coercive if and only if for each unit vector $\boldsymbol{u} \in \mathbb{R}^d$ there exists at least one index i such that $y_i = 0 < \boldsymbol{z}_i^\top \boldsymbol{u}$ or $1 - y_i = 0 > \boldsymbol{z}_i^\top \boldsymbol{u}$.

In other words, the function f fails to be coercive if and only if there exists a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that

$$\{\boldsymbol{z}_i : y_i = 0\} \subset \{\boldsymbol{x} : \boldsymbol{u}^\top \boldsymbol{x} \le 0\} \quad \text{and} \quad \{\boldsymbol{z}_i : y_i = 1\} \subset \{\boldsymbol{x} : \boldsymbol{u}^\top \boldsymbol{x} \ge 0\}.$$

Exercise 3.65 (Logistic regression, I). Verify formulae (3.17) and (3.18).

Exercise 3.66 (Least absolute deviations regression). For given observations $(\boldsymbol{z}_1, y_1)$, $(\boldsymbol{z}_2, y_2)$, $\ldots$, $(\boldsymbol{z}_n, y_n)$ in $\mathbb{R}^d \times \mathbb{R}$ let $f : \mathbb{R}^d \to \mathbb{R}$ be given by

$$f(\boldsymbol{x}) := \sum_{i=1}^n |\boldsymbol{z}_i^\top \boldsymbol{x} - y_i|.$$

(a) Show that f is convex.

(b) Show that f is coercive if and only if

$$\operatorname{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \mathbb{R}^d.$$

3.5.4 Convex conjugates

Definition 3.67 (Fenchel–Lagrange transform). Let $f : \mathbb{R}^d \to (-\infty, \infty]$ with $\operatorname{dom}(f) \neq \emptyset$. The *Fenchel–Legendre transform of f* (also called *convex conjugate of f*) is defined as the function $f^* : \mathbb{R}^d \to (-\infty, \infty]$,

$$f^*(\boldsymbol{y}) := \sup_{\boldsymbol{x} \in \mathbb{R}^d} (\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x})) = \sup_{\boldsymbol{x} \in \operatorname{dom}(f)} (\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x})).$$

It follows from Lemma 3.20 (c) and Exercise 3.54 that the convex conjugate f^* of f is convex and lower semicontinuous. Moreover, the definition of f^* implies that

$$\boldsymbol{x}^\top \boldsymbol{y} \le f(\boldsymbol{x}) + f^*(\boldsymbol{y}) \quad \text{for arbitrary } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Now consider an affine function $A : \mathbb{R}^d \to \mathbb{R}$, $A(\boldsymbol{x}) := \boldsymbol{b}^\top \boldsymbol{x} - c$, with a certain vector $\boldsymbol{b} \in \mathbb{R}^d$ and a real number c. Then $A \le f$ if and only if

$$\boldsymbol{x}^\top \boldsymbol{b} - f(\boldsymbol{x}) \;\le\; c \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d,$$

and this is equivalent to

$$f^*(\boldsymbol{b}) \;\le\; c.$$

Consequently, if $\mathcal{A}$ denotes the set of all affine functions A on $\mathbb{R}^d$ with $A \le f$, then

$$\begin{aligned} \underline{f}(\boldsymbol{x}) := \sup_{A \in \mathcal{A}} A(\boldsymbol{x}) \;&=\; \sup_{(\boldsymbol{b},c) \in \mathbb{R}^d \times \mathbb{R} \,:\, c \ge f^*(\boldsymbol{b})} (\boldsymbol{b}^\top \boldsymbol{x} - c) \\ &=\; \sup_{\boldsymbol{b} \in \mathbb{R}^d \,:\, f^*(\boldsymbol{b}) < \infty} (\boldsymbol{b}^\top \boldsymbol{x} - f^*(\boldsymbol{b})) \\ &=\; (f^*)^*(\boldsymbol{x}). \end{aligned}$$

These considerations and Theorem 3.56 yield the following result:

Theorem 3.68. *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be convex and lower semicontinuous with $\mathrm{dom}(f) \neq \emptyset$. Then f^* has these properties, too, and*

$$(f^*)^* \;\equiv\; f.$$

Example 3.69. Let $f(\boldsymbol{x}) := \|\boldsymbol{x}\|^2/2$. Then

$$\boldsymbol{x}^\top \boldsymbol{y} - f(\boldsymbol{x}) \;=\; \|\boldsymbol{y}\|^2/2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2/2 \;\le\; f(\boldsymbol{y})$$

with equality if and only if $\boldsymbol{x} = \boldsymbol{y}$. Thus

$$f^* \;\equiv\; f.$$

Exercise 3.70 (Conjugates of quadratic functions). Let

$$f(\boldsymbol{x}) \;:=\; c + \boldsymbol{b}^\top \boldsymbol{x} + 2^{-1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$$

with $c \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ symmetric and positive definite. Show that

$$f^*(\boldsymbol{y}) \;=\; c_* + \boldsymbol{b}_*^\top \boldsymbol{y} + 2^{-1} \boldsymbol{y}^\top \boldsymbol{A}_* \boldsymbol{y}$$

with $c_* := 2^{-1} \boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b} - c$, $\boldsymbol{b}_* := -\boldsymbol{A}^{-1} \boldsymbol{b}$ and $\boldsymbol{A}_* := \boldsymbol{A}^{-1}$.

The following two exercises are about the standard p-norms on $\mathbb{R}^d$: For a vector $\boldsymbol{x} \in \mathbb{R}^d$ and $p \in [1, \infty)$ we write

$$\|\boldsymbol{x}\|_p \;:=\; \Big(\sum_{i=1}^d |x_i|^p \Big)^{1/p},$$

and

$$\|\boldsymbol{x}\|_\infty \;:=\; \max_{i=1,\ldots,d} |x_i|.$$

Exercise 3.71 (Dual norms, I, and Young's inequality). Let

$$f(\boldsymbol{x}) \;:=\; \|\boldsymbol{x}\|_p^p/p$$

with $1 < p < \infty$. Show that

$$f^*(\boldsymbol{y}) \;=\; \|\boldsymbol{y}\|_q^q/q$$

with $q := p/(p-1)$, so $p^{-1} + q^{-1} = 1$. Then show that

$$|\boldsymbol{x}^\top \boldsymbol{y}| \;\le\; f(\boldsymbol{x}) + f^*(\boldsymbol{y}) \quad \text{and} \quad |\boldsymbol{x}^\top \boldsymbol{y}| \;\le\; \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q$$

for arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Exercise 3.72 (Dual norms, II). Show that for $p \in [1, \infty]$, the convex conjugate of the function $f(\boldsymbol{x}) := \|\boldsymbol{x}\|_p$ is given by

$$f^*(\boldsymbol{y}) \;=\; \begin{cases} 0 & \text{if } \|\boldsymbol{y}\|_q \le 1, \\ \infty & \text{else,} \end{cases}$$

where

$$q \;:=\; \begin{cases} \infty & \text{if } p = 1, \\ p/(p-1) & \text{if } 1 < p < \infty, \\ 1 & \text{if } p = \infty. \end{cases}$$

Remark 3.73 (Support functions). Let $\boldsymbol{C}$ be a nonvoid convex and closed subset of $\mathbb{R}^d$, and define

$$f_{\boldsymbol{C}}(\boldsymbol{x}) \;:=\; \begin{cases} 0 & \text{if } \boldsymbol{x} \in \boldsymbol{C}, \\ \infty & \text{else.} \end{cases}$$

Then

$$f_{\boldsymbol{C}}^*(\boldsymbol{y}) \;=\; \sup_{\boldsymbol{x} \in \boldsymbol{C}} \boldsymbol{x}^\top \boldsymbol{y} \;=\; h_{\boldsymbol{C}}(\boldsymbol{y}).$$

In case of $\boldsymbol{C}$ being a closed and convex cone,

$$f_{\boldsymbol{C}}^* \;=\; f_{\boldsymbol{C}^*}$$

with the polar cone

$$\boldsymbol{C}^* \;=\; \big\{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{x}^\top \boldsymbol{y} \le 0 \text{ for all } \boldsymbol{x} \in \boldsymbol{C}\big\}$$

introduced in Remark 2.26.

Minimizing sums of convex functions. Sometimes one wants to minimize the sum of two convex functions $f, g : \mathbb{R}^d \to (-\infty, \infty]$, where we assume that $\mathrm{dom}(f) \cap \mathrm{dom}(g) \neq \emptyset$. It may happen that the minimization of $\boldsymbol{y} \mapsto f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})$ is easier. Under the slightly stronger constraint that $\mathrm{interior}(\mathrm{dom}(f)) \cap \mathrm{dom}(g) \neq \emptyset$, both minimization problems are strongly related:

Theorem 3.74 (Fenchel). *Let $f, g : \mathbb{R}^d \to (-\infty, \infty]$ be convex functions such that the intersection of $\mathrm{interior}(\mathrm{dom}(f))$ and $\mathrm{dom}(g)$ is nonvoid. Then*

$$\inf_{\boldsymbol{x} \in \mathbb{R}^d} \big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) \;=\; -\inf_{\boldsymbol{y} \in \mathbb{R}^d} \big(f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})\big).$$

The latter infimum is a minimum if it is real.

A special case of this theorem is the minimization of a convex function f over a closed convex cone $\boldsymbol{C}$. With its polar cone $\boldsymbol{C}^*$,

$$\inf_{\boldsymbol{x}\in\boldsymbol{C}} f(\boldsymbol{x}) \;=\; -\inf_{\boldsymbol{y}\in-\boldsymbol{C}^*} f^*(\boldsymbol{y}).$$

Proof of Theorem 3.74. Let $\alpha := \inf(f+g)$. Then

$$\begin{aligned}\alpha \;&=\; -\sup_{\boldsymbol{x}\in\mathbb{R}^d}\big(\boldsymbol{0}^\top\boldsymbol{x} - f(\boldsymbol{x}) - g(\boldsymbol{x})\big)\\ &=\; -\inf_{\boldsymbol{y}\in\mathbb{R}^d}\sup_{\boldsymbol{x}\in\mathbb{R}^d}\big(\boldsymbol{y}^\top\boldsymbol{x} - f(\boldsymbol{x}) + (-\boldsymbol{y})^\top\boldsymbol{x} - g(\boldsymbol{x})\big)\\ &\geq\; -\inf_{\boldsymbol{y}\in\mathbb{R}^d}(f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})).\end{aligned}$$

Hence it suffices to show that

$$\inf_{\boldsymbol{y}\in\mathbb{R}^d}(f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})) \;\leq\; -\alpha. \tag{3.19}$$

Since $\mathrm{interior}(\mathrm{dom}(f)) \cap \mathrm{dom}(g) \neq \emptyset$, we know that $\alpha < \infty$. In case of $\alpha = -\infty$, inequality (3.19) is trivial, so it suffices to consider the case $\alpha \in \mathbb{R}$. Since $f + g \geq \alpha$, the two sets

$$\boldsymbol{D}_1 \;:=\; \big\{(\boldsymbol{x}, s) : \boldsymbol{x} \in \mathbb{R}^d, s > f(\boldsymbol{x})\big\}$$

and

$$\boldsymbol{D}_2 \;:=\; \big\{(\boldsymbol{y}, t) : \boldsymbol{y} \in \mathbb{R}^d, t \leq \alpha - g(\boldsymbol{y})\big\}$$

are convex and disjoint. Thus there exists a nonzero vector $(\boldsymbol{b}, c) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\boldsymbol{b}^\top\boldsymbol{x} + cs \;\leq\; \boldsymbol{b}^\top\boldsymbol{y} + ct \quad \text{for arbitrary } (\boldsymbol{x}, s) \in \boldsymbol{D}_1, (\boldsymbol{y}, t) \in \boldsymbol{D}_2.$$

Taking $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{x}_o \in \mathrm{interior}(\mathrm{dom}(f)) \cap \mathrm{dom}(g)$ and $f(\boldsymbol{x}_o) < s \to \infty$, $\alpha - g(\boldsymbol{x}_o) \geq t \to -\infty$ shows that $c \leq 0$. If $c = 0$, we could take $\boldsymbol{x} = \boldsymbol{x}_o + \delta\boldsymbol{b}$ and $\boldsymbol{y} = \boldsymbol{x}_o$ with $\delta > 0$ sufficiently small such that $\boldsymbol{x} \in \mathrm{dom}(f)$. This would lead to the contradiction $\boldsymbol{b} = \boldsymbol{0}$. Hence $c < 0$, and we may assume without loss of generality that $c = -1$. Then we may conclude that

$$\boldsymbol{b}^\top\boldsymbol{x} - f(\boldsymbol{x}) \;\leq\; \boldsymbol{b}^\top\boldsymbol{y} - \alpha + g(\boldsymbol{y}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Rewriting the latter inequality as $\boldsymbol{b}^\top\boldsymbol{x} - f(\boldsymbol{x}) + (-\boldsymbol{b})^\top\boldsymbol{y} - g(\boldsymbol{y}) \leq -\alpha$, we realize that

$$f^*(\boldsymbol{b}) + g^*(-\boldsymbol{b}) \;\leq\; -\alpha \;\leq\; \inf_{\boldsymbol{y}\in\mathbb{R}^d}\big(f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})\big).$$

Consequently, inequality (3.19) is an equality, and $\boldsymbol{b}$ is a minimizer of the function $\boldsymbol{y} \mapsto f^*(\boldsymbol{y}) + g^*(-\boldsymbol{y})$. □

Chapter 4

Multivariate Optimization

In this chapter we modify methods from Chapter 1 for multivariate functions.

4.1 Newton's Method

Let $\boldsymbol{f} = (f_i)_{i=1}^d$ be a continuously differentiable mapping from an open set $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}^d$. We assume that the Jacobian matrix

$$D\boldsymbol{f}(\boldsymbol{x}) \;:=\; \Big(\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\Big)_{i,j=1}^d \;\in\; \mathbb{R}^{d\times d}$$

is nonsingular for all $\boldsymbol{x} \in \mathcal{X}$. Now we are looking for a zero $\boldsymbol{x}_*$ of $\boldsymbol{f}$, that means, a point $\boldsymbol{x}_* \in \boldsymbol{f}^{-1}(\boldsymbol{0})$.

Let $\boldsymbol{x}_o$ be a first candidate for a zero. We assume that we are already rather close to a zero $\boldsymbol{x}_*$ and approximate $\boldsymbol{f}$ by an affine function

$$\boldsymbol{x} \;\mapsto\; \boldsymbol{f}(\boldsymbol{x}_o) + D\boldsymbol{f}(\boldsymbol{x}_o)(\boldsymbol{x} - \boldsymbol{x}_o).$$

The latter function has a unique zero at

$$\boldsymbol{\psi}(\boldsymbol{x}_o) \;:=\; \boldsymbol{x}_o - D\boldsymbol{f}(\boldsymbol{x}_o)^{-1}\boldsymbol{f}(\boldsymbol{x}_o),$$

and we hope that $\boldsymbol{\psi}(\boldsymbol{x}_o)$ is even closer to $\boldsymbol{x}_*$ than $\boldsymbol{x}_o$.

Indeed, iterating this mapping leads to a sequence converging rapidly to a zero of $\boldsymbol{f}$, provided that we start sufficiently close to the latter. In what follows we work with the Euclidean norm

$$\|\boldsymbol{w}\| \;:=\; \sqrt{\boldsymbol{w}^\top \boldsymbol{w}},$$

of vectors $\boldsymbol{w} \in \mathbb{R}^\ell$, and for a matrix $\boldsymbol{A} \in \mathbb{R}^{k\times d}$ let

$$\|\boldsymbol{A}\| \;:=\; \sup_{\boldsymbol{v}\in\mathbb{R}^d\,:\,\|\boldsymbol{v}\|\le 1} \|\boldsymbol{A}\boldsymbol{v}\| \;=\; \sup_{\boldsymbol{v}\in\mathbb{R}^d\,:\,\|\boldsymbol{v}\|\le 1,\ \boldsymbol{w}\in\mathbb{R}^k\,:\,\|\boldsymbol{w}\|\le 1} \boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{v},$$

the so-called operator norm of $\boldsymbol{A}$. The definition implies that

$$\|\boldsymbol{A}\boldsymbol{v}\| \;\le\; \|\boldsymbol{A}\|\|\boldsymbol{v}\| \quad \text{for all } \boldsymbol{v} \in \mathbb{R}^d.$$

An important fact is that any identity matrix has operator norm one, and for arbitrary matrices $\boldsymbol{A}, \boldsymbol{B}$ whose product $\boldsymbol{AB}$ is well-defined,

$$\|\boldsymbol{AB}\| \;\leq\; \|\boldsymbol{A}\|\|\boldsymbol{B}\|.$$

These are the main advantages of the operator norm over the Frobenius norm. But these two matrix norms are equivalent:

Exercise 4.1. Show that the operator norm $\|\boldsymbol{A}\|$ and Frobenius norm $\|\boldsymbol{A}\|_F$ of a matrix $\boldsymbol{A} \in \mathbb{R}^{k\times d}$ satisfy the inequalities

$$\|\boldsymbol{A}\| \;\leq\; \|\boldsymbol{A}\|_F \;\leq\; \sqrt{\min(k,d)}\,\|\boldsymbol{A}\|.$$

Theorem 4.2 (Local convergence of Newton's method). *Let $\boldsymbol{x}_*$ be a zero of $\boldsymbol{f}$.*

(a) *Then*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_*,\boldsymbol{x}\neq\boldsymbol{x}_*} \frac{\|\boldsymbol{\psi}(\boldsymbol{x})-\boldsymbol{x}_*\|}{\|\boldsymbol{x}-\boldsymbol{x}_*\|} \;=\; 0.$$

(b) *Let $D\boldsymbol{f}$ be Lipschitz continuous in a neighborhood U of $\boldsymbol{x}_*$ with constant L, that means,*

$$\|D\boldsymbol{f}(\boldsymbol{x})-D\boldsymbol{f}(\boldsymbol{y})\| \;\leq\; L\|\boldsymbol{x}-\boldsymbol{y}\| \quad \text{for all } \boldsymbol{x},\boldsymbol{y}\in U.$$

Then

$$\limsup_{\boldsymbol{x}\to\boldsymbol{x}_*,\boldsymbol{x}\neq\boldsymbol{x}_*} \frac{\|\boldsymbol{\psi}(\boldsymbol{x})-\boldsymbol{x}_*\|}{\|\boldsymbol{x}-\boldsymbol{x}_*\|^2} \;\leq\; \frac{L\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|}{2}.$$

Consequently, with the multivariate version of Newton's method we achieve superlinear or even quadratic convergence. By means of Exercise 4.1 one can verify that a sufficient condition for Lipschitz-continuity of $D\boldsymbol{f}$ on a convex set $U\subset\mathcal{X}$ is that all second derivatives $\partial^2 f_i/(\partial x_k \partial x_\ell)$ are bounded on U. In the proof of Theorem 4.2 we use the following inequalities for matrices:

Lemma 4.3. *Let $\boldsymbol{A}$ be a nonsingular matrix in $\mathbb{R}^{d\times d}$, and let $\boldsymbol{B}=\boldsymbol{A}+\Delta$ for some $\Delta\in\mathbb{R}^{d\times d}$. In case of $\|\Delta\|<\|\boldsymbol{A}^{-1}\|^{-1}$, the matrix $\boldsymbol{B}$ is nonsingular, too, and*

$$\|\boldsymbol{B}^{-1}\| \;\leq\; \frac{\|\boldsymbol{A}^{-1}\|}{1-\|\boldsymbol{A}^{-1}\|\|\Delta\|},$$
$$\|\boldsymbol{B}^{-1}-\boldsymbol{A}^{-1}\| \leq\; \frac{\|\boldsymbol{A}^{-1}\|^2\|\Delta\|}{1-\|\boldsymbol{A}^{-1}\|\|\Delta\|}.$$

We also need an inequality for vector-valued integrals:

Exercise 4.4. Let $\boldsymbol{g}:[a,b]\to\mathbb{R}^d$ have integrable components $g_j:[a,b]\to\mathbb{R}$. Show that

$$\left\|\int_a^b \boldsymbol{g}(t)\,dt\right\| \;\leq\; \int_a^b \|\boldsymbol{g}(t)\|\,dt \;<\; \infty,$$

where the integral on the left hand side is defined component-wise.

Proof of Lemma 4.3. We rewrite $\boldsymbol{B}$ as

$$\boldsymbol{B} \;=\; \boldsymbol{A}+\Delta \;=\; \boldsymbol{A}(\boldsymbol{I}+\boldsymbol{A}^{-1}\Delta).$$

A matrix of type $\boldsymbol{I} + \boldsymbol{M}$ is nonsingular with inverse

$$(\boldsymbol{I} + \boldsymbol{M})^{-1} \;=\; \sum_{i=0}^{\infty} (-1)^i \boldsymbol{M}^i,$$

provided that $\|\boldsymbol{M}\| < 1$. In our context, we may apply this to $\boldsymbol{M} = \boldsymbol{A}^{-1}\Delta$ with norm $\|\boldsymbol{M}\| \le \|\boldsymbol{A}^{-1}\|\|\Delta\|$. Hence $\boldsymbol{B}$ is nonsingular with inverse

$$\boldsymbol{B}^{-1} \;=\; (\boldsymbol{I} + \boldsymbol{A}^{-1}\Delta)^{-1}\boldsymbol{A}^{-1} \;=\; \sum_{i=0}^{\infty} (-1)^i (\boldsymbol{A}^{-1}\Delta)^i \boldsymbol{A}^{-1} \;=\; \boldsymbol{A}^{-1} + \sum_{i=1}^{\infty} (-1)^i (\boldsymbol{A}^{-1}\Delta)^i \boldsymbol{A}^{-1},$$

provided that $\|\boldsymbol{A}^{-1}\|\|\Delta\| < 1$. The asserted inequalities follow from the fact that

$$\Big\| \sum_{i=k}^{\infty} (-1)^i (\boldsymbol{A}^{-1}\Delta)^i \boldsymbol{A}^{-1} \Big\| \;\le\; \sum_{i=k}^{\infty} \|\boldsymbol{A}^{-1}\|^{i+1} \|\Delta\|^i \;=\; \frac{\|\boldsymbol{A}^{-1}\|^{k+1}\|\Delta\|^k}{1 - \|\boldsymbol{A}^{-1}\|\|\Delta\|}$$

for $k = 0, 1$. □

Proof of Theorem 4.2. It follows from continuity of $\boldsymbol{x} \mapsto D\boldsymbol{f}(\boldsymbol{x})$ that

$$\rho(\delta) \;:=\; \sup_{\boldsymbol{x} \in \mathcal{X} \,:\, \|\boldsymbol{x} - \boldsymbol{x}_*\| \le \delta} \big\| D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_*) \big\| \;\to\; 0 \quad \text{as } \delta \downarrow 0.$$

For sufficiently small $\delta > 0$, the closed ball $B_\delta(\boldsymbol{x}_*)$ with center $\boldsymbol{x}_*$ and radius δ is contained in $\mathcal{X}$, and for any vector $\boldsymbol{x} \in B_\delta(\boldsymbol{x}_*)$,

$$\begin{aligned}
\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{x}_* \;&=\; \boldsymbol{x} - \boldsymbol{x}_* - D\boldsymbol{f}(\boldsymbol{x})^{-1}\boldsymbol{f}(\boldsymbol{x}) \\
&=\; \boldsymbol{x} - \boldsymbol{x}_* - D\boldsymbol{f}(\boldsymbol{x})^{-1} \int_0^1 D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*))(\boldsymbol{x} - \boldsymbol{x}_*)\, dt \\
&=\; D\boldsymbol{f}(\boldsymbol{x})^{-1} \int_0^1 \big(D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) \big)(\boldsymbol{x} - \boldsymbol{x}_*)\, dt.
\end{aligned}$$

This implies that

$$\begin{aligned}
\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{x}_*\| \;&\le\; \|D\boldsymbol{f}(\boldsymbol{x})^{-1}\| \Big\| \int_0^1 \big(D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) \big)(\boldsymbol{x} - \boldsymbol{x}_*)\, dt \Big\| \\
&\le\; \|D\boldsymbol{f}(\boldsymbol{x})^{-1}\| \int_0^1 \big\| \big(D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) \big)(\boldsymbol{x} - \boldsymbol{x}_*) \big\|\, dt \\
&\le\; \|D\boldsymbol{f}(\boldsymbol{x})^{-1}\| \int_0^1 \big\| D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) \big\|\, dt \, \|\boldsymbol{x} - \boldsymbol{x}_*\|, \qquad (4.1)
\end{aligned}$$

where we used the inequality from Exercise 4.4 in the second step. But it follows from Lemma 4.3 that

$$\|D\boldsymbol{f}(\boldsymbol{x})^{-1}\| \;\le\; \frac{\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|}{1 - \|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|\rho(\delta)},$$

provided that $\rho(\delta) < \|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|^{-1}$, and

$$\begin{aligned}
&\big\| D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) \big\| \\
&\quad\le\; \big\| D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_*) \big\| + \big\| D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*)) - D\boldsymbol{f}(\boldsymbol{x}_*) \big\| \\
&\quad\le\; 2\rho(\delta).
\end{aligned}$$

Hence for sufficiently small $\delta > 0$ and $\boldsymbol{x} \neq \boldsymbol{x}_*$,

$$\frac{\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{x}_*\|}{\|\boldsymbol{x} - \boldsymbol{x}_*\|} \;\leq\; \frac{2\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|\rho(\delta)}{1 - \|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|\rho(\delta)}.$$

Since the latter bound tends to 0 as $\delta \downarrow 0$, this proves part (a).

For part (b) we choose $\delta > 0$ small enough such that $B_\delta(\boldsymbol{x}_*) \subset \mathcal{X}$, $\rho(\delta) < \|D\boldsymbol{f}^{-1}(\boldsymbol{x}_*)\|^{-1}$ and $D\boldsymbol{f}$ is Lipschitz continuous on $B_\delta(\boldsymbol{x}_*)$ with constant L. Then for $\boldsymbol{x} \in B_\delta(\boldsymbol{x}_*)$ and $t \in [0,1]$,

$$\big\|D\boldsymbol{f}(\boldsymbol{x}) - D\boldsymbol{f}(\boldsymbol{x}_* + t(\boldsymbol{x} - \boldsymbol{x}_*))\big\| \;\leq\; L\|\boldsymbol{x} - \boldsymbol{x}_* - t(\boldsymbol{x} - \boldsymbol{x}_*)\| \;=\; L(1-t)\|\boldsymbol{x} - \boldsymbol{x}_*\|,$$

so the right hand side of (4.1) is not greater than

$$\|D\boldsymbol{f}(\boldsymbol{x})^{-1}\|L\int_0^1 (1-t)\,dt\;\|\boldsymbol{x} - \boldsymbol{x}_*\|^2 \;=\; L\|D\boldsymbol{f}(\boldsymbol{x})^{-1}\|\|\boldsymbol{x} - \boldsymbol{x}_*\|^2/2.$$

Hence for $\boldsymbol{x} \neq \boldsymbol{x}_*$,

$$\frac{\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{x}_*\|}{\|\boldsymbol{x} - \boldsymbol{x}_*\|^2} \;\leq\; \frac{L\|D\boldsymbol{f}(\boldsymbol{x})^{-1}\|}{2} \;\leq\; \frac{L\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|}{2\big(1 - \|D\boldsymbol{f}(\boldsymbol{x})^{-1}\|\rho(\delta)\big)},$$

and the latter bound converges to $L\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|/2$ as $\delta \downarrow 0$. □

Remark 4.5 (Range of attraction). Let $\boldsymbol{x}_*$ be a zero of $\boldsymbol{f}$, and let $\mathcal{X}(\boldsymbol{x}_*)$ be the set of all starting points $\boldsymbol{x}_0 \in \mathcal{X}$ such that the Newton sequence $(\boldsymbol{x}_n)_{n\geq 0} = \big(\boldsymbol{\psi}^n(\boldsymbol{x}_0)\big)_{n\geq 0}$ is well-defined and converges to $\boldsymbol{x}_*$. (Here and throughout the sequel, $\boldsymbol{\psi}^n$ denotes the n-fold iteration of $\boldsymbol{\psi}$.) The set $\mathcal{X}(\boldsymbol{x}_*)$ is always an open subset of $\mathcal{X}$ containing $\boldsymbol{x}_*$.

To prove this, note first that by Theorem 4.2, $\mathcal{X}(\boldsymbol{x}_*)$ contains an open neighborhood U_0 of $\boldsymbol{x}_*$. A point $\boldsymbol{x}_0 \in \mathcal{X} \setminus U_0$ belongs to $\mathcal{X}(\boldsymbol{x}_*)$ if and only if

$$\text{for some } n \in \mathbb{N}, \quad \boldsymbol{x}_j \;\in\; \mathcal{X} \quad \text{for } 1 \leq j < n \quad \text{and} \quad \boldsymbol{x}_n \;\in\; U_0.$$

Indeed, once the sequence has reached a point in U_0, it will certainly converge to $\boldsymbol{x}_*$. Defining

$$U_n \;:=\; \boldsymbol{\psi}^{-1}(U_{n-1}) \quad \text{for } n = 1, 2, 3, \ldots$$

we may reformulate this as $\boldsymbol{x}_n \in U_n$. Hence

$$\mathcal{X}(\boldsymbol{x}_*) \;=\; \bigcup_{n\geq 0} U_n.$$

But continuity of $\boldsymbol{\psi} : \mathcal{X} \to \mathbb{R}^d$ and openness of U_0 imply that all sets $U_0, U_1, U_2, \ldots$ are open. Consequently, $\mathcal{X}(\boldsymbol{x}_*)$ is an open set, too.

Example 4.6 (Julia[1] set). Let $f : \mathbb{C} \to \mathbb{C}$ be given by

$$f(z) \;:=\; z^3 - 1.$$

[1] Gaston Maurice Julia, 1893-1978: French mathematician working on rational functions and fractal sets

We may identify a complex number z with the vector $(\operatorname{Re} z, \operatorname{Im} z)^\top \in \mathbb{R}^2$. Then f is continuously differentiable with (complex) derivative

$$f'(z) \;=\; 3z^2$$

corresponding to the Jacobian matrix

$$\begin{bmatrix} \operatorname{Re} f'(z) & -\operatorname{Im} f'(z) \\ \operatorname{Im} f'(z) & \operatorname{Re} f'(z) \end{bmatrix}.$$

Hence we consider $\mathcal{X} = \mathbb{C} \setminus \{0\}$. The zeros of f are precisely the third unit roots

$$y_0 \;:=\; 1, \quad y_1 \;:=\; \exp(2\pi \boldsymbol{i}/3), \quad y_2 \;:=\; \exp(4\pi \boldsymbol{i}/3).$$

Newton's method is based on the mapping

$$\psi(z) \;:=\; z - \frac{f(z)}{f'(z)} \;=\; \frac{2z}{3} + \frac{1}{3z^2}.$$

According to Remark 4.5, for $j = 0, 1, 2$ the set

$$\mathcal{X}_j \;:=\; \big\{ z \in \mathcal{X} : \text{Newton's method with starting point } z \text{ converges to } y_j \big\}$$

is an open neighborhood of the point y_j. One can also verify directly that

$$\{ z \in \mathbb{C} : |z - y_j| \le 1/3 \} \;\subset\; \mathcal{X}_j.$$

The sets $\mathcal{X}_j$ have a very interesting shape, as shown in Figure 4.1. There $\mathcal{X}_0$, $\mathcal{X}_1$ and $\mathcal{X}_2$ are depicted in white, light gray and dark gray, respectively. A remarkable property of these sets is their *fractal* nature: On arbitrarily small scales one can find the same structures as on the original scale. All this is elaborated in the following four exercises.

Exercise 4.7 (Julia set, I). Let f, $\mathcal{X}$ and ψ be as in Example 4.6 with $f^{-1}(0) = \{y_0, y_1, y_2\}$, where $y_j = \exp(2\pi \boldsymbol{i} j/3)$.

(a) Show that $f(y_j z) = f(z)$ and $\psi(y_j z) = y_j \psi(z)$ for arbitrary $z \in \mathcal{X}$ and $j = 0, 1, 2$.

(b) Verify that for any $z \in \mathcal{X}$, the three equations $\psi(z) = z$, $\psi'(z) = 0$ and $f(z) = 0$ are equivalent.

(c) Determine explicit constants $C > 0$ and $D \in (0, 1)$ such that for $j = 0, 1, 2$,

$$|\psi(z) - y_j| \;\le\; D|z - y_j| \quad \text{if } |z - y_j| \le C.$$

For instance, there exists a solution (C, D) with $C = 1/3$.

(c') Determine a maximal constant $C' > 0$ such that for $j = 0, 1, 2$,

$$|\psi(z) - y_j| \;<\; |z - y_j| \quad \text{if } |z - y_j| < C'.$$

Exercise 4.8 (Julia set, II). Now we investigate the inverse image $\psi^{-1}(y)$ for various $y \in \mathbb{C}$.

(a) Determine the set $\psi^{-1}(0) \subset \mathcal{X}$.

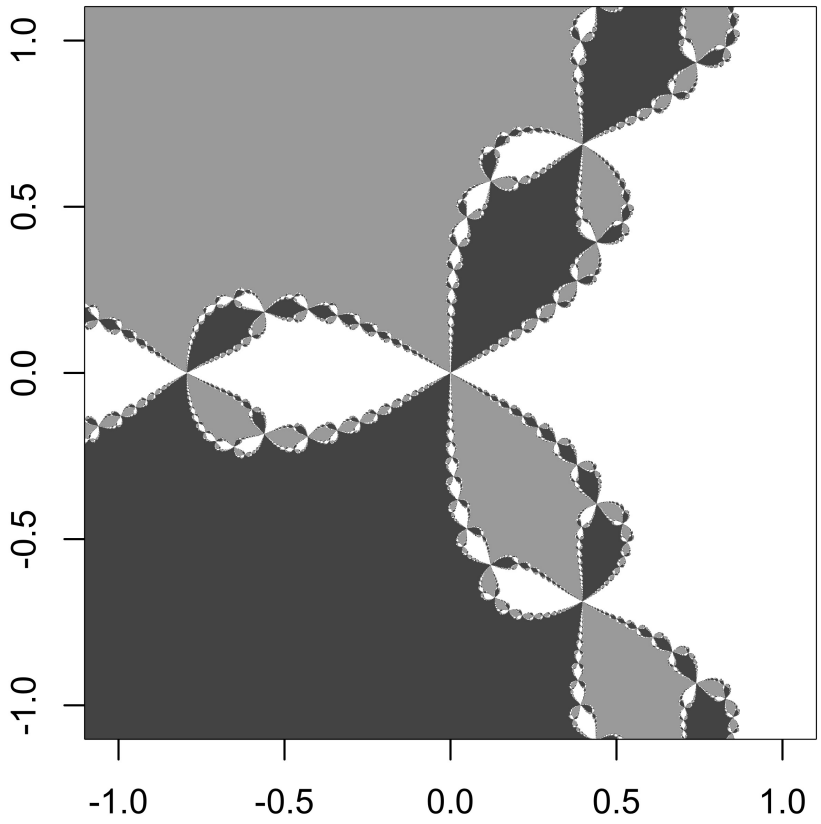

Figure 4.1: Ranges of attraction for Example 4.6.

(b) Show that $\psi^{-1}(y_j) = \{y_j, -y_j/2\}$ for $j = 0, 1, 2$.

(c) Show that for any $y \in \mathbb{C} \setminus \{y_0, y_1, y_2\}$ the set $\psi^{-1}(y)$ consists of three different points $z_1, z_2, z_3 \in \mathcal{X}$, where

$$\max_{i=1,2,3} |z_i| \;\geq\; |y|/2 \quad \text{and} \quad \min_{i=1,2,3} |z_i| \;\leq\; |y|^{-1/2}.$$

Exercise 4.9 (Julia set, III). From now on we consider the iterates ψ^n, $n \in \mathbb{N}_0$, of the mapping ψ. Show that there exist real numbers $0 = t_0 > t_1 > t_2 > t_3 > \cdots$ with the following properties:

(i) $\psi^n(t_n) = 0$ for $n \in \mathbb{N}$.

(ii) $\lim_{n\to\infty} t_n = -\infty$.

(iii) $\lim_{n\to\infty} \psi^n(t) = 1$ for any starting point $t \in \mathbb{R} \setminus \{0, t_1, t_2, t_3, \ldots\}$.

Exercise 4.10 (Julia set, IV). Now we combine the results from the previous three exercises.

(a) Deduce from the previous two exercises that 0 is contained in $\overline{\mathcal{X}(y_0)} \cap \overline{\mathcal{X}(y_1)} \cap \overline{\mathcal{X}(y_2)}$.

(b) Let $\psi^{-n}(0)$ be the inverse image $(\psi^n)^{-1}(0)$, and let $\mathcal{X}_o := \bigcup_{n\geq 1} \psi^{-n}(0)$. That means, $\mathcal{X}_o$ is the set of problematic starting points for Newton's method.

(b.i) Verify that for arbitrary integers $1 \leq m < n$, $\psi^{-m}(0) \cap \psi^{-n}(0) = \emptyset$, where $\psi^{-m}(0)$ consists of 3^m different points in $\mathcal{X}$.

(b.ii) Show that $\mathcal{X}_o = y_j \mathcal{X}_o$ for $j = 1, 2$.

(b.iii) Show that $\mathcal{X}_o$ contains points z with arbitrarily small and arbitrarily large modulus $|z|$.

(c) Show that for each integer $n \geq 1$ and any point $x \in \psi^{-n}(\mathbb{C})$,

$$(\psi^n)'(x) \;=\; \prod_{k=0}^{n-1} \psi'(\psi^k(x))$$

(with $\psi^0(x) = x$). Deduce from that formula that $(\psi^n)'(x) \neq 0$ for $x \in \psi^{-n}(0)$. Then apply the inverse mapping theorem to show that each point $x \in \mathcal{X}_o$ belongs to the intersection $\overline{\mathcal{X}(y_0)} \cap \overline{\mathcal{X}(y_1)} \cap \overline{\mathcal{X}(y_2)}$. More generally, the intersection of $\mathcal{X}(y_j)$ with a small neighborhood of x looks approximately like the intersection of $\mu(x)^{-1}\mathcal{X}(y_j)$ with a small neighborhood of 0.

Exercise 4.11 (Logistic regression, II). As in Example 3.64 let

$$f(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} \big(-y_i\, \boldsymbol{z}_i^\top \boldsymbol{x} + \log(1 + e^{\boldsymbol{z}_i^\top \boldsymbol{x}})\big)$$

with vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}^d$ such that $\operatorname{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \mathbb{R}^d$ and numbers $y_1, \ldots, y_n \in \{0, 1\}$. Write a program for the minimization of f via Newton's method.

Variations. Sometimes one replaces the exact derivative $D\boldsymbol{f}(\boldsymbol{x})$ by an approximation $\boldsymbol{H}(\boldsymbol{x})$. Two examples for this are:

(A) One approximates the derivative of $\boldsymbol{f}$ by difference ratios. That means, for a small number $\delta = \delta(\boldsymbol{x}) > 0$, the Jacobian matrix $D\boldsymbol{f}(\boldsymbol{x})$ is replaced with

$$\boldsymbol{H}(\boldsymbol{x}) \;:=\; \Big(\frac{f_i(\boldsymbol{x} + \delta \boldsymbol{e}_j) - f_i(\boldsymbol{x} - \delta \boldsymbol{e}_j)}{2\delta}\Big)_{i,j=1}^{d}.$$

(B) Quite often, the function $\boldsymbol{f}$ considered here is the gradient of a real-valued and twice continuously differentiable function $\tilde{f}$ to be minimized. In this case, $D\boldsymbol{f}(\boldsymbol{x})$ is the Hessian matrix of $\tilde{f}$ at the point $\boldsymbol{x}$ and symmetric. Often the function $\tilde{f}$ is even strictly convex with $D\boldsymbol{f}(\boldsymbol{x})$ positive definite for any $\boldsymbol{x}$. But the condition number of $D\boldsymbol{f}(\boldsymbol{x})$, the ratio of its largest to its smallest eigenvalue, can be very large, which causes numerical problems when computing $\boldsymbol{\psi}(\boldsymbol{x})$. To avoid this, people often replace $D\boldsymbol{f}(\boldsymbol{x})$ with

$$\boldsymbol{H}(\boldsymbol{x}) \;:=\; D\boldsymbol{f}(\boldsymbol{x}) + \epsilon(\boldsymbol{x})\boldsymbol{I}$$

with some $\epsilon(\boldsymbol{x}) > 0$.

Running Newton's method with a surrogate $\boldsymbol{H}(\boldsymbol{x})$ for $D\boldsymbol{f}(\boldsymbol{x})$ is called a *quasi-Newton method*. An obvious question is how well it performs. Let us assume that $\boldsymbol{H}(\boldsymbol{x})$ is always nonsingular and continuous in $\boldsymbol{x} \in \mathcal{X}$. Now we consider the algorithmic mapping

$$\tilde{\boldsymbol{\psi}}(\boldsymbol{x}) \;:=\; \boldsymbol{x} - \boldsymbol{H}(\boldsymbol{x})^{-1}\boldsymbol{f}(\boldsymbol{x}).$$

The proof of the following lemma is left to the reader as an exercise:

Lemma 4.12. *Let $\boldsymbol{x}_* \in \mathcal{X}$ be a zero of $\boldsymbol{f}$. Then the algorithmic mapping $\tilde{\boldsymbol{\psi}}$ defined above has the following property:*

$$\tilde{\boldsymbol{\psi}}(\boldsymbol{x}) - \boldsymbol{x}_* \;=\; \boldsymbol{A}(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{x}_*)$$

with a matrix $\boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ such that

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_*} \boldsymbol{A}(\boldsymbol{x}) \;=\; \boldsymbol{I} - \boldsymbol{H}(\boldsymbol{x}_*)^{-1} D\boldsymbol{f}(\boldsymbol{x}_*).$$

This lemma shows that superlinear convergence gets lost in general. Instead,

$$\limsup_{\boldsymbol{x}\to\boldsymbol{x}_*} \frac{\|\tilde{\boldsymbol{\psi}}(\boldsymbol{x}) - \boldsymbol{x}_*\|}{\|\boldsymbol{x} - \boldsymbol{x}_*\|} \;=\; \|\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{x}_*)^{-1} D\boldsymbol{f}(\boldsymbol{x}_*)\|.$$

If the right hand side is strictly smaller than 1, we still have linear convergence. One can deduce from Lemma 4.3 that

$$\begin{aligned}\|\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{x}_*)^{-1} D\boldsymbol{f}(\boldsymbol{x}_*)\| \;&\le\; \|\boldsymbol{H}(\boldsymbol{x}_*)^{-1}\|\|\boldsymbol{H}(\boldsymbol{x}_*) - D\boldsymbol{f}(\boldsymbol{x}_*)\| \\ &\le\; \frac{\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|\|\boldsymbol{H}(\boldsymbol{x}_*) - D\boldsymbol{f}(\boldsymbol{x}_*)\|}{1 - \|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|\|\boldsymbol{H}(\boldsymbol{x}_*) - D\boldsymbol{f}(\boldsymbol{x}_*)\|},\end{aligned}$$

and the right hand side is strictly smaller than 1 if

$$\|\boldsymbol{H}(\boldsymbol{x}_*) - D\boldsymbol{f}(\boldsymbol{x}_*)\| \;<\; \frac{1}{2\|D\boldsymbol{f}(\boldsymbol{x}_*)^{-1}\|}.$$

For the special setting (B) with a symmetric and positive definitie matrix $D\boldsymbol{f}(\boldsymbol{x}_*)$ and $\boldsymbol{H}(\boldsymbol{x}_*) = D\boldsymbol{f}(\boldsymbol{x}_*) + \epsilon(\boldsymbol{x}_*)\boldsymbol{I}$ one can deduce from the spectral representation of $D\boldsymbol{f}(\boldsymbol{x}_*)$ that

$$\|\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{x}_*)^{-1} D\boldsymbol{f}(\boldsymbol{x}_*)\| \;=\; \frac{\epsilon(\boldsymbol{x}_*)}{\epsilon(\boldsymbol{x}_*) + \lambda_{\min}(D\boldsymbol{f}(\boldsymbol{x}_*))} \tag{4.2}$$

with $\lambda_{\min}(\boldsymbol{A})$ denoting the smallest real eigenvalue of a matrix $\boldsymbol{A}$. Thus we may guarantee linear convergence here.

Exercise 4.13. Prove Lemma 4.12.

Exercise 4.14. Verify equation (4.2).

4.2 Minimization Problems

Now let $(\mathcal{X}, d)$ be a metric space and $f : \mathcal{X} \to \mathbb{R}$ be a given function to be minimized. To this end we consider algorithmic mappings $\psi : \mathcal{X} \to \mathcal{X}$ we intend to iterate. With $\psi^0(x) := x$ and $\psi^n(x) := \psi(\psi^{n-1}(x))$ for $n \in \mathbb{N}$, we want to find conditions on f and ψ such that

$$\mathcal{X}_* \;:=\; \arg\min_{\mathcal{X}} f$$

is compact (and thus nonvoid) and any starting point $x_0 \in \mathcal{X}$ will lead to a sequence $(x_n)_{n\ge0} := (\psi^n(x_0))_{n\ge0}$ converging to the set $\mathcal{X}_*$. That means,

$$d(x_n, \mathcal{X}_*) \;:=\; \min_{x_*\in\mathcal{X}_*} d(x_n, x_*) \;\to\; 0 \quad \text{as } n \to \infty.$$

4.2.1 A general criterion for convergence

Suppose that f is lower semicontinuous and satisfies the following condition:

(4.3) $\{x \in \mathcal{X} : f(x) \le f(x_o)\}$ is compact for arbitrary $x_o \in \mathcal{X}$.

For instance, this condition is satisfied if $f : \mathbb{R}^d \to \mathbb{R}$ is convex and coercive. Condition (4.3) implies that $\mathcal{X}_* = \arg\min_{\mathcal{X}}(f)$ is a compact (and thus nonvoid) subset of $\mathcal{X}$, see Exercise 3.55.

Now we consider a mapping $\psi : \mathcal{X} \to \mathcal{X}$ with the following two properties:

$$f(\psi(x)) \;=\; f(x) \quad \text{for } x \in \mathcal{X}_*, \tag{4.4}$$

$$f(\psi(x)) \;<\; f(x) \quad \text{for } x \notin \mathcal{X}_*. \tag{4.5}$$

These conditions imply that for any starting point $x_0 \in \mathcal{X}$, the sequence $\big(f(\psi^n(x_0))\big)_{n\ge 0}$ is non-increasing. Unfortunately, however, (4.5) is too weak to guarantee convergence of $(\psi^n(x_0))_{n\ge 0}$ to $\mathcal{X}_*$.

Example 4.15. Let $\mathcal{X} = [0,\infty)$ and $f(x) = x$, so $\mathcal{X}_* = \{0\}$. For any $c \in (0,1)$, the algorithmic mapping

$$\psi(x) \;:=\; \begin{cases} 0 & \text{if } x \le 1 \\ 1 + c(x-1) & \text{if } x > 1 \end{cases}$$

satisfies conditions (4.4) and (4.5). But for any starting point $x_0 > 1$, $\psi^n(x_0) = 1 + c^n(x_0 - 1) \to 1$ as $n \to \infty$.

Here is a stronger version of condition (4.5) which is sufficient for our purposes:

$$\limsup_{x \to y} f(\psi(x)) \;<\; f(y) \quad \text{for all } y \in \mathcal{X} \setminus \mathcal{X}_*. \tag{4.6}$$

For instance, if f and ψ are continuous, and if ψ satisfies (4.5), then condition (4.6) is satisfied as well.

Theorem 4.16. *Suppose that ψ satisfies conditions* (4.4) *and* (4.6). *Then for any starting point $x_0 \in \mathcal{X}$, the sequence $(x_n)_{n\ge 0} = (\psi^n(x_0))_{n\ge 0}$ converges to $\mathcal{X}_*$.*

Proof of Theorem 4.16. As mentioned already, properties (4.4) and (4.5) of ψ imply that the sequence of numbers $f(x_n)$ is non-increasing. In particular, all points x_n are contained in the compact set $\mathcal{X}_o := \big\{x \in \mathcal{X} : f(x) \le f(x_0)\big\}$. Suppose that $(x_n)_n$ does not converge to $\mathcal{X}_*$. That means, for some $\delta > 0$, $d(x_n, \mathcal{X}_*) \ge \delta$ for infinitely many indices n. Since $\big\{x \in \mathcal{X}_o : d(x, \mathcal{X}_*) \ge \delta\big\}$ is compact, too, there exist indices $n(1) < n(2) < n(3) < \cdots$ such that $\lim_{k\to\infty} x_{n(k)} = y \in \mathcal{X}_o$ with $d(y, \mathcal{X}_*) \ge \delta$. By lower semicontinuity of f and monotonicity of $(f(x_n))_{n=0}^{\infty}$,

$$\begin{aligned} f(y) \;&\le\; \lim_{k\to\infty} f(x_{n(k)}) \\ &=\; \lim_{k\to\infty} f(x_{n(k)+1}) \\ &=\; \lim_{k\to\infty} f(\psi(x_{n(k)})) \\ &\le\; \limsup_{\boldsymbol{x\to y}} f(\psi(x)). \end{aligned}$$

Because of condition (4.6), this would imply that $y \in \mathcal{X}_*$, a contradiction. □

Example 4.17 (Coordinatewise descent and LASSO). Let $\mathcal{X} = \mathbb{R}^d$, and consider the function $f_\lambda(\boldsymbol{x}) := \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{x}\|^2/2 + \lambda \sum_{j=1}^d |x_j|$ with a given vector $\boldsymbol{y} \in \mathbb{R}^n$, a matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ whose columns have Euclidean norm 1 and some $\lambda > 0$. This function is convex and coercive. As shown in Exercise 3.39, for any point $\boldsymbol{x} \in \mathbb{R}^d$ and index $k \in \{1, \dots, d\}$ there exists a unique vector

$$\boldsymbol{\psi}^{(k)}(\boldsymbol{x}) \;:=\; \arg\min\bigl\{f_\lambda(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \in \mathbb{R}^d, \tilde{x}_j = x_j \text{ whenever } j \neq k\bigr\},$$

and this defines a continuous mapping $\boldsymbol{\psi}^{(k)} : \mathbb{R}^d \to \mathbb{R}^d$. Moreover, $\boldsymbol{x}$ is a minimizer of f_λ if and only if $\boldsymbol{\psi}^{(k)}(\boldsymbol{x}) = \boldsymbol{x}$ for all $k \in \{1, \dots, d\}$. This implies that the algorithmic mapping

$$\boldsymbol{\psi} \;:=\; \boldsymbol{\psi}^{(d)} \circ \boldsymbol{\psi}^{(d-1)} \circ \cdots \circ \boldsymbol{\psi}^{(1)}$$

is continuous and satisfies conditions (4.4) and (4.5). Since f_λ is continuous, too, condition (4.6) is satisfied as well. Consequently, iterating the mapping $\boldsymbol{\psi}$, also called iterative coordinatewise descent, yields a sequence converging to $\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} f_\lambda(\boldsymbol{x})$.

Example 4.18 (Spatial median and Weiszfeld's algorithm). As in Exercise 3.37, let $f : \mathbb{R}^d \to \mathbb{R}$ be given by

$$f(\boldsymbol{x}) \;:=\; \sum_{i=1}^n \|\boldsymbol{x} - \boldsymbol{x}_i\|$$

with given points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$. As shown earlier,

$$Df(\boldsymbol{x}, \boldsymbol{v}) \;=\; n(\boldsymbol{x})\|\boldsymbol{v}\| + \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{v}$$

with

$$n(\boldsymbol{x}) \;:=\; \#\{i : \boldsymbol{x}_i = \boldsymbol{x}\} \quad \text{and} \quad \boldsymbol{g}(\boldsymbol{x}) \;:=\; \sum_{i:\boldsymbol{x}_i\neq\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{x}_i\|^{-1}(\boldsymbol{x} - \boldsymbol{x}_i).$$

In particular,

$$\min_{\boldsymbol{v}:\|\boldsymbol{v}\|\le 1} Df(\boldsymbol{x}, \boldsymbol{v}) \;=\; n(\boldsymbol{x}) - \|\boldsymbol{g}(\boldsymbol{x})\|,$$

so

$$\boldsymbol{x} \;\in\; \mathcal{X}_* \quad \text{if and only if} \quad \|\boldsymbol{g}(\boldsymbol{x})\| \;\le\; n(\boldsymbol{x}).$$

An explicit algorithm for the computation of a point in $\mathcal{X}_*$ has been proposed by Weiszfeld (1937)[2] For a given candidate $\boldsymbol{x} \in \mathbb{R}^d \setminus \mathcal{X}_*$ we approximate f by a function $f_{\boldsymbol{x}}$ which is easier to minimize explicitly. Precisely, since $\|\boldsymbol{a} + \boldsymbol{v}\| \le \|\boldsymbol{a}\| + (\|\boldsymbol{a} + \boldsymbol{v}\|^2 - \|\boldsymbol{a}\|^2)/(2\|\boldsymbol{a}\|)$ for arbitrary vectors $\boldsymbol{a} \neq \boldsymbol{0}$ and $\boldsymbol{v}$ in $\mathbb{R}^d$,

$$\begin{aligned}
f(\boldsymbol{x} + \boldsymbol{v}) \;&=\; \sum_{i\,:\,\boldsymbol{x}_i\neq\boldsymbol{x}} \|\boldsymbol{x} + \boldsymbol{v} - \boldsymbol{x}_i\| + n(\boldsymbol{x})\|\boldsymbol{v}\| \\
&\le\; \sum_{i\,:\,\boldsymbol{x}_i\neq\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{x}_i\| + \sum_{i\,:\,\boldsymbol{x}_i\neq\boldsymbol{x}} \frac{\|\boldsymbol{x} + \boldsymbol{v} - \boldsymbol{x}_i\|^2 - \|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\|\boldsymbol{x} - \boldsymbol{x}_i\|} + n(\boldsymbol{x})\|\boldsymbol{v}\| \\
&=\; f(\boldsymbol{x}) + \sum_{i\,:\,\boldsymbol{x}_i\neq\boldsymbol{x}} \frac{2(\boldsymbol{x} - \boldsymbol{x}_i)^\top \boldsymbol{v} + \|\boldsymbol{v}\|^2}{2\|\boldsymbol{x} - \boldsymbol{x}_i\|} + n(\boldsymbol{x})\|\boldsymbol{v}\| \\
&=\; f(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{v} + n(\boldsymbol{x})\|\boldsymbol{v}\| + w(\boldsymbol{x})\,\frac{\|\boldsymbol{v}\|^2}{2} \\
&=:\; f_{\boldsymbol{x}}(\boldsymbol{x} + \boldsymbol{v})
\end{aligned}$$

[2]Endre Weiszfeld, alias Andrew Vázsonyi (1916-2003): Hungarian mathematician and operations researcher, who emigrated in the 1940's via Paris to the USA.

with $n(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ as before, and

$$w(\boldsymbol{x}) \;:=\; \sum_{i\,:\,\boldsymbol{x}_i \neq \boldsymbol{x}} \frac{1}{\|\boldsymbol{x}-\boldsymbol{x}_i\|}.$$

Note that

$$f_{\boldsymbol{x}} \;\geq\; f, \quad f_{\boldsymbol{x}}(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \quad \text{and} \quad Df_{\boldsymbol{x}}(\boldsymbol{x},\cdot) \;\equiv\; Df(\boldsymbol{x},\cdot).$$

Writing $\boldsymbol{v} = t\boldsymbol{a}$ with a unit vector $\boldsymbol{a} \in \mathbb{R}^d$ and a scalar $t > 0$,

$$\begin{aligned} f_{\boldsymbol{x}}(\boldsymbol{x}+t\boldsymbol{a}) \;&=\; f(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{a}t + n(\boldsymbol{x})t + w(\boldsymbol{x})\,\frac{t^2}{2} \\ &\geq\; f(\boldsymbol{x}) - \big(\|\boldsymbol{g}(\boldsymbol{x})\| - n(\boldsymbol{x})\big)t + w(\boldsymbol{x})\,\frac{t^2}{2} \end{aligned}$$

with equality if and only if $\boldsymbol{a} \;=\; -\|\boldsymbol{g}(\boldsymbol{x})\|^{-1}\boldsymbol{g}(\boldsymbol{x})$. Then the unique minimizer with respect to $t > 0$ is given by

$$t \;=\; -\,\frac{\|\boldsymbol{g}(\boldsymbol{x})\| - n(\boldsymbol{x})}{w(\boldsymbol{x})},$$

and the improvement in the value of $f_{\boldsymbol{x}}$ equals

$$f_{\boldsymbol{x}}(\boldsymbol{x}) - f_{\boldsymbol{x}}(\boldsymbol{x}+t\boldsymbol{a}) \;=\; \frac{\big(\|\boldsymbol{g}(\boldsymbol{x})\| - n(\boldsymbol{x})\big)^2}{2w(\boldsymbol{x})}.$$

Since $f_{\boldsymbol{x}} \;\geq\; f$ and $f_{\boldsymbol{x}}(\boldsymbol{x}) \;=\; f(\boldsymbol{x})$, the latter quantity is automatically a lower bound for the improvement in the value of f.

For general $\boldsymbol{x} \in \mathbb{R}^d$ we define

$$\boldsymbol{\psi}(\boldsymbol{x}) \;:=\; \begin{cases} \boldsymbol{x} & \text{if } \|\boldsymbol{g}(\boldsymbol{x})\| \leq n(\boldsymbol{x}), \\ \boldsymbol{x} - \dfrac{1 - n(\boldsymbol{x})/\|\boldsymbol{g}(\boldsymbol{x})\|}{w(\boldsymbol{x})}\,\boldsymbol{g}(\boldsymbol{x}) & \text{if } \|\boldsymbol{g}(\boldsymbol{x})\| > n(\boldsymbol{x}). \end{cases}$$

Then for $\boldsymbol{x} \in \mathbb{R}^d \setminus \mathcal{X}_*$,

$$f(\boldsymbol{x}) - f(\boldsymbol{\psi}(\boldsymbol{x})) \;\geq\; \frac{\big(\|\boldsymbol{g}(\boldsymbol{x})\| - n(\boldsymbol{x})\big)^2}{2w(\boldsymbol{x})}.$$

Consequently, Weiszfeld's algorithmic mapping $\boldsymbol{\psi}$ satisfies (4.4) and (4.5).

Unfortunately, (4.6) is violated in general. But one can easily verify that $\boldsymbol{\psi}$ is continuous on $\mathbb{R}^d \setminus \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Indeed, for $\boldsymbol{x} \notin \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$,

$$\begin{aligned} n(\boldsymbol{x}) \;&=\; 0, \\ \boldsymbol{g}(\boldsymbol{x}) \;&=\; \sum_{i=1}^n \|\boldsymbol{x}-\boldsymbol{x}_i\|^{-1}(\boldsymbol{x}-\boldsymbol{x}_i), \\ w(\boldsymbol{x}) \;&=\; \sum_{i=1}^n \|\boldsymbol{x}-\boldsymbol{x}_i\|^{-1}, \\ \boldsymbol{\psi}(\boldsymbol{x}) \;&=\; w(\boldsymbol{x})^{-1} \sum_{i=1}^n \|\boldsymbol{x}-\boldsymbol{x}_i\|^{-1}\boldsymbol{x}_i, \\ f(\boldsymbol{x}) - f(\boldsymbol{\psi}(\boldsymbol{x})) \;&\geq\; \frac{\|\boldsymbol{g}(\boldsymbol{x})\|^2}{2w(\boldsymbol{x})}. \end{aligned}$$

Thus a possible algorithm with guaranteed convergence runs as follows:

Phase 1: We compute $f(\boldsymbol{x}_i)$ for $1 \le i \le n$ and determine a minimizer $\boldsymbol{x}^0$ of f on $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

Phase 2: If

$$\|\boldsymbol{g}(\boldsymbol{x}^0)\| \;\le\; n(\boldsymbol{x}^0),$$

then $\boldsymbol{x}^0$ is already a minimizer of f. Otherwise, $\boldsymbol{\psi}(\boldsymbol{x}^0)$ defines a first point in the open convex set $\mathcal{X}_0 := \big\{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) < f(\boldsymbol{x}^0)\big\}$, which is contained in $\mathbb{R}^d \setminus \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Since $\boldsymbol{\psi}$ is continuous on the latter set and satisfies (4.4) and (4.5), the sequence $(\boldsymbol{x}^k)_{k\ge0} := \big(\psi^k(\boldsymbol{x}^0)\big)_{k\ge0}$ converges automatically to $\mathcal{X}_*$.

In the next exercise it will be shown that $\mathcal{X}_*$ consists of one point $\boldsymbol{x}_*$, unless all points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are lying on a straight line in $\mathbb{R}^d$.

Exercise 4.19 (Spatial median, II).

(a) For $\boldsymbol{x} \in \mathbb{R}^d$ let $g(\boldsymbol{x}) := \|\boldsymbol{x}\|$. Show that for $\boldsymbol{x} \ne \boldsymbol{0}$,

$$\nabla g(\boldsymbol{x}) \;=\; \boldsymbol{u} \quad \text{and} \quad D^2 g(\boldsymbol{x}) \;=\; \|\boldsymbol{x}\|^{-1}(\boldsymbol{I}_d - \boldsymbol{u}\boldsymbol{u}^\top)$$

with $\boldsymbol{u} := \|\boldsymbol{x}\|^{-1}\boldsymbol{x}$.

(b) As in Exercise 3.37 let $f(\boldsymbol{x}) := \sum_{i=1}^n \|\boldsymbol{x} - \boldsymbol{x}_i\|$ with given points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$. Determine the gradient $\nabla f(\boldsymbol{x})$ and the Hessian matrix $D^2 f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Show that $D^2 f(\boldsymbol{x})$ is positive definite, unless all points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are lying on a straight line in $\mathbb{R}^d$ running through $\boldsymbol{x}$.

(c) Show that the function f is strictly convex on $\mathbb{R}^d$, unless all points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are lying on a straight line in $\mathbb{R}^d$ (which is always the case if $d = 1$).

4.2.2 Gradient, Newton and quasi-Newton procedures

Now we consider the special case of a continuously differentiable function f on an open subset $\mathcal{X}$ of $\mathbb{R}^d$. We assume condition (4.3) and that

$$\nabla f(\boldsymbol{x}) \;\ne\; 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*. \tag{4.7}$$

Conditions (4.3) and (4.7) are satisfied if, for instance, $\mathcal{X}$ is convex and f is convex such that

$$f(\boldsymbol{x}) \;\to\; \infty \quad \text{as } \boldsymbol{x} \to \partial\mathcal{X} \text{ or } \|\boldsymbol{x}\| \to \infty.$$

A model algorithm. In the present context we discuss algorithmic mappings $\boldsymbol{\psi} : \mathcal{X} \to \mathcal{X}$ of the following type:

$$\boldsymbol{\psi}(\boldsymbol{x}) \;=\; \boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x})$$

with a "candidate step function" $\Delta : \mathcal{X} \to \mathbb{R}^d$ and a "step size function" $\lambda : \mathcal{X} \to [0, 1]$. The mapping $\boldsymbol{\psi}$ satisfies the conditions of Theorem 4.16, provided that the following two conditions are fulfilled:

(i) Δ is continuous, and

$$\begin{aligned} \Delta(\boldsymbol{x}) \;&=\; \mathbf{0} \quad \text{for } \boldsymbol{x} \in \mathcal{X}_*, \\ \nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x}) \;&<\; 0 \quad \text{for } \boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*. \end{aligned}$$

(ii) For $x \in \mathcal{X} \setminus \mathcal{X}_*$ define

$$\begin{aligned} C(\boldsymbol{x}) \;&:=\; \frac{f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))}{\max_{t\in[0,1]}\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))\big)} \qquad (f := \infty \text{ on } \mathbb{R}^d \setminus \mathcal{X}), \\ \tilde{C}(\boldsymbol{x}) \;&:=\; \frac{f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))}{-\nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x})}. \end{aligned}$$

Then for any $\boldsymbol{y} \in \mathcal{X} \setminus \mathcal{X}_*$,

$$\liminf_{\boldsymbol{x}\to\boldsymbol{y}} C(\boldsymbol{x}) \;>\; 0 \quad \text{or} \quad \liminf_{\boldsymbol{x}\to\boldsymbol{y}} \tilde{C}(\boldsymbol{x}) \;>\; 0$$

Concerning the maximum in $C(\boldsymbol{x})$, note that $\{t \in \mathbb{R} : \boldsymbol{x} + t\Delta(\boldsymbol{x}) \in \mathcal{X}\}$ is an open subset of $\mathbb{R}$ containing 0. Compactness of $\{f \le f(\boldsymbol{x})\}$ implies that the set $\big\{t \in [0,1] : f(\boldsymbol{x} + t\Delta(\boldsymbol{x})) \le f(\boldsymbol{x})\big\}$ is compact, and by condition (4.5), $\max_{t\in[0,1]}\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))\big) > 0$.

To verify (4.6), note that for $\boldsymbol{y} \in \mathcal{X} \setminus \mathcal{X}_*$,

$$\limsup_{\boldsymbol{x}\to\boldsymbol{y}} f(\boldsymbol{\psi}(\boldsymbol{x})) \;=\; f(\boldsymbol{y}) - \liminf_{\boldsymbol{x}\to\boldsymbol{y}}\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))\big),$$

and for a suitable fixed $t_o = t_o(\boldsymbol{y}) > 0$,

$$\begin{aligned} \liminf_{\boldsymbol{x}\to\boldsymbol{y}}\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))\big) \;&\ge\; \liminf_{\boldsymbol{x}\to\boldsymbol{y}} C(\boldsymbol{x})\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + t_o\Delta(\boldsymbol{x}))\big) \\ &=\; \liminf_{\boldsymbol{x}\to\boldsymbol{y}} C(\boldsymbol{x})\big(f(\boldsymbol{y}) - f(\boldsymbol{y} + t_o\Delta(\boldsymbol{y}))\big) \\ &>\; 0 \end{aligned}$$

or

$$\begin{aligned} \liminf_{\boldsymbol{x}\to\boldsymbol{y}}\big(f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))\big) \;&=\; \liminf_{\boldsymbol{x}\to\boldsymbol{y}} \tilde{C}(\boldsymbol{x})\big(-\nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x})\big) \\ &=\; \liminf_{\boldsymbol{x}\to\boldsymbol{y}} \tilde{C}(\boldsymbol{x})\big(-\nabla f(\boldsymbol{y})^\top \Delta(\boldsymbol{y})\big) \\ &>\; 0. \end{aligned}$$

4.2.3 Examples for the candidate step function Δ

Gradient descent. In the simplest case one chooses

$$\Delta(\boldsymbol{x}) \;:=\; -\nabla f(\boldsymbol{x}).$$

More generally one could take $\Delta(\boldsymbol{x}) := -\kappa(\boldsymbol{x})\nabla f(\boldsymbol{x})$ with a continuous mapping $\kappa : \mathcal{X} \to (0, \infty)$.

Newton proposal. Suppose that f is twice continuously differentiable on $\mathcal{X}$ with positive definite Hessian matrix $D^2 f(\boldsymbol{x})$. Then a promising choice of $\Delta(\boldsymbol{x})$ seems to be

$$\Delta(\boldsymbol{x}) \;:=\; -D^2 f(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}),$$

see also Section 4.1.

Quasi-Newton proposal. Both preceding variants may be generalized and interpreted as follows: For given $\boldsymbol{x} \in \mathcal{X}$ we approximate f by a quadratic function

$$\boldsymbol{y} \;\mapsto\; f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + 2^{-1} (\boldsymbol{y} - \boldsymbol{x})^\top \boldsymbol{A}(\boldsymbol{x}) (\boldsymbol{y} - \boldsymbol{x})$$

with a symmetric, positive definite matrix $\boldsymbol{A}(\boldsymbol{x})$ depending continuously on $\boldsymbol{x} \in \mathcal{X}$. The unique minimizer of this quadratic function equals $\boldsymbol{x} + \Delta(\boldsymbol{x})$ with

$$\Delta(\boldsymbol{x}) \;:=\; -\boldsymbol{A}(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}).$$

In case of gradient descent, $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{I}$ or $\boldsymbol{A}(\boldsymbol{x}) = \kappa(\boldsymbol{x})^{-1} \boldsymbol{I}$, respectively. In case of Newton's proposal, $\boldsymbol{A}(\boldsymbol{x}) = D^2 f(\boldsymbol{x})$. Another possible variant is

$$\boldsymbol{A}(\boldsymbol{x}) \;:=\; \operatorname{diag}\Big(\Big(\frac{\partial^2 f}{\partial x_i^2}(\boldsymbol{x}) \Big)_{i=1}^d \Big),$$

provided that the second derivatives on the right hand side exist, are strictly positive and continuous in $\boldsymbol{x} \in \mathcal{X}$.

Condition (i) of our model algorithm is always satisfied. Obviously, Δ is a continuous function on $\mathcal{X}$, and $\Delta(\boldsymbol{x}) = 0$ if and only if $\nabla f(\boldsymbol{x}) = 0$ which is equivalent to $\boldsymbol{x} \in \mathcal{X}_*$. Moreover,

$$\nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x}) \;=\; -\nabla f(\boldsymbol{x})^\top \boldsymbol{A}(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}) \;<\; 0$$

for $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$.

4.2.4 Examples for the step size function λ

An ideal choice for $\lambda(\boldsymbol{x})$ seems to be

$$\lambda(\boldsymbol{x}) \;\in\; \operatorname*{arg\,min}_{t \in [0,1]} \; f(\boldsymbol{x} + t \Delta(\boldsymbol{x})).$$

Obviously condition (ii) in Section 4.2.2 would be satisfied with $C(\cdot) \equiv 1$. But in practice one could compute these step sizes $\lambda(\boldsymbol{x})$ only approximately, for instance with one of the methods in Chapter 1. Hence one would run two nested iterative algorithms which could be too time-consuming.

With the "improvement function" $H_{\boldsymbol{x}} : \mathbb{R} \to [-\infty, \infty)$ given by

$$H_{\boldsymbol{x}}(t) \;:=\; f(\boldsymbol{x}) - f(\boldsymbol{x} + t \Delta(\boldsymbol{x}))$$

one could rewrite the previous goal as

$$\lambda(\boldsymbol{x}) \;\in\; \operatorname*{arg\,max}_{t\in[0,1]} H_{\boldsymbol{x}}(t).$$

Now we introduce two alternative methods for picking $\lambda(\boldsymbol{x})$. Note that

$$H_{\boldsymbol{x}}(0) \;=\; 0 \;<\; H'_{\boldsymbol{x}}(0) \;=\; -\nabla f(\boldsymbol{x})^{\top}\Delta(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x}\in\mathcal{X}\setminus\mathcal{X}_*.$$

With $H_{\boldsymbol{x}}$ one may rewrite $C(\boldsymbol{x})$ and $\tilde{C}(\boldsymbol{x})$ in condition (ii) in Section 4.2.2 as

$$C(\boldsymbol{x}) \;=\; \frac{H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))}{\sup_{[0,1]} H_{\boldsymbol{x}}} \quad\text{and}\quad \tilde{C}(\boldsymbol{x}) \;=\; \frac{H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))}{H'_{\boldsymbol{x}}(0)}.$$

The method of Goldstein–Armijo. We fix a constant $a\in(0,1)$ and define

$$\lambda(\boldsymbol{x}) \;:=\; 2^{-n(\boldsymbol{x})} \quad\text{with}\quad n(\boldsymbol{x}) \;:=\; \min\bigl\{n\in\mathbb{N}_0 : H_{\boldsymbol{x}}(2^{-n}) \geq a2^{-n}H'_{\boldsymbol{x}}(0)\bigr\}$$

for $\boldsymbol{x}\in\mathcal{X}\setminus\mathcal{X}_*$. Since

$$\lim_{n\to\infty} \frac{H_{\boldsymbol{x}}(2^{-n})}{2^{-n}H'_{\boldsymbol{x}}(0)} \;=\; 1 \;>\; a,$$

the number $n(\boldsymbol{x})$ is well-defined in $\mathbb{N}_0$. To verify Condition (ii) in Section 4.2.2, it suffices to show that for any $\boldsymbol{y}\in\mathcal{X}\setminus\mathcal{X}_*$,

$$\limsup_{\boldsymbol{x}\to\boldsymbol{y}} n(\boldsymbol{x}) \;<\; \infty,$$

because

$$\tilde{C}(\boldsymbol{x}) \;\geq\; 2^{-n(\boldsymbol{x})}a$$

by definition of $n(\boldsymbol{x})$. Indeed, it follows from $2^nH_{\boldsymbol{y}}(2^{-n}) \to H'_{\boldsymbol{y}}(0)$ that for a suitable $n_o = n_o(\boldsymbol{y})\in\mathbb{N}_0$,

$$H_{\boldsymbol{y}}(2^{-n_o}) > 2^{-n_o}aH'_{\boldsymbol{y}}(0).$$

Since $H_{\boldsymbol{x}}(2^{-n_o}) \to H_{\boldsymbol{y}}(2^{-n_o})$ and $H'_{\boldsymbol{x}}(0) \to H'_{\boldsymbol{y}}(0)$ as $\boldsymbol{x}\to\boldsymbol{y}$, this implies that $n(\boldsymbol{x}) \leq n_o$ if $\boldsymbol{x}$ is sufficiently close to $\boldsymbol{y}$.

The definition of $\lambda(\boldsymbol{x})$ is illustrated in Figure 4.2. One sees two different examples for a function $H_{\boldsymbol{x}}$ with $H_{\boldsymbol{x}}(0)=0$ and $H'_{\boldsymbol{x}}(0)=1$. In addition the linear functions $t\mapsto H'_{\boldsymbol{x}}(0)t$, $t\mapsto aH'_{\boldsymbol{x}}(0)t$ with $a=1/3$ and the resulting value $\lambda(\boldsymbol{x})$ are depicted.

An obvious question is what particular value a we should choose. To this end, suppose for the moment that the function $H=H_{\boldsymbol{x}}$ is a quadratic and concave function:

$$H(t) \;=\; bt - ct^2/2$$

with constants $b>0$ and $c\geq 0$. Here $H'(t)=b-ct$, so

$$t_* := \operatorname*{arg\,max}_{[0,1]} H \;=\; \min(b/c,1).$$

Suppose first that $t_*=1$, i.e. $c\leq b$. Here H is strictly increasing on $[0,1]$. Hence it would be desirable to have $\lambda=\lambda(\boldsymbol{x})=1$, because otherwise $H(\lambda)\leq H(1/2)<H(1)$. Note that

$$H(1) = b-c/2 \;\geq\; b/2 \quad\text{and}\quad aH'(0) \;=\; ab.$$

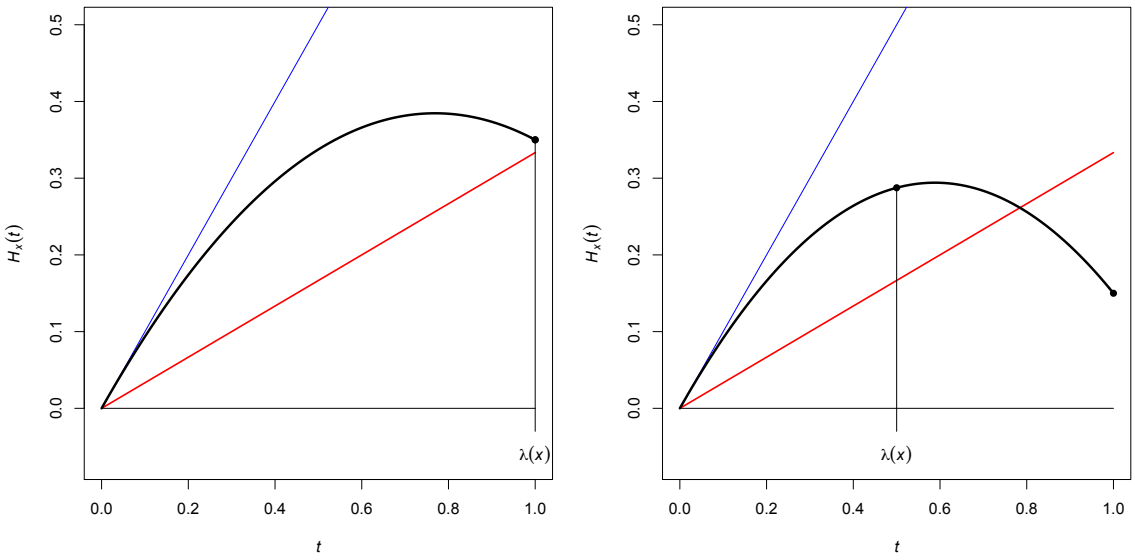

Figure 4.2: Step size function à la Goldstein–Armijo.

Hence to guarantee $H(1) \geq aH'(0)$ whenever $c \leq b$, we should choose

$$a \;\leq\; 1/2.$$

Now suppose that $t_* < 1$, i.e. $c > b$. Then $t_* = b/c$ and $H(t_*) = b^2/(2c)$, so

$$\frac{H(\lambda)}{H(t_*)} \;=\; \frac{b\lambda - c\lambda^2/2}{b^2/(2c)} \;=\; 2\lambda/t_* - (\lambda/t_*)^2 \;=\; 1 - (\lambda/t_* - 1)^2.$$

If $\lambda < t_*$, then $2\lambda \leq 1$, and the inequality $H(2\lambda) < a2\lambda H'(0)$ reads

$$2b\lambda - 2c\lambda^2 \;<\; ab2\lambda.$$

Dividing both sides by $2b\lambda$ yields the inequality $1 - \lambda/t_* < a$, so $0 < 1 - \lambda/t_* < a$ and

$$\frac{H(\lambda)}{H(t_*)} \;\geq\; 1 - a^2.$$

If $\lambda > t_*$, then the inequality $H(\lambda) \geq a\lambda H'(0)$ reads

$$b\lambda - c\lambda^2/2 \;\geq\; ab\lambda.$$

Dividing both sides by $b\lambda$ yields the inequality $1 - (\lambda/t_*)/2 \geq a$, so $1 < \lambda/t_* \leq 2(1-a)$ and

$$\frac{H(\lambda)}{H(t_*)} \;\geq\; 1 - (1-2a)^2.$$

Note that $1 - a^2$ is decreasing and $1 - (1-2a)^2$ is increasing in $a \in (0, 1/2]$. Equating both bounds yields

$$a \;=\; 1/3 \quad \text{and} \quad \frac{H(\lambda)}{H(t_*)} \;\geq\; 8/9.$$

More generally, for any value $a \in [1/4, 1/2]$ we could guarantee that $H(\lambda)/H(t_*) \geq 3/4$.

Coarse binary search for convex functions. Suppose that f is convex on the convex set $\mathcal{X}$, so all functions $H_{\boldsymbol{x}}$ are concave. A simple version of the step size function $\lambda(\cdot)$ is as follows: For $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$ let

$$\lambda(\boldsymbol{x}) \;:=\; 2^{-n(\boldsymbol{x})},$$

with $n(\boldsymbol{x})$ being the smallest number $n \in \mathbb{N}_0$ satisfying the two inequalities

$$0 \;<\; H_{\boldsymbol{x}}(2^{-n}/2) \;\leq\; H_{\boldsymbol{x}}(2^{-n}).$$

Figure 4.3 illustrates this definition of $\lambda(\boldsymbol{x})$. In both panels one sees $H_{\boldsymbol{x}}(t)$ for $t \in [0,1]$ as well as $\lambda(\boldsymbol{x})$ and $\lambda(\boldsymbol{x})/2$. In the left panel, $\lambda(\boldsymbol{x}) = 1$, because $0 < H_{\boldsymbol{x}}(1/2) \leq H_{\boldsymbol{x}}(1)$. In the right panel, $H_{\boldsymbol{x}}(1) < H_{\boldsymbol{x}}(1/2) \geq H_{\boldsymbol{x}}(1/4) > 0$, whence $\lambda(\boldsymbol{x}) = 1/2$.

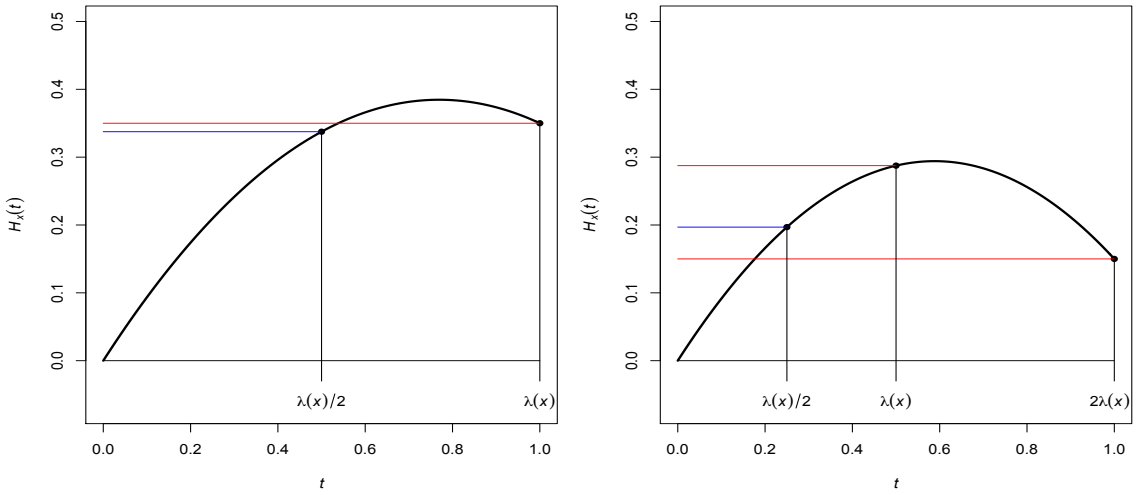

Figure 4.3: Step size function via coarse binary search.

This step size function fulfills condition (ii) in Section 4.2.2 with $C(\cdot) \geq 1/2$: Note first that by concavity of $H_{\boldsymbol{x}}$ and construction of $\lambda(\boldsymbol{x})$, the point $t(\boldsymbol{x}) := \max\bigl(\arg\max_{[0,1]} H_{\boldsymbol{x}}\bigr)$ has to fulfill

$$\lambda(\boldsymbol{x})/2 \;\leq\; t(\boldsymbol{x}) \;<\; 2\lambda(\boldsymbol{x}),$$

because $0 < H_{\boldsymbol{x}}(\lambda(\boldsymbol{x})/2) \leq H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))$ and $H_{\boldsymbol{x}}(\lambda(\boldsymbol{x})) > H_{\boldsymbol{x}}(2\lambda(\boldsymbol{x}))$ in case of $\lambda(\boldsymbol{x}) < 1$.

In case of $\lambda(\boldsymbol{x})/2 \leq t(\boldsymbol{x}) < \lambda(\boldsymbol{x})$, concavity of $H_{\boldsymbol{x}}$ implies that

$$\begin{aligned}
H_{\boldsymbol{x}}(\lambda(\boldsymbol{x})) \;\geq\; H_{\boldsymbol{x}}(\lambda(\boldsymbol{x})/2) \;&=\; H_{\boldsymbol{x}}\Bigl(\Bigl(1 - \frac{\lambda(\boldsymbol{x})/2}{t(\boldsymbol{x})}\Bigr) \cdot 0 + \frac{\lambda(\boldsymbol{x})/2}{t(\boldsymbol{x})}\, t(\boldsymbol{x})\Bigr) \\
&\geq\; \Bigl(1 - \frac{\lambda(\boldsymbol{x})/2}{t(\boldsymbol{x})}\Bigr) H_{\boldsymbol{x}}(0) + \frac{\lambda(\boldsymbol{x})/2}{t(\boldsymbol{x})}\, H_{\boldsymbol{x}}(t(\boldsymbol{x})) \\
&=\; \frac{\lambda(\boldsymbol{x})}{2t(\boldsymbol{x})}\, H_{\boldsymbol{x}}(t(\boldsymbol{x})) \\
&\geq\; \frac{H_{\boldsymbol{x}}(t(\boldsymbol{x}))}{2},
\end{aligned}$$

whence $H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))/H_{\boldsymbol{x}}(t(\boldsymbol{x})) \geq 1/2$.

In case of $\lambda(\boldsymbol{x}) < t(\boldsymbol{x}) < 2\lambda(\boldsymbol{x})$, concavity of $H_{\boldsymbol{x}}$ yields

$$\begin{aligned} H_{\boldsymbol{x}}(\lambda(\boldsymbol{x})) &= H_{\boldsymbol{x}}\Big(\Big(1 - \frac{\lambda(\boldsymbol{x})}{t(\boldsymbol{x})}\Big) \cdot 0 + \frac{\lambda(\boldsymbol{x})}{t(\boldsymbol{x})}\, t(\boldsymbol{x})\Big) \\ &\geq \Big(1 - \frac{\lambda(\boldsymbol{x})}{t(\boldsymbol{x})}\Big) H_{\boldsymbol{x}}(0) + \frac{\lambda(\boldsymbol{x})}{t(\boldsymbol{x})}\, H_{\boldsymbol{x}}(t(\boldsymbol{x})) \\ &= \frac{\lambda(\boldsymbol{x})}{t(\boldsymbol{x})}\, H_{\boldsymbol{x}}(t(\boldsymbol{x})) \\ &> \frac{H_{\boldsymbol{x}}(t(\boldsymbol{x}))}{2}, \end{aligned}$$

so again $H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))/H_{\boldsymbol{x}}(t(\boldsymbol{x})) \geq 1/2$.

It is also interesting to consider once more the special case of a quadratic and concave function $H = H_{\boldsymbol{x}}$, i.e. $H(t) = bt - ct^2/2$ with $b > 0$ and $c \geq 0$. Let $\lambda = \lambda(\boldsymbol{x})$ and $t_* := \min\{b/c, 1\}$, the unique maximizer of H on $[0, 1]$. In case of $c \leq b$, the function H is strictly increasing on $[0, 1]$, so $\lambda = t_*$. In case of $c > b$, $t_* \in (0, 1)$, and

$$\frac{H(\lambda)}{H(t_*)} = 1 - \Big(\frac{\lambda}{t_*} - 1\Big)^2,$$

see the considerations for the Goldstein–Armijo method. But the symmetry of H around t_* together with the inequalities $H(\lambda/2) \leq H(\lambda)$ and $H(2\lambda) < H(\lambda)$ in case of $\lambda < 1$ imply that even

$$\frac{3}{4}\lambda \leq t_* \leq \frac{3}{2}\lambda.$$

Hence $2/3 \leq \lambda/t_* \leq 4/3$ and

$$\frac{H(\lambda)}{H(t_*)} = 1 - \Big(\frac{\lambda}{t_*} - 1\Big)^2 \geq \frac{8}{9}.$$

Implementation of the quasi-Newton method with stepsize correction. The actual implementation of a quasi-Newton method with one of the former two stepsize corrections is rather simple, and one does not need to work with the auxiliary index $n(\boldsymbol{x})$ explicitly. Here is pseudo-code for a quasi-Newton method with the Goldstein–Armijo stepsize correction, where we assume that we have access to the functions $f : \mathcal{X} \to \mathbb{R}$, $\nabla f : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{A} : \mathcal{X} \to \mathbb{R}^{d\times d}_{\mathrm{sym},+}$, and $\delta_o > 0$ is a small threshold for the stopping criterion:

$\boldsymbol{x} \leftarrow \boldsymbol{x}_o$
$f \leftarrow f(\boldsymbol{x})$
$\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})$
$\Delta \leftarrow \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{g}$
$\delta \leftarrow \boldsymbol{g}^\top \Delta$
while $\delta > \delta_o$ **do**
    $\boldsymbol{x}_{\text{new}} \leftarrow \boldsymbol{x} - \Delta$
    $f_{\text{new}} \leftarrow f(\boldsymbol{x}_{\text{new}})$
    while $f - f_{\text{new}} < \delta/3$ **do**
        $\boldsymbol{x}_{\text{new}} \leftarrow (\boldsymbol{x} + \boldsymbol{x}_{\text{new}})/2$
        $\delta \leftarrow \delta/2$
        $f_{\text{new}} \leftarrow f(\boldsymbol{x}_{\text{new}})$
    end while
    $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{new}}$
    $f \leftarrow f_{\text{new}}$
    $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})$
    $\Delta \leftarrow \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{g}$
    $\delta \leftarrow \boldsymbol{g}^\top \Delta$
end while.

And here is pseudo-code for a variant with coarse binary search:

$\boldsymbol{x} \leftarrow \boldsymbol{x}_o$
$f \leftarrow f(\boldsymbol{x})$
$\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})$
$\Delta \leftarrow \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{g}$
$\delta \leftarrow \boldsymbol{g}^\top \Delta$
while $\delta > \delta_o$ **do**
    $\boldsymbol{x}_{\text{new}} \leftarrow \boldsymbol{x} - \Delta$
    $f_{\text{new}} \leftarrow f(\boldsymbol{x}_{\text{new}})$
    $\boldsymbol{x}'_{\text{new}} \leftarrow (\boldsymbol{x} + \boldsymbol{x}_{\text{new}})/2$
    $f'_{\text{new}} \leftarrow f(\boldsymbol{x}'_{\text{new}})$
    while $f_{\text{new}} \geq f$ **or** $f'_{\text{new}} < f_{\text{new}}$ **do**
        $\boldsymbol{x}_{\text{new}} \leftarrow \boldsymbol{x}'_{\text{new}}$
        $f_{\text{new}} \leftarrow f'_{\text{new}}$
        $\boldsymbol{x}'_{\text{new}} \leftarrow (\boldsymbol{x} + \boldsymbol{x}_{\text{new}})/2$
        $f'_{\text{new}} \leftarrow f(\boldsymbol{x}'_{\text{new}})$
    end while
    $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{new}}$
    $f \leftarrow f_{\text{new}}$
    $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})$
    $\Delta \leftarrow \boldsymbol{A}(\boldsymbol{x})^{-1}\boldsymbol{g}$
    $\delta \leftarrow \boldsymbol{g}^\top \Delta$
end while.

In practice, one should secure the inner while loop with a counter to make sure that it is not repeated endlessly due to numerical errors. And right after the inner while loop, one should include an extra check whether really $f_{\text{new}} < f$. If not, one should just set $\delta \leftarrow 0$ and thus stop the outer while loop.

Exercise 4.20 (Another step size function, I). Suppose that f is convex and continuously differentiable on the convex open set $\mathcal{X}$, and let $\Delta : \mathcal{X} \to \mathbb{R}^d$ be a candidate step function with the stated properties. For $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$ define

$$\lambda(\boldsymbol{x}) \;:=\; 2^{-n(\boldsymbol{x})}$$

with

$$n(\boldsymbol{x}) \;:=\; \min\bigl\{n \in \mathbb{N}_0 : H_{\boldsymbol{x}}(2 \cdot 2^{-n}) \geq 0\bigr\}.$$

(a) Show that this stepfunction satisfies the inequality

$$\frac{H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))}{\max_{[0,1]} H_{\boldsymbol{x}}} \;\geq\; \frac{1}{4}.$$

(b) Show that in case of a quadratic function $H_{\boldsymbol{x}}$,

$$\frac{H_{\boldsymbol{x}}(\lambda(\boldsymbol{x}))}{\max_{[0,1]} H_{\boldsymbol{x}}} \;\geq\; \frac{3}{4}.$$

4.2.5 Performance in connection with quasi-Newton methods

If we determine $\Delta(\boldsymbol{x})$ via a quasi-Newton method with good local convergence properties, it is desirable to have $\lambda(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$ sufficiently close to $\mathcal{X}_*$. Suppose that f is twice continuously differentiable and convex. Then $H_{\boldsymbol{x}}$ is real-valued and twice differentiable on $[0,1]$ whenever $\boldsymbol{x}$ is sufficiently close to $\mathcal{X}_*$. In that case, both the Goldstein–Armijo method with $a \leq 1/3$ and the rough binary search yield

$$\lambda(\boldsymbol{x}) \;=\; 1 \quad \text{whenever} \quad H''_{\boldsymbol{x}} \;\geq\; -\frac{4}{3}\, H'_{\boldsymbol{x}}(0) \;\text{ on } [0,1].$$

To verify this claim, note that

$$\begin{aligned} H_{\boldsymbol{x}}(1) \;&=\; H'_{\boldsymbol{x}}(0) + \int_0^1 (H'_{\boldsymbol{x}}(t) - H'_{\boldsymbol{x}}(0))\, dt \\ &\geq\; H'_{\boldsymbol{x}}(0) - \frac{4}{3}\, H'_{\boldsymbol{x}}(0) \int_0^1 t\, dt \\ &=\; \frac{H'_{\boldsymbol{x}}(0)}{3} \end{aligned}$$

and

$$\begin{aligned} H_{\boldsymbol{x}}(1) - H_{\boldsymbol{x}}(1/2) \;&=\; \frac{H'_{\boldsymbol{x}}(0)}{2} + \int_{1/2}^1 (H'_{\boldsymbol{x}}(t) - H'_{\boldsymbol{x}}(0))\, dt \\ &\geq\; \frac{H'_{\boldsymbol{x}}(0)}{2} - \frac{4}{3}\, H'_{\boldsymbol{x}}(0) \int_{1/2}^1 t\, dt \\ &=\; 0, \end{aligned}$$

because $H'_{\boldsymbol{x}}(t) - H'_{\boldsymbol{x}}(0) = H''_{\boldsymbol{x}}(\xi(t))t \geq -(4/3)H'_{\boldsymbol{x}}(0)t$ for some point $\xi(t) \in [0,t]$.

Specifically let $\Delta(\boldsymbol{x}) = -\boldsymbol{A}(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$ with a symmetric and positive definite matrix $\boldsymbol{A}(\boldsymbol{x})$ depending continuously on $\boldsymbol{x} \in \mathcal{X}$. For $\boldsymbol{x}_* \in \mathcal{X}_*$ and $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$ sufficiently close to $\boldsymbol{x}_*$, the whole line segment $\{\boldsymbol{x} + t\Delta(\boldsymbol{x}) : t \in [0,1]\}$ is contained in $\mathcal{X}$, and

$$\begin{aligned} H'_{\boldsymbol{x}}(0) &= -\nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x}) \\ &= \Delta(\boldsymbol{x})^\top A(\boldsymbol{x})\Delta(\boldsymbol{x}), \\ H''_{\boldsymbol{x}}(t) &= -\Delta(\boldsymbol{x})^\top D^2 f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))\Delta(\boldsymbol{x}). \end{aligned}$$

Hence

$$H''_{\boldsymbol{x}}(t) + \frac{4}{3}H'_{\boldsymbol{x}}(0) = \Delta(\boldsymbol{x})^\top \Big(\frac{4}{3}\boldsymbol{A}(\boldsymbol{x}) - D^2 f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))\Big)\Delta(\boldsymbol{x}) \geq 0,$$

provided that $(4/3)\boldsymbol{A}(\boldsymbol{x}) - D^2 f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))$ is positive semidefinite. Consequently, $\lambda(\boldsymbol{x}) = 1$ for $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_*$ sufficiently close to $\boldsymbol{x}_*$, provided that

$$\boldsymbol{A}(\boldsymbol{x}_*) - \frac{3}{4} D^2 f(\boldsymbol{x}_*) \quad \text{is positive definite.}$$

In particular, let f be twice continuously differentiable on the convex open set $\mathcal{X}$ with positive definite Hessian matrix $D^2 f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, and let $\Delta(\boldsymbol{x})$ be the Newton step $-D^2 f(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})$. Then iterating the algorithmic mapping $\boldsymbol{\psi}$ with arbitrary starting point $\boldsymbol{x}_0$ will yield a sequence $(\boldsymbol{x}_n)_{n\geq 0}$ with limit $\mathcal{X}_* = \{\boldsymbol{x}_*\}$, and $\lambda(\boldsymbol{x}_n) < 1$ for at most finitely many n. Hence the step size correction yields guaranteed convergence without sacrificing the rapid convergence of Newton's method.

Exercise 4.21 (Another step size function, II). Consider the step size function $\lambda(\cdot)$ introduced in Exercise 4.20. Suppose that $H_{\boldsymbol{x}}$ is twice differentiable on $[0,2]$ with $H''_{\boldsymbol{x}} \geq -CH'_{\boldsymbol{x}}(0)$ for some constant $C > 0$. Which value of C guarantees that $\lambda(\boldsymbol{x}) = 1$? Would you recommend this step size function in connection with the usual Newton candidate step function?

Exercise 4.22 (Logistic regression, III). Complement your Newton procedure for logistic regression in Exercise 4.11 with a step size function of your choice.

Exercise 4.23 (Smooth approximations). Several of our examples involved convex but non-differentiable functions $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \boldsymbol{x}_o\|$ or $\boldsymbol{x} \mapsto |\boldsymbol{x}_o^\top \boldsymbol{x} - r_o|$. A simple way to avoid the problems resulting from this lack of smoothness is to approximate these functions by smooth ones.

(a) For $\epsilon > 0$ let $h_\epsilon : \mathbb{R} \to \mathbb{R}$ be given by

$$h_\epsilon(x) := \sqrt{x^2 + \epsilon^2}.$$

Show that

$$|x| \leq h_\epsilon(x) \leq |x| + \min\Big(\epsilon, \frac{\epsilon^2}{2|x|}\Big),$$

and

$$\begin{aligned} h'_\epsilon(r) &= \frac{r}{\sqrt{\epsilon^2 + r^2}}, \\ h''_\epsilon(r) &= \frac{\epsilon^2}{(\epsilon^2 + r^2)^{3/2}}. \end{aligned}$$

(b) Now let $h_\epsilon(\boldsymbol{x}) := \sqrt{\|x\|^2 + \epsilon^2}$. Determine the gradient and Hessian matrix of h_ϵ. Show that the Hessian matrix is always positive definite.

(c) Design and implement a Newton method with regularization of the Hessian matrix and some step size correction for the (approximate) minimization of

$$f(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} |\boldsymbol{z}_i^\top \boldsymbol{x} - y_i|$$

with given numbers $y_1, \ldots, y_n \in \mathbb{R}$ and vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}^d$. Choose a parameter $\epsilon > 0$ which is related to the numbers $y_1, \ldots, y_n$ in a reasonable fashion.

Exercise 4.24. Let $X_1, X_2, \ldots, X_n$ be stochastically independent random variables with values in $\mathbb{R}$ and density function f_{μ_*,σ_*}. Here $\mu_* \in \mathbb{R}$ and $\sigma_* > 0$ are unknown parameters, and

$$f_{\mu,\sigma}(x) \;=\; \frac{1}{\sigma} f_{0,1}\Big(\frac{x-\mu}{\sigma}\Big) \quad \text{with} \quad f_{0,1}(y) \;:=\; \frac{1}{\pi(1+y^2)}.$$

Now we'd like to compute a *maximum-likelihood estimator* for the unknown parameter vector $\boldsymbol{\theta}_* := (\mu_*, \sigma_*)^\top$. That is a parameter vector $\boldsymbol{\theta} = (\mu, \sigma)^\top \in \mathbb{R} \times (0, \infty)$ minimizing the negative log-likelihood

$$\widehat{L}(\boldsymbol{\theta}) \;:=\; -\sum_{i=1}^{n} \log f_{\mu,\sigma}(X_i).$$

Implement a Newton method with step size correction for this task.

Chapter 5

Constrained Optimization

5.1 Lagrange Multipliers

5.1.1 The general principle

We consider functions $f : \mathcal{X} \to (-\infty, \infty]$ and $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^q$ on an arbitrary set $\mathcal{X}$. For a given vector $\boldsymbol{c} \in \mathbb{R}^q$ we would like to minimize f under the constraint that $\boldsymbol{g} = \boldsymbol{c}$ or that $\boldsymbol{g} \leq \boldsymbol{c}$ component-wise.

In many cases it is much easier to minimize the function $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ on $\mathcal{X}$ for an arbitrary vector $\boldsymbol{\lambda} \in \mathbb{R}^q$ and then to study the minimizer as a function of $\boldsymbol{\lambda}$. The next theorem explains to what extent this approach works.

Theorem 5.1. *For some $\boldsymbol{\lambda} \in \mathbb{R}^q$ let*

$$x_\lambda \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \big(f(x) - \boldsymbol{\lambda}^\top \boldsymbol{g}(x)\big).$$

Then

$$x_\lambda \in \operatorname*{arg\,min}_{x \in \mathcal{X} \,:\, \boldsymbol{g}(x) = \boldsymbol{g}(x_\lambda)} f(x).$$

More generally, let

$$\mathcal{X}_\lambda := \big\{x \in \mathcal{X} : \boldsymbol{\lambda}^\top \boldsymbol{g}(x) \geq \boldsymbol{\lambda}^\top \boldsymbol{g}(x_\lambda)\big\}.$$

Then

$$x_\lambda \in \operatorname*{arg\,min}_{x \in \tilde{\mathcal{X}}} f(x)$$

for any set $\tilde{\mathcal{X}}$ such that $x_\lambda \in \tilde{\mathcal{X}} \subset \mathcal{X}_\lambda$.

If x_λ is the unique minimizer of $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ on $\mathcal{X}$, then x_λ is also the unique minimizer of f on any set $\tilde{\mathcal{X}}$ such that $x_\lambda \in \tilde{\mathcal{X}} \subset \mathcal{X}_\lambda$.

Remark 5.2 (Equality constraints). For the original problem, minimizing f under the constraint that $\boldsymbol{g} = \boldsymbol{c}$, we obtain the following procedure: For $\boldsymbol{\lambda} \in \mathbb{R}^q$ we determine a minimizer $\boldsymbol{x}_\lambda$ of $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ over $\mathcal{X}$. Then we try to find $\boldsymbol{\lambda} \in \mathbb{R}^q$ such that $\boldsymbol{g}(\boldsymbol{x}_\lambda) = \boldsymbol{c}$. Then the corresponding point $\boldsymbol{x}_\lambda$ solves the original problem.

Remark 5.3 (Inequality constraints). Suppose for a vector $\boldsymbol{c} = (c_j)_{j=1}^q \in \mathbb{R}^q$ and some index $q_o \in \{1, \ldots, q\}$ we want to minimize $f(x)$ under the following constraints:

$$g_j(x) \begin{cases} \leq c_j & \text{for } 1 \leq j \leq q_o, \\ = c_j & \text{for } q_o < j \leq q. \end{cases} \tag{5.1}$$

For this purpose we try to find a vector $\boldsymbol{\lambda} \in \mathbb{R}^q$ such that the corresponding minimizer x_λ of $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ satisfies (5.1), and

$$\lambda_j \begin{cases} \leq 0 & \text{if } j \leq q_o, \\ = 0 & \text{if } j \leq q_o \text{ and } g_j(x_\lambda) < c_\lambda. \end{cases}$$

If such a vector $\boldsymbol{\lambda}$ exists, the corresponding x_λ solves the given constrained optimization problem, because any point $x \in \mathcal{X}$ satisfying (5.1) satisfies also $\lambda_j g_j(x) \geq \lambda_j g_j(x_\lambda)$ for $1 \leq j \leq q$, so $x \in \mathcal{X}_\lambda$.

Proof of Theorem 5.1. Let y be an arbitrary point in $\mathcal{X}_\lambda$ such that $f(y) \leq f(x_\lambda)$. Then

$$f(x_\lambda) - \boldsymbol{\lambda}^\top \boldsymbol{g}(x_\lambda) \;\geq\; f(y) - \boldsymbol{\lambda}^\top \boldsymbol{g}(x_\lambda) \;\geq\; f(y) - \boldsymbol{\lambda}^\top \boldsymbol{g}(y),$$

where the last inequality follows from the definition of $\mathcal{X}_\lambda$. Optimality of x_λ implies that both inequalities are equalities, so $f(y) = f(x_\lambda)$ and $\boldsymbol{\lambda}^\top \boldsymbol{g}(y) = \boldsymbol{\lambda}^\top \boldsymbol{g}(x_\lambda)$. If x_λ is even the unique minimizer of $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ on $\mathcal{X}$, then the inequality $f(y) \leq f(x_\lambda)$ implies that $y = x_\lambda$. □

Exercise 5.4 (Varying Lagrange multipliers). For two different vectors $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathbb{R}^q$ let

$$x_i \;\in\; \operatorname*{arg\,min}_{x \in \mathcal{X}} (f - \boldsymbol{\lambda}_i^\top \boldsymbol{g})(x), \quad i = 1, 2.$$

Show that

$$(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^\top (\boldsymbol{g}(x_1) - \boldsymbol{g}(x_2)) \;\geq\; 0.$$

Show also that this inequality is strict whenever the two minimizers x_1 and x_2 are unique and different.

5.1.2 Examples for Lagrange's Method

Optimal linear combinations of estimators. Let $X_1, X_2, \ldots, X_d$ be stochastically independent random variables with one and the same unknown expected value $\mu \in \mathbb{R}$ and known standard deviations $\sigma_i = \operatorname{Std}(X_i) > 0$. (An example are d different measurement devices with known imprecision, and with each of them one measures a sample from one and the same substance.) Now we want to estimate μ by a weighted average of the X_i,

$$\widehat{\mu} \;:=\; \sum_{i=1}^d w_i X_i$$

with certain weights $w_i \in \mathbb{R}$. A natural constraint on $\boldsymbol{w} = (w_i)_{i=1}^d$ is that

$$g(\boldsymbol{w}) := \sum_{i=1}^d w_i \;=\; 1,$$

because then $\mathbb{E}(\widehat{\mu}) = \mu$. Now we would like to minimize the mean quadratic error

$$\mathbb{E}\left((\widehat{\mu} - \mu)^2\right) \;=\; f(\boldsymbol{w}) := \sum_{i=1}^{d} w_i^2 \sigma_i^2$$

under that constraint. To this end we minimize for an arbitrary number $\lambda \in \mathbb{R}$ the function

$$f(\boldsymbol{w}) - \lambda g(\boldsymbol{w}) \;=\; \sum_{i=1}^{d} \big(w_i^2 \sigma_i^2 - \lambda w_i\big).$$

This minimization may be performed component-wise, and the unique minimizer is given by

$$\boldsymbol{w}_\lambda \;:=\; \Big(\frac{\lambda}{2\sigma_i^2}\Big)_{i=1}^{d}.$$

The constraint $g(\boldsymbol{w}) = 1$ is satisfied if and only if $\lambda = 2C$ with

$$C \;:=\; \Big(\sum_{i=1}^{d} \frac{1}{\sigma_i^2}\Big)^{-1}.$$

Hence optimal weights are given by

$$w_i \;:=\; C/\sigma_i^2,$$

and the resulting mean squared error equals C.

Minimizing a quadratic function under linear constraints. The previous example may be considered as a special case of the following optimization problem: Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be symmetric and positive definite. Now we would like to minimize

$$f(\boldsymbol{x}) \;:=\; 2^{-1}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$$

over all $\boldsymbol{x} \in \mathbb{R}^d$ satisfying the constraint

$$\boldsymbol{B}^\top \boldsymbol{x} \;=\; \boldsymbol{c}.$$

Here $\boldsymbol{B}$ is a given matrix in $\mathbb{R}^{d\times q}$ of rank $q < d$, and $\boldsymbol{c}$ is a given vector in $\mathbb{R}^q$.

To this end we minimize for $\boldsymbol{\lambda} \in \mathbb{R}^q$ the function $f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{B}^\top \boldsymbol{x} = f(\boldsymbol{x}) - (\boldsymbol{B}\boldsymbol{\lambda})^\top \boldsymbol{x}$. Note that the gradient and Hessian matrix of the latter function are given by $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}$ and $\boldsymbol{A}$, respectively. Thus its unique minimizer is given by

$$\boldsymbol{x}_\lambda \;:=\; \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda}.$$

Alternatively one could argue with quadratic completion: $f(\boldsymbol{x}) - (\boldsymbol{B}\boldsymbol{\lambda})^\top \boldsymbol{x}$ equals

$$\begin{aligned} 2^{-1}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - (\boldsymbol{B}\boldsymbol{\lambda})^\top \boldsymbol{x} \;&=\; 2^{-1}\Big(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - 2(\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda})^\top \boldsymbol{A}\boldsymbol{x}\Big) \\ &=\; 2^{-1}(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda})^\top \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda}) - 2^{-1}\boldsymbol{\lambda}^\top \boldsymbol{B}^\top \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda}. \end{aligned}$$

Furthermore,

$$\boldsymbol{B}^\top \boldsymbol{x}_\lambda \;=\; \boldsymbol{B}^\top \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{\lambda},$$

and this equals $\boldsymbol{c}$ if and only if

$$\boldsymbol{\lambda} \;=\; (\boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B})^{-1} \boldsymbol{c}.$$

Consequently the original constrained optimization problem has the unique solution

$$\boldsymbol{x}_* \;:=\; \boldsymbol{A}^{-1} \boldsymbol{B} (\boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B})^{-1} \boldsymbol{c}$$

with

$$f(\boldsymbol{x}_*) \;=\; 2^{-1} \boldsymbol{c}^\top (\boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B})^{-1} \boldsymbol{c}.$$

Exercise 5.5 (Minimizing a quadratic function, I). Generalize the previous considerations to the case of $f(\boldsymbol{x}) = \gamma + \boldsymbol{\beta}^\top \boldsymbol{x} + 2^{-1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$ with $\gamma \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$.

Exercise 5.6 (Minimizing a quadratic function, II). **(a)** Minimize

$$f(\boldsymbol{x}) \;=\; \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} + \boldsymbol{\beta}^\top \boldsymbol{x}$$

over all $\boldsymbol{x} \in \mathbb{R}^3$ satisfying

$$\boldsymbol{b}^\top \boldsymbol{x} \;=\; 4, \tag{5.2}$$

where

$$\boldsymbol{A} \;=\; \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} \;=\; \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{b} \;=\; \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

(b) What happens to the solution in part (a) if the constraint (5.2) is replaced with

$$\boldsymbol{b}^\top \boldsymbol{x} \;\leq (\geq)\; 4\,?$$

Density functions with maximal entropy. For a probability density p with respect to Lebesgue measure on a Borel set $\mathcal{X} \subset \mathbb{R}^d$ with $\mathrm{Leb}(\mathcal{X}) > 0$, the (differential Shannon) entropy

$$H(p) \;:=\; -\int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \, d\boldsymbol{x}$$

measures how "diffuse" the corresponding distribution is; higher values of $H(p)$ indicate a more diffuse distribution. An obvious question is what the most diffuse distribution looks like. Note that $H(p)$ is well-defined in $[-\infty, \infty)$ for arbitrary integrable functions $p : \mathcal{X} \to [0, \infty)$, with the convention that $0 \log 0 := 0$, because $t \mapsto t \log t$ is bounded from below on $[0, \infty)$.

More formally, we consider the set $\mathcal{P}$ of all nonnegative functions $p \in L^1(\mathcal{X})$ and the functional $f(p) := -H(p) \in (0, \infty]$. The additional constraint that $p \in \mathcal{P}$ is a probability density corresponds to the requirement

$$g_0(p) := \int_{\mathcal{X}} p(\boldsymbol{x}) \, d\boldsymbol{x} \;\stackrel{!}{=}\; 1.$$

Case 1: No further constraints, but $\mathrm{Leb}(\mathcal{X}) < \infty$. We want to maximize $H(p) = -f(p)$ among all $p \in \mathcal{X}$ with $g_0(p) = 1$. To this end we minimize

$$f(p) - \lambda g_0(p) \;=\; \int_{\mathcal{X}} \Big(p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \lambda p(\boldsymbol{x}) \Big)\, d\boldsymbol{x}.$$

Since

$$\frac{d}{dt}\,(t \log t - \lambda t) \;=\; \log t + 1 - \lambda,$$

the minimizing density is unique (almost everywhere) and given by

$$p_\lambda \;\equiv\; \exp(\lambda - 1).$$

With $\lambda := 1 - \log \mathrm{Leb}(\mathcal{X})$ we obtain the probability density $p_* \equiv \mathrm{Leb}(\mathcal{X})^{-1}$, the density of the uniform distribution on $\mathcal{X}$ with $H(p) = \log \mathrm{Leb}(\mathcal{X})$.

Case 2: Conditions on the first and second moments. In case of $\mathrm{Leb}(\mathcal{X}) = \infty$, there exists no maximizer $p \in \mathcal{P}$ of H under the constraint $g_0(p) = 1$. This can be verified by considering the uniform distribution on $\mathcal{X}^{(R)} := \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x}\| \le R\}$ for sufficiently large radius $R > 0$ with density $p^{(R)}(\boldsymbol{x}) := \mathrm{Leb}(\mathcal{X}^{(R)})^{-1} 1_{\mathcal{X}^{(R)}}(\boldsymbol{x})$ and entropy $H(p^{(R)}) = \log \mathrm{Leb}(\mathcal{X}^{(R)}) \to \infty$ as $R \to \infty$.

Suppose that $\mathcal{X} = \mathbb{R}^d$, and let $\mathcal{P}$ be the set of all measurable functions $p : \mathbb{R}^d \to [0, \infty)$ such that $\int p(\boldsymbol{x})(1 + \|\boldsymbol{x}\|^2)\, d\boldsymbol{x} < \infty$. Now we want to maximize $H(p)$ over all probability densities $p \in \mathcal{P}$ satisfying

$$\int \boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x} \;=\; \boldsymbol{\mu}_o \quad \text{and} \quad \int (\boldsymbol{x} - \boldsymbol{\mu}_o)(\boldsymbol{x} - \boldsymbol{\mu}_o)^\top p(\boldsymbol{x})\, d\boldsymbol{x} \;=\; \boldsymbol{\Sigma}_o$$

for a given vector $\boldsymbol{\mu}_o \in \mathbb{R}^d$ and a given symmetric and positive definite matrix $\boldsymbol{\Sigma}_o \in \mathbb{R}^{d \times d}$. Such a density p describes the distribution of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ with $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}_o$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_o$.

In other words, we want to minimize $f(p) = -H(p)$ under the constraints that

$$\begin{aligned} g_0(p) &:= \int p(\boldsymbol{x})\, d\boldsymbol{x} \;\overset{!}{=}\; 1, \\ \boldsymbol{g}_1(p) &:= \int \boldsymbol{x} p(\boldsymbol{x})\, d\boldsymbol{x} \;\overset{!}{=}\; \boldsymbol{\mu}_o, \\ \boldsymbol{G}_2(p) &:= \int (\boldsymbol{x} - \boldsymbol{\mu}_o)(\boldsymbol{x} - \boldsymbol{\mu}_o)^\top p(\boldsymbol{x})\, d\boldsymbol{x} \;\overset{!}{=}\; \boldsymbol{\Sigma}_o. \end{aligned}$$

To this end we consider for a number $\lambda_0 \in \mathbb{R}$, a vector $\boldsymbol{\lambda}_1 \in \mathbb{R}^d$ and a symmetric matrix $\boldsymbol{\Lambda}_2 \in \mathbb{R}^{d \times d}$ the quantity

$$\begin{aligned} &-H(p) - \lambda_0 g_0(p) - \boldsymbol{\lambda}_1^\top \boldsymbol{g}_1(p) - \mathrm{trace}(\boldsymbol{\Lambda}_2 \boldsymbol{G}_2(p)) \\ &\quad = \int \Big(\log p(\boldsymbol{x}) - \lambda_0 - \boldsymbol{\lambda}_1^\top \boldsymbol{x} - (\boldsymbol{x} - \boldsymbol{\mu}_o)^\top \boldsymbol{\Lambda}_2 (\boldsymbol{x} - \boldsymbol{\mu}_o) \Big) p(\boldsymbol{x})\, d\boldsymbol{x}. \end{aligned}$$

The minimizer of that auxiliary functional is unique (almost everywhere) and given by

$$p(\boldsymbol{x}) \;=\; \exp\big(-1 + \lambda_0 + \boldsymbol{\lambda}_1^\top \boldsymbol{x} + (\boldsymbol{x} - \boldsymbol{\mu}_o)^\top \boldsymbol{\Lambda}_2 (\boldsymbol{x} - \boldsymbol{\mu}_o)\big),$$

provided that $-\mathbf{\Lambda}_2$ is positive definite. Note that this is always a positive multiple of a Gaussian density. If we choose $\mathbf{\Lambda}_2 := -\mathbf{\Sigma}_o^{-1}/2$, $\boldsymbol{\lambda}_1 := \mathbf{0}$ and the constant $\lambda_0 := 1 - 2^{-1}d\log(2\pi) - 2^{-1}\log\det(\mathbf{\Sigma}_o)$, then

$$p(\boldsymbol{x}) \;=\; (2\pi)^{-d/2}\det(\mathbf{\Sigma}_o)^{-1/2}\exp\bigl(-(\boldsymbol{x}-\boldsymbol{\mu}_o)^\top\mathbf{\Sigma}_o^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_o)/2\bigr),$$

which is the density function of the d-variate Gaussian distribution $\mathcal{N}_d(\boldsymbol{\mu}_o, \mathbf{\Sigma}_o)$.

This shows that the Gaussian distribution $\mathcal{N}_d(\boldsymbol{\mu}_o, \mathbf{\Sigma}_o)$ is the most "diffuse" distribution (in the sense of entropy) among all probability distributions with given mean vector $\boldsymbol{\mu}_o$ and covariance matrix $\mathbf{\Sigma}_o$.

Exercise 5.7. Determine a probability density p on $(0,\infty)$ such that its entropy $H(p)$ is maximized under the constraint

$$\int_0^\infty xp(x)\,dx \;=\; \mu > 0.$$

Exercise 5.8 (Discrete Shannon entropy, I). For a vector $\boldsymbol{p} \in [0,\infty)^d$ let

$$H(\boldsymbol{p}) \;:=\; -\sum_{i=1}^d p_i \log p_i.$$

Maximize $H(\boldsymbol{p})$ under the constraint that $\sum_{i=1}^d p_i = 1$.

Exercise 5.9 (Discrete Shannon entropy, II). Let $\boldsymbol{P} \in [0,1]^{r\times s}$ be a matrix with given row and column sums:

$$\sum_{j=1}^s P_{ij} \;=\; a_i \quad \text{for } i = 1,\ldots,r$$

and

$$\sum_{i=1}^r P_{ij} \;=\; b_j \quad \text{for } j = 1,\ldots,s$$

with given probability vectors $\boldsymbol{a} \in (0,1)^r$ and $\boldsymbol{b} \in (0,1)^s$. Determine such a matrix $\boldsymbol{P}$ with maximal entropy

$$H(\boldsymbol{P}) \;:=\; -\sum_{i,j} P_{ij}\log P_{ij}.$$

Interpretation: The matrix $\boldsymbol{P}$ describes the joint distribution of two random variables X and Y with values in $\{1,\ldots,r\}$ and $\{1,\ldots,s\}$, respectively. Their respective marginal distributions are given by $\boldsymbol{a}$ and $\boldsymbol{b}$, and we are looking for a maximally diffuse joint distribution.

Kullback–Leibler divergence. Let P and Q be probability distributions on a measurable space $(\mathcal{X},\mathcal{B})$. The Kullback–Leibler divergence of Q with respect to P is defined as

$$K(Q\,|\,P) \;:=\; \int \log(q)\,dQ \;\in\; [0,\infty]$$

if Q has a density q with respect to P,[1] and $K(Q\,|\,P) := \infty$ otherwise. Note that

$$\int \log(q)\,dQ \;=\; -\int_{\{q>0\}} \log(1/q)q\,dP \;\geq\; -\int_{\{q>0\}} (1/q-1)q\,dP \;=\; 1 - P\{q>0\} \;\geq\; 0$$

[1] $Q(B) = \int_B q\,dP$ for arbitrary $B \in \mathcal{B}$

with equality if and only if $q = 1$ P-almost everywhere. Thus, $K(Q \,|\, P) \geq 0$ with equality if and only if $Q \equiv P$.

By means of Lagrange's method one can show that

$$K(Q \,|\, P) \;=\; \sup_{g \in \mathcal{G}(P)} \int g \, dQ,$$

where $\mathcal{G}(P)$ is the set of all measurable functions $g : \mathcal{X} \to [-\infty, \infty)$ such that $\int g \, dQ$ exists in $[-\infty, \infty]$ and

$$\int e^g \, dP \;=\; 1.$$

Suppose first that $Q(B) > 0 = P(B)$ for some $B \in \mathcal{B}$. Then for $r > 0$, $g_r := 1_B \cdot r$ defines a function in $\mathcal{G}(P)$ such that $\int g_r \, dQ = rQ(B) \to \infty$ as $r \to \infty$. Note also that the existence of such a set B implies that Q does not admit a density with respect to P, whence $K(Q \,|\, P) = \infty$.

If no such set B exists, it follows from the theorem of Radon–Nikodym that Q does admit a density q with respect to P. Then we may write $\int g \, dQ = \int gq \, dP$, and we consider the larger family $\bar{\mathcal{G}}(P)$ of all measurable functions $g : \mathcal{X} \to [-\infty, \infty)$ such that $\int g \, dQ$ exists in $[-\infty, \infty]$ and $\int e^g \, dP < \infty$. Instead of maximising $\int g \, dQ$ over all $g \in \mathcal{G}(P)$, we try to maximise

$$\int g \, dQ - \int e^g \, dP \;=\; \int (gq - e^g) \, dP$$

over all functions $g \in \bar{\mathcal{G}}(P)$, which is Lagrange's method with $\lambda = 1$. Indeed, for any $x \in \mathcal{X}$,

$$\operatorname*{arg\,max}_{h \in [-\infty, \infty)} (hq(x) - e^h) \;=\; \log q(x)$$

(with $\log 0 := -\infty$), so

$$\int (gq - e^g) \, dP \;\leq\; \int (\log(q) - 1) q \, dP \;=\; \int \log(q) \, dQ - 1$$

with equality if and only if $g = g_* := \log(q)$ P-almost everywhere. Since $g_* \in \mathcal{G}(P)$, this shows that $K(Q \,|\, P) = \int g_* \, dQ = \max_{g \in \mathcal{G}(P)} \int g \, dQ$.

Exercise 5.10. Given a strictly positive probability density p_o on $(0, 1)$ and numbers $\mu_o, \sigma_o \in (0, 1)$, determine a probability density p with minimal Kullback–Leibler divergence $K(p \,|\, p_o) := \int_0^1 \log(p(x)/p_o(x)) p(x) \, dx$ under the constraints that

$$\mu(p) \;:=\; \int_0^1 x p(x) \, dx \;=\; \mu_o \quad \text{and} \quad \sigma^2(p) \;:=\; \int_0^1 (x - \mu_o)^2 p(x) \, dx \;=\; \sigma_o^2.$$

With a suitable Lagrange ansatz you'll see immediately how the density p should look like.

(For which values of μ_o and σ_o does a solution exist?)

Optimal kernel functions. In connection with kernel density estimation as well as in connection with Wilcoxon's rank tests one encounters the following optimization problem: Find a measurable

function $K : \mathbb{R} \to [0, \infty)$ such that

$$g_0(K) := \int K(x)\,dx \stackrel{!}{=} 1,$$
$$g_1(K) := \int K(x)x\,dx \stackrel{!}{=} 0,$$
$$g_2(K) := \int K(x)x^2\,dx \stackrel{!}{=} 1$$

with minimal value of

$$f(K) := \int K(x)^2\,dx.$$

To solve this problem, let $\mathcal{K}$ be the set of all measurable functions $K : \mathbb{R} \to [0, \infty)$ such that

$$\int K(x)(1 + x^2)\,dx < \infty.$$

We just ignore the constraint $g_1(K) = 0$, hoping that the minimizer $K \in \mathcal{K}$ of $f(K)$ under the constraints $g_0(K) = g_2(K) = 1$ will automatically satisfy $g_1(K) = 0$.

The Lagrange ansatz leads to the auxiliary function

$$f(K) - \lambda_0 g_0(K) - \lambda_2 g_2(K) = \int \big(K(x)^2 - K(x)(\lambda_0 + \lambda_2 x^2)\big)\,dx$$

with a certain vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_2)^\top \in \mathbb{R}^2$. Now

$$K(x)^2 - K(x)(\lambda_0 + \lambda_2 x^2) = \big(K(x) - (\lambda_0 + \lambda_2 x^2)/2\big)^2 - 4^{-1}(\lambda_0 + \lambda_2 x^2)^2,$$

so an optimal function K would be given by

$$K_\lambda(x) = (\lambda_0/2 + \lambda_2 x^2/2)^+.$$

This defines a nontrivial function in $\mathcal{K}$ if and only if $\lambda_0 > 0 > \lambda_2$. With constants $a, b > 0$ we may also consider

$$K_{a,b}(x) := a(1 - (x/b)^2)^+.$$

Note first that with the substition $z = x/b$, $dx = b\,dz$,

$$\begin{aligned} g_0(K_{a,b}) &= a\int_{-b}^{b}(1 - (x/b)^2)\,dx \\ &= ab\int_{-1}^{1}(1 - z^2)\,dz = ab \cdot 2(z - z^3/3)\big|_{z=0}^{1} = \frac{4ab}{3}, \end{aligned}$$

so

$$a \stackrel{!}{=} \frac{3}{4b}.$$

Furthermore, since $K_{a,b}$ is an even function, $g_1(K_{a,b}) = 0$. Finally,

$$\begin{aligned} g_2(K_{a,b}) &= a\int_{-b}^{b}(1 - (x/b)^2)x^2\,dx \\ &= \frac{3b^2}{4}\int_{-1}^{1}(1 - z^2)z^2\,dz = ab^3 \cdot 2(z^3/3 - z^5/5)\big|_{z=0}^{1} = \frac{b^2}{5}, \end{aligned}$$

so

$$b \stackrel{!}{=} \sqrt{5}.$$

All in all, the original optimization problem has the (almost everywhere) unique solution

$$K(x) \;=\; \frac{3}{4\sqrt{5}}\Big(1-\frac{x^2}{5}\Big)^+,$$

a so-called Epanechnikov kernel.

The shape of a hanging chain. Let us consider a thin chain of length $2L$ which is hanging freely from two points at identical height and with distance $2M$, where $0 < M < L$. The question is which shape of the chain will materialize. We talk about chains rather than strings or ropes to emphasize that within the chain there are no elastic forces at work. Hence the only driving forces result from gravitation.

Formulating the problem mathematically, we are looking for a curve

$$[-L,L] \ni t \;\mapsto\; \big[x(t),y(t)\big]^\top \in \mathbb{R}^2$$

with the following properties:

- The functions x and y are continuously differentiable[2] on $[-L,L]$ with derivatives x' and y' satisfying

$$x'(t)^2 + y'(t)^2 \;=\; 1$$

for arbitrary $t \in [-L,L]$. That means, our curve is parametrized in arc length.

- The curve starts in $(-M,0)^\top$ and ends in $(M,0)^\top$, that means,

$$x(\pm L) \;=\; \pm M \quad \text{and} \quad y(\pm L) \;=\; 0.$$

Under these constraints we would like to minimize the potential energy

$$E \;:=\; \int_{-L}^{L} y(t)\,dt.$$

This definition comes from the consideration that for $t \in [-L,L)$ and a small $\delta > 0$, the section $\big\{(x(s),y(s)) : t \le s < t+\delta\big\}$ of our chain has mass proportional to δ and is hanging at height $y(t)$.

To solve this minimization problem we introduce

$$\boldsymbol{v}(t) \;:=\; \big[x'(t),y'(t)\big]^\top,$$

defining a measurable mapping $\boldsymbol{v}$ from $[-L,L]$ to the unit sphere of $\mathbb{R}^2$. Now we rewrite E and the constraints in terms of $\boldsymbol{v}$: On the one hand, it follows from $y(-L) = 0$ and $y(L) - y(-L) =$

[2] it would even suffice to assume that x and y are absolutely continuous.

$\int_{-L}^{L} y'(t)\,dt = 0$ that

$$\begin{aligned} E(\boldsymbol{v}) &= \int_{-L}^{L}\int_{-L}^{t} y'(s)\,ds\,dt \\ &= \int_{-L}^{L}\int_{s}^{L} y'(s)\,dt\,ds \\ &= \int_{-L}^{L} y'(s)(L-s)\,ds \\ &= \int_{-L}^{L} y'(s)(-s)\,ds. \end{aligned}$$

On the other hand, the constraint $x(L) - x(-L) = 2M$ is equivalent to

$$G(\boldsymbol{v}) := \int_{-L}^{L} x'(s)\,ds \;=\; 2M.$$

Consequently we are now trying the following ansatz: For some $\lambda \in \mathbb{R}$ we want to minimize the function

$$E(\boldsymbol{v}) - \lambda G(\boldsymbol{v}) \;=\; \int_{-L}^{L} \big(-y'(s)s - \lambda x'(s)\big)\,ds \;=\; -\int_{-L}^{L} \big\langle \boldsymbol{v}(s), (\lambda, s)^\top \big\rangle\,ds$$

with respect to $\boldsymbol{v}$. The solution may be determined point-wise: By means of the Cauchy–Schwarz inequality, the optimal $\boldsymbol{v}$ is given by

$$\boldsymbol{v}_\lambda(s) \;:=\; \Big(\frac{\lambda}{\sqrt{\lambda^2+s^2}}, \frac{s}{\sqrt{\lambda^2+s^2}}\Big)^\top,$$

obviously a continuous function of $s \in [-L, L]$. Only $\lambda > 0$ yields a reasonable solution because for $\lambda \le 0$ we would have $G(\boldsymbol{v}_\lambda) \le 0$. For the curve (x, y) itself this yields the following solution: Note that

$$\int \frac{s}{\sqrt{\lambda^2+s^2}}\,ds \;=\; C + \sqrt{\lambda^2+s^2}.$$

Combining this with the boundary conditions that $y_\lambda(\pm L) = 0$, we obtain

$$y_\lambda(s) \;=\; \sqrt{\lambda^2+s^2} - \sqrt{\lambda^2+L^2}.$$

Recall the hyperbolic functions $\cosh, \sinh : \mathbb{R} \to \mathbb{R}$ with $\cosh(r) = (e^r + e^{-r})/2$ and $\sinh(r) = (e^r - e^{-r})/2$. They satisfy the equation

$$\cosh \;=\; \sqrt{1+\sinh^2},$$

while their derivatives are given by $\cosh' = \sinh$ and $\sinh' = \cosh > 0$. Moreover, $\sinh$ is bijective with inverse function $\operatorname{arsinh} : \mathbb{R} \to \mathbb{R}$ such that

$$\operatorname{arsinh}'(r) \;=\; \frac{1}{\sqrt{1+r^2}}$$

for arbitrary $r \in \mathbb{R}$. From this and the boundary conditions $x_\lambda(\pm L) = \pm M$ one can easily derive that

$$x_\lambda(s) \;=\; \lambda \operatorname{arsinh}(s/\lambda),$$

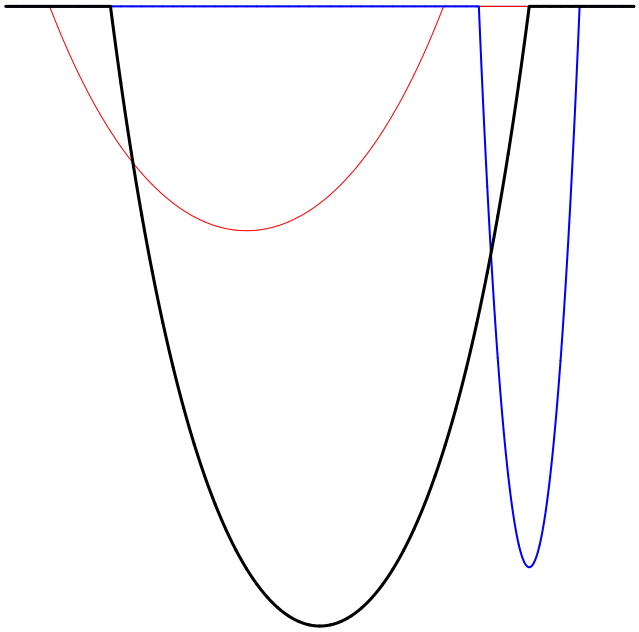

Figure 5.1: Some catenaries.

and we may rewrite this as

$$s \;=\; \lambda \sinh(x_\lambda(s)/\lambda).$$

Now the boundary condition $x_\lambda(\pm L) \;=\; \pm M$ is equivalent to the requirement that $L/\lambda \;=\; \sinh(M/\lambda)$, that is,

$$L/M \;=\; \frac{\sinh(M/\lambda)}{M/\lambda} \;=\; 1 + \sum_{k=1}^{\infty} \frac{(M/\lambda)^{2k}}{(2k+1)!}. \tag{5.3}$$

Since the right hand side is continuous and strictly decreasing in $\lambda > 0$ with limits ∞ as $\lambda \to 0$ and 1 as $\lambda \to \infty$, there exists a unique solution $\lambda > 0$ of the equation (5.3). For this particular $\lambda > 0$, the functions $x(\cdot) := x_\lambda(\cdot)$ and $y(\cdot) := y_\lambda(\cdot)$ are the desired solutions with

$$\begin{aligned} y \;&=\; \lambda\sqrt{1+\sinh(x/\lambda)^2} - \lambda\sqrt{1+(L/\lambda)^2} \\ &=\; \lambda\sqrt{1+\sinh(x/\lambda)^2} - \lambda\sqrt{1+\sinh(M/\lambda)^2} \\ &=\; \lambda\cosh(x/\lambda) - \lambda\cosh(M/\lambda). \end{aligned}$$

Hence the chain configuration is given by the set

$$\Big\{\big(x,\, \lambda\cosh(x/\lambda) - \lambda\cosh(M/\lambda)\big)^\top : x \in [-M, M]\Big\}.$$

Such a curve is also called a *catenary*. One of the first people to propose catenaries in architecture was the English natural philosopher, architect and polymath Robert Hooke (1635–1703). The famous Catalan architect Antoni Gaudi (1852–1926) used chain models to design a (never completed) church at Santa Coloma de Cervello. Figure 5.1 shows three different catenaries.

Exercise 5.11 (A smooth ride with the subway). **(a)** Determine for a given parameter $c > 0$ a twice continuously differentiable function $h : [0, 1] \to \mathbb{R}$ such that

$$h(0) = 0, \quad h(1) = 1, \quad h'(0) = h'(1) = 0, \quad |h''| \leq c$$

with minimal value of

$$U(h) \ := \ \int_0^1 h''(t)^2 \, dt.$$

For which parameters c does a solution exist at all?

(b) One may view $h(t)$ as the position of a subway along a given route at time t. At time 0 the subway starts at station A, and at time 1 it should arrive at station B. The cost functional $U(h)$ quantifies how problematic the ride would be for the material. For passengers, however, sudden changes in h'' or values $h''(0), h''(1) \neq 0$ would be considered as rather unpleasant. How would the solution to (a) change if we impose that h'' should be Lipschitz-continuous with constant $c_3 > 0$ and satisfy $h''(0) = h''(1) = 0$?

The Neyman–Pearson lemma. Suppose we observe a random variable X with unknown distribution P on a measurable space $(\mathcal{X}, \mathcal{B})$. We would like to test the null hypothesis "$P = P_0$" versus the alternative hypothesis "$P = P_1$" at a given test level $\alpha \in (0, 1)$, where P_0, P_1 are given probability distributions on $(\mathcal{X}, \mathcal{B})$.

That means, we are looking for a statistical test ϕ, i.e. a measurable function $\phi : \mathcal{X} \to [0, 1]$, where we reject "$P = P_0$" with probability $\phi(X)$. The probability of an error of the first kind equals

$$g(\phi) \ := \ \int_{\mathcal{X}} \phi \, dP_0$$

and should not exceed α, while the probability of an error of the second kind equals

$$f(\phi) \ := \ 1 - \int_{\mathcal{X}} \phi \, dP_1$$

and should be as small as possible.

To solve this optimization problem, let P_0 and P_1 have densities p_0 and p_1, respectively, with respect to some measure μ on $(\mathcal{X}, \mathcal{B})$. This is not a restriction: One could choose $\mu = P_0 + P_1$ and deduce from the Radon–Nikodym theorem the existence of such densities p_0 and p_1. Now we make the Lagrange ansatz and consider for $\lambda \in \mathbb{R}$ the functional

$$\phi \ \mapsto \ f(\phi) + \lambda g(\phi) \ = \ 1 + \int_{\mathcal{X}} \phi(\lambda p_0 - p_1) \, d\mu.$$

(Using $+\lambda g(\phi)$ instead of $-\lambda g(\phi)$ will be convenient notationally.) This functional may be minimized by minimizing the integrand pointwise: A test ϕ minimizes $f + \lambda g$ if and only if

$$\phi \ = \ \begin{cases} 1 & \mu\text{-almost everywhere on } \{p_1 > \lambda p_0\}, \\ 0 & \mu\text{-almost everywhere on } \{p_1 < \lambda p_0\}. \end{cases}$$

Specifically, let

$$\phi_{\lambda,\gamma}(x) \;:=\; \begin{cases} 1 & \text{if } p_1(x) > \lambda p_0(x) \\ \gamma & \text{if } p_1(x) = \lambda p_0(x) \\ 0 & \text{if } p_1(x) < \lambda p_0(x) \end{cases}$$

with $\gamma \in [0,1]$. Then

$$\int_{\mathcal{X}} \phi_{\lambda,\gamma}\, dP_0 \;=\; P_0\{p_1 > \lambda p_0\} + \gamma P_0\{p_1 = \lambda p_0\}.$$

Note that $H(\lambda) := P_0\{p_1 > \lambda p_0\}$ is monotone decreasing and right-continuous in $\lambda \in \mathbb{R}$ with

$$H(\lambda -) - H(\lambda) \;=\; P_0\{p_1 = \lambda p_0\},$$

so

$$\int_{\mathcal{X}} \phi_{\lambda,\gamma}\, dP_0 \;=\; H(\lambda) + \gamma\big(H(\lambda) - H(\lambda -)\big).$$

Moreover, $H(0-) = 1$ and $\lim_{\lambda\to\infty} H(\lambda) = 0$. Hence an optimal test of "$P = P_0$" versus "$P = P_1$" at level α is given by $\phi_{\lambda,\gamma}$, where

$$\lambda \;:=\; \min\{\lambda \in \mathbb{R} : H(\lambda) \le \alpha\} \;\ge\; 0,$$
$$\gamma \;:=\; \begin{cases} 0 & \text{if } H(\lambda) = \alpha, \\ \dfrac{\alpha - H(\lambda)}{H(\lambda -) - H(\lambda)} & \text{if } H(\lambda) > \alpha. \end{cases}$$

This optimal test was developed by Neyman [3] and Pearson [4] (1933).

5.1.3 Justification of Lagrange's method

At first glance, Lagrange's method looks like a cute trick which may work or fail, but we don't know when and why. By means of our results from convex analysis we can show that for convex functions $f : \mathbb{R}^d \to \mathbb{R}$ and linear functions $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^q$ the method *has* to work.

Theorem 5.12. *Let $\mathcal{X}$ be an open convex subset of $\mathbb{R}^d$, let $f : \mathcal{X} \to \mathbb{R}$ be a convex and $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^q$ be a linear function. Furthermore let $\boldsymbol{C}$ be a convex subset of $\mathbb{R}^q$. Suppose there exists a point*

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}\,:\,\boldsymbol{g}(\boldsymbol{x})\in\boldsymbol{C}} f(\boldsymbol{x}).$$

Then there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^q$ such that

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \big(f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\big)$$

and

$$\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) \;\ge\; \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_*) \quad \textit{whenever } \boldsymbol{x} \in \mathcal{X}, \boldsymbol{g}(\boldsymbol{x}) \in \boldsymbol{C}.$$

[3] Jerzy Neyman (1894–1981): Polish mathematician and statistician; pioneered hypothesis testing and confidence intervals.

[4] Egon S. Pearson (1895–1980): British statistician; son of the mathematician and biostatistician Karl Pearson (1857–1936).

Finally we formulate an important special case of Theorems 5.1 and 5.12 concerning optimization problems with linear inequality and equality constraints:

Theorem 5.13 (Karush–Kuhn–Tucker conditions). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set, $f : \mathcal{X} \to \mathbb{R}$ a convex function and $\boldsymbol{g} = (g_j)_{j=1}^q : \mathbb{R}^d \to \mathbb{R}^q$ a linear function. For certain numbers $c_1, c_2, \ldots, c_q \in \mathbb{R}$ and $q_o \in \{0, 1, \ldots, q\}$ let*

$$\mathcal{K} \;:=\; \big\{\boldsymbol{x} \in \mathcal{X} : g_j(\boldsymbol{x}) \le c_j \text{ if } j \le q_o, g_j(\boldsymbol{x}) = c_j \text{ if } j > q_o\big\}.$$

***(a)** Suppose there exist $\boldsymbol{\lambda} \in \mathbb{R}^q$ and*

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \big(f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\big)$$

such that $\boldsymbol{x}_* \in \mathcal{K}$ and

$$\lambda_j \begin{cases} \le\; 0 & \text{if } j \le q_o, \\ =\; 0 & \text{if } j \le q_o \text{ and } g_j(\boldsymbol{x}_*) < c_j. \end{cases} \tag{5.4}$$

Then

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{K}} f(\boldsymbol{x}).$$

***(b)** Suppose that $\mathcal{X}$ is open, and let*

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{K}} f(\boldsymbol{x}).$$

Then there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^q$ such that

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \big(f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\big)$$

and (5.4) *is satisfied.*

Proof of Theorem 5.12. Let $\mathcal{K} = \boldsymbol{g}^{-1}(\boldsymbol{C}) \cap \mathcal{X}$. By linearity of $\boldsymbol{g}$, convexity of $\boldsymbol{C}$ and $\mathcal{X}$ implies that $\mathcal{K}$ is a convex set, too. Thus

$$\mathcal{D} \;:=\; \mathcal{K} \times (-\infty, f(\boldsymbol{x}_*)) \quad \text{and} \quad \operatorname{epi}(f) \;=\; \big\{(\boldsymbol{x}, r) \in \mathcal{X} \times \mathbb{R} : f(\boldsymbol{x}) \le r\big\}$$

are convex and disjoint subsets of $\mathbb{R}^d \times \mathbb{R}$. According to Theorem 2.22, there exists a nonzero pair $(\boldsymbol{b}, t) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\boldsymbol{b}^\top \boldsymbol{y} + ts \;\le\; \boldsymbol{b}^\top \boldsymbol{x} + tr \quad \text{for all } (\boldsymbol{y}, s) \in \mathcal{D}, (\boldsymbol{x}, r) \in \operatorname{epi}(f).$$

Setting $\boldsymbol{y} = \boldsymbol{x} = \boldsymbol{x}_*$ and $s = f(\boldsymbol{x}_*) - 1, r = f(\boldsymbol{x}_*)$ shows that $t \ge 0$. Moreover, $t = 0$ would imply that $\boldsymbol{b} \ne \boldsymbol{0}$ and $\boldsymbol{b}^\top \boldsymbol{x} \ge \boldsymbol{b}^\top \boldsymbol{x}_*$ for all $\boldsymbol{x} \in \mathcal{X}$, a contradiction to $\mathcal{X}$ being an open set and $\boldsymbol{x}_* \in \mathcal{X}$. Hence we may assume without loss of generality that $t = 1$. Then the separation inequality is equivalent to

$$\boldsymbol{b}^\top \boldsymbol{y} + f(\boldsymbol{x}_*) \;\le\; \boldsymbol{b}^\top \boldsymbol{x} + f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{y} \in \mathcal{K}, \boldsymbol{x} \in \mathcal{X}. \tag{5.5}$$

Setting $\boldsymbol{x} = \boldsymbol{x}_*$ in (5.5) leads to

$$\boldsymbol{b}^\top \boldsymbol{y} \;\le\; \boldsymbol{b}^\top \boldsymbol{x}_* \quad \text{for all } \boldsymbol{y} \in \mathcal{K}. \tag{5.6}$$

But for sufficiently small $\delta > 0$, all points $\boldsymbol{y} = \boldsymbol{x}_* \pm \boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^d$, $\boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{0}$ and $\|\boldsymbol{w}\| < \delta$ belong to the set $\mathcal{K}$. Hence it follows from (5.6) that $\boldsymbol{b}^\top \boldsymbol{w} = 0$ for all $\boldsymbol{w} \in \mathbb{R}^d$ such that $\boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{0}$. If we write $\boldsymbol{g}(\boldsymbol{w}) = \boldsymbol{B}^\top \boldsymbol{w}$ with a matrix

$$\boldsymbol{B} \;=\; [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_q] \;\in\; \mathbb{R}^{d\times q},$$

then $\boldsymbol{b}$ is perpendicular to all vectors in $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q\}^\perp$. In other words, $\boldsymbol{b}$ has to be a linear combination of the vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$, see Exercise 5.14. Hence $\boldsymbol{b} = -\sum_{j=1}^q \lambda_j \boldsymbol{b}_j = -\boldsymbol{B}\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^q$, so $\boldsymbol{b}^\top \boldsymbol{x} = -\boldsymbol{\lambda}^\top \boldsymbol{B}^\top \boldsymbol{x} = -\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})$. Consequently we may deduce from (5.5) that

$$f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) \;\ge\; f(\boldsymbol{x}_*) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_*) \quad \text{for all } \boldsymbol{x} \in \mathcal{X},$$

while (5.6) is equivalent to

$$\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{y}) \;\ge\; \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_*) \quad \text{for all } \boldsymbol{y} \in \mathcal{K}.$$

□

Exercise 5.14 (Bi-orthogonal sets). Show that for any nonvoid set $\boldsymbol{M} \subset \mathbb{R}^d$,

$$(\boldsymbol{M}^\perp)^\perp \;=\; \operatorname{span}(\boldsymbol{M}),$$

where $\boldsymbol{M}^\perp := \{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{x}^\top \boldsymbol{y} = 0 \text{ for all } \boldsymbol{x} \in \boldsymbol{M}\}$.

Proof of Theorem 5.13. Part (a) follows immediately from Theorem 5.1 and Remarks 5.2 and 5.3.

It remains to prove part (b). According to Theorem 5.12 there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^q$ such that $\boldsymbol{x}_*$ minimizes $f - \boldsymbol{\lambda}^\top \boldsymbol{g}$ on $\mathcal{X}$ and

$$\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{y}) \;\ge\; \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_*) \quad \text{for all } \boldsymbol{y} \in \mathcal{K}.$$

In other words, $\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{v}) \ge 0$ for any vector $\boldsymbol{v} \in \mathbb{R}^q$ such that $\boldsymbol{x}_* + t\boldsymbol{v} \in \mathcal{K}$ for some $t > 0$. Since $\boldsymbol{x}_*$ is an interior point of $\mathcal{X}$, one may reformulate this statement as follows: For $\boldsymbol{v} \in \mathbb{R}^d$,

$$\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{v}) \;\ge\; 0 \quad \text{whenever} \quad g_i(\boldsymbol{v}) \begin{cases} \le\; 0 & \text{if } i \le q_o \text{ and } g_i(\boldsymbol{x}_*) = c_i, \\ =\; 0 & \text{if } i > q_o. \end{cases} \tag{5.7}$$

Unfortunately, the latter condition on $\boldsymbol{\lambda}$ alone does not guarantee the KKT conditions (5.4). To achieve the latter, we have to "sparsify" $\boldsymbol{\lambda}$ as follows: We write $g_j(\boldsymbol{v}) = \boldsymbol{b}_j^\top \boldsymbol{v}$ with certain vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q \in \mathbb{R}^d$ and define the index set

$$J = J(\boldsymbol{\lambda}) \;:=\; \big\{j \in \{1, \ldots, q\} : \lambda_j \ne 0 \text{ or } g_j(\boldsymbol{x}_*) = c_j\big\}.$$

Suppose that for some index j,

$$\lambda_j \;\ne\; 0 \quad \text{and} \quad \boldsymbol{b}_j \;\in\; \operatorname{span}\big(\boldsymbol{b}_i : i \in J \setminus \{j\}\big).$$

Then there exist real numbers μ_i, $i \in J \setminus \{j\}$ such that $\boldsymbol{b}_j = \sum_{i \in J \setminus \{j\}} \mu_i \boldsymbol{b}_i$. But then

$$\boldsymbol{\lambda}^\top \boldsymbol{g} \;\equiv\; \tilde{\boldsymbol{\lambda}}^\top \boldsymbol{g}$$

with

$$\tilde{\lambda}_i \;:=\; \begin{cases} \lambda_i + \mu_i \lambda_j & \text{if } i \in J \setminus \{j\}, \\ 0 & \text{else.} \end{cases}$$

Hence we may modify the vector $\boldsymbol{\lambda}$ step by step without changing $\boldsymbol{\lambda}^\top \boldsymbol{g}$ until finally its index set J has the following property:

$$\boldsymbol{b}_j \;\notin\; \operatorname{span}\bigl(\boldsymbol{b}_i : i \in J \setminus \{j\}\bigr) \quad \text{whenever } \lambda_j \;\neq\; 0. \tag{5.8}$$

This vector $\boldsymbol{\lambda}$ does satisfy the KKT conditions (5.4). Suppose the contrary. Then there exists an index $j \leq q_o$ such that either $\lambda_j > 0$ or $\lambda_j < 0$ and $g_j(\boldsymbol{x}_*) < c_j$. By property (5.8) there exists a vector $\boldsymbol{w} \in \mathbb{R}^d$ such that

$$\boldsymbol{b}_j^\top \boldsymbol{w} = 1 \quad \text{and} \quad \boldsymbol{b}_i^\top \boldsymbol{w} = 0 \text{ for all } i \in J \setminus \{j\}.$$

But then the vector $\boldsymbol{v} := -|\lambda_j| \boldsymbol{w}$ would satisfy the inequalities for $g_i(\boldsymbol{v})$ in (5.7), because $g_i(\boldsymbol{v}) = -|\lambda_j| 1_{[i=j]}$ for $i \in J$. But $\boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{v}) = -\lambda_j^2 < 0$. This contradiction shows that $\boldsymbol{\lambda}$ has to satisfy (5.4). □

Exercise 5.15 (Empirical likelihood). Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be pairwise different vectors in $\mathbb{R}^d$ which are not lying on a common hyperplane. Show that for any interior point $\boldsymbol{\mu}$ of $\operatorname{conv}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ there exists at least one vector $\boldsymbol{p} \in (0,1)^n$ such that

$$\sum_{i=1}^n p_i \;=\; 1 \quad \text{and} \quad \sum_{i=1}^n p_i \boldsymbol{x}_i \;=\; \boldsymbol{\mu}.$$

(a) Show that among all such vectors $\boldsymbol{p}$ there is precisely one minimizer of

$$f(\boldsymbol{p}) \;:=\; -\sum_{i=1}^n \log p_i.$$

Show that this minimizer has the form

$$p_i \;=\; (a + \boldsymbol{b}^\top \boldsymbol{x}_i)^{-1}$$

for suitable parameters $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^d$.

(b) Show that a vector $\boldsymbol{p} \in (0,1)^n$ solves the optimization problem in part (a) if and only if for some vector $\boldsymbol{b} \in \mathbb{R}^d$,

$$p_i \;=\; \frac{1}{n + \boldsymbol{b}^\top (\boldsymbol{x}_i - \boldsymbol{\mu})} \quad \text{with} \quad n + \boldsymbol{b}^\top (\boldsymbol{x}_i - \boldsymbol{\mu}) \;>\; 0 \quad \text{for } 1 \leq i \leq n$$

and

$$\sum_{i=1}^n \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{n + \boldsymbol{b}^\top (\boldsymbol{x}_i - \boldsymbol{\mu})} \;=\; \boldsymbol{0}.$$

(c) Show that the vector $\boldsymbol{b}$ in part (b) is the unique minimizer of

$$\tilde{f}(\boldsymbol{b}) \;:=\; -\sum_{i=1}^{n} \log\bigl(n + \boldsymbol{b}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})\bigr)$$

on $\mathbb{R}^d$, where $\log(r) := -\infty$ for $r \le 0$.

(d) Describe and implement a procedure to solve the optimization problem in part (a) by means of parts (b-c).

5.1.4 Lagrange duality

In connection with Theorem 5.12, there is a so-called dual approach to optimization under linear constraints. We elaborate on this in the following setting: Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and coercive, and let $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^q$ be linear and surjective with $q < d$. The goal is to minimize f over $N(\boldsymbol{g}) := \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}\}$.

By coercivity of f, there exists a point

$$\boldsymbol{x}_* \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x} \in N(\boldsymbol{g})} f(\boldsymbol{x}),$$

and by Theorem 5.12, there exists a vector $\boldsymbol{\lambda}_* \in \mathbb{R}^q$ such that

$$\boldsymbol{x}_* \in \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathbb{R}^d} \bigl(f(\boldsymbol{x}) - \boldsymbol{\lambda}_*^\top \boldsymbol{g}(\boldsymbol{x})\bigr).$$

We define an auxiliary function $L : \mathbb{R}^q \to [-\infty, \infty)$ via

$$L(\boldsymbol{\lambda}) \;:=\; \inf_{\boldsymbol{x} \in \mathbb{R}^d} \bigl(f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\bigr).$$

This function L is concave and upper semicontinuous, that is, $-L$ is convex and lower semicontinuous. Indeed, if $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^d$, for some matrix $\boldsymbol{A} \in \mathbb{R}^{q \times d}$, then

$$L(\boldsymbol{\lambda}) \;=\; -f^*(\boldsymbol{A}^\top \boldsymbol{\lambda})$$

with the Fenchel-Laplace transform f^* of f; see Chapter 3. An important fact is that

$$f(\boldsymbol{x}_*) \;=\; L(\boldsymbol{\lambda}_*) \;=\; \max_{\boldsymbol{\lambda} \in \mathbb{R}^q} L(\boldsymbol{\lambda}).$$

Indeed, since $\boldsymbol{x}_* \in N(\boldsymbol{g})$,

$$L(\boldsymbol{\lambda}_*) \;=\; f(\boldsymbol{x}_*) - \boldsymbol{\lambda}_*^\top \boldsymbol{g}(\boldsymbol{x}_*) \;=\; f(\boldsymbol{x}_*),$$

and for any $\boldsymbol{\lambda} \in \mathbb{R}^q$,

$$L(\boldsymbol{\lambda}) \;\le\; f(\boldsymbol{x}_*) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_*) \;=\; f(\boldsymbol{x}_*).$$

These findings suggest the following strategy to minimize f over $N(\boldsymbol{g})$:

Step 1. Find a vector

$$\widehat{\boldsymbol{\lambda}} \;\in\; \operatorname*{arg\,max}_{\boldsymbol{\lambda} \in \mathbb{R}^q} L(\boldsymbol{\lambda}).$$

Step 2. Determine a point

$$\widehat{\boldsymbol{x}} \;\in\; \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d}\bigl(f(\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x})\bigr)$$

and check whether $\widehat{\boldsymbol{x}} \in N(\boldsymbol{g})$. If yes, this point $\widehat{\boldsymbol{x}}$ minimizes f over $N(\boldsymbol{g})$.

Remark 5.16. If L is strictly concave on $\{L > -\infty\}$, then Step 1 yields automatically $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_*$, and Step 2 is possible in the sense that some minimizer of $f(\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x})$ over all $\boldsymbol{x} \in \mathbb{R}^q$ does belong to $N(\boldsymbol{g})$. More generally, let $\boldsymbol{g}^*$ be the conjugate mapping of $\boldsymbol{g}$, that is, the mapping $\boldsymbol{\lambda} \mapsto \boldsymbol{A}^\top\boldsymbol{\lambda}$ if $\boldsymbol{g}$ is given by $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$. Since

$$f(\boldsymbol{x}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) - \boldsymbol{g}^*(\boldsymbol{\lambda})^\top \boldsymbol{x},$$

one can write

$$L(\boldsymbol{\lambda}) \;=\; \tilde{L}(\boldsymbol{g}^*(\boldsymbol{\lambda}))$$

with a concave function $\tilde{L} : N(\boldsymbol{g})^\perp \to [-\infty, \infty)$. Here we utilise the fact that $N(\boldsymbol{g})^\perp = \boldsymbol{g}^*(\mathbb{R}^q)$. If the latter function is strictly concave on $\{\tilde{L} > -\infty\}$, the set $\widehat{\mathcal{X}}$ coincides with the set $\mathcal{X}_* = \operatorname{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^q}(f(\boldsymbol{x}) - \boldsymbol{\lambda}_*^\top \boldsymbol{x})$, which contains $\boldsymbol{x}_*$.

Remark 5.17. Suppose that Step 1 yields some vector $\widehat{\boldsymbol{\lambda}}$, and let $\widehat{\boldsymbol{x}}$ be any minimizer of $f(\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x})$ over all $\boldsymbol{x} \in \mathbb{R}^q$. If f is strictly convex, then $\widehat{\boldsymbol{x}} = \boldsymbol{x}_*$. To see this, note that

$$f((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) - \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) \;=\; f((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) - u\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\widehat{\boldsymbol{x}})$$

is a convex function of $u \in \mathbb{R}$ with value $L(\widehat{\boldsymbol{\lambda}}) = L(\boldsymbol{\lambda}_*) = f(\boldsymbol{x}_*)$ for $u \in \{0, 1\}$, whence

$$f((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) \;\le\; f(\boldsymbol{x}_*) + u\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\widehat{\boldsymbol{x}}) \quad \text{for } u \in [0, 1].$$

On the other hand, by definition of $L(\widehat{\boldsymbol{\lambda}})$,

$$f(\boldsymbol{x}_*) = L(\widehat{\boldsymbol{\lambda}}) \;\le\; f((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) - u\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\widehat{\boldsymbol{x}})$$

for arbitrary $u \in \mathbb{R}$, whence

$$f((1-u)\boldsymbol{x}_* + u\widehat{\boldsymbol{x}}) \;=\; f(\boldsymbol{x}_*) + u\widehat{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\widehat{\boldsymbol{x}}) \quad \text{for } u \in [0, 1].$$

If f is strictly convex, this implies that $\widehat{\boldsymbol{x}} = \boldsymbol{x}_*$.

Example 5.18 (Least squares with ℓ^1-penalties). For a given vector $\boldsymbol{y} \in \mathbb{R}^n$ and a matrix $\boldsymbol{D} \in \mathbb{R}^{q\times n}$, consider the function $H : \mathbb{R}^n \to \mathbb{R}$ with

$$H(\boldsymbol{\beta}) \;:=\; \|\boldsymbol{y} - \boldsymbol{\beta}\|^2/2 + \|\boldsymbol{D}\boldsymbol{\beta}\|_1,$$

where $\|\cdot\|$ is standard Euclidean norm, and $\|\boldsymbol{z}\|_1 := \sum_{i=1}^q |z_i|$. Minimizing H over all vectors $\boldsymbol{\beta} \in \mathbb{R}^n$ is equivalent to minimizing

$$f(\boldsymbol{\beta}, \boldsymbol{z}) \;:=\; \|\boldsymbol{y} - \boldsymbol{\beta}\|^2/2 + \|\boldsymbol{z}\|_1$$

over all pairs $(\boldsymbol{\beta}, \boldsymbol{z}) \in \mathbb{R}^n \times \mathbb{R}^q$ under the constraint that $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{z}) := \boldsymbol{z} - \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{0}$. We may identify $\mathbb{R}^n \times \mathbb{R}^q$ with $\mathbb{R}^d$, where $d = n + q$, and apply the previous considerations. Note first that for $\boldsymbol{\lambda} \in \mathbb{R}^q$,

$$f(\boldsymbol{\beta}, \boldsymbol{z}) - \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{z}) \;=\; \|\boldsymbol{y} - \boldsymbol{\beta}\|^2/2 + \boldsymbol{\lambda}^\top \boldsymbol{D}\boldsymbol{\beta} + \sum_{i=1}^q (|z_i| - \lambda_i z_i).$$

For fixed $\boldsymbol{\beta}$,

$$\inf_{\boldsymbol{z}\in\mathbb{R}^q}\sum_{i=1}^{q}(|z_i|-\lambda_i z_i) \;=\; \begin{cases}0 & \text{if } \|\boldsymbol{\lambda}\|_\infty \le 1\\ -\infty & \text{if } \|\boldsymbol{\lambda}\|_\infty > 1\end{cases}$$

with $\|\boldsymbol{\lambda}\|_\infty := \max_{i=1,\ldots,q}|\lambda_i|$. In case of $\|\boldsymbol{\lambda}\|_\infty \le 1$, the minimum is attained at $\boldsymbol{z}$ if and only if

$$\text{for } i=1,\ldots,q,\quad z_i \begin{cases}= 0 & \text{if } |\lambda_i|<1,\\ \ge 0 & \text{if } \lambda_i = 1,\\ \le 0 & \text{if } \lambda_i = -1.\end{cases} \tag{5.9}$$

Moreover,

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{\beta}\|^2/2+\boldsymbol{\lambda}^\top\boldsymbol{D}\boldsymbol{\beta} \;&=\; 2^{-1}\big(\boldsymbol{\beta}^\top\boldsymbol{\beta}-2\boldsymbol{\beta}^\top(\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda})+\|\boldsymbol{y}\|^2\big)/2\\ &\ge\; \|\boldsymbol{y}\|^2/2-\|\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}\|^2/2\end{aligned}$$

with equality if and only if

$$\boldsymbol{\beta} \;=\; \boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}.$$

Hence,

$$\begin{aligned}L(\boldsymbol{\lambda}) \;&=\; \inf_{(\boldsymbol{\beta},\boldsymbol{z})\in\mathbb{R}^n\times\mathbb{R}^q}\big(f(\boldsymbol{\beta},\boldsymbol{z})-\boldsymbol{\lambda}^\top\boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{z})\big)\\ &=\; \begin{cases}-\infty & \text{if } \|\boldsymbol{\lambda}\|_\infty > 1,\\ \|\boldsymbol{y}\|^2-\|\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}\|^2/2 & \text{if } \|\boldsymbol{\lambda}\|_\infty \le 1,\end{cases}\end{aligned}$$

and in case of $\|\boldsymbol{\lambda}\|_\infty \le 1$,

$$\operatorname*{arg\,min}_{(\boldsymbol{\beta},\boldsymbol{z})\in\mathbb{R}^n\times\mathbb{R}^q}\big(f(\boldsymbol{\beta},\boldsymbol{z})-\boldsymbol{\lambda}^\top\boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{z})\big) \;=\; \big\{(\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda},\boldsymbol{z}) : \boldsymbol{z}\in\mathcal{Z}(\boldsymbol{\lambda})\big\}$$

with $\mathcal{Z}(\boldsymbol{\lambda})$ denoting the set of all vectors $\boldsymbol{z}\in\mathbb{R}^q$ satisfying (5.9).

Consequently, to obtain the unique minimizer $\boldsymbol{\beta}_*$ of H, one may proceed as follows: Find a vector

$$\widehat{\boldsymbol{\lambda}} \;\in\; \operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\mathbb{R}^q:\|\boldsymbol{\lambda}\|_\infty\le 1}\|\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}\|^2.$$

Then,

$$\widehat{\boldsymbol{\beta}} \;:=\; \boldsymbol{y}-\boldsymbol{D}^\top\widehat{\boldsymbol{\lambda}}$$

coincides with the unique minimizer $\boldsymbol{\beta}_*$ of $H(\boldsymbol{\beta})=\|\boldsymbol{y}-\boldsymbol{\beta}\|^2/2+\|\boldsymbol{D}\boldsymbol{\beta}\|_1$ over all $\boldsymbol{\beta}\in\mathbb{R}^n$. The reason is that if we would replace $\widehat{\boldsymbol{\lambda}}$ with the Lagrange vector $\boldsymbol{\lambda}_*$ corresponding to $\boldsymbol{\beta}_*$, then $\boldsymbol{\lambda}_*$ would also minimize $\|\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}\|^2$ over all $\boldsymbol{\lambda}\in\mathbb{R}^q$ satisfying $\|\boldsymbol{\lambda}\|_\infty\le 1$, and $\boldsymbol{\beta}_*=\boldsymbol{y}-\boldsymbol{D}^\top\boldsymbol{\lambda}_*$. But $\|\cdot\|^2$ is a strictly convex function on $\mathbb{R}^n$, so $\boldsymbol{D}^\top\widehat{\boldsymbol{\lambda}}=\boldsymbol{D}^\top\boldsymbol{\lambda}_*$, whence $\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}_*$.

5.2 Special Algorithms

In what follows we consider a continuous and convex function $f:\mathbb{R}^d\to(-\infty,\infty]$ which is continuously differentiable and strictly convex on $\mathcal{X}:=\operatorname{dom}(f)=\{x\in\mathbb{R}^d : f(x)<\infty\}$. Furthermore let $\mathcal{K}$ be a closed convex subset of $\mathbb{R}^d$ such that

$$\mathcal{X}\cap\mathcal{K} \;\neq\; \emptyset,$$

and we assume that $\{\boldsymbol{x} \in \mathcal{K} : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_o)\}$ is compact for arbitrary $\boldsymbol{x}_o \in \mathcal{X} \cap \mathcal{K}$. These conditions imply existence of a unique minimizer

$$\boldsymbol{x}_* := \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{K}} f(\boldsymbol{x})$$

Subsequently we describe algorithms to compute $\boldsymbol{x}_*$. Our general results about convex functions show that $\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K}$ equals $\boldsymbol{x}_*$ if and only if

$$\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \geq 0 \quad \text{for arbitrary } \boldsymbol{y} \in \mathcal{K}. \tag{5.10}$$

This condition will be encountered repeatedly.

Exercise 5.19. Suppose that $\mathcal{K} = [0, \infty)^d$. Show that $\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K}$ equals $\boldsymbol{x}_*$ if and only if for $1 \leq j \leq d$,

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j} \begin{cases} \geq 0, \\ = 0 \quad \text{if } x_j > 0. \end{cases}$$

5.2.1 Iterative algorithms

As in Section 4.2 we iterate a mapping $\boldsymbol{\psi} : \mathcal{X} \cap \mathcal{K} \to \mathcal{X} \cap \mathcal{K}$ of type

$$\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x} + \lambda(\boldsymbol{x}) \Delta(\boldsymbol{x})$$

with a candidate step function $\Delta : \mathcal{X} \cap \mathcal{K} \to \mathbb{R}^d$ and a step size function $\lambda : \mathcal{X} \cap \mathcal{K} \to [0, 1]$. The regularity conditions on Δ and λ are essentially identical with the ones in Section 4.2.2:

(i) Δ is continuous with $\Delta(\boldsymbol{x}_*) = \boldsymbol{0}$ while

$$\boldsymbol{x} + \Delta(\boldsymbol{x}) \in \mathcal{K} \quad \text{and} \quad \nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x}) < 0 \quad \text{for } \boldsymbol{x} \in \mathcal{X} \cap \mathcal{K} \setminus \{\boldsymbol{x}_*\}.$$

(ii) With

$$C(\boldsymbol{x}) := \frac{f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))}{\max_{t \in [0,1]} \big[f(\boldsymbol{x}) - f(\boldsymbol{x} + t\Delta(\boldsymbol{x}))\big]} \quad \text{and} \quad \tilde{C}(\boldsymbol{x}) := \frac{f(\boldsymbol{x}) - f(\boldsymbol{x} + \lambda(\boldsymbol{x})\Delta(\boldsymbol{x}))}{-\nabla f(\boldsymbol{x})^\top \Delta(\boldsymbol{x})}$$

for $\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K} \setminus \{\boldsymbol{x}_*\}$, we assume that for any fixed $\boldsymbol{y} \in \mathcal{X} \cap \mathcal{K} \setminus \{\boldsymbol{x}_*\}$,

$$\liminf_{\mathcal{X} \cap \mathcal{K} \ni \boldsymbol{x} \to \boldsymbol{y}} C(\boldsymbol{x}) > 0 \quad \text{or} \quad \liminf_{\mathcal{X} \cap \mathcal{K} \ni \boldsymbol{x} \to \boldsymbol{y}} \tilde{C}(\boldsymbol{x}) > 0.$$

With the same arguments as in Section 4.2.2 one can show that iteration of $\boldsymbol{\psi}$ with arbitrary starting point $\boldsymbol{x}_0 \in \mathcal{X} \cap \mathcal{K}$ yields a sequence converging to $\{\boldsymbol{x}_*\}$.

A special procedure for $\mathcal{K} = [0, \infty)^d$

In several settings one may parametrize the optimization problem such that the constraints correspond to the convex cone $\mathcal{K} = [0, \infty)^d$. Now we derive an explicit example for the candidate

step function Δ: As in case of quasi-Newton procedures we approximate f locally by a quadratic function. Precisely let

$$f_{\boldsymbol{x}}(\boldsymbol{y}) \;:=\; f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + 2^{-1}(\boldsymbol{y}-\boldsymbol{x})^\top \boldsymbol{A}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$$

for $\boldsymbol{x} \in \mathcal{X} \cap \mathcal{K}$ with a diagonal matrix

$$\boldsymbol{A}(\boldsymbol{x}) \;=\; \operatorname{diag}\big(a_1(\boldsymbol{x}), \ldots, a_d(\boldsymbol{x})\big)$$

and continuous diagonal elements $a_j : \mathcal{X} \cap \mathcal{K} \to (0,\infty)$. Explicit examples are

$$a_j(\boldsymbol{x}) \;:=\; 1, \quad a_j(\boldsymbol{x}) \;:=\; \epsilon(\boldsymbol{x}) \quad \text{or} \quad a_j(\boldsymbol{x}) \;:=\; \frac{\partial^2 f(\boldsymbol{x})}{\partial x_j^2}.$$

In the latter case we assume that these partial derivatives exist and are continuous and strictly positive on $\mathcal{X} \cap \mathcal{K}$.

With $g_j(\boldsymbol{x}) := \nabla f(\boldsymbol{x})_j$ one may write

$$\begin{aligned} f_{\boldsymbol{x}}(\boldsymbol{y}) \;&=\; f(\boldsymbol{x}) + \sum_{j=1}^{d} \Big(g_j(\boldsymbol{x})(y_j - x_j) + 2^{-1} a_j(\boldsymbol{x})(y_j - x_j)^2 \Big) \\ &=\; c(\boldsymbol{x}) + 2^{-1} \sum_{j=1}^{d} a_j(\boldsymbol{x}) \Big(y_j^2 - 2y_j \Big(x_j - \frac{g_j(\boldsymbol{x})}{a_j(\boldsymbol{x})} \Big) \Big) \end{aligned}$$

with $c(\boldsymbol{x})$ not depending on $\boldsymbol{y}$. As a function von $\boldsymbol{y} \in \mathcal{K}$ this is minimal if and only if

$$y_j \;=\; \Big(x_j - \frac{g_j(\boldsymbol{x})}{a_j(\boldsymbol{x})} \Big)^+$$

for $1 \le j \le d$. Hence we define

$$\Delta(\boldsymbol{x}) \;:=\; \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathcal{K}} f_{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x} \;=\; \Big(\Big(x_j - \frac{g_j(\boldsymbol{x})}{a_j(\boldsymbol{x})} \Big)^+ - x_j \Big)_{j=1}^{d}.$$

One can easily verify that this candidate step function Δ does fulfill Condition (i).

An inverse problem from Physics

In a basic experiment in particle physics a random number $Y \ge 0$ of certain particles is generated, and these particles fly off independently into different directions, the directions being uniformly distributed on the unit sphere of $\mathbb{R}^3$. The goal is to estimate the unknown weight function $q : \mathbb{N}_0 \to [0,\infty)$ with

$$q(y) \;:=\; \mathbb{P}(Y = y),$$

but the total number Y cannot be measured directly. Instead one uses a detector which counts the number $X \ge 0$ of particles flying off into a direction which is covered by the detector. The relative size $\rho \in (0,1)$ of these detectable directions is given. So one measurement yields a random variable X such that

$$\mathbb{P}(X = x \,|\, Y = y) \;=\; K(x \,|\, y) := \binom{y}{x} \rho^x (1-\rho)^{y-x},$$

and

$$p(x) := \mathbb{P}(X = x) \;=\; Kq(x) := \sum_{y=0}^{\infty} K(x \,|\, y) q(y).$$

Suppose we repeat this basic experiments $n \gg 1$ times, resulting in independent random counts $X_1, X_2, \ldots, X_n$ with distribution p. We estimate the weight function p by the empirical weight function $\widehat{p}$ with

$$\widehat{p}(x) \;=\; \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i = x]}.$$

As shown in Exercise 5.21, the mapping $q \mapsto p = Kq$ is linear and invertible, but the inverse mapping is ill-conditioned in the sense that small deviations of $\widehat{p}$ from p may lead to rather large errors in the reconstruction of q.

An alternative approach is to minimize the functional

$$f(q) \;:=\; -\sum_{x=0}^{\infty} \widehat{p}(x) \log Kq(x) + \sum_{y=0}^{\infty} q(y)$$

over all weight functions $q : \mathbb{N}_0 \to [0, \infty)$. Indeed, one can show that in case of $\widehat{p} \equiv p = Kq$, the unique minimizer of f is given by q. In general, any minimizer q of f satisfies automatically the constraint $\sum_y q(y) = 1$, because for any weight function q with $f(q) < \infty$ and arbitrary numbers $t > 0$,

$$\frac{d}{dt} f(tq) \;=\; \sum_{y=0}^{\infty} q(y) - t^{-1}.$$

Note that for integers $0 \le x \le y$,

$$\frac{K(x \,|\, y+1)}{K(x \,|\, y)} \;=\; \frac{1-\rho}{1 - x/(y+1)} \;<\; 1 \quad \text{if and only if} \quad y \;>\; x/\rho - 1.$$

Hence, if we define $\widehat{x} := \max(X_1, \ldots, X_n)$ and

$$\widehat{y} \;:=\; \begin{cases} 0 & \text{if } \widehat{x} = 0 \\ \lfloor \widehat{x}/\rho \rfloor - 1 & \text{if } \widehat{x} > 0 \end{cases}$$

for any weight function q and $x \le \widehat{x}$, the value $Kq(x)$ gets strictly larger if we replace

$$\big(q(\widehat{y}), q(\widehat{y}+1), q(\widehat{y}+2), q(\widehat{y}+3), \ldots\big) \quad \text{with} \quad \Big(\sum_{y \ge \widehat{y}} q(y), 0, 0, 0, \ldots\Big),$$

unless $q(y) = 0$ for all $y > \widehat{y}$. Consequently, we have to solve a finite-dimensional minimization problem:

With the dimensions $c := \widehat{x} + 1$ and $d := \widehat{y} + 1$, the vector $\widehat{\boldsymbol{p}} := \big(\widehat{p}(i-1)\big)_{i=1}^{c}$ and the matrix $\boldsymbol{K} := \big(K(i-1 \,|\, j-1)\big)_{1 \le i \le c, 1 \le j \le d}$ we want to minimize

$$f(\boldsymbol{q}) \;:=\; -\sum_{i=1}^{c} \widehat{p}_i \log(\boldsymbol{K}\boldsymbol{q})_i + \sum_{j=1}^{d} q_j$$

over all vectors $\boldsymbol{q} \in [0,\infty)^d$, where $\boldsymbol{q}$ corresponds to $\big(q(j-1)\big)_{j=1}^d$. Note that with the convention $\log(r) := -\infty$ for $r \le 0$, the function $f : \mathbb{R}^d \to (-\infty,\infty]$ is continuous with $f(\boldsymbol{q}) < \infty$ if and only if

$$(\boldsymbol{K}\boldsymbol{q})_i \;>\; 0 \quad \text{whenever} \quad \widehat{p}_i \;>\; 0.$$

Moreover, for $\boldsymbol{q} = t\boldsymbol{u}$ with $t \ge 0$ and $\boldsymbol{u} \in \Sigma_d := \big\{\boldsymbol{v} \in [0,1]^d : \sum_{j=1}^d v_j = 1\big\}$,

$$f(\boldsymbol{q}) \;=\; -\log t + t - 1 + f(\boldsymbol{u}) \;\ge\; -\log t + t - 1 + \min_{\boldsymbol{v}\in\Sigma_d} f(\boldsymbol{v}) \;\to\; \infty \quad \text{as } t \to \{0,\infty\}.$$

Hence the set $\big\{\boldsymbol{x} \in [0,\infty)^d : f(\boldsymbol{x}) \le f(\boldsymbol{q})\big\}$ is compact for any $\boldsymbol{q} \in [0,\infty)^d \cap \mathrm{dom}(f)$. Finally, f is twice continuously differentiable on $\mathrm{dom}(f)$ with

$$\begin{aligned} g_j(\boldsymbol{q}) := \frac{\partial}{\partial q_j} f(\boldsymbol{q}) \;&=\; 1 - \sum_{i=1}^c \frac{\widehat{p}_i K_{ij}}{(\boldsymbol{K}\boldsymbol{q})_i}, \\ a_j(\boldsymbol{q}) := \frac{\partial^2}{\partial q_j^2} f(\boldsymbol{q}) \;&=\; \sum_{i=1}^c \frac{\widehat{p}_i K_{ij}^2}{(\boldsymbol{K}\boldsymbol{q})_i^2}. \end{aligned}$$

Hence we may apply the general optimization algorithm described before.

Remark 5.20 (EM algorithm). For the particular problem here, many statisticians would use a so-called expectation-minimization algorithm. Suppose we would observe independent pairs $(X_1,Y_1),\dots,(X_n,Y_n)$ with Y_i having distribution q and $\mathbb{P}(X_i = x \,|\, Y_i = y) = K(x \,|\, y)$. Then one could estimate q by the unobserved empirical weight function $\check{q}$ with

$$\check{q}(y) \;:=\; \frac{1}{n}\sum_{i=1}^n 1_{[Y_i=y]}.$$

Indeed, $\check{q}$ is the unique minimizer of

$$\check{f}(q) \;:=\; -\sum_{y=0}^\infty \check{q}(y)\log q(y) + \sum_{y=0}^\infty q(y)$$

Since we observe only $X_1,\dots,X_n$, the idea is to replace a current estimator q_k for the true q by

$$\begin{aligned} q_{k+1}(y) \;&:=\; \mathbb{E}_{q_k}\big(\check{q}(y) \,\big|\, X_1,\dots,X_n\big) \\ &=\; \sum_{x=0}^{\widehat{x}} \widehat{p}(x)\,\mathbb{P}_{q_k}(Y_1 = y \,|\, X_1 = x) \\ &=\; \sum_{x=0}^{\widehat{x}} \frac{\widehat{p}(x)K(x \,|\, y)q_k(y)}{\sum_{z\ge 0} K(x \,|\, z)q_k(z)}. \end{aligned}$$

Here $\mathbb{E}_{q_k}$ and $\mathbb{P}_{q_k}$ denote expectations and probabilites in case of $q \equiv q_k$. Although this approach is rather intuitive and easy to implement, the resulting algorithm converges slower than the algorithm we derived before.

Exercise 5.21. Show that for $s \in \mathbb{R}$ and $k \in \mathbb{N}_0$,

$$\mathbb{E}\big([X]_k s^{X-k}\big) \;=\; \rho^k\,\mathbb{E}\big([Y]_k(1-\rho+\rho s)^{Y-k}\big),$$

provided that both sides are well-defined in $\mathbb{R}$. Here $[z]_0 := 1$ and $[z]_k := \prod_{i=0}^{k-1}(z-i)$ for $z \in \mathbb{R}$ and $k \in \mathbb{N}$.

Deduce from the previous identities that

$$\mathbb{E}(Y) \;=\; \rho^{-1}\,\mathbb{E}(X) \quad \text{and} \quad \mathrm{Var}(Y) \;=\; \rho^{-2}\,\mathrm{Var}(X) - \rho^{-2}(1-\rho)\,\mathbb{E}(X).$$

Furthermore, show that for arbitrary $k \in \mathbb{N}_0$,

$$q(k) \;=\; \frac{1}{\rho^k k!}\,\mathbb{E}\big([X]_k(1-1/\rho)^{X-k}\big),$$

provided that $\mathbb{E}(s_o^Y) < \infty$ for some $s_o > 2(1-\rho)$.

5.2.2 Active set methods

In this section we concentrate on a polyhedral cone

$$\mathcal{K} \;=\; \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \le 0 \text{ for } i = 1,\dots,q\big\}$$

with $q \le d$ *linearly independent* vectors $\boldsymbol{v}_1,\dots,\boldsymbol{v}_q \in \mathbb{R}^d$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ let

$$J(\boldsymbol{x}) \;:=\; \big\{i \in \{1,\dots,q\} : \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle = 0\big\}.$$

If $\boldsymbol{x} \in \mathcal{K}$, then $J(\boldsymbol{x})$ specifies the "set of active constraints (active set) at $\boldsymbol{x}$", that is, the set of inequality constraints which are even equalities for the point $\boldsymbol{x}$. For an arbitrary index set $J \subset \{1,\dots,q\}$ let

$$\mathcal{X}_J \;:=\; \mathcal{X} \cap \big\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{v}_i\rangle = 0 \text{ for all } i \in J\big\},$$

i.e. the intersection of the open set $\mathcal{X} = \mathrm{dom}(f)$ with a $(d - \#J)$-dimensional linear subspace of $\mathbb{R}^d$. That is, $\boldsymbol{x} \in \mathcal{X}_J$ if and only if $J(\boldsymbol{x}) \supset J$. Moreover, $\boldsymbol{x} \in \mathcal{X}_J$ belongs to $\mathcal{K}$ if and only if

$$\langle \boldsymbol{x}, \boldsymbol{v}_i\rangle \;\le\; 0 \quad \text{for all } i \in \{1,\dots,q\} \setminus J.$$

The idea of active set methods is to determine for an index set $J \subset \{1,\dots,q\}$ the point

$$\boldsymbol{x}_J \;:=\; \operatorname*{arg\,min}_{\mathcal{X}_J} f$$

(if possible) and to modify the set J finitely many times until $\boldsymbol{x}_J = \{\boldsymbol{x}_*\}$ eventually. Here we assume that minimizing f over such a set $\mathcal{X}_J$ is sufficiently easy. In case of a quadratic function f this is certainly the case, see Section 5.1.2.

An obvious question is how to infer for a given index set J that the vector $\boldsymbol{x}_J$ is equal to $\boldsymbol{x}_*$. To this end we chose a "dual basis" $\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_d$ of $\mathbb{R}^d$ in the sense that for $i \in \{1,\dots,d\}$ and $j \in \{1,\dots,q\}$,

$$\langle \boldsymbol{b}_i, \boldsymbol{v}_j\rangle \;=\; 1_{[i=j]}.$$

(Such a dual basis always exists. For instance, one may augment the given vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ to a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ of $\mathbb{R}^d$, and then the rows of the matrix $[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d]^\top := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d]^{-1}$ have the desired property.) With these basis vectors $\boldsymbol{b}_i$ we may write

$$\mathcal{K} \;=\; \Big\{\sum_{i=1}^{d} \lambda_i \boldsymbol{b}_i : \boldsymbol{\lambda} \in \mathbb{R}^d, \lambda_i \le 0 \text{ for } i = 1, \ldots, q\Big\},$$

$$\mathcal{X}_J \;=\; \mathcal{X} \cap \Big\{\sum_{i=1}^{d} \lambda_i \boldsymbol{b}_i : \boldsymbol{\lambda} \in \mathbb{R}^d, \lambda_i = 0 \text{ for all } i \in J\Big\}.$$

Lemma 5.22 (Characterizing $\boldsymbol{x}_J$ and $\boldsymbol{x}_*$). *Let $J \subset \{1, \ldots, q\}$ and $\boldsymbol{x} \in \mathcal{X}_J$.*

(a) *The vector $\boldsymbol{x}$ equals $\boldsymbol{x}_J$ if and only if*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle \;=\; 0 \quad \text{for all } i \in \{1, \ldots, d\} \setminus J. \tag{5.11}$$

(b) *If $\boldsymbol{x} \in \mathcal{K}$ and satisfies* (5.11)*, then it equals $\boldsymbol{x}_*$ if and only if*

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_j \rangle \;\le\; 0 \quad \text{for all } j \in J. \tag{5.12}$$

Proof of Lemma 5.22. Part (a) is an easy consequence of convexity of f plus the fact that $\mathcal{X}_J$ is the intersection of the open set $\mathcal{X}$ with the linear subspace $\operatorname{span}\big\{\boldsymbol{b}_i : i \in \{1, \ldots, d\} \setminus J\big\} \subset \mathbb{R}^d$.

Concerning part (b), suppose first that $\boldsymbol{x} = \boldsymbol{x}_*$. For arbitrary $j \in J$, the vector $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{b}_j$ lies in $\mathcal{K}$, so by (5.10),

$$0 \;\le\; \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \;=\; -\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_j \rangle.$$

Hence $\boldsymbol{x}$ has to fulfill (5.12).

On the other hand, suppose that $\boldsymbol{x}$ lies within $\mathcal{K}$ and satisfies both (5.11) and (5.12). Any vector $\boldsymbol{y} \in \mathcal{K}$ may be written as $\boldsymbol{x} + \sum_{i=1}^{d} \lambda_i \boldsymbol{b}_i$, where $\lambda_i \le 0$ in case of $\langle \boldsymbol{x}, \boldsymbol{b}_i \rangle = 0$. In particular, $\lambda_i \le 0$ for all $i \in J$. Thus,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \;=\; \sum_{i \in J} \underbrace{\lambda_i}_{\le 0} \underbrace{\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle}_{\le 0} + \sum_{i \in \{1, \ldots, d\} \setminus J} \lambda_i \underbrace{\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle}_{=0} \;\ge\; 0.$$

Consequently, $\boldsymbol{x}$ satisfies (5.10) and hence is equal to $\boldsymbol{x}_*$. □

An active set algorithm for "quadratic programming"

We consider the special case of a quadratic function f with positive definite Hessian matrix. The active set algorithm alternates finitely many times between two basic procedures. Both precedures return a pair $(\boldsymbol{x}, F)$ consisting of a candidate $\boldsymbol{x} \in \mathcal{K}$ for $\boldsymbol{x}_*$ and the value $F = f(\boldsymbol{x})$.

An important notion of the algorithm is *local optimality* of a candidate $\boldsymbol{x} \in \mathcal{K}$ for $\boldsymbol{x}_*$. This means that

$$\boldsymbol{x} \;=\; \boldsymbol{x}_{J(\boldsymbol{x})}.$$

That is, $\boldsymbol{x}$ lies in $\mathcal{K}$ and minimizes f over the linear subspace $\mathcal{X}_{J(\boldsymbol{x})}$ of all points $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ such that $J(\tilde{\boldsymbol{x}}) \supset J(\boldsymbol{x})$.

Starting point. A natural starting point is $\boldsymbol{x}_{\{1,\ldots,\}}$, because $\mathcal{X}_{\{1,\ldots,q\}} \subset \mathcal{K}$. This point is clearly locally optimal.

Basic procedure 1 (checking optimality; deactivating constraints). Suppose that $\boldsymbol{x} \in \mathcal{K}$ is locally optimal. Then we check Condition (5.12) in Lemma 5.22: If $\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle \leq 0$ for all $i \in J(\boldsymbol{x})$, we know that $\boldsymbol{x} = \boldsymbol{x}_*$ and stop the algorithm. Otherwise we replace $\boldsymbol{x}$ with some vector $\boldsymbol{x}_{\mathrm{new}} \in \mathcal{K}$ such that $f(\boldsymbol{x}_{\mathrm{new}}) < f(\boldsymbol{x})$.

Here is a particular variant for $\boldsymbol{x}_{\mathrm{new}}$: For some nonvoid set $\tilde{J}(\boldsymbol{x}) \subset J(\boldsymbol{x})$ such that $\langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle > 0$ for all $i \in \tilde{J}(\boldsymbol{x})$, we compute

$$\boldsymbol{x}_{\mathrm{new}} \;=\; \boldsymbol{x} - \sum_{i \in \tilde{J}(\boldsymbol{x})} \langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle \, \boldsymbol{b}_i .$$

This vector is certainly in $\mathcal{K}$. Then we check whether $f(\boldsymbol{x}_{\mathrm{new}}) < f(\boldsymbol{x})$. As long as this is not the case, we replace $\boldsymbol{x}_{\mathrm{new}}$ with $2^{-1}(\boldsymbol{x} + \boldsymbol{x}_{\mathrm{new}}) \in \mathcal{K}$. After finitely many steps, $f(\boldsymbol{x}_{\mathrm{new}}) < f(\boldsymbol{x})$, because

$$\lim_{t \to 0+} t^{-1}\big(f((1-t)\boldsymbol{x} + t\boldsymbol{x}_{\mathrm{new}}) - f(\boldsymbol{x})\big) \;=\; \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}_{\mathrm{new}} - \boldsymbol{x} \rangle \;=\; -\sum_{i \in \tilde{J}(\boldsymbol{x})} \langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_i \rangle^2$$

is strictly negative. The fact that a finite number of updates $\boldsymbol{x}_{\mathrm{new}} \leftarrow (\boldsymbol{x} + \boldsymbol{x}_{\mathrm{new}})/2$ is sufficient is true in theory. In practice, if a moderate number of updates does not yield a better candidate $\boldsymbol{x}_{\mathrm{new}}$, we conclude that $\boldsymbol{x}$ is equal to $\boldsymbol{x}_*$ up to numerical errors.

Basic procedure 2 (local search). Starting from an arbitrary candidate $\boldsymbol{x} \in \mathcal{K}$ for $\boldsymbol{x}_*$, we want to replace it with a point $\boldsymbol{x}_{\mathrm{new}} \in \mathcal{K}$ which is locally optimal and satisfies

$$J(\boldsymbol{x}_{\mathrm{new}}) \;\supset\; J(\boldsymbol{x}), \tag{5.13}$$

$$f(\boldsymbol{x}_{\mathrm{new}}) \;\leq\; f(\boldsymbol{x}). \tag{5.14}$$

To this end, we determine $J := J(\boldsymbol{x})$ and compute $\boldsymbol{x}_{\mathrm{new}} := \boldsymbol{x}_J$. In case of $f(\boldsymbol{x}_{\mathrm{new}}) = f(\boldsymbol{x})$ we know that $\boldsymbol{x}_{\mathrm{new}} = \boldsymbol{x}$ is locally optimal, and requirements (5.13), (5.14) are obviously satisfied.

If $f(\boldsymbol{x}_{\mathrm{new}}) < f(\boldsymbol{x})$, we know that $f((1-t)\boldsymbol{x} + t\boldsymbol{x}_{\mathrm{new}})$ is strictly decreasing in $t \in [0,1]$. Note that by definition of $\boldsymbol{x}_{\mathrm{new}}$,

$$\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \;=\; 0 \;=\; \langle \boldsymbol{x}_{\mathrm{new}}, \boldsymbol{v}_i \rangle \quad \text{for all } i \in J,$$

while the definition of J implies that

$$\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle \;<\; 0 \quad \text{whenever } i \in \{1,\ldots,q\} \setminus J.$$

Thus, $\boldsymbol{x}_{\mathrm{new}} \in \mathcal{K}$ if and only if

$$V(J, \boldsymbol{x}_{\mathrm{new}}) \;:=\; \big\{ i \in \{1,\ldots,q\} \setminus J : \langle \boldsymbol{x}_{\mathrm{new}}, \boldsymbol{v}_i \rangle > 0 \big\}$$

$(\boldsymbol{x}, F) \leftarrow \mathbf{BasicProcedure2}(\boldsymbol{x})$	
$F \leftarrow f(\boldsymbol{x})$	
$J \leftarrow J(\boldsymbol{x})$	**(**)**
$\boldsymbol{x}_{\rm new} \leftarrow \boldsymbol{x}_J$	
$F_{\rm new} \leftarrow f(\boldsymbol{x}_{\rm new})$	
while $F_{\rm new} < F$ **do**	
$\quad V \leftarrow V(J, \boldsymbol{x}_{\rm new})$	**(**)**
$\quad$ **if** $V = \emptyset$ **then**	
$\quad\quad \boldsymbol{x} \leftarrow \boldsymbol{x}_{\rm new}$	
$\quad\quad F \leftarrow F_{\rm new}$	
$\quad$ **else**	
$\quad\quad t_o \leftarrow t_o(J, \boldsymbol{x}_{\rm new})$	
$\quad\quad \boldsymbol{x}_{\rm new} \leftarrow (1 - t_o)\boldsymbol{x} + t_o \boldsymbol{x}_{\rm new}$	
$\quad\quad F_{\rm new} \leftarrow f(\boldsymbol{x}_{\rm new})$	
$\quad\quad$ **if** $F_{\rm new} < F$ **then**	**(*)**
$\quad\quad\quad \boldsymbol{x} \leftarrow \boldsymbol{x}_{\rm new}$	
$\quad\quad\quad F \leftarrow F_{\rm new}$	
$\quad\quad\quad J \leftarrow J(\boldsymbol{x})$	**(**)**
$\quad\quad\quad \boldsymbol{x}_{\rm new} \leftarrow \boldsymbol{x}_J$	
$\quad\quad\quad F_{\rm new} \leftarrow f(\boldsymbol{x}_{\rm new})$	
$\quad\quad$ **end if**	
$\quad$ **end if**	
end while.	

Table 5.1: Basic procedure 2 for quadratic f.

is empty. In that case, it follows from $J(\boldsymbol{x}_{\rm new}) \supset J$ that $\boldsymbol{x}_{\rm new}$ is locally optimal, satisfies (5.13) and satisfies (5.14) with strict inequality.

If $V(J, \boldsymbol{x}_{\rm new}) \neq \emptyset$, the largest value $t \in [0, 1]$ such that $(1 - t)\boldsymbol{x} + t\boldsymbol{x}_{\rm new} \in \mathcal{K}$ is given by

$$t_o(J, \boldsymbol{x}_{\rm new}) \;=\; \min_{i \in V(J, \boldsymbol{x}_{\rm new})} \frac{-\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{x}_{\rm new}, \boldsymbol{v}_i \rangle - \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle} \;\in\; (0, 1).$$

If we set $t_o := t_o(J, \boldsymbol{x}_{\rm new})$ and replace $\boldsymbol{x}_{\rm new}$ with $(1 - t_o)\boldsymbol{x} + t_o \boldsymbol{x}_{\rm new}$, then the new point $\boldsymbol{x}_{\rm new}$ belongs to $\mathcal{K}$, satisfies (5.13) with strict inclusion and (5.14) with strict inequality. But it may still fail to be locally optimal. Thus we replace $\boldsymbol{x}$ with $\boldsymbol{x}_{\rm new}$ and repeat the previous steps.

After a finite number of iterations, the point $\boldsymbol{x}_{\rm new}$ will be locally optimal and satisfy the requirements (5.13) and (5.14). Indeed, whenever we replace $\boldsymbol{x}$ with a new vector, the value $f(\boldsymbol{x})$ decreases strictly, the set $J(\boldsymbol{x})$ increases strictly, so eventually the point $\boldsymbol{x}_{\rm new} = \boldsymbol{x}_{J(\boldsymbol{x})}$ will have to satisfy $V(\boldsymbol{x}, \boldsymbol{x}_{\rm new}) = \emptyset$.

Table 5.1 provides pseudo-code for the whole procedure. Checking the inequality $F_{\rm new} < F$ in line (*) is superflous in theory, but in practice it could be violated due to rounding errors. In that case we conclude that $\boldsymbol{x}$ is already locally optimal up to rounding errors. Whenever one determines the set $J = J(\boldsymbol{x})$ or the set $V(J, \boldsymbol{x}_{\rm new})$ in lines (**), one should be wary or numerical errors and replace conditions such as $\pm\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle > 0$ with $\pm\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle > \tau$ with a very small parameter $\tau > 0$. Without such precautions, there is a risk of creating an endless loop.

$(\boldsymbol{x}, F) \leftarrow$ **ActiveSet**

$\boldsymbol{x} \leftarrow \boldsymbol{x}_{\{1,\ldots,q\}}$
$F \leftarrow f(\boldsymbol{x})$
$\gamma \leftarrow \max_{j\in\{1,\ldots,q\}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_j \rangle$
if $\gamma \leq 0$ **then**
  return $\boldsymbol{x}$
end if
while $\gamma > 0$ **do**
  $(\boldsymbol{x}_{\rm new}, F_{\rm new}) \leftarrow$ **BasicProcedure1**$(\boldsymbol{x})$
  if $F_{\rm new} \geq F$ **then**
  $\gamma \leftarrow 0$
  else
  $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\rm new}$
  $F \leftarrow F_{\rm new}$
  $(\boldsymbol{x}, F) \leftarrow$ **BasicProcedure2**$(\boldsymbol{x})$
  $\gamma \leftarrow \max_{j\in J(\boldsymbol{x})} \langle \nabla f(\boldsymbol{x}), \boldsymbol{b}_j \rangle$
  end if
end while.

Table 5.2: Active set algorithm for quadratic f.

Complete algorithm. Table 5.2 describes the whole algorithm as described here.

Example: Concave regression

Suppose that for given data vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$, we are looking for a concave function $\widehat{\mu} : \mathbb{R} \to \mathbb{R}$ such that

$$\sum_{\ell=1}^{n} (Y_\ell - \widehat{\mu}(X_\ell))^2$$

is minimimal. The minimizer $\widehat{\mu}$ is not unique, but it is unique on the set $\{X_1, \ldots, X_n\}$. Precisely, let

$$t_1 < t_2 < \cdots < t_d$$

be the different elements of $\{X_1, \ldots, X_n\}$, and let

$$w_i \ := \ \#\{\ell \leq n : X_\ell = x_i\}.$$

Then for any function $\widehat{\mu} : \mathbb{R} \to \mathbb{R}$,

$$\sum_{\ell=1}^{n} (Y_\ell - \widehat{\mu}(X_\ell))^2 \ = \ S_0^2 + \sum_{i=1}^{d} w_i (y_i - \widehat{\mu}(t_i))^2,$$

where

$$S_0^2 \ := \ \sum_{i=1}^{d} \sum_{\ell : X_\ell = x_i} (Y_\ell - y_i)^2 \ = \ \|\boldsymbol{Y}\|^2 - \sum_{i=1}^{d} w_i y_i^2.$$

Thus, we are looking for a vector $\boldsymbol{m} \in \mathbb{R}^d$, interpreted as $(\widehat{\mu}(t_i))_{i=1}^d$, with minimal value of

$$f(\boldsymbol{m}) \ := \ \sum_{i=1}^{d} w_i (y_i - m_i)^2$$

under the constraints that

$$\frac{m_k - m_{k-1}}{t_k - t_{k-1}} \;\geq\; \frac{m_{k+1} - m_k}{t_{k+1} - t_k} \quad \text{for } 1 < k < d. \tag{5.15}$$

A possible extension of $\boldsymbol{m}$ to a concave function $\widehat{\mu}$ on $\mathbb{R}$ is to require that $\widehat{\mu}$ is affine on each interval $[t_i, t_{i+1}]$, $1 \leq i < d$, and on $(-\infty, t_2]$ as well as on $[t_{d-1}, \infty)$. (It is tacitly assumed that $d \geq 3$.)

In this particular example it is convenient to label the constraint vectors for $\mathcal{K}$ with indices $k \in \{2, \ldots, d-1\}$. Precisely, let

$$\begin{aligned}
\boldsymbol{v}_k \;&:=\; \Big(\frac{1_{[i=k+1]} - 1_{[i=k]}}{t_{k+1} - t_k} - \frac{1_{[i=k]} - 1_{[i=k-1]}}{t_k - t_{k-1}} \Big)_{i=1}^{d} \\
&=\; \Big(\frac{(t_{k+1} - t_k) 1_{[i=k-1]} - (t_{k+1} - t_{k-1}) 1_{[i=k]} + (t_k - t_{k-1}) 1_{[i=k+1]}}{(t_{k+1} - t_k)(t_k - t_{k-1})} \Big)_{i=1}^{d}
\end{aligned}$$

for $2 \leq k \leq d-1$. Then (5.15) is equivalent to

$$\langle \boldsymbol{m}, \boldsymbol{v}_k \rangle \;\leq\; 0 \quad \text{for } 2 \leq k \leq d-1.$$

A suitable dual basis is given, for instance, by the vectors

$$\begin{aligned}
\boldsymbol{b}_1 \;&:=\; (1)_{i=1}^{d}, \\
\boldsymbol{b}_k \;&:=\; \big((t_k - t_i)^+ \big)_{i=1}^{d}, \quad 2 \leq k \leq d.
\end{aligned}$$

For a set $J \subset \{2, \ldots, d-1\}$, the set

$$\mathcal{X}_J \;=\; \big\{ \boldsymbol{m} \in \mathbb{R}^d : \langle \boldsymbol{m}, \boldsymbol{v}_k \rangle = 0 \text{ for all } k \in J \big\}$$

corresponds to the set of all continuous and piecewise linear functions $\mu : [t_1, t_d] \to \mathbb{R}$ with potential changes of slope only at the points t_i, $i \in \{1, \ldots, d\} \setminus J$. These points will be referred to as *knots*.

Concerning the set $\tilde{J}(\boldsymbol{m})$ for basic procedure 1, in our particular implementation, we go through all pairs of consecutive knots x_k and x_ℓ of $\boldsymbol{m}$ such that $k + 1 < \ell$. Then we choose an index

$$i(k, \ell) \;\in\; \operatorname*{arg\,max}_{i \in \{k+1, \ldots, \ell-1\}} \langle \nabla f(\boldsymbol{m}), \boldsymbol{b}_i \rangle.$$

If $\langle \nabla f(\boldsymbol{m}), \boldsymbol{b}_{i(k,\ell)} \rangle$ is larger than 0.5 times $\max_{i=1,\ldots,d} \langle \nabla f(\boldsymbol{m}), \boldsymbol{b}_i \rangle$, then $i(k, \ell) \in \tilde{J}(\boldsymbol{m})$.

One important fact is that we don't have to generate and store the vectors $\boldsymbol{v}_k$ and $\boldsymbol{b}_i$. Indeed, fitting a continuous and piecewise linear function with specified knots (a linear spline) is a standard problem from linear regression. Moreover, for

$$\boldsymbol{g} \;:=\; \nabla f(\boldsymbol{m}) \;=\; -2\big(w_i (y_i - m_i) \big)_{i=1}^{d}$$

and $k \in \{2, \ldots, d-1\}$ one may write

$$\begin{aligned}\langle \boldsymbol{g}, \boldsymbol{b}_k \rangle &= \sum_{i=1}^{d} g_i (t_k - t_i)^+ \\ &= \sum_{i=1}^{k-1} g_i \sum_{j=i}^{k-1} (t_{j+1} - t_j) \\ &= \sum_{j=1}^{k-1} (t_{j+1} - t_j) \sum_{i=1}^{j} g_i.\end{aligned}$$

Hence one may compute all directional derivatives $\langle \boldsymbol{g}, \boldsymbol{b}_k \rangle$, $2 \le k < d$, by a simple summation scheme.

Figures 5.2 and 5.3 illustrate this procedure with a data example involving $n = 1000$ raw observations (X_ℓ, Y_ℓ), leading to $d = 101$ triples (t_i, y_i, w_i). The raw observations (X_ℓ, Y_ℓ) and pairs (t_i, y_i) are shown in the upper panel of Figure 5.2. The lower panel shows the final fit. Figure 5.3 shows intermediate stages of the active set algorithm. The current fit is shown as a blue line with the knots (changes of slope) highlighted. On sees the current fit after one, three, five and seven iterations (i.e. basic procedure 1, followed by basic procedure 2). That is, the current fit is locally optimal. The vertical lines (magenta) indicate positions at which the currently active constraint will be deactivated, leading to a new knot.

Active set algorithms for general f

For the general setting, suppose that we work with quasi-Newton steps for the miminization of f over $\mathcal{X}_J$. That is, in addition to f and ∇f there is a continuous mapping $A : \mathcal{X} \to \mathbb{R}^{d \times d}_{\mathrm{sym},+}$, and for a candidate $\boldsymbol{x}_o \in \mathcal{K} \cap \mathcal{X}$ for x_*, the function $f(\boldsymbol{x})$ is approximated temporarily by the quadratic function

$$\tilde{f}(\boldsymbol{x} \,|\, \boldsymbol{x}_o) := f(\boldsymbol{x}_o) + \nabla f(\boldsymbol{x}_o)^\top (\boldsymbol{x} - \boldsymbol{x}_o) + 2^{-1} (\boldsymbol{x} - \boldsymbol{x}_o)^\top A(\boldsymbol{x}_o)(\boldsymbol{x} - \boldsymbol{x}_o).$$

Note that $\boldsymbol{x}_o = \boldsymbol{x}_*$ if and only if $\boldsymbol{x}_o$ minimizes $\tilde{f}(\cdot \,|\, \boldsymbol{x}_o)$ over $\mathcal{K}$.

One strategy, used by numerous authors, is to apply an active set algorithm for quadratic functions to this surrogate function $\tilde{f}(\cdot \,|\, \boldsymbol{x}_o)$ to obtain a new candidate $\boldsymbol{\psi}_o(\boldsymbol{x}_o) \in \mathcal{K}$ for $\boldsymbol{x}_*$. Then one applies some stepsize correction and determines a point $\boldsymbol{\psi}(\boldsymbol{x}_o) = (1 - \lambda(\boldsymbol{x}_o))\boldsymbol{x}_o + \lambda(\boldsymbol{x}_o)\boldsymbol{\psi}_o(\boldsymbol{x}_o)$ to make sure that this point does belong to $\mathcal{K} \cap \mathcal{X}$ and yields a strictly smaller value of f, unless $\boldsymbol{x}_o$ was already the minimizer $\boldsymbol{x}_*$.

5.2.3 Isotonic least squares regression

The present and subsequent subsection is about the special cone

$$\mathcal{K}_\uparrow := \big\{ \boldsymbol{m} \in \mathbb{R}^d : m_1 \le m_2 \le \cdots \le m_d \big\}.$$

At first we'll discuss a particular optimization problem on this set $\mathcal{K}_\uparrow$ and describe geometrical properties of the solution. In the next subsection we consider a rather general optimization problem

and provide an exact algorithm for its solution which may be viewed as a special active set method. All this material is based on ideas of Robertson and Waltman (1968) and Robertson et al. (1988).

For a fixed data vector $\boldsymbol{y} \in \mathbb{R}^d$ and a weight vector $\boldsymbol{w} \in (0,\infty)^d$ we consider the function

$$f(\boldsymbol{m}) \;:=\; \sum_{i=1}^{d} w_i(y_i - m_i)^2.$$

Obviously this quadratic function is strictly convex and coercive on $\mathbb{R}^d$, i.e. $f(\boldsymbol{m}) \to \infty$ as $\|\boldsymbol{m}\| \to \infty$. Consequently, $\boldsymbol{x}_* = \arg\min_{\boldsymbol{m}\in\mathcal{K}_\uparrow} f(\boldsymbol{m})$ exists. Here

$$\nabla f(\boldsymbol{m}) \;=\; 2\big(w_i(m_i - y_i)\big)_{i=1}^{d},$$

so it follows from (5.10) that the minimizer of f over $\mathcal{K}_\uparrow$ is the unique point $\boldsymbol{m} \in \mathcal{K}_\uparrow$ such that

$$\sum_{i=1}^{d} w_i(m_i - y_i)\Delta_i \;\geq\; 0 \quad \text{whenever } \Delta \in \mathbb{R}^d \text{ such that } \boldsymbol{m} + t\Delta \in \mathcal{K}_\uparrow \text{ for some } t > 0. \tag{5.16}$$

Plugging in suitable test vectors Δ in the previous inequality yieds the following characterization of $\boldsymbol{x}_*$.

Theorem 5.23. *A vector $\boldsymbol{m} \in \mathcal{K}_\uparrow$ minimizes the function f above over $\mathcal{K}_\uparrow$ if and only if for arbitrary $j \in \{1, 2, \ldots, d\}$,*

$$\sum_{i=1}^{j} w_i m_i \;\leq\; \sum_{i=1}^{j} w_i y_i$$

with equality if $j = d$ or $m_j < m_{j+1}$.

Remark 5.24 (Convex minorants). Before proving the latter theorem, let us interpret it geomatrically. An arbitrary vector $\boldsymbol{v} \in \mathbb{R}^d$ may be identified with the picewise linear path $\mathcal{S}(\boldsymbol{v}) \subset \mathbb{R} \times \mathbb{R}$, connecting the $d+1$ points $(0,0)$ and

$$\Big(\sum_{i=1}^{j} w_i, \sum_{i=1}^{j} w_i v_i\Big), \quad 1 \leq j \leq d.$$

A vector $\boldsymbol{m}$ belongs to $\mathcal{K}_\uparrow$ if and only if its path $\mathcal{S}(\boldsymbol{m})$ is the graph of a convex function on the interval $\big[0, \sum_{i=1}^{f} w_i\big]$. According to Theorem 5.23, this path has to be below the path $\mathcal{S}(\boldsymbol{y})$ with the same starting point and end point. Moreover, $m_j < m_{j+1}$ is equivalent to saying that the path $\mathcal{S}(\boldsymbol{m})$ changes slope at the point $\big(\sum_{i=1}^{j} w_i, \sum_{i=1}^{j} w_i m_i\big)$. Consequently, at any such point the path $\mathcal{S}(\boldsymbol{m})$ has to touch the path $\mathcal{S}(\boldsymbol{y})$. Now one can verify easily that this amounts to the following description: The path $\mathcal{S}(\boldsymbol{m})$ corresponds to the largest (pointwise) convex function on the interval $\big[0, \sum_{i=1}^{f} w_i\big]$ which is less than or equal to the function corresponding to the path $\mathcal{S}(\boldsymbol{y})$.

This fact is illustrated in Figure 5.4. In the upper panel one sees a scatter plot of $d = 40$ pairs $(t_i, y_i) = (i, y_i)$ as well as a line plot of the optimal vector $\boldsymbol{m} \in \mathcal{K}_\uparrow$, where we chose $w_i = 1$. In the lower panel one sees the corresponding paths $\mathcal{S}(\boldsymbol{y})$ and $\mathcal{S}(\boldsymbol{m})$.

Proof of Theorem 5.23. Let $\boldsymbol{m}$ be the minimizer of f on $\mathcal{K}_\uparrow$. Clearly the vector $\Delta := \pm(1)_{i=1}^d$ satisfies the condition in (5.16). Hence $\pm\sum_{i=1}^d w_i(m_i - y_i) \geq 0$ which is equivalent to

$$\sum_{i=1}^d w_i m_i \;=\; \sum_{i=1}^d w_i y_i.$$

For $1 \leq j < d$ we consider the vector $\Delta := (-1_{[i\leq j]})_{i=1}^d$. It satisfies the condition in (5.16), too, whence $-\sum_{i\leq j} w_i(m_i - y_i) \geq 0$ or, equivalently,

$$\sum_{i=1}^j w_i m_i \;\leq\; \sum_{i=1}^j w_i y_i.$$

In case of $m_j < m_{j+1}$ the vector $\Delta := (1_{[i\leq j]})_{i=1}^d$ fulfils the same requirement, whence

$$\sum_{i=1}^j w_i m_i \;=\; \sum_{i=1}^j w_i y_i.$$

For the reverse direction, let $\boldsymbol{m} \in \mathcal{K}_\uparrow$ such that $\sum_{i=1}^j w_i m_i \leq \sum_{i=1}^j w_i y_i$ for all $j \in \{1, \ldots, d\}$ with equality in case of $j = d$ or $m_j < m_{j+1}$. Now let $\Delta \in \mathbb{R}^d$ be a vector such that $\boldsymbol{m} + t\Delta \in \mathcal{K}_\uparrow$ for some $t > 0$. This is equivalent to saying that

$$\Delta_j \leq \Delta_{j+1} \quad \text{if } 1 \leq j < d \text{ and } m_j = m_{j+1}. \tag{5.17}$$

Since $\sum_{i=1}^d w_i(m_i - y_i) = 0$, we may write

$$\begin{aligned}
\sum_{i=1}^d w_i(m_i - y_i)\Delta_i \;&=\; \sum_{i=1}^d w_i(m_i - y_i)(\Delta_i - \Delta_d) \\
&=\; \sum_{i=1}^{d-1} w_i(m_i - y_i) \sum_{j=i}^{d-1} (\Delta_j - \Delta_{j+1}) \\
&=\; \sum_{j=1}^{d-1} (\Delta_j - \Delta_{j+1}) \sum_{i=1}^{j} w_i(m_i - y_i) \\
&\geq\; 0.
\end{aligned}$$

In the latter step we used the assumptions on $\boldsymbol{m}$ and (5.17). □

5.2.4 The pool-adjacent-violators algorithm (PAVA)

Now we consider a generalization of the function f in the previous subsection, namely,

$$f(\boldsymbol{m}) \;:=\; \sum_{i=1}^d h_i(m_i)$$

with certain functions $h_i : \mathbb{R} \mapsto (-\infty, \infty]$. Explicit examples for the h_i, which are relevant in statistical applications, are:

- $h_i(r) := w_i(y_i - r)^2$,
- $h_i(r) := |y_i - r|$,
- $h_i(r) := w_i\big(y_i L(r) + (1-y_i)L(1-r)\big)$, where $y_i \in [0,1]$ and

$$L(r) \;:=\; \begin{cases} -\log(s) & \text{if } s \in (0,1], \\ \infty & \text{else.} \end{cases}$$

Assumption on the functions h_i. For any nonvoid index set $Q \subset \{1,2,\ldots,d\}$ let

$$h_Q(r) \;:=\; \sum_{i \in Q} h_i(r).$$

We assume that there exists a number $\xi_Q \in \mathbb{R}$ with the following properties: $h_Q(\xi_Q) < \infty$, and

$$h_Q \text{ is } \begin{cases} \text{isotonic on } [\xi_Q, \infty), \\ \text{antitonic on } (-\infty, \xi_Q]. \end{cases} \tag{5.18}$$

In the explicit examples mentioned before, this assumption is always satisfied:

- In case of $h_i(r) = w_i(y_i - r)^2$ or $h_i(r) := w_i\big(y_i L(r) + (1-y_i)L(1-r)\big)$,

$$\xi_Q \;=\; \sum_{i \in Q} w_i y_i \Big/ \sum_{i \in Q} w_i.$$

- In case of $h_i(r) = |y_i - r|$ we may choose

$$\xi_Q \;\in\; \mathrm{Median}(y_i : i \in Q).$$

Abstract description of PAVA. In principle the algorithm operates on partitions $\mathcal{P}$ of the index set $\{1,2,\ldots,d\}$ into blocks, that is, into sets of the form $\{a,\ldots,b\}$ with integers $1 \le a \le b \le d$. For each such partition $\mathcal{P}$ we define the vector

$$\boldsymbol{m}_{\mathcal{P}} \;:=\; \Big(\sum_{P \in \mathcal{P}} 1_{[i \in P]} \xi_P\Big)_{i=1}^{d} \;\in\; \operatorname*{arg\,min}_{\boldsymbol{m} \in \mathcal{X}_{\mathcal{P}}} f(\boldsymbol{m}),$$

where

$$\mathcal{X}_{\mathcal{P}} \;:=\; \Big\{ \Big(\sum_{P \in \mathcal{P}} 1_{[i \in P]} v_P\Big)_{i=1}^{d} : v_P \in \mathbb{R} \text{ for } P \in \mathcal{P} \Big\}.$$

Initialization: We start with the finest partition

$$\mathcal{P} \;=\; \big\{\{1\},\{2\},\ldots,\{n\}\big\}.$$

Induction step: Suppose that the vector $\boldsymbol{m}_{\mathcal{P}}$ is not in $\mathcal{K}_{\uparrow}$. More generally, suppose that the partition $\mathcal{P}$ contains two neighboring blocks $Q = \{a,\ldots,b-1\}$ and $R = \{b,\ldots,c\}$ such that $\xi_Q > \xi_R$, that is, Q and R are two "adjacent violators". Then we replace $\mathcal{P}$ with a new partition by replacing Q and R with their union $Q \cup R = \{a,\ldots,c\}$ ("pool adjacent violators").

End: The induction step is repeated until finally $\boldsymbol{m}_{\mathcal{P}} \in \mathcal{K}_{\uparrow}$. As shown next, this vector $\boldsymbol{m}_{\mathcal{P}}$ is automatically a minimizer of f over $\mathcal{K}_{\uparrow}$.

Validity of PAVA. Since the cardinality of $\mathcal{P}$ decreases by one in each instance of the induction step, the algorithm will terminate after at most $d-1$ repetitions of the induction step. Obviously the final vector $\boldsymbol{m}_{\mathcal{P}}$ belongs to $\mathcal{K}_{\uparrow}$. That it minimizes f follows from the following observation:

Lemma 5.25. *Let $\mathcal{P}$ be a partition of $\{1,2,\dots,d\}$ into blocks, among them the two neighbors $Q=\{a,\dots,b-1\}$ and $R=\{b,\dots,c\}$. Further let $\widetilde{\mathcal{P}}$ be the partition resulting from pooling Q and R. Suppose that $\xi_Q \geq \xi_R$. Then for each vector $\boldsymbol{m} \in \mathcal{K}_{\uparrow} \cap \mathcal{X}_{\mathcal{P}}$ there exists a vector $\tilde{\boldsymbol{m}} \in \mathcal{K}_{\uparrow} \cap \mathcal{X}_{\widetilde{\mathcal{P}}}$ such that $f(\tilde{\boldsymbol{m}}) \leq f(\boldsymbol{m})$.*

This lemma implies that the final partition $\mathcal{P}$ produced by PAVA satisfies

$$\min_{\boldsymbol{m}\in\mathcal{K}_{\uparrow}} f(\boldsymbol{m}) \;=\; \min_{\boldsymbol{m}\in\mathcal{K}_{\uparrow}\cap\mathcal{X}_{\mathcal{P}}} f(\boldsymbol{m}) \;\geq\; \min_{\boldsymbol{m}\in\mathcal{X}_{\mathcal{P}}} f(\boldsymbol{m}) \;=\; f(\boldsymbol{m}_{\mathcal{P}}).$$

And since the vector $\boldsymbol{m}_{\mathcal{P}}$ belongs to $\mathcal{K}_{\uparrow}$, the previous inequality is even an equality, so $\boldsymbol{m}_{\mathcal{P}}$ minimizes f over $\mathcal{K}_{\uparrow}$.

Proof of Lemma 5.25. Let $\boldsymbol{m} \in \mathcal{K}_{\uparrow} \cap \mathcal{X}_{\mathcal{P}}$, that means,

$$\boldsymbol{m} \;=\; \Big(\sum_{P\in\mathcal{P}} 1_{[i\in P]} m_P\Big)_{i=1}^{d}$$

with certain numbers m_P, $P\in\mathcal{P}$ such that $m_P \leq m_{P'}$ whenever $\max(P) < \min(P')$. Then

$$f(\boldsymbol{m}) \;=\; \sum_{P\in\mathcal{P}} h_P(m_P).$$

In case of $\xi_Q > m_Q$ one could replace both numbers m_Q and m_R with the value

$$\min(\xi_Q, m_R) \;\in\; [m_Q, m_R]$$

without increasing $f(\boldsymbol{m})$, because h_Q is antitonic on $(-\infty,\xi_Q]$ while h_R is isotonic on $[\xi_R,\infty) \supset [\xi_Q,\infty)$.

In case of $\xi_Q \leq m_Q$ it follows from $\xi_Q \geq \xi_R$ that $\xi_R \leq m_Q$, so we could replace m_R with m_Q without increasing $f(\boldsymbol{m})$.

In both cases the resulting vector belongs to the smaller set $\mathcal{K}_{\uparrow} \cap \mathcal{X}_{\widetilde{\mathcal{P}}}$. □

Explicit implementations. In practice, it is even advisable to pool adjacent weak violators, that is, to pool two adjacent blocks $Q=\{a,\dots,b-1\}$ and $R=\{b,\dots,c\}$ in $\mathcal{P}$ whenever $\xi_Q \geq \xi_R$. This is justified by Lemma 5.25, too.

We introduce an auxiliary integer variable c running from 2 through d, and in each step we compute a vector in

$$\operatorname*{arg\,min}_{\boldsymbol{m}\in\mathbb{R}^c : m_1\leq\dots\leq m_c} \sum_{i=1}^{c} h_i(m_i).$$

To this end we use an additional tuple $\boldsymbol{\ell} = (\ell_0,\ell_1,\dots,\ell_k)$ of variable length $k+1$ with integer components $0=\ell_0<\ell_1<\ell_2<\dots<\ell_k=c$. This tuple corresponds to the partition

$$\mathcal{P} \;=\; \Big(\underbrace{\{1,\dots,\ell_1\}}_{=:P_1}, \underbrace{\{\ell_1+1,\dots,\ell_2\}}_{=:P_2}, \dots, \underbrace{\{\ell_{k-1}+1,\dots,c\}}_{=:P_k}, \{c+1\},\{c+2\},\dots,\{d\}\Big)$$

in the abstract definition. In addition we use the tuple $\boldsymbol{M} = (M_0, M_1, \ldots, M_k)$ with components $M_0 := -\infty$ and $M_j := \xi_{P_j}$ for $1 \le j \le k$. Then the algorithm works as in Table 5.3.

Algorithm $\boldsymbol{m} \leftarrow$ **PAVA**$(h_1, \ldots, h_d)$
$\boldsymbol{\ell} \leftarrow (0, 1)$
$\boldsymbol{M} \leftarrow (-\infty, \xi_{\{1\}})$
$k \leftarrow 1$
for $c \leftarrow 2$ **to** d **do**
  $\boldsymbol{\ell} \leftarrow (\boldsymbol{\ell}, c)$
  $\boldsymbol{M} \leftarrow (\boldsymbol{M}, \xi_{\{c\}})$
  $k \leftarrow k + 1$
  while $M_{k-1} \ge M_k$ **do**
    $\boldsymbol{\ell} \leftarrow (\ell_0, \ldots, \ell_{k-2}, c)$
    $\boldsymbol{M} \leftarrow \big(M_0, \ldots, M_{k-2}, \xi_{\{\ell_{k-2}+1, \ldots, c\}}\big)$
    $k \leftarrow k - 1$
  end while
end for
$\boldsymbol{m} \leftarrow (0)_{i=1}^{d}$
for $j \leftarrow 1$ **to** k **do**
  for $i \leftarrow \ell_{j-1} + 1$ **to** ℓ_j **do**
    $m_i \leftarrow M_j$
  end for
end for.

Table 5.3: Explicit version of PAVA.

• **The special case** $h_i(r) = w_i(y_i - r)^2$**.** Here one should use an additional tuple $\boldsymbol{W} = (W_1, \ldots, W_k)$ with the weights $W_j := \sum_{i \in P_j} w_i$. For then

$$\xi_{\{\ell_{k-2}+1, \ldots, c\}} \;=\; \frac{W_{k-1} M_{k-1} + W_k M_k}{W_{k-1} + W_k} \quad \text{and} \quad \sum_{i \in P_{k-1} \cup P_k} w_i \;=\; W_{k-1} + W_k.$$

Thus each instance of the while–loop in Table 5.3 requires only $O(1)$ operations. Consequently, the whole algorithm requires $O(d)$ memory and $O(d)$ operations.

• **The special case** $h_i(r) = |y_i - r|$**.** Here one should use an additional vector $\boldsymbol{z} = (z_i)_{i=1}^{d}$ which is equal to $\boldsymbol{y}$ initially. In general, $(z_i)_{i \in P_j}$ contains the components of $(y_i)_{i \in P_j}$ in non-decreasing order. If two blocks $P_{k-1} = \{\ell_{k-1}, \ldots, \ell_k - 1\}$ and $P_k = \{\ell_k, \ldots, c\}$ have to be pooled, the entries $z_{\ell_{k-1}}, \ldots, z_c$ need to be sorted which is possible within $O(d)$ steps (similarly as in "MergeSort"). All in all the whole algorithm requires $O(d)$ memory and $O(d^2)$ operations.

Refinement. Suppose that the local minimizers ξ_Q, $\emptyset \ne Q \subset \{1, \ldots, d\}$, are unique. This implies that for arbitrary disjoint and nonempty sets $Q, R \subset \{1, \ldots, d\}$,

$$\min(\xi_Q, \xi_R) \;\le\; \xi_{Q \cup R} \;\le\; \max(\xi_Q, \xi_R).$$

Consequently, instead of starting the PAVA (abstract description) with the finest possible partition $\mathcal{P} = \big(\{1\}, \{2\}, \ldots, \{d\}\big)$, one can start with a partition into maximal blocks on which $i \mapsto \xi_{\{i\}}$ is non-increasing.

For the special case of $h_i(r) = w_i(y_i - r)^2$, Table 5.4 contains pseudocode for a refined version of the PAVA. Table 5.5 illustrates this procedure for a vector $\boldsymbol{y} \in \mathbb{R}^{12}$ and $\boldsymbol{w} = \mathbf{1}_{12}$.

Algorithm $\boldsymbol{m} \leftarrow$ **PAVA**$(\boldsymbol{y}, \boldsymbol{w})$
$\boldsymbol{\ell} \leftarrow (0)$ **% initialize**
$\boldsymbol{W} \leftarrow (0)$
$\boldsymbol{M} \leftarrow (-\infty)$
$c \leftarrow 0$
$k \leftarrow 0$
while $c < d$ **do**
  $c \leftarrow c + 1$ **% add new block**
  $W_{\text{new}} \leftarrow w_c$
  $S_{\text{new}} \leftarrow w_c y_c$
  $c' \leftarrow c + 1$
  while $c' \leq d$ **and** $y_{c'} \leq y_c$ **do**
    $c \leftarrow c'$
    $W_{\text{new}} \leftarrow W_{\text{new}} + w_c$
    $S_{\text{new}} \leftarrow S_{\text{new}} + w_c y_c$
    $c' \leftarrow c + 1$
  end while
  $\boldsymbol{\ell} \leftarrow (\boldsymbol{\ell}, c)$
  $\boldsymbol{W} \leftarrow (\boldsymbol{W}, W_{\text{new}})$
  $\boldsymbol{M} \leftarrow (\boldsymbol{M}, S_{\text{new}}/W_{\text{new}})$
  $k \leftarrow k + 1$
  while $M_{k-1} \geq M_k$ **do**
    $\boldsymbol{\ell} \leftarrow (\ell_0, \ldots, \ell_{k-2}, c)$ **% pool adjacent**
    $W_{\text{new}} \leftarrow W_{k-1} + W_k$ **% weak violators**
    $\boldsymbol{M} \leftarrow \big(M_0, \ldots, M_{k-2}, (W_{k-1}M_{k-1} + W_k M_k)/W_{\text{new}}\big)$
    $\boldsymbol{W} \leftarrow \big(W_0, \ldots, W_{k-1}, W_{\text{new}}\big)$
    $k \leftarrow k - 1$
  end while
end for
$\boldsymbol{m} \leftarrow (0)_{i=1}^{d}$ **% generate** $\boldsymbol{m}$
for $j \in \{1, \ldots, k\}$ **do**
  for $i \in \{\ell_{j-1} + 1, \ldots, \ell_j\}$ **do**
    $m_i \leftarrow M_j$
  end for
end for.

Table 5.4: Explicit version of PAVA for isotonic weighted least squares.

The next three exercises show that the PAVA leads to an explicit nonparametric density estimator introduced by U. Grenander (1956). Instead of minimizing a convex function f over the cone $\mathcal{K}_\uparrow$ one has to use an appropriate variant of the PAVA for the cone $\mathbb{R}^n_\downarrow = \{\boldsymbol{m} \in \mathbb{R}^n : m_1 \geq \cdots \geq m_n\}$.

Exercise 5.26 (Silverman's trick). Let $\mathcal{F}$ be a family of probability densities on a measure space $(\mathcal{X}, \mathcal{B}, \mu)$, and let $\mathcal{P} := \{tf : t > 0, f \in \mathcal{F}\}$ be the cone generated by $\mathcal{F}$. Show that for given

	$\boldsymbol{y}^\top$	$=$		(	2	0	3	4	1	0	3	5	6	1	7	3	)
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2	)										
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2	)										
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1	)										
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2	3	)									
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2	1	)									
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1	3	)									
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2	3			6	)						
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2	1			3	)						
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1	3			$1.\overline{6}$	)						
pool	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	)						
adj.	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	)						
viol.	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	)						
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	7	)					
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	1	)					
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	3	)					
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	7	8	)				
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	1	1	)				
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	3	5	)				
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	7	8		10	)		
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	1	1		2	)		
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	3	5		3.5	)		
pool	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	7			10	)		
adj.	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	1			3	)		
viol.	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	3			4	)		
add	$\boldsymbol{\ell}$	$\leftarrow$	(	0		2				6	7			10		12	)
new	$\boldsymbol{W}$	$\leftarrow$	(	0		2				4	1			3		2	)
block	$\boldsymbol{M}$	$\leftarrow$	(	$-\infty$		1				2	3			4		5	)
	$\boldsymbol{m}^\top$	$=$		(	1	1	2	2	2	2	3	4	4	4	5	5	)

Table 5.5: Numerical example for the PAVA with $h_i(r) = (y_i - r)^2$.

$n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$,

$$\operatorname*{arg\,max}_{f \in \mathcal{F}} \ell(f) \;=\; \operatorname*{arg\,min}_{p \in \mathcal{P}} \Big(-\ell(p) + n \int p \, d\mu \Big)$$

where $\ell(p) := \sum_{i=1}^n \log p(x_i)$.

Exercise 5.27 (Grenander's estimator, I). Let $\mathcal{F}$ be the family of monotone decreasing probability densitites f on $(0, \infty)$ with respect to Lebesgue measure. Let $0 = x_0 < x_1 < x_2 < \cdots x_n$ be given numbers. Viewing $x_1, \ldots, x_n$ as the order statistics of n independent random variables with unknown density $f \in \mathcal{F}$, the nonparametric maximum likelihood estimator of f is defined as

$$\widehat{f} \;:=\; \operatorname*{arg\,max}_{f \in \mathcal{F}} \ell(f) \quad \text{with} \quad \ell(f) \;:=\; \sum_{i=1}^n \log f(x_i).$$

The goal of this and the next exercise is to show that this estimator is well-defined and can be computed by means of the pool-adjacent-violators algorithm.

(a) Show that for each $f \in \mathcal{F}$ there exists a density $\bar{f} \in \mathcal{F}$ such that

$$\bar{f} \;\equiv\; \begin{cases} \bar{f}(x_i) & \text{on} (x_{i-1}, x_i] \text{ for } 1 \le i \le n, \\ 0 & \text{on } (x_n, \infty), \end{cases}$$

and $\ell(\bar{f}) \ge \ell(f)$ with equality if and only if $f \equiv \bar{f}$.

(b) Show by means of the previous part and Exercise 5.26 that $\widehat{f}$ is uniquely determined by

$$\widehat{\boldsymbol{p}} \;:=\; \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathbb{R}^n_{\downarrow}} \sum_{i=1}^{n} h_i(p_i) \quad \text{with} \quad h_i(s) \;:=\; -\log(s) + n(x_i - x_{i-1})s.$$

Exercise 5.28 (Grenander's estimator, II). **(a)** Show that the vector $\widehat{\boldsymbol{p}}$ in Exercise 5.27 is identical with

$$\operatorname*{arg\,min}_{\boldsymbol{m} \in \mathbb{R}^n_{\downarrow}} \sum_{i=1}^{n} w_i (y_i - m_i)^2 \quad \text{with} \quad w_i := x_i - x_{i-1} \text{ and } y_i := (n w_i)^{-1}.$$

(b) Let $\widehat{F}_{\mathrm{emp}} : [0, \infty) \to [0, 1]$ be the empirical distribution function of the points $x_1, \ldots, x_n$, that means, $\widehat{F}_{\mathrm{emp}}(x) = i/n$ for $x_i \le x < x_{i+1}$, $0 \le i \le n$, with $x_{n+1} := \infty$. Further let $\widehat{F}(x) := \int_0^x \widehat{f}(t)\, dt$ with Grenander's estimator $\widehat{f}$. Verify that $\widehat{F}$ is the smallest concave function which is bounded from below by $\widehat{F}_{\mathrm{emp}}$.

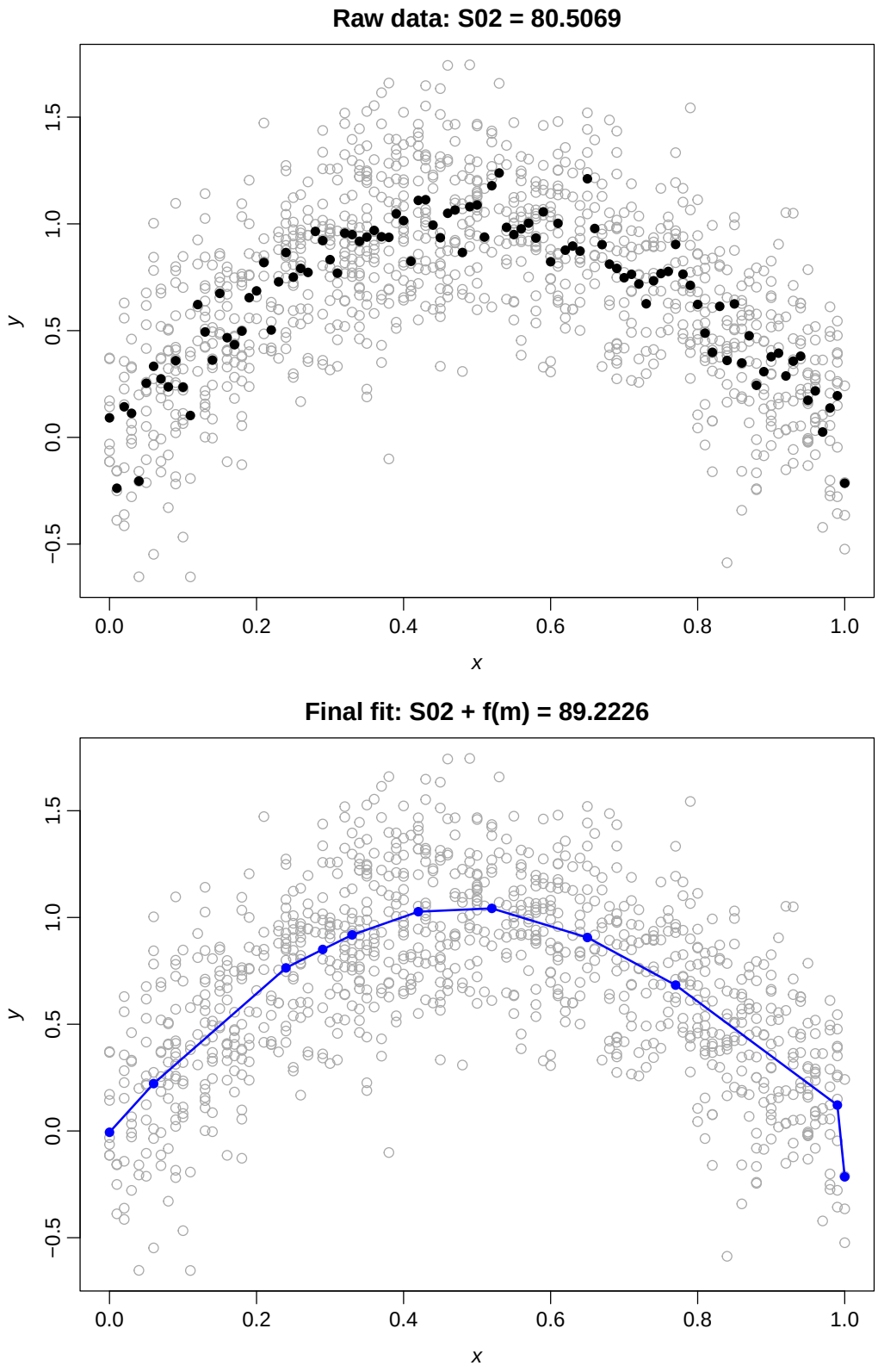

Figure 5.2: Example for concave regression.
Upper panel: data pairs (X_ℓ, Y_ℓ) ($\circ$) and preprocessed pairs (t_i, y_i) ($\bullet$).
Lower panel: data pairs and final fit.

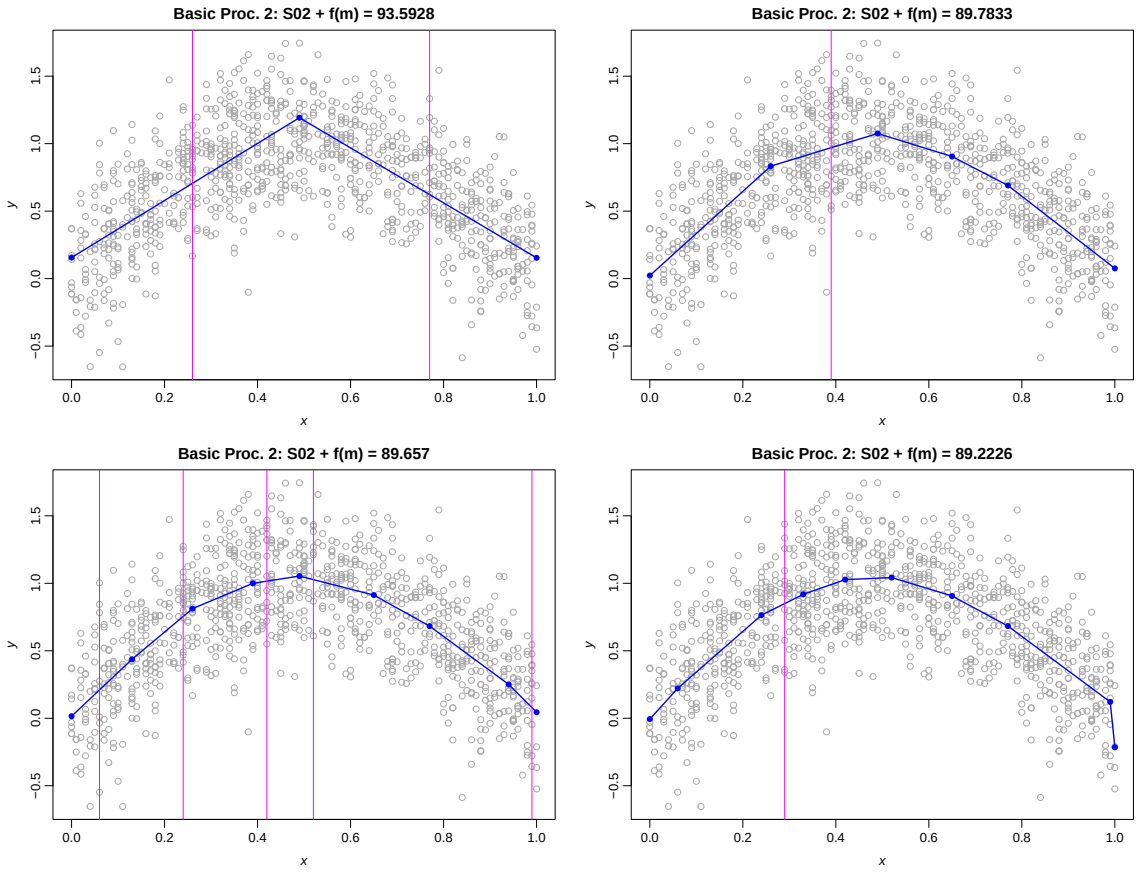

Figure 5.3: Example for concave regression: Some intermediate steps of the active set algorithm.

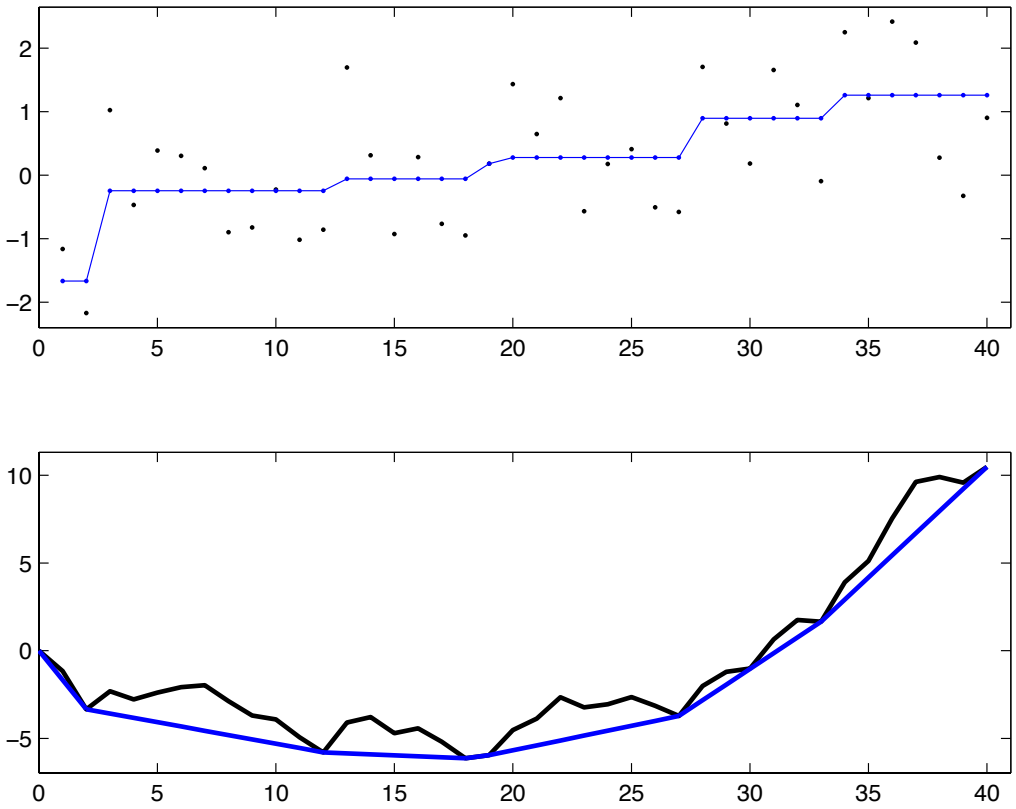

Figure 5.4: Isotonic regression and convex minorants.

Chapter 6

Conjugate Gradients

The main part of this chapter is based on the textbook of G. Opfer (1994).

6.1 The Task

Let $(\boldsymbol{V}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be a Euclidean space[1] with dimension d, and let $\boldsymbol{A} : \boldsymbol{V} \to \boldsymbol{V}$ be a self-adjoint, positive definite linear operator. That $\boldsymbol{A}$ is self-adjoint means that

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\rangle \;=\; \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle \quad \text{for arbitrary } \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{V}.$$

And a self-adjoint linear operator is called positive definite if

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle \;>\; 0 \quad \text{for arbitrary } \boldsymbol{x} \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}.$$

In particular, $\boldsymbol{A}$ is invertible. Now, for a given vector $\boldsymbol{b} \in \boldsymbol{V}$, our goal is to determine the vector

$$\boldsymbol{x}_* \;:=\; \boldsymbol{A}^{-1}\boldsymbol{b}.$$

This looks like a rather trivial task. If we choose an orthonormal basis $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ of $\boldsymbol{V}$, then $\boldsymbol{A}$ corresponds to a symmetric, positive definite matrix in $\mathbb{R}^{d\times d}$, $\boldsymbol{b}$ may be viewed as a vector in $\mathbb{R}^d$, and $\boldsymbol{x}_*$ is the solution of the linear equation system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The reason for considering an operator rather than a matrix is that in some applications it is easy to compute the point $\boldsymbol{A}\boldsymbol{x}$ for any $\boldsymbol{x} \in \boldsymbol{V}$, but the storage or inversion of the corresponding matrix $\boldsymbol{A}$ would require too much space or computation time. Then it is not obvious how to determine $\boldsymbol{x}_*$, and in what follows, a special iterative scheme will be derived.

Note first that the vector $\boldsymbol{x}_*$ is the unique minimizer of the function $f : \boldsymbol{V} \to \mathbb{R}$ with

$$f(\boldsymbol{x}) \;:=\; 2^{-1}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle.$$

Indeed, for arbitrary $\boldsymbol{x}, \boldsymbol{h} \in \boldsymbol{V}$,

$$f(\boldsymbol{x}+\boldsymbol{h}) \;=\; f(\boldsymbol{x}) - \langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{h}\rangle + 2^{-1}\langle \boldsymbol{h}, \boldsymbol{A}\boldsymbol{h}\rangle \tag{6.1}$$

[1] a finite-dimensional real Hilbert space

with the "residual"

$$\boldsymbol{r}(\boldsymbol{x}) \;:=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \;=\; \boldsymbol{A}(\boldsymbol{x}_* - \boldsymbol{x}).$$

This expansion shows that $\nabla f(\boldsymbol{x}) = -\boldsymbol{r}(\boldsymbol{x})$ and $D^2 f(\boldsymbol{x}) = \boldsymbol{A}$. In particular,

$$f(\boldsymbol{x}_* + \boldsymbol{h}) \;=\; f(\boldsymbol{x}_*) + 2^{-1}\langle \boldsymbol{h}, \boldsymbol{A}\boldsymbol{h}\rangle \;\geq\; f(\boldsymbol{x}_*)$$

with equality if and only if $\boldsymbol{h} = \boldsymbol{0}$. Precisely, for arbitrary $\boldsymbol{x} \in \boldsymbol{V}$,

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \;&=\; 2^{-1}\langle \boldsymbol{x} - \boldsymbol{x}_*, \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_*)\rangle \\ &=\; 2^{-1}\langle \boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{r}(\boldsymbol{x})\rangle \qquad (6.2) \\ &=\; 2^{-1}\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x} - \boldsymbol{x}_*)\big\|^2. \qquad (6.3) \end{aligned}$$

6.2 The Gradient Method

A first approach is to minimize f iteratively in the direction of its residual. That means, for a vector $\boldsymbol{x} \in \boldsymbol{V} \setminus \{\boldsymbol{x}_*\}$ we define

$$\boldsymbol{\psi}(\boldsymbol{x}) \;:=\; \boldsymbol{x} + t(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x})$$

with

$$t(\boldsymbol{x}) \;:=\; \operatorname*{arg\,min}_{t\in\mathbb{R}} \; f(\boldsymbol{x} + t\boldsymbol{r}(\boldsymbol{x})).$$

Starting from an arbitrary point $\boldsymbol{x}_0 \in \boldsymbol{V}$, we define the sequence $(\boldsymbol{x}_k)_{k\geq 0}$ inductively via $\boldsymbol{x}_{k+1} := \boldsymbol{\psi}(\boldsymbol{x}_k)$, where $\boldsymbol{\psi}(\boldsymbol{x}_*) := \boldsymbol{x}_*$. As shown later, this sequence will always converge to the minimizer $\boldsymbol{x}_*$.

Since

$$f(\boldsymbol{x} + t\boldsymbol{r}(\boldsymbol{x})) \;=\; f(\boldsymbol{x}) - t\|\boldsymbol{r}(\boldsymbol{x})\|^2 + 2^{-1}t^2\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle,$$

the number $t(\boldsymbol{x})$ is given by

$$t(\boldsymbol{x}) \;=\; \frac{\|\boldsymbol{r}(\boldsymbol{x})\|^2}{\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle},$$

and

$$f(\boldsymbol{\psi}(\boldsymbol{x})) \;=\; f(\boldsymbol{x}) - \frac{\|\boldsymbol{r}(\boldsymbol{x})\|^4}{2\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle}.$$

Combining this equation and (6.2) leads to

$$\frac{f(\boldsymbol{\psi}(\boldsymbol{x})) - f(\boldsymbol{x}_*)}{f(\boldsymbol{x}) - f(\boldsymbol{x}_*)} \;=\; 1 - \frac{\|\boldsymbol{r}(\boldsymbol{x})\|^4}{\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle\,\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{x})\rangle}.$$

The ratio on the right hand side may be bounded as follows.

Lemma 6.1. *Let $\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^d \lambda_i\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle\boldsymbol{u}_i$ for $\boldsymbol{x} \in \boldsymbol{V}$ with an orthonormal basis $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ of $\boldsymbol{V}$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. For arbitrary $\boldsymbol{y} \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}$,*

$$1 \;\leq\; \frac{\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{y}\rangle\,\langle \boldsymbol{y}, \boldsymbol{A}^{-1}\boldsymbol{y}\rangle}{\|\boldsymbol{y}\|^4} \;\leq\; \frac{(\lambda_1 + \lambda_d)^2}{4\lambda_1\lambda_d}.$$

The second last inequality is an equality if and only if $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{A}$. The last inequality is an equality if $\boldsymbol{y}$ is a multiple of $\boldsymbol{u}_1 + \boldsymbol{u}_d$ or $\boldsymbol{u}_1 - \boldsymbol{u}_d$.

For the algorithmic mapping $\boldsymbol{\psi}$, this lemma leads to the following inequality:

Theorem 6.2. *With the same notation as in Lemma 6.1,*

$$\frac{f(\boldsymbol{\psi}(\boldsymbol{x})) - f(\boldsymbol{x}_*)}{f(\boldsymbol{x}) - f(\boldsymbol{x}_*)} \;\leq\; \Big(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\Big)^2$$

for arbitrary $\boldsymbol{x} \in \boldsymbol{V} \setminus \{\boldsymbol{x}_*\}$. Equality holds if $\boldsymbol{r}(\boldsymbol{x})$ is a multiple of $\boldsymbol{u}_1 + \boldsymbol{u}_d$ or $\boldsymbol{u}_1 - \boldsymbol{u}_d$. In that case, $\boldsymbol{r}(\boldsymbol{\psi}(\boldsymbol{x}))$ is a multiple of $\boldsymbol{u}_1 - \boldsymbol{u}_d$ or $\boldsymbol{u}_1 + \boldsymbol{u}_d$, respectively.

This theorem implies that the sequence $(\boldsymbol{x}_k)_{k\geq 0}$ just introduced converges to $\boldsymbol{x}_*$ for any starting point $\boldsymbol{x}_0 \in \boldsymbol{V}$. More precisely, the inequality is equivalent to

$$\frac{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{x}_*)\big\|}{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x} - \boldsymbol{x}_*)\big\|} \;\leq\; \frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d} \;=\; \frac{1 - \lambda_d/\lambda_1}{1 + \lambda_d/\lambda_1} \;=\; 1 - \frac{2}{\lambda_1/\lambda_d + 1}$$

by virtue of (6.3). Hence,

$$\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_k - \boldsymbol{x}_*)\big\| \;\leq\; \Big(\frac{1 - \lambda_d/\lambda_1}{1 + \lambda_d/\lambda_1}\Big)^k \big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_0 - \boldsymbol{x}_*)\big\|$$

for arbitrary $k \geq 0$ with equality in case of $\boldsymbol{r}(\boldsymbol{x}_0)$ being a multiple of $\boldsymbol{u}_1 \pm \boldsymbol{u}_d$. However, the speed of convergence can be rather low whenever the condition number λ_d/λ_1 of the operator $\boldsymbol{A}$ is close to zero.

Figure 6.1 illustrates the gradient method in case of $\boldsymbol{V} = \mathbb{R}^2$ with the usual inner product, $\boldsymbol{A} = \operatorname{diag}(1, 3)$, $\boldsymbol{b} = (1, 1)^\top$ whence $\boldsymbol{x}_* = (5, 5/3)^\top$, and $\boldsymbol{x}_0 = \boldsymbol{0}$. Here $\boldsymbol{r}(\boldsymbol{x}_k)$ is always a positive multiple of $(1, 1)^\top$ or $(1, -1)^\top$, and $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)$ decreases in each step by the same factor $\big((3-1)/(3+1)\big)^2 = 1/4$.

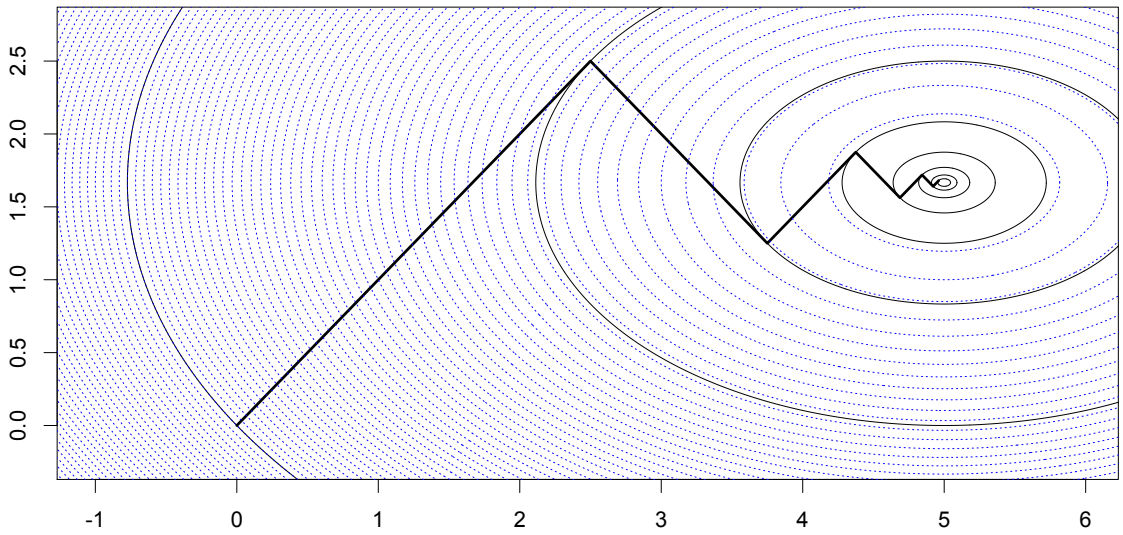

Figure 6.1: The gradient method.

Proof of Lemma 6.1. The asserted inequalities are equivalent to the following ones: For arbitrary vectors $\boldsymbol{p} \in [0, 1]^d$ with $\sum_{i=1}^d p_i = 1$,

$$1 \;\leq\; \sum_{i=1}^d p_i \lambda_i \sum_{j=1}^d p_j \lambda_j^{-1} \;\leq\; \frac{(\lambda_1 + \lambda_d)^2}{4\lambda_1 \lambda_d}.$$

Indeed, if we write $\boldsymbol{y} = \sum_{i=1}^d a_i \boldsymbol{u}_i$ with $\boldsymbol{a} \in \mathbb{R}^d$, then

$$\|\boldsymbol{y}\|^{-2} \langle \boldsymbol{y}, \boldsymbol{A}^s \boldsymbol{y} \rangle \;=\; \sum_{i=1}^d p_i \lambda_i^s \quad \text{with } p_i := \|\boldsymbol{y}\|^{-2} a_i^2.$$

Furthermore, any probability vector $\boldsymbol{p}$ may be obtained in this fashion from some vector $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

In the center of the new inequalities we see $\mathbb{E}(X)\,\mathbb{E}(X^{-1})$ with a random variable $X > 0$ having distribution $\sum_{i=1}^d p_i \delta_{\lambda_i}$. Since the function $x \mapsto x^{-1}$ is strictly convex on $(0, \infty)$, Jensen's inequality implies that $\mathbb{E}(X^{-1}) \geq \mathbb{E}(X)^{-1}$. Equality holds if and only if X has a degenerate distribution. In the original inequality, this means that $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{A}$.

On the other hand, $\lambda_d \leq X \leq \lambda_1$, so convexity of $x \mapsto x^{-1}$ yields the following inequality:

$$X^{-1} \;\leq\; \frac{\lambda_1 - X}{\lambda_1 - \lambda_d}\, \lambda_d^{-1} + \frac{X - \lambda_d}{\lambda_1 - \lambda_d}\, \lambda_1^{-1}.$$

Consequently,

$$\begin{aligned}
\mathbb{E}(X)\,\mathbb{E}(X^{-1}) \;&\leq\; \mathbb{E}(X)\Big(\frac{\lambda_1 - \mathbb{E}(X)}{\lambda_1 - \lambda_d}\, \lambda_d^{-1} + \frac{\mathbb{E}(X) - \lambda_d}{\lambda_1 - \lambda_d}\, \lambda_1^{-1}\Big) \\
&=\; \frac{\mathbb{E}(X)}{(\lambda_1 - \lambda_d)\lambda_1\lambda_d}\big(\lambda_1^2 - \lambda_d^2 - (\lambda_1 - \lambda_d)\,\mathbb{E}(X)\big) \\
&=\; \frac{\mathbb{E}(X)(\lambda_1 + \lambda_d - \mathbb{E}(X))}{\lambda_1\lambda_d} \\
&\leq\; \frac{(\lambda_1 + \lambda_d)^2}{4\lambda_1\lambda_d}.
\end{aligned}$$

Equality holds if $p_1 = p_d = 1/2$, and in the original inequality, this means that $\boldsymbol{y}$ is proportional to $\boldsymbol{u}_1 \pm \boldsymbol{u}_d$. □

Proof of Theorem 6.2. According to Lemma 6.1,

$$\frac{f(\boldsymbol{\psi}(\boldsymbol{x})) - f(\boldsymbol{x}_*)}{f(\boldsymbol{x}) - f(\boldsymbol{x}_*)} \;=\; 1 - \frac{\|\boldsymbol{r}(\boldsymbol{x})\|^4}{\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle\, \langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{x})\rangle} \;\leq\; 1 - \frac{4\lambda_1\lambda_d}{(\lambda_1 + \lambda_d)^2} \;=\; \Big(\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\Big)^2.$$

In general,

$$\boldsymbol{x}_* - \boldsymbol{\psi}(\boldsymbol{x}) \;=\; \boldsymbol{x}_* - \boldsymbol{x} - t(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x}) \;=\; (\boldsymbol{A}^{-1} - t(\boldsymbol{x})\boldsymbol{I})\boldsymbol{r}(\boldsymbol{x}).$$

Consequently, if $\boldsymbol{r}(\boldsymbol{x}) = \gamma(\boldsymbol{u}_1 \pm \boldsymbol{u}_d)$ for some $\gamma \neq 0$, then

$$\begin{aligned}
\boldsymbol{r}(\boldsymbol{\psi}(\boldsymbol{x})) \;=\; \boldsymbol{A}(\boldsymbol{x}_* - \boldsymbol{\psi}(\boldsymbol{x})) \;&=\; (\boldsymbol{I} - t(\boldsymbol{x})\boldsymbol{A})\boldsymbol{r}(\boldsymbol{x}) \\
&=\; \Big(\boldsymbol{I} - \frac{\|\boldsymbol{r}(\boldsymbol{x})\|^2}{\langle \boldsymbol{r}(\boldsymbol{x}), \boldsymbol{A}\boldsymbol{r}(\boldsymbol{x})\rangle}\boldsymbol{A}\Big)\boldsymbol{r}(\boldsymbol{x}) \\
&=\; \gamma\Big(\boldsymbol{u}_1 \pm \boldsymbol{u}_d - \frac{2}{\lambda_1 + \lambda_d}(\lambda_1\boldsymbol{u}_1 \pm \lambda_d\boldsymbol{u}_d)\Big) \\
&=\; \gamma\Big(\frac{\lambda_d - \lambda_1}{\lambda_1 + \lambda_d}\,\boldsymbol{u}_1 \pm \frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\,\boldsymbol{u}_d\Big) \\
&=\; -\,\frac{\lambda_1 - \lambda_d}{\lambda_1 + \lambda_d}\,\gamma(\boldsymbol{u}_1 \mp \boldsymbol{u}_d).
\end{aligned}$$

□

6.3 Conjugate Directions

As we shall see soon, the minimization of f would be rather simple if we had a basis of $\boldsymbol{V}$ which is "conjugate with respect to $\boldsymbol{A}$".

Definition 6.3 (Conjugate vectors). Vectors $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_k \in \boldsymbol{V}$ are called *conjugate with respect to* $\boldsymbol{A}$ if

$$\langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_j \rangle \;=\; 0 \quad \text{for } 1 \le i < j \le k.$$

In other words, the vectors $\boldsymbol{A}^{1/2}\boldsymbol{h}_1, \boldsymbol{A}^{1/2}\boldsymbol{h}_2, \ldots, \boldsymbol{A}^{1/2}\boldsymbol{h}_k$ are pairwise orthogonal.

Lemma 6.4. *Let $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_\ell \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}$ be conjugate with respect to $\boldsymbol{A}$. For an arbitrary point $\boldsymbol{x}_0 \in \mathbb{R}^d$ and any integer $k \in \{1, 2, \ldots, \ell\}$ let*

$$\begin{aligned} \boldsymbol{V}_k \;&:=\; \operatorname{span}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k), \\ \boldsymbol{x}_k \;&:=\; \operatorname*{arg\,min}_{\boldsymbol{x} \in \boldsymbol{x}_0 + \boldsymbol{V}_k} f(\boldsymbol{x}). \end{aligned}$$

Then

$$\boldsymbol{x}_k \;=\; \boldsymbol{x}_0 + \sum_{i=1}^{k} t_i \boldsymbol{h}_i \;=\; \boldsymbol{x}_{k-1} + t_k \boldsymbol{h}_k$$

with

$$t_i \;:=\; \frac{\langle \boldsymbol{r}(\boldsymbol{x}_0), \boldsymbol{h}_i \rangle}{\langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_i \rangle} \;=\; \frac{\langle \boldsymbol{r}(\boldsymbol{x}_{i-1}), \boldsymbol{h}_i \rangle}{\langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_i \rangle}.$$

Moreover, $\boldsymbol{r}(\boldsymbol{x}_k) \perp \boldsymbol{V}_k$.

Proof of Lemma 6.4. For $\boldsymbol{t} \in \mathbb{R}^k$, it follows from (6.1) that

$$\begin{aligned} f\Big(\boldsymbol{x}_0 + \sum_{i=1}^{k} t_i \boldsymbol{h}_i\Big) \;&=\; f(\boldsymbol{x}_0) - \sum_{i=1}^{k} t_i \langle \boldsymbol{r}(\boldsymbol{x}_0), \boldsymbol{h}_i \rangle + 2^{-1} \sum_{i,j=1}^{k} t_i t_j \langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_j \rangle \\ &=\; f(\boldsymbol{x}_0) + \sum_{i=1}^{k} \big(2^{-1} t_i^2 \langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_i \rangle - t_i \langle \boldsymbol{r}(\boldsymbol{x}_0), \boldsymbol{h}_i \rangle\big). \end{aligned}$$

Obviously, this is minimal in $\boldsymbol{t}$ if and only if

$$t_i \;=\; \frac{\langle \boldsymbol{r}(\boldsymbol{x}_0), \boldsymbol{h}_i \rangle}{\langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_i \rangle}.$$

In case of $i \ge 2$,

$$\boldsymbol{r}(\boldsymbol{x}_{i-1}) \;=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{i-1} \;=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}_{i-1}) \;=\; \boldsymbol{r}(\boldsymbol{x}_0) + \boldsymbol{A}(\boldsymbol{x}_0 - \boldsymbol{x}_{i-1}),$$

and this implies that

$$\langle \boldsymbol{r}(\boldsymbol{x}_0), \boldsymbol{h}_i \rangle \;=\; \langle \boldsymbol{r}(\boldsymbol{x}_{i-1}), \boldsymbol{h}_i \rangle + \langle \boldsymbol{x}_{i-1} - \boldsymbol{x}_0, \boldsymbol{A}\boldsymbol{h}_i \rangle \;=\; \langle \boldsymbol{r}(\boldsymbol{x}_{i-1}), \boldsymbol{h}_i \rangle,$$

because $\boldsymbol{x}_{i-1} - \boldsymbol{x}_0 \in \operatorname{span}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{i-1})$ and $\langle \boldsymbol{h}_j, \boldsymbol{A}\boldsymbol{h}_i \rangle = 0$ for $1 \le j < i$. This proves the alternative expression for the optimal t_i.

That $\boldsymbol{r}(\boldsymbol{x}_k)$ is perpendicular to $\boldsymbol{V}_k$ follows from the fact that for arbitrary $\boldsymbol{h} \in \boldsymbol{V}_k$,

$$0 \;=\; \frac{d}{dt}\Big|_{t=0} f(\boldsymbol{x}_k + t\boldsymbol{h}) \;=\; -\langle \boldsymbol{r}(\boldsymbol{x}_k), \boldsymbol{h} \rangle.$$

□

A Gram–Schmidt type procedure. The preceding considerations suggest the following procedure: We start with a vector $\boldsymbol{x}_0 \in \boldsymbol{V}$ such that $\boldsymbol{r}(\boldsymbol{x}_0) \neq \boldsymbol{0}$ which is equivalent to $\boldsymbol{x}_0 \neq \boldsymbol{x}_*$. Then we set

$$\boldsymbol{h}_1 \;:=\; \boldsymbol{r}(\boldsymbol{x}_0).$$

Suppose that for some $k \in \mathbb{N}$ we have chosen conjugate vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k \in \boldsymbol{V} \setminus \{\boldsymbol{0}\}$ with respect to $\boldsymbol{A}$. Let

$$\boldsymbol{V}_k \;:=\; \operatorname{span}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k) \quad \text{and} \quad \boldsymbol{x}_k \;:=\; \operatorname*{arg\,min}_{\boldsymbol{x} \,\in\, \boldsymbol{x}_0 + \boldsymbol{V}_k} f(\boldsymbol{x}).$$

If $\boldsymbol{r}(\boldsymbol{x}_k) = \boldsymbol{0}$, we know that $\boldsymbol{x}_k = \boldsymbol{x}_*$. Otherwise we define

$$\boldsymbol{h}_{k+1} \;:=\; \boldsymbol{r}(\boldsymbol{x}_k) - \sum_{i=1}^{k} \beta_{k,i} \boldsymbol{h}_i$$

with certain coefficients $\beta_{k,i} \in \mathbb{R}$ yet to be specified. Since $\boldsymbol{r}(\boldsymbol{x}_k) \perp \boldsymbol{V}_k$, the vector $\boldsymbol{h}_{k+1}$ is nonzero for any choice of the $\beta_{k,i}$. The vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k+1}$ are conjugate with respect to $\boldsymbol{A}$ if and only if for $1 \leq j \leq k$,

$$0 \;=\; \langle \boldsymbol{h}_{k+1}, \boldsymbol{A}\boldsymbol{h}_j \rangle \;=\; \langle \boldsymbol{r}(\boldsymbol{x}_k), \boldsymbol{A}\boldsymbol{h}_j \rangle - \beta_{k,j} \langle \boldsymbol{h}_j, \boldsymbol{A}\boldsymbol{h}_j \rangle,$$

that means,

$$\beta_{k,j} \;=\; \frac{\langle \boldsymbol{r}(\boldsymbol{x}_k), \boldsymbol{A}\boldsymbol{h}_j \rangle}{\langle \boldsymbol{h}_j, \boldsymbol{A}\boldsymbol{h}_j \rangle}.$$

All in all, if we start from a point $\boldsymbol{x}_0 \neq \boldsymbol{x}_*$, we obtain linearly independent vectors $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k_o} \in \boldsymbol{V}$ and points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k_o} \in \boldsymbol{V}$ such that for $1 \leq k \leq k_o$,

$$\boldsymbol{V}_k := \operatorname{span}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k) \;=\; \operatorname{span}\big(\boldsymbol{r}(\boldsymbol{x}_0), \ldots, \boldsymbol{r}(\boldsymbol{x}_{k-1})\big) \quad \text{and} \quad \boldsymbol{x}_k \;=\; \operatorname*{arg\,min}_{\boldsymbol{x} \in \boldsymbol{x}_0 + \boldsymbol{V}_k} f(\boldsymbol{x}).$$

Moreover, for $1 \leq k < k_o$,

$$\boldsymbol{r}(\boldsymbol{x}_k) \;\perp\; \boldsymbol{V}_k \quad \text{and} \quad \boldsymbol{h}_{k+1} \begin{cases} \in\; \boldsymbol{r}(\boldsymbol{x}_k) + \boldsymbol{V}_k, \\ \perp\; \boldsymbol{A}\boldsymbol{V}_k, \end{cases}$$

while $\boldsymbol{r}(\boldsymbol{x}_{k_o}) = \boldsymbol{0}$ and

$$\boldsymbol{x}_{k_o} \;=\; \boldsymbol{x}_*.$$

6.4 The Conjugate Gradient (CG) Algorithm

At first glance, it looks as if we have solved the original problem already. But recall that we have situations in mind in which the storage of a matrix for $\boldsymbol{A}$ would be too involved. But the Gram–Schmidt type procedure just described would require to keep track of all vectors $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{k_o}$. Fortunately, as shown in the next theorem, all coefficients $\beta_{k,j}$ with $1 \leq j < k$ are zero, so we only need $\boldsymbol{h}_k$ and $\boldsymbol{r}(\boldsymbol{x}_k)$ to compute the next direction $\boldsymbol{h}_{k+1}$.

Theorem 6.5. *Starting from $\boldsymbol{x}_0 \neq \boldsymbol{x}_*$, let $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k_o}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k_o}$ be constructed as just described. For $1 \leq k \leq k_o$, the space $\boldsymbol{V}_k = \operatorname{span}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_k) = \operatorname{span}\big(\boldsymbol{r}(\boldsymbol{x}_0), \ldots, \boldsymbol{r}(\boldsymbol{x}_{k-1})\big)$ is equal to the so-called Krylow space*

$$\operatorname{span}(\boldsymbol{A}^{i-1}\boldsymbol{r}(\boldsymbol{x}_0) : 1 \leq i \leq k).$$

Moreover, for $1 \leq k \leq k_o$,

$$\boldsymbol{x}_k \;=\; \boldsymbol{x}_{k-1} + t_k \boldsymbol{h}_k \quad \textit{and} \quad \boldsymbol{r}(\boldsymbol{x}_k) \;=\; \boldsymbol{r}(\boldsymbol{x}_{k-1}) - t_k \boldsymbol{A}\boldsymbol{h}_k \quad \textit{with} \quad t_k \;:=\; \frac{\|\boldsymbol{r}(\boldsymbol{x}_{k-1})\|^2}{\langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle},$$

while in case of $k < k_o$,

$$\boldsymbol{h}_{k+1} \;=\; \boldsymbol{r}(\boldsymbol{x}_k) + \frac{\|\boldsymbol{r}(\boldsymbol{x}_k)\|^2}{\|\boldsymbol{r}(\boldsymbol{x}_{k-1})\|^2}\, \boldsymbol{h}_k.$$

This theorem shows that $\boldsymbol{h}_{k+1} = \boldsymbol{r}(\boldsymbol{x}_k) - \sum_{i=1}^{k} \beta_{k,i} \boldsymbol{h}_i$ is just a linear combination of $\boldsymbol{r}(\boldsymbol{x}_k)$ and $\boldsymbol{h}_k$. The explicit formulae in Theorem 6.5 yield a relatively simple algorithm for the computation (approximation) of $\boldsymbol{x}_*$:

Algorithmus $\boldsymbol{x}_k \leftarrow \mathbf{CG}(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{x}_0, \epsilon)$

$\boldsymbol{r}_0 \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$
$\boldsymbol{h}_1 \leftarrow \boldsymbol{r}_0$
$k \leftarrow 0$

while $\|\boldsymbol{r}_k\| > \epsilon$ **do**
  $k \leftarrow k + 1$
  $t_k \leftarrow \|\boldsymbol{r}_{k-1}\|^2 / \langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle$
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + t_k \boldsymbol{h}_k$
  $\boldsymbol{r}_k \leftarrow \boldsymbol{r}_{k-1} - t_k \boldsymbol{A}\boldsymbol{h}_k$
  $\beta_k \leftarrow \|\boldsymbol{r}_k\|^2 / \|\boldsymbol{r}_{k-1}\|^2$
  $\boldsymbol{h}_{k+1} \leftarrow \boldsymbol{r}_k + \beta_k \boldsymbol{h}_k$
end while

After a finite number of steps, this algorithm yields a vector $\boldsymbol{x}_k$ such that $\|\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}\| \leq \epsilon$, where $\epsilon \geq 0$ is an arbitrary given number. Here is a different description:

Algorithmus $\boldsymbol{x} \leftarrow \mathbf{CG}(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{x}_0, \epsilon)$

$\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
$\boldsymbol{r} \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}$
$\gamma \leftarrow \|\boldsymbol{r}\|^2$
$\boldsymbol{h} \leftarrow \boldsymbol{r}$
while $\sqrt{\gamma} > \epsilon$ **do**
  $\boldsymbol{y} \leftarrow \boldsymbol{A}\boldsymbol{h}$
  $t \leftarrow \gamma / \langle \boldsymbol{h}, \boldsymbol{y} \rangle$
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + t\boldsymbol{h}$
  $\boldsymbol{r} \leftarrow \boldsymbol{r} - t\boldsymbol{y}$
  $\gamma_{\text{new}} \leftarrow \|\boldsymbol{r}\|^2$
  $\boldsymbol{h} \leftarrow \boldsymbol{r} + (\gamma_{\text{new}}/\gamma)\boldsymbol{h}$
  $\gamma \leftarrow \gamma_{\text{new}}$
end while

Proof of Theorem 6.5. We write $\boldsymbol{r}_\ell := \boldsymbol{r}(\boldsymbol{x}_\ell)$ and $\boldsymbol{W}_k := \mathrm{span}(\boldsymbol{A}^{i-1}\boldsymbol{r}_0 : 1 \le i \le k)$. Since $\dim(\boldsymbol{V}_k) = k$ and $\dim(\boldsymbol{W}_k) \le k$, it suffices to show that $\boldsymbol{V}_k \subset \boldsymbol{W}_k$ for $1 \le k \le k_o$. Obviously, $\boldsymbol{W}_1 = \boldsymbol{V}_1$. Suppose that $\boldsymbol{V}_k \subset \boldsymbol{W}_k$ for some $1 \le k < k_o$. Then there exist real coefficients $\lambda_1, \ldots, \lambda_k$ such that

$$\boldsymbol{x}_k \;=\; \boldsymbol{x}_0 + \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^{i-1}\boldsymbol{r}_0.$$

But then

$$\boldsymbol{r}_k \;=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k \;=\; \boldsymbol{r}_0 - \sum_{i=1}^{k} \lambda_i \boldsymbol{A}^{i}\boldsymbol{r}_0 \;\in\; \boldsymbol{W}_{k+1},$$

whence $\boldsymbol{V}_{k+1} \subset \boldsymbol{W}_{k+1}$.

It remains to verify the recursion formulae for $1 \le k \le k_o$. We know already from Lemma 6.4 that

$$\boldsymbol{x}_k \;=\; \boldsymbol{x}_{k-1} + t_k \boldsymbol{h}_k \quad \text{with} \quad t_k \;=\; \frac{\langle \boldsymbol{r}_{k-1}, \boldsymbol{h}_k \rangle}{\langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle}.$$

But $\boldsymbol{h}_k \in \boldsymbol{r}_{k-1} + \boldsymbol{V}_{k-1}$ and $\boldsymbol{r}_{k-1} \perp \boldsymbol{V}_{k-1}$, whence $\langle \boldsymbol{r}_{k-1}, \boldsymbol{h}_k \rangle = \langle \boldsymbol{r}_{k-1}, \boldsymbol{r}_{k-1} \rangle = \|\boldsymbol{r}_{k-1}\|^2$, and

$$t_k \;=\; \frac{\|\boldsymbol{r}_{k-1}\|^2}{\langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle}.$$

Now the simplified formula for $\boldsymbol{r}_k$ follows from

$$\boldsymbol{r}_k \;=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k \;=\; \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{A}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \;=\; \boldsymbol{r}_{k-1} - t_k \boldsymbol{A}\boldsymbol{h}_k.$$

Concerning $\boldsymbol{h}_{k+1}$, recall that $\boldsymbol{h}_{k+1} = \boldsymbol{r}_k - \sum_{i=1}^{k} \beta_{k,i}\boldsymbol{h}_i$ with

$$\beta_{k,i} \;=\; \frac{\langle \boldsymbol{r}_k, \boldsymbol{A}\boldsymbol{h}_i \rangle}{\langle \boldsymbol{h}_i, \boldsymbol{A}\boldsymbol{h}_i \rangle}.$$

But for $1 \le i < k$, the vector $\boldsymbol{h}_i$ lies in $\boldsymbol{V}_i = \mathrm{span}(\boldsymbol{A}^{s-1}\boldsymbol{r}_0 : 1 \le s \le i)$, whence $\boldsymbol{A}\boldsymbol{h}_i \in \mathrm{span}(\boldsymbol{A}^{s}\boldsymbol{r}_0 : 1 \le s \le i) \subset \boldsymbol{V}_k \perp \boldsymbol{r}_k$, and thus $\beta_{k,i} = 0$. Consequently, we know that

$$\boldsymbol{h}_{k+1} \;=\; \boldsymbol{r}_k - \frac{\langle \boldsymbol{r}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle}{\langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle}\, \boldsymbol{h}_k.$$

Finally, it follows from $\boldsymbol{r}_k = \boldsymbol{r}_{k-1} - t_k \boldsymbol{A}\boldsymbol{h}_k$ that

$$\boldsymbol{A}\boldsymbol{h}_k \;=\; t_k^{-1}(r_{k-1} - r_k) \;=\; \frac{\langle \boldsymbol{h}_k, \boldsymbol{A}\boldsymbol{h}_k \rangle}{\|\boldsymbol{r}_{k-1}\|^2}(\boldsymbol{r}_{k-1} - \boldsymbol{r}_k),$$

and this leads to

$$\boldsymbol{h}_{k+1} \;=\; \boldsymbol{r}_k - \frac{\langle \boldsymbol{r}_k, \boldsymbol{r}_{k-1} - \boldsymbol{r}_k \rangle}{\|\boldsymbol{r}_{k-1}\|^2}\, \boldsymbol{h}_k \;=\; \boldsymbol{r}_k + \frac{\|\boldsymbol{r}_k\|^2}{\|\boldsymbol{r}_{k-1}\|^2}\, \boldsymbol{h}_k,$$

because $\boldsymbol{r}_{k-1} \in \boldsymbol{V}_k \perp \boldsymbol{r}_k$. □

6.5 Bounding the running time and approximation error

In many applications of the CG algorithm, it turns out that the norm of the residual $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$ is rather small or even zero after $k \ll d$ steps. This is related to the aforementioned Krylow spaces

$$\boldsymbol{V}_k \;=\; \operatorname{span}(\boldsymbol{A}^{i-1}\boldsymbol{r}_0 : 1 \le i \le k)$$

with $\boldsymbol{r}_0 = \boldsymbol{r}(\boldsymbol{x}_0)$. As shown already,

$$\begin{aligned}
\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_k)\big\|^2 \;&=\; \min_{\boldsymbol{x}\in\boldsymbol{x}_0+\boldsymbol{V}_k} \big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x})\big\|^2 \\
&=\; \min_{\boldsymbol{v}\in\boldsymbol{V}_k} \big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0) - \boldsymbol{A}^{1/2}\boldsymbol{v}\big\|^2 \\
&=\; \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \Big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0) - \boldsymbol{A}^{1/2}\sum_{j=1}^{k}\beta_j\boldsymbol{A}^{j-1}\boldsymbol{r}_0\Big\|^2 \\
&=\; \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \Big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0) - \sum_{j=1}^{k}\beta_j\boldsymbol{A}^{j}\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0)\Big\|^2.
\end{aligned}$$

Now we consider a spectral representation of $\boldsymbol{A}$, that means, $\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{d}\lambda_i\langle\boldsymbol{u}_i,\boldsymbol{x}\rangle\boldsymbol{u}_i$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_d > 0$ and an orthonormal basis $\boldsymbol{u}_1,\dots,\boldsymbol{u}_d$ of $\boldsymbol{V}$. If we write $\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0) = \sum_{i=1}^{d}\gamma_i\boldsymbol{u}_i$ for some $\boldsymbol{\gamma}\in\mathbb{R}^d$, then

$$\begin{aligned}
\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_k)\big\|^2 \;&=\; \min_{\boldsymbol{\beta}\in\mathbb{R}^k} \Big\|\sum_{i=1}^{d}\Big(1 - \sum_{j=1}^{d}\beta_j\lambda_i^j\Big)\gamma_i\boldsymbol{u}_i\Big\|^2 \\
&=\; \min_{Q\in\mathcal{Q}_k} \sum_{i=1}^{d}\gamma_i^2(1 - Q(\lambda_i))^2,
\end{aligned}$$

where $\mathcal{Q}_k$ denotes the set of all real polynomials Q of order k with $Q(0) = 0$.

Early stopping. If the set $\{\lambda_1,\lambda_2,\dots,\lambda_d\}$ consists of the $k(\boldsymbol{A}) \le d$ different numbers $\lambda_{(1)} > \lambda_{(2)} > \cdots > \lambda_{(k(\boldsymbol{A}))}$, then

$$Q(t) \;:=\; 1 - \prod_{j=1}^{k(\boldsymbol{A})}(1 - t/\lambda_{(j)})$$

defines a polynomial from $\mathcal{Q}_{k(\boldsymbol{A})}$ such that $Q = 0$ on $\{\lambda_1,\lambda_2,\dots,\lambda_d\}$. Consequently,

$$\boldsymbol{x}_{k(\boldsymbol{A})} \;=\; \boldsymbol{x}_*$$

for arbitrary starting point $\boldsymbol{x}_0$! This result should be taken with a pinch of salt, though. Numerical errors may result in more than $k(\boldsymbol{A})$ iterations, sometimes even more than d. Indeed, one should take precautions to avoid endless loops.

Tshebyshev polynomials. The subsequent error bounds use the well-known Tshebyshev polynomials. For $k\in\mathbb{N}_0$ and $s\in\mathbb{R}$ let

$$T_k(s) \;:=\; \begin{cases} (-1)^k\cosh\big(k\operatorname{arcosh}(-s)\big) & \text{if } s \le -1, \\ \cos\big(k\arccos(s)\big) & \text{if } s\in[-1,1], \\ \cosh\big(k\operatorname{arcosh}(s)\big) & \text{if } s \ge 1. \end{cases}$$

Then $T_0(s) = 1$, $T_1(s) = s$, and

$$T_{k+1}(s) \;=\; 2sT_k(s) - T_{k-1}(s) \quad \text{for } k \in \mathbb{N}. \tag{6.4}$$

The claims about T_0 and T_1 are obvious. The recursion formula (6.4) can be verified as follows: Note first that

$$\cosh(z_1 + z_2) + \cosh(z_1 - z_2) \;=\; 2\cosh(z_1)\cosh(z_2) \quad \text{for } z_1, z_2 \in \mathbb{C}.$$

Since $\cos(z) = \cosh(\boldsymbol{i}z)$, the latter equation is true with $\cos(\cdot)$ in place of $\cosh(\cdot)$. Thus, for $s \geq 1$,

$$\begin{aligned}
T_{k+1}(\pm s) \;&= (\pm 1)^{k+1}\cosh\bigl(\operatorname{arcosh}(s) + k\operatorname{arcosh}(s)\bigr) \\
&= 2(\pm 1)^{k+1}\cosh\bigl(\operatorname{arcosh}(s)\bigr)\cosh\bigl(k\operatorname{arcosh}(s)\bigr) \\
&\quad -\; (\pm 1)^{k+1}\cosh\bigl(\operatorname{arcosh}(s) - k\operatorname{arcosh}(s)\bigr) \\
&= 2(\pm s)(\pm 1)^{k}\cosh\bigl(k\operatorname{arcosh}(s)\bigr) - (\pm 1)^{k-1}\cosh\bigl((k-1)\operatorname{arcosh}(s)\bigr) \\
&= 2(\pm s)T_k(\pm s) - T_{k-1}(\pm s).
\end{aligned}$$

The case $s \in [-1, 1]$ can be treated analogously.

The formulae $T_0(s) = 1$, $T_1(s) = s$ and the recursion formula (6.4) imply that $T_k(s)$ is equal to $\sum_{j=0}^{k} \alpha_{kj} s^j$ with certain coefficients $\alpha_{kj} \in \mathbb{R}$, where $\alpha_{kk} = 2^{k-1}$ for $k \in \mathbb{N}$. Moreover, one can show that

$$\begin{aligned}
|T_k| \;&\leq\; 1 \quad \text{on } [-1, 1], \\
|T_k| \;&>\; 1 \quad \text{on } \mathbb{R} \setminus [-1, 1], \\
T_k\Bigl(\cos\Bigl(\frac{2j-1}{k}\,\pi\Bigr)\Bigr) \;&=\; 0 \quad \text{for } j = 1, \ldots, k, \\
T_k\Bigl(\cos\Bigl(\frac{2j}{k}\,\pi\Bigr)\Bigr) \;&=\; (-1)^j \quad \text{for } j = 0, 1, \ldots, k.
\end{aligned}$$

Another useful representation of T_k results from the known formulae

$$\operatorname{arcosh}(s) \;=\; \log\bigl(s + \sqrt{s^2 - 1}\bigr) \;=\; -\log\bigl(s - \sqrt{s^2 - 1}\bigr) \quad \text{for } s \geq 1.$$

This implies that

$$T_k(\pm s) \;=\; (-1)^k\,\frac{\bigl(s + \sqrt{s^2-1}\bigr)^k + \bigl(s - \sqrt{s^2-1}\bigr)^k}{2} \quad \text{for } s \geq 1.$$

And from this one can deduce that

$$T_k(s) \;=\; \sum_{\ell=0}^{\lfloor k/2 \rfloor} \binom{k}{2\ell} s^{k-2\ell}(s^2-1)^\ell.$$

Error bounds. With the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d > 0$ and eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ of $\boldsymbol{A}$, let $\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0) = \sum_{i=1}^d \gamma_i \boldsymbol{u}_i$. With the probability weights $p_i := \|\boldsymbol{\gamma}\|^{-2}\gamma_i^2$,

$$\frac{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_k)\big\|}{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0)\big\|} \;=\; \min_{Q\in\mathcal{Q}_k} \Big(\sum_{i=1}^d p_i (1 - Q(\lambda_i))^2\Big)^{1/2} \;\leq\; \min_{Q\in\mathcal{Q}_k} \max_{t\in[\lambda_d,\lambda_1]} \big|1 - Q(t)\big|.$$

As shown later, the minimum is attained for a special polynomial Q_k such that

$$\max_{t\in[\lambda_d,\lambda_1]} \big|1 - Q_k(t)\big| \;=\; \big|1 - Q_k(\lambda_d)\big| \;=\; \big|1 - Q_k(\lambda_1)\big|.$$

Hence, if $\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0)$ fulfils

$$p_i \;=\; 0 \quad \text{whenever} \quad \big|1 - Q_k(\lambda_i)\big| \;<\; \max_{t\in[\lambda_d,\lambda_1]} \big|1 - Q_k(t)\big|,$$

then

$$\frac{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_k)\big\|}{\big\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_0)\big\|} \;=\; \max_{t\in[\lambda_d,\lambda_1]} \big|1 - Q_k(t)\big|.$$

Theorem 6.6. *For $k \in \mathbb{N}$ and real numbers $0 < a < b$ let*

$$Q_k(t \,|\, a, b) \;:=\; 1 - T_k\Big(\frac{2t - a - b}{b - a}\Big)\Big/T_k\Big(-\frac{a+b}{b-a}\Big),$$

where T_k is the Tshebyshev polynomial of order k. Then $Q_k(\cdot \,|\, a, b) \in \mathcal{Q}_k$, and

$$\max_{t\in[a,b]} \big|1 - Q_k(t \,|\, a, b)\big| \;=\; \delta_k(a/b)$$

with

$$\delta_k(u) \;:=\; \frac{2\,(1-u)^k}{\big(1+\sqrt{u}\big)^{2k} + \big(1-\sqrt{u}\big)^{2k}} \quad \textit{for} \quad u \in [0,1].$$

Furthermore, for arbitrary $Q \in \mathcal{Q}_k$,

$$\max_{t\in[a,b]} \big|1 - Q(t)\big| \;\geq\; \delta_k(a/b)$$

with equality if and only if $Q \equiv Q_k(\cdot \,|\, a, b)$.

Figure 6.2 shows for $a = 1$, $b = 4$ and $k = 2, 3, 6, 7$ the optimal polynomial $Q_k(\cdot \,|\, a, b)$.

Remark 6.7 (Inequalities for $\delta_k(\cdot)$). Obviously,

$$\delta_k(u) \;<\; 2\frac{(1-u)^k}{\big(1+\sqrt{u}\big)^{2k}} \;=\; 2\Big(\frac{1-\sqrt{u}}{1+\sqrt{u}}\Big)^k \quad \text{for } 0 \leq u < 1.$$

For $k \geq 2$ and $0 < u < 1$, strict convexity of $s \mapsto s^k$ on $(0, \infty)$ implies that

$$\frac{\big(1+\sqrt{u}\big)^{2k} + \big(1-\sqrt{u}\big)^{2k}}{2} = \frac{\big(1+u+2\sqrt{u}\big)^{k} + \big(1+u-2\sqrt{u}\big)^{k}}{2} \;>\; (1+u)^k.$$

Hence,

$$\delta_k(u) \;<\; \delta_1(u)^k \quad \text{for } k \geq 2 \text{ and } 0 < u < 1.$$

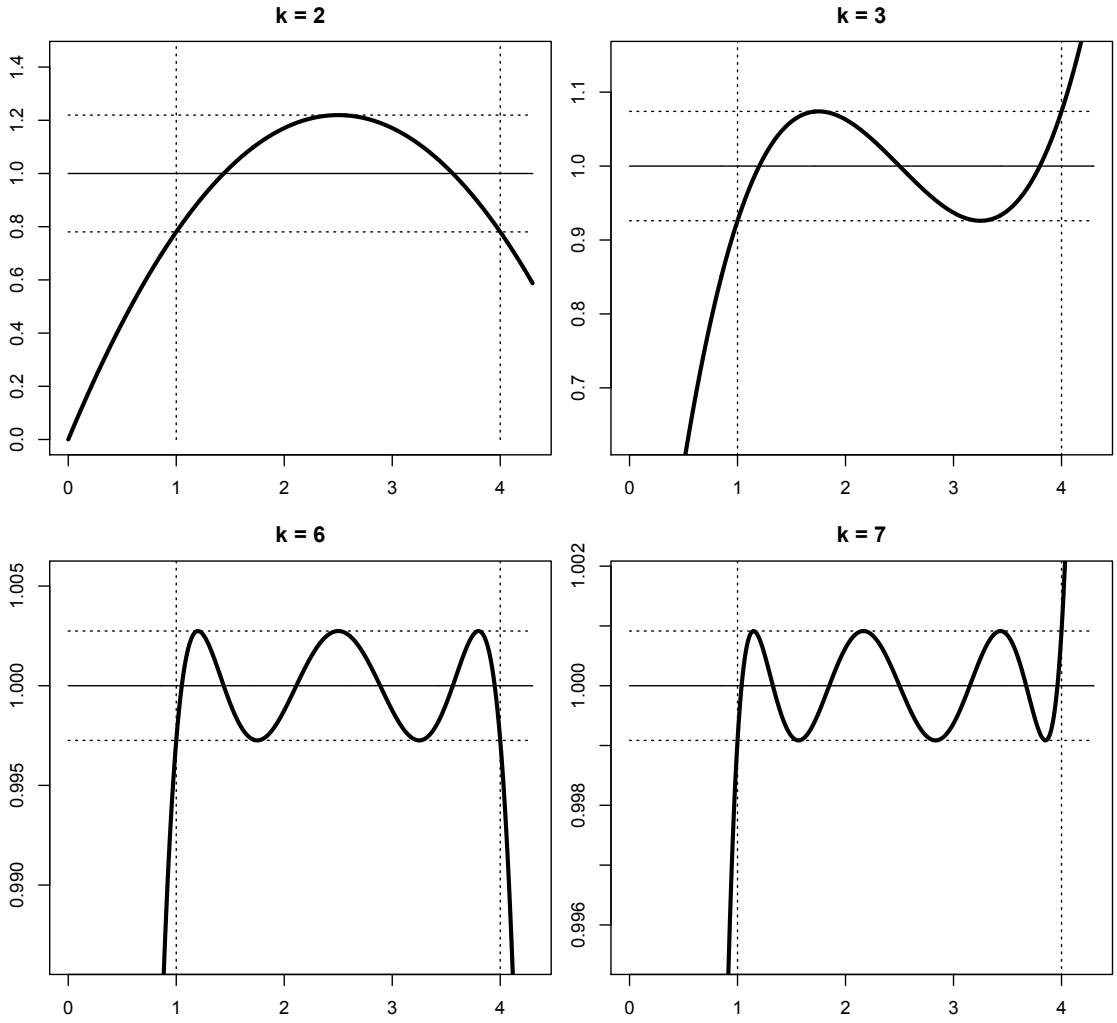

Figure 6.2: Some optimal polynomials $Q_k(\cdot \mid a, b)$ for $a = 1$ and $b = 4$.

Remark 6.8 (Expansions for $\delta_k(\cdot)$). By virtue of the binomial formula,

$$\frac{\left(1+\sqrt{u}\right)^{2k} + \left(1-\sqrt{u}\right)^{2k}}{2} \;=\; \sum_{\ell=0}^{k} \binom{2k}{2\ell} u^{\ell} \;=\; \begin{cases} 1+u & \text{for } k = 1, \\ 1 + k(2k-1)u + O(u^2) & \text{for } k \geq 1. \end{cases}$$

In particular,

$$\delta_1(u) \;=\; \frac{1-u}{1+u} \;=\; 1 - 2u + O(u^2),$$

and

$$\begin{aligned} \delta_1(u)^k &\;=\; 1 - 2ku + O(u^2), \\ \delta_k(u) &\;=\; 1 - 2k^2u + O(u^2). \end{aligned}$$

Remark 6.9 (Comparison of gradient and CG method). Theorem 6.6 and Remark 6.7 imply that k steps of the CG algorithm result in a reduction of the approximation error $\left\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x})\right\|$ by a factor $\delta_k(\lambda_d/\lambda_1)$ at least. This factor is strictly smaller than the factor $\delta_1(\lambda_d/\lambda_1)^k$ which could result from k iterations of the gradient method.

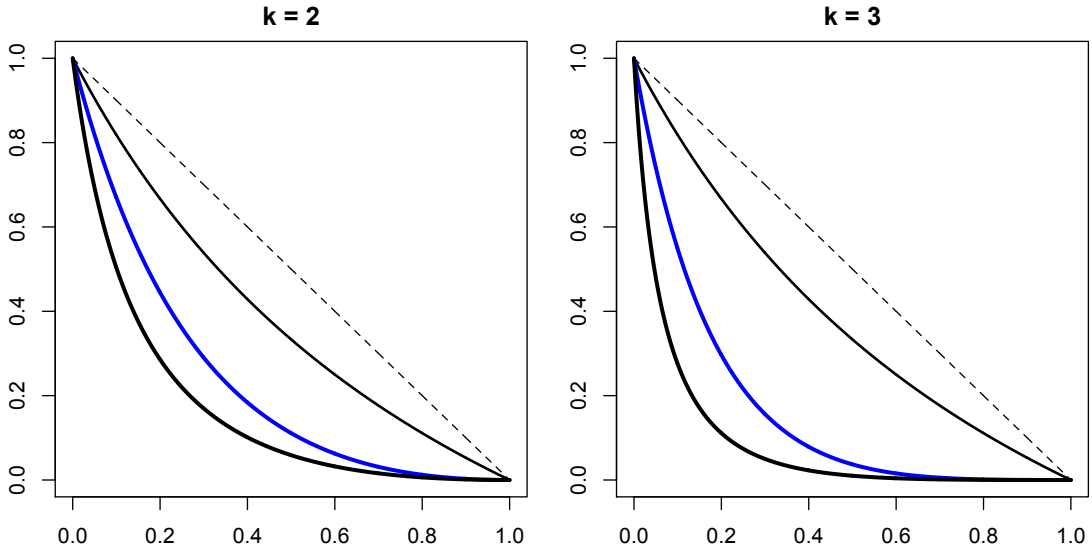

Figure 6.3: Approximation errors $\delta_1(\cdot) > \delta_1(\cdot)^k > \delta_k(\cdot)$.

Figure 6.3 shows the functions $\delta_1(\cdot) > \delta_1(\cdot)^k > \delta_k(\cdot)$ for $k = 2, 3$.

In the proof of Theorem 6.6 we use an elementary lemma about real polynomials.

Lemma 6.10. *For $k \in \mathbb{N}_0$ let $\Delta : \mathbb{R} \to \mathbb{R}$ be given by $\Delta(s) = \sum_{j=0}^{k} \alpha_j s^j$ with real coefficients $\alpha_0, \alpha_1, \ldots, \alpha_k$. Suppose that there exist points $s_0 < s_1 < \cdots < s_{k+1}$ such that*

$$(-1)^i \Delta(s_i) \;\geq\; 0 \quad \text{for } 0 \leq i \leq k+1.$$

Then, $\Delta \equiv 0$.

Exercise 6.11. Prove Lemma 6.10.

Proof of Theorem 6.6. One can easily verify that $Q_k(\cdot) = Q_k(\cdot \,|\, a, b)$ defines a polynomial in $\mathcal{Q}_k$. Further, the properties of the Tshebyshev polynomials imply that

$$\begin{aligned} \delta_k \;&:=\; \max_{t \in [a,b]} \big|1 - Q_k(t)\big| \;=\; \big|1 - Q_k(a)\big| \;=\; \frac{|T_k(-1)|}{\big|T_k\big(-(a+b)/(b-a)\big)\big|} \\ &= \frac{1}{\cosh\big(k \operatorname{arcosh}(1/\delta)\big)} \end{aligned}$$

with $\delta := (b-a)/(a+b) = (1-u)/(1+u)$. In particular, $\delta_1 = \delta$. For general $k \geq 1$,

$$\begin{aligned}
\delta_k &= \frac{2}{\big(\delta^{-1} + \sqrt{\delta^{-2}-1}\big)^k + \big(\delta^{-1} - \sqrt{\delta^{-2}-1}\big)^k} \\
&= \frac{2\,\delta^k}{\big(1 + \sqrt{1-\delta^2}\big)^k + \big(1 - \sqrt{1-\delta^2}\big)^k} \\
&= \frac{2\,(1-u)^k}{\big(1+u + \sqrt{(1+u)^2-(1-u)^2}\big)^k + \big(1+u - \sqrt{(1+u)^2-(1-u)^2}\big)^k} \\
&= \frac{2\,(1-u)^k}{\big(1+u+2\sqrt{u}\big)^k + \big(1+u-2\sqrt{u}\big)^k} \\
&= \frac{2\,(1-u)^k}{\big(1+\sqrt{u}\big)^{2k} + \big(1-\sqrt{u}\big)^{2k}}.
\end{aligned}$$

It remains to prove optimality and uniqueness of Q_k. Suppose there exists a polynomial $Q \in \mathcal{Q}_k$ such that

$$\max_{t\in[a,b]} \big|1 - Q(t)\big| \;\leq\; \delta_k.$$

Let $a = t_1 < t_2 < \cdots < t_{k+1} = b$ be the points $t_i \in [a,b]$ such that $Q_k(t_i) = 1 - (-1)^{i-1}\delta_k$. Then the polynomial $\Delta := Q_k - Q$ satisfies $\Delta(0) = 0$ and $(-1)^i \Delta(t_i) \geq 0$ for $1 \leq i \leq k+1$. Hence, it follows from Lemma 6.10 that $\Delta \equiv 0$. □

Example 6.12. We consider the matrix $\boldsymbol{A} = \operatorname{diag}(2^{1-i} : i = 1,\ldots,d)$ and the vector $\boldsymbol{b} = (1,1,\ldots,1)^\top$ with $d = 8$. For the starting point $\boldsymbol{x}_0 = \boldsymbol{0}$, the CG algorithm yields the following approximations $\boldsymbol{x}_k = (x_{k,i})_{i=1}^d$ and errors (all numbers rounded to three digits):

k	$x_{k,1}$	$x_{k,2}$	$x_{k,3}$	$x_{k,4}$	$x_{k,5}$	$x_{k,6}$	$x_{k,7}$	$x_{k,8}$	$\|\boldsymbol{r}_k\|$
0	0	0	0	0	0	0	0	0	2.828
1	4.016	4.016	4.016	4.016	4.016	4.016	4.016	4.016	3.674
2	−0.423	6.287	9.643	11.320	12.159	12.578	12.788	12.893	3.265
3	1.271	−0.984	9.273	17.248	21.947	24.474	25.782	26.448	2.560
4	0.978	2.610	−0.494	14.748	28.789	37.692	42.649	45.258	1.807
5	1.001	1.953	4.859	2.395	24.889	46.736	60.991	69.043	1.128
6	1.000	2.001	3.947	8.852	10.429	43.122	74.498	95.047	0.589
7	1.000	2.000	4.001	7.969	16.492	28.442	74.842	117.072	0.222
8	1.000	2.000	4.000	8.000	16.000	32.000	64.000	128.000	0.000

Note that $f(\boldsymbol{x}_k)$ and $\|\boldsymbol{A}^{1/2}(\boldsymbol{x}_* - \boldsymbol{x}_k)\| = \|\boldsymbol{A}^{-1/2}\boldsymbol{r}_k\|$ are non-decreasing in k, but $\|\boldsymbol{r}_k\|$ can increase every now and then.

Example 6.13 (Hilbert matrix). A notorious example for a symmetric and positive definite, but ill-conditioned matrix $\boldsymbol{A}$ is the Hilbert matrix

$$\boldsymbol{A} \;:=\; \Big(\frac{1}{i+j-1}\Big)_{i,j=1}^d.$$

Obviously, $\boldsymbol{A}$ is symmetric. That $\boldsymbol{A}$ is positive definite can be deduced from the representation

$$\frac{1}{i+j-1} \;=\; \int_0^1 t^{i-1} t^{j-1}\,dt.$$

Thus,

$$\boldsymbol{h}^\top \boldsymbol{A}\boldsymbol{h} \;=\; \int_0^1 \Big(\sum_{i=1}^{d} h_i t^{i-1}\Big)^2 dt \;>\; 0 \quad \text{for } \boldsymbol{h} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}.$$

However, the ratio λ_d/λ_1 tends to be very small, even for moderate dimension d. For instance, if $d = 8$, then $(\lambda_1, \lambda_d) \approx (1.696, 1.112 \cdot 10^{-10})$. For $\boldsymbol{b} = (1, 1, \ldots, 1)^\top$, the equation system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can not be solved with standard methods such as the R procedure `qr.solve`. But the CG algorithm with starting point $\boldsymbol{x}_0 = \boldsymbol{0}$ yields after $k = 19$ steps an approximation $\boldsymbol{x}_k$ of $\boldsymbol{x}_*$ such that $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k\| < 10^{-8}$.

Example 6.14 (Poisson problem). For integers $d_x, d_y > 2$, we consider the vector space $\boldsymbol{V} = \mathbb{R}^{d_x \times d_y}$ with inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \;:=\; \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} x_{ij} y_{ij} \;=\; \mathrm{trace}(\boldsymbol{x}^\top \boldsymbol{y}).$$

Occasionally we interpret a matrix $\boldsymbol{x} \in \boldsymbol{V}$ as a function $(i, j) \mapsto \boldsymbol{x}(i, j) = x_{ij}$ on the index set $\mathcal{D} := \{1, 2, \ldots, d_x\} \times \{1, 2, \ldots, d_y\}$.

Here is a discrete version of the classical Poisson problem (heat equation): Let

$$\begin{aligned} \mathcal{I} \;&:=\; \{2, \ldots, d_x - 1\} \times \{2, \ldots, d_y - 1\} \quad \text{and} \\ \mathcal{R} \;&:=\; \big\{(i, j) \in \mathcal{D} : i \in \{1, d_x\} \text{ oder } j \in \{1, d_y\}\big\} \end{aligned}$$

be the "interior" and the "boundary" of $\mathcal{D}$, respectively. For a given matrix $\boldsymbol{b} \in \boldsymbol{V}$ with

$$\boldsymbol{b} \;=\; 0 \;\text{ on }\; \mathcal{I}$$

we look for a matrix $\boldsymbol{x} \in \boldsymbol{V}$ such that

$$\boldsymbol{x} \;=\; \boldsymbol{b} \;\text{ on }\; \mathcal{R}$$

and

$$x_{ij} \;=\; \sum_{u,v=-1,0,1} p_{u,v} x_{i+u,j+v} \quad \text{for all } (i, j) \in \mathcal{I}. \tag{6.5}$$

Here

$$p_{u,v} \;:=\; \begin{cases} 3/16 & \text{if } |u| + |v| = 1, \\ 1/16 & \text{if } |u| + |v| = 2, \\ 0 & \text{if } u = v = 0. \end{cases}$$

In other words, we are looking for a solution $\boldsymbol{x} \in \boldsymbol{V}$ of the equation

$$\boldsymbol{L}\boldsymbol{x} \;=\; \boldsymbol{b}$$

with the linear operator $\boldsymbol{L} : \boldsymbol{V} \to \boldsymbol{V}$ given by

$$(\boldsymbol{L}\boldsymbol{x})_{ij} \;:=\; \begin{cases} x_{ij} & \text{for } (i, j) \in \mathcal{R}, \\ x_{ij} - \sum_{u,v=-1,0,1} p_{u,v} x_{i+u,j+v} & \text{for } (i, j) \in \mathcal{I}. \end{cases}$$

One may imagine $\boldsymbol{x}(i,j)$ as the temperature at the point (i,j) on a rectangular metal plate $\mathcal{D}$. By heating or cooling, the temperature is fixed as $b(i,j)$ at the boundary points $(i,j) \in \mathcal{R}$.

The operator $\boldsymbol{L}$ is easily implemented, and the computation of $\boldsymbol{Lx}$ requires $O(d_x d_y)$ steps. However, representating $\boldsymbol{L}$ by a matrix would result in a matrix of size $d_x d_y \times d_x d_y$. This can be very large, even for moderate values d_x and d_y.

The operator $\boldsymbol{L}$ is nonsingular. Indeed, suppose that $\boldsymbol{Lx} = \boldsymbol{0}$ for some $\boldsymbol{x} \in \boldsymbol{V}$. Let $(i,j) \in \mathcal{D}$ with $|x_{ij}| = \max_{i',j'} |x_{i'j'}|$. Since $\boldsymbol{x} = \boldsymbol{Lx} = 0$ on $\mathcal{R}$, we may assume without loss of generality that $(i,j) \in \mathcal{I}$. It follows from (6.5) that $x_{i+u,j+v} = x_{ij}$ for all $u, v \in \{-1,0,1\}$. Now one may deduce inductively that $\boldsymbol{x}$ is a constant function (matrix), and since $\boldsymbol{x} = \boldsymbol{Lx} = 0$ on $\mathcal{R}$, this implies that $\boldsymbol{x} = \boldsymbol{0}$.

Solution via CG algorithm. Unfortunately, $\boldsymbol{L}$ is not self-adjoint, but the adjoint operator $\boldsymbol{L}^* : \boldsymbol{V} \to \boldsymbol{V}$, given by

$$\langle \boldsymbol{x}, \boldsymbol{Ly} \rangle \;=\; \langle \boldsymbol{L}^* \boldsymbol{x}, \boldsymbol{y} \rangle \quad \text{for arbitrary } \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{V},$$

is given by

$$(\boldsymbol{L}^* \boldsymbol{x})_{ij} \;=\; x_{ij} - \sum_{u,v=-1,0,1} 1_{[(i+u,j+v)\in\mathcal{I}]} p_{u,v} x_{i+u,j+v}.$$

Thus we consider

$$\tilde{f}(\boldsymbol{x}) \;:=\; 2^{-1} \|L\boldsymbol{x} - \boldsymbol{b}\|^2 \;=\; 2^{-1} \langle \boldsymbol{x}, L^* L \boldsymbol{x} \rangle - \langle \boldsymbol{x}, L^* \boldsymbol{b} \rangle + 2^{-1} \|\boldsymbol{b}\|^2$$

and apply the CG algorithm with the self-adjoined and positive definite operator $\boldsymbol{A} := \boldsymbol{L}^* \boldsymbol{L}$ and $\boldsymbol{L}^* \boldsymbol{b}$ in place of $\boldsymbol{b}$.

Figure 6.4 depicts a gray scale image of the approximate solution $\boldsymbol{x}_k$ in case of $d_x = d_y = 50$ and

$$b_{ij} \;=\; \begin{cases} +1 & \text{if } (i,j) \in \mathcal{R} \text{ and } \min\{|i-1| + |j-1|, |i-50| + |j-50|\} \le 9, \\ -1 & \text{if } (i,j) \in \mathcal{R} \text{ and } \min\{|i-1| + |j-50|, |i-50| + |j-1|\} \le 9, \\ 0 & \text{else.} \end{cases}$$

This corresponds to the temperature on a metal plate which is heated at to opposite corners and cooled down at the other two corners. The solution depicted here resultet from the starting point $\boldsymbol{x}_0 = \boldsymbol{b}$ and the stopping criterion $\|\boldsymbol{L}^* \boldsymbol{b} - \boldsymbol{L}^* \boldsymbol{L} \boldsymbol{x}_k\| < \epsilon := 10^{-9}$ after $k = 761$ steps, which is substantially smaller than $d_x d_y = 2500$. Figure 6.5 shows $\log_{10} \|\boldsymbol{b} - \boldsymbol{L}\boldsymbol{x}_s\|$ as a function of $s \in \{0, 1, \ldots, k\}$.

Solution via fixed point iteration. The solution $\boldsymbol{x}_*$ may be interpreted stochastically. let $(Z_t)_{t \ge 0}$ be a random walk on $\mathbb{Z} \times \mathbb{Z}$ with independent increments

$$Z_{t+1} - Z_t \;\sim\; \sum_{u,v=-1,0,1} p_{u,v} \delta_{(u,v)}$$

and fixed starting value $Z_0 \in \mathcal{D}$. Further let T be the stopping time

$$T \;:=\; \min\{t \in \mathbb{N}_0 : Z_t \in \mathcal{R}\}.$$

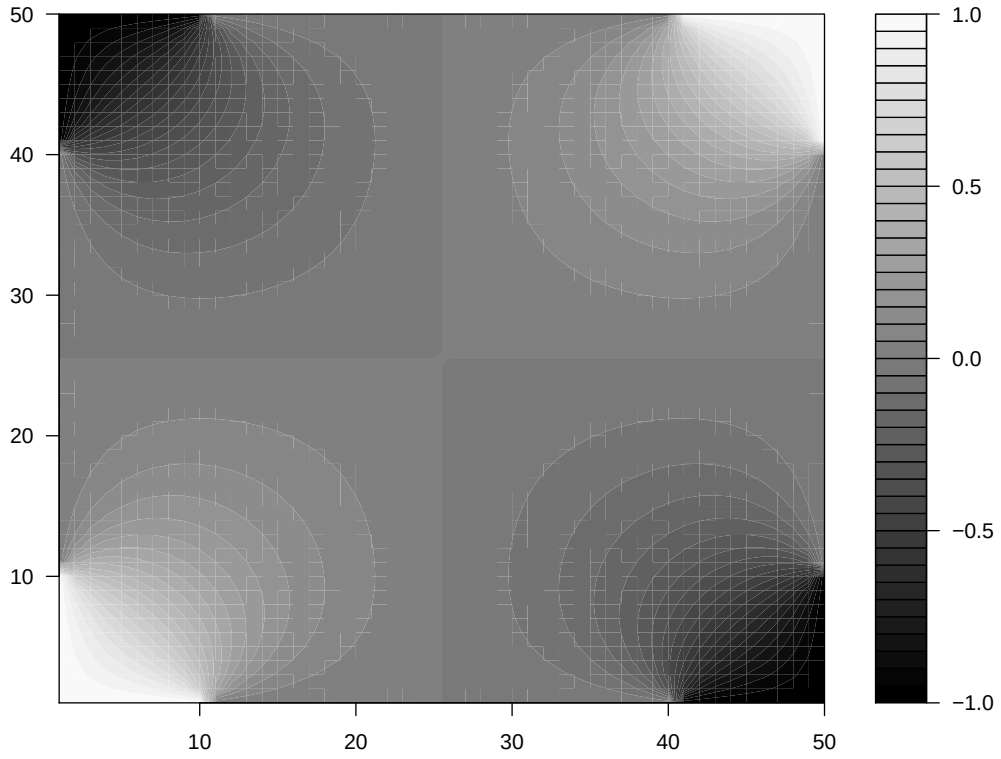

Figure 6.4: Approximate solution of the Poisson problem (Example 6.14).

Then,

$$\boldsymbol{x}_*(i,j) \;=\; \mathbb{E}\big(\boldsymbol{b}(Z_T) \,\big|\, Z_0 = (i,j)\big) \quad \text{for } (i,j) \in \mathcal{D}.$$

If we define for $s \in \mathbb{N}_0$ the matrix $\boldsymbol{y}_s \in \boldsymbol{V}$ via

$$\boldsymbol{y}_s(i,j) \;:=\; \mathbb{E}\big(\boldsymbol{b}(Z_{\min(T,s)}) \,\big|\, Z_0 = (i,j)\big),$$

then one can show that

$$\begin{aligned} \boldsymbol{y}_0 \;&=\; \boldsymbol{b}, \\ \boldsymbol{y}_{s+1} \;&=\; M\boldsymbol{y}_s \quad \text{for } s = 0,1,2,\ldots, \\ \lim_{s\to\infty} \boldsymbol{y}_s \;&=\; \boldsymbol{x}_* \;=\; M\boldsymbol{x}_*. \end{aligned}$$

Here $M : \boldsymbol{V} \to \boldsymbol{V}$ is the linear operator with

$$(M\boldsymbol{x})_{ij} \;:=\; \begin{cases} x_{ij} & \text{for } (i,j) \in \mathcal{R}, \\ \sum_{u,v=-1,0,1} p_{u,v} x_{i+u,j+v} & \text{for } (i,j) \in \mathcal{I}. \end{cases}$$

For our particular example, Figure 6.5 shows in addition the log-approximation errors $\log_{10} \|\boldsymbol{b} - L\boldsymbol{y}_s\|$, $0 \le s \le k = 761$. For quite some time, $\|\boldsymbol{b} - L\boldsymbol{y}_s\|$ is smaller than $\|\boldsymbol{b} - L\boldsymbol{x}_s\|$, but eventually, as s approaches k, $\|\boldsymbol{b} - L\boldsymbol{x}_s\|$ is substantially smaller than $\|\boldsymbol{b} - L\boldsymbol{y}_s\|$.

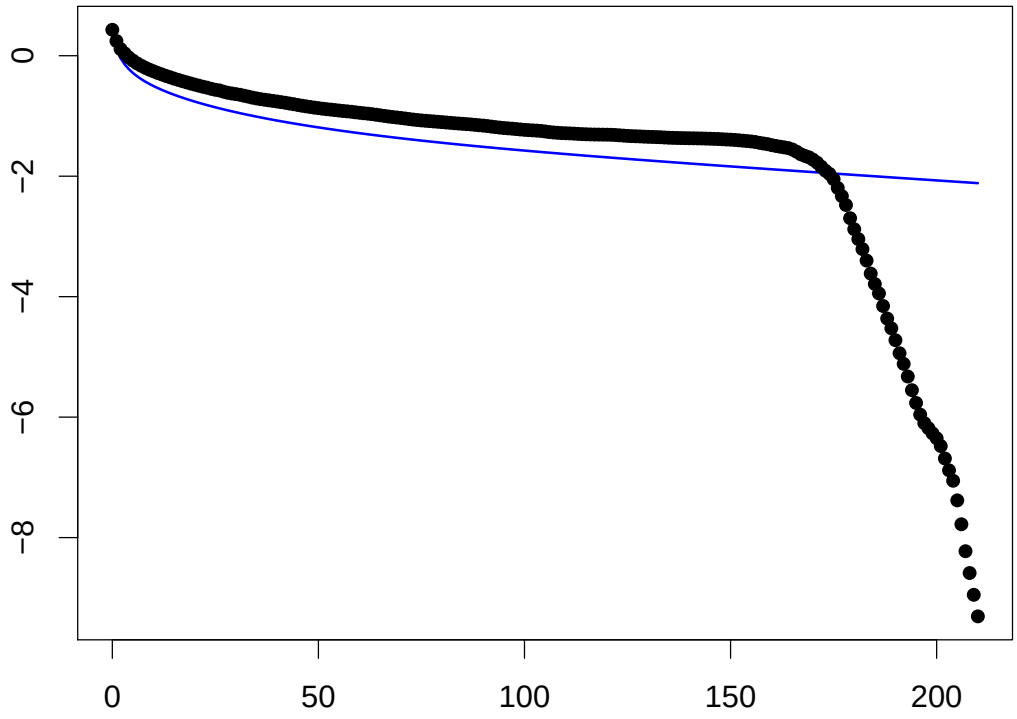

Figure 6.5: Log-approximation errors in Poisson problem (Example 6.14): CG algorithm (•) and fixed point iterations (blue line).

6.6 Minimizing a Smooth Convex Function

Let $\mathcal{X} \subset \mathbb{R}^d$ be an open, convex set, and let $f : \mathcal{X} \to \mathbb{R}$ be twice continuously differentiable such that its Hessian matrix $\boldsymbol{A}(\boldsymbol{x}) := D^2 f(\boldsymbol{x})$ is positive definite for any $\boldsymbol{x} \in \mathcal{X}$. Suppose that f has a unique minimizer $\boldsymbol{x}_* \in \mathcal{X}$. Now we describe an iterative procedure for the approximation of $\boldsymbol{x}_*$ which is inspired by the CG algorithm. To this end we write $\boldsymbol{r}(\boldsymbol{x}) := -\nabla f(\boldsymbol{x})$. The idea is to approximate $f(\boldsymbol{x} + \boldsymbol{h})$ for given $\boldsymbol{x} \in \mathcal{X}$ and "small" $\boldsymbol{h} \in \mathbb{R}^d$ by

$$\tilde{f}(\boldsymbol{x} + \boldsymbol{h} \,|\, \boldsymbol{x}) \;:=\; f(\boldsymbol{x}) - \boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{h} + 2^{-1} \boldsymbol{h}^\top \boldsymbol{A}(\boldsymbol{x}) \boldsymbol{h}.$$

We start with an arbitrary point $\boldsymbol{x}_0 \in \mathcal{X}$ and set

$$\tilde{\boldsymbol{h}}_1 \;:=\; \operatorname*{arg\,min}_{\boldsymbol{h} \in \operatorname{span}(\boldsymbol{r}(\boldsymbol{x}_0))} \tilde{f}(\boldsymbol{x}_0 + \boldsymbol{h} \,|\, \boldsymbol{x}_0) \;=\; \frac{\|\boldsymbol{r}(\boldsymbol{x}_0)\|^2}{\boldsymbol{r}(\boldsymbol{x}_0)^\top \boldsymbol{A}(\boldsymbol{x}_0) \boldsymbol{r}(\boldsymbol{x}_0)}\, \boldsymbol{r}(\boldsymbol{x}_0).$$

Suppose that for some number $k \in \mathbb{N}$, we have already chosen points $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1} \in \mathcal{X}$ and directions $\tilde{\boldsymbol{h}}_1, \tilde{\boldsymbol{h}}_2, \ldots, \tilde{\boldsymbol{h}}_k \in \mathbb{R}^d$ such that $\boldsymbol{r}(\boldsymbol{x}_{k-1})^\top \tilde{\boldsymbol{h}}_k > 0$. Then we choose

$$\boldsymbol{x}_k \;:=\; \boldsymbol{x}_{k-1} + \lambda_k \tilde{\boldsymbol{h}}_k$$

for a certain step size $\lambda_k = \lambda(\boldsymbol{x}_{k-1}, \tilde{\boldsymbol{h}}_k) \in (0, 1]$ such that $\boldsymbol{x}_k \in \mathcal{X}$ and $f(\boldsymbol{x}_k) < f(\boldsymbol{x}_{k-1})$. Then we define

$$\tilde{\boldsymbol{h}}_{k+1} \;:=\; \operatorname*{arg\,min}_{\boldsymbol{h} \in \operatorname{span}\left(\tilde{\boldsymbol{h}}_k, \boldsymbol{r}(\boldsymbol{x}_k)\right)} \tilde{f}(\boldsymbol{x}_k + \boldsymbol{h} \,|\, \boldsymbol{x}_k).$$

If f is a quadratic function on $\mathcal{X} = \mathbb{R}^d$, then this algorithm coincides essentially with the CG algorithm, provided that $\lambda_k = 1$ for all k. In general, the computation of $\boldsymbol{h}_{k+1}$ amounts to minimizing a uni- or bivariate quadratic function which is much easier than the computation of a usual Newton step $A(\boldsymbol{x}_k)^{-1}\boldsymbol{r}(\boldsymbol{x}_k)$.

Chapter 7

Dynamic Programming

Dynamic programming is a strategy to solve discrete optimization problems. More precisely, one wants to minimize a target function f on a *finite* but huge set $\mathcal{X}$. The idea of dynamic programming is to identify a sequence of minimization problems $(\mathcal{X}_1, f_1), (\mathcal{X}_2, f_2), \ldots, (\mathcal{X}_n, f_n) = (\mathcal{X}, f)$ such that

- minimizing f_1 over $\mathcal{X}_1$ is easy,
- for $2 \leq k \leq n$, minimizing f_k over $\mathcal{X}_k$ is easy, provided that we have already minimized f_j over $\mathcal{X}_j$ for $1 \leq j < k$.

We illustrate this paradigm with three particular examples. For a more thorough treatment we refer to the textbook of Cormen et al. (1990).

7.1 Dykstra's Algorithm

We are given $n \geq 3$ "locations" which we label with the numbers $1, 2, \ldots, n$. For two locations a, b let $D(a, b) \in [0, \infty]$ be their "direct distance", where we assume that $D(a, a) = 0$. Specifically one may think of geographic locations, and $D(a, b)$ could measure the distance, time or costs when travelling from a to b directly. The value $D(a, b) = \infty$ could indicate that there is no direct route from a to b.

For two locations a and b we are now interested in a path

$$\vec{x} \;=\; (x_0, x_1, \ldots, x_m)$$

of arbitrary length $m \geq 1$ consisting of locations $x_0, x_1, \ldots, x_m$ such that $x_0 = a$ and $x_m = b$ with minimal total distance

$$f(\vec{x}) \;:=\; \sum_{j=1}^{m} D(x_{j-1}, x_j).$$

Since $D \geq 0$, it suffices to consider paths of length at most $n-1$, because constant pieces or loops would never diminish total distance. The assumption that $D(c, c) = 0$ for all c implies that any path of lenth $m < n - 1$ may be extended by a constant piece without changing its total distance.

Hence we consider for $m = 1, 2, \ldots, n-1$ the minimal total distances

$$G_m(a,b) \;:=\; \min_{\vec{x} \in \mathcal{X}_m(a,b)} f(\vec{x})$$

with $\mathcal{X}_m(a,b)$ denoting the set of all paths $\vec{x} = (x_0, x_1, \ldots, x_m)$ of length m connecting a and b.

In principle one could determine $G_{n-1}(a,b)$ by considering all n^{n-2} paths in $\mathcal{X}_{n-1}(a,b)$. Or one goes through all $(n-1)!$ paths $(x_0, x_1, \ldots, x_{n-1})$ such that $x_0 = a$ and $\{x_0, x_1, \ldots, x_{n-1}\} = \{1, 2, \ldots, n\}$ and computes for each $b \neq a$ the sum $\sum_{j=1}^{j(b)} D(x_{j-1}, x_j)$ with $j(b)$ denoting the unique index $j \geq 1$ such that $x_j = b$. But even for moderate numbers n this would be way too demanding.

Instead one may solve the problem in $O(n^3)$ steps with memory $O(n^2)$ for any fixed location a and simultaneously for all destinations b: Note first that

$$G_1(a,b) \;=\; D(a,b).$$

For $m = 2, 3, \ldots, n-1$ we have the recursion

$$G_m(a,b) \;=\; \min_{z=1,\ldots,n} \big(G_{m-1}(a,z) + D(z,b)\big) \;\leq\; G_{m-1}(a,b),$$

because the part $(x_0, \ldots, x_{m-1})$ of an optimal path $(x_0, x_1, \ldots, x_m) \in \mathcal{X}_m(a,b)$ has to be an optimal path in $\mathcal{X}_{m-1}(a, x_{m-1})$. These facts are utilized in Dykstra's algorithm described in Table 7.1: For a given location a we determine the matrix

$$\boldsymbol{G} \;:=\; \big(G_m(a,y)\big)_{1 \leq m \leq n-1, 1 \leq y \leq n} \in [0,\infty]^{(n-1)\times n}$$

inductively for $m = 2, 3, \ldots, n-1$. Thereafter, we determine for an arbitrary b an optimal path $(x_0, x_1, \ldots, x_{n-1}) \in \mathcal{X}_{n-1}(a,b)$. To ease the latter task, the algorithm utilizes a second matrix $\boldsymbol{J} = \big(J_m(y)\big)_{1 \leq m \leq n-1, 1 \leq y \leq n}$ with $J_1(y) := a$ and

$$J_m(y) \;\in\; \operatorname*{arg\,min}_{z=1,\ldots,n} \big(G_{m-1}(a,z) + D(z,y)\big) \quad \text{for } 2 \leq m \leq n-1, 1 \leq y \leq n.$$

Then $(x_0, x_1, \ldots, x_{n-1})$ with $x_{n-1} := b$ and $x_{m-1} := x_{J_m(x_m)}$ for $m = n-1, n-2, \ldots, 1$ yields an optimal path in $\mathcal{X}_{n-1}(a,b)$.

If we also require that $J_m(y) = y$ whenever possible, then the optimal path $(x_0, x_1, \ldots, x_{n-1}) \in \mathcal{X}_{n-1}(a,b)$ just constructed will have at most one constant piece at the end and can be shortened easily, see Exercise 7.2.

Remark 7.1. Dykstra's algorithm is often formulated as an algorithm acting on weighted and directed graphs, see for instance Cormen et al. (1990). Depending on the underlying data structure and distance function, it may be worthwhile to utilize the regions $E(y) := \{z : D(z,y) < \infty\}$ and write

$$J_{m,y} \;\leftarrow\; z_o \in \operatorname*{arg\,min}_{z \in E(y)} \big(G_{m-1,z} + D(z,y)\big)$$

in Table 7.1.

Algorithm $\vec{x} \leftarrow$ **Dykstra**$\big(D(\cdot,\cdot), a, b\big)$

$\boldsymbol{J} \leftarrow (a)_{1\le m\le n-1, 1\le y\le n}$
$\boldsymbol{G} \leftarrow (\infty)_{1\le m\le n-1, 1\le y\le n}$
$\big(G_{1,y}\big)_{1\le y\le n} \leftarrow \big(D(a,y)\big)_{1\le y\le n}$
for $m \leftarrow 2$ **to** $(n-1)$ **do**
    for $y \in \{1,\dots,n\}$ **do**
        $J_{m,y} \leftarrow z_o \in \underset{1\le z\le n}{\arg\min}\big(G_{m-1,z} + D(z,y)\big)$
        $G_{m,y} \leftarrow G_{m-1,z_o} + D(z_o, y)$
    end for
end for

$\vec{x} = (x_0,\dots,x_{n-2},x_{n-1}) \leftarrow (*,\dots,*,b)$
for $m \leftarrow (n-1)$ **downto** 1 **do**
    $x_{m-1} \leftarrow J_{m,x_m}$
end for

Table 7.1: Dykstra's Algorithm.

Exercise 7.2. Suppose that for arbitrary $y \in \{1,\dots,n\}$ and $m \in \{2,\dots,n-1\}$, $J_m(y) = y$ whenever possible. Show that Dykstra's algorithm produces an optimal path $(x_0, x_1, \dots, x_{n-1})$ in $\mathcal{X}_{n-1}(a,b)$ such that for some $M \in \{0,1,\dots,n-1\}$, the points x_j, $0 \le j \le M$ are distinct while $x_M = \cdots = x_{n-1} = b$.

Exercise 7.3 (Negative costs). If one considers the planning of a bike route and if $D(a,b)$ measures strain to get from a to b, it could be reasonable to consider negative numbers $D(a,b)$, for instance, if the direct road from a to b goes mostly downhill. Formulate conditions on a cost function $D : \{1,\dots,n\} \times \{1,\dots,n\} \to (-\infty,\infty]$ such that Dykstra's algorithm works without modifications.

Exercise 7.4 (A small world). Consider n people which we label with $1, 2, \dots, n$. Let $\boldsymbol{K} \in \{0,1\}^{n\times n}$ be a matrix such that $K_{a,b} = 1$ indicates that person a knows person b. Now let $\boldsymbol{U} \in \{0,1\}^{n\times n}$ have entry $U_{a,b} = 1$ if and only if there is a tuple $(x_j)_{j=0}^m$ of $m+1 \le n$ persons such that $x_0 = a$, $x_m = b$ and $K_{x_{j-1},x_j} = 1$ for $1 \le j \le m$. How could one determine this matrix $\boldsymbol{U}$ efficiently from $\boldsymbol{K}$?

7.2 Alignment of Sequences

We are given two sequences $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ of "letters" x_i and y_j in an alphabet $\mathcal{A}$. For instance, the sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ could represent amino acids of two proteins, i.e. $\mathcal{A}$ consists of the 20 possible amino acids. Alternatively, $\boldsymbol{x}$ and $\boldsymbol{y}$ could be two DNA sequences, represented by letters in the alphabet $\mathcal{A}$ of 4 possible nucleic acids.

To judge how similar these sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ are, we want to determine an optimal 'alignment' of them. That means, we are looking for a matrix

$$\boldsymbol{H} = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,N} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,N} \end{bmatrix}$$

with entries $H_{i,j} \in \bar{\mathcal{A}} := \mathcal{A} \cup \{\diamond\}$, where $\diamond \notin \mathcal{A}$ represents a blank (space), such that the two rows $(H_{1,j})_{j=1}^N$ and $(H_{2,j})_{j=1}^N$ contain the entries of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, in their original order, augmented by several blanks $\diamond$ at arbitrary places. The number N of columns is arbitrary but will certainly satisfy $N \geq \max(m, n)$. Our goal is to minimize

$$f(\boldsymbol{H}) \;:=\; \sum_{j=1}^{N} c(H_{1,j}, H_{2,j})$$

over all alignments $\boldsymbol{H}$ of $\boldsymbol{x}$ and $\boldsymbol{y}$, where $c : \bar{\mathcal{A}} \times \bar{\mathcal{A}} \to [0, \infty)$ is a certain cost function satisfying $c(\diamond, \diamond) = 0$, $c(a, a) \leq 0$ for all $a \in \mathcal{A}$, and $c(a, b) > 0$ for $a, b \in \bar{\mathcal{A}}$ with $a \neq b$. Obviously it suffices to consider alignments $\boldsymbol{H}$ with $N \leq m + n$ columns, because columns with two blanks $\diamond$ could be deleted without increasing $f(\boldsymbol{H})$.

To construct an optimal alignment $\boldsymbol{H}$ via dynamic programming we consider the subsequences $\boldsymbol{x}^{(k)} := (x_1, x_2, \ldots, x_k)$ and $\boldsymbol{y}^{(\ell)} := (y_1, y_2, \ldots, y_\ell)$, including the empty sequences $\boldsymbol{x}^{(0)}$ and $\boldsymbol{y}^{(0)}$. Now one can easily verify that an optimal alignment $\boldsymbol{H} \in \bar{\mathcal{A}}^{2\times N}$ of $\boldsymbol{x}^{(k)}$ and $y^{(\ell)}$ with $k, \ell \geq 1$ needs to satisfy one of the following three conditions:

(i) $H_{1,N} = x_k$, $H_{2,N} = y_\ell$,
and $(H_{i,j})_{1\leq i\leq 2, 1\leq j<N}$ is an optimal alignment of $\boldsymbol{x}^{(k-1)}$ and $\boldsymbol{y}^{(\ell-1)}$;

(ii) $H_{1,N} = x_k$, $H_{2,N} = \diamond$,
and $(H_{i,j})_{1\leq i\leq 2, 1\leq j<N}$ is an optimal alignment of $\boldsymbol{x}^{(k-1)}$ and $\boldsymbol{y}^{(\ell)}$;

(iii) $H_{1,N} = \diamond$, $H_{2,N} = y_\ell$,
and $(H_{i,j})_{1\leq i\leq 2, 1\leq j<N}$ is an optimal alignment of $\boldsymbol{x}^{(k)}$ and $\boldsymbol{y}^{(\ell-1)}$.

Hence we construct a matrix $\boldsymbol{G} = (G_{k,\ell})_{0\leq k\leq m, 0\leq \ell\leq n}$ as follows:

$$\begin{aligned} G_{0,0} \;&:=\; 0, \\ G_{k,0} \;&:=\; \sum_{i=1}^{k} c(x_i, \diamond) \quad \text{for } 1 \leq k \leq m, \\ G_{0,\ell} \;&:=\; \sum_{j=1}^{\ell} c(\diamond, y_j) \quad \text{for } 1 \leq \ell \leq n, \end{aligned}$$

and for $1 \leq k \leq m$, $1 \leq \ell \leq n$ let

$$\begin{aligned} G_{k,\ell} \;&:=\; \min\Big\{ \sum_{j=1}^{N} c(H_{1,j}, H_{2,j}) : \boldsymbol{H} \text{ Alignment of } \boldsymbol{x}^{(k)} \text{ and } \boldsymbol{y}^{(\ell)} \Big\} \\ &=\; \min\big(G_{k-1,\ell-1} + c(x_k, y_\ell), G_{k-1,\ell} + c(x_k, \diamond), G_{k,\ell-1} + c(\diamond, y_\ell) \big). \end{aligned}$$

This matrix may be computed inductively in $O(mn)$ steps, and one obtains an optimal alignment by backtracking the matrix $\boldsymbol{G}$ as described in Table 7.2.

Remark 7.5. Alignment algorithms as the one described here are a standard tool in many genome data bases. Instead of minimizing a cost function, these programs often maximize a similarity function, i.e. $c(a, b)$ measures similarity of two letters $a, b \in \bar{\mathcal{A}}$. In this case one has to use the assignment

$$G_{k,\ell} \;\leftarrow\; \max\big(G_{k-1,\ell-1} + c(x_k, y_\ell), G_{k-1,\ell} + c(x_k, \diamond), G_{k,\ell-1} + c(\diamond, y_\ell) \big)$$

in Table 7.2.

Exercise 7.6 (Longest monotone subsequences). Let $\boldsymbol{x} = (x_j)_{j=1}^n$ be a given vector in $\mathbb{R}^n$. We are looking for a subvector $(x_{j(i)})_{i=1}^m$ satisfying

$$j(1) < j(2) < \cdots < j(m) \quad \text{and} \quad x_{j(1)} \leq x_{j(2)} \leq \cdots \leq x_{j(m)}$$

such that m is as large as possible. Design and implement a dynamic program for this task.

Algorithmus $\boldsymbol{H} \leftarrow \textbf{Align}\big(\boldsymbol{x}, \boldsymbol{y}, c(\cdot,\cdot)\big)$

$(m,n) \leftarrow \big(\mathrm{length}(\boldsymbol{x}), \mathrm{length}(\boldsymbol{y})\big)$
$\boldsymbol{G} \leftarrow \big(\infty\big)_{0\le k\le m, 0\le \ell\le n}$
$G_{0,0} \leftarrow 0$
$\big(G_{k,0}\big)_{1\le k\le m} \leftarrow \Big(\sum_{i=1}^{k} c(x_i,\diamond)\Big)_{1\le k\le m}$
$\big(G_{0,\ell}\big)_{1\le \ell\le n} \leftarrow \Big(\sum_{j=1}^{\ell} c(\diamond,y_j)\Big)_{1\le \ell\le n}$
for $k \leftarrow 1$ **to** m **do**
    for $\ell \leftarrow 1$ **to** n **do**
        $G_{k,\ell} \leftarrow \min\big(G_{k-1,\ell-1} + c(x_k,y_\ell), G_{k-1,\ell} + c(x_k,\diamond), G_{k,\ell-1} + c(\diamond,y_\ell)\big)$
    end for
end for

$\boldsymbol{H} \leftarrow \big(\diamond\big)_{1\le i\le 2, 1\le j\le m+n}$
$(k,\ell) \leftarrow (m,n)$
$T \leftarrow m+n$
while $k>0$ **and** $\ell>0$ **do**
    if $G_{k,\ell} = G_{k-1,\ell-1} + c(x_k,y_\ell)$ **then**
        $(H_{1,T}, H_{2,T}) \leftarrow (x_k, y_\ell)$
        $(k,\ell) \leftarrow (k-1, \ell-1)$
    else if $G_{k,\ell} = G_{k-1,\ell} + c(x_k,\diamond)$ **then**
        $(H_{1,T}, H_{2,T}) \leftarrow (x_k, \diamond)$
        $k \leftarrow k-1$
    else
        $(H_{1,T}, H_{2,T}) \leftarrow (\diamond, y_\ell)$
        $\ell \leftarrow \ell-1$
    end if
    $T \leftarrow T-1$
end while
if $k>0$ **then**
    $(H_{1,T-k+1}, \ldots, H_{1,T}) \leftarrow (x_1, \ldots, x_k)$
    $T \leftarrow T-k$
else if $\ell>0$ **then**
    $(H_{2,T-\ell+1}, \ldots, H_{2,T}) \leftarrow (y_1, \ldots, y_\ell)$
    $T \leftarrow T-\ell$
end if
$\boldsymbol{H} \leftarrow \big(H_{i,j}\big)_{1\le i\le 2, T<j\le m+n}$

Table 7.2: Alignment of two sequences.

7.3 Bimonotone Regression

Our third example combines ideas from previous chapters, notably convex polyhedral cones and the PAVA, with dynamic programming to solve a particular regression problem. This material is taken from Beran and Dümbgen [2].

For integers $m, n > 1$, let $\boldsymbol{y} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{w} \in (0,\infty)^{m\times n}$ be given matrices, and consider the function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ given by

$$f(\boldsymbol{x}) \;:=\; \sum_{i,j} w_{ij}(y_{ij} - x_{ij})^2,$$

where $\sum_{i,j}\cdots$ is shorthand notation for $\sum_{i=1}^m \sum_{j=1}^n \cdots$. As in previous chapters, we identify $\mathbb{R}^{m\times n}$ with the Euclidean space $\mathbb{R}^d$, where $d = mn$, so the standard inner product is given by $\langle \boldsymbol{x}, \tilde{\boldsymbol{x}}\rangle := \mathrm{trace}(\boldsymbol{x}^\top \tilde{\boldsymbol{x}})$, and $\|\boldsymbol{x}\| = \mathrm{trace}(\boldsymbol{x}^\top \boldsymbol{x})^{1/2} = \big(\sum_{i,j} x_{ij}^2\big)^{1/2}$ is the Frobenius norm.

Obviously, f is strictly convex and coercive, so there exists a unique minimizer $\boldsymbol{x}^* \in \mathcal{K}$ of f over any closed, convex subset $\mathcal{K}$ of $\mathbb{R}^{r\times s}$. Note also that

$$\nabla f(\boldsymbol{x}) \;=\; 2\big(w_{ij}(x_{ij} - y_{ij})\big)_{i\le r, j\le s}$$

in the sense that

$$f(\boldsymbol{x} + \boldsymbol{v}) \;=\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle + \sum_{i,j} w_{ij} v_{ij}^2.$$

Thus a point $\boldsymbol{x} \in \mathcal{K}$ is equal to $\boldsymbol{x}^*$ if and only if

$$\langle \nabla f(\boldsymbol{x}), \tilde{\boldsymbol{x}} - \boldsymbol{x}\rangle \;\ge\; 0 \quad \text{for all } \tilde{\boldsymbol{x}} \in \mathcal{K}. \tag{7.1}$$

Now we consider the special set $\mathcal{K}$ of all matrices $\boldsymbol{x} \in \mathbb{R}^{r\times s}$ which are bimonotone in the sense that

$$\begin{aligned} x_{ij} &\le x_{i+1,j} \;\text{ for } 1 \le i < r \text{ and } 1 \le j \le s,\\ x_{ij} &\le x_{i,j+1} \;\text{ for } 1 \le i \le r \text{ and } 1 \le j < s. \end{aligned}$$

Note that $\mathcal{K}$ is a closed, concex cone (and a convex polyhedron). Consequently, characterization (7.1) of $\boldsymbol{x}^*$ is equivalent to the following two conditions:

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}\rangle \;=\; 0, \tag{7.2}$$

$$\langle \nabla f(\boldsymbol{x}), \tilde{\boldsymbol{x}}\rangle \;\ge\; 0 \quad \text{for all } \tilde{\boldsymbol{x}} \in \mathcal{K}. \tag{7.3}$$

Concerning (7.3), we may apply Exercise 2.37 to show that $\mathcal{K}$ consists of all matrices

$$\boldsymbol{x} \;=\; \sum_{\boldsymbol{e}\in\mathcal{E}} \lambda_{\boldsymbol{e}} \boldsymbol{e},$$

where

$$\mathcal{E} \;:=\; \mathcal{K} \cap \{0,1\}^{r\times s},$$

and $\lambda_{\boldsymbol{e}} \geq 0$ whenever $\boldsymbol{e} \neq \mathbf{1} := (1)_{i\leq r, j\leq s}$. This representation of $\mathcal{K}$ implies that (7.3) is equivalent to the following two conditions:

$$\langle \nabla f(\boldsymbol{x}), \mathbf{1}\rangle \;=\; 0, \tag{7.4}$$

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{e}\rangle \;\geq\; 0 \quad \text{for all } \boldsymbol{e} \in \mathcal{E}. \tag{7.5}$$

Now we are ready to describe an explicit active set algorithm for the minimization of f over $\mathcal{K}$. It consists of alternating between the following two basic procedures finitely many times.

Basic procedure 1: Finding a locally optimal matrix. For $\boldsymbol{x} \in \mathcal{K}$ consider the set

$$\mathcal{K}(\boldsymbol{x}) \;:=\; \big\{\tilde{\boldsymbol{x}} \in \mathbb{R}^{m\times n} : \tilde{x}_{ij} \leq \tilde{x}_{k\ell} \text{ whenever } x_{ij} \leq x_{k\ell}\big\}.$$

One can easily verify that $\mathcal{K}(\boldsymbol{x})$ is a subcone of $\mathcal{K}$. More importantly, if $\sigma(1), \sigma(2), \ldots, \sigma(mn)$ is a list of all index pairs $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$ such that $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \cdots \leq x_{\sigma(mn)}$, then minimizing $f(\cdot)$ over $\mathcal{K}(\boldsymbol{x})$ is equivalent to minimizing

$$\sum_{i=1}^{mn} \boldsymbol{w}_{\sigma(i)}(y_{\sigma(i)} - z_i)^2$$

over all vectors $\boldsymbol{z} \in \mathbb{R}^{mn}$ such that $z_i \leq z_j$ whenever $i \leq j$ and $z_i = z_j$ whenever $x_{\sigma(i)} = x_{\sigma(j)}$. This can be done via a suitable modification of the pool-adjacent-violators algorithm. Finally, the resulting minimizer $\boldsymbol{z}$ corresponds to a matrix $\boldsymbol{x}^{\text{new}} \in \mathcal{K}(\boldsymbol{x})$ with components $x^{\text{new}}_{\sigma(i)} = z_i$ for $1 \leq i \leq mn$. This matrix satisfies $f(\boldsymbol{x}^{\text{new}}) \leq f(\boldsymbol{x})$ with equality if and only if $\boldsymbol{x}^{\text{new}} = \boldsymbol{x}$. Finally, we replace $\boldsymbol{x}$ with $\boldsymbol{x}^{\text{new}}$ and note that the new matrix $\boldsymbol{x}$ is locally optimal in the sense that $f(\tilde{\boldsymbol{x}}) \geq f(\boldsymbol{x})$ for all $\tilde{\boldsymbol{x}} \in \mathcal{K}(\boldsymbol{x})$.

Basic procedure 2: Checking for global optimality and moving on, if necessary. Suppose that $\boldsymbol{x} \in \mathcal{K}$ is locally optimal. Applying (7.2) and (7.4) with $\mathcal{K}(\boldsymbol{x})$ in place of $\mathcal{K}$ shows that $\boldsymbol{x}$ satisfies these two conditions. Consequently, $\boldsymbol{x}$ is equal to the global minimizer $\boldsymbol{x}^*$ if and only if it satisfies (7.5). To this end, we determine a matrix

$$\boldsymbol{e}(\boldsymbol{x}) \;\in\; \operatorname*{arg\,min}_{\boldsymbol{e}\in\mathcal{E}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{e}\rangle.$$

If $\langle \nabla f(\boldsymbol{x}), \boldsymbol{e}\rangle = 0$, we know that $\boldsymbol{x} = \boldsymbol{x}^*$. Otherwise, we replace $\boldsymbol{x}$ with

$$\boldsymbol{x} + t(\boldsymbol{x})\boldsymbol{e}(\boldsymbol{x})$$

with

$$t(\boldsymbol{x}) \;:=\; \operatorname*{arg\,min}_{t\geq 0} \; f(\boldsymbol{x} + t\boldsymbol{e}(\boldsymbol{x})) \;=\; -2^{-1}\big\langle \nabla f(\boldsymbol{x}), \boldsymbol{e}(\boldsymbol{x})\big\rangle \big/ \sum_{i,j} e_{ij}(\boldsymbol{x}).$$

The new vector $\boldsymbol{x}$ has a strictly smaller value $f(\boldsymbol{x})$ than its predecessor. Then we run basic procedure 1 with this new vector $\boldsymbol{x}$.

Convergence of the algorithm. Since the number of locally optimal matrices $\boldsymbol{x} \in \mathcal{K}$ is finite, after finitely many repetitions of basic procedures 1 and 2, the resulting matrix $\boldsymbol{x}$ will be the global optimum $\boldsymbol{x}^*$.

In high-dimensional settings, one could also stop earlier when $-\big\langle \nabla f(\boldsymbol{x}), \boldsymbol{e}(\boldsymbol{x}) \big\rangle$ is smaller than a small constant $\delta > 0$.

Minimizing $\langle \boldsymbol{a}, \cdot \rangle$ over $\mathcal{E}$. With $\boldsymbol{a} := \nabla f(\boldsymbol{x})$, basic procedure 2 involves the minimization of $\langle \boldsymbol{a}, \boldsymbol{e} \rangle$ with respect to $\boldsymbol{e} \in \mathcal{E}$. Note that a matrix $\boldsymbol{e} \in \{0,1\}^{m \times n}$ belongs to $\mathcal{K}$ if and only if

$$e_{ij} \;=\; 1_{[j > f(i)]}$$

for some non-increasing function $f : \{1, \dots, m\} \to \{0, 1, \dots, n\}$, and a simple combinatorial argument shows that

$$\#\mathcal{E} \;=\; \binom{m+n}{m}.$$

Since this grows exponentially in $\min(m,n)$, computing $\langle \boldsymbol{a}, \boldsymbol{e} \rangle$ for all $\boldsymbol{e} \in \mathcal{E}$ is not feasible in general. Fortunately, the optimization problem can be solved via dynamic programming as follows.

Let $\boldsymbol{H} = (H_{k,\ell})_{1 \le k \le m+1, 1 \le \ell \le n+1}$ be given by

$$\begin{aligned}
H_{m+1,\cdot} \;&:=\; 0, \\
H_{k,n+1} \;&:=\; \min\Big\{ \sum_{i=k}^{m} \sum_{j=1}^{n} a_{ij} e_{ij} : \boldsymbol{e} \in \mathcal{E} \Big\} \quad \text{for } 1 \le k \le m, \\
H_{k,\ell} \;&:=\; \min\Big\{ \sum_{i=k}^{m} \sum_{j=1}^{n} a_{ij} e_{ij} : \boldsymbol{e} \in \mathcal{E}, e_{k\ell} = 1 \Big\} \quad \text{for } 1 \le k \le m, 1 \le \ell \le n,
\end{aligned}$$

so

$$H_{1,n+1} \;=\; \min_{\boldsymbol{e} \in \mathcal{E}} \langle \boldsymbol{a}, \boldsymbol{e} \rangle.$$

To compute $\boldsymbol{H}$ via dynamic programming, we use the auxiliary matrix $\boldsymbol{B} = (B_{i,\ell})_{1 \le i \le m, 1 \le \ell \le n+1}$ with components $B_{i,n+1} = 0$ and

$$B_{i,\ell} := \sum_{j=\ell}^{n} a_{ij} \;\text{ for } 1 \le i \le m, 1 \le \ell \le n.$$

Then for $1 \le k \le m$ and $2 \le \ell \le n+1$,

$$\begin{aligned}
H_{k,1} \;&=\; B_{k,1} + H_{k+1,1}, \\
H_{k,\ell} \;&=\; \min\big(H_{k,\ell-1}, B_{k,\ell} + H_{k,\ell}\big).
\end{aligned}$$

Consequently, $\boldsymbol{B}$ and $\boldsymbol{H}$ can be computed in time $O(mn)$. A minimizer $\boldsymbol{e} \in \mathcal{E}$ of $\langle \boldsymbol{a}, \boldsymbol{e} \rangle$ can be

determined by backtracking:

```
e ← 0
k ← 1
ℓ ← n
while k ≤ m and ℓ ≥ 1 do
        if H_{k,ℓ} = H_{k,ℓ+1} then
                (e_{iℓ})_{k≤i≤m} ← 1
                ℓ ← ℓ − 1
        else
                k ← k + 1
        end if
end while
```

Numerical example. Figure 7.1 shows a numerical example for bimonotone regression with $m = n = 100$. The upper panel shows the data matrix $\boldsymbol{y}$ on a colour scale from light yellow to dark purple for $-5 \leq y_{ij} \leq 5$. The lower panel depicts the matrix $\boldsymbol{x}^*$ on a colour scale from light yellow to dark purple for $-2.2 \leq x_{ij} \leq 2.7$.

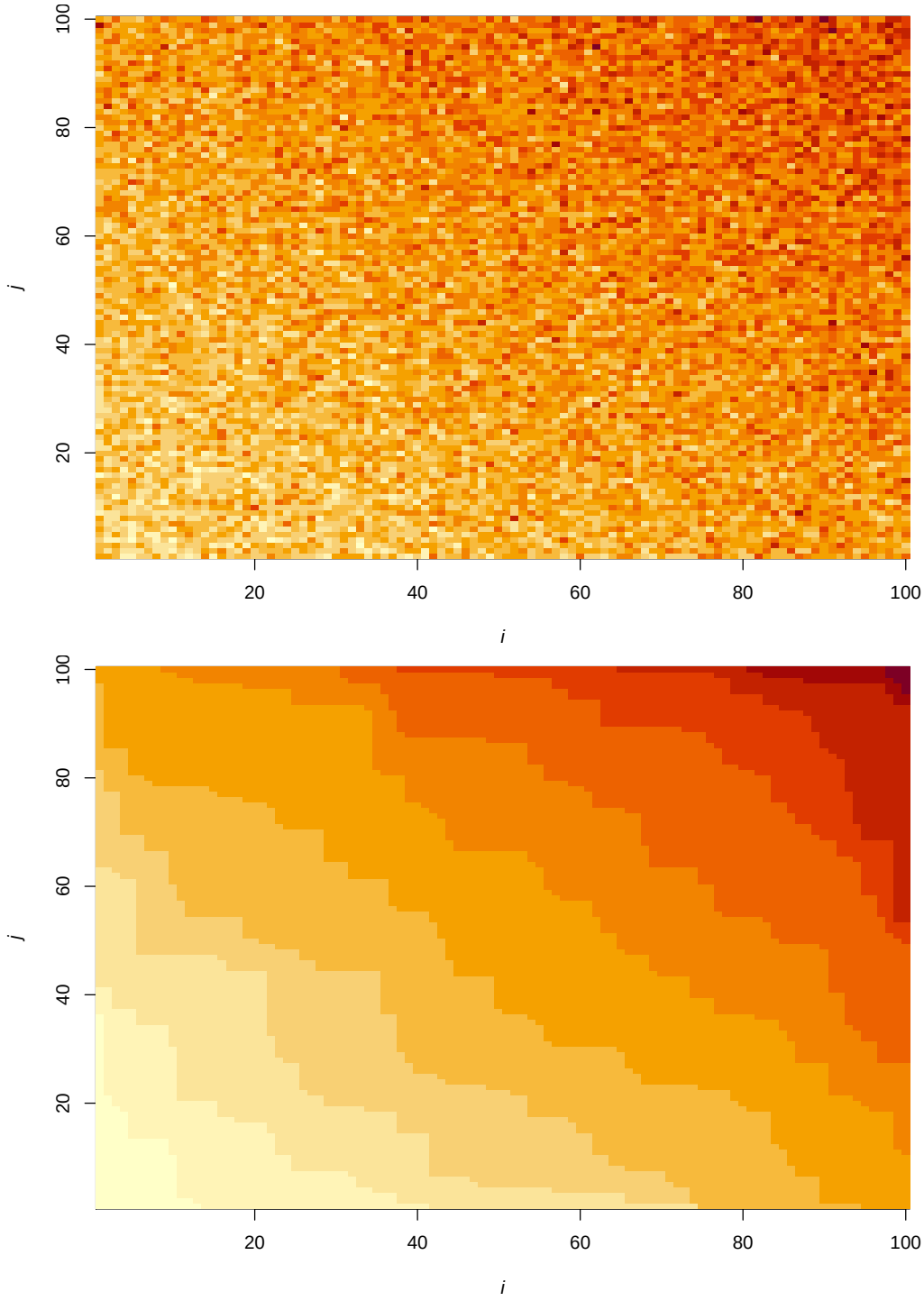

Figure 7.1: Example for bimonotone regression.