Lecture Notes
Nonlinear Optimization

Summer Semester 2024

Evelyn Herberg*

2024-04-26

*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany (evelyn.herberg@iwr.uni-heidelberg.de, https://scoop.iwr.uni-heidelberg.de/team/eherberg).

These lecture notes are based on previous lecture notes by Roland Herzog and Gerd Wachsmuth (BTU Cottbus).

Please send comments to evelyn.herberg@iwr.uni-heidelberg.de.

Contents

Chapter 1 Introduction

Mathematical optimization is about solving problems of the form

$$\left.\begin{array}{rlll} \text{Minimize} & f(x) & \text{where } x \in \Omega & \textbf{(objective function)} \\ \text{subject to} & g_i(x) \leq 0 & \text{for } i = 1, \ldots, n_{\text{ineq}} & \textbf{(inequality constraints)} \\ \text{and} & h_j(x) = 0 & \text{for } j = 1, \ldots, n_{\text{eq}}. & \textbf{(equality constraints)} \end{array}\right\} \qquad \textbf{(P)}$$

$\Omega \subseteq \mathbb{R}^n$ is the **basic set** and x is the **optimization variable** or simply the **variable** of the problem. We will assume that

- the functions $f, g_i, h_j \colon \mathbb{R}^n \to \mathbb{R}$ are sufficiently smooth (C^2 functions),
- we have a finite number (possibly zero) of inequality and equality constraints, i. e., n_{ineq} and n_{eq} are in $\mathbb{N}_0$.

We will assume $\Omega = \mathbb{R}^n$, i. e., we consider only **continuous optimization** problems and without implicit constraints.

§ 1 Elementary Notions

Definition 1.1 (Elementary notions)**.**

(*i*) *The set*

$$F := \left\{ x \in \mathbb{R}^n \,\middle|\, g_i(x) \leq 0 \text{ for all } i = 1, \ldots, n_{\text{ineq}},\ h_j(x) = 0 \text{ for all } j = 1, \ldots, n_{\text{eq}} \right\} \tag{1.1}$$

associated with an optimization problem (**P**) *is termed the* ***feasible set****. Any $x \in F$ is termed a* ***feasible point****.*

(*ii*) *The inequality $g_i(x) \leq 0$ is called* ***active*** *at a point x if $g_i(x) = 0$ holds. It is called* ***inactive*** *in case $g_i(x) < 0$. It is called* ***violated*** *if $g_i(x) > 0$ holds.*

(*iii*) *The value*

$$f^* := \inf \{ f(x) \mid x \in F \}$$

is termed the ***infimal value*** *of problem* (**P**)*.*

(*iv*) *In case $F = \emptyset$, the problem* (**P**) *is said to be* ***infeasible****. In that case, we have $f^* = +\infty$. In case $f^* = -\infty$, the problem is said to be* ***unbounded****.*

(v) *A point $x^* \in F$ is a* ***global minimizer*** *or* ***globally optimal solution*** *of* (**P**) *if*

$$f(x^*) \leq f(x) \text{ for all } x \in F$$

holds. Equivalently, $x^* \in F$ is a global minimizer if $f(x^*) = f^*$ holds. In this case, the infimal value f^* is also referred to as the ***global minimum*** *or* ***globally optimal value*** *of* (**P**).

(vi) *A global minimizer x^* is* ***strict*** *in case*

$$f(x^*) < f(x) \text{ for all } x \in F,\ x \neq x^*.$$

(vii) *A point $x^* \in F$ is a* ***local minimizer*** *or* ***locally optimal solution*** *of* (**P**) *if there exists a neighborhood $U(x^*)$ such that*

$$f(x^*) \leq f(x) \text{ for all } x \in F \cap U(x^*)$$

holds. In this case, $f(x^*)$ is also referred to as a ***local minimum*** *or a* ***locally optimal value*** *of* (**P**).

(viii) *A local minimizer x^* is* ***strict*** *in case*

$$f(x^*) < f(x) \text{ for all } x \in F \cap U(x^*), \quad x \neq x^*.$$

(ix) *An optimization problem* (**P**) *is* ***solvable*** *if it has at least one global minimizer, i. e., if the optimal value is attained at some point. Otherwise, the problem is* ***unsolvable****.*

Definition 1.2 (Classification of optimization problems)**.**

(i) *An optimization problem* (**P**) *is said to be* ***unconstrained*** *in case $n_{\text{ineq}} = n_{\text{eq}} = 0$. Otherwise, it is said to be* ***equality constrained*** *and/or* ***inequality constrained****.*

(ii) *Inequality constraints of the simple kind*

$$\ell_i \leq x_i \leq u_i, \quad i = 1, \dots, n$$

with bounds $\ell_i \in \mathbb{R} \cup \{-\infty\}$ and $u_i \in \mathbb{R} \cup \{\infty\}$ are called ***bound constraints*** *or* ***box constraints****.*

(iii) *When f, g and h are (affine) linear functions, then* (**P**) *is called a* ***linear optimization problem*** *or a* ***linear program*** *(**LP**).*

(iv) *When f is a quadratic polynomial and g and h are affine linear functions, then* (**P**) *is called a* ***quadratic optimization problem*** *or a* ***quadratic program*** *(**QP**).*

(v) *In the general case, i. e., when* (**P**) *is not a linear or quadratic program, we refer to* (**P**) *as a* ***nonlinear optimization problem*** *or* ***nonlinear program*** *(**NLP**).*

The emphasis in this class is on numerical techniques for unconstrained and constrained nonlinear programs. We will see that fast algorithms take into account the optimality conditions of the respective problem. Therefore we will also discuss optimality conditions.

We will begin in Chapter 2 with algorithms for unconstrained optimization. Some of the content was already part of the class *Grundlagen der Optimierung* (Herzog, 2022), but we will revisit the material in more detail here. The theory for constrained problems is relatively involved and merits its own chapter (Chapter 3). We will subsequently discuss major algorithmic ideas for constrained problems in Chapter 4. Finally, we will review in Chapter 5 some computer-aided techniques to obtain derivatives of functions, which the algorithms under consideration generally require.

Throughout the class, we will emphasize the connections between optimization and numerical linear algebra.

Chapter 2 Numerical Techniques for Unconstrained Optimization Problems

We discuss in this chapter numerical methods for the unconstrained version of (**P**), i. e.,

$$\text{Minimize} \quad f(x) \quad \text{where } x \in \mathbb{R}^n. \tag{UP}$$

The reason for discussing the unconstrained problem first is that we can introduce the essential algorithmic techniques without the difficulties of any constraints present.

Up front, we mention that we can only hope to find *local* minimizers. Determining *global* minimizers is generally much harder and only possible under additional assumptions on the objective, and generally only in relatively small dimensions $n \in \mathbb{N}$. A notable case of an additional assumption is that of a *convex* objective f. In this case, every local minimizer is already a global minimizer. Morever, the first-order optimality condition is already sufficient for optimality (see homework problem 1.2), and we do not require a second-order condition.

§ 2 Optimality Conditions

We suppose you have seen the following first- and second-order optimality conditions, so we only briefly recall them; see Herzog, 2022 for more details.

Theorem 2.1 (First-order necessary optimality condition).
Suppose that x^* is a local minimizer of (**UP**) *and that f is differentiable at x^*. Then $f'(x^*) = 0$.*

Proof. Suppose that $d \in \mathbb{R}^n$ is arbitrary. We consider the curve $\gamma\colon (-\delta, \delta) \to \mathbb{R}^n$, $\gamma(t) \coloneqq x^* + t\,d$. For sufficiently small $\delta > 0$, this curve runs within the neighborhood of local optimality of x^*. This implies that $f \circ \gamma$ has a local minimizer at $t = 0$.

From this local optimality, we infer that the difference quotient satisfies

$$\frac{f(\gamma(t)) - f(\gamma(0))}{t} = \frac{f(x^* + t\,d) - f(x^*)}{t} \begin{cases} \geq 0 & \text{for } t > 0, \\ \leq 0 & \text{for } t < 0. \end{cases}$$

On the other hand, this difference quotient converges to $f'(x^*)\,d$ as $t \to 0$. Consequently, we must have $f'(x^*)\,d = 0$. Since $d \in \mathbb{R}^n$ was arbitrary, this means $f'(x^*) = 0$. □

A point $x \in \mathbb{R}^n$ with the property $f'(x) = 0$ is termed a **stationary point** of f.

Theorem 2.2 (**Second-order necessary optimality condition**).
Suppose that x^* is a local minimizer of (**UP**) *and that f is twice differentiable at x^*. Then the Hessian $f''(x^*)$ is positive semidefinite.*[1]

Proof. Suppose that $d \in \mathbb{R}^n$ is arbitrary. Wie in Theorem 2.1 we define $\gamma(t) := x^* + t\,d$ and again consider the objective along the curve, i. e., $\varphi := f \circ \gamma$, which has a local minimizer at $t = 0$. Since φ is twice differentiable at $t = 0$, Theorem A.3 implies the following: for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left|\varphi(t) - \varphi(0) - \varphi'(0)\,t - \frac{1}{2}\varphi''(0)\,t^2\right| \leq \varepsilon\,t^2$$

holds for all $|t| < \delta$. In view of Theorem 2.1, $\varphi'(0) = 0$, and the local optimality implies $\varphi(0) \leq \varphi(t)$ for all $|t|$ sufficiently small. We thus obtain

$$-\frac{1}{2}\varphi''(0)\,t^2 \leq \varphi(t) - \varphi(0) - \frac{1}{2}\varphi''(0)\,t^2 \leq \varepsilon\,t^2$$

for all $|t|$ sufficiently small, whence

$$\frac{1}{2}\varphi''(0) \geq -\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we conclude $\varphi''(0) = d^\mathsf{T} f''(x^*)\,d \geq 0$. And since $d \in \mathbb{R}^n$ was arbitrary, we have shown $f''(x^*)$ to be positive semidefinite. □

Theorem 2.3 (**Second-order sufficient optimality condition**).
Suppose that f is twice differentiable at x^* and

(*i*) $f'(x^*) = 0$ *and*

(*ii*) $f''(x^*)$ *is positive definite*[2]*, with minimal eigenvalue $\alpha > 0$.*

Then for every $\beta \in (0, \alpha)$, there exists a neighborhood $U(x^*)$ of x^* such that

$$f(x) \geq f(x^*) + \frac{\beta}{2}\|x - x^*\|^2 \quad \text{for all } x \in U(x^*). \tag{2.1}$$

In particular, x^* is a strict local minimizer of f.

Proof. Here we use Theorem A.3 directly for f (not along a curve). For every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left|f(x^* + d) - f(x^*) - f'(x^*)\,d - \frac{1}{2}d^\mathsf{T} f''(x^*)d\right| \leq \varepsilon\,\|d\|^2$$

holds for all $\|d\| < \delta$. According to the assumptions, $f'(x^*) = 0$ holds. Therefore,

$$-\varepsilon\,\|d\|^2 \leq f(x^* + d) - f(x^*) - \frac{1}{2}d^\mathsf{T} f''(x^*)d$$

[1] Due to the symmetry of $f''(x^*)$ this is equivalent to all eigenvalues of $f''(x^*)$ being non-negative.
[2] Due to the symmetry of $f''(x^*)$ this is equivalent to all eigenvalues of $f''(x^*)$ being positive.

holds for all $\|d\| < \delta$. This implies

$$f(x^* + d) \geq f(x^*) + \frac{1}{2} d^\mathsf{T} f''(x^*)\, d - \varepsilon \|d\|^2$$

for all $\|d\| < \delta$.

From (A.12) (with $M = \mathrm{Id}$), the values of the Rayleigh quotient associated with the symmetric matrix $f''(x^*)$ are bounded above and below by the extremal eigenvalues of $f''(x^*)$. In particular, we have

$$d^\mathsf{T} f''(x^*)\, d \geq \alpha \|d\|^2 \quad \text{for all } d \in \mathbb{R}^n.$$

We can now finalize the proof: for $\beta \in (0, \alpha)$, choose $\varepsilon := (\alpha - \beta)/2 > 0$ and an appropriate value of $\delta > 0$. Then we have

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{1}{2} d^\mathsf{T} f''(x^*)\, d - \varepsilon \|d\|^2 \\ &\geq f(x^*) + \frac{\alpha}{2} \|d\|^2 - \varepsilon \|d\|^2 \\ &= f(x^*) + \frac{\beta}{2} \|d\|^2 \end{aligned}$$

for all $\|d\| < \delta$. □

Property (2.1) means that f has at least **quadratic growth** near x^*. Equivalently, f is locally strongly convex with parameter $\beta \in (0, \alpha)$.

§ 3 Minimization of Quadratic Functions

In this section we consider the simplest reasonable class of unconstrained optimization problems, namely the minimization of quadratic polynomials:

$$\text{Minimize} \quad \phi(x) := \frac{1}{2} x^\mathsf{T} A\, x - b^\mathsf{T} x + c \quad \text{where } x \in \mathbb{R}^n. \tag{3.1}$$

The data of the problem is $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. We can assume w.l.o.g. that A is symmetric.
Quiz 3.1: Why?

If we knew a spectral decomposition of $A = V \Lambda V^\mathsf{T}$ (which of course we usually don't), we could represent the objective as $\phi(x) = \frac{1}{2} x^\mathsf{T} V\, \Lambda\, V^\mathsf{T} x - b^\mathsf{T} V\, V^\mathsf{T} x + c$. After a substitution of variables $x = V^\mathsf{T} y$, this becomes $\widetilde{\phi}(y) = \frac{1}{2} y^\mathsf{T}\, \Lambda\, y - b^\mathsf{T} V\, y + c$. Consequently, in these coordinates, the problem decomposes into a sum of n independent quadratic minimization problems in the components y_i.

Being able to solve (3.1) is an essential building block for subsequent tasks.

Lemma 3.1 (Solvability and global solutions of (3.1)[3]). *Suppose that $A \in \mathbb{R}^{n \times n}$ is symmetric, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the following holds:*

[3] compare Nocedal, Wright, 2006, Lemma 4.7

(*i*) *If A is positive semidefinite, then the objective in (3.1) is convex. In this case, the following are equivalent:*

(a) The problem (3.1) possesses at least one (global) minimizer.

(b) The objective ϕ is bounded below.

(c) $Ax = b$ is solvable.

The global minimizers of (3.1) are precisely the solutions of the linear system $Ax = b$.

(*ii*) *In case A is not positive semidefinite[4], the objective ϕ is not bounded below, thus problem (3.1) is unbounded.*

Proof. The proof is part of homework problem 2.1. □

Corollary 3.2 (Unique solvability of (3.1)[5])**.** *Problem (3.1) possesses a unique (global) solution x^* if and only if A is s. p. d. In this case, $x^* = A^{-1}b$, and the optimal value is*

$$\phi(x^*) = c - \frac{1}{2}\|x^*\|_A^2 = c - \frac{1}{2}\|A^{-1}b\|_A^2 = c - \frac{1}{2}\|b\|_{A^{-1}}^2.$$

We will assume for the remainder of § 3 that A is symmetric and positive definite (s. p. d.). Hence, the solution of (3.1) is equivalent to the solution of the linear system $Ax = b$. We denote that solution by $x^* = A^{-1}b$. Of course, we could be using a **direct solver**, such as Gaussian elimination, which computes an LU decomposition of A, or rather its s. p. d. variant without pivoting, which computes the Cholesky decomposition $A = LL^\mathsf{T}$ with the lower triangular matrix L.[6] However, when the problem is high-dimensional (such as $n \geq 10\,000$), then the generic $\sim n^3$ effort for solving the linear system becomes prohibitive. Even when A is sparse, as is often the case for high-dimensional problems, and a direct solver which exploits this is used[7], this is no longer feasible for very high dimension n.

This is where **iterative solvers** for linear systems come into play. They do not solve the problem at once, but rather generate a sequence $\left(x^{(k)}\right)$ which converges to the solution. Beyond the ability to deal with very high-dimensional problems, iterative solvers have another advantage: Any iterate $x^{(k)}$ of the method can be viewed as an approximate solution of $Ax = b$ (or an approximate solution of (3.1)), and we can stop the iteration as soon as the desired tolerance is reached, when the time budget is used up, or when something unexpected happens, e. g., A turns out not to be positive definite after all. Recall that direct solvers do not yield any usable approximate solutions of the system while they are running; they have to carry through to the end, and only then return a solution, which is exact up to the influence of floating-point error. Iterative solvers have the additional advantage that they do not require access to the matrix A entry by entry. Rather they only require matrix-vector products,

[4] The matrix A possesses at least one negative eigenvalue.
[5] compare Nocedal, Wright, 2006, Lemma 4.7
[6] We assume you have seen these methods, e. g., in the class *Einführung in die Numerik*.
[7] such as a sparse Cholesky decomposition

i. e., a function which evaluates $x \mapsto Ax$. **Quiz 3.2:** Can you think of an example where matrix-vector products are available, but you typically don't have access to the entries of the underlying matrix?

Our objective ϕ from (3.1) satisfies

$$\begin{aligned}\phi(x) &= \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x + c\\ \nabla\phi(x) &= Ax - b =: r.\end{aligned}$$

We call $r = \nabla\phi(x)$ the **residual** of the linear system $Ax = b$ at x.[8] Independently of any method we might be using to solve $Ax = b$ (or minimize ϕ), we have the following relation between the values of the objective, the **error** $x - x^*$ at a point x, and the residual at x:

Lemma 3.3. *We have*

$$\phi(x) - \phi(x^*) = \frac{1}{2}\|x - x^*\|_A^2 = \frac{1}{2}\|r\|_{A^{-1}}^2 = \frac{1}{2}\|\nabla\phi(x)\|_{A^{-1}}^2. \tag{3.2}$$

Proof. Direct calculation shows

$$\begin{aligned}\phi(x) - \phi(x^*) &= \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x + c - \frac{1}{2}(x^*)^\mathsf{T}Ax^* + b^\mathsf{T}x^* - c\\ &= \frac{1}{2}x^\mathsf{T}Ax - (x^*)^\mathsf{T}Ax - \frac{1}{2}(x^*)^\mathsf{T}Ax^* + (x^*)^\mathsf{T}Ax^* \quad \text{since } b = Ax^*\\ &= \frac{1}{2}x^\mathsf{T}Ax - (x^*)^\mathsf{T}Ax + \frac{1}{2}(x^*)^\mathsf{T}Ax^*\\ &= \frac{1}{2}\|x - x^*\|_A^2\\ &= \frac{1}{2}(x - x^*)^\mathsf{T}r = \frac{1}{2}r^\mathsf{T}A^{-1}r \qquad \text{since } r = A(x - x^*)\\ &= \frac{1}{2}\|r\|_{A^{-1}}^2\\ &= \frac{1}{2}\|\nabla\phi(x)\|_{A^{-1}}^2.\end{aligned}$$

□

We will discuss in the remainder of this section two different iterative methods for the solution of (3.1), and equivalently the solution of the linear system $Ax = b$, where A is s. p. d.[9] These methods are the **gradient descent method** (also known as **steepest descent method**), and the **conjugate gradient method**.

We begin with the gradient descent method, which is based on the following simple
Idea: from the current iterate $x^{(k)}$, move a bit along the direction of steepest descent of the objective, and take the point reached as the next iterate $x^{(k+1)}$.

[8] Sometimes the residual is defined in the literature with opposite sign. We do not write $r(x)$ to keep the notation concise. It will be clear from the context which vector x the residual is associated with.

[9] You can learn more about iterative solvers for more general linear systems (not related to optimization) in the class *Numerische lineare Algebra*.

§ 3.1 Direction of Steepest Descent

We first need to clarify what **descent directions** and the **directions of steepest descent** of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ at a point x are.

Definition 3.4 (Descent direction)**.**
*Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$. A vector $d \in \mathbb{R}^n$ is termed a **descent direction** for f at x if*

$$f'(x)\, d < 0. \tag{3.3}$$

holds.

By definition, the direction of steepest descent minimizes the directional derivative $f'(x)\, d$ over all vectors $d \in \mathbb{R}^n$ of constant length. What we mean by "length" is defined through the inner product M in use:

$$\begin{aligned} &\text{Minimize} && f'(x)\, d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M = 1. \end{aligned} \tag{3.4}$$

We note that we could be considering the equivalent problem

$$\begin{aligned} &\text{Minimize} && f'(x)\, d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M \leq 1. \end{aligned} \tag{3.5}$$

The normalization to unit length is, by the way, arbitrary.

Problems (3.4), (3.5) are constrained problems, but we can solve them without an elaborated theory. We rewrite the objective so that the directional derivative is expressed using the M-inner product[10]

$$f'(x)\, d = \nabla f(x)^{\mathsf{T}} d = \nabla f(x)^{\mathsf{T}} M^{-1} M\, d = \left(M^{-1}\nabla f(x)\right)^{\mathsf{T}} M\, d,$$

where we used the symmetry of M (actually of M^{-1}) in the last step. The Cauchy-Schwarz inequality w.r.t. the M-inner product shows that this expression is minimal precisely when d is antiparallel to $M^{-1}\nabla f(x)$.

We summarize these findings:

Definition 3.5 (M-gradient, direction of steepest descent w.r.t. the M-inner product)**.**
Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$ and that $f'(x) \neq 0$ holds.

(*i*) *The vector*

$$\nabla_M f(x) := M^{-1}\nabla f(x) \tag{3.6}$$

*is termed the **gradient of f at x w.r.t. the M-inner product** or briefly: the **M-gradient**.*

(*ii*) *The vector $-\nabla_M f(x)$ and all of its positive multiples are termed the **directions of steepest descent** of f at x w.r.t. the M-inner product.*

[10] In case this means something to you, we determine the Riesz representer of $f'(x)$ w.r.t. the M-inner product.

We evaluate the negative M-gradient (direction of steepest descent) by solving the linear system

$$M\,d^* = -\nabla f(x). \tag{3.7}$$

When using the Euclidean inner product ($M = \mathrm{Id}$), we continue to write $\nabla f(x)$ instead of $\nabla_{\mathrm{Id}} f(x)$. Sometimes, the use of $\nabla_M f(x)$ instead of the Euclidean gradient direction $\nabla f(x)$ is referred to as **preconditioning**.

§ 3.2 Gradient Descent Method with Cauchy Step Sizes

The direction of steepest descent at x used by the gradient method is thus[11]

$$d = -\nabla_M \phi(x) = -M^{-1} r.$$

Now that the choice of direction is clear, let us analyze the choice of the step size. We have the following expression for the difference of function values before and after a step:

$$\begin{aligned}\phi(x+\alpha\,d) - \phi(x) &= \frac{1}{2}(x+\alpha\,d)^{\mathsf{T}} A\,(x+\alpha\,d) - b^{\mathsf{T}}(x+\alpha\,d) + c - \frac{1}{2}x^{\mathsf{T}} A\,x + b^{\mathsf{T}} x - c \\ &= \frac{1}{2}(d^{\mathsf{T}} A\,d)\,\alpha^2 + (A\,x - b)^{\mathsf{T}} d\,\alpha \\ &= \frac{1}{2}(d^{\mathsf{T}} A\,d)\,\alpha^2 + (r^{\mathsf{T}} d)\,\alpha. \end{aligned} \tag{3.8}$$

Note: This formula holds for arbitrary directions d and step sizes α.

When $d \neq 0$, then the one-dimensional quadratic polynomial $\alpha \mapsto \phi(x+\alpha\,d)$ is strongly convex. It is therefore an obvious idea to choose α such that $\phi(x+\alpha\,d)$ is minimized. According to (3.8), we have

$$\begin{aligned}\frac{\partial}{\partial\alpha}\phi(x+\alpha\,d) &= (d^{\mathsf{T}} A\,d)\,\alpha + r^{\mathsf{T}} d, \\ \frac{\partial^2}{\partial\alpha^2}\phi(x+\alpha\,d) &= d^{\mathsf{T}} A\,d > 0.\end{aligned}$$

End of Week 1

Due to the positivity of the second derivative, the second-order sufficient condition (Theorem 2.3) is satisfied when $\frac{\partial}{\partial\alpha}\phi(x+\alpha\,d) = 0$, which amounts to

$$\alpha^* = -\frac{r^{\mathsf{T}} d}{d^{\mathsf{T}} A\,d}. \tag{3.9}$$

Note: $\alpha^* = 0$ holds if and only if $r = 0$, i. e., the solution has been found.

[11] We avoid iteration indices for now in order to avoid cluttered notation.

This "optimal" step size is also known as the **Cauchy step size**. For this choice, the difference of function values (3.8) before and after a step becomes

$$\begin{aligned}\phi(x+\alpha^* d)-\phi(x) &= \frac{1}{2}(d^\mathsf{T} A\, d)\,(\alpha^*)^2 + (r^\mathsf{T} d)\,\alpha^* \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\left(\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d}\right)^2 - (r^\mathsf{T} d)\,\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d} \\ &= -\frac{1}{2}\frac{(r^\mathsf{T} d)^2}{d^\mathsf{T} A\, d}. \end{aligned} \tag{3.10}$$

Note: This formula holds for arbitrary directions $d \neq 0$ but it uses the Cauchy step size α^*.

We can now state the steepest descent method w.r.t. the M-inner product and the Cauchy step size (3.9) for the iterative solution of the unconstrained quadratic minimization problem (3.1) with s. p. d. A. This method, with $M = \mathrm{Id}$, was already published by Cauchy, 1847.

Algorithm 3.6 (Gradient descent method for (3.1) w.r.t. the M-inner product with Cauchy step size)**.**
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *right-hand side* $b \in \mathbb{R}^n$
Input: *s. p. d. matrix* A *(or matrix-vector products with* A*)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Output: *approximate solution of* (3.1)*, i. e., of* $Ax = b$

1: Set $k := 0$
2: Set $r^{(0)} := Ax^{(0)} - b$ // *evaluate the initial residual*
3: Set $d^{(0)} := -M^{-1}r^{(0)}$ // *evaluate the initial negative M-gradient*
4: Set $\delta^{(0)} := -(r^{(0)})^\mathsf{T} d^{(0)}$ // $\delta^{(0)} = \|\nabla_M \phi(x^{(0)})\|_M^2 = \|r^{(0)}\|_{M^{-1}}^2$
5: ***while*** *stopping criterion not met* ***do***
6: Set $q^{(k)} := Ad^{(k)}$
7: Set $\theta^{(k)} := (q^{(k)})^\mathsf{T} d^{(k)}$
8: Set $\alpha^{(k)} := \delta^{(k)}/\theta^{(k)}$ // *evaluate the Cauchy step size*
9: Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$ // *update the iterate*
10: Set $r^{(k+1)} := r^{(k)} + \alpha^{(k)} q^{(k)}$ // *update the residual*
11: Set $d^{(k+1)} := -M^{-1}r^{(k+1)}$ // *evaluate the negative M-gradient*
12: Set $\delta^{(k+1)} := -(r^{(k+1)})^\mathsf{T} d^{(k+1)}$ // $\delta^{(k+1)} = \|\nabla_M \phi(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}^2$
13: Set $k := k+1$
14: ***end while***
15: ***return*** $x^{(k)}$

The following can be said about Algorithm 3.6.

Remark 3.7 (on Algorithm 3.6)**.**

(*i*) *Algorithm 3.6 is an iterative solver for the unconstrained quadratic minimization problem* (3.1) *with s. p. d.* A*, and simultaneously an iterative solver for the linear system* $Ax = b$*.*

(*ii*) *We do not require access to the matrix* A *entry by entry, matrix-vector products with* A *are enough.*

(*iii*) *The user gets to choose the inner product M. This is known as* ***preconditioning****, and therefore Algorithm 3.6 is often termed a* ***preconditioned gradient descent method****. The case $M = \mathrm{Id}$ corresponds to the classical gradient descent method (without preconditioning).*

(*iv*) *We also do not require access to the inner product matrix M entry by entry, matrix-vector products with M^{-1} (i. e., solutions of linear systems with M) are enough.*

(*v*) *Algorithm 3.6 requires the storage of four vectors, which are iteratively overwritten: iterates $x^{(k)}$, residuals $r^{(k)}$, negative gradient directions $d^{(k)}$, and vectors $q^{(k)} = A\,d^{(k)}$.*

(*vi*) *Every iteration requires one matrix-vector product with A and one application of the preconditioner, i. e., one matrix-vector product with M^{-1}.*

(*vii*) *In order to mitigate the accumulation of round-off error, it is advisable to evaluate the residual every, say, 50 iterations according to $r^{(k)} := A\,x^{(k)} - b$, rather than update it.*

(*viii*) *The Cauchy step sizes satisfy*

$$0 < \lambda_{\min}(A;M) \leq \frac{1}{\alpha^{(k)}} = \frac{(d^{(k)})^{\mathsf{T}} A\,d^{(k)}}{(d^{(k)})^{\mathsf{T}} M\,d^{(k)}} \leq \lambda_{\max}(A;M), \tag{3.11}$$

as long as $d^{(k)} \neq 0$ holds, i. e., as long as $x^{(k)} \neq x^*$. Consequently, the Cauchy step sizes generated can be used to obtain estimates on the eigenvalues of A w.r.t. M.

(*ix*) *When Algorithm 3.6 is provided with the value of c, the following recursion can be added to the algorithm to keep track of the value of the objective:*

$$\phi(x^{(0)}) = c + \frac{1}{2}(r^{(0)} - b)^{\mathsf{T}}(x^{(0)}) \quad \textit{initialization} \tag{3.12a}$$

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \frac{1}{2}\,\alpha^{(k)}\delta^{(k)} \quad \textit{update.} \tag{3.12b}$$

This does not incur noticeable computational overhead and does not require the storage of extra vectors. Alternatively, the value of $\phi(x^{(0)})$ can be provided.

We now seek to estimate the speed of convergence of Algorithm 3.6. The function values at the iterates satisfy

$$\begin{aligned}
&\phi(x^{(k+1)}) - \phi(x^*) \\
&\quad = \frac{1}{2}\|r^{(k+1)}\|^2_{A^{-1}} && \text{by (3.2)} \\
&\quad = \frac{1}{2}\|r^{(k)} + \alpha^{(k)} A\,d^{(k)}\|^2_{A^{-1}} \\
&\quad = \frac{1}{2}\|r^{(k)}\|^2_{A^{-1}} + \alpha^{(k)}(r^{(k)})^{\mathsf{T}} d^{(k)} + \frac{1}{2}\left[\alpha^{(k)}\right]^2 (d^{(k)})^{\mathsf{T}} A\,d^{(k)}.
\end{aligned}$$

This formula so far holds for any choice of step size $\alpha^{(k)}$ and any choice of direction $d^{(k)}$. We now insert the Cauchy step size $\alpha^{(k)} = -\frac{(r^{(k)})^\mathsf{T} d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}}$ and obtain

$$
\begin{aligned}
&= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} + \frac{1}{2}\frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} \\
&= \left(1 - \frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{\left[(d^{(k)})^\mathsf{T} A\, d^{(k)}\right]\left[(r^{(k)})^\mathsf{T} A^{-1} r^{(k)}\right]}\right)\left(\phi(x^{(k)}) - \phi(x^*)\right) \qquad \text{by (3.2).}
\end{aligned}
$$

The directions $d^{(k)}$ are still arbitrary. Inserting the relationship $d^{(k)} = -M^{-1} r^{(k)} = -\nabla_M \phi(x^{(k)})$ characteristic for gradient descent, in the form $r^{(k)} = -M\, d^{(k)}$, we obtain

$$
= \left(1 - \frac{\left[(d^{(k)})^\mathsf{T} M\, d^{(k)}\right]^2}{\left[(d^{(k)})^\mathsf{T} A\, d^{(k)}\right]\left[(d^{(k)})^\mathsf{T} M A^{-1} M\, d^{(k)}\right]}\right)\left(\phi(x^{(k)}) - \phi(x^*)\right).
$$

The fraction is precisely the type of expression estimated by the generalized Kantorovich inequality (A.19), where $\alpha := \lambda_{\min}(A; M)$ and $\beta := \lambda_{\max}(A; M)$ are the extremal generalized eigenvalues of A w.r.t. M. This yields

$$
\begin{aligned}
&\phi(x^{(k+1)}) - \phi(x^*) \\
&\quad \le \left(1 - \frac{4\,\alpha\,\beta}{(\alpha+\beta)^2}\right)\left(\phi(x^{(k)}) - \phi(x^*)\right) \\
&\quad = \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^2 \left(\phi(x^{(k)}) - \phi(x^*)\right) \\
&\quad = \left(\frac{\kappa-1}{\kappa+1}\right)^2 \left(\phi(x^{(k)}) - \phi(x^*)\right) \qquad \text{since } \kappa = \beta/\alpha.
\end{aligned}
$$

We have thus shown the following classical convergence result for Algorithm 3.6:

Theorem 3.8 (Convergence of Algorithm 3.6)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ are both s. p. d., $\alpha := \lambda_{\min}(A; M)$ and $\beta := \lambda_{\max}(A; M)$ are the extremal generalized eigenvalues of A w.r.t. M. Then for any choice of the initial guess $x^{(0)}$, the gradient descent method with Cauchy step sizes converges to the unique solution $x^* = A^{-1}b$ of (3.1). In terms of the generalized condition number $\kappa = \beta/\alpha$, we have the estimates*

$$
\phi(x^{(k+1)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^2 \left(\phi(x^{(k)}) - \phi(x^*)\right) \tag{3.13a}
$$

$$
\|x^{(k+1)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right) \|x^{(k)} - x^*\|_A \tag{3.13b}
$$

and consequently

$$
\phi(x^{(k)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{3.13c}
$$

$$
\|x^{(k)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \|x^{(0)} - x^*\|_A. \tag{3.13d}
$$

Moreover, the objective values $\phi(x^{(k)})$ and thus the norm of the error $\|x^{(k)} - x^*\|_A$ are monotonically decreasing.

As an immediate consequence of this theorem, we can estimate the maximal number of iterations required until the left-hand terms in (3.13c) and (3.13d) have been decreased relative to their initial values.

Corollary 3.9 (Maximal number of iterations required in Algorithm 3.6)**.** *Given positive numbers ε_1 and ε_2, it takes*

$$k \leq \left\lceil \frac{\kappa}{4} \ln\left(\frac{1}{\varepsilon_1}\right) \right\rceil \text{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \leq \varepsilon_1,$$
$$k \leq \left\lceil \frac{\kappa}{2} \ln\left(\frac{1}{\varepsilon_2}\right) \right\rceil \text{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \leq \varepsilon_2.$$

Proof. For $\kappa = 1$ we have $\frac{\kappa-1}{\kappa+1} = 0$, i.e. $k \leq 1$. We now assume $\kappa > 1$.

(1) We first show that

$$-\ln\left(\frac{\kappa-1}{\kappa+1}\right) > \frac{2}{\kappa} > 0$$

holds for all $\kappa > 1$. At $\kappa = \frac{e+1}{e-1}$, we have

$$-\ln\left(\frac{\kappa-1}{\kappa+1}\right) = -\ln\left(\frac{1}{e}\right) = 1 > \frac{2}{\kappa} = 2\,\frac{e-1}{e+1} \approx 0.92.$$

Furthermore, we observe that

$$\lim_{\kappa\to\infty}\left[-\ln\left(\frac{\kappa-1}{\kappa+1}\right) - \frac{2}{\kappa}\right] = 0,$$

and for $\kappa > 1$

$$\frac{\partial}{\partial\kappa}\left[-\ln\left(\frac{\kappa-1}{\kappa+1}\right) - \frac{2}{\kappa}\right] = \frac{2}{\kappa^2-\kappa^4} < 0.$$

Hence, we can conclude that $\left(-\ln\left(\frac{\kappa-1}{\kappa+1}\right) - \frac{2}{\kappa}\right)$ for $\kappa > 1$ is approaching zero from above, which proves the claim.

(2) Taking the reciprocal of the inequality shown above, we obtain

$$0 < \frac{-1}{\ln\left(\frac{\kappa-1}{\kappa+1}\right)} \leq \frac{\kappa}{2} \tag{$*$}$$

for all $\kappa > 1$.

(3) Given $\kappa > 1$, we easily infer that $\left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \leq \varepsilon_1$ holds if and only if

$$k \geq \frac{1}{2}\frac{-\ln\varepsilon_1}{-\ln\left(\frac{\kappa-1}{\kappa+1}\right)} = \frac{1}{2}\frac{-1}{\ln\left(\frac{\kappa-1}{\kappa+1}\right)}\ln\left(\frac{1}{\varepsilon_1}\right). \tag{$**$}$$

In view of the inequality (∗) shown above, we obtain that

$$k \geq \left\lceil \frac{\kappa}{4} \ln\left(\frac{1}{\varepsilon_1}\right) \right\rceil \geq \frac{\kappa}{4} \ln\left(\frac{1}{\varepsilon_1}\right)$$

implies (∗∗), which proves the first claim.

The second claim follows similarly. □

Remark 3.10 (on Theorem 3.8)**.**

(*i*) (3.13b) *shows the Q-linear convergence of* $\left(x^{(k)}\right)$ *to the solution* x^* *in the A-norm.*

(*ii*) *The contraction factor is* $0 \leq \frac{\kappa-1}{\kappa+1} < 1$*, i. e., the convergence estimate depends on the ratio* κ *between the largest and the smallest generalized eigenvalue of A w.r.t. M. It is the purpose of the preconditioner/inner product M to keep this ratio small.*

(*iii*) *In the extreme case* $\kappa = 1$ *we obtain convergence in one step. This happens precisely when M is a multiple of A. However, we need a solve a linear system with M in every iteration. If we were able to do that, we might as well solve* $Ax = b$ *directly.*

(*iv*) *A good preconditioner is a compromise between a moderate generalized condition number* κ *and the effort in applying* M^{-1}*. Finding a good preconditioner generally requires knowledge about the problem at hand.*

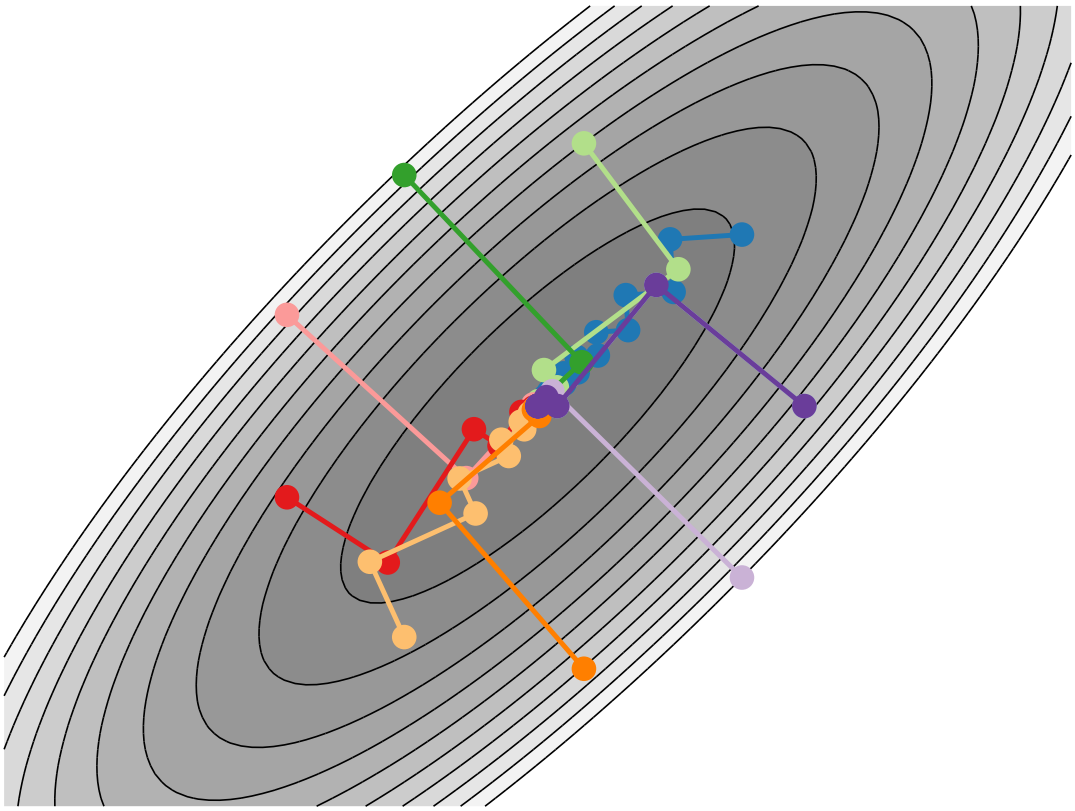
(*v*) *It is natural to measure convergence of the method in the A-norm of the error because, due to* (3.2)*, that is the quantity being minimized.*

(*vi*) *The estimates of Theorem 3.8 are worst-case estimates since they do not depend on the initial guess* $x^{(0)}$*. In fact, as can be seen in Figure 3.1c, the actual contraction factor for the objective values can be significantly smaller for some initial guesses than the estimate* (3.13c) *suggests.*

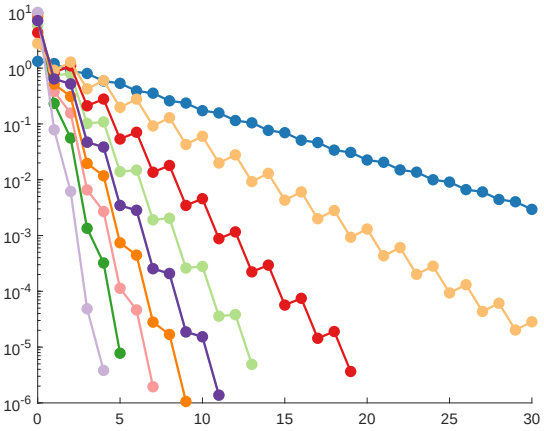
Figure 3.1 illustrates the convergence behavior of Algorithm 3.6 for a 2-dimensional example problem from a number of different initial guesses $x^{(0)}$. We observe the typical "zig-zagging" behavior of the iterates as they converge to the solution. This happens for any initial guess, except when $x^{(0)} - x^*$ happens to be a generalized eigenvector of A w.r.t. M, in which case convergence occurs in one step due to $x^{(1)} = x^*$. (Such a case is not shown in Figure 3.1).

The zig-zagging behavior of the iterates $x^{(k)}$, as well as the non-monotone behavior of $\|r^{(k)}\|_{M^{-1}}$ have been analyzed in detail in the literature; see for instance Akaike, 1959; Forsythe, 1968; Nocedal, Sartenaer, Zhu, 2002. Essentially what happens is that, asymptotically, the error $x^{(k)} - x^*$ alternates between elements of the eigenspaces belonging to the smallest and the largest eigenvalues of A w.r.t. M. This is ultimately a consequence of the fact that gradient descent is a memoryless method.

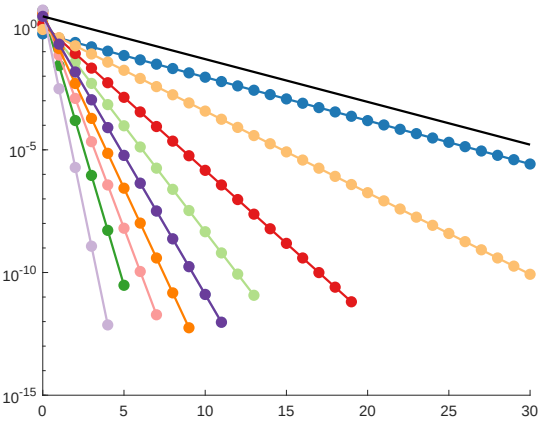
It has also been shown that a necessary condition in order for the norm of the gradient $\|r^{(k)}\|_{M^{-1}}$ to converge non-monotonically is that the condition number satisfy $\kappa > 3 + 2\sqrt{2} \approx 5.83$.



(a) Iterates $\left(x^{(k)}\right)$ of the method. Each color corresponds to a different initial guess $x^{(0)}$.



(b) The norm of the gradient $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_M = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.



(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (3.13c).

Figure 3.1: Illustration of the convergence behavior of Algorithm 3.6 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \mathrm{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the condition number is $\kappa = 10$.

It remains to discuss stopping criteria. Several quantities may be of interest in this respect:

(*i*) Are we happy with a point $x^{(k)}$ which is almost stationary, i. e., where $\|r^{(k)}\|_{M^{-1}}$ is small?

(*ii*) Are we happy with a point $x^{(k)}$ whose objective value is near the optimal value, i. e., where $\phi(x^{(k)}) - \phi(x^*)$ is small, or equivalently, where $\|x^{(k)} - x^*\|_A$ is small?

(*iii*) Are we happy with a point $x^{(k)}$ whose distance from the minimizer is small in the preconditioner-induced norm M, i. e., where $\|x^{(k)} - x^*\|_M$ is small?

Note: These criteria do not necessarily imply one another. Try to think of examples.

The only of these three quantities which we can evaluate without knowing x^* or $\phi(x^*)$ is $\delta^{(k)} = \|r^{(k)}\|^2_{M^{-1}}$. Therefore, many implementations use one of the following combinations of a relative and an absolute criterion based on $\|r^{(k)}\|_{M^{-1}}$:

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon^2_{\text{rel}}\, \delta^{(0)}, \tag{3.14a}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{abs}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon^2_{\text{abs}}, \tag{3.14b}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}} + \varepsilon_{\text{abs}}, \qquad \text{i. e., } (\delta^{(k)})^{1/2} \le \varepsilon_{\text{rel}}\, (\delta^{(0)})^{1/2} + \varepsilon_{\text{abs}}, \tag{3.14c}$$

$$\|r^{(k)}\|_{M^{-1}} \le \max\bigl\{\varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}},\ \varepsilon_{\text{abs}}\bigr\}, \qquad \text{i. e., } \delta^{(k)} \le \max\bigl\{\varepsilon^2_{\text{rel}}\, \delta^{(0)},\ \varepsilon^2_{\text{abs}}\bigr\}. \tag{3.14d}$$

Let us see which consequences either of the implementable stopping criteria (3.14) has on the other two quantities of interest:

Lemma 3.11 (Implications). *The criteria from* (3.14) *imply, respectively,*

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A \\ \|x^{(k)} - x^*\|_M &\le \kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M \end{aligned}\right\} \tag{3.15a}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le (1/\alpha)\, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{3.15b}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A + (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le \kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M + (1/\alpha)\, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{3.15c}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \max\bigl\{\sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A,\ (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}}\bigr\} \\ \|x^{(k)} - x^*\|_M &\le \max\bigl\{\kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M,\ (1/\alpha)\, \varepsilon_{\text{abs}}\bigr\} \end{aligned}\right\} \tag{3.15d}$$

Proof. The proof is part of homework problem 2.3. □

§ 3.3 Gradient Descent Method with Constant Step Sizes

We can show that the gradient descent method continues to converge Q-linearly when, in place of the Cauchy step sizes, we choose constant step sizes $\alpha^{(k)} \equiv \overline{\alpha}$ within a certain range. We obtain as

above

$$\begin{aligned}\phi(x^{(k+1)}) - \phi(x^*) \\ &= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\,(r^{(k)})^\mathsf{T} d^{(k)} + \frac{1}{2}\overline{\alpha}^2 (d^{(k)})^\mathsf{T} A\, d^{(k)}.\end{aligned}$$

We leave $\overline{\alpha}$ open for now and insert the gradient descent relation $r^{(k)} = -M\, d^{(k)}$ to obtain

$$\begin{aligned}&= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \overline{\alpha}\,(d^{(k)})^\mathsf{T} M\, d^{(k)} + \frac{1}{2}\overline{\alpha}^2 (d^{(k)})^\mathsf{T} A\, d^{(k)} \\ &\leq \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \overline{\alpha}\,(d^{(k)})^\mathsf{T} M\, d^{(k)} + \frac{1}{2}\overline{\alpha}^2 \beta\, (d^{(k)})^\mathsf{T} M\, d^{(k)} \quad \text{since } d^\mathsf{T} A\, d \leq \beta\, d^\mathsf{T} M\, d \\ &= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)(d^{(k)})^\mathsf{T} M\, d^{(k)}.\end{aligned}$$

Here we need to convert the last term into $d^\mathsf{T} M\, A^{-1} M\, d$, which is equal to $r^\mathsf{T} A^{-1} r$, so that it can be combined with the first term. We require that the coefficient $\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)$ is negative to obtain convergence. Consequently, we use the first estimate in (A.15a):

$$\begin{aligned}&\leq \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\,(d^{(k)})^\mathsf{T} M\, A^{-1} M\, d^{(k)} \quad \text{provided that } \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right) < 0 \\ &= \left[1 + 2\,\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\right]\frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 \\ &= \left[1 + 2\,\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\right]\big(\phi(x^{(k)}) - \phi(x^*)\big).\end{aligned}$$

The condition that $\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)$ is negative amounts to $\overline{\alpha} \in (0, \frac{2}{\beta})$.

Remark 3.12 (on the convergence of Algorithm 3.6 with constant step sizes)**.**

(*i*) *We have shown that Algorithm 3.6, where Line 8 is replaced by* $\alpha^{(k)} := \overline{\alpha}$*, still converges, provided that* $\overline{\alpha} \in (0, \frac{2}{\beta})$.

(*ii*) *From a practical perspective, we therefore need to know at least an upper bound for the largest eigenvalue* β *of the generalized eigenvalue problem* $A\,x = \lambda\, M\, x$*. When we have* $\beta \leq \beta_{\text{estimate}}$ *and choose* $\overline{\alpha} \in (0, \frac{2}{\beta_{\text{estimate}}})$*, we also have* $\overline{\alpha} \in (0, \frac{2}{\beta})$.

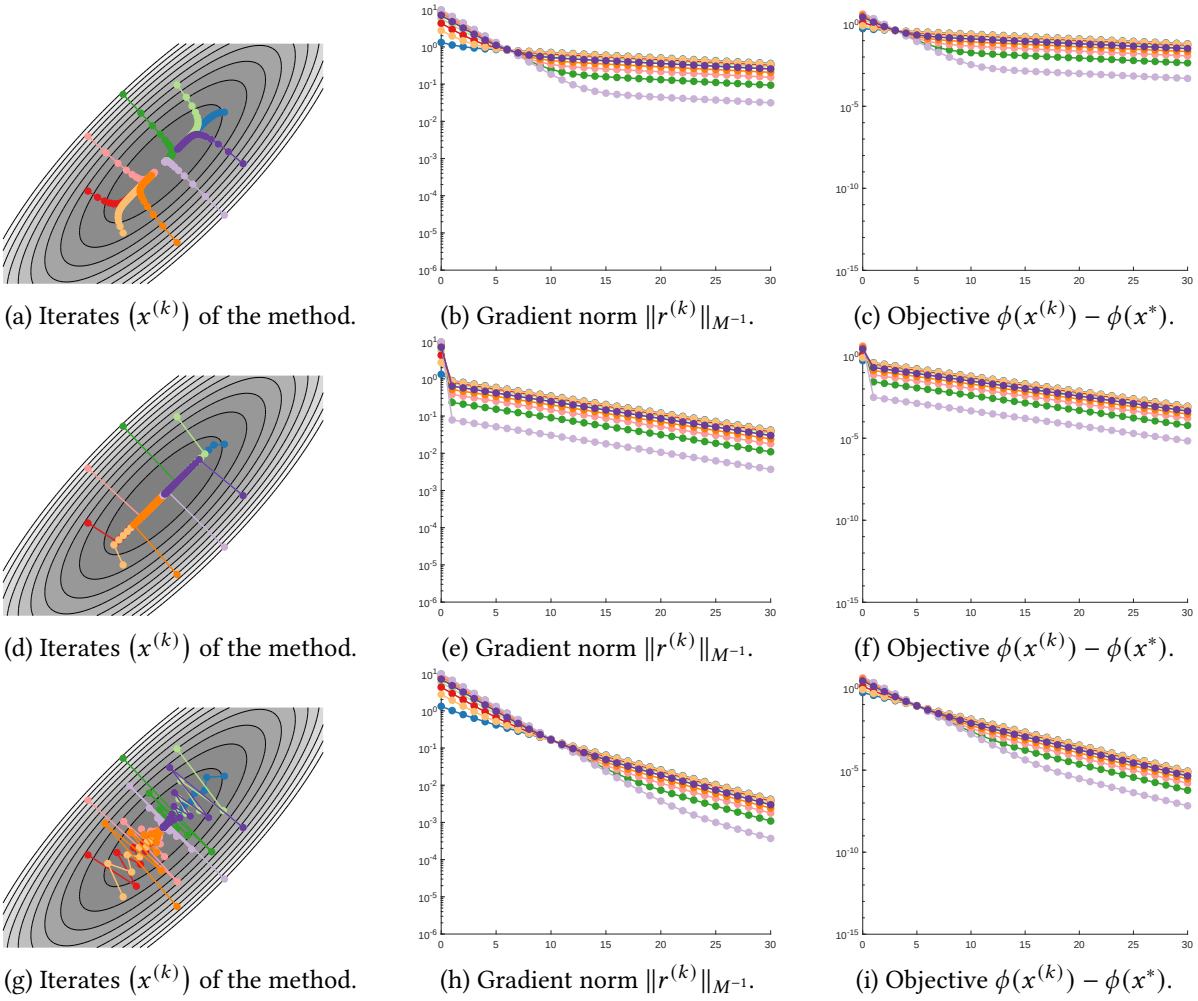
(*iii*) *The choice* $\overline{\alpha} = \frac{1}{\beta}$ *yields the optimal estimate. In this case, we obtain*

$$\phi(x^{(k+1)}) - \phi(x^*) \leq \left(\frac{\kappa - 1}{\kappa}\right)\big(\phi(x^{(k)}) - \phi(x^*)\big).$$

Since for all $\kappa \geq 1$*, we have* $\left(\frac{\kappa-1}{\kappa+1}\right)^2 \leq \frac{\kappa-1}{\kappa}$*, the contraction factor in the bound we obtained with constant step sizes is worse than the one for the Cauchy step sizes; see (3.13a). Consequently, there is no reason to prefer the gradient descent method with constant step sizes over the version with Cauchy step sizes.*

(*iv*) *The Kantorovich inequality was not needed in the proof.*

Figure 3.2 illustrates the convergence behavior of Algorithm 3.6 with constant step sizes for a 2-dimensional example problem from a number of different initial guesses $x^{(0)}$.



(a) Iterates $\left(x^{(k)}\right)$ of the method. (b) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (c) Objective $\phi(x^{(k)}) - \phi(x^*)$.

(d) Iterates $\left(x^{(k)}\right)$ of the method. (e) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (f) Objective $\phi(x^{(k)}) - \phi(x^*)$.

(g) Iterates $\left(x^{(k)}\right)$ of the method. (h) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (i) Objective $\phi(x^{(k)}) - \phi(x^*)$.

Figure 3.2: Illustration of the convergence behavior of Algorithm 3.6 with various constant step sizes instead of the Cauchy step size. The step sizes, from top to bottom, are $\overline{\alpha} \in \{0.03, 0.10, 0.17\}$. No preconditioning ($M = \mathrm{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the admissible range of constant step sizes is $\overline{\alpha} \in (0, \frac{2}{\beta}) = (0, 0.2)$.

§ 3.4 Gradient Descent Method with Other Step Size Rules

Step size rules other than the Cauchy step sizes and constant step sizes have been proposed and analyzed in the literature with the goal of breaking the non-efficient zig-zaggging pattern; among them Barzilai, Borwein, 1988; De Asmundis, di Serafino, Riccio, et al., 2013; De Asmundis, di Serafino, Hager, et al., 2014; Gonzaga, Schneider, 2015. We do not go into the details here but mention one remarkable result from Gonzaga, 2016, Theorem 1. Suppose that $\alpha := \lambda_{\min}(A;M)$ and $\beta := \lambda_{\max}(A;M)$ are the extremal generalized eigenvalues of A w.r.t. M, and $\kappa := \frac{\beta}{\alpha}$ is the generalized condition number. Suppose that $\kappa \geq 1.06$ and that

$$k := \left\lceil \sqrt{\kappa} \ln\left(\frac{2}{\varepsilon_1}\right) \right\rceil.$$

holds. Consider the set of mutually distinct, precomputed step sizes

$$\left\{ \alpha^{(j)} := \frac{1}{\omega^{(j)}} \,\middle|\, \omega^{(j)} := \frac{\beta-\alpha}{2} \cos\left(\frac{1+2j}{2k}\pi\right) + \frac{\beta+\alpha}{2},\ j = 0,1,\dots,k-1 \right\}.$$

Then the gradient descent method Algorithm 3.6 with step sizes $\alpha^{(k)}$, applied in any order, requires at most

$$k \text{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \leq \varepsilon_1.$$

The interesting fact is that, compared to the estimate of Corollary 3.9 for the Cauchy step size, the bound on the iteration numbers is proportional only to $\sqrt{\kappa}$, not to κ. The result can be modified so that it is not required to know the extremal eigenvalues exactly, but knowledge of an interval containing them is sufficient.

We are going to obtain a similar complexity result for the conjgate gradient method in § 3.6.

§ 3.5 Gradient Descent Method as Discretized Gradient Flow

We conclude the discussion of the gradient descent method by interpreting it in another way. Consider the differential equation

$$\begin{aligned} \dot{x}(t) &= -\nabla_M f(x(t)), \quad t \geq 0 \\ x(0) &= x^{(0)}. \end{aligned} \tag{3.16}$$

This is known as the **gradient flow** associated with f. Its stationary points are precisely the stationary points of f. Due to

$$\frac{\partial}{\partial t} f(x(t)) = f'(x(t))\,\dot{x}(t) = -f'(x(t))\,M^{-1}\nabla f(x(t)) = -\|\nabla f(x(t))\|^2_{M^{-1}} = -\|\nabla_M f(x(t))\|^2_M, \tag{3.17}$$

the value of f is decreasing along the path $x(t)$.

When we discretize (3.16) by the explicit (forward) Euler method with time step size $\Delta t^{(k)}$, we obtain

$$\frac{x^{(k+1)} - x^{(k)}}{\Delta t^{(k)}} = -M^{-1}\nabla f(x^{(k)}),$$

or equivalently,

$$x^{(k+1)} = x^{(k)} - \Delta t^{(k)}\, M^{-1}\nabla f(x^{(k)}). \tag{3.18}$$

This is precisely a step of the gradient descent method with step size $\Delta t^{(k)}$. Therefore, we can interpret the gradient descent method as a discretization of the continuous gradient flow equation.

§ 3.6 Conjugate Gradient Method

The typical inefficient zig-zaggging pattern of the directions $d^{(k)}$ is a consequence of the fact that gradient descent is a memoryless method. That is, we could restart the method at any iterate and it would produce the same iterates, whether restarted or not. This is where the **conjugate gradient method** (**CG method**, introduced in Hestenes, Stiefel, 1952) takes a different turn. It works with search directions $d^{(k)}$ which are pairwise A-orthogonal (also known as A-conjugate), and builds a memory of previously visited directions.

Definition 3.13 (Conjugate directions)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ is s. p. d. A set of non-zero vectors $\{d^{(0)}, \dots, d^{(k)}\} \subset \mathbb{R}^n$ is termed A-**conjugate** if*

$$(d^{(i)})^{\mathsf{T}} A\, d^{(j)} = 0 \quad \text{for } 0 \le i, j \le k, \quad i \ne j.$$

In other words, A-conjugate vectors are pairwise orthogonal w.r.t. the A-inner product. In particular, $\{d^{(0)}, \dots, d^{(k)}\}$ is a linearly independent set. (**Quiz 3.3:** Can you prove that?)

The CG method is a member of the class of **conjugate direction methods**. We begin by describing the properties of a generic conjugate direction method first before we particularize to the CG method. A conjugate direction method chooses its search directions $d^{(0)}, d^{(1)}, \dots$ so that they are A-conjugate, and the iterates satisfy

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}. \tag{3.19}$$

The step size $\alpha^{(k)}$ is the Cauchy step size, which minimizes the one-dimensional quadratic polynomial

$$\alpha \mapsto \phi(x^{(k)} + \alpha\, d^{(k)}).$$

That is, we have

$$\alpha^{(k)} := -\frac{(r^{(k)})^{\mathsf{T}} d^{(k)}}{(d^{(k)})^{\mathsf{T}} A\, d^{(k)}}, \tag{3.20}$$

compare (3.9). As in the gradient descent method, the residuals satisfy the recursion

$$r^{(k+1)} = r^{(k)} + \alpha^{(k)} A\, d^{(k)}. \tag{3.21}$$

End of Week 2

Conjugate direction methods have the remarkable property that the sequence of one-dimensional minimizations in the A-conjugate directions $d^{(0)}, d^{(1)}, \dots$ is equivalent to the minimization over the entire affine subspace $x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots\}$. This is shown in the following result.

Lemma 3.14 (Properties of conjugate direction methods). *Suppose that $A \in \mathbb{R}^{n\times n}$ is s. p. d. Given an initial guess $x^{(0)}$ and a set $\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$, $k \geq 1$ of A-conjugate search directions, suppose that the iterates $x^{(0)}, \dots, x^{(k)}$ are generated according to (3.19) with Cauchy step size (3.20). Then the following holds.*

(i)
$$(r^{(k)})^\mathsf{T} d^{(i)} = 0 \quad \text{for all } i = 0, 1, \dots, k-1. \tag{3.22}$$

(ii) $x^{(k)}$ *minimizes ϕ over the affine subspace* $x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$.

Proof. We can show Statement (*i*) via induction over k. For $k = 1$,

$$\begin{aligned}
(r^{(1)})^\mathsf{T} d^{(0)} &= (A x^{(1)} - b)^\mathsf{T} d^{(0)} && \text{by definition of the residual} \\
&= (A x^{(0)} + \alpha^{(0)} A d^{(0)} - b)^\mathsf{T} d^{(0)} && \text{by (3.19)} \\
&= (r^{(0)})^\mathsf{T} d^{(0)} + \alpha^{(0)} (d^{(0)})^\mathsf{T} A d^{(0)} && \text{by definition of the residual} \\
&= 0 && \text{since } \alpha^{(0)} \text{ is the Cauchy step size (3.20).}
\end{aligned}$$

The induction step assumes $(r^{(k-1)})^\mathsf{T} d^{(i)} = 0$ for all $i = 0, 1, \dots, k-2$ and proceeds as follows.

$$\begin{aligned}
(r^{(k)})^\mathsf{T} d^{(k-1)} &= (r^{(k-1)} + \alpha^{(k-1)} A d^{(k-1)})^\mathsf{T} d^{(k-1)} && \text{by the residual recursion (3.21)} \\
&= 0 && \text{since } \alpha^{(k-1)} \text{ is the Cauchy step size (3.20).}
\end{aligned}$$

For the remaining search directions $d^{(i)}$, $i = 0, 1, \dots, k-2$ we have

$$\begin{aligned}
(r^{(k)})^\mathsf{T} d^{(i)} &= \left(r^{(k-1)} + \alpha^{(k-1)} A d^{(k-1)}\right)^\mathsf{T} d^{(i)} && \text{by the residual recursion (3.21)} \\
&= \underbrace{(r^{(k-1)})^\mathsf{T} d^{(i)}}_{=0 \text{ by assumption}} + \alpha^{(k-1)} \underbrace{(d^{(k-1)})^\mathsf{T} A d^{(i)}}_{=0 \text{ due to } A\text{-conjugacy}} \\
&= 0.
\end{aligned}$$

For Statement (*ii*) we consider the function $h\colon \mathbb{R}^k \to \mathbb{R}$

$$h(\sigma) := \phi\left(x^{(0)} + \sum_{j=0}^{k-1} \sigma_j d^{(j)}\right).$$

h is strongly convex (**Quiz 3.4:** Why?), and the unique minimizer σ^* is characterized by

$$\frac{\partial h(\sigma^*)}{\partial \sigma_i} = \nabla\phi\left(x^{(0)} + \sum_{j=0}^{k-1} \sigma_j^* d^{(j)}\right)^\mathsf{T} d^{(i)} = 0, \quad i = 0, \dots, k-1. \tag{$*$}$$

However, we already know that it is the iterate

$$x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)} d^{(j)} \in x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$$

which satisfies (∗), since

$$\nabla\phi\Big(x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)}\, d^{(j)}\Big)^{\mathsf{T}} d^{(i)} = \nabla\phi(x^{(k)})^{\mathsf{T}} d^{(i)} = (r^{(k)})^{\mathsf{T}} d^{(i)} = 0$$

holds for all $i = 0, \dots, k-1$, as shown in Statement (i). □

Corollary 3.15 (Properties of conjugate direction methods)**.** *Any iterative method* (3.19) *using A-conjugate directions* $d^{(k)}$ *and Cauchy step sizes* (3.20) *converges to the unique solution of* (3.1) *in at most n steps.*

Proof. The search directions $d^{(k)}$ are A-conjugate and thus linearly independent. Therefore,

$$\operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(n-1)}\}$$

is all of $\mathbb{R}^n$, so that $x^{(n)}$ minimizes ϕ over all of $\mathbb{R}^n$ by Lemma 3.14. □

In practice, the statement of Corollary 3.15 is weakened by floating point error. Moreover, the result of Corollary 3.15 is not really relevant for high-dimensional problems since performing n iterations is prohibitively expensive. We will later see more practical convergence estimates.

There are many possibilities to generate pairwise A-conjugate directions $d^{(k)}$, each of which leads to a different conjugate direction method. The **conjugate gradient method** (**CG method**) determines the current direction $d^{(k)}$ as a linear combination of the previous direction $d^{(k-1)}$ and the current steepest descent direction $-M^{-1}r^{(k)}$:[12]

$$\begin{aligned} d^{(0)} &:= -M^{-1}r^{(0)} && \text{for } k = 0, \\ d^{(k)} &:= -M^{-1}r^{(k)} + \beta^{(k)}\, d^{(k-1)} && \text{for } k \geq 1. \end{aligned} \tag{3.23}$$

The coefficient $\beta^{(k)}$ is determined in such a way that at least $d^{(k)}$ and $d^{(k-1)}$ are A-conjugate:

$$\beta^{(k)} := \frac{(r^{(k)})^{\mathsf{T}} M^{-1} A\, d^{(k-1)}}{(d^{(k-1)})^{\mathsf{T}} A\, d^{(k-1)}}. \tag{3.24}$$

Interestingly, the algorithm obtained in this way generates search directions which are fully A-conjugate, as shown in the following result.

Lemma 3.16 (Properties of the iterates in the CG algorithm, see Nocedal, Wright, 2006, Theorem 5.3)**.** *Suppose that* $x^{(0)} \in \mathbb{R}^n$ *is given and that the search directions* $\{d^{(0)}, d^{(1)}, \dots, d^{(k)}\}$ *and the subsequent iterates* $x^{(1)}, \dots, x^{(k)}$, $k \geq 1$, *are generated according to* (3.19)–(3.20), (3.23)–(3.24)*, where* $\alpha^{(k)} \neq 0$.[13]

$$\operatorname{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k)}\} = \operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \dots, (A\,M^{-1})^k\, r^{(0)}\}, \tag{3.25}$$

$$\operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k)}\} = M^{-1} \operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \dots, (A\,M^{-1})^k\, r^{(0)}\}, \tag{3.26}$$

$$(d^{(k)})^{\mathsf{T}} A\, d^{(i)} = 0 \quad \textit{for all } i = 0, 1, \dots, k-1, \tag{3.27}$$

$$(r^{(k)})^{\mathsf{T}} M^{-1} r^{(i)} = 0 \quad \textit{for all } i = 0, 1, \dots, k-1. \tag{3.28}$$

[12] With $\beta^{(k)} = 0$, we obtain again the steepest descent method (Algorithm 3.6).

[13] $\alpha^{(k)} = 0$ would mean that $x^{(k)}$ is the unique solution x^*. Due to the form of the Cauchy step (3.20), this is clear for $k = 0$, as the nominator is $\|r^{(k)}\|_{M^{-1}}$. (3.22) shows that this is also true for $k > 0$.

The subspace

$$\mathcal{K}^{(k+1)}(A\,M^{-1}; r^{(0)}) := \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} \tag{3.29}$$

is termed the **Krylov subspace** (of order $k+1$) of the matrix $A\,M^{-1}$ with initial vector $r^{(0)}$. Therefore, the CG method is a representative of the class of **Krylov subspace methods**. The properties (3.25) and (3.26) imply that the method creates, simultaneously, an expanding sequence of M^{-1}-orthogonal basis vectors of the spaces $\mathcal{K}^{(k+1)}(A\,M^{-1}; r^{(0)})$, as well as an expanding sequence of A-orthogonal basis vectors of the spaces $M^{-1}\mathcal{K}^{(k+1)}(A\,M^{-1}; r^{(0)})$.

Proof. We first prove (3.25)–(3.27), by induction. For $k = 0$, statement (3.25) holds trivially. Statement (3.26) holds since the CG method starts with $d^{(0)} = -M^{-1}r^{(0)}$. Statement (3.27) is void for $k = 0$.

Suppose now that (3.25) and (3.26) have been shown up to some $k \geq 0$. We need to show that they also hold for $k+1$. By hypothesis,

$$\begin{aligned} r^{(k)} &\in \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\}, \\ d^{(k)} &\in M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\}, \\ \text{hence } A\,d^{(k)} &\in A\,M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} \\ &= \operatorname{span}\{(A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}. \end{aligned}$$

Due to the residual recursion (3.21), we therefore have

$$\begin{aligned} r^{(k+1)} &= r^{(k)} + \alpha^{(k)} A\,d^{(k)} \\ &\in \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} + \operatorname{span}\{(A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\} \\ &= \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}. \end{aligned} \tag{$*$}$$

Due to the induction hypothesis for (3.25), the same statement ($*$) holds when $k+1$ is replaced by a smaller index. Therefore, we have shown that

$$\operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}\} \subseteq \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}$$

holds. Now for the reverse inequality. By the induction hypothesis for (3.26), we find

$$A\,M^{-1}(A\,M^{-1})^k r^{(0)} \in A\,\operatorname{span}\{d^{(0)}, d^{(1)}, \ldots, d^{(k)}\} = \operatorname{span}\{A\,d^{(0)}, A\,d^{(1)}, \ldots, A\,d^{(k)}\}.$$

By the residual recursion (3.21), specifically

$$A\,d^{(i)} = \frac{1}{\alpha^{(i)}}\left(r^{(i+1)} - r^{(i)}\right) \in \operatorname{span}\{r^{(i)}, r^{(i+1)}\}$$

for $i = 0, 1, \ldots, k$, it follows that

$$A\,M^{-1}(A\,M^{-1})^k r^{(0)} \in \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}\}.$$

When combined with the induction hypothesis for (3.25), i. e.,

$$\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} = \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k)}\},$$

we find the desired reverse inequality

$$\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\} \subseteq \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}\}.$$

Thus the induction step for (3.25) is complete.

To see (3.26),

$$\begin{aligned}
&\operatorname{span}\{d^{(0)}, \ldots, d^{(k)}, d^{(k+1)}\} \\
&\quad = \operatorname{span}\{d^{(0)}, \ldots, d^{(k)}, M^{-1}r^{(k+1)}\} && \text{by (3.23)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k}\,r^{(0)}, r^{(k+1)}\} && \text{by (3.26)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k)}, r^{(k+1)}\} && \text{by (3.25)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k}\,r^{(0)}, (A\,M^{-1})^{k+1}\,r^{(0)}\} && \text{by (3.25) for } k+1.
\end{aligned}$$

This concludes the induction step for (3.26).

Next we address the A-conjugacy of search directions, (3.27). By the induction hypothesis, the directions $d^{(0)}, \ldots, d^{(k)}$ are pairwise A-conjugate. Consider

$$(d^{(k+1)})^{\mathsf{T}} A\,d^{(i)} = (-M^{-1}r^{(k+1)} + \beta^{(k+1)}\,d^{(k)})^{\mathsf{T}} A\,d^{(i)} \qquad (**)$$

for $i = 0, \ldots, k$. In case $i = k$, we have

$$(d^{(k+1)})^{\mathsf{T}} A\,d^{(k)} = 0$$

by construction of the search direction $d^{(k+1)}$, see (3.23) and (3.24). When $i \leq k-1$, we argue as follows. From (3.26), we obtain

$$\begin{aligned}
M^{-1}A\,d^{(0)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, d^{(1)}\}, \\
M^{-1}A\,d^{(1)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, d^{(1)}, d^{(2)}\}, \\
&\vdots \qquad \vdots && \vdots \\
M^{-1}A\,d^{(k-1)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}, \ldots, (A\,M^{-1})^{k-1}\,r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, \ldots, d^{(k)}\}.
\end{aligned}$$

We thus find that, for any $i \leq k-1$, the term $(r^{(k+1)})^{\mathsf{T}} M^{-1} A\,d^{(i)}$ in (**) belongs to

$$(r^{(k+1)})^{\mathsf{T}}\operatorname{span}\{d^{(0)}, \ldots, d^{(i+1)}\} = \operatorname{span}\{(r^{(k+1)})^{\mathsf{T}}d^{(0)}, \ldots, (r^{(k+1)})^{\mathsf{T}}d^{(i+1)}\}.$$

By (3.22), however, $(r^{(k+1)})^{\mathsf{T}}d^{(j)} = 0$ for $j = 0, \ldots, k$. Therefore, (**) reduces to

$$(d^{(k+1)})^{\mathsf{T}} A\,d^{(i)} = \beta^{(k+1)}\,(d^{(k)})^{\mathsf{T}} A\,d^{(i)}. \qquad (***)$$

By the induction hypothesis, this is equal to zero, which concludes the induction step for (3.27).

Finally, we consider the M^{-1}-conjugacy of residuals, (3.28), for $k \geq 1$. We do not need an induction argument for this. We consider two cases for $(r^{(k)})^{\mathsf{T}} M^{-1} r^{(i)}$:

(1) In case $i = k-1$, we have

$$(r^{(k)})^\mathsf{T} M^{-1} r^{(k-1)} = \begin{cases} \overbrace{(r^{(k-1)} + \alpha^{(k-1)} A\, d^{(k-1)})^\mathsf{T}(-d^{(k-1)} + \beta^{(k-1)}\, d^{(k-2)})}^{(\square)} & \text{for } k \geq 2 \\ \underbrace{(r^{(k-1)} + \alpha^{(k-1)} A\, d^{(k-1)})^\mathsf{T}(-d^{(k-1)})}_{(\square)} & \text{for } k = 1 \end{cases}$$

by the residual recursion (3.21) and the construction of search directions (3.23). Since the Cauchy step size satisfies $\alpha^{(k-1)} = -\frac{(d^{(k-1)})^\mathsf{T} r^{(k-1)}}{(d^{(k-1)})^\mathsf{T} A\, d^{(k-1)}}$, the term $(\square)$ is equal to zero for all $k \geq 1$. Let us consider the remaining terms when $k \geq 2$. We obtain

$$\begin{aligned} (r^{(k-1)})^\mathsf{T} d^{(k-2)} &= 0 \quad \text{due to (3.22)}, \\ (A\, d^{(k-1)})^\mathsf{T}(d^{(k-2)}) &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

Therefore we conclude that $(r^{(k)})^\mathsf{T} M^{-1} r^{(k-1)} = 0$ holds for all $k \geq 1$.

(2) in case $i < k-1$, we have

$$(r^{(k)})^\mathsf{T} M^{-1} r^{(i)} = \begin{cases} (r^{(k-1)} + \alpha^{(k-1)} A\, d^{(k-1)})^\mathsf{T}(-d^{(i)} + \beta^{(i)}\, d^{(i-1)}) & \text{for } i \geq 1 \\ (r^{(k-1)} + \alpha^{(k-1)} A\, d^{(k-1)})^\mathsf{T}(-d^{(i)}) & \text{for } i = 0 \end{cases}$$

When expanding, we obtain terms of the types (note $i < k-1$)

$$\begin{aligned} (r^{(k-1)})^\mathsf{T} d^{(i)} &= 0 \quad \text{due to (3.22)}, \\ (A\, d^{(k-1)})^\mathsf{T} d^{(i)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions,} \\ (r^{(k-1)})^\mathsf{T} d^{(i-1)} &= 0 \quad \text{due to (3.22)}, \\ (A\, d^{(k-1)})^\mathsf{T} d^{(i-1)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

Therefore we conclude that $(r^{(k)})^\mathsf{T} M^{-1} r^{(i)} = 0$ holds for all $k \geq 1$ and $0 \leq i < k-1$. □

Using the properties of the iterates shown above, the equations (3.20) for $\alpha^{(k)}$ as well as (3.24) for $\beta^{(k)}$ in the CG method can be equivalently formulated as follows:

$$\begin{aligned} \alpha^{(k)} &= -\frac{(r^{(k)})^\mathsf{T} d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the Cauchy step size formula (3.20)} \\ &= \frac{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} - \beta^{(k)} \frac{(r^{(k)})^\mathsf{T} d^{(k-1)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the search direction recursion (3.23)} \\ &= \frac{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by (3.22)} \qquad (3.20') \end{aligned}$$

and

$$\begin{aligned} \beta^{(k+1)} &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} A\, d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the orthogonalization coefficient (3.24)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} (r^{(k+1)} - r^{(k)})}{(d^{(k)})^\mathsf{T} (r^{(k+1)} - r^{(k)})} && \text{by the residual recursion (3.21)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} (r^{(k+1)} - r^{(k)})}{(-M^{-1} r^{(k)} + \beta^{(k)}\, d^{(k-1)})^\mathsf{T} (r^{(k+1)} - r^{(k)})} && \text{by the construction of search directions (3.23)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} r^{(k+1)}}{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}} && \text{by (3.22) and (3.25).} \qquad (3.24') \end{aligned}$$

The relations (3.20') and (3.24') are also true for $k = 0$.

We have now obtained the common form of the CG method w.r.t. the M-inner product, commonly referred to as the **preconditioned conjugate gradient method**.

Algorithm 3.17 (Conjugate gradient method for (3.1) w.r.t. the M-inner product)**.**
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *right-hand side* $b \in \mathbb{R}^n$
Input: *s. p. d. matrix* A *(or matrix-vector products with* A*)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Output: *approximate solution of (3.1), i. e., of* $Ax = b$

1: Set $k := 0$
2: Set $r^{(0)} := Ax^{(0)} - b$ // evaluate the initial residual
3: Set $d^{(0)} := -M^{-1}r^{(0)}$ // evaluate the initial negative M-gradient
4: Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$ // $\delta^{(0)} = \|\nabla_M \phi(x^{(0)})\|_M^2 = \|r^{(0)}\|_{M^{-1}}^2$
5: ***while*** *stopping criterion not met* ***do***
6: Set $q^{(k)} := Ad^{(k)}$
7: Set $\theta^{(k)} := (q^{(k)})^\intercal d^{(k)}$
8: Set $\alpha^{(k)} := \delta^{(k)}/\theta^{(k)}$ // evaluate the Cauchy step size
9: Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$ // update the iterate
10: Set $r^{(k+1)} := r^{(k)} + \alpha^{(k)} q^{(k)}$ // update the residual
11: Set $d^{(k+1)} := -M^{-1}r^{(k+1)}$ // evaluate the negative M-gradient
12: Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d^{(k+1)}$ // $\delta^{(k+1)} = \|\nabla_M \phi(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}^2$
13: Set $\beta^{(k+1)} := \delta^{(k+1)}/\delta^{(k)}$ // evaluate the A-orthogonalization coefficient
14: Set $d^{(k+1)} := d^{(k+1)} + \beta^{(k+1)} d^{(k)}$ // make $d^{(k+1)}$ A-orthogonal w.r.t. $d^{(k)}$
15: Set $k := k + 1$
16: ***end while***
17: ***return*** $x^{(k)}$

Remark 3.18 (on Algorithm 3.17)**.**

(*i*) *From Lemma 3.16 we know that the CG method generates pairwise A-orthogonal directions, although it only needs to orthogonalize any new direction* $d^{(k+1)}$ *against the most recent one,* $d^{(k)}$*. This phenomenon, known as* ***short-term recurrence****, is possible due to the symmetry of A.*

(*ii*) *The conjugate thus keeps a memory of previously visited directions, although this memory is mainly implicit. As shown in Algorithm 3.17, we can implement the method with a constant amount of storage.*

(*iii*) *The implementation of the CG method is very similar to the steepest descent method (Algorithm 3.6). The only (but significant!) difference lies in the fact that we A-orthogonalize the steepest descent direction against* $d^{(k)}$ *before we use it as the new search direction* $d^{(k+1)}$*. The initial search direction* $d^{(0)}$ *is the steepest descent direction for* ϕ *at* $x^{(0)}$*. Consequently, the iterate* $x^{(1)}$ *is the same for the conjugate gradient method and the steepest descent method with Cauchy step size (Algorithm 3.6).*

(*iv*) *The name* ***conjugate gradient method*** *is a bit of a misnomer, since it is not the gradients which are A-conjugate, but rather the search directions $d^{(k)}$.*

(*v*) *Remark 3.7 remains valid for the conjugate gradient method as well, with minor modifications. We need to store one additional vector since $d^{(k)}$ and $d^{(k+1)}$ are needed simultaneously.*

(*vi*) *The stopping criteria (3.14) and their consequences (3.15) continue to hold since they depend on the same computable quantity $\|r^{(k)}\|_{M^{-1}}$ as in the steepest descent method.*

Our next goal is to establish a convergence result for the conjugate gradient method, and to compare it to Theorem 3.8 for the steepest descent method with Cauchy step size. A major difference is that we will not obtain a result about the reduction of the error from iteration to iteration, but rather a result about the reduction of the error compared with its initial value.

Theorem 3.19 (Convergence of Algorithm 3.17, compare Theorem 3.8)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times n}$ are both s. p. d., $\alpha := \lambda_{\min}(A;M)$ and $\beta := \lambda_{\max}(A;M)$ are the extremal generalized eigenvalues of A w.r.t. M. Then for any choice of the initial guess $x^{(0)}$, the conjugate gradient method converges to the unique solution $x^* = A^{-1}b$ of (3.1). In terms of the generalized condition number $\kappa = \beta/\alpha$, we have the estimates*[14]

$$\phi(x^{(k)}) - \phi(x^*) \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{3.30a}$$

$$\|x^{(k)} - x^*\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k} \|x^{(0)} - x^*\|_A, \tag{3.30b}$$

Moreover, the objective values $\phi(x^{(k)})$ and thus the norm of the error $\|x^{(k)} - x^*\|_A$ are monotonically decreasing.

Proof. Since the search directions, by (3.26), span $M^{-1}\mathcal{K}^{(k)}(AM^{-1}; r^{(0)})$, we have

$$x^{(k)} - x^{(0)} \in M^{-1}\mathcal{K}^{(k)}(AM^{-1}; r^{(0)}).$$

In other words, we have

$$x^{(k)} - x^{(0)} = q^{(k-1)}(M^{-1}A)\, M^{-1}r^{(0)}$$

for some polynomial $q^{(k-1)}$ in the matrix $M^{-1}A$ of degree at most $k-1$. Abbreviating $e^{(k)} := x^{(k)} - x^*$ and using $Ae^{(0)} = Ax^{(0)} - Ax^* = r^{(0)}$, we can manipulate this equation into

$$\begin{aligned} e^{(k)} &= e^{(0)} + q^{(k-1)}(M^{-1}A)\, M^{-1}r^{(0)} \\ &= e^{(0)} + q^{(k-1)}(M^{-1}A)\, M^{-1}A\, e^{(0)} \\ &= \left[\mathrm{Id} + q^{(k-1)}(M^{-1}A)\, M^{-1}A\right] e^{(0)} \\ &= p^{(k)}(M^{-1}A)\, e^{(0)}, \end{aligned}$$

where now $p^{(k)}$ is a polynomial of degree at most k satisfying $p^{(k)}(0) = 1$.

[14] compare (3.13c), (3.13d)

By construction, the conjugate gradient method minimizes $\|e^{(k)}\|_A$ in every iteration. We can now express this in terms of a minimization over the vector space Π_k of polynomials of degree $\leq k$:

$$\|e^{(k)}\|_A = \min\left\{\|p(M^{-1}A)\, e^{(0)}\|_A \,\middle|\, p \in \Pi_k,\ p(0) = 1\right\}. \tag{3.31}$$

We expand the initial error $e^{(0)}$ in terms of the basis of eigenvectors of A w.r.t. M; see (A.10), (A.11). Suppose we denote the generalized eigenpairs by $(\lambda^{(j)}, v^{(j)})$, we can write

$$e^{(0)} = \sum_{j=1}^{n} \gamma^{(j)} v^{(j)}$$

with some coefficients $\gamma^{(j)}$ determined by $e^{(0)}$. We can thus manipulate the objective in the minimization problem above as follows:

$$\begin{aligned}\|p(M^{-1}A)\, e^{(0)}\|_A &= \left\|p(M^{-1}A)\left(\sum_{j=1}^{n} \gamma^{(j)} v^{(j)}\right)\right\|_A \\ &= \left\|\sum_{j=1}^{n} \gamma^{(j)} p(M^{-1}A)\, v^{(j)}\right\|_A\end{aligned}$$

In view of $A\, v^{(j)} = \lambda^{(j)} M\, v^{(j)}$ and thus $M^{-1}A\, v^{(j)} = \lambda^{(j)} v^{(j)}$, this is

$$= \left\|\sum_{j=1}^{n} \gamma^{(j)} p(\lambda^{(j)})\, v^{(j)}\right\|_A.$$

By pulling the maximal value of $|p(\lambda^{(j)})|$ out of the sum, we can estimate this quantity further:

$$\begin{aligned}&\leq \max_{j=1,\ldots,n} |p(\lambda^{(j)})| \left\|\sum_{j=1}^{n} \gamma^{(j)} v^{(j)}\right\|_A \\ &= \max_{j=1,\ldots,n} |p(\lambda^{(j)})| \left\|e^{(0)}\right\|_A.\end{aligned}$$

In detail, the inequality above can be seen as follows

$$
\begin{aligned}
&\Big\|\sum_{j=1}^{n} \gamma^{(j)} p(\lambda^{(j)})\, v^{(j)}\Big\|_A^2 \\
&\quad= \sum_{i,j=1}^{n} \gamma^{(i)}\gamma^{(j)} p(\lambda^{(i)})\, p(\lambda^{(j)})\, (v^{(i)})^{\mathsf{T}} A\, v^{(j)} \\
&\quad= \sum_{i,j=1}^{n} \gamma^{(i)}\gamma^{(j)} p(\lambda^{(i)})\, p(\lambda^{(j)})\, \lambda^{(j)} (v^{(i)})^{\mathsf{T}} M\, v^{(j)} && \text{using } A\, v^{(j)} = \lambda^{(j)} M\, v^{(j)} \\
&\quad= \sum_{j=1}^{n} \big[p(\lambda^{(j)})\big]^2 \big[\gamma^{(j)}\big]^2 \lambda^{(j)} (v^{(j)})^{\mathsf{T}} M\, v^{(j)} && \text{using the } M\text{-orthogonality of the eigenvectors} \\
&\quad\le \max_{j=1,\dots,n} \big[p(\lambda^{(j)})\big]^2 \sum_{j=1}^{n} \big[\gamma^{(j)}\big]^2 \lambda^{(j)} (v^{(j)})^{\mathsf{T}} M\, v^{(j)} && \text{using the positive definiteness of } M \\
&\quad= \max_{j=1,\dots,n} \big[p(\lambda^{(j)})\big]^2 \sum_{i,j=1}^{n} \gamma^{(i)}\gamma^{(j)}\, \lambda^{(j)} (v^{(i)})^{\mathsf{T}} M\, v^{(j)} \\
&\quad= \max_{j=1,\dots,n} \big[p(\lambda^{(j)})\big]^2 \sum_{i,j=1}^{n} \gamma^{(i)}\gamma^{(j)}\, (v^{(i)})^{\mathsf{T}} A\, v^{(j)} \\
&\quad= \max_{j=1,\dots,n} \big[p(\lambda^{(j)})\big]^2 \Big\|\sum_{j=1}^{n} \gamma^{(j)} v^{(j)}\Big\|_A^2
\end{aligned}
$$

Combining $\|p(M^{-1}A)\, e^{(0)}\|_A \le \max_{j=1,\dots,n}|p(\lambda^{(j)})|\, \big\|e^{(0)}\big\|_A$ with (3.31), we see

$$
\begin{aligned}
\|e^{(k)}\|_A &\le \min\Big\{ \max_{j=1,\dots,n} |p(\lambda^{(j)})|\, \|e^{(0)}\|_A \,\Big|\, p \in \Pi_k,\ p(0) = 1 \Big\} \\
&= \min\Big\{ \max_{j=1,\dots,n} |p(\lambda^{(j)})| \,\Big|\, p \in \Pi_k,\ p(0) = 1 \Big\}\, \|e^{(0)}\|_A
\end{aligned}
$$

and since the eigenvalues lie in the interval $[\alpha, \beta]$,

$$
\|e^{(k)}\|_A \le \min\Big\{ \max_{z\in[\alpha,\beta]} |p(z)| \,\Big|\, p \in \Pi_k,\ p(0) = 1 \Big\}\, \|e^{(0)}\|_A. \tag{3.32}
$$

We have thus estimated $\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A}$ by the smallest maximal absolute value any polynomial $p \in \Pi_k$ with $p(0) = 1$ can attain on the interval $[\alpha, \beta]$ spanning all generalized eigenvalues of A w.r.t. M.

The question about the *optimal* polynomial in (3.32) can be answered by Chebyshev polynomials; we refer you to Elman, Silvester, Wathen, 2014, Theorem 2.4 if you want to know more details. It turns out that the optimal value

$$
\min\Big\{ \max_{z\in[\alpha,\beta]} |p(z)| \,\Big|\, p \in \Pi_k,\ p(0) = 1 \Big\}
$$

depends only on $\kappa = \beta/\alpha$ and it is given by

$$\begin{aligned} &= 2 \left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \right]^{-1} \\ &\leq 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k. \end{aligned}$$

From there, we finally obtain

$$\|e^{(k)}\|_A \leq 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \|e^{(0)}\|_A,$$

which is precisely (3.30b). Squaring both sides and dividing by 2, we also obtain (3.30a). □

Corollary 3.20 (Maximal number of iterations required in Algorithm 3.17, compare Corollary 3.9)**.** *Given positive numbers ε_1 and ε_2, it takes*

$$k \leq \left\lceil \frac{\sqrt{\kappa}}{4} \ln\left(\frac{2}{\varepsilon_1}\right) \right\rceil \text{ iterations until } 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^{2k} \leq \varepsilon_1,$$
$$k \leq \left\lceil \frac{\sqrt{\kappa}}{2} \ln\left(\frac{2}{\varepsilon_2}\right) \right\rceil \text{ iterations until } 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^{k} \leq \varepsilon_2.$$

Proof. The proof is similar to Corollary 3.9 and it uses that

$$-\ln\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right) > \frac{2}{\sqrt{\kappa}} > 0$$

holds for all $\kappa \geq 1$. □

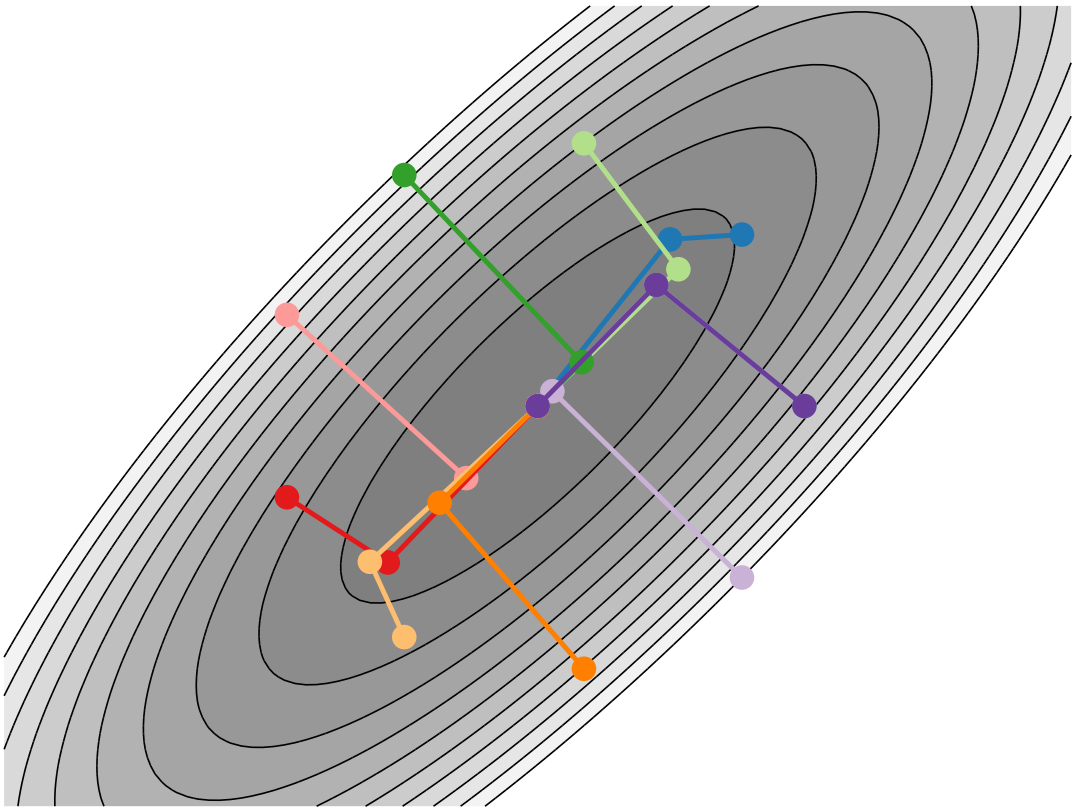
Remark 3.21 (on Theorem 3.19)**.**

(*i*) *The estimates* (3.30a) *and* (3.30b) *establish the R-linear convergence of the respective quantities to zero.*

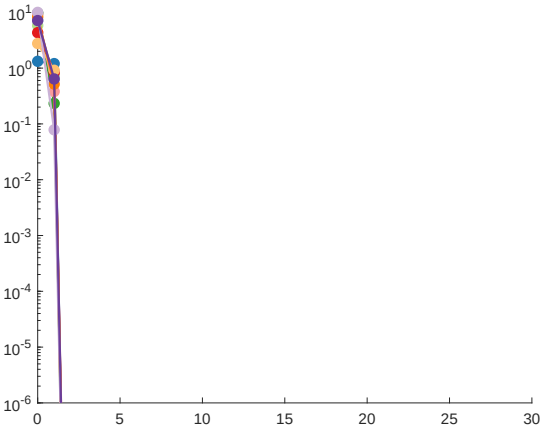
(*ii*) *Compared to the estimates* (3.13c) *and* (3.13d) *for the gradient descent method, we obtain the reduction factor* $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$ *in place of* $\left(\frac{\kappa-1}{\kappa+1}\right)^k$*, which is generally much better.*

(*iii*) *The superiority of the CG method compared to the gradient descent method is also reflected in the estimates for the maximal iteration numbers to achieve a certain reduction in the quantities* $\phi(x^{(k)}) - \phi(x^*)$ *and* $\|x^{(k)} - x^*\|_A$*, respectively. The bounds for the maximal iteration numbers are proportional to* $\sqrt{\kappa}$ *for the CG method, not proportional to* κ*.*

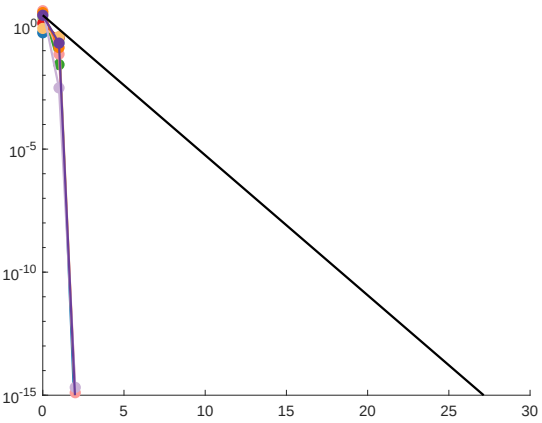
(*iv*) *As was the case for Theorem 3.8, the estimates of Theorem 3.19 are worst-case estimates since they do not depend on the initial guess* $x^{(0)}$*. In fact, as can be seen in Figure 3.3c and Figure 3.4b, the actual contraction factor for the objective values can be significantly smaller for some initial guesses than the estimate* (3.30a) *suggests.*



(a) Iterates $\left(x^{(k)}\right)$ of the method. Each color corresponds to a different initial guess $x^{(0)}$.

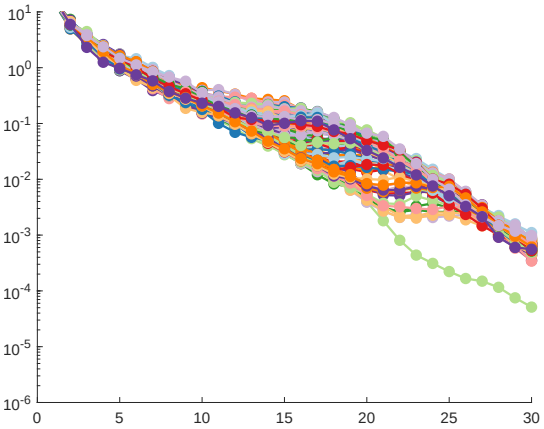


(b) The norm of the gradient $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_M = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.

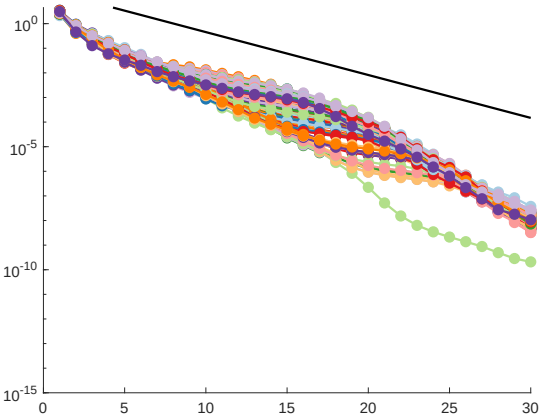


(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (3.30a).

Figure 3.3: Illustration of the convergence behavior of Algorithm 3.17 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \mathrm{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the condition number is $\kappa = 10$.



(a) The norm of the gradient $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_M = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.



(b) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (3.30a).

Figure 3.4: Illustration of the convergence behavior of Algorithm 3.17 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \mathrm{Id}$) is used. Here A is a random matrix of dimension 100×100 with eigenvalues in the interval $[\alpha, \beta] = [1, 100]$ so that the condition number is $\kappa = 100$.

(*v*) *Other informative error bounds than* (3.30) *and* (3.30b) *and convergence results can be obtained by proceeding as in the proof of Theorem 3.19 and choosing other polynomials to bound the error with.*

The iterates of the conjugate gradient method have a further remarkable property, which we will exploit later on:

Lemma 3.22 (Growth of the distance from the initial guess[15])**.** *Consider the iterates $x^{(k)}$ of the conjugate gradient method (Algorithm 3.17). As long as $x^{(k)} \neq x^*$ holds, the sequence $\|x^{(k)} - x^{(0)}\|_M$ is strictly increasing.*

Note: The steepest descent method does not have this property.

Proof. Statement (*i*) in Lemma 3.14 implies that

$$(r^{(k)})^\mathsf{T}(x^{(k)} - x^{(0)}) = \sum_{i=0}^{k-1} \alpha_i \underbrace{(r^{(k)})^\mathsf{T} d^{(i)}}_{=0} = 0 \quad \text{for all } k \geq 0. \tag{$*$}$$

We now show by induction that $(x^{(k)} - x^{(0)})^\mathsf{T} M d^{(k)} > 0$ holds for $k \geq 1$. Initially, for $k = 1$,

[15] In the literature, we find this result often only for the case $x^{(0)} = 0$, see for instance Nocedal, Wright, 2006, Theorem 7.3.

Statement (i) in Lemma 3.14 once again yields

$$\begin{aligned}(x^{(1)} - x^{(0)})^\mathsf{T} M\, d^{(1)} &= \alpha^{(0)} \overbrace{(d^{(0)})^\mathsf{T} M\,(-M^{-1} r^{(1)}}^{=0} + \beta^{(1)} d^{(0)}) \\ &= \underbrace{\alpha^{(0)}}_{>0}\ \underbrace{\beta^{(1)}}_{>0}\ \underbrace{(d^{(0)})^\mathsf{T} M\, d^{(0)}}_{>0} \\ &> 0.\end{aligned}$$

We now proceed with the step from index k to $k+1$:

$$\begin{aligned}(x^{(k+1)} - x^{(0)})^\mathsf{T} M\, d^{(k+1)} &= (x^{(k+1)} - x^{(0)})^\mathsf{T} M\,(-M^{-1} r^{(k+1)} + \beta^{(k+1)}\, d^{(k)}) \\ &= \beta^{(k+1)}\,(x^{(k+1)} - x^{(0)})^\mathsf{T} M\, d^{(k)} && \text{by } (*) \\ &= \beta^{(k+1)}\,(x^{(k)} + \alpha^{(k)} d^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} \\ &= \beta^{(k+1)}\,(x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} + \alpha^{(k)} \beta^{(k+1)}\,(d^{(k)})^\mathsf{T} M\, d^{(k)} \\ &> 0. && (**)\end{aligned}$$

Due to the induction hypothesis as well as $\alpha^{(k)} > 0$, $\beta^{(k+1)} > 0$ and $(d^{(k)})^\mathsf{T} M\, d^{(k)} > 0$, the entire expression is positive.

The desired result now easily follows from

$$\begin{aligned}\|x^{(k+1)} - x^{(0)}\|_M^2 &= \|x^{(k)} + \alpha^{(k)} d^{(k)} - x^{(0)}\|_M^2 \\ &= \|x^{(k)} - x^{(0)}\|_M^2 + 2\ \underbrace{\alpha^{(k)}}_{>0} \underbrace{(x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)}}_{>0} + \underbrace{(\alpha^{(k)})^2\, \|d^{(k)}\|_M^2}_{>0}. \qquad (***)\end{aligned}$$

□

The relations $(**)$ and $(***)$ allow us to compute the informative quantities

$$\omega^{(k)} := \|x^{(k)} - x^{(0)}\|_M^2 \tag{3.33a}$$
$$\xi^{(k)} := (x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} \tag{3.33b}$$
$$\gamma^{(k)} := \|d^{(k)}\|_M^2 \tag{3.33c}$$

on the side without any noticeable effort. This can be achieved by inserting, at the appropriate positions in Algorithm 3.17 (**Quiz 3.5:** Where?), the relations

$$\omega^{(0)} := 0, \qquad \omega^{(k+1)} := \omega^{(k)} + 2\,\alpha^{(k)} \xi^{(k)} + (\alpha^{(k)})^2\, \gamma^{(k)} \qquad \text{see } (***) \tag{3.34a}$$
$$\xi^{(0)} := 0, \qquad \xi^{(k+1)} := \beta^{(k+1)}\,(\xi^{(k)} + \alpha^{(k)} \gamma^{(k)}) \qquad \text{see } (**) \tag{3.34b}$$
$$\gamma^{(0)} := \delta^{(0)}, \qquad \gamma^{(k+1)} := \delta^{(k+1)} + (\beta^{(k+1)})^2\, \gamma^{(k)} \qquad \text{(confirm for yourself).} \tag{3.34c}$$

The remarkable fact about this is the possibility to keep track of (3.33) without requiring access to the matrix M, or even matrix-vector products with M. Notice that we usually do not have the latter since we only need matrix-vector products with M^{-1} in Algorithm 3.17.

Chapter 3 Theory for Constrained Optimization Problems

Chapter 4 Numerical Techniques for Constrained Optimization Problems

Chapter 5 Differentiation Techniques

Appendix A. Notation and Background Material

In these lecture notes we use color codes for **definitions** and highlights. The natural numbers are $\mathbb{N} = \{1, 2, \ldots\}$, and we write $\mathbb{N}_0$ for $\mathbb{N} \cup \{0\}$. We denote open intervals by (a, b) and closed intervals by $[a, b]$. We usually use Latin capital letters for matrices, Latin lowercase letters for vectors and Greek or Latin lowercase letters for scalars. We use Id for the identity matrix. We distinguish the vector space $\mathbb{R}^n$ of column vectors from the vector space $\mathbb{R}_n$ of row vectors.

A.1 Vector Norms

An **inner product** $(\cdot, \cdot)$ on $\mathbb{R}^n$ is a symmetric and positive definite bilinear form, i. e., a map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:

$$(x, y) = (y, x) \qquad \text{(symmetry)} \tag{A.1a}$$
$$(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_2, y) \qquad \text{(bilinearity part 1)} \tag{A.1b}$$
$$(x, \beta_1 y_1 + \beta_2 y_2) = \beta_1 (x, y_1) + \beta_2 (x, y_2) \qquad \text{(bilinearity part 2)} \tag{A.1c}$$
$$(x, x) \geq 0 \quad \text{and} \quad x \neq 0 \Rightarrow (x, x) > 0 \qquad \text{(positive definiteness)} \tag{A.1d}$$

for all $x, x_1, x_2, y, y_1, y_2 \in \mathbb{R}^n$ and all $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$.

Inner products on $\mathbb{R}^n$ are in one-to-one correspondence with symmetric and positive definite (s. p. d.) $n \times n$ matrices. That is, every s. p. d. matrix $M \in \mathbb{R}^{n \times n}$ induces an inner product

$$(x, y)_M := x^\mathsf{T} M\, y,$$

and, on the other hand, every inner product $(\cdot, \cdot)$ on $\mathbb{R}^n$ is induced by an s. p. d. matrix M. For simplicity, we will refer to M itself as the inner product it induces, or use the term "M-inner product".

Every inner product $(\cdot, \cdot)_M$ induces a norm[1] by way of

$$\|x\|_M := \sqrt{x^\mathsf{T} M\, x}. \tag{A.2}$$

In particular, the Euclidean inner product $x^\mathsf{T} y$ corresponds to the identity matrix $M = \mathrm{Id}$, and we denote the associated norm by $\|x\|$. We won't be writing $\langle x\,, y \rangle$ or $x \cdot y$ for the Euclidean inner product.

[1] We are only considering norms induced by inner products.

Notice that for vectors $x, y \in \mathbb{R}^n$, we have

$$\begin{aligned} a^\mathsf{T} b &= a^\mathsf{T} M^{-1} M\, b \\ &\leq \|M^{-1}a\|_M \|b\|_M \quad \text{by the Cauchy-Schwarz inequality w.r.t. the } M\text{-inner product} \\ &= \|a\|_{M^{-1}} \|b\|_M. \end{aligned} \tag{A.3}$$

A.2 Matrix Norms

A matrix $A \in \mathbb{R}^{m\times n}$ represents a linear map by way of $\mathbb{R}^n \ni x \mapsto A\,x \in \mathbb{R}^m$. When $\mathbb{R}^n$ is equipped with the M_1-inner product and $\mathbb{R}^m$ is equipped with the M_2-inner product, we define the **matrix norm** or **operator norm** of A as

$$\|A\|_{M_2 \leftarrow M_1} := \max_{x\neq 0} \frac{\|A\,x\|_{M_2}}{\|x\|_{M_1}}. \tag{A.4}$$

We thus have

$$\|A\,x\|_{M_2} \leq \|A\|_{M_2\leftarrow M_1}\, \|x\|_{M_1} \quad \text{for all } x \in \mathbb{R}^n. \tag{A.5}$$

When M_1 and M_2 are both the Euclidean inner products, $\|A\|_{\mathrm{Id}\leftarrow\mathrm{Id}}$ or simply $\|A\|$ is the largest singular value of A. In the general case, $\|A\|_{M_2\leftarrow M_1}$ is the largest singular value of a suitably generalized singular value decomposition.

There are matrix norm which are not operator norms. The most prominent one is induced by the inner product

$$A : B := \operatorname{trace}(A^\mathsf{T} B) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\, b_{ij}. \tag{A.6}$$

The associated norm

$$\|A\|_F := \Big(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\Big)^{1/2} \tag{A.7}$$

is termed the **Frobenius norm** of A.

A.3 Eigenvalues and Eigenvectors

Every symmetric matrix $A \in \mathbb{R}^{n\times n}$ possesses an orthogonal transformation to a diagonal matrix, known as **eigen decomposition** or **spectral decomposition**. That is, there exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$, such that

$$A\,V = V\Lambda, \quad \text{i. e.,} \quad A = V\Lambda V^\mathsf{T} \tag{A.8}$$

holds. The diagonal of Λ contains the eigenvalues λ_i, and the columns v_i of V are the corresponding eigenvectors. This decomposition yields the complete solution to the **eigenvalue problem**

$$A\,v = \lambda\, v. \tag{A.9}$$

We also work with the **generalized eigenvalue problem**

$$A\,v = \lambda\,M\,v \tag{A.10}$$

for the particular case where A is still symmetric and the second matrix $M \in \mathbb{R}^{n\times n}$ is s. p. d. There exists an analogous **generalized spectral decomposition**

$$A\,V = M\,V\Lambda, \quad \text{i. e.,} \quad A = M\,V\Lambda V^{\mathsf{T}}M, \tag{A.11}$$

where now V is orthogonal w.r.t. the M-inner product, i. e., $V^{\mathsf{T}}M\,V = \mathrm{Id}$ holds. We also refer to the solutions of (A.10) as the **eigenvalues/eigenvectors of A w.r.t. M** or **eigenvalues/eigenvectors of the pair $(A; M)$.**

In view of the Courant-Fischer theorem for (generalized) eigenvalues of symmetric matrices, the **generalized Rayleigh quotient** of A w.r.t. M satisfies

$$\lambda_{\min}(A;M) \le \frac{x^{\mathsf{T}}A\,x}{x^{\mathsf{T}}M\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0. \tag{A.12}$$

The eigenvectors associated with the smallest and largest generalized eigenvalues $\lambda_{\min}(A;M)$ and $\lambda_{\max}(A;M)$ satisfy the first respectively the second inequality with equality. Using (A.3) and (A.5), we also have

$$-\|A\|_{M^{-1}\leftarrow M} \le -\frac{\|x\|_M\|A\,x\|_{M^{-1}}}{\|x\|_M^2} \le \frac{x^{\mathsf{T}}A\,x}{\|x\|_M^2} \le \frac{\|x\|_M\|A\,x\|_{M^{-1}}}{\|x\|_M^2} \le \|A\|_{M^{-1}\leftarrow M}$$

and thus

$$\lambda_{\max}(H;M) \le \|H\|_{M^{-1}\leftarrow M} \quad \text{and} \quad -\lambda_{\min}(H;M) \le \|H\|_{M^{-1}\leftarrow M}. \tag{A.13}$$

Notice that the generalized eigenvalue problems (A.10) and

$$M\,v = \lambda\,M\,A^{-1}M\,v \tag{A.14a}$$

as well as

$$A\,M^{-1}A\,v = \lambda\,A\,v \tag{A.14b}$$

have the same eigenvalues and eigenvectors (provided that A is not only symmetric but also invertible) since $M\,v = \lambda\,M\,A^{-1}M\,v \Leftrightarrow v = \lambda\,A^{-1}M\,v \Leftrightarrow A\,v = \lambda\,M\,v$ and $A\,M^{-1}A\,v = \lambda\,A\,v \Leftrightarrow M^{-1}A\,v = \lambda\,v \Leftrightarrow A\,v = \lambda\,M\,v$. Consequently, we obtain the following estimate for the generalized Rayleigh quotients associated with (A.14):

$$\lambda_{\min}(A;M) \le \frac{x^{\mathsf{T}}M\,x}{x^{\mathsf{T}}M\,A^{-1}M\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0, \tag{A.15a}$$

$$\lambda_{\min}(A;M) \le \frac{x^{\mathsf{T}}A\,M^{-1}A\,x}{x^{\mathsf{T}}A\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0. \tag{A.15b}$$

Every s. p. d. matrix $A \in \mathbb{R}^{n\times n}$ possesses a unique s. p. d. **matrix square root** $A^{1/2}$. When $A = V\Lambda V^{\mathsf{T}}$ is a spectral decomposition of A with orthogonal V, then

$$A^{1/2} = V\Lambda^{1/2}V^{\mathsf{T}} \tag{A.16}$$

holds. Herein, $\Lambda^{1/2}$ is the elementwise square root of the diagonal matrix Λ.

A.4 Kantorovich Inequality

Suppose that A is an s. p. d. matrix. Let us denote the extremal eigenvalues by $\alpha := \lambda_{\min}(A)$ and $\beta := \lambda_{\max}(A)$. Moreover, since A is s. p. d., it follows that its **condition number**[2] is given by

$$\kappa := \frac{\beta}{\alpha}. \tag{A.17}$$

Notice that a condition number always satisfies $\kappa \geq 1$. From the Rayleigh quotient estimate (A.12) (with $M = \mathrm{Id}$), we have

$$\frac{x^\mathsf{T} A\, x}{\|x\|^2} \leq \beta.$$

Moreover, since the eigenvalues of A^{-1} are the reciprocals of those of A, we have $\lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A) = 1/\alpha$ and thus

$$\frac{x^\mathsf{T} A^{-1}\, x}{\|x\|^2} \leq \frac{1}{\alpha}.$$

These inequalities hold for all $x \in \mathbb{R}^n \setminus \{0\}$, and they imply

$$\frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1}\, x)}{\|x\|^4} \leq \frac{\beta}{\alpha}.$$

This estimate, however, is not sharp in general. (**Quiz A.1:** Can you explain why not?) The Kantorovich inequality improves this estimate.

Lemma A.1 (Kantorovich inequality)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ is s. p. d., $\alpha := \lambda_{\min}(A)$ and $\beta := \lambda_{\max}(A)$ are its extremal eigenvalues, and $\kappa = \beta/\alpha$ is its condition number. Then*

$$1 \leq \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|^4} \leq \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \leq \frac{\beta}{\alpha} \tag{A.18a}$$

holds for all $x \in \mathbb{R}^n \setminus \{0\}$, or equivalently, in terms of the condition number $\kappa = \beta/\alpha$,

$$1 \leq \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|^4} \leq \frac{(\kappa+1)^2}{4\,\kappa} \leq \kappa. \tag{A.18b}$$

Proof. The Cauchy-Schwarz inequality implies

$$\|x\|^2 = x^\mathsf{T} x = x^\mathsf{T} A^{-1/2}\, A^{1/2} x \leq \|A^{-1/2} x\|\, \|A^{1/2} x\|.$$

By squaring this, we obtain

$$\|x\|^4 \leq \|A^{-1/2} x\|^2\, \|A^{1/2} x\|^2 = (x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)$$

and thus the lower bound in (A.18).

[2] Generally, the condition of an invertible matrix A is $\kappa = \|A\|\,\|A^{-1}\|$. This is equal to $\sigma_{\max}(A)/\sigma_{\min}(A)$ with the extremal singular values $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$. Since A is symmetric, its singular values are just the absolute values of its eigenvalues, and since A is also positive definite, we have $\sigma_{\max}(A) = \lambda_{\max}(A) = \beta$ and $\sigma_{\min}(A) = \lambda_{\min}(A) = \alpha$.

From here on, the proof follows Anderson, 1971, as reproduced in the Master's thesis Alpargu, 1996, Section 1.2.2. Let $\lambda_1, \dots, \lambda_n > 0$ be the eigenvalues of A (in any order), and let $v_1, \dots, v_n$ be an orthonormal set of associated eigenvectors. We represent $x \in \mathbb{R}^n \setminus \{0\}$ as $x = \sum_{i=1}^n \gamma_i v_i$. Suppose, w.l.o.g., that $\|x\|^2 = \sum_{i=1}^n \gamma_i^2 = 1$ holds. Inserting the representation of x yields

$$\frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|^4} = \underbrace{\Big[\sum_{i=1}^n \lambda_i\, \gamma_i^2\Big]}_{=\mathbb{E}(T)}\ \underbrace{\Big[\sum_{i=1}^n \frac{1}{\lambda_i}\, \gamma_i^2\Big]}_{=\mathbb{E}(1/T)}.$$

It is helpful to think about the two factors on the right-hand side as expected values of a "random variable" T and $1/T$, respectively. Here T takes the values $\lambda_i \in [\alpha, \beta]$ with "probability" γ_i^2. For any $0 < \alpha \le T \le \beta$, we can estimate

$$0 \le (\beta - T)\,(T - \alpha) = (\beta + \alpha - T)\,T - \alpha\,\beta,$$

and thus

$$\frac{1}{T} \le \frac{\alpha + \beta - T}{\alpha\,\beta}.$$

Taking the expected value, this implies

$$\begin{aligned}\mathbb{E}(T)\,\mathbb{E}(1/T) &\le \mathbb{E}(T)\,\frac{\alpha + \beta - \mathbb{E}(T)}{\alpha\,\beta} \\ &= \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} - \frac{1}{\alpha\,\beta}\Big[\mathbb{E}(T) - \frac{1}{2}(\alpha+\beta)\Big]^2 \\ &\le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta}.\end{aligned}$$

This shows that essential upper bound in (A.18). The remaining inequality follows directly from $0 < \alpha \le \beta$. □

Instead of the Euclidean norm, we can also use the norm induced by the M-inner product.

Corollary A.2 (Generalized Kantorovich inequality)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ and M are both s. p. d., $\alpha := \lambda_{\min}(A; M)$ and $\beta := \lambda_{\max}(A; M)$ are the extremal generalized eigenvalues of A w.r.t. M. Then*

$$1 \le \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} M\, A^{-1} M\, x)}{\|x\|_M^4} \le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{A.19a}$$

holds for all $x \in \mathbb{R}^n \setminus \{0\}$, or equivalently, in terms of the ***generalized condition number*** *$\kappa = \beta/\alpha$,*

$$1 \le \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|_M^4} \le \frac{(\kappa+1)^2}{4\,\kappa} \le \kappa. \tag{A.19b}$$

We do not give a proof of Corollary A.2 here; see for instance Herzog, 2022.

A.5 Functions and Derivatives

- Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$, the derivative of the partial function $t \mapsto f(x + t\,e^{(i)})$ at $t = 0$ is the i-th **partial derivative** of f at x, briefly: $\frac{\partial}{\partial x_i} f(x)$. Here $e^{(i)} = (0, \dots, 0, 1, 0, \dots, 0)^\mathsf{T}$ is one of the standard basis vectors of $\mathbb{R}^n$. In other words,

$$\frac{\partial}{\partial x_i} f(x) = \lim_{t \to 0} \frac{f(x + t\,e^{(i)}) - f(x)}{t}.$$

- More generally, the derivative of the function $t \mapsto f(x + t\,d)$ at $t = 0$ is the **(two-sided) directional derivative** of f at x in the direction $d \in \mathbb{R}^n$, briefly:

$$\frac{\partial}{\partial d} f(x) = \lim_{t \to 0} \frac{f(x + t\,d) - f(x)}{t}.$$

- The right-sided derivative of the function $t \mapsto f(x + t\,d)$ at $t = 0$ is the **(one-sided) directional derivative** of f at x in the direction $d \in \mathbb{R}^n$, briefly:

$$f'(x; d) = \lim_{t \searrow 0} \frac{f(x + t\,d) - f(x)}{t}.$$

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **differentiable** at $x \in \mathbb{R}^n$ if there exists a row vector $v \in \mathbb{R}_n$ such that

$$\frac{f(x + d) - f(x) - v\,d}{\|d\|} \to 0 \quad \text{for } d \to 0.$$

  In this case, the vector v is the **(total) derivative** of f at x, and it is denoted by $f'(x)$.

- When $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then

$$f'(x) = \left(\frac{\partial f(x)}{\partial x_1}, \quad \cdots, \quad \frac{\partial f(x)}{\partial x_n} \right) \in \mathbb{R}_n.$$

  The transposed vector (a column vector)

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = f'(x)^\mathsf{T} \in \mathbb{R}^n$$

  is the **gradient** (w.r.t. the Euclidean inner product) of f at x.

- When $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then

$$f'(x; d) = \frac{\partial}{\partial d} f(x) = f'(x)\,d$$

  holds for all $d \in \mathbb{R}^n$. That is, the one-sided and two-sided directional derivatives of f at x agree, and they can be evaluated by applying the derivative $f'(x)$ to the direction d.

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **continuously partially differentiable** or briefly: $C^1(\mathbb{R}^n, \mathbb{R})$, if all partial derivatives $\frac{\partial f(x)}{\partial x_i}$, as functions of x, are continuous. C^1-functions are differentiable, and the derivative f' is continuous.

- A vector-valued function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $x \in \mathbb{R}^n$ if all component function $F_1, \dots, F_m$ are differentiable at x. In this case, the derivative $F'(x)$ is given by the **Jacobian** of F at x, i. e., by
$$\begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \dots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} & \dots & \frac{\partial F_m(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

- F is **continuously partially differentiable** if all entries of the Jacobian are continuous as functions of x. C^1-functions are differentiable, and the derivative F' is continuous.

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **twice differentiable** at $x \in \mathbb{R}^n$ if f is differentiable in a neighborhood of x and the derivative $x \mapsto f'(x) \in \mathbb{R}^n$ is differentiable at x. In this case, the second derivative $f''(x)$ is given by the **Hessian** of f at x, i. e., by the matrix of second-order partial derivatives
$$\left(\frac{\partial^2 f(x)}{\partial x_i\, \partial x_j}\right)_{i,j=1}^n = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$
When f is twice differentiable at x, then the Hessian is symmetric by Schwarz' theorem.[3]

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **twice continuously partially differentiable** or briefly: $C^2(\mathbb{R}^n, \mathbb{R})$, if all entries of the Hessian are continuous as functions of x. C^2-functions are twice differentiable.

A.6 Taylor's Theorem

We are going to state Taylor's theorem in two variants:

Theorem A.3 (Taylor, see Cartan, 1971, Theorem 5.6.3). *Suppose that $G \subseteq \mathbb{R}^n$ open, $k \in \mathbb{N}_0$ and $f\colon G \to \mathbb{R}$ k times differentiable, and $(k+1)$ times differentable at $x^{(0)} \in G$. Then for all $\varepsilon > 0$, there exists $\delta > 0$ such that*
$$\text{in case } k = 0: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\, d\right| \le \varepsilon \|d\|,$$
$$\text{in case } k = 1: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\, d - \frac{1}{2} d^\top f''(x^{(0)}) d\right| \le \varepsilon \|d\|^2.$$
for all $\|d\| < \delta$.

[3] See for instance Cartan, 1971, Proposition 5.2.2

Theorem A.4 (Taylor, see Geiger, Kanzow, 1999, Satz A.2 or Heuser, 2002, Satz 168.1)**.**
Suppose that $G \subseteq \mathbb{R}^n$ is open, $k \in \mathbb{N}_0$ and $f : G \to \mathbb{R}$ $(k+1)$ times continuously partially differentiable, briefly a $C^{k+1}(G, \mathbb{R})$ function. Suppose that $x^{(0)}$ and $x^{(0)} + d$ and the entire line segment between them lie in G. Then there exists $\xi \in (0,1)$ such that

$$\text{in case } k = 0: \quad f(x^{(0)} + d) = f(x^{(0)}) + f'(x^{(0)} + \xi\, d)\, d \quad (\textbf{\textit{mean value theorem}}),$$
$$\text{in case } k = 1: \quad f(x^{(0)} + d) = f(x^{(0)}) + f'(x^{(0)})\, d + \frac{1}{2} d^{\mathsf{T}} f''(x^{(0)} + \xi\, d)\, d.$$

A.7 Convergence Rates

We denote (vector-valued) sequences $\mathbb{N} \to \mathbb{R}^n$ by $\left(x^{(k)}\right)$ and not (x_k) etc., in order to avoid a conflict of notation with the components of a vector $x = (x_1, \ldots, x_n)^{\mathsf{T}} \in \mathbb{R}^n$. The **subsequence** of $\left(x^{(k)}\right)$ obtained by the strictly increasing sequence $\mathbb{N} \ni \ell \mapsto k^{(\ell)} \in \mathbb{N}$ is denoted by $\left(x^{(k^{(\ell)})}\right)$.

We introduce various convergence rates for sequences in order to characterize the speed of convergence, e. g., of iterates in an algorithm.

Definition A.5 (Q-convergence rates[4])**.**
Suppose that $\left(x^{(k)}\right) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let M be an inner product on $\mathbb{R}^n$.

(i) $\left(x^{(k)}\right)$ *converges to x^* (at least)* ***Q-linearly*** *w.r.t. the M-norm if there exists $c \in (0,1)$ such that*

$$\|x^{(k+1)} - x^*\|_M \leq c\, \|x^{(k)} - x^*\|_M \quad \text{for all } k \in \mathbb{N} \text{ sufficiently large.}$$

(ii) $\left(x^{(k)}\right)$ *converges to x^* (at least)* ***Q-superlinearly*** *w.r.t. the M-norm if there exists a null sequence $\left(\varepsilon^{(k)}\right)$ such that*

$$\|x^{(k+1)} - x^*\|_M \leq \varepsilon^{(k)}\, \|x^{(k)} - x^*\|_M \quad \text{for all } k \in \mathbb{N}.$$

(iii) Suppose that $x^{(k)} \to x^*$. $\left(x^{(k)}\right)$ converges to x^* (at least) ***Q-quadratically*** *w.r.t. the M-norm if there exists $C > 0$ such that*

$$\|x^{(k+1)} - x^*\|_M \leq C\, \|x^{(k)} - x^*\|_M^2 \quad \text{for all } k \in \mathbb{N}.$$

Note: Q-superlinear and Q-quadratic convergence of a sequence are independent of the norm (inner product) M. However, the property of Q-linear convergence can be lost when changing the norm.

Definition A.6 (R-convergence rates[5])**.**
Suppose that $\left(x^{(k)}\right) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let M be an inner product on $\mathbb{R}^n$.

[4] "Q" stands for "quotient".
[5] "R" stands for "root".

(*i*) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-linearly*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \leq \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-linearly w.r.t.* $|\cdot|$.

(*ii*) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-superlinearly*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \leq \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-superlinearly w.r.t.* $|\cdot|$.

(*iii*) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-quadratically*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \leq \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-quadratically w.r.t.* $|\cdot|$.

Note: The R-convergence modes are slightly weaker than the respective Q-convergence rates. Q-convergence considers the decrease in the distance to the limit $\|x^{(k)} - x^*\|_M$ in every step of the sequence. By contrast, R-convergence considers the decrease overall.

A.8 Convexity

Convexity plays a very important role in optimization in general. In this class, however, we will rely on it only scarcely. We briefly recall here some elements of convexity. You may study Herzog, 2022 if you wish to have more background information.

Definition A.7 (Convex set)**.**
A set $C \subseteq \mathbb{R}^n$ *is termed* ***convex*** *if* $x, y \in C$ *and* $\alpha \in [0,1]$ *imply* $\alpha x + (1-\alpha)\, y \in C$.

The condition in Definition A.7 means that the entire line segment between x and y belongs to C.

Definition A.8 (Convex function)**.**
A function $f : \mathbb{R}^n \to \mathbb{R}$ *is termed*

(*i*) ***convex*** *in case*

$$f(\alpha x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{A.20}$$

holds for all $x, y \in \mathbb{R}^n$ *and* $\alpha \in [0,1]$.

(*ii*) ***strictly convex*** *in case*

$$f(\alpha x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{A.21}$$

holds for all $x, y \in \mathbb{R}^n$ *and* $\alpha \in (0,1)$.

(*iii*) *μ-**strongly convex** or **strongly convex** with parameter $\mu > 0$ in case*

$$f(\alpha\, x + (1-\alpha)\, y) + \frac{\mu}{2}\,\alpha\,(1-\alpha)\|x-y\|^2 \leq \alpha f(x) + (1-\alpha) f(y) \tag{A.22}$$

holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

(*iv*) ***concave** (concave) or **strictly concave** or **constrly concave** if $-f$ is convex or strictly convex or strongly convex, respectively.*

Theorem A.9 (Characterization of convexity via first-order derivatives)**.**
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable.

(a) The following are equivalent:

(*i*) *f is convex.*

(*ii*) *For all $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) \geq f'(y)(x-y) \tag{A.23}$$

holds.

(*iii*) *For all $x, y \in \mathbb{R}^n$,*

$$\big(f'(x) - f'(y)\big)(x-y) \geq 0 \tag{A.24}$$

*holds. Equation (A.24) means that f' is a **monotone operator**.*

(b) The following are equivalent:

(*i*) *f ist strictly convex.*

(*ii*) *For all $x, y \in \mathbb{R}^n$ such that $x \neq y$,*

$$f(x) - f(y) > f'(y)(x-y) \tag{A.25}$$

holds.

(*iii*) *For all $x, y \in \mathbb{R}^n$ such that $x \neq y$,*

$$\big(f'(x) - f'(y)\big)(x-y) > 0. \tag{A.26}$$

*Equation (A.26) means that f' is a **strictly monotone operator**.*

(c) The following are equivalent:

(*i*) *f ist strongly convex.*

(*ii*) *There exists $\mu > 0$ such that for all $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) \geq f'(y)(x-y) + \frac{\mu}{2}\,\|x-y\|^2 \tag{A.27}$$

holds.

(iii) *There exists $\mu > 0$ such that for all $x, y \in \mathbb{R}^n$,*

$$\big(f'(x) - f'(y)\big)(x - y) \geq \mu \|x - y\|^2. \tag{A.28}$$

Equation (A.28) means that f' is a ***strongly monotone operator****.*

Theorem A.10 (Characterization of convexity via second-order derivatives)**.**
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.

(a) *The following are equivalent:*

(i) *f ist convex.*

(ii) *f'' is everywhere positive semidefinite (has only non-negative eigenvalues).*

(b) *When f'' is everywhere positive definite, then f is strictly convex.*

(c) *The following are equivalent:*

(i) *f is strongly convex with parameter $\mu > 0$.*

(ii) *The smallest eigenvalue of $f''(x)$ satisfies $\lambda_{\min}(f''(x)) \geq \mu > 0$ for all $x \in \mathbb{R}^n$.*

A.9 Hyperplanes and Half Spaces

Suppose that $a \in \mathbb{R}^n$, $a \neq 0$ and $\beta \in \mathbb{R}$. Then the set

$$H(a, \beta) := \{x \in \mathbb{R}^n \mid a^\mathsf{T} x = \beta\} \tag{A.29}$$

is termed a **hyperplane** in $\mathbb{R}^n$ with **normal vector** a.

A hyperplane separates $\mathbb{R}^n$ into two closed **half spaces**

$$\begin{aligned} H^-(a, \beta) &:= \{x \in \mathbb{R}^n \mid a^\mathsf{T} x \leq \beta\} \quad \textbf{negative half space}, \\ H^+(a, \beta) &:= \{x \in \mathbb{R}^n \mid a^\mathsf{T} x \geq \beta\} \quad \textbf{positive half space}. \end{aligned} \tag{A.30}$$

A.10 Miscellanea

We denote the **interior** of a set $S \subseteq \mathbb{R}^n$ by $\operatorname{int} S$ and its **closure** by $\operatorname{cl} S$.

Given $\varepsilon > 0$ and $x \in \mathbb{R}^n$,

$$B_\varepsilon^M(\overline{x}) := \left\{x \in \mathbb{R}^n \,\middle|\, \|x - \overline{x}\|_M < \varepsilon\right\}$$

denotes the **open ε-ball** w.r.t. the M-norm about $\overline{x}$ (centered at $\overline{x}$). Similarly, the **closed ε-ball** is

$$\operatorname{cl} B_\varepsilon^M(\overline{x}) := \left\{ x \in \mathbb{R}^n \,\middle|\, \|x - \overline{x}\|_M \leq \varepsilon \right\}.$$

A **neighborhood** of a point $\overline{x} \in \mathbb{R}$ is a set containing some open ball centered at $\overline{x}$. We often write $U(\overline{x})$ for such a neighborhood.

The **ceiling function** $\lceil x \rceil$ returns the smallest integer $\geq x$.

Bibliography

Akaike, H. (1959). "On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method". *Annals of the Institute of Statistical Mathematics* 11, pp. 1–16. DOI: 10.1007/bf01831719.

Alpargu, G. (1996). "The Kantorovich Inequality, with Some Extensions and with Some Statistical Applications". MA thesis. Department of Mathematics and Statistics, McGill University, Montreal, Canada.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley & Sons, Inc., New York-London-Sydney. DOI: 10.1002/9781118186428.

Barzilai, J.; J. M. Borwein (1988). "Two-point step size gradient methods". *IMA Journal of Numerical Analysis* 8.1, pp. 141–148. DOI: 10.1093/imanum/8.1.141.

Cartan, H. (1971). *Differential Calculus*. Translated from the French. Hermann, Paris; Houghton Mifflin Co., Boston, Massachusetts.

Cauchy, A.-L. (1847). "Méthode générale pour la résolution des systemes d'équations simultanées". *Comptes Rendus de l'Académie des Sciences Paris* 25, pp. 536–538.

De Asmundis, R.; D. di Serafino; F. Riccio; G. Toraldo (2013). "On spectral properties of steepest descent methods". *IMA Journal of Numerical Analysis* 33.4, pp. 1416–1435. DOI: 10.1093/imanum/drs056.

De Asmundis, R.; D. di Serafino; W. W. Hager; G. Toraldo; H. Zhang (2014). "An efficient gradient method using the Yuan steplength". *Computational Optimization and Applications* 59.3, pp. 541–563. DOI: 10.1007/s10589-014-9669-5.

Elman, H. C.; D. J. Silvester; A. J. Wathen (2014). *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. 2nd ed. Numerical Mathematics and Scientific Computation. Oxford University Press. DOI: 10.1093/acprof:oso/9780199678792.001.0001.

Forsythe, G. E. (1968). "On the asymptotic directions of the s-dimensional optimum gradient method". *Numerische Mathematik* 11, pp. 57–76. DOI: 10.1007/BF02165472.

Geiger, C.; C. Kanzow (1999). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-58582-1.

Gonzaga, C. C. (2016). "On the worst case performance of the steepest descent algorithm for quadratic functions". *Mathematical Programming Series A* 160, pp. 307–320. DOI: 10.1007/s10107-016-0984-8.

Gonzaga, C. C.; R. M. Schneider (2015). "On the steepest descent algorithm for quadratic functions". *Computational Optimization and Applications* 63.2, pp. 523–542. DOI: 10.1007/s10589-015-9775-z.

Herzog, R. (2022). *Grundlagen der Optimierung*. Lecture notes. URL: https://tinyurl.com/scoop-gdo.

Hestenes, M. R.; E. Stiefel (1952). "Methods of conjugate gradients for solving linear systems". *Journal of Research of the National Bureau of Standards* 49, 409–436 (1953). DOI: 10.6028/jres.049.044.

Heuser, H. (2002). *Lehrbuch der Analysis. Teil 2*. 12th ed. Stuttgart: B.G.Teubner. DOI: 10.1007/978-3-322-96826-5.

Nocedal, J.; A. Sartenaer; C. Zhu (2002). "On the behavior of the gradient norm in the steepest descent method". *Computational Optimization and Applications. An International Journal* 22.1, pp. 5–35. DOI: 10.1023/A:1014897230089.

Nocedal, J.; S. J. Wright (2006). *Numerical Optimization*. 2nd ed. New York: Springer. DOI: 10.1007/978-0-387-40065-5.