Lecture Notes
Nonlinear Optimization

Spring Semester 2023

Roland Herzog*

2023-05-13

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

These lecture notes are partly based on content from the books Nocedal, Wright, 2006; Ulbrich, Ulbrich, 2012.

Material for 14 weeks.

Please send comments to roland.herzog@iwr.uni-heidelberg.de.

Contents

Chapter 0 Introduction

§ 1 Elementary Notions

Mathematical optimization is about solving problems of the form

$$\left.\begin{array}{rlll} \text{Minimize} & f(x) & \text{where } x \in \Omega & \text{(\textbf{objective function})} \\ \text{subject to} & g_i(x) \leq 0 & \text{for } i = 1, \ldots, n_{\text{ineq}} & \text{(\textbf{inequality constraints})} \\ \text{and} & h_j(x) = 0 & \text{for } j = 1, \ldots, n_{\text{eq}}. & \text{(\textbf{equality constraints})} \end{array}\right\} \tag{1.1}$$

$\Omega \subseteq \mathbb{R}^n$ 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.

Definition 1.1 (Elementary notions)**.**

(*i*) *The set*

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

associated with an optimization problem (1.1) *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* (1.1)*.*

(*iv*) *In case $F = \emptyset$, the problem* (1.1) *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* (1.1) *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* (1.1).

(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* (1.1) *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* (1.1).

(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* (1.1) *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* (1.1) *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*

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

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

(iii) *When f is a quadratic polynomial and g and h are affine linear functions, then* (1.1) *is called a* ***quadratic optimization problem*** *or a* ***quadratic program*** *(****QP****).*

(iv) *In the general case, i. e., when* (1.1) *is not a quadratic program, we refer to* (1.1) *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 1 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 2). We will subsequently discuss major algorithmic ideas for constrained problems in Chapter 3. Finally, we will review in Chapter 4 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.

§ 2 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.

§ 2.1 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:

$$(x, y) = (y, x) \qquad \text{(symmetry)} \tag{2.1a}$$
$$(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_2, y) \qquad \text{(bilinearity part 1)} \tag{2.1b}$$
$$(x, \beta_1 y_1 + \beta_2 y_2) = \beta_1 (x, y_1) + \beta_2 (x, y_2) \qquad \text{(bilinearity part 2)} \tag{2.1c}$$
$$(x, x) \geq 0 \quad \text{and} \quad x \neq 0 \Rightarrow (x, x) > 0 \qquad \text{(positive definiteness)} \tag{2.1d}$$

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

$$\|x\|_M := \sqrt{x^\mathsf{T} M\, x}. \tag{2.2}$$

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.

§ 2.2 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

$$\|A\|_{M_2\leftarrow M_1} := \max_{x\neq 0} \frac{\|A\,x\|_{M_2}}{\|x\|_{M_1}}. \tag{2.3}$$

We thus have

$$\|A\,x\|_{M_2} \le \|A\|_{M_2\leftarrow M_1}\,\|x\|_{M_1} \quad \text{for all } x \in \mathbb{R}^n. \tag{2.4}$$

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.

§ 2.3 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

$$A\,V = V\Lambda, \quad \text{i. e.,} \quad A = V\Lambda V^\mathsf{T} \tag{2.5}$$

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**

$$A\,v = \lambda\,v. \tag{2.6}$$

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

$$A\,v = \lambda\,M\,v \tag{2.7}$$

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**

$$A\,V = M\,V\Lambda, \quad \text{i. e.,} \quad A = M\,V\Lambda V^\mathsf{T} M, \tag{2.8}$$

where now V is orthogonal w.r.t. the M-inner product, i. e., $V^\mathsf{T} M\,V = \mathrm{Id}$ holds. This implies $VV^\mathsf{T} = M^{-1}$. We also refer to the solutions of (2.7) 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

$$\lambda_{\min}(A;M) \le \frac{x^\mathsf{T} A\,x}{x^\mathsf{T} M\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0. \tag{2.9}$$

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.

Notice that the generalized eigenvalue problems (2.7) and

$$M v = \lambda M A^{-1} M v \tag{2.10a}$$

as well as

$$A M^{-1} A v = \lambda A v \tag{2.10b}$$

have the same eigenvalues and eigenvectors (provided ~~in case of (2.10a)~~ 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 (2.10):

$$\lambda_{\min}(A;M) \leq \frac{x^\mathsf{T} M x}{x^\mathsf{T} M A^{-1} M x} \leq \lambda_{\max}(A;M) \quad \text{for all } x \neq 0, \tag{2.11a}$$

$$\lambda_{\min}(A;M) \leq \frac{x^\mathsf{T} A M^{-1} A x}{x^\mathsf{T} A x} \leq \lambda_{\max}(A;M) \quad \text{for all } x \neq 0. \tag{2.11b}$$

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

$$A^{1/2} = V \Lambda^{1/2} V^\mathsf{T} \tag{2.12}$$

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

§ 2.4 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

$$\kappa := \frac{\beta}{\alpha}. \tag{2.13}$$

Notice that a condition number always satisfies $\kappa \geq 1$. From the Rayleigh quotient estimate (2.9) (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

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

[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 $\sigma_{\min}(A) = \lambda_{\min}(A) = \alpha$.

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 2.1:** Can you explain why not?) The Kantorovich inequality improves this estimate.

Lemma 2.1 (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*

$$1 \leq \frac{(x^\mathsf{T} A\, x)\, (x^\mathsf{T} A^{-1} x)}{\|x\|^4} \leq \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \leq \frac{\beta}{\alpha} \tag{2.14a}$$

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

$$1 \leq \frac{(x^\mathsf{T} A\, x)\, (x^\mathsf{T} A^{-1} x)}{\|x\|^4} \leq \frac{(\kappa+1)^2}{4\,\kappa} \leq \kappa. \tag{2.14b}$$

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 (2.14).

From here on, the proof follows Anderson, 1971, as reproduced in the Master's thesis Alpargu, 1996, Section 1.2.2. Let $\lambda_1, \ldots, \lambda_n > 0$ be the eigenvalues of A (in any order), and let $v_1, \ldots, v_n$ 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 $0 < \alpha \leq T \leq \beta$, we can estimate

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

and thus

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

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 (2.14). 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 2.2 (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*

$$1 \le \frac{(x^{\mathsf{T}}A\,x)\,(x^{\mathsf{T}}M\,A^{-1}M\,x)}{\|x\|_M^4} \le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{2.15a}$$

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

$$1 \le \frac{(x^{\mathsf{T}}A\,x)\,(x^{\mathsf{T}}A^{-1}x)}{\|x\|_M^4} \le \frac{(\kappa+1)^2}{4\,\kappa} \le \kappa. \tag{2.15b}$$

We do not give a proof of Corollary 2.2 here; see for instance Herzog, 2022, Folgerung 4.14.

§ 2.5 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 $F_1, \ldots, F_m$ are differentiable at x. In this case, the derivative $F'(x)$ is given by the **Jacobian** of F at x, i. e., by

$$\begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x)}{\partial x_1} & \cdots & \frac{\partial F_m(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

- 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

$$\left(\frac{\partial^2 f(x)}{\partial x_i\,\partial x_j}\right)_{i,j=1}^{n} = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

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.

§ 2.6 Taylor's Theorem

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

Theorem 2.3 (Taylor, see Cartan, 1967, 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*

$$\text{in case } k = 0: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\,d\right| \le \varepsilon \|d\|,$$
$$\text{in case } k = 1: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\,d - \frac{1}{2} d^{\mathsf{T}} f''(x^{(0)}) d\right| \le \varepsilon \|d\|^2.$$

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

Theorem 2.4 (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 $f\colon G \to \mathbb{R}$ $(k+1)$ 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

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

§ 2.7 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.

[3] See for instance Cartan, 1967, Proposition 5.2.2

Definition 2.5 (Q-convergence rates[4]).
Suppose that $(x^{(k)}) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let M be an inner product on $\mathbb{R}^n$.

(*i*) $(x^{(k)})$ *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*) $(x^{(k)})$ *converges to x^* (at least)* ***Q-superlinearly*** *w.r.t. the M-norm if there exists a null sequence $(\varepsilon^{(k)})$ 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 $x^{(k)} \to x^*$. $(x^{(k)})$ 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 2.6 (R-convergence rates[5]).
Suppose that $(x^{(k)}) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let M be an inner product on $\mathbb{R}^n$.

(*i*) $(x^{(k)})$ *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.

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

§ 2.8 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, § 13 if you wish to have more background information.

Definition 2.7 (Convex function)**.**
A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is termed

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

$$f(\alpha\, x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{2.16}$$

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

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

$$f(\alpha\, x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{2.17}$$

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*

$$f(\alpha\, x + (1-\alpha)\, y) + \frac{\mu}{2}\, \alpha\, (1-\alpha) \|x - y\|^2 \leq \alpha f(x) + (1-\alpha) f(y) \tag{2.18}$$

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

(a) The following are equivalent:

(*i*) *f is convex.*

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

$$f(x) - f(y) \geq f'(y)(x - y) \tag{2.19}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x - y) \geq 0 \tag{2.20}$$

holds. Equation (2.20) 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 $x \neq y$,*

$$f(x) - f(y) > f'(y)(x - y) \tag{2.21}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x - y) > 0. \tag{2.22}$$

*Equation (2.22) 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 $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) \geq f'(y)(x - y) + \frac{\mu}{2} \|x - y\|^2 \tag{2.23}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x - y) \geq \mu \|x - y\|^2. \tag{2.24}$$

*Equation (2.24) means that f' is a **strongly monotone operator**.*

Theorem 2.9 (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 $x \in \mathbb{R}^n$.*

§ 2.9 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 $x \in \mathbb{R}^n$,

$$B_\varepsilon^M(x^{(0)}) \coloneqq \left\{x \in \mathbb{R}^n \,\middle|\, \|x - x^{(0)}\|_M < \varepsilon\right\}$$

denotes the **open ε-ball** w.r.t. the M-norm about $x^{(0)}$. Similarly, the **closed ε-ball** is

$$\operatorname{cl} B_\varepsilon^M(x^{(0)}) \coloneqq \left\{x \in \mathbb{R}^n \,\middle|\, \|x - x^{(0)}\|_M \leq \varepsilon\right\}.$$

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

Chapter 1 Numerical Techniques for Unconstrained Optimization Problems

We discuss in this chapter numerical methods for the unconstrained version of (1.1), 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, first-order optimality conditions are already sufficient for optimality, and we do not require second-order conditions.

§ 3 Optimality Conditions

We assume you have seen the following first- and second-order optimality conditions, so we only briefly recall them; see Herzog, 2022, § 3 for more details.

Theorem 3.1 (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 3.2 (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 3.1 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 2.3 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 3.1, $\varphi'(0) = 0$, 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 3.3 (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 $\mu > 0$.*

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

$$f(x) \geq f(x^*) + \frac{\beta}{2}\|x - x^*\|^2 \quad \text{for all } x \in U(x^*). \tag{3.1}$$

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

Proof. Here we use Theorem 2.3 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 (2.9) (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

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

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

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

for all $\|d\| < \delta$. □

Property (3.1) means that f has at least **quadratic growth** near x^*. Equivalently, f is locally strongly convex with parameter $\beta \in (0, \mu)$.

End of Week 1

§ 4 Minimization of Quadratic Functions

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

$$\text{Minimize} \quad \phi(x) := \frac{1}{2} x^\mathsf{T} A\, x - b^\mathsf{T} x + c \quad \text{where } x \in \mathbb{R}^n. \tag{4.1}$$

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 4.1:** 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 (4.1) is an essential building block for subsequent tasks.

Lemma 4.1 (Solvability and global solutions of (4.1)[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:*

(*i*) *If A is positive semidefinite, then the objective in (4.1) is convex. In this case, the following are equivalent:*

  (*a*) *The problem (4.1) possess at least one (global) minimizer.*

  (*b*) *The objective ϕ is bounded below.*

  (*c*) *$Ax = b$ is solvable.*

  The global minimizers of (4.1) 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 (4.1) is unbounded.*

Proof. □

Corollary 4.2 (Unique solvability of (4.1)). *Problem (4.1) 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 § 4 that A is symmetric and positive definite (s. p. d.). Hence, the solution of (4.1) 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 $A = LL^\intercal$ with the lower triangular matrix L.[5] 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[6], 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 (4.1)), 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

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

[4] The matrix A possesses at least one negative eigenvalue.

[5] We assume you have seen these methods, e. g., in the class *Einführung in die Numerik*.

[6] such as a sparse Cholesky decomposition

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, i. e., a function which evaluates $x \mapsto Ax$. **Quiz 4.2:** 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 (4.1) 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.[7] 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 4.3. *We have*

$$\phi(x) - \phi(x^*) = \frac{1}{2}\|x - x^*\|_A^2 = \frac{1}{2}\|r\|_{A^{-1}}^2 = \frac{1}{2}\|\nabla\phi(x)\|_{A^{-1}}^2. \tag{4.2}$$

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 (4.1), and equivalently the solution of the linear system $Ax = b$, where A is s. p. d.[8] These methods are the **gradient descent method** (also known as **steepest descent method**), and the **conjugate gradient method**.

[7] 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.

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

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 $x^{(k+1)}$.

§ 4.1 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 4.4 (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*

$$f'(x)\, d < 0. \tag{4.3}$$

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:

$$\begin{aligned} &\text{Minimize} && f'(x)\, d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M = 1. \end{aligned} \tag{4.4}$$

We note that we could be considering the equivalent problem

$$\begin{aligned} &\text{Minimize} && f'(x)\, d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M \leq 1. \end{aligned} \tag{4.5}$$

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

Problems (4.4), (4.5) 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[9]

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

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 4.5 (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.

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

(*i*) *The vector*

$$\nabla_M f(x) := M^{-1}\nabla f(x) \tag{4.6}$$

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.*

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

$$M\,d^* = -\nabla f(x). \tag{4.7}$$

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**.

§ 4.2 Gradient Descent Method with Cauchy Step Sizes

The direction of steepest descent at x used by the gradient method is thus[10]

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

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:

$$\begin{aligned}\phi(x+\alpha\,d) - \phi(x) &= \frac{1}{2}(x+\alpha\,d)^{\mathsf{T}}A\,(x+\alpha\,d) - b^{\mathsf{T}}(x+\alpha\,d) + c - \frac{1}{2}x^{\mathsf{T}}A\,x + b^{\mathsf{T}}x - c\\ &= \frac{1}{2}(d^{\mathsf{T}}A\,d)\,\alpha^2 + (A\,x-b)^{\mathsf{T}}d\,\alpha\\ &= \frac{1}{2}(d^{\mathsf{T}}A\,d)\,\alpha^2 + (r^{\mathsf{T}}d)\,\alpha.\end{aligned} \tag{4.8}$$

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 (4.8), we have

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

Due to the positivity of the second derivative, the second-order sufficient condition (Theorem 3.3) is satisfied when $\frac{\mathrm{d}}{\mathrm{d}\alpha}\phi(x+\alpha\,d) = 0$, which amounts to

$$\alpha^* = -\frac{r^{\mathsf{T}}d}{d^{\mathsf{T}}A\,d}. \tag{4.9}$$

[10] 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 (4.8) before and after a step becomes

$$\begin{aligned}\phi(x+\alpha^* d) - \phi(x) &= \frac{1}{2}(d^\mathsf{T} A\, d)\,(\alpha^*)^2 + (r^\mathsf{T} d)\,\alpha^* \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\left(\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d}\right)^2 - (r^\mathsf{T} d)\,\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d} \\ &= -\frac{1}{2}\frac{(r^\mathsf{T} d)^2}{d^\mathsf{T} A\, d}. \end{aligned} \tag{4.10}$$

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 (4.9) for the iterative solution of the unconstrained quadratic minimization problem (4.1) with s. p. d. A. This method, with $M = \mathrm{Id}$, was already published by Cauchy, 1847.

Algorithm 4.6 (Gradient descent method for (4.1) 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* (4.1)*, i. e., of* $A\,x = b$

1: Set $k := 0$
2: Set $r^{(0)} := A\,x^{(0)} - b$ // *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 $q^{(k)} := A\,d^{(k)}$
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 4.6.

Remark 4.7 (on Algorithm 4.6)**.**

(*i*) *Algorithm 4.6 is an iterative solver for the unconstrained quadratic minimization problem* (4.1) *with s. p. d.* A*, and simultaneously an iterative solver for the linear system* $A\,x = b$*.*

(*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 4.6 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 4.6 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*

$$0 < \lambda_{\min}(A;M) \le \frac{1}{\alpha^{(k)}} = \frac{(d^{(k)})^{\mathsf{T}} A\,d^{(k)}}{(d^{(k)})^{\mathsf{T}} M\,d^{(k)}} \le \lambda_{\max}(A;M), \tag{4.11}$$

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 4.6 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:*

$$\phi(x^{(0)}) = c + \frac{1}{2}(r^{(0)} - b)^{\mathsf{T}}(x^{(0)}) \quad \textit{initialization} \tag{4.12a}$$

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \frac{1}{2}\,\alpha^{(k)}\delta^{(k)} \quad \textit{update.} \tag{4.12b}$$

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 4.6. The function values at the iterates satisfy

$$\begin{aligned}
&\phi(x^{(k+1)}) - \phi(x^*) \\
&\quad = \frac{1}{2}\|r^{(k+1)}\|^2_{A^{-1}} && \text{by (4.2)} \\
&\quad = \frac{1}{2}\|r^{(k)} + \alpha^{(k)} A\,d^{(k)}\|^2_{A^{-1}} \\
&\quad = \frac{1}{2}\|r^{(k)}\|^2_{A^{-1}} + \alpha^{(k)}(r^{(k)})^{\mathsf{T}} d^{(k)} + \frac{1}{2}\left[\alpha^{(k)}\right]^2 (d^{(k)})^{\mathsf{T}} A\,d^{(k)}.
\end{aligned}$$

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

$$
\begin{aligned}
&= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} + \frac{1}{2}\frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} \\
&= \left(1 - \frac{\left[(r^{(k)})^\mathsf{T} d^{(k)}\right]^2}{\left[(d^{(k)})^\mathsf{T} A\, d^{(k)}\right]\left[(r^{(k)})^\mathsf{T} A^{-1} r^{(k)}\right]}\right)\left(\phi(x^{(k)}) - \phi(x^*)\right) \qquad \text{by (4.2).}
\end{aligned}
$$

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 (2.15). 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 4.6:

Theorem 4.8 (Convergence of Algorithm 4.6)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ are 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 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 (4.1). In terms of the generalized condition number $\kappa = \beta/\alpha$, we have the estimates*

$$
\phi(x^{(k+1)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^2 \left(\phi(x^{(k)}) - \phi(x^*)\right) \tag{4.13a}
$$

$$
\|x^{(k+1)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right) \|x^{(k)} - x^*\|_A \tag{4.13b}
$$

and consequently

$$
\phi(x^{(k)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{4.13c}
$$

$$
\|x^{(k)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \|x^{(0)} - x^*\|_A. \tag{4.13d}
$$

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 (4.13c) and (4.13d) have been decreased relative to their initial values.

Corollary 4.9 (Maximal number of iterations required in Algorithm 4.6)**.** *Given positive numbers ε_1 and ε_2, it takes*

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

Proof. (1) We first show that

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

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

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

We now show that

$$\frac{\mathrm{d}}{\mathrm{d}\kappa}\left[-\ln\left(\frac{\kappa-1}{\kappa+1}\right)\right] \ge \frac{\mathrm{d}}{\mathrm{d}\kappa}\frac{2}{\kappa}$$

holds for all $\kappa > 1$, which proves the claim. The derivative on the left is $\frac{-2}{(\kappa-1)(\kappa+1)}$, while the derivative on the right is $\frac{-2}{\kappa^2}$. In view of $0 < \kappa^2 - 1 < \kappa^2$ for all $\kappa > 1$, we conclude

$$\frac{-2}{(\kappa-1)(\kappa+1)} < \frac{-2}{\kappa^2} < 0 \quad \text{for all } \kappa > 1.$$

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

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

for all $\kappa > 1$.

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

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

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

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

implies ($**$), which proves the first claim.

The second claim follows similarly. □

Remark 4.10 (on Theorem 4.8)**.**

(*i*) (4.13b) *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* (4.2)*, that is the quantity being minimized.*

(*vi*) *The estimates of Theorem 4.8 are worst-case estimates since they do not depend on the initial guess* $x^{(0)}$*. In fact, as can be seen in Figure 4.1c, the actual contraction factor for the objective values can be significantly smaller for some initial guesses than the estimate* (4.13c) *suggests.*

Figure 4.1 illustrates the convergence behavior of Algorithm 4.6 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 4.1). **Quiz 4.3:** Suppose A, b and M are given and you consider a random distribution of initial values $x^{(0)}$ in $\mathbb{R}^n$, which has a probability density. What is the probability of hitting an initial value such that convergence happens in one step?

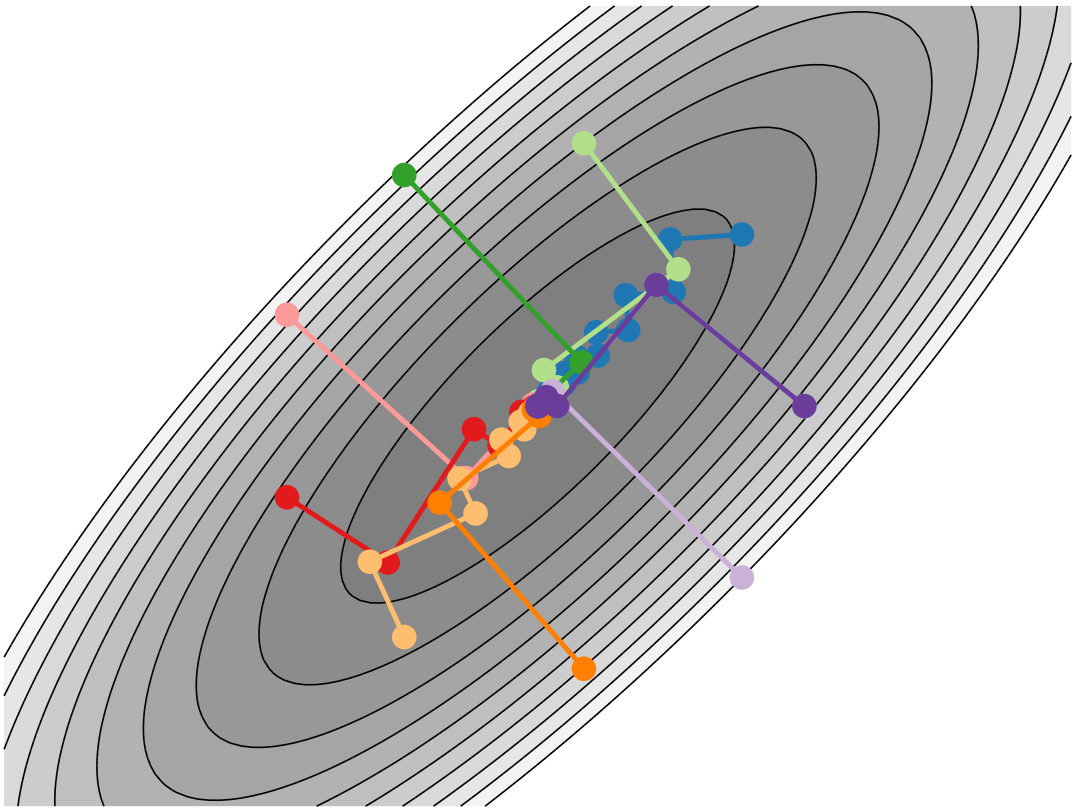
The zig-zagging behavior of the iterates $x^{(k)}$, as well as the non-monotone behavior of $\|r^{(k)}\|_{M^{-1}}$ have been analyzed in detail in the literature; see for instance Akaike, 1959; Forsythe, 1968; Nocedal, Sartenaer, Zhu, 2002. Essentially what happens is that, asymptotically, the error $x^{(k)} - x^*$ 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 $\kappa > 3 + 2\sqrt{2} \approx 5.83$.

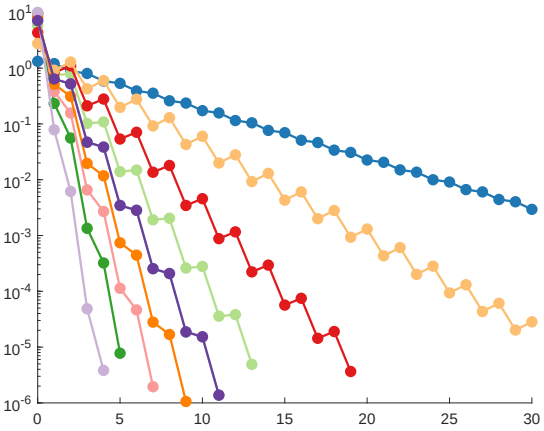
It remains to discuss stopping criteria. Several quantities may be of interest in this respect:

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

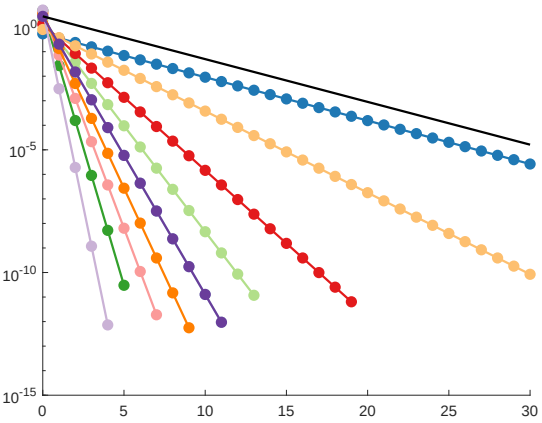
(*ii*) Are we happy with a point $x^{(k)}$ whose objective value is near the optimal value, i. e., where $\phi(x^{(k)}) - \phi(x^*)$ is small, or equivalently, where $\|x^{(k)} - x^*\|_A$ is small?



(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 (4.13c).

Figure 4.1: Illustration of the convergence behavior of Algorithm 4.6 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 $\kappa = 10$.

(*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?

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 $\|r^{(k)}\|_{M^{-1}}$:

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon^2_{\text{rel}}\, \delta^{(0)}, \tag{4.14a}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{abs}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon^2_{\text{abs}}, \tag{4.14b}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}} + \varepsilon_{\text{abs}}, \qquad \text{i. e., } (\delta^{(k)})^{1/2} \le \varepsilon_{\text{rel}} (\delta^{(0)})^{1/2} + \varepsilon_{\text{abs}}, \tag{4.14c}$$

$$\|r^{(k)}\|_{M^{-1}} \le \max\{\varepsilon_{\text{rel}} \|r^{(0)}\|_{M^{-1}},\ \varepsilon_{\text{abs}}\}, \qquad \text{i. e., } \delta^{(k)} \le \max\{\varepsilon^2_{\text{rel}}\, \delta^{(0)},\ \varepsilon^2_{\text{abs}}\}. \tag{4.14d}$$

Let us see which consequences either of the implementable stopping criteria (4.14) has on the other two quantities of interest:

Lemma 4.11 (Implications). *The criteria from* (4.14) *imply, respectively,*

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A \\ \|x^{(k)} - x^*\|_M &\le \kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M \end{aligned}\right\} \tag{4.15a}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le (1/\alpha)\, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.15b}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A + (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le \kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M + (1/\alpha)\, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.15c}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \max\{\sqrt{\kappa}\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_A,\ (1/\sqrt{\alpha})\, \varepsilon_{\text{abs}}\} \\ \|x^{(k)} - x^*\|_M &\le \max\{\kappa\, \varepsilon_{\text{rel}} \|x^{(0)} - x^*\|_M,\ (1/\alpha)\, \varepsilon_{\text{abs}}\} \end{aligned}\right\} \tag{4.15d}$$

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

§ 4.3 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

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

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

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

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 (2.11a):

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

The condition that $\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)$ is negative amounts to $\overline{\alpha} \in (0, \frac{2}{\beta})$. It is precisely the midpoint $\overline{\alpha} = 1/\beta$ of this interval which minimizes this term and yields the optimal estimate, and the expression in $[\cdots]$ becomes $\frac{\kappa-1}{\kappa}$ in this case.

Remark 4.12 (on the convergence of Algorithm 4.6 with constant step sizes)**.**

(*i*) *We have shown that Algorithm 4.6, 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* $\beta \le \beta_{\text{estimate}}$ *and choose* $\overline{\alpha} \in (0, \frac{2}{\beta_{\text{estimate}}})$*, we also have* $\overline{\alpha} \in (0, \frac{2}{\beta})$.

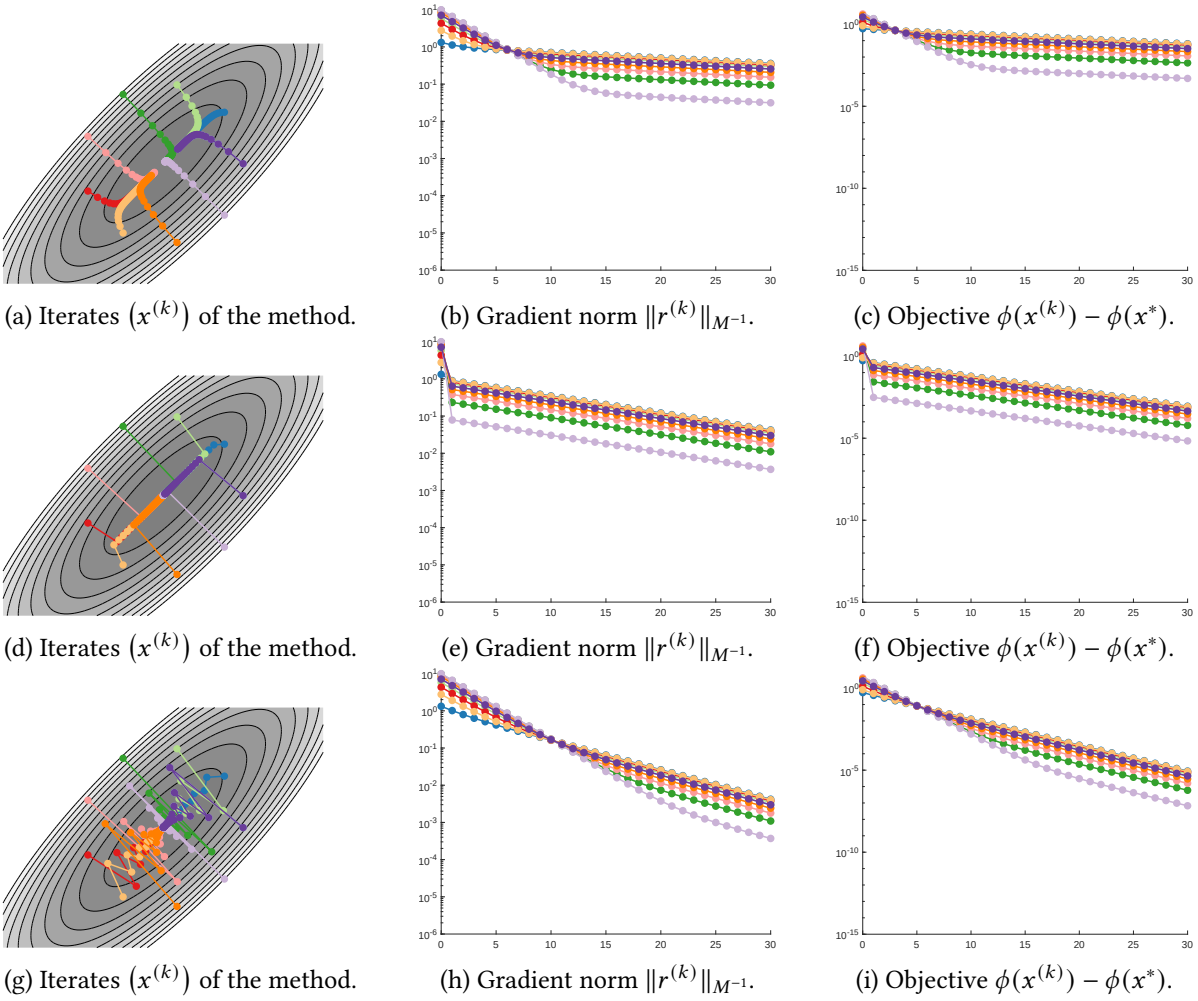
(*iii*) *The choice* $\overline{\alpha} = \frac{1}{\beta}$ *yields the optimal estimate. In this case, we obtain*

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

Since for all $\kappa \ge 1$*, we have* $\left(\frac{\kappa-1}{\kappa+1}\right)^2 \le \frac{\kappa-1}{\kappa}$*, the contraction factor in the bound we obtained with constant step sizes is worse than the one for the Cauchy step sizes; see (4.13a). 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 4.2 illustrates the convergence behavior of Algorithm 4.6 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 4.2: Illustration of the convergence behavior of Algorithm 4.6 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 $\overline{\alpha} \in (0, \frac{2}{\beta}) = (0, 0.2)$.

§ 4.4 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

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

Then the gradient descent method Algorithm 4.6 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 4.9 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 § 4.6.

§ 4.5 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

$$\begin{aligned} \dot{x}(t) &= -\nabla_M f(x(t)), \quad t \geq 0 \\ x(0) &= x^{(0)}. \end{aligned} \tag{4.16}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = f'(x(t))\,\dot{x}(t) = -f'(x(t))\,M^{-1}\nabla f(x(t)) = -\|\nabla f(x(t))\|^2_{M^{-1}} = -\|\nabla_M f(x(t))\|^2_M, \tag{4.17}$$

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

When we discretize (4.16) 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,

$$x^{(k+1)} = x^{(k)} - \Delta t^{(k)}\, M^{-1}\nabla f(x^{(k)}). \tag{4.18}$$

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.

End of Week 2

§ 4.6 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 4.13 (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 4.4:** 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

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}. \tag{4.19}$$

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

$$\alpha^{(k)} := -\frac{(r^{(k)})^{\mathsf{T}} d^{(k)}}{(d^{(k)})^{\mathsf{T}} A\, d^{(k)}}, \tag{4.20}$$

compare (4.9). As in the gradient descent method, the residuals satisfy the recursion

$$r^{(k+1)} = r^{(k)} + \alpha^{(k)} A\, d^{(k)}. \tag{4.21}$$

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 4.14 (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 (4.19) with Cauchy step size (4.20). Then the following holds.*

(*i*)
$$(r^{(k)})^\mathsf{T} d^{(i)} = 0 \quad \text{for all } i = 0, 1, \dots, k-1. \tag{4.22}$$

(*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 $k = 1$,

$$\begin{aligned}
(r^{(1)})^\mathsf{T} d^{(0)} &= (A x^{(1)} - b)^\mathsf{T} d^{(0)} && \text{by definition of the residual}\\
&= (A x^{(0)} + \alpha^{(0)} A d^{(0)} - b)^\mathsf{T} d^{(0)} && \text{by (4.19)}\\
&= (r^{(0)})^\mathsf{T} d^{(0)} + \alpha^{(0)} (d^{(0)})^\mathsf{T} A d^{(0)} && \text{by definition of the residual}\\
&= 0 && \text{since } \alpha^{(0)} \text{ is the Cauchy step size (4.20).}
\end{aligned}$$

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

$$\begin{aligned}
(r^{(k)})^\mathsf{T} d^{(k-1)} &= (r^{(k-1)} + \alpha^{(k-1)} A d^{(k-1)})^\mathsf{T} d^{(k-1)} && \text{by the residual recursion (4.21)}\\
&= 0 && \text{since } \alpha^{(k-1)} \text{ is the Cauchy step size (4.20).}
\end{aligned}$$

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

$$\begin{aligned}
(r^{(k)})^\mathsf{T} d^{(i)} &= \left(r^{(k-1)} + \alpha^{(k-1)} A d^{(k-1)}\right)^\mathsf{T} d^{(i)} && \text{by the residual recursion (4.21)}\\
&= \underbrace{(r^{(k-1)})^\mathsf{T} d^{(i)}}_{=0 \text{ by assumption}} + \alpha^{(k-1)} \underbrace{(d^{(k-1)})^\mathsf{T} A d^{(i)}}_{=0 \text{ due to } A\text{-conjugacy}}\\
&= 0.
\end{aligned}$$

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 4.5:** 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 4.15 (Properties of conjugate direction methods)**.** *Any iterative method* (4.19) *using A-conjugate directions* $d^{(k)}$ *and Cauchy step sizes* (4.20) *converges to the unique solution of* (4.1) *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 4.14. □

In practice, the statement of Corollary 4.15 is weakened by floating point error. Moreover, the result of Corollary 4.15 is not really relevant for high-dimensional problems since performing n iterations is prohibitively expensive. We will later see more practical converge 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 $-M^{-1}r^{(k)}$:[11]

$$\begin{aligned} d^{(0)} &:= -M^{-1}r^{(0)} && \text{for } k = 0, \\ d^{(k)} &:= -M^{-1}r^{(k)} + \beta^{(k)}\, d^{(k-1)} && \text{for } k \geq 1. \end{aligned} \tag{4.23}$$

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

$$\beta^{(k)} := \frac{(r^{(k)})^{\mathsf{T}} M^{-1} A\, d^{(k-1)}}{(d^{(k-1)})^{\mathsf{T}} A\, d^{(k-1)}}. \tag{4.24}$$

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

Lemma 4.16 (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* (4.19)–(4.20), (4.23)–(4.24), *where* $\alpha^{(k)} \neq 0$.[12]

$$\operatorname{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k)}\} = \operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \dots, (A\,M^{-1})^k\, r^{(0)}\}, \tag{4.25}$$

$$\operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k)}\} = M^{-1} \operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \dots, (A\,M^{-1})^k\, r^{(0)}\}, \tag{4.26}$$

$$(d^{(k)})^{\mathsf{T}} A\, d^{(i)} = 0 \quad \textit{for all } i = 0, 1, \dots, k-1, \tag{4.27}$$

$$(r^{(k)})^{\mathsf{T}} M^{-1} r^{(i)} = 0 \quad \textit{for all } i = 0, 1, \dots, k-1. \tag{4.28}$$

[11] With $\beta^{(k)} = 0$, we obtain again the steepest descent method (Algorithm 4.6).

[12] $\alpha^{(k)} = 0$ would mean that $x^{(k)}$ is the unique solution x^*. Due to the form of the Cauchy step (4.20), this is clear for $k = 0$, as the nominator is $\|r^{(k)}\|_{M^{-1}}$. (4.22) shows that this is also true for $k > 0$.

The subspace

$$\mathcal{K}^{(k+1)}(A\,M^{-1}; r^{(0)}) := \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \dots, (A\,M^{-1})^k\,r^{(0)}\} \tag{4.29}$$

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 (4.25) and (4.26) imply that the method creates, simultaneously, an expanding sequence of M^{-1}-orthogonal basis vectors of the spaces $\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$, as well as an expanding sequence of A-orthogonal basis vectors of the spaces $M^{-1}\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$.

Proof. We first prove (4.25)–(4.27), by induction. For $k = 0$, statement (4.25) holds trivially. Statement (4.26) holds since the CG method starts with $d^{(0)} = -M^{-1}r^{(0)}$. Statement (4.27) is void for $k = 0$.

Suppose now that (4.25) and (4.26) have been shown up to some $k \geq 0$. We need to show that they also hold for $k+1$. By hypothesis,

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

Due to the residual recursion (4.21), we therefore have

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

Due to the induction hypothesis for (4.25), the same statement (∗) holds when $k+1$ is replaced by a smaller index. Therefore, we have shown that

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

holds. Now for the reverse inequality. By the induction hypothesis for (4.26), we find

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

By the residual recursion (4.21), specifically

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

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

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

When combined with the induction hypothesis for (4.25), i. e.,

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

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 (4.25) is complete.

To see (4.26),

$$\begin{aligned}
&\operatorname{span}\{d^{(0)}, \ldots, d^{(k)}, d^{(k+1)}\} \\
&\quad = \operatorname{span}\{d^{(0)}, \ldots, d^{(k)}, M^{-1}r^{(k+1)}\} && \text{by (4.23)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k}\,r^{(0)}, r^{(k+1)}\} && \text{by (4.26)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k)}, r^{(k+1)}\} && \text{by (4.25)} \\
&\quad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k}\,r^{(0)}, (A\,M^{-1})^{k+1}\,r^{(0)}\} && \text{by (4.25) for } k+1.
\end{aligned}$$

This concludes the induction step for (4.26).

Next we address the A-conjugacy of search directions, (4.27). 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 (4.23) and (4.24). When $i \leq k-1$, we argue as follows. From (4.26), 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 (4.22), 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 (4.27).

Finally, we consider the M^{-1}-conjugacy of residuals, (4.28), 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 (4.21) and the construction of search directions (4.23). 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

$$\begin{aligned} (r^{(k-1)})^\mathsf{T} d^{(k-2)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\mathsf{T} (d^{(k-2)}) &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

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 $i < k-1$)

$$\begin{aligned} (r^{(k-1)})^\mathsf{T} d^{(i)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\mathsf{T} d^{(i)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions,} \\ (r^{(k-1)})^\mathsf{T} d^{(i-1)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\mathsf{T} d^{(i-1)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

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 (4.20) for $\alpha^{(k)}$ as well as (4.24) for $\beta^{(k)}$ in the CG method can be equivalently formulated as follows:

$$\begin{aligned} \alpha^{(k)} &= -\frac{(r^{(k)})^\mathsf{T} d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the Cauchy step size formula (4.20)} \\ &= \frac{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} - \beta^{(k)} \frac{(r^{(k)})^\mathsf{T} d^{(k-1)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the search direction recursion (4.23)} \\ &= \frac{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by (4.22)} \qquad (4.20') \end{aligned}$$

and

$$\begin{aligned} \beta^{(k+1)} &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} A\, d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} && \text{by the orthogonalization coefficient (4.24)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} (r^{(k+1)} - r^{(k)})}{(d^{(k)})^\mathsf{T} (r^{(k+1)} - r^{(k)})} && \text{by the residual recursion (4.21)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} (r^{(k+1)} - r^{(k)})}{(-M^{-1} r^{(k)} + \beta^{(k)}\, d^{(k-1)})^\mathsf{T} (r^{(k+1)} - r^{(k)})} && \text{by the construction of search directions (4.23)} \\ &= \frac{(r^{(k+1)})^\mathsf{T} M^{-1} r^{(k+1)}}{(r^{(k)})^\mathsf{T} M^{-1} r^{(k)}} && \text{by (4.22) and (4.25).} \qquad (4.24') \end{aligned}$$

The relations (4.20') and (4.24') 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 4.17 (Conjugate gradient method for (4.1) 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* (4.1)*, 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 4.18 (on Algorithm 4.17)**.**

(*i*) *From Lemma 4.16 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 4.17, 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 4.6). 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 4.6).*

(*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 4.7 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 (4.14) and their consequences (4.15) 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 4.8 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 4.19 (Convergence of Algorithm 4.17, compare Theorem 4.8)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ are 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 for any choice of the initial guess $x^{(0)}$, the conjugate gradient method converges to the unique solution $x^* = A^{-1}b$ of (4.1). In terms of the generalized condition number $\kappa = \beta/\alpha$, we have the estimates*[13]

$$\phi(x^{(k)}) - \phi(x^*) \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{4.30a}$$

$$\|x^{(k)} - x^*\|_A \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k} \|x^{(0)} - x^*\|_A, \tag{4.30b}$$

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 (4.26), span $M^{-1}\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$, we have

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

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 $A\,e^{(0)} = A\,x^{(0)} - A\,x^* = r^{(0)}$, 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 $p^{(k)}(0) = 1$.

[13] compare (4.13c), (4.13d)

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 $\leq k$:

$$\|e^{(k)}\|_A = \min\Big\{\|p(M^{-1}A)\,e^{(0)}\|_A \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}. \tag{4.31}$$

We expand the initial error $e^{(0)}$ in terms of the basis of eigenvectors of A w.r.t. M; see (2.7), (2.8). 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:

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

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

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

By pulling the maximal value of $|p(\lambda^{(j)})|$ out of the sum (**Quiz 4.6:** Can you fill in the details why this is possible?), we can estimate this quantity further:

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

Combining this with (4.31), we see

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

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

$$\|e^{(k)}\|_A \leq \min\Big\{\max_{z\in[\alpha,\beta]} |p(z)| \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}\,\|e^{(0)}\|_A. \tag{4.32}$$

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 (4.32) 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

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

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

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

From there, we finally obtain

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

which is precisely (4.32). Squaring both sides and dividing by 2, we also obtain (4.30a). □

Corollary 4.20 (Maximal number of iterations required in Algorithm 4.17, compare Corollary 4.9)**.** *Given positive numbers ε_1 and ε_2, it takes*

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

Proof. The proof is similar to Corollary 4.9 and it uses that

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

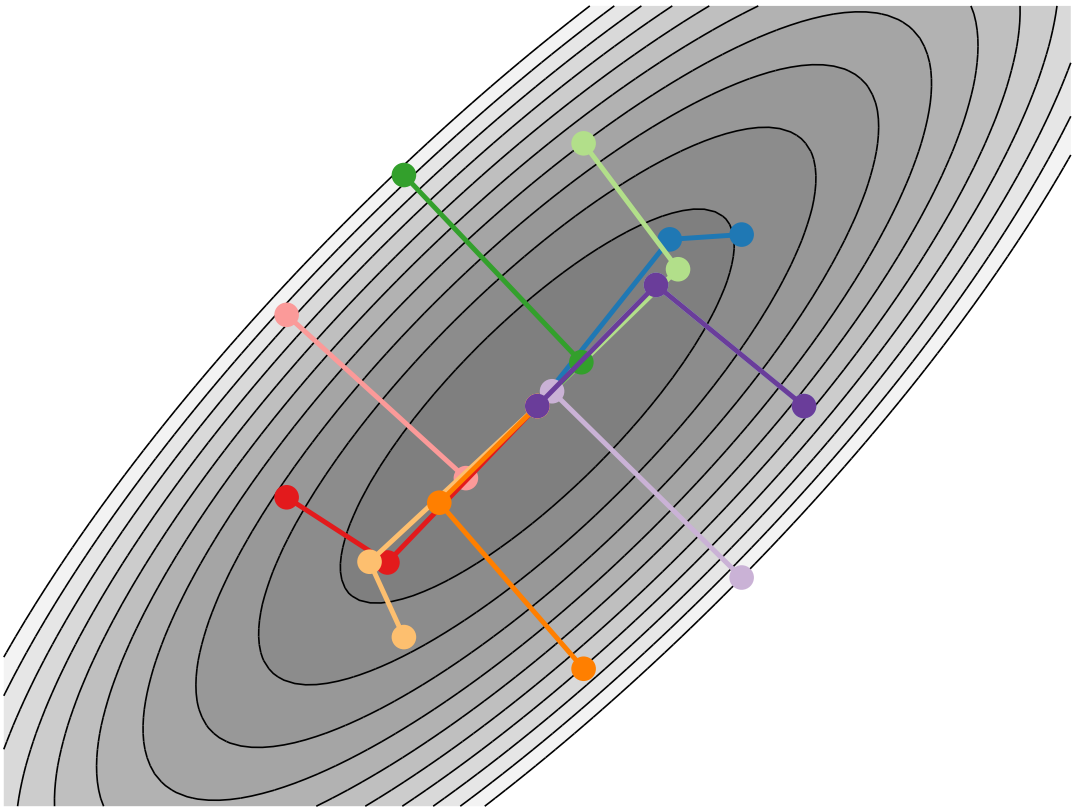
holds for all $\kappa \ge 1$. □

Remark 4.21 (on Theorem 4.19)**.**

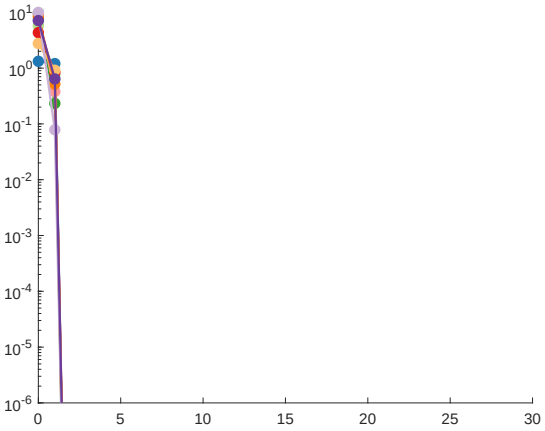
(*i*) *The estimates* (4.30a) *and* (4.32) *establish the R-linear convergence of the respective quantities to zero.*

(*ii*) *Compared to the estimates* (4.13c) *and* (4.13d) *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.*

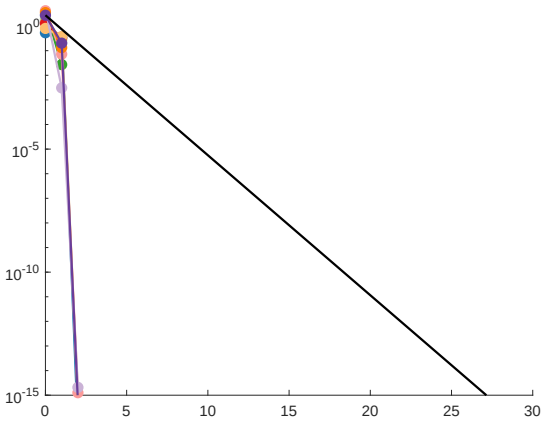
(*iii*) *The superiority of the CG method compared to the gradient descent method is also reflected in the estimates for the maximal iteration numbers to achieve a certain reduction in the quantities* $\phi(x^{(k)}) - \phi(x^*)$ *and* $\|x^{(k)} - x^*\|_A$*, respectively. The bounds for the maximal iteration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



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

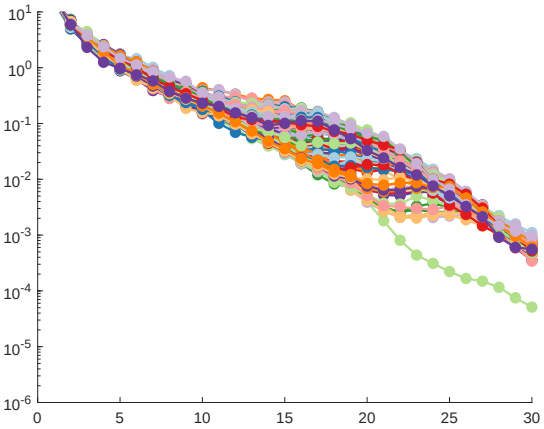


(b) The norm of the gradient $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_M = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.

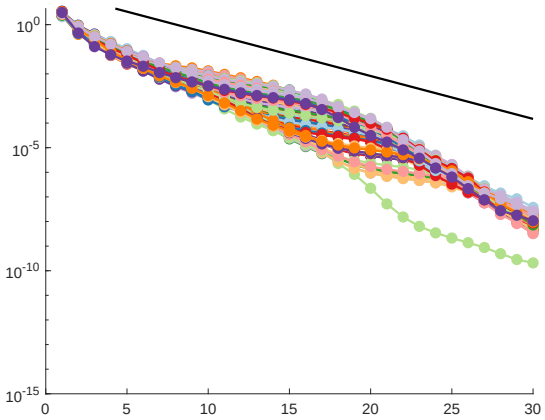


(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (4.30a).

Figure 4.3: Illustration of the convergence behavior of Algorithm 4.17 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 (4.30a).

Figure 4.4: Illustration of the convergence behavior of Algorithm 4.17 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 $\kappa = 100$.

(*iv*) *As was the case for Theorem 4.8, the estimates of Theorem 4.19 are worst-case estimates since they do not depend on the initial guess $x^{(0)}$. In fact, as can be seen in Figure 4.3c and Figure 4.4b, the actual contraction factor for the objective values can be significantly smaller for some initial guesses than the estimate (4.30a) suggests.*

(*v*) *Other informative error bounds than (4.30) and (4.32) and convergence results can be obtained by proceeding as in the proof of Theorem 4.19 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 4.22 (Growth of the distance from the initial guess[14])**.** *Consider the iterates $x^{(k)}$ of the conjugate gradient method (Algorithm 4.17). 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 4.14 implies that

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

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

We now show by induction that $(x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} > 0$ holds for $k \geq 1$. Initially, for $k = 1$, Statement (i) in Lemma 4.14 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

$$\omega^{(k)} := \|x^{(k)} - x^{(0)}\|_M^2 \tag{4.33a}$$
$$\xi^{(k)} := (x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} \tag{4.33b}$$
$$\gamma^{(k)} := \|d^{(k)}\|_M^2 \tag{4.33c}$$

on the side without any noticeable effort. This can be achieved by inserting, at the appropriate positions in Algorithm 4.17 (**Quiz 4.7:** Where?), the relations

$$\omega^{(0)} := 0, \qquad \omega^{(k+1)} := \omega^{(k)} + 2\, \alpha^{(k)} \xi^{(k)} + (\alpha^{(k)})^2\, \gamma^{(k)} \qquad \text{see } (***) \tag{4.34a}$$
$$\xi^{(0)} := 0, \qquad \xi^{(k+1)} := \beta^{(k+1)}\, (\xi^{(k)} + \alpha^{(k)} \gamma^{(k)}) \qquad \text{see } (**) \tag{4.34b}$$
$$\gamma^{(0)} := \delta^{(0)}, \qquad \gamma^{(k+1)} := \delta^{(k+1)} + (\beta^{(k+1)})^2\, \gamma^{(k)} \qquad \text{(confirm).} \tag{4.34c}$$

The remarkable fact about this is the possibility to keep track of (4.33) 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 4.17.

End of Week 3

§ 5 Line Search Methods for Nonlinear Unconstrained Problems

We consider in this section a large class of methods to solve general, nonlinear unconstrained problems

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

The methods we consider are so-called **line search methods**. In every iteration, a line search method first determines a **search direction** and subsequently finds a **step size** (or **step length**) $\alpha^{(k)}$, that leads to the next iterate via

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

Assumption 5.1. *Throughout § 5 we are assuming that $f\colon \mathbb{R}^n \to \mathbb{R}$ is a C^1 function.*

Most line search methods, in particular the ones we consider, require that $d^{(k)}$ is a **descent direction** for the objective f at the current iterate $x^{(k)}$, i. e., that

$$f'(x^{(k)})\, d^{(k)} < 0 \tag{5.1}$$

holds, see Definition 4.4. This implies that we have descent at least for sufficiently small positive step sizes $\alpha^{(k)}$,

$$f(x^{(k+1)}) = f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$$

and it motivates the term **descent method**.

Most methods[15] we are discussing in § 5 determine the search direction $d^{(k)}$ by considering a local **quadratic model** of the ojective:

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^{\mathsf{T}} H^{(k)} d. \tag{5.2}$$

This model uses the data $f(x^{(k)})$ and $f'(x^{(k)})$ at the iterate $x^{(k)}$ and it agrees with f regarding that data at $d = 0$:

$$q^{(k)}(0) = f(x^{(k)})$$
$$\text{and} \quad (q^{(k)})'(0) = f'(x^{(k)})$$

The matrix $H^{(k)}$ is the Hessian of the model, briefly: the **model Hessian**. In case $H^{(k)} = f''(x^{(k)})$, the model $q^{(k)}$ is the second-order Taylor polynomial of f at $x^{(k)}$. However, in general, the model Hessian is chosen to be any symmetric and possibly positive definite matrix. In fact, different line search methods differ w.r.t. their choice of the model Hessians $H^{(k)}$, and thus with respect to the search directions they use.

The search direction $d^{(k)}$ is obtained by minimizing (possibly only to a certain accuracy) the quadratic polynomial $q^{(k)}$:

$$\text{Minimize} \quad q^{(k)}(d), \quad d \in \mathbb{R}^n. \tag{5.3}$$

As we know from Lemma 4.1, the following cases can occur:

[15] with the exception of nonlinear conjugate gradient methods

(*i*) When $H^{(k)}$ is s. p. d., then the unique solution of (5.3) is given by the unique solution of the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.4}$$

(*ii*) When $H^{(k)}$ is symmetric and only positive semidefinite, then (5.3) is either unbounded, or else has infinitely many minimizers. In any case, the minimizers of (5.3) are precisely the solutions of the linear system (5.4).[16]

(*iii*) When $H^{(k)}$ is symmetric but not positive semidefinite (i. e., at least one eigenvalue of $H^{(k)}$ is negative), then (5.3) is an unbounded problem. However, the linear system (5.4) may still be uniquely solvable, or solvable with multiple solutions, or not solvable. The solutions of the linear systems (if any) are either all saddle points[17] of $q^{(k)}$, or they are all global maximizers. (**Quiz 5.1:** Is this statement clear?)

To solve (5.3) and (5.4), respectively, we can employ the conjugate gradient (CG) method from § 4.6. However, it would be useful to enhance it so that it checks and reacts to the potential occurrence of non-positive eigenvalues in the model Hessian $H^{(k)}$. We will see more details on that later.

§ 5.1 A Generic Descent Method

We begin by considering the following model algorithm of a generic line-search descent method:

Algorithm 5.2 (Generic line-search descent method)**.**
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* f *and* f' *(or* ∇f*)*
Output: *approximate stationary point of* **(UP)**

1: *Set* $k := 0$
2: ***while*** *stopping criterion not met* ***do***
3: *Determine a search direction* $d^{(k)}$ *such that* $f'(x^{(k)})\, d^{(k)} < 0$ // *descent direction*
4: *Choose a step size* $\alpha^{(k)} > 0$ *such that* $f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$ // *obtain descent*
5: *Set* $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$ // *take the step*
6: *Set* $k := k + 1$
7: ***end while***
8: ***return*** $x^{(k)}$

In order to analyze the convergence properties of this generic algorithm and to determine further requirements for the descent directions and step sizes, we ignore the stopping criterion for now, so that Algorithm 5.2 produces infinite sequences of iterates $x^{(k)}$, search directions $d^{(k)}$ and step sizes $\alpha^{(k)}$. In practice, of course, we will use a stopping criterion to be discussed later.

[16] The solution set of the linear system (5.4) is either the empty set or an affine subspace of $\mathbb{R}^n$ whose dimension agrees with the dimension of $\ker H^{(k)}$.

[17] A stationary point x of f is called a **saddle point** of f if the Hessian $f''(x)$ is indefinite, i. e., has at least one positive and at least one negative eigenvalue.

We will see that, in general, we cannot expect the iterates $x^{(k)}$ to converge overall, but there may be convergent subsequences with different limit points (although this rarely occurs in practice). We recall that the limit points of convergent subsequences $\left(x^{(k^{(\ell)})}\right)$ are precisely the **accumulation points** of $\left(x^{(k)}\right)$.

We would like the accumulation points of the sequence of iterates$\{x^{(k)}\}$ to be "special" points. Therefore, it would be desirable to have the following property:

$$\text{When } x^* \text{ is an accumulation of } \left(x^{(k)}\right), \text{ then } f'(x^*) = 0, \text{ i. e., } x^* \text{ is stationary.} \tag{5.5}$$

The relatively weak property (5.5) is often referred to as the **global convergence** of an algorithm. In particular, global convergence does not mean that one obtains a global minimizer. By contrast, it means that one obtains a convergence result (5.5) that is valid for arbitrary initial guesses $x^{(0)}$. Notice that (5.5) does not assert that an accumulation point even exists.[18] It turns out that, in general, we cannot expect more. Under additional assumptions on f, one may be able to show stronger results, for instance

$$\|\nabla f(x^{(k)})\| \text{ has an accumulation point at } 0. \tag{5.6a}$$
$$\text{The entire sequence } \|\nabla f(x^{(k)})\| \text{ converges to } 0. \tag{5.6b}$$
$$\text{Accumulation points of } \left(x^{(k)}\right) \text{ are stationary.} \tag{5.6c}$$
$$\text{The entire sequence } \left(x^{(k)}\right) \text{ converges to a stationary point.} \tag{5.6d}$$
$$\text{The entire sequence } \left(x^{(k)}\right) \text{ converges to a local miminizer.} \tag{5.6e}$$

We will now investigate the minimal requirements on the search directions $d^{(k)}$ and step sizes $\alpha^{(k)}$ in Algorithm 5.2 that ensure global convergence in the sense of (5.5). To this end, two properties are essential:

(1) The search directions $d^{(k)}$ are "good descent directions".

(2) The step sizes $\alpha^{(k)}$ are chosen so that the achievable descent along the search direction $d^{(k)}$ is "sufficiently exploited".

We use the user-defined M-inner product in the space of optimization variables and search directions $\mathbb{R}^n$. Since all norms in $\mathbb{R}^n$ are equivalent, all concepts and properties of algorithms in the remainder of § 5 are *qualitatively independent* of the choice of M. However, the choice of M is still important through its impact on the convergence properties and stopping criteria.

Requirements on the Descent Directions

Definition 5.3 (Admissible search directions)**.** *Suppose that $x^{(k)}$ and $d^{(k)}$ the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. The sequence $d^{(k)}$ of search*

[18] Indeed, an example such as $f(x) = x$ for $x \in \mathbb{R}$ shows that any algorithm with the global convergence property (5.5) couldn't produce an accumulation point, since f has no stationary point.

directions is termed ***admissible*** *in case*

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \to 0 \quad \Rightarrow \quad f'(x^{(k)}) \to 0. \tag{5.7}$$

Note: The admissibility is a property that the sequence of search directions generated by a particular algorithm, applied to a particular problem (objective), started from a particular initial guess may or may not possess. One is, of course, interested in designing algorithms which generate admissible search directions for arbitrary objectives f and initial guesses $x^{(0)}$.

The expression $\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M}$ is the directional derivative of f at $x^{(k)}$ in the direction $d^{(k)}$ normalized. Therefore, we can interpret the condition (5.7) as follows: when the directional derivatives in the normalized search directions converge to zero, then it is due to the derivatives converging to zero and not due to the search directions becoming inefficient (which would be the case if they become essentially M-orthogonal to the steepest descent direction $-\nabla_M f$). This reflects our first goal (item (1) above) that the search directions are "good descent directions".

Condition (5.7) is purely qualitative. By contrast, the **angle condition**

$$\cos \measuredangle \Big(\underbrace{-\nabla_M f(x^{(k)})}_{\text{steepest descent direction}}, \overbrace{d^{(k)}}^{\text{chosen search direction}} \Big) = \frac{(-\nabla_M f(x^{(k)}), d^{(k)})_M}{\|\nabla_M f(x^{(k)})\|_M \, \|d^{(k)}\|_M} = \frac{-f'(x^{(k)})\, d^{(k)}}{\|f'(x^{(k)})\|_{M^{-1}} \|d^{(k)}\|_M} \geq \eta \tag{5.8}$$

with some $\eta \in (0,1)$ is a stronger, quantitative condition, which is moreover easy to verify. It means that the angles (as measured in the M-inner product) between the chosen search directions $d^{(k)}$ and the directions of steepest descent $-\nabla_M f(x^{(k)})$ are uniformly bounded away from 90°.

Lemma 5.4 (Angle condition implies admissibility)**.** *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. If the angle condition (5.8) holds with some $\eta \in (0,1)$, then the sequence $d^{(k)}$ of search directions is admissible.*

Proof. We have

$$f'(x^{(k)})\, d^{(k)} = \big(\nabla f(x^{(k)}), d^{(k)}\big) = \big(\nabla_M f(x^{(k)}), d^{(k)}\big)_M.$$

The angle condition (5.8) implies

$$-\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \geq \eta \, \|\nabla_M f(x^{(k)})\|_M = \eta \, \|f'(x^{(k)})^{\mathsf{T}}\|_{M^{-1}} \geq 0.$$

When the left-hand term goes to zero, then $f'(x^{(k)})$ must go to zero as well. □

As we already mentioned, almost all of the algorithms we will discuss in detail determine their search directions from the solutions of linear systems

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) \tag{5.4}$$

with a symmetric and possibly positive definite matrix $H^{(k)}$, the model Hessian. In the s. p. d. case, in view of

$$f'(x^{(k)})\, d^{(k)} = -f'(x^{(k)})\left[(H^{(k)})^{-1}\, \nabla f(x^{(k)})\right] = -\nabla f(x^{(k)})^{\mathsf{T}}\, (H^{(k)})^{-1}\, \nabla f(x^{(k)}) < 0, \tag{5.9}$$

$d^{(k)}$ is a descent direction as long as $f'(x^{(k)}) \neq 0$ holds. However, when $H^{(k)}$ is not positive definite, then $d^{(k)}$ may fail to be a descent direction.

In the s. p. d. case, we can show that as long as the sequence of model Hessians remains "well behaved", the sequence of search directions satisfies the angle condition (5.8) and thus is admissible as well.

Lemma 5.5 (Bounded condition numbers imply the angle condition[19]). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. Suppose that the search directions are obtained from (5.4), where $H^{(k)} \in \mathbb{R}^{n\times n}$ is a sequence of s. p. d. model Hessians. Suppose, moreover, that the generalized condition numbers of $H^{(k)}$ w.r.t. M satisfy*

$$\kappa(H^{(k)}; M) := \frac{\lambda_{\max}(H^{(k)}; M)}{\lambda_{\min}(H^{(k)}; M)} \leq \overline{\kappa}.$$

Then the sequence of search directions $d^{(k)}$ satisfies the angle condition (5.8) with

$$\eta = \frac{2\sqrt{\overline{\kappa}}}{\overline{\kappa}+1} \geq \frac{1}{\sqrt{\overline{\kappa}}}.$$

Proof. We perform a couple of equivalent reformulations of the claim to obtain

$$\begin{aligned}
& -\nabla f(x^{(k)})^{\mathsf{T}} d^{(k)} \geq \frac{2\sqrt{\kappa}}{\kappa+1} \|\nabla_M f(x^{(k)})\|_M \, \|d^{(k)}\|_M \\
\Leftrightarrow\quad & (d^{(k)})^{\mathsf{T}} H^{(k)} d^{(k)} \geq \frac{2\sqrt{\kappa}}{\kappa+1} \|M^{-1} H^{(k)} d^{(k)}\|_M \, \|d^{(k)}\|_M \qquad \text{since } H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) \\
\Leftrightarrow\quad & ((d^{(k)})^{\mathsf{T}} H^{(k)} d^{(k)})^2 \geq \frac{4\kappa}{(\kappa+1)^2} \|M^{-1} H^{(k)} d^{(k)}\|_M^2 \, \|d^{(k)}\|_M^2 \\
\Leftrightarrow\quad & \frac{((d^{(k)})^{\mathsf{T}} H^{(k)} M^{-1} H^{(k)} d^{(k)})\, ((d^{(k)})^{\mathsf{T}} M\, d^{(k)})}{((d^{(k)})^{\mathsf{T}} H^{(k)} d^{(k)})^2} \leq \frac{(\kappa+1)^2}{4\kappa}.
\end{aligned}$$

The statement in the previous line, however, is true due to the generalized Kantorovich inequality (Corollary 2.2). □

We summarize our findings on search directions:

$$\begin{aligned}
& \text{the model Hessians } H^{(k)} \text{ have bounded condition numbers} \\
\Rightarrow\quad & \text{the angle condition (5.8) holds} \\
\Rightarrow\quad & \text{the search directions are admissible (5.7).}
\end{aligned}$$

[19] In the literature, one often finds this result only in the case $M = \mathrm{Id}$, and with the non-optimal bound $\eta = \frac{1}{\kappa}$; see for instance Ulbrich, Ulbrich, 2012, S.32 or Nocedal, Wright, 2006, eq.(3.19).

Requirements on the Step Sizes

We now address the step sizes $\alpha^{(k)}$. The following example shows that the mere requirement

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$$

is not sufficient to obtain a reasonable convergence behavior.

Example 5.6 (Too small step sizes[20]). *Consider the objective $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, initial guess $x^{(0)} = 1$, search directions $d^{(k)} = -1$ and the Euclidean inner product $M = 1$. With step sizes $\alpha^{(k)} = \left(\frac{1}{2}\right)^{k+2}$, we obtain the sequences of iterates according to*

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)}\,(-1) = x^{(0)} - \sum_{i=0}^{k} \left(\tfrac{1}{2}\right)^{i+2} = \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^{k+2}.$$

This implies $x^{(k+1)} < x^{(k)}$ and $f(x^{(k+1)}) < f(x^{(k)})$. However, $x^{(k)} \to x^* = 1/2$, which is not a stationary point of f.

The step sizes in the previous example are too small and thus they violate our second goal (item (2) above) since they do not exploit the achievable descent sufficiently well. We therefore introduce the following qualitative condition on the step sizes.

Definition 5.7 (Admissible step sizes). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. The sequence $\alpha^{(k)}$ of step sizes is termed **admissible** in case*

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) \quad \text{for all } k \in \mathbb{N}_0, \tag{5.10a}$$

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \to 0 \quad \Rightarrow \quad \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \to 0. \tag{5.10b}$$

We can interpret (5.10b) as follows: when the progress in the objective values converges to zero, then it is due to the normalized directional derivatives converging to zero and not due to the step sizes becoming too small. In other words, admissible step sizes do make sufficient use of the descent available in the direction $d^{(k)}$.

Condition (5.10) is purely qualitative. By contrast, the condition that the step sizes be **efficient**, i. e., there exists $\theta > 0$ such that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) - \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \tag{5.11}$$

for all $k \in \mathbb{N}_0$ is a stronger, quantitative condition, which is moreover easy to verify.

[20] from Alt, 2002, Beispiel 4.4.1

Lemma 5.8 (Efficiency implies admissibility). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. If the sequence of step sizes $\alpha^{(k)}$ is efficient, then it is also admissible.*

Proof. Suppose that $\alpha^{(k)}$ is efficient, i. e.,

$$0 \le \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \le f(x^{(k)}) - f(x^{(k)} + \alpha^{(k)} d^{(k)})$$

Therefore (5.10a) is clear. To show (5.10b), suppose

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \to 0.$$

Since θ is strictly positive, this implies

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \to 0,$$

which confirms (5.10b). □

Using the assumptions of admissible search directions and admissible step sizes, we will obtain a theorem (see Theorem 5.9 below) on the global convergence of Algorithm 5.2. However, in view of the expected convergence result (5.5), we will have to work with accumulation points (limits of subsequences) of the iterates. This means that we should refine the notion of admissible search directions (5.7), the notions of admissible step sizes (5.10) as well as efficient step sizes (5.11) to subsequences. We denote such subsequences here with $\left(x^{(k)}\right)_{k\in K}$, where $K \subseteq \mathbb{N}_0$ is an infinite subset of the index set $\mathbb{N}_0$. (**Quiz 5.2:** How does this notation relate to the notation for subsequences $\left(x^{(k^{(\ell)})}\right)$ introduced in § 2.7?)

In detail, the refined conditions on subsequences read as follows:

admissible search directions:

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0 \quad \Rightarrow \quad f'(x^{(k)}) \xrightarrow{k\in K} 0, \tag{5.7'}$$

angle condition:

$$\frac{-f'(x^{(k)})\, d^{(k)}}{\|\nabla_M f(x^{(k)})\|_M \, \|d^{(k)}\|_M} \ge \eta \quad \text{for all } k \in K \tag{5.8'}$$

admissible step sizes:

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) \quad \text{for all } k \in \mathbb{N}_0, \tag{5.10a'}$$

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \xrightarrow{k\in\mathbb{N}_0} 0 \quad \Rightarrow \quad \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0, \tag{5.10b'}$$

efficient step sizes:

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \leq f(x^{(k)}) - \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \quad \text{for all } k \in K. \tag{5.11'}$$

The statements of Lemma 5.4 and Lemma 5.5 continue to hold when restricted to subsequences. For the analog of Lemma 5.8, we have to make (5.10a') an assumption rather than a conclusion.

We now show a global convergence theorem for the model Algorithm 5.2.

Theorem 5.9 (Global convergence of model Algorithm 5.2)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates $x^{(k)}$, search directions $d^{(k)} \neq 0$ and step sizes $\alpha^{(k)}$. Suppose that x^* is an accumulation point of $x^{(k)}$ and that $\left(x^{(k)}\right)_{k \in K}$ is a subsequence converging to x^*. Finally, suppose that the subsequences $\left(d^{(k)}\right)_{k \in K}$ and $\left(\alpha^{(k)}\right)_{k \in K}$ of search directions and step sizes are both admissible. Then $f'(x^*) = 0$.*

Note: In other words, when a generic descent algorithm (Algorithm 5.2) produces admissible search directions and admissible step sizes, then any accumulation point of the iterates is stationary.

Quiz 5.3: What goes wrong in Example 5.6?

Proof. Due to the continuity of f, we have $f(x^{(k)}) \xrightarrow{k \in K} f(x^*)$. Moreover, by admissibility of the step sizes (5.10a'), the entire sequence $f(x^{(k)})$ is monotone decreasing. Therefore, the entire sequence in fact converges: $f(x^{(k)}) \to f(x^*)$. Consequently, we also have

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \to 0.$$

The admissibility of step sizes along the subsequence, (5.10b'), implies

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|} \xrightarrow{k \in K} 0.$$

Since the search directions along the subsequence are in turn admissible, (5.7'), we can conclude

$$f'(x^{(k)}) \xrightarrow{k \in K} 0.$$

On the other hand, since f is of class C^1, we also have

$$f'(x^{(k)}) \xrightarrow{k \in K} f'(x^*).$$

This shows $f'(x^*) = 0$. □

§ 5.2 Step Size Strategies

In this section we will see how efficient step sizes (5.11) or at least admissible step sizes (5.10) can be found in general.

Armijo Backtracking Line Search

The Armijo backtracking line search is the simplest step size strategy and it is sufficient in many situations. Suppose that $d^{(k)}$ is a descent direction for f at $x^{(k)}$. In order to obtain sufficient decrease, the **Armijo condition** requires that the step size α satisfy

$$f(x^{(k)} + \alpha\, d^{(k)}) \leq f(x^{(k)}) + \sigma\, \alpha\, f'(x^{(k)})\, d^{(k)} \tag{5.12}$$

holds. Here $\sigma \in (0,1)$ is the given **Armijo parameter**. Using the auxiliary function (**line search function**)

$$\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$$

to simplify notation, we can write the Armijo condition (5.12) equivalently in the form

$$\varphi(\alpha) \leq \varphi(0) + \sigma\, \alpha\, \varphi'(0). \tag{5.12}$$

Step sizes $\alpha \geq 0$ which satify (5.12) are termed **Armijo step sizes**. Condition (5.12) requires that the step size α realizes at least the σ-fraction of the first-order descent suggested by the tangent of φ at $\alpha = 0$.

Notice that due to the chain rule, φ inherits the C^1 property of f, and we have

$$\varphi'(\alpha) = f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)} \tag{5.13a}$$

$$\text{and, in particular,} \quad \varphi'(0) = f'(x^{(k)})\, d^{(k)}. \tag{5.13b}$$

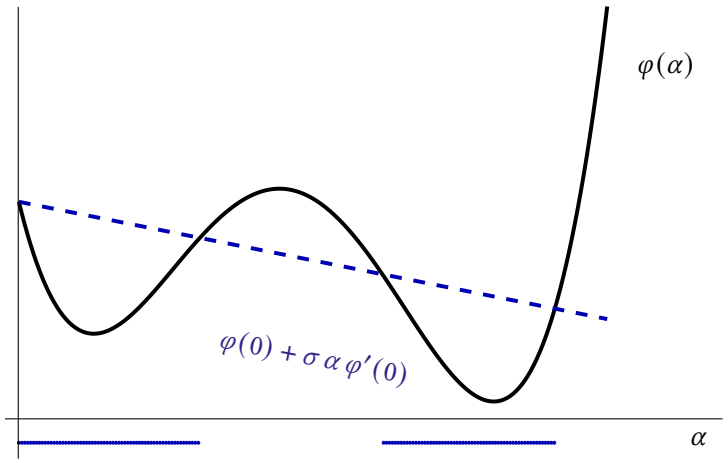


Figure 5.1: Illustration of step sizes $\alpha \geq 0$ satisfying the Armijo condition (5.12) (blue). As an example, the Armijo parameter is chosen as $\sigma = 0.05$.

We will now answer the question whether Armijo step sizes exist, and how to find them.

Lemma 5.10 (Existence of Armijo step sizes)**.** *Suppose that d is a descent direction for f at x, and that the Armijo parameter satisfies $\sigma \in (0,1)$. Then there exists $\overline{\alpha} > 0$ such that (5.12) holds for all $\alpha \in [0, \overline{\alpha}]$.*

Proof. φ' is continuous at 0, which implies that there exists $\overline{\alpha} > 0$ such that

$$\varphi'(\alpha) < \sigma\,\varphi'(0) \quad \text{holds for all } \alpha \in [0, \overline{\alpha}].$$

From Taylor's theorem 2.4 we obtain that there exists $\xi \in [0, \alpha]$ such that

$$\begin{aligned}\varphi(\alpha) &= \varphi(0) + \alpha\,\varphi'(\xi)\\ &\leq \varphi(0) + \sigma\,\alpha\,\varphi'(0).\end{aligned}$$

Therefore, the Armijo condition (5.12) holds for all $\alpha \in [0, \overline{\alpha}]$. □

We have seen that the Armijo condition is always satisfied in an interval starting at $\alpha = 0$. However, we need to select a step size which is not too small, as demonstrated by Example 5.6. This can be achieved by a **backtracking strategy**: run through a sequence of trial step sizes from large to small until the Armijo conditon (5.12) is satisfied for the first time.

Algorithm 5.11 (Armijo backtracking line search)**.**

Input: *initial trial step size* α
Input: *routine to evaluate* φ
Input: *pre-computed function values* $\varphi(0)$ *and* $\varphi'(0)$
Input: *Armijo parameter* $\sigma \in (0,1)$
Input: *backtracking parameter* $\beta \in (0,1)$
Output: *step size* α *satisfying the Armijo condition* (5.12)

1: *Set* $\ell := 0$
2: ***while*** *Armijo condition* (5.12) *does not hold for* α ***do***
3: *Set* $\alpha := \beta\,\alpha$ // *new trial step size*
4: *Set* $\ell := \ell + 1$
5: ***end while***
6: ***return*** α

Remark 5.12 (on Algorithm 5.11)**.**

(*i*) *In Algorithm 5.11, we did not number the trial step sizes* $\alpha^{(0)}, \alpha^{(1)}, \ldots$ *by an index in order to avoid confusion with the step size* $\alpha^{(k)}$ *which eventually gets used in the* k*-th iteration of the outer algorithm (Algorithm 5.2).*

(*ii*) *Every trial step size that fails to satisfy the Armijo condition "costs" one additional evaluation of* φ*, i. e., one additional evaluation of* f*.*

(*iii*) *The Armijo parameter is often chosen to be small, e. g.,* $\sigma = 10^{-2}$ *or even* $\sigma = 10^{-4}$*. A typical value for the backtracking parameter is* $\beta = 1/2$*.*

(*iv*) *It follows from Lemma 5.10 that Algorithm 5.11 terminates after finitely many iterations with a successful trial step size* $\alpha \geq \overline{\alpha}\,\beta$*. (Recall that* $\overline{\alpha}$ *is the upper bound of any interval* $[0, \overline{\alpha}]$ *containing only Armijo step sizes.)*

(v) *In a practical implementation, one often adds further checks and stopping criteria to Algorithm 5.11. For instance, we need to safeguard against $\varphi'(0) \geq 0$ (d is not a descent direction) and against too many unsuccessful trial steps.*

Suitable values for the initial trial step size α in Algorithm 5.11 depend on how the search directions $d^{(k)}$ are generated in the outer method. We will see more on that when we discuss concrete instances of Algorithm 5.2. Since the backtracking strategy only shortens the initial trial step size, we need to ensure that the initial trial step size is sufficiently large in order to obtain admissible step sizes that exploit the achievable descent sufficiently well. This is what the following result is about.

Lemma 5.13 (Armijo backtracking line search produces admissible step sizes)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates $x^{(k)}$ and search (descent) directions $d^{(k)} \neq 0$. Suppose moreover that the step sizes $\alpha^{(k)}$ are obtained by the Armijo backtracking line search (Algorithm 5.11) with initial trial step size $\alpha^{(k,0)}$. Assume that $K \subseteq \mathbb{N}_0$ is an infinite index set such that the subsequence $\left(x^{(k)}\right)_{k\in K}$ is bounded. Finally, suppose that $\psi\colon [0,\infty) \to [0,\infty)$ is any monotone increasing function and that the initial trial step sizes satisfy*

$$\alpha^{(k,0)} \, \|d^{(k)}\|_M \geq \psi\Big(\frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M}\Big) \quad \text{for all } k \in K. \tag{5.14}$$

Then the step sizes $\left(\alpha^{(k)}\right)_{k\in K}$ are admissible.

Proof. We need to show (5.10a') and (5.10b'). The first condition is a direct consequence of the Armijo condition holding at $\alpha^{(k)} > 0$

$$f(x^{(k)} + \alpha^{(k)}\, d^{(k)}) \leq f(x^{(k)}) + \sigma\, \alpha^{(k)} \underbrace{f'(x^{(k)})\, d^{(k)}}_{<0},$$

the fact that $d^{(k)}$ is a descent direction and that σ is positive. It remains to verify (5.10b').

By assumption, the sequence $\left(x^{(k)}\right)_{k\in K}$ is bounded. Therefore, it has a convergent subsequence with index set K'. By continuity of f, $\left(f(x^{(k)})\right)_{k\in K}$ converges. Due to the Armijo condition (5.12), the sequence $f(x^{(k)})$ is monotone decreasing, so that in fact the entire sequence $f(x^{(k)})$ converges. From there and the Armijo condition (5.12) we conclude

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \leq \sigma\, \alpha^{(k)} f'(x^{(k)})\, d^{(k)} < 0.$$

The left-hand side converges to 0, therefore we must have

$$\alpha^{(k)} f'(x^{(k)})\, d^{(k)} \to 0. \tag{$*$}$$

In order to verify (5.10b'), we need to show

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0.$$

In the remainder of the proof, we distinguish indices $k \in K$ according to the following cases:

When $\alpha^{(k)}\|d^{(k)}\|_M$ is "large", then $\dfrac{\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)}\|d^{(k)}\|_M}$ is small.

When $\alpha^{(k)}\|d^{(k)}\|_M$ is "small", then $\begin{cases} \text{we use the assumption (5.14)} & \text{in case } \alpha^{(k)} = \alpha^{(k,0)}. \\ \text{we use the Armijo condition (5.12)} & \text{in case } \alpha^{(k)} < \alpha^{(k,0)}. \end{cases}$

By assumption, the sequence $\big(x^{(k)}\big)_{k\in K}$ is bounded, hence the continuous function f' is uniformly continuous "near the $\big(x^{(k)}\big)_{k\in K}$". More precisely, suppose that $R > 0$ is any fixed number, then f' is uniformly continuous on the compact set

$$A_R := \operatorname{cl} \bigcup_{k\in K} B_R^M(x^{(k)}).$$

(**Quiz 5.4:** Why is this set compact?) Now suppose that $\varepsilon > 0$ is given. Then there exists $\overline{\delta} > 0$ such that

$$\|f'(y) - f'(z)\|_{M^{-1}} \le (1-\sigma)\,\varepsilon$$

holds for all $y, z \in A_R$ such that $\|y - z\|_M \le \overline{\delta}$. Possibly by making $\overline{\delta}$ smaller, we can assume $\overline{\delta} \le R$. Thus, in particular, we obtain

$$\Big\|f'(\underbrace{x^{(k)} + e}_{\in A_R}) - f'(\underbrace{x^{(k)}}_{\in A_R})\Big\|_{M^{-1}} \le (1-\sigma)\,\varepsilon \quad \text{for all } k \in K,\ \|e\|_M \le \overline{\delta}. \tag{$**$}$$

We now set

$$\delta := \min\big\{\overline{\delta}\,\beta,\ \psi(\varepsilon)\big\} \in (0, \overline{\delta}).$$

Due to the convergence in ($*$), there exists an index $k_0 \in \mathbb{N}_0$ such that

$$\alpha^{(k)}\big|f'(x^{(k)})\, d^{(k)}\big| \le \varepsilon\,\delta \quad \text{holds for all } k \ge k_0. \tag{$***$}$$

From now on, let $k \in K$, $k \ge k_0$, be arbitrary. We are going to show that

$$0 \le \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon$$

holds, which proves (5.10b'). We distinguish the following cases, as anticipated above:

Case 1: $\alpha^{(k)}\|d^{(k)}\|_M \ge \delta$
In this case we immediately conclude

$$\begin{aligned} 0 &\le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} && \text{since } d^{(k)} \text{ is a descent direction} \\ &= \frac{-\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)}\|d^{(k)}\|_M} \\ &\le \frac{\varepsilon\,\delta}{\delta} && \text{by } (***) \text{ and the assumption in case 1} \\ &= \varepsilon. \end{aligned}$$

Case 2: $\alpha^{(k)}\|d^{(k)}\|_M < \delta$ and $\alpha^{(k)} = \alpha^{(k,0)}$
We obtain

$$\begin{aligned}\psi\Big(\frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M}\Big) &\le \alpha_{k,0}\,\|d^{(k)}\|_M && \text{by assumption (5.14)}\\ &< \delta && \text{by the assumption in case 2}\\ &\le \psi(\varepsilon) && \text{by the choice of } \delta.\end{aligned}$$

Since ψ is monotone increasing, we conclude

$$0 \le \frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

Case 3: $\alpha^{(k)}\|d^{(k)}\|_M < \delta$ and $\alpha^{(k)} < \alpha^{(k,0)}$
The assumption $\alpha^{(k)} < \alpha^{(k,0)}$ means that the initial trial step size (and possibly some of the subsequent trial step sizes) did not satisfy the Armijo condition. Since $\alpha^{(k)}$ was the first trial step size to satisfy the Armijo condition (5.12), the previous trial step size, $\beta^{-1}\alpha^{(k)}$, violated it:

$$\sigma\,\beta^{-1}\alpha^{(k)} f'(x^{(k)})d^{(k)} < f\big(x^{(k)} + \beta^{-1}\alpha^{(k)}d^{(k)}\big) - f(x^{(k)}).$$

By Taylor's theorem 2.4, there exists $\xi^{(k)} \in (0,1)$ such that

$$\sigma\,\beta^{-1}\alpha^{(k)} f'(x^{(k)})d^{(k)} < \beta^{-1}\alpha^{(k)}\, f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\,\xi^{(k)}d^{(k)}\big)\,d^{(k)}$$

and thus

$$\begin{aligned}\sigma\, f'(x^{(k)})\,d^{(k)} &< f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\,\xi^{(k)}d^{(k)}\big)\,d^{(k)}\\ &= f'(x^{(k)})\,d^{(k)} + \big[f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\,\xi^{(k)}d^{(k)}\big) - f'(x^{(k)})\big]d^{(k)}\\ &\le f'(x^{(k)})\,d^{(k)} + \big\|f'\big(x^{(k)} + \underbrace{\beta^{-1}\alpha^{(k)}\,\xi^{(k)}d^{(k)}}_{=:e^{(k)}}\big) - f'(x^{(k)})\big\|_{M^{-1}}\,\|d^{(k)}\|_M.\end{aligned}$$

The vector $e^{(k)}$ satisfies

$$\begin{aligned}\|e^{(k)}\|_M &= \beta^{-1}\alpha^{(k)}\xi^{(k)}\|d^{(k)}\|_M\\ &< \beta^{-1}\delta && \text{by the assumption in case 3 and since } \xi^{(k)} \in (0,1)\\ &\le \overline{\delta} && \text{by the choice of } \delta.\end{aligned}$$

We may thus apply estimate (**) to the inequality above to obtain

$$\sigma\, f'(x^{(k)})\,d^{(k)} \le f'(x^{(k)})\,d^{(k)} + (1-\sigma)\,\varepsilon\,\|d^{(k)}\|_M.$$

Sorting terms and dividing by $\|d^{(k)}\|_M$ finally yields

$$0 \le \frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

□

Remark 5.14 (Armijo backtracking line search produces efficient step sizes). *When we choose $\psi(z) = c\,z$ with some $c > 0$, i. e., when we use initial trial step sizes satisfying*

$$\alpha_{k,0}\,\|d^{(k)}\|_M \geq c\,\frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M}, \tag{5.15}$$

and if f' is Lipschitz continuous on the sublevel set $\mathcal{M}_f(x^{(0)}) := \{x \in \mathbb{R}^n \mid f(x) \leq f(x^{(0)})\}$, then one can show that Algorithm 5.11 produces not only admissible, but efficient step sizes.

To conclude the presentation of Armijo backtracking strategies, we consider a modification of Algorithm 5.11 which often produces trial step sizes more effectively than simple backtracking $\alpha \rightsquigarrow \beta\,\alpha$ in case the Armijo condition fails on the initial trial step size.

The modification is based on the fact that we have available the data of the line search function φ

$$\varphi(0), \quad \varphi'(0) < 0 \quad \text{and} \quad \varphi(\alpha)$$

for the current trial step size α. Using this data, we can fit a quadratic polynomial

$$p(\alpha) = a + b\,\alpha + c\,\alpha^2.$$

The conditions[21] $p(0) = \varphi(0)$, $p'(0) = \varphi'(0)$ and $p(\alpha) = \varphi(\alpha)$ uniquely define the coefficients

$$a = \varphi(0), \qquad b = \varphi'(0), \qquad c = \frac{1}{\alpha^2}\big(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha\big). \tag{5.16}$$

Naturally, this quadratic model of φ will be used only when the Armijo condition (5.12) failed at the trial step size α, i. e., in case

$$\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha > \varphi(\alpha) - \varphi(0) - \sigma\,\varphi'(0)\,\alpha > 0$$

holds, which implies $c > 0$. This in turn means that the unique global minimizer $\alpha^* = -\frac{b}{2c}$ of p satisfies

$$\alpha^* = \frac{-\varphi'(0)\,\alpha^2}{2\,\big(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha\big)} > 0.$$

We then choose α^* as the next trial step size α^+, but in order to avoid drastic changes or even an increase from α to α^+, we clip α^* to the interval $[\underline{\beta}\,\alpha, \overline{\beta}\,\alpha]$ according to

$$\alpha^+ := \min\big\{\max\{\alpha^*, \underline{\beta}\,\alpha\}, \overline{\beta}\,\alpha\big\} = \begin{cases} \underline{\beta}\,\alpha, & \text{if } \alpha^* < \underline{\beta}\,\alpha, \\ \alpha^*, & \text{if } \underline{\beta}\,\alpha \leq \alpha^* \leq \overline{\beta}\,\alpha, \\ \overline{\beta}\,\alpha, & \text{if } \alpha^* > \overline{\beta}\,\alpha, \end{cases}$$

where $0 < \underline{\beta} < \overline{\beta} < 1$ are the clipping parameters.[22] This modified Armijo backtracking line search maintains the essential properties of the simple Armijo backtracking line search. In particular, the admissibility (and potentially efficiency) of the accepted step sizes (see Lemma 5.13 and Remark 5.14) continue to hold.

For completeness, we present the modified Armijo backtracking line search procedure in Algorithm 5.15.

[21] Fitting a polynomial using function values and derivatives is known as **Hermite interpolation**. Using function values only is known as **Lagrange interpolation**.

[22] Using $\underline{\beta} = \overline{\beta} = \beta$ we get back our previous simple backtracking strategy where $\alpha^+ = \beta\,\alpha$.

Algorithm 5.15 (Modified Armijo backtracking line search with interpolation).
Input: *initial trial step size* α
Input: *routine to evaluate* φ
Input: *pre-computed function values* $\varphi(0)$ *and* $\varphi'(0)$
Input: *Armijo parameter* $\sigma \in (0,1)$
Input: *backtracking parameters* $0 < \underline{\beta} < \overline{\beta} < 1$
Output: *step size* α *satisfying the Armijo condition* (5.12)

1: *Set* $\ell := 0$
2: ***while*** *Armijo condition* (5.12) *does not hold for* α ***do***
3: *Set* $\alpha^* := \dfrac{-\varphi'(0)\,\alpha^2}{2\,(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha)}$ *// minimizer of quadratic polynomial*
4: *Set* $\alpha := \min\{\max\{\alpha^*, \underline{\beta}\,\alpha\}, \overline{\beta}\,\alpha\}$ *// clip it and use as new trial step size*
5: *Setze* $\ell := \ell + 1$
6: ***end while***
7: ***return*** α

Wolfe-Powell Line Search

Recall from Lemma 5.10 that the Armijo condition

$$f(x^{(k)} + \alpha\, d^{(k)}) \le f(x^{(k)}) + \sigma\,\alpha\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi(\alpha) \le \varphi(0) + \sigma\,\alpha\,\varphi'(0) \tag{5.12}$$

always holds in some interval $[0, \overline{\alpha}]$. Therefore, we combined the Armijo condition with backtracking, where we generate trial step sizes from large to small, in order to avoid overly small step sizes.

Alternatively, we could require, in addition to (5.12), the **curvature condition**

$$f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)} \ge \tau\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi'(\alpha) \ge \tau\,\varphi'(0) \tag{5.17}$$

or even the **strong curvature condition**

$$|f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)}| \le -\tau\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad |\varphi'(\alpha)| \le -\tau\,\varphi'(0) \tag{5.18}$$

to hold, where $\tau \in (\sigma, 1)$ is the **curvature parameter**. The curvature condition (5.17) demands that the derivative of φ at α is not too negative, namely that it is larger (has less descent) than at $\alpha = 0$. However, it would be fine for φ to increase near α; see Figure 5.2. This curvature condition already avoids too small step sizes α near 0.

The strong curvature condition (5.18) demands that, in addition, the derivative of φ at α it not too positive either. The condition can be interpreted as the requirement that α be an approximately stationary point of φ. **Note:** When α is a local minimizer of φ, then (5.18) holds with $\tau = 0$.

The Armijo condition (5.12) and the curvature condition (5.17) together are referred to as the **Wolfe-Powell conditions**. The Armijo condition (5.12) and the strong curvature condition (5.18) together are referred to as the **strong Wolfe-Powell conditions**. Consequently, step sizes $\alpha \ge 0$ which satisfy the above conditions are referred to as **Wolfe-Powell step sizes** and **strong Wolfe-Powell step sizes**, respectively.

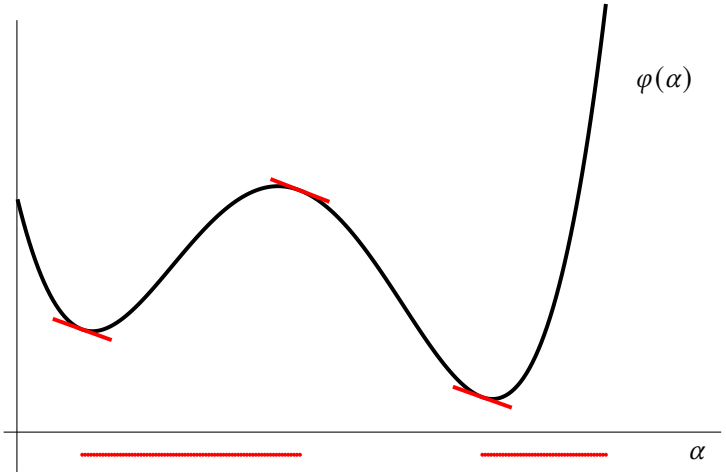


Figure 5.2: Illustration of step sizes $\alpha \geq 0$ satisfying the curvature condition (5.17) (red). As an example, the curvature parameter is chosen as $\tau = 0.1$.

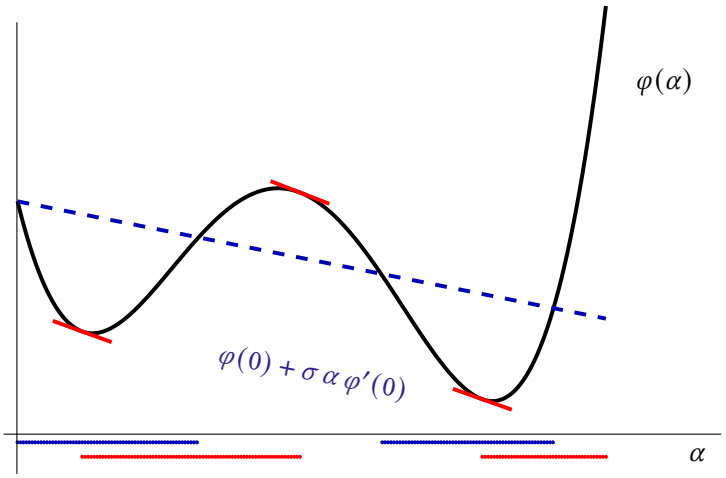


Figure 5.3: Illustration of step sizes $\alpha \geq 0$ satisfying both the Armijo condition (5.12) (blue) and the curvature condition (5.17) (red). As an example, the Armijo parameter is chosen as $\sigma = 0.05$ and the curvature parameter is chosen as $\tau = 0.1$.

A simple example such as $\varphi(\alpha) = -\alpha$ shows that the curvature condition may not be satisfiable without further assumptions on f. The following result gives a sufficient condition for strong Wolfe-Powell step sizes to exist.

Lemma 5.16 (Existence of (strong) Wolfe-Powell step sizes)**.** *Suppose that d is a descent direction for f at x and that the Armijo and curvature parameters satisfy $0 < \sigma < \tau < 1$. Suppose, moreover, that f is bounded below on the ray $\{x + \alpha d \mid \alpha \geq 0\}$. Then there exists a step size $\alpha_2 > 0$ such that the strong Wolfe-Powell conditions (5.12) and (5.18) (and thus also the regular Wolfe-Powell conditions (5.12) and (5.17)) hold in a neighborhood of α_2.*

Proof. We abbreviate as usual $\varphi(\alpha) := f(x + \alpha d)$. Since by assumption, φ is bounded below on $\mathbb{R}_{\geq 0}$, φ

intersects the Armijo line

$$\alpha \mapsto \varphi(0) + \sigma \underbrace{\varphi'(0)}_{<0} \alpha,$$

which is unbounded below, in at least one positive point. Suppose that α_1 is the smallest positive point of intersection (**Quiz 5.5:** Why does α_1 exist?). Then we have

$$\varphi(\alpha_1) = \varphi(0) + \sigma\, \varphi'(0)\, \alpha_1.$$

In view of $\varphi'(0) < 0$, the Armijo condition (5.12) holds for all $\alpha \in [0, \alpha_1]$, i. e., the Armijo line lies below φ on this interval. From the mean value theorem 2.4, we infer the existence of $\alpha_2 \in (0, \alpha_1)$ such that

$$\varphi'(\alpha_2) = \frac{\varphi(\alpha_1) - \varphi(0)}{\alpha_1} = \sigma\, \varphi'(0).$$

And thus we obtain the strong curvature condition (5.18) at α_2:

$$|\varphi'(\alpha_2)| = -\sigma\, \varphi'(0) < -\tau\, \varphi'(0).$$

Due to the continuity of φ', the strong curvature condition (5.18) and thus also the regular curvature condition (5.17) continue to hold for all α in a neighborhood of α_2. □

We now address an algorithm to find a Wolfe-Powell step size. To simplify notation, we introduce the auxiliary function

$$\psi(\alpha) := \varphi(\alpha) - \varphi(0) - \sigma\, \varphi'(0)\, \alpha$$

so that we can write

$$\begin{aligned}
\text{the Armijo condition (5.12)} &\iff \psi(\alpha) \le 0, && \text{(5.12')}\\
\text{the curvature condition (5.17)} &\iff -(\tau - \sigma)\, |\varphi'(0)| \le \psi'(\alpha), && \text{(5.17')}\\
\text{the strong curvature condition (5.18)} &\iff -\underbrace{(\tau - \sigma)}_{>0}\, |\varphi'(0)| \le \psi'(\alpha) \le (\tau + \sigma)\, |\varphi'(0)|. && \text{(5.18')}
\end{aligned}$$

We restrict the discussion to the regular Wolfe-Powell condition, i. e., (5.12) and (5.17). See for instance Geiger, Kanzow, 1999, Kapitel 6.3 for the strong Wolfe-Powell condition.

Lemma 5.17 (Inclusion of Wolfe-Powell step sizes, see Geiger, Kanzow, 1999, Lemma 6.1)**.** *Suppose that $0 \le a < b$ are chosen such the conditions*

$$\psi(a) \le 0 \quad \text{and} \quad \psi'(a) < 0 \tag{5.19a}$$

$$\text{as well as} \quad \psi(b) \ge 0 \tag{5.19b}$$

hold; see Figure 5.4. Then there exists $\alpha^* \in (a, b)$ such that

$$\psi(\alpha^*) < 0 \quad \text{and} \quad \psi'(\alpha^*) = 0$$

holds. In particular, the Wolfe-Powell conditions (5.12') and (5.17') hold in a neighborhood of α^*.

Proof. Let us denote by α^* a global minimizer of

$$\text{Minimize } \psi(\alpha) \text{ on the compact interval } [a, b].$$

The assumptions on a and b imply that α^* belongs to the open interval (a, b). Consequently, α^* is also a local minimizer of the unconstrained problem "Minimize $\psi(\alpha)$ where $\alpha \in \mathbb{R}$", and thus we have $\psi'(\alpha^*) = 0$. From $\psi(a) \le 0$ and $\psi'(a) < 0$ we infer $\psi(\alpha^*) < 0$. Since both (5.12') and (5.17') hold with strict inequalities at α^*, continuity implies that they hold in a neighborhood of α^*. □

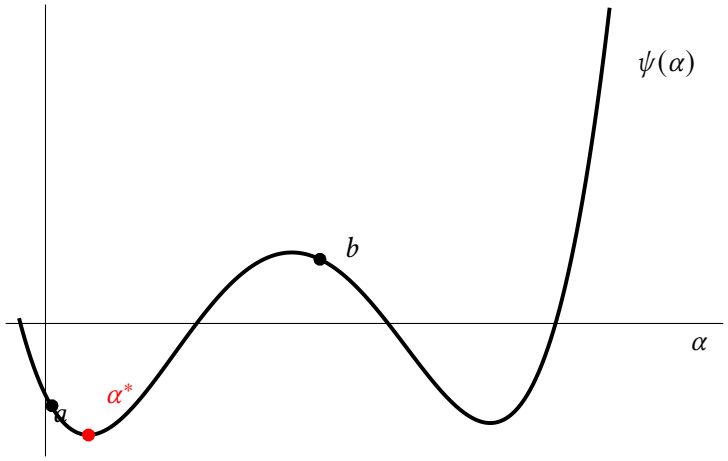


Figure 5.4: Illustration of the condition (5.19) and the statement of Lemma 5.17.

Note: The condition (5.19a) is readily seen to hold at $a = 0$. This motivates the strategy to first find a right boundary b so that (5.19b) holds as well, and then to approximate α^* by nesting intervals.

Algorithm 5.18 (Wolfe-Powell line search)**.**
Input: *initial trial step size* α
Input: *routine to evaluate* φ *and* φ'
Input: *pre-computed function values* $\varphi(0)$ *and* $\varphi'(0)$
Input: *Armijo and curvature parameters* $0 < \sigma < \tau < 1$
Input: *expansion parameter* $\gamma > 1$
Input: *nesting parameters* $\underline{\gamma}, \overline{\gamma} \in (0, 1/2]$
Output: *step size* α *satisfying the Wolfe-Powell conditions* (5.12) *and* (5.17)

1: *Set* $a := 0$ *and* $b := \alpha$
2: *Set* $\ell := 0$
3: ***while*** $\varphi(b) < \varphi(0) + \sigma\,\varphi'(0)\,b$ *and* $\varphi'(b) < \tau\,\varphi'(0)$ ***do*** *// phase 1 repeatedly expands* $[0, b]$ *until* (5.19) *holds*
4: *Set* $b := \gamma\, b$ *// expand the right boundary b*
5: *Set* $\ell := \ell + 1$
6: ***end while*** *// now we have* (5.19)
7: *Set* $\alpha := b$
8: ***while*** *Armijo condition* (5.12) *or curvature condition* (5.17) *is violated at* α ***do*** *// phase 2 repeatedly shrinks* $[a, b]$ *until* (5.12) *and* (5.17) *hold*
9: *Choose* $\alpha \in [a + \underline{\gamma}\,(b - a), b - \overline{\gamma}\,(b - a)]$ *// for instance, choose the midpoint*

10: **if** $\varphi(\alpha) \geq \varphi(0) + \sigma\,\varphi'(0)\,\alpha$ **then** // *Armijo condition is violated at* α
11: Set $b := \alpha$ // *reduce the right boundary* b
12: **else**
13: Set $a := \alpha$ // *increase the left boundary* a
14: **end if**
15: Set $\ell := \ell + 1$
16: **end while**
17: **return** α

Remark 5.19 (on Algorithm 5.18, compare Remark 5.12)**.**

(*i*) *The Armijo parameter is often chosen to be small, e. g.,* $\sigma = 10^{-2}$ *or even* $\sigma = 10^{-4}$*. Depending on the characteristics of the outer method (which determines the search directions), the curvature parameter* $\tau > \sigma$ *should be chosen "small" as well, e. g.,* $\tau = 0.1$*, or otherwise "large", e. g.,* $\tau = 0.9$*.*

(*ii*) *Each iteration of phase 1 "costs" one additional evaluation of* φ *and* φ'*, i. e., one additional evaluation of* f *and* f'*, or rather the directional derivative of* f *in the direction of the current search direction; compare (5.13). Each iteration of phase 2 "costs" one additional evaluation of* φ*.*

(*iii*) *Using Lemma 5.17, it is not difficult to see that Algorithm 5.18 terminates after finitely many steps under the conditions of Lemma 5.16:*

- *The while loop beginning at Line 3 terminates, since for* b *sufficiently large, the Armijo condition (5.12) is violated. For such* b*, we have* $\psi(b) > 0$*, i. e., (5.19b) holds.*
- *At the first iteration of the while loop beginning at Line 8, the conditions (5.19) of Lemma 5.17 are satisfied. Consequently, they continue to hold also in all subsequent iterations.*
- *The length of the intervals* $[a, b]$ *in phase 2 goes to zero if infinitely many iterations of the while loop beginning at Line 8 were performed. However, as shown in Lemma 5.17, there is an open set of points which satisfy both the Armijo condition (5.12) and the curvature condition (5.17) inside any of the intervals* $[a, b]$ *considered in phase 2. Therefore, phase 2 must terminate.*

(*iv*) *The step size accepted by Algorithm 5.18 may be larger or smaller than the initial trial step size provided by the user.*

(*v*) *As was already noted for the Armijo backtracking line search (Algorithm 5.11) in Remark 5.12, in a practical implementation, one often adds further checks and stopping criteria to Algorithm 5.11. For instance, we need to safeguard against* $\varphi'(0) \geq 0$ *(*d *is not a descent direction) and against too many unsuccessful trial steps.*

(*vi*) *An algorithm for the strong Wolfe-Powell line search can be found in Geiger, Kanzow, 1999, Kapitel 6.3.*

The admissibility of step sizes generated by the Wolfe-Powell line search algorithm is shown in the

following result. Clearly, this result also applies to step sizes satisfying the strong Wolfe-Powell conditions.

Lemma 5.20 (Wolfe-Powell line search produces admissible step sizes)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates $x^{(k)}$ and search (descent) directions $d^{(k)} \neq 0$. Suppose moreover that the step sizes $\alpha^{(k)}$ are chosen so that they satisfy the Wolfe-Powell conditions (5.12) and (5.17) (for instance by Algorithm 5.18).*[23] *Assume that $K \subseteq \mathbb{N}_0$ is an infinite index set such that the subsequence $\left(x^{(k)}\right)_{k\in K}$ is bounded. Then the step sizes $\left(\alpha^{(k)}\right)_{k\in K}$ are admissible.*

Proof. As in the proof of Lemma 5.13 we obtain the result

$$-\alpha^{(k)} f'(x^{(k)})\, d^{(k)} \to 0. \tag{$*$}$$

It remains to show

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|} \xrightarrow{k\in K} 0.$$

To this end, let $\varepsilon > 0$.

Just like in the proof of Lemma 5.13, we can argue that the boundedness of $\left(x^{(k)}\right)_{k\in K}$ entails that the continuous function f' is uniformly continuous "near the $\left(x^{(k)}\right)_{k\in K}$". More precisely, there exists $\delta > 0$ such that

$$\left\|f'(x^{(k)} + e) - f'(x^{(k)})\right\|_{M^{-1}} \leq (1 - \tau)\,\varepsilon \quad \text{for all } k \in K,\ \|e\|_M \leq \delta.$$

Because of ($*$), there exists an index $k_0 \in \mathbb{N}$ such that

$$\alpha^{(k)} |f'(x^{(k)})\, d^{(k)}| \leq \varepsilon\,\delta \quad \text{for all } k \geq k_0. \tag{$**$}$$

From now on, let $k \in K$, $k \geq k_0$, be arbitrary. Similarly as in the proof of Lemma 5.13, we consider the following cases:

Case 1: $\alpha^{(k)}\|d^{(k)}\|_M \geq \delta$
Precisely as in the proof of Lemma 5.13, we obtain

$$\begin{aligned}
0 &\leq \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} && \text{since } d^{(k)} \text{ is a descent direction} \\
&= \frac{-\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)}\|d^{(k)}\|_M} \\
&\leq \frac{\varepsilon\,\delta}{\delta} && \text{by } (**) \text{ and the assumption in case 1} \\
&= \varepsilon.
\end{aligned}$$

[23] Notice that, in contrast to condition (5.14) in Lemma 5.13, there is no lower bound on the initial trial step size necessary to be observed.

Case 2: $\alpha^{(k)}\|d^{(k)}\|_M < \delta$
In this we argue with the satisfaction of the curvature conditon (5.17) for $\alpha^{(k)}$:

$$\tau\, f'(x^{(k)})\, d^{(k)} \le f'(x^{(k)} + \alpha^{(k)} d^{(k)})\, d^{(k)}.$$

The addition of $|f'(x^{(k)})\, d^{(k)}| = -f'(x^{(k)})\, d^{(k)}$ on both sides yields

$$\begin{aligned}(1-\tau)\,|f'(x^{(k)})\, d^{(k)}| &\le f'(x^{(k)} + \alpha^{(k)} d^{(k)})\, d^{(k)} - f'(x^{(k)})\, d^{(k)} \\ &\le \left|f'\left(x^{(k)} + \alpha^{(k)} d^{(k)}\right) d^{(k)} - f'(x^{(k)})\, d^{(k)}\right| \\ &\le \left\|f'\left(x^{(k)} + \alpha^{(k)} d^{(k)}\right) - f'(x^{(k)})\right\|_{M^{-1}} \|d^{(k)}\|_M.\end{aligned}$$

Invoking now the uniform continuity, we obtain

$$(1-\tau)\,|f'(x^{(k)})\, d^{(k)}| \le (1-\tau)\,\varepsilon\,\|d^{(k)}\|_M,$$

and hence

$$0 \le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

□

Analogously as with the Armijo backtracking line search (Remark 5.14), one can also show the efficiency of step sizes when f' is Lipschitz continuous on the sublevel set $\mathcal{M}_f(x^{(0)}) := \{x \in \mathbb{R}^n \,|\, f(x) \le f(x^{(0)})\}$.

In concluding, we also remark that Line 9 in phase 2 of Algorithm 5.18 leaves some freedom in the choice of the next trial step size α. The available data $\varphi(a)$, $\varphi'(a)$, $\varphi(b)$ and $\varphi'(b)$ lends itself to a cubic Hermite interpolation, using the model

$$p(\alpha) = a + b\,\alpha + c\,\alpha^2 + d\,\alpha^3.$$

Provided that a unique local minimizuer α^* of p exists, we can calculate it explicitly and subsequently clip it to the interval $[a, b]$:

$$\alpha := \max\{a, \min\{b, \alpha^*\}\}.$$

One needs to pay attention to the fact that not all of the data $\varphi'(a)$ and $\varphi'(b)$ is necessarily available in the current iteration of Algorithm 5.18. In this case one may proceed with a quadratic polynomial as in the modified Armijo backtracking line search method.

Remark 5.21 (Scaling invariance of the Armijo and curvature conditions)**.** *The Armijo and curvature conditions* (5.12), (5.17) *and* (5.18) *are invariant w.r.t. affine scaling in the domain and codomain spaces. Suppose that we consider, besides the objective f, another objective g related via*

$$f(x) \quad \rightsquigarrow \quad g(x) := \gamma\, f(A\,x + b) + \delta,$$

where $A \in \mathbb{R}^{n\times n}$ is non-singular, $b \in \mathbb{R}^n$, $\gamma > 0$ and $\delta \in \mathbb{R}$.

Then the following holds: a step size α that satisfies any of the conditions (5.12), (5.17) *or* (5.18) *for g at x with search direction d, satisfies the same conditions for f at $A\,x + b$ with the search direction $A\,d$. Since the scaling of an optimization problem is often arbitrary, this is a desirable property.*

§ 5.3 Gradient Descent Method

In the remainder of § 5 we consider different concrete realizations of the generic descent method Algorithm 5.2. The methods differ w.r.t. the way the search directions $d^{(k)}$ are generated and w.r.t. the choice of the line search method (Armijo or Wolfe-Powell) to determine the step sizes $\alpha^{(k)}$. As was already mentioned, the methods discussed here obtain the search drection at an iterate $x^{(k)}$ by minimizing a quadratic model of the objective

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^{\mathsf{T}} H^{(k)} d. \tag{5.2}$$

When the model Hessian $H^{(k)}$ is s. p. d., this is equivalent to the solution of the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.4}$$

The **gradient descent method** (also known as **steepest descent method**) for our generic unconstrained linear problem

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

generates its search directions in the same way we already know from § 4.2, when f was a quadratic polynomial. That is, we use

$$M\, d^{(k)} = -\nabla f(x^{(k)}) \quad \text{or} \quad d^{(k)} = -M^{-1} \nabla f(x^{(k)}) = -\nabla_M f(x^{(k)}). \tag{5.20}$$

This corresponds to using a constant model Hessian $H^{(k)} \equiv M$ in the model (5.2):

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^{\mathsf{T}} M\, d.$$

The choice of the inner product M is due to the user. As was already mentioned in Remark 4.7, one refers to the case $M = \mathrm{Id}$ as the classical **gradient descent method** without preconditioning. Otherwise one speaks of a **preconditioned gradient descent method** with **preconditioner** M.

The particular choice of $d^{(k)}$ in the gradient descent method clearly implies the angle condition (5.8) with the maximal possible value, $\eta = 1$. In particular, the search direction $d^{(k)}$ is a descent direction for f at $x^{(k)}$, as long as $f'(x^{(k)}) \neq 0$ holds.

A simple strategy is sufficient to determine admissible step sizes (5.10). One typically employs the Armijo backtracking line search (Algorithm 5.11) or the version with interpolation (Algorithm 5.15).

The efficiency condition (5.15) requires that the initial trial step size satisfy

$$\begin{aligned}
\alpha^{(k,0)} &\geq c\, \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M^2} \\
&= c\, \frac{-\big(\nabla_M f(x^{(k)}), d^{(k)}\big)_M}{\|d^{(k)}\|_M^2} \\
&= c\, \frac{\|d^{(k)}\|_M^2}{\|d^{(k)}\|_M^2} && \text{since } d^{(k)} = -\nabla_M f(x^{(k)}) \\
&= c
\end{aligned}$$

with some constant $c > 0$. This simply suggests to impose a lower bound on the initial trial step sizes in gradient descent methods. We will re-label c as $\underline{\alpha}$ in Algorithm 5.22 below.

In addition to observing this bound, it is useful to construct initial trial step sizes using information from past iterations. Assuming that the descent achievable in the current step is equal (to first order) to the descent in the previous step (when the accepted step size was $\alpha^{(k-1)}$), we obtain the following proposal for an initial trial step size $\alpha^{(k,0)}$ at iteration $k \geq 1$:

$$\begin{aligned} & \alpha^{(k,0)} f'(x^{(k)})\, d^{(k)} = \alpha^{(k-1)} f'(x^{(k-1)})\, d^{(k-1)} \\ \Rightarrow \quad & \alpha^{(k,0)} = \alpha^{(k-1)} \frac{f'(x^{(k-1)})\, d^{(k-1)}}{f'(x^{(k)})\, d^{(k)}}. \end{aligned}$$

Plugging in the descent directions used in the gradient descent method, this becomes

$$\alpha^{(k,0)} = \alpha^{(k-1)} \frac{\|\nabla_M f(x^{(k-1)})\|_M^2}{\|\nabla_M f(x^{(k)})\|_M^2} = \alpha^{(k-1)} \frac{\|d^{(k-1)}\|_M^2}{\|d^{(k)}\|_M^2}.$$

Alternatively, we could use the actual descent achieved in the previous step instead of its linearization, which would result in

$$\alpha^{(k,0)} = \frac{f(x^{(k-1)}) - f(x^{(k)})}{\|\nabla_M f(x^{(k)})\|_M^2} = \frac{f(x^{(k-1)}) - f(x^{(k)})}{\|d^{(k)}\|_M^2}.$$

We state the full gradient descent method in Algorithm 5.22, using the above considerations for the initial trial step size. As was the case for our methods in § 4 addressing the minimization of quadratic polynomials, we refer to the value of the derivative of f at an iterate $x^{(k)}$ as the **residual** $r^{(k)}$.

The global convergence of Algorithm 5.22, in the sense that every accumulation point of the sequence of iterates $x^{(k)}$ is a stationary point, follows directly from the global convergence theorem 5.9.

Algorithm 5.22 (Gradient descent method for (**UP**) w.r.t. the M-inner product and Armijo backtracking line search)**.**

Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* f *and* f' *(or* ∇f*)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Input: *Armijo parameter* $\sigma \in (0,1)$ *// to be passed through to the Armijo backtracking line search*
Input: *backtracking parameter* $\beta \in (0,1)$ *// to be passed through to the Armijo backtracking line search*
Input: *lower bound* $\underline{\alpha} > 0$ *for the initial trial step sizes*
Output: *approximate stationary point of* (**UP**)

1: *Set* $k := 0$
2: *Set* $f^{(0)} := f(x^{(0)})$ *// evaluate the initial objective value*
3: *Set* $r^{(0)} := f'(x^{(0)})^\mathsf{T} = \nabla f(x^{(0)})$ *// evaluate the initial residual*
4: *Set* $d^{(0)} := -M^{-1} r^{(0)}$
5: *Set* $\delta^{(0)} := -(r^{(0)})^\mathsf{T} d^{(0)}$ *//* $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d^{(0)}\|_M^2$
6: ***while*** *stopping criterion not met* ***do***
7: ***if*** $k = 0$ ***then***

8: *Set* $\alpha^{(k,0)} := \underline{\alpha}$ *// no information from previous iteration available*
9: **else**
10: *Set* $\alpha^{(k,0)} := \max\{\underline{\alpha}, \frac{f^{(k)} - f^{(k-1)}}{\delta^{(k)}}\}$
11: **end if**
12: *Determine a step size* $\alpha^{(k)} > 0$ *from an Armijo backtracking line search procedure (Algorithm 5.11), applied to* $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$*, with initial trial step size* $\alpha^{(k,0)}$*, Armijo parameter* σ *and backtracking parameter* β *//* $\varphi(0) = f^{(k)}$ *and* $\varphi'(0) = -\delta^{(k)}$ *are already known*
13: *Set* $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
14: *Set* $f^{(k+1)} := f(x^{(k+1)})$ *// can be returned by the Armijo backtracking line search routine*
15: *Set* $r^{(k+1)} := f'(x^{(k+1)})^\mathsf{T} = \nabla f(x^{(k+1)})$
16: *Set* $d^{(k+1)} := -M^{-1} r^{(k+1)}$
17: *Set* $\delta^{(k+1)} := -(r^{(k+1)})^\mathsf{T} d^{(k+1)}$ *//* $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d^{(k+1)}\|_M^2$
18: *Set* $k := k + 1$
19: **end while**
20: **return** $x^{(k)}$

In Line 12, we could also invoke the modified Armijo backtracking method (Algorithm 5.15), with the backtracking parameter β replaced by the pair of parameters $0 < \underline{\beta} < \overline{\beta} < 1$.

As a stopping criterion, we can choose again any of the conditions from (4.14), i. e., stop on the relative or absolute magnitude of the derivative or gradient

$$\|r^{(k)}\|_{M^{-1}} = \|f'(x^{(k)})\|_{M^{-1}} = \|\nabla_M f(x^{(k)})\|_M = \|d^{(k)}\|_M = (\delta^{(k)})^{1/2}.$$

These quantities are already available in the algorithm. A limited interpretation in the sense of Lemma 4.11 is also possible. In case the sequence $x^{(k)}$ converges to a local minimizer that satisfies the second-order sufficient optimality conditions (Theorem 3.3), then we have: for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|x^{(k)} - x^*\|_M \le \delta \quad \text{and} \quad \|f'(x^{(k)})\|_{M^{-1}} \le \varepsilon_{\text{abs}} \quad \Rightarrow \quad \|x^{(k)} - x^*\|_M \le \underbrace{\left(\frac{1}{\alpha} + \varepsilon\right)}_{\approx 1/\alpha} \varepsilon_{\text{abs}},$$

where $\alpha = \lambda_{\min}(f''(x^*); M)$ is the smallest eigenvalue of the Hessian at the solution w.r.t. M. In other words, when we are sufficiently close to a local minimizer satisfying the second-order sufficient optimality conditions, then the norm of the derivative (or the gradient) is — up to the factor $1/\alpha$ — a useful measure of the distance to the solution.

Other often used stopping criteria are

$$\|x^{(k)} - x^{(k-1)}\|_M \le \varepsilon^x_{\text{abs}} + \varepsilon^x_{\text{rel}} \|x^{(k)} - x^{(0)}\|_M,$$
$$|f(x^{(k)}) - f(x^{(k-1)})| \le \varepsilon^f_{\text{abs}} + \varepsilon^f_{\text{rel}} |f(x^{(k)}) - f(x^{(0)})|.$$

These are triggered by slow progress in the iterates or the objective values, respectively. One typically sets $\varepsilon^f_{\text{rel}} = (\varepsilon^x_{\text{rel}})^2$.

It is remarkable that it is possible to monitor the quantities $\|x^{(k)} - x^{(k-1)}\|_M$ and $\|x^{(k)} - x^{(0)}\|_M$, although the matrix M (or matrix-vector products with M) may not be available. Matrix-vector

products with M^{-1} are sufficient. The following quantities are useful for this purpose and can be recursively updated, compare (4.33):

$$\omega^{(k)} := \|x^{(k)} - x^{(0)}\|_M^2 \tag{5.21a}$$

$$\xi^{(k)} := (x^{(k)} - x^{(0)})^\mathsf{T} M\, d^{(k)} = -(x^{(k)} - x^{(0)})^\mathsf{T} r^{(k)} \tag{5.21b}$$

$$\delta^{(k)} := \|d^{(k)}\|_M^2 \tag{5.21c}$$

The details are left as an exercise.

End of Week 4

§ 5.4 Newton's Method

Newton's method is known as a method to solve a (nonlinear) equation $F(x) = 0$, where $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function. For optimization purposes, we apply it to the first-order necessary optimality conditions, i. e., we have $F(x) = \nabla f(x) = 0$, and thus f is assumed to be of class C^2.

The idea of Newton's method to find a zero (root) of F is as follows. Suppose $x^{(0)}$ is an initial guess. We replace F by its linear Taylor model at $x^{(0)}$ and determine the zero of this model instead. This results in

$$F(x^{(0)}) + F'(x^{(0)})(x - x^{(0)}) = 0 \quad \Leftrightarrow \quad x = x^{(0)} - F'(x^{(0)})^{-1} F(x^{(0)}),$$

provided that the Jacobian $F'(x^{(0)})$ is non-singular. This zero of the linear model is used as the next iterate $x^{(1)}$, etc. This procedure is known as the **(local) Newton's method**.

Algorithm 5.23 (Local Newton's method for $F(x) = 0$).
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* F *and* F'
Output: *approximate zero of* F

1: *Set* $k := 0$
2: ***while*** *stopping criterion not met* ***do***
3: *Determine the* ***Newton direction*** *by solving*

$$F'(x^{(k)})\, d^{(k)} = -F(x^{(k)})$$

4: *Set* $x^{(k+1)} := x^{(k)} + d^{(k)}$
5: *Set* $k := k + 1$
6: ***end while***
7: ***return*** $x^{(k)}$

Auxiliary Results

We recall some auxiliary results, which you may know from *Grundlagen der Optimierung* (Herzog, 2022) or other classes. As usual, we equip $\mathbb{R}^n$ with the M-inner product. Recall from § 2.2 that the

operator norm of a matrix $K \in \mathbb{R}^{n\times n}$ that represents a map $K\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\|K\|_{M\leftarrow M} := \max_{x\neq 0} \frac{\|K\,x\|_M}{\|x\|_M}.$$

Although in finite dimensions all norms are equivalent, the above norm is not always the most appropriate choice: some matrices $A \in \mathbb{R}^{n\times n}$ actually represent maps $A\colon \mathbb{R}^n \to (\mathbb{R}^n)^*$, where $(\mathbb{R}^n)^*$ is the dual space of $\mathbb{R}^n$. The appropriate inner product in the dual space is the M^{-1}-inner product, leading to

$$\|A\|_{M^{-1}\leftarrow M} := \max_{x\neq 0} \frac{\|A\,x\|_{M^{-1}}}{\|x\|_M}.$$

Consequently, we would use

$$\|A^{-1}\|_{M\leftarrow M^{-1}} := \max_{r\neq 0} \frac{\|A^{-1}\,r\|_M}{\|r\|_{M^{-1}}}$$

for the inverse of A. We also need the case $B\colon (\mathbb{R}^n)^* \to \mathbb{R}^n$.

Lemma 5.24 (Banach' lemma)**.**

(*i*) *Suppose that $K \in \mathbb{R}^{n\times n}$ is a matrix $\|K\|_{M\leftarrow M} < 1$. Then $\mathrm{Id} - K$ is non-singular, and we have the following estimate on the norm of its inverse:*

$$\|(\mathrm{Id} - K)^{-1}\|_{M\leftarrow M} \leq \frac{1}{1 - \|K\|_{M\leftarrow M}}.$$

(*ii*) *Suppose that $A, B \in \mathbb{R}^{n\times n}$ are such that $\|\mathrm{Id} - BA\|_{M\leftarrow M} < 1$. Then A and B are both non-singular, and we have*

$$\|B^{-1}\|_{M^{-1}\leftarrow M} \leq \frac{\|A\|_{M^{-1}\leftarrow M}}{1 - \|\mathrm{Id} - BA\|_{M\leftarrow M}} \quad \textit{und} \quad \|A^{-1}\|_{M\leftarrow M^{-1}} \leq \frac{\|B\|_{M\leftarrow M^{-1}}}{1 - \|\mathrm{Id} - BA\|_{M\leftarrow M}}.$$

Note: Statement (*i*) states that "small" perturbations of the identity matrix are still invertible. Statement (*ii*) states that $\mathrm{Id} - BA$ "small", i. e., $B \approx A^{-1}$, entails that A and B are both necessarily invertible.

Proof. Statement (*i*): For $x \in \mathbb{R}^n$, we have

$$\begin{aligned}
\|(\mathrm{Id} - K)\,x\|_M &= \|x - Kx\|_M \\
&\geq \|x\|_M - \|Kx\|_M && \text{by the triangle inequality} \\
&\geq \underbrace{\left(1 - \|K\|_{M\leftarrow M}\right)}_{>0} \|x\|_M && \text{since } \|Kx\|_M \leq \|K\|_{M\leftarrow M}\,\|x\|_M.
\end{aligned}$$

This implies $(\mathrm{Id} - K)\,x \neq 0$ for $x \neq 0$, ie, $\mathrm{Id} - K$ is injective and thus non-singular.

Now let $y \in \mathbb{R}^n$ be arbitrary and $x := (\mathrm{Id} - K)^{-1}y$. Then the above estimate shows

$$\begin{aligned}
&\|y\|_M \geq (1 - \|K\|_{M\leftarrow M})\|(\mathrm{Id} - K)^{-1}y\|_M \\
\Rightarrow\quad &\|(\mathrm{Id} - K)^{-1}\|_{M\leftarrow M} = \max_{y\neq 0} \frac{\|(\mathrm{Id} - K)^{-1}y\|_M}{\|y\|_M} \leq \frac{1}{1 - \|K\|_{M\leftarrow M}}.
\end{aligned}$$

Statement (*ii*): We set $K := \mathrm{Id} - BA$, whence $\|K\|_{M\leftarrow M} < 1$ holds. Due to Statement (*i*), we find that $\mathrm{Id} - K = \mathrm{Id} - (\mathrm{Id} - BA) = BA$ is non-singular, i. e., A and B are both non-singular. Moreover,

$$\begin{aligned}
& (\mathrm{Id} - K)^{-1} = (BA)^{-1} = A^{-1}B^{-1} \\
\Rightarrow \qquad & B^{-1} = A\,(\mathrm{Id} - K)^{-1} \\
\Rightarrow \qquad & \|B^{-1}\|_{M^{-1}\leftarrow M} \le \|A\|_{M^{-1}\leftarrow M}\|(\mathrm{Id} - K)^{-1}\|_{M\leftarrow M} \\
& \qquad\qquad \le \frac{\|A\|_{M^{-1}\leftarrow M}}{1 - \|K\|_{M\leftarrow M}} \quad \text{by Statement } (i) \\
& \qquad\qquad = \frac{\|A\|_{M^{-1}\leftarrow M}}{1 - \|\mathrm{Id} - BA\|_{M\leftarrow M}}.
\end{aligned}$$

The remaining inequality follows similarly. □

Lemma 5.25 (Implications of the invertibility of the Jacobian). *Suppose tat $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^* \in \mathbb{R}^n$ is arbitrary with non-singular Jacobian $F'(x^*)$.*

(*i*) *Then there exists a neighborhood $B_\delta^M(x^*)$ and a constant $c > 0$ such that $F'(x)$ is invertible for all $x \in B_\delta^M(x^*)$. Moreover,*

$$\|F'(x)^{-1}\|_{M\leftarrow M^{-1}} \le c \quad \text{holds for all } x \in B_\delta^M(x^*). \tag{5.22}$$

(*ii*) *Suppose now in addition that $F(x^*) = 0$ holds. Then there exist a neighborhood $B_\delta^M(x^*)$ and a constant $\beta > 0$ such that*

$$\|x - x^*\|_M \le \beta\,\|F(x)\|_{M^{-1}} \quad \text{for all } x \in B_\delta^M(x^*). \tag{5.23}$$

β can be chosen as $2\,\|F'(x^*)^{-1}\|_{M\leftarrow M^1}$.

Note: Statement (*i*) is an instance of the fact from functional analysis that the set of boundedly invertible linear operators between two Banach spaces is open. Statement (*ii*) allows us to estimate the norm of the error $\|x - x^*\|_M$ from the norm of the residual $\|F(x)\|_{M^{-1}}$.

Proof. Statement (*i*): Since F' is continuous at x^*, there exists $\delta > 0$ such that

$$\|F'(x^*) - F'(x)\|_{M^{-1}\leftarrow M} \le \varepsilon := \frac{1}{2\,\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}}$$

holds for all $x \in B_\delta^M(x^*)$. Consequently,

$$\begin{aligned}
\|\mathrm{Id} - F'(x^*)^{-1}F'(x)\|_{M\leftarrow M} &= \|F'(x^*)^{-1}\big(F'(x^*) - F'(x)\big)\|_{M\leftarrow M} \\
&\le \|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}\|F'(x^*) - F'(x)\|_{M^{-1}\leftarrow M} \\
&\le \frac{1}{2} < 1.
\end{aligned}$$

By Statement (*ii*) of Lemma 5.24 [with $A = F'(x)$ and $B = F'(x^*)^{-1}$], we can conclude that $F'(x)$ is non-singular for all $x \in B_\delta^M(x^*)$ with

$$\|F'(x)^{-1}\|_{M\leftarrow M^{-1}} \leq \frac{\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}}{1 - \|\mathrm{Id} - F'(x^*)^{-1}F'(x)\|_{M\leftarrow M}} \leq 2\,\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}} =: c.$$

Statement (*ii*): Since F is differentiable in x^*, there exists — for the same $\varepsilon > 0$ as above — a $\delta > 0$ such that

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} \leq \varepsilon\,\|x - x^*\|_M \quad \text{for all } x \in B_\delta^M(x^*).$$

Therefore, for all $x \in B_\delta^M(x^*)$,

$$\begin{aligned}&\|F(x)\|_{M^{-1}}\\ &\quad\geq \|F'(x^*)(x - x^*)\|_{M^{-1}} - \|F(x) - \overbrace{F(x^*)}^{=0} - F'(x^*)(x - x^*)\|_{M^{-1}} \quad \text{by the triangle inequality.}\end{aligned}$$

In view of $\|x - x^*\|_M = \|F'(x^*)^{-1}F'(x^*)(x - x^*)\|_M \leq \|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}\|F'(x^*)(x - x^*)\|_{M^{-1}}$, we can estimate this by

$$\begin{aligned}&\|F(x)\|_{M^{-1}}\\ &\quad\geq \frac{1}{\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}}\|x - x^*\|_M - \varepsilon\,\|x - x^*\|_M\\ &\quad= \varepsilon\,\|x - x^*\|_M \quad \text{by the definition of } \varepsilon,\end{aligned}$$

and the claim follows with $\beta = \varepsilon^{-1}$. □

Lemma 5.26 (Auxiliary estimate). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and $x^* \in \mathbb{R}^n$. For all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}} < \varepsilon\,\|x - x^*\|_M$$

holds for all $x \in B_\delta^M(x^*)$.[24]

Proof. Take $\varepsilon > 0$. The triangle inequality implies

$$\begin{aligned}&\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}}\\ &\quad\leq \|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} + \|F'(x^*) - F'(x)\|_{M^{-1}\leftarrow M}\|x - x^*\|_M.\end{aligned}$$

Since by assumption, F is differentiable in x^*, there exists $\delta_1 > 0$ uch that

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} < \frac{\varepsilon}{2}\,\|x - x^*\|_M$$

holds for all $x \in B_{\delta_1}^M(x^*)$. On the other hand, F' is continuous in x^*, which implies the existence of $\delta_2 > 0$ such that

$$\|F'(x^*) - F'(x)\|_{M^{-1}} < \frac{\varepsilon}{2}$$

holds for all $x \in B_{\delta_2}^M(x^*)$. The conclusion follows with $\delta := \min\{\delta_1, \delta_2\}$. □

[24] Briefly, we can also denote this result as $\|F(x) - F(x^*) - F'(x)(x - x^*)\|_M = o\left(\|x - x^*\|_M\right)$.

Local Newton's Method for $F(x) = 0$

We are now in a position to prove a convergence theorem for local Newton's method.

Theorem 5.27 (Convergence of local Newton's method). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^* \in \mathbb{R}^n$ is a point where $F(x^*) = 0$ and $F'(x^*)$ is non-singular. Then there exists a neighborhood $B_\delta^M(x^*)$ such that*

(*i*) *x^* is the unique zero of F in $B_\delta^M(x^*)$.*

(*ii*) *For any initial guess $x^{(0)} \in B_\delta^M(x^*)$, the local Newton's method is well-defined, and it generates a sequence $x^{(k)}$ which converges to x^*.*

(*iii*) *$\big(x^{(k)}\big)$ converges to x^* Q-superlinearly w.r.t. the M-norm.*

(*iv*) *If F' is Lipschitz continuous in $B_\delta^M(x^*)$, then this convergence is even Q-quadratic.*

Proof. Statement (*i*): By Statement (*ii*) of Lemma 5.25, there exists $\delta_0 > 0$ such that x^* is the only zero of F in $B_\delta^M(x^*)$.

Statement (*ii*): By Statement (*i*) of Lemma 5.25, there exist $\delta_1 > 0$ and $c > 0$ such that $F'(x)$ is non-singular for all $x \in B_{\delta_1}^M(x^*)$ and

$$\|F'(x)^{-1}\|_{M\leftarrow M^{-1}} \le c := 2\,\|F(x^*)^{-1}\|_{M\leftarrow M^{-1}}. \tag{$*$}$$

By Lemma 5.26, given $\varepsilon = 1/(2c)$, there exists $\delta_2 > 0$ such that

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}} \le \frac{1}{2c}\|x - x^*\|_M$$

holds for all $x \in B_{\delta_2}^M(x^*)$. Now set $\delta := \min\{\delta_0, \delta_1, \delta_2\}$ and choose $x^{(0)} \in B_\delta^M(x^*)$ arbitrarily. Then the next iterate $x^{(1)} := x^{(0)} - F'(x^{(0)})^{-1}F(x^{(0)})$ is well-defined, and we have

$$\begin{aligned}
\|x^{(1)} - x^*\|_M &= \|x^{(0)} - x^* - F'(x^{(0)})^{-1}F(x^{(0)})\|_{M^{-1}} \\
&= \|F'(x^{(0)})^{-1}\big[F'(x^{(0)})(x^{(0)} - x^*) - F(x^{(0)}) + \overbrace{F(x^*)}^{=0}\big]\|_{M^{-1}} \\
&\le \|F'(x^{(0)})^{-1}\|_{M\leftarrow M^{-1}}\|F(x^{(0)}) - F(x^*) - F'(x^{(0)})(x^{(0)} - x^*)\|_{M^{-1}} \\
&\le c\,\frac{1}{2c}\|x^{(0)} - x^*\|_M \\
&= \frac{1}{2}\|x^{(0)} - x^*\|_M,
\end{aligned}$$

and thus $x^{(1)}$ again belongs to $B_\delta^M(x^*)$. By induction, $x^{(k)}$ is well-defined, it belongs to $B_\delta^M(x^*)$, and $x^{(k)} \to x^*$ Q-linearly w.r.t. the M-norm.

Statement (*iii*): We begin by setting up an equation for the error:

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - F'(x^{(k)})^{-1}\big(F(x^{(k)}) - F(x^*)\big) \\ &= F'(x^{(k)})^{-1}\big[F'(x^{(k)})(x^{(k)} - x^*) - \big(F(x^{(k)}) - F(x^*)\big)\big] \\ &= F'(x^{(k)})^{-1}\Big[F'(x^{(k)})(x^{(k)} - x^*) - \int_0^1 F'(x^{(k)} + t\,(x^* - x^{(k)}))(x^{(k)} - x^*)\,\mathrm{d}t\Big] \\ &= F'(x^{(k)})^{-1}\Big[\int_0^1 F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\,\mathrm{d}t\Big](x^{(k)} - x^*). \end{aligned}$$

This gives us the following fundamental estimate:

$$\begin{aligned} &\|x^{(k+1)} - x^*\|_M \\ &\quad \le \big\|F'(x^{(k)})^{-1}\big\|_{M\leftarrow M^{-1}} \underbrace{\int_0^1 \overbrace{\big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|}^{=:D^{(k)}(t)}_{M^{-1}\leftarrow M}\,\mathrm{d}t}_{=:I^{(k)}} \|x^{(k)} - x^*\|_M. \quad (**) \end{aligned}$$

Due to $x^{(k)} \to x^*$, we infer that $x^{(k)} + t\,(x^* - x^{(k)}) \to x^*$ uniformly for $t \in [0,1]$. Moreover, F' is continuous, and thus for any $\varepsilon > 0$, there exists an index $k_0 \in \mathbb{N}$ such that

$$\|D^{(k)}(t)\|_{M^{-1}\leftarrow M} \le \varepsilon \quad \text{for all } k \ge k_0 \text{ and all } t \in [0,1].$$

This implies

$$0 \le I^{(k)} = \int_0^1 \|D^{(k)}(t)\|_{M^{-1}\leftarrow M}\,\mathrm{d}t \le \varepsilon \quad \text{for all } k \ge k_0.$$

This in turn gives $I^{(k)} \to 0$. But now (*) and (**) give us

$$\|x^{(k+1)} - x^*\|_M \le c\,I^{(k)}\,\|x^{(k)} - x^*\|_M \le c\,\varepsilon\,\|x^{(k)} - x^*\|_M$$

for all $k \ge k_0$, which is the Q-superlinear convergence.

Statement (*iv*): Since $x^{(k)}$ and $x^{(k)} + t\,(x^* - x^{(k)})$ belong to $B_\delta^M(x^*)$ for all $t \in [0,1]$, we can estimate the integral in a better way, using the stronger assumptions:

$$I^{(k)} = \int_0^1 \big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|_{M^{-1}\leftarrow M}\,\mathrm{d}t \le \int_0^1 L\,t\,\|x^* - x^{(k)}\|_M\,\mathrm{d}t = \frac{L}{2}\|x^{(k)} - x^*\|_M.$$

From (**) we now obtain

$$\|x^{(k+1)} - x^*\|_M \le c\,\frac{L}{2}\|x^{(k)} - x^*\|_M^2.$$

□

Remark 5.28 (on local Newton's method (Algorithm 5.23))**.**

(*i*) *Local Newton's method (Algorithm 5.23) may break down since $F'(x^{(k)})$ is not guaranteed to be invertible, in case the initial guess $x^{(0)}$ lies outside the unknown neighborhood of local convergence $B_\delta^M(x^*)$.*

(*ii*) *The **simplified Newton's method**, which uses the fixed matrix $F'(x^{(0)})$ (assumed to be invertible) instead of $F'(x^{(k)})$, still converges Q-linearly w.r.t. the M-norm.*

Local Newton's Method in Optimization

Newton's method in optimization can be motivated in one of two ways:

(*i*) The first-order necessary optimality condition for (**UP**) reads

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

see Theorem 3.1. When we employ Newton's mmethod to solve this (generally nonlinear) equation $F(x) = \nabla f(x)$ with Jacobian $F'(x) = f''(x)$, we obtain the iteration

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1} \nabla f(x^{(k)}). \tag{5.24}$$

(*ii*) At the current iterate $x^{(k)}$, we replace (**UP**) by the minimization of the quadratic model

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^{\mathsf{T}} H^{(k)}\, d \tag{5.2}$$

where the model Hessian is the symmetric matrix $H^{(k)} = f''(x^{(k)})$. That is, (5.2) becomes the second-order Taylor polynomial. If $H^{(k)}$ is positive definite, then the unique solution of (5.2) is characterized by the linear system

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)}).$$

When one uses the fixed step size $\alpha^{(k)} = 1$ and sets

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)} = x^{(k)} + d^{(k)},$$

we obtain again the iteration (5.24).

Remark 5.29 (on local Newton's method for (**UP**))**.**

(*i*) *Theorem 5.27 proves the local Q-superlinear (or local Q-quadratic) of local Newton's method towards a stationary point x^* of f, provided that $f''(x^*)$ is non-singular. The point x^* may be a local minimizer, a local maximizer, or a saddle point of f, unless we make an assumption or have knowledge about the definiteness of $f''(x^*)$.*

(*ii*) *If $f''(x^{(k)})$ is s. p. d., then the **Newton direction** $d^{(k)}$ obtained from the **Newton system***

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)}) \tag{5.25}$$

is a descent direction for f at $x^{(k)}$, as long as $f'(x^{(k)}) \neq 0$; compare (5.9):

$$f'(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)})^{\mathsf{T}} f''(x^{(k)})^{-1}\, \nabla f(x^{(k)}) < 0.$$

Due to the fixed step size $\alpha^{(k)} = 1$ (instead of line search), descent from iterate to iterate, i. e., $f(x^{(k+1)}) < f(x^{(k)})$, is not guaranteed when $x^{(k)}$ is still "far" from the local minimizer x^*.

(*iii*) *Local Newton's method is invariant w.r.t. affine scaling. This is in contrast to the steepest descent method.*

(*iv*) *$f''(x)$ is a bilinear form accepting two directions and returning a number. Consequently, when we specify only a single direction, the resulting object becomes a linear form. It is thus appropriate to view the Hessian $f''(x)$ as a map $\mathbb{R}^n \to (\mathbb{R}^n)^*$ and to use the associated operator norm.*

A Globalized Newton's Method in Optimization

We now seek to globalize the local Newton's method. In order to be able to apply the global convergence theorem 5.9, we require the search directions and the step sizes to be admissible. We will realize these requirements via a (generalized) angle condition and an Armijo backtracking line search. In addition, we pay attention not to disturb the local Q-superlinear convergence.

Algorithm 5.30 (Globalized Newton method for (**UP**))**.**

Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* f *and* f' *(or* ∇f*)*
Input: *routine to evaluate* f'' *(or matrix-vector products with* f''*)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Input: *globalization parameters* $\eta \in (0,1)$, $\rho > 0$ *and exponent* $p > 0$
Input: *Armijo parameter* $\sigma \in (0, 1/2)$ *// to be passed through to the Armijo backtracking line search*
Input: *backtracking parameter* $\beta \in (0,1)$ *// to be passed through to the Armijo backtracking line search*
Output: *approximate stationary point of* (**UP**)

1: *Set* $k := 0$
2: *Set* $f^{(0)} := f(x^{(0)})$ *// evaluate the initial objective value*
3: *Set* $r^{(0)} := f'(x^{(0)})^\mathsf{T} = \nabla f(x^{(0)})$ *// evaluate the initial residual*
4: *Set* $d_G^{(0)} := -M^{-1} r^{(0)}$ *// evaluate the negative M-gradient*
5: *Set* $\delta^{(0)} := -(r^{(0)})^\mathsf{T} d_G^{(0)}$ *//* $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$
6: ***while*** *stopping criterion not met* ***do***
7: *Attempt to solve the Newton system*

$$f''(x^{(k)})\, d_N^{(k)} = -r^{(k)} \tag{5.26}$$

8: ***if*** *the Newton system is not solvable or not uniquely solvable* ***then***
9: *Set* $d^{(k)} := d_G^{(k)}$ *// use the steepest descent direction as fallback*
10: ***else*** *// Newton direction* $d_N^{(k)}$ *available*
11: *Evaluate the generalized angle condition for the Newton direction*

$$f'(x^{(k)})\, d_N^{(k)} \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d_N^{(k)}\|_M \tag{5.27}$$

12: ***if*** *true* ***then***
13: *Set* $d^{(k)} := d_N^{(k)}$ *// use the Newton direction*
14: ***else***
15: *Set* $d^{(k)} := d_G^{(k)}$ *// use the steepest descent direction as fallback*
16: ***end if***
17: ***end if***
18: *Determine a step size* $\alpha^{(k)} > 0$ *from an Armijo backtracking line search procedure (Algorithm 5.11), applied to* $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$*, with initial trial step size* $\alpha^{(k,0)} = 1$*, Armijo parameter* σ *and backtracking parameter* β *//* $\varphi(0) = f^{(k)}$ *and* $\varphi'(0) = -\delta^{(k)}$ *in case of* $d^{(k)} = d_G^{(k)}$*, or* $\varphi'(0) = f'(x^{(k)})\, d_N^{(k)}$ *in case of* $d^{(k)} = d_N^{(k)}$*, are already known*
19: *Set* $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
20: *Set* $f^{(k+1)} := f(x^{(k+1)})$ *// can be returned by the Armijo backtracking line search routine*

Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$ // *evaluate the negative M-gradient*
Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d_G^{(k+1)}$ // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
Set $k := k+1$
end while
return $x^{(k)}$

So the basic idea of Algorithm 5.30 is to use the negative M-gradient $d_G^{(k)}$ in case the Newton direction $d_N^{(k)}$ is either not available, or in case it is not a good descent direction. To decide the latter, we verify its angle with the direction of steepest descent. We know that the steepest descent direction $d = d_G^{(k)}$ satisfies the angle condition (5.8), i. e.,

$$f'(x^{(k)})\, d \le -\eta\, \|\nabla_M f(x^{(k)})\|_M\, \|d\|_M = -\eta\, \|d_G^{(k)}\|_M\, \|d\|_M$$

with the maximal possible value, $\eta = 1$. In (5.27), we require qualitatively the same condition for the Newton direction, with some smaller value $\eta \in (0,1)$. Moreover, as the norm of the gradient $\|d_G^{(k)}\|_M$ becomes smaller, the convergence proof will exhibit that we can be even less strict and we can replace η by a term going to zero. To be concrete, we use the term $\rho\, \|d_G^{(k)}\|_M^p$ for this purpose. This explains the condition

$$f'(x^{(k)})\, d \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d\|_M$$

that we employ the check the descent quality of the Newton direction $d_N^{(k)}$ in (5.27). A range of similar conditions achieving the same goal is also conceivable; see for instance Geiger, Kanzow, 1999, Kapitel 9.2 or Ulbrich, Ulbrich, 2012, S.49.

Remark 5.31 (on globalized Newton's method (Algorithm 5.30)).

(*i*) *The parameters ρ and p are often chosen relatively small, e. g.,*

$$\rho = 10^{-6} \quad \textit{and} \quad p = 10^{-1}.$$

(*ii*) *As in our previous algorithms, we may have available the preconditioner only in the form of matrix-vector products with M^{-1}. In order to evaluate (5.27), however, we need to be able to compute $\|d_N^{(k)}\|_M$ as well, which appears to be unavailable.*

There is, however, an elegant way out. If we solve the Newton system (5.24) using the CG method (Algorithm 4.17) with preconditioner M and initial guess 0, we have available by (4.33)–(4.34) the M-norm of the iterates and thus also the M-norm of the solution $d_N^{(k)}$.

Moreover, the CG method can be easily modified to accommodate the situation that the Newton system is not solvable, or not uniquely solvable. This is the case when a direction of non-positive curvature is encountered during the CG iterations, i. e., when the quantity θ in Algorithm 4.17 becomes ≤ 0. We describe these modifications below (Algorithm 5.41) in the context of inexact Newton methods, where we also take advantage of the fact that it may not be necessary to solve (5.24) exactly.

(iii) *The approach to globalization taken in Algorithm 5.30 is to reject the Newton direction if it does not exist or does not offer a sufficiently negative directional derivative, and to replace it by the steepest descent direction. There are other approaches that modify the Newton direction so that it always exists and offers sufficient descent. One can, for instance, add a multiple of the identity matrix (or rather a multiple of the preconditioner) to $f''(x^{(k)})$ when the latter is found not to be "sufficiently positive definite". The modified Newton system then reads*

$$\left[f''(x^{(k)}) + \tau M\right] d^{(k)} = -\nabla f(x^{(k)})$$

with some $\tau > 0$; see for instance Geiger, Kanzow, 1999, S.93 and Nocedal, Wright, 2006, S.51.

We now proceed to show the global convergence of Algorithm 5.30.

Theorem 5.32 (Convergence of globalized Newton's method)**.** *Suppose that f is of class C^2. Suppose that x^* is an accumulation point of $x^{(k)}$ and that $\left(x^{(k)}\right)_{k\in K}$ is a subsequence converging to x^*. Then the search directions $\left(d^{(k)}\right)_{k\in K}$ and step sizes $\left(\alpha^{(k)}\right)_{k\in K}$ are admissible. Consequently, we have $f'(x^*) = 0$.*

Proof. We verify the prerequisites of the global convergence theorem 5.9, which then implies $f'(x^*) = 0$. To this end, we set

$$\begin{aligned} K_N &:= \{k \in K : d^{(k)} = d_N^{(k)}\} && \text{(index set of Newton steps)} \\ K_G &:= K \setminus K_N && \text{(index set of gradient steps).} \end{aligned}$$

Step (1) Wir first show the admissibility of the search directions. That is, we have to show that

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0 \quad \text{implies} \quad f'(x^{(k)}) \xrightarrow{k\in K} 0. \tag{5.7'}$$

For indices $k \in K_G$ we have $d^{(k)} = -M^{-1}\nabla f(x^{(k)})$ and thus

$$-\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} = \frac{\|\nabla_M f(x^{(k)})\|_M^2}{\|\nabla_M f(x^{(k)})\|_M} = \|\nabla_M f(x^{(k)})\|_M.$$

The left-hand side of (5.7') thus implies $\|\nabla_M f(x^{(k)})\|_M \xrightarrow{k\in K_G} 0$, which is equivalent to $f'(x^{(k)}) \xrightarrow{k\in K_G} 0$.

For the complementary indices $k \in K_N$, the generalized angle condition (5.27) reads

$$-\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \geq \min\left\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\right\} \|d_G^{(k)}\|_M \geq 0.$$

The left-hand side of (5.7') thus implies $\|d_G^{(k)}\|_M = \|\nabla_M f(x^{(k)})\|_M \xrightarrow{K_N} 0$, which is the same as $f'(x^{(k)}) \xrightarrow{k\in K_N} 0$.

Step (2) The convergence of $(x^{(k)})_{k\in K}$ and the C^2-property of the objective imply that the subsequence of Hessians $f''(x^{(k)})$ converges as well, and consequently the subsequence $f''(x^{(k)})$ is bounded (in any norm we might impose on the space of n-by-n matrices), so that we have $\|f''(x^{(k)})\|_{M^{-1}\leftarrow M} \le C$ for $k \in K$. For the Newton steps, we recall $f''(x)\,d = -\nabla f(x)$, which we can also write as $-M^{-1}f''(x)\,d = \nabla_M f(x)$. By the definition of matrix norms, see (2.4), we find

$$\|d^{(k)}\|_M \ge \frac{1}{\|f''(x^{(k)})\|_{M^{-1}\leftarrow M}}\|\nabla_M f(x^{(k)})\|_M \ge \frac{1}{C}\|\nabla_M f(x^{(k)})\|_M \quad \text{for } k \in K_N,$$

and clearly

$$\|d^{(k)}\|_M = 1\,\|\nabla_M f(x^{(k)})\|_M \qquad \text{for } k \in K_G,$$

so overall we have

$$\|d^{(k)}\|_M \ge \min\Big\{\frac{1}{C},1\Big\}\,\|\nabla_M f(x^{(k)})\|_M \ge \min\Big\{\frac{1}{C},1\Big\}\,\frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \tag{5.28}$$

for all $k \in K$. In view of the initial Armijo trial step size being $\alpha^{(k,0)} = 1$, we satisfy condition (5.14) of Lemma 5.13 with $\psi(t) = \min\{t, t/C\}$, which in turn implies the admissibility of the step sizes along the subsequence.

□

Next we show that, under appropriate assumptions, Algorithm 5.30 eventually becomes identical to the local Newton's method, which means that

$$d^{(k)} = d_N^{(k)} \quad \text{and} \quad \alpha^{(k)} = 1 \tag{5.29}$$

holds for all k sufficiently large. Consequently, the local convergence theorem 5.27 applies, which yields the fast (at least Q-superlinear) convergence of the *entire sequence* of iterates, as soon as it is sufficiently close to a local minimizer satisfying second-order sufficient optimality conditions.

Theorem 5.33 (Transition to fast local convergence in Algorithm 5.30, see Ulbrich, Ulbrich, 2012, Satz 10.14)**.** *Suppose that f is of class C^2. Suppose that x^* is an accumulation point of $x^{(k)}$ and that $(x^{(k)})_{k\in K}$ is a subsequence converging to x^*. Assume, moreover, that the Hessian $f''(x^*)$ is s. p. d. Then the following holds:*

(*i*) *$f'(x^*) = 0$ holds, i. e., x^* satisfies the second-order sufficient optimality conditions.*

(*ii*) *The entire sequence $x^{(k)}$ converges to x^*.*

(*iii*) *There exists an index $k_0 \in \mathbb{N}_0$ such that (5.29) holds for all $k \ge k_0$. Consequently, $x^{(k)}$ converges to x^* Q-superlinearly w.r.t. the M-norm.*

(*iv*) *If f'' is Lipschitz continuous in a neighorbood of x^*, then the convergence is Q-quadratic.*

Proof. We do not provide the proof but refer the interested reader to Ulbrich, Ulbrich, 2012, Satz 10.14 for the time being. □

§ 5.5 Newton-Like Methods

From the point of convergence analysis, the globalized Newton's method (Algorithm 5.30) is superior to the steepest descent method (Algorithm 5.22) since it offers a Q-superlinear convergence phase. However, Newton's method has a number of drawbacks as well:

(1) The Hessian $f''(x)$ may be expensive to evaluate, and it is needed in addition to the first-order derivative $f'(x)$ of the objective.

(2) The solution of the Newton systems

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)}) \tag{5.25}$$

is often more expensive compared to the evaluation of the gradient direction

$$M\, d^{(k)} = -\nabla f(d^{(k)}).$$

After all, M is constant and can be factorized using the Cholesky decomposition when the number of optimization variables is moderate.

We will address both issues simultaneously. To this end, we consider methods which allow us to

(1) replace the Hessian $f''(x^{(k)})$ by a (s. p. d.) model Hessian $H^{(k)}$ and

(2) solve the linear system

$$H^{(k)}\, d^{(k)} = -\nabla f(x^{(k)}) \tag{5.30}$$

iteratively, and possibly only inexactly.

The latter means that effectively we are solving a linear system

$$H^{(k)}\, d^{(k)} = -\nabla f(x^{(k)}) + \zeta^{(k)} \tag{5.31}$$

with an implicitly defined residual $\zeta^{(k)}$. To this end, we will typically specify a tolerance of the form $\|\zeta^{(k)}\|_{M^{-1}} \leq \varepsilon^{(k)}$.

As a starting point, we consider a generic local Newton-like method with no line search.

Algorithm 5.34 (Generic Newton-like method for (**UP**))**.**
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* f *and* f' *(or* ∇f*)*
Input: *symmetric model Hessian* $H^{(0)} \in \mathbb{R}^{n\times n}$ *(possibly s. p. d.)*
Input: *routine to determine the symmetric model Hessians* $H^{(k)}$ *(possibly s. p. d.)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Output: *approximate stationary point of* (**UP**)
1: *Set* $k := 0$
2: ***while*** *stopping criterion not met* ***do***

Determine a search direction $d^{(k)}$ by (inexactly) solving $H^{(k)}d^{(k)} = -\nabla f(x^{(k)})$
// $H^{(k)}d^{(k)} = -\nabla f(x^{(k)}) + \zeta^{(k)}$ with some residual $\zeta^{(k)}$
Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
Determine the next model Hessian $H^{(k+1)}$
Set $k := k+1$
end while
return $x^{(k)}$

The following questions arise:

(1) What are the requirements for $H^{(k)}$ and $\zeta^{(k)}$ in order to obtain fast ("Newton-like", i. e., Q-superlinear) convergence?

(2) What practical approaches exist to choose the matrices $H^{(k)}$ and to impose a bound for residual norm $\zeta^{(k)}$, with an eye to reducing the numerical effort?

As we did for Newton's method (§ 5.4), we begin by considering an analog of Algorithm 5.34 to find a zero of a C^1 function $F\colon \mathbb{R}^n \to \mathbb{R}^n$. In place of the exact Jacobians $F'(x^{(k)})$, we use model Jacobians $H^{(k)}$, which are supposed to be non-singular but not necessarily symmetric or positive definite.

Algorithm 5.35 (Generic Newton-like method for $F(x) = 0$).
Input: *initial guess $x^{(0)} \in \mathbb{R}^n$*
Input: *routine to evaluate F*
Input: *routine to determine the non-singular model Jacobians $H^{(k)}$*
Input: *s. p. d. matrix M (or matrix-vector products with M^{-1})*
Output: *approximate zero of F*
Set $k := 0$
while *stopping criterion not met* ***do***
Determine a search direction $d^{(k)}$ by (inexactly) solving $H^{(k)}\,d^{(k)} = -F(x^{(k)})$
// $H^{(k)}\,d^{(k)} = -\nabla f(x^{(k)}) + \zeta^{(k)}$ with some residual $\zeta^{(k)}$
Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
Set $k := k+1$
end while
return $x^{(k)}$

In a nutshell, the sequence generated by Algorithm 5.35 is governed by

$$\begin{aligned} H^{(k)}d^{(k)} &= -F(x^{(k)}) + \zeta^{(k)} \\ x^{(k+1)} &= x^{(k)} + d^{(k)}. \end{aligned} \tag{5.32}$$

The following lemma shows that the fast local convergence of any sequence $x^{(k)}$ converging to a zero of F is related to the question how well the elements of that sequence satisfy the true Newton systems $F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)}) = 0$.

Lemma 5.36 (Characterization of fast local convergence). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^{(k)}$ is any sequence in $\mathbb{R}^n$ converging to x^* with non-singular Jacobian $F'(x^*)$. Then the following are equivalent:*

(*i*) *$x^{(k)}$ converges Q-superlinearly w.r.t. the M-norm, and we have $F(x^*) = 0$.*

(*ii*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[25]

$$\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.33a}$$

(*iii*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[26]

$$\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\,\|x^{(k)} - x^*\|_M \quad \text{for all } k \ge k_0. \tag{5.33b}$$

(*iv*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[27]

$$\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.33c}$$

Proof. We begin with some preliminary estimates. Since F is of class C^1 and $F'(x^*)$ is non-singular, there exists a neighborhood $B_\delta^M(x^*)$ and constants $c, C > 0$ such that $\|F'(x)\|_{M^{-1}\leftarrow M}$ and $\|F'(x)^{-1}\|_{M\leftarrow M^{-1}}$ hold for all $x \in B_\delta^M(x^*)$; compare Lemma 5.25. The mean value theorem 2.4 gives us

$$F(x^{(k+1)}) - F(x^*) = F'\big(x^* + \xi^{(k)}\,(x^{(k+1)} - x^*)\big)\,(x^{(k+1)} - x^*)$$

with some $\xi^{(k)} \in (0,1)$. We thus conclude

$$c\,\|x^{(k+1)} - x^*\|_M \le \|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} \le C\,\|x^{(k+1)} - x^*\|_M. \tag{$*$}$$

for sufficiently large $k \in \mathbb{N}_0$.

Another application of the mean value theorem 2.4 yields

$$\begin{aligned} &F(x^{(k+1)}) - F(x^{(k)}) \\ &\quad = F'\big(x^{(k)} + \widehat{\xi}^{(k)}\,(x^{(k+1)} - x^{(k)})\big)(x^{(k+1)} - x^{(k)}) \quad \text{where } \widehat{\xi}^{(k)} \in (0,1) \\ &\quad = F'(x^{(k)})(x^{(k+1)} - x^{(k)}) + \big[F'\big(x^{(k)} + \widehat{\xi}^{(k)}\,(x^{(k+1)} - x^{(k)})\big) - F'(x^{(k)})\big](x^{(k+1)} - x^{(k)}) \end{aligned}$$

and thus

$$\begin{aligned} &\|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\qquad\qquad \le \|F'\big(x^{(k)} + \widehat{\xi}_k\,(x^{(k+1)} - x^{(k)})\big) - F'(x^{(k)})\|_{M^{-1}\leftarrow M}\,\|x^{(k+1)} - x^{(k)}\|_M. \end{aligned}$$

[25] briefly: $\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} = o(\|x^{(k+1)} - x^{(k)}\|_M)$
[26] briefly: $\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} = o(\|x^{(k)} - x^*\|_M)$
[27] briefly: $\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} = o(\|x^{(k+1)} - x^{(k)}\|_M)$

As in the proof of Lemma 5.20 we can now exploit the uniform continuity of F' "near the $(x^{(k)})$". This entails that, for any $\varepsilon > 0$, there exists an index $k_0 \in \mathbb{N}_0$ such that

$$\|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon \|x^{(k+1)} - x^{(k)}\|_M \qquad (**)$$

holds for all $k \ge k_0$.

Statement (i) $\Rightarrow$ Statement (ii) and Statement (iii): The triangle inequality and the Q-superlinear convergence imply that, for sufficiently large k, we have

$$\|x^{(k)} - x^*\|_M \le \|x^{(k)} - x^{(k+1)}\|_M + \|x^{(k+1)} - x^*\|_M \le \|x^{(k)} - x^{(k+1)}\|_M + \frac{1}{2}\|x^{(k)} - x^*\|_M,$$

and thus

$$\|x^{(k)} - x^*\|_M \le 2\|x^{(k+1)} - x^{(k)}\|_M. \qquad (***)$$

On the other hand, the triangle inequality and the Q-superlinear convergence also imply that

$$\|x^{(k+1)} - x^{(k)}\|_M \le \|x^{(k+1)} - x^*\|_M + \|x^* - x^{(k)}\|_M \le 1\|x^{(k)} - x^*\|_M + \|x^* - x^{(k)}\|_M,$$

holds for sufficiently large k, whence

$$\|x^{(k+1)} - x^{(k)}\|_M \le 2\|x^{(k)} - x^*\|_M. \qquad (****)$$

In other words, the quantities $\|x^{(k+1)} - x^{(k)}\|_M$ and $\|x^{(k)} - x^*\|_M$ "control each other" for k sufficiently large.

Let $\varepsilon > 0$ be arbitrary. We can estimate

$$\begin{aligned}
&\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\quad \le \|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|F(x^{(k+1)}) - \underbrace{F(x^*)}_{=0}\|_{M^{-1}} \\
&\quad \le \varepsilon\|x^{(k+1)} - x^{(k)}\|_M + \|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} \quad \text{by } (**)
\end{aligned}$$

for k sufficiently large. We need to address the second term in the previous inequality:

$$\begin{aligned}
\|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} &\le C\|x^{(k+1)} - x^*\|_M \qquad \text{by } (*) \\
&\le C\|x^{(k)} - x^*\|_M \qquad \text{by the Q-superlinear convergence.}
\end{aligned}$$

for k sufficiently large. Plugging tihs estimate into the previous inequality is Statement (ii).

Moreover, as we demonstrated in $(****)$, $\|x^{(k)} - x^*\|_M$ and $\|x^{(k+1)} - x^{(k)}\|_M$ are different by at most a constant factor, we also have proved Statement (iii).

Statement (ii) or Statement (iii) $\Rightarrow$ Statement (i): We estimate

$$\begin{aligned}
&\|F(x^{(k+1)})\|_{M^{-1}} \\
&\quad \le \|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|F(x^{(k)}) + F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\quad \le \varepsilon\|x^{(k+1)} - x^{(k)}\|_M + \|F(x^{(k)}) + F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \quad \text{by } (**).
\end{aligned}$$

By Statement (*ii*) or Statement (*iii*), the second term can be bounded by $\varepsilon \|x^{(k+1)} - x^{(k)}\|_M$ or $\varepsilon \|x^{(k)} - x^*\|_M$, respectively. In any case, we have

$$\|F(x^{(k+1)})\|_{M^{-1}} \leq \varepsilon \|x^{(k+1)} - x^{(k)}\|_M + 2\,\varepsilon \|x^{(k+1)} - x^{(k)}\|_M = 3\,\varepsilon \|x^{(k+1)} - x^{(k)}\|_M$$

for sufficiently large k. In view of $x^{(k)} \to x^*$, we find $F(x^{(k+1)}) \to 0$ and thus $F(x^*) = 0$.

It remains to show the Q-superlinear convergence of $x^{(k)}$. For any $\varepsilon \in (0, c)$, we have

$$\begin{aligned} c \|x^{(k+1)} - x^*\|_M &\leq \|F(x^{(k+1)})\|_{M^{-1}} && \text{by } (*) \\ &\leq 3\,\varepsilon \|x^{(k+1)} - x^{(k)}\|_M \\ &\leq 6\,\varepsilon \|x^{(k+1)} - x^*\|_M && \text{by } (****), \end{aligned}$$

and thus

$$\|x^{(k+1)} - x^*\|_M \leq \frac{6\,\varepsilon}{c} \|x^{(k)} - x^*\|_M$$

for sufficiently large k. This shows the Q-superlinear convergence of $x^{(k)}$ to x^* w.r.t. the M-norm.

Statement (*ii*) $\Leftrightarrow$ Statement (*iv*): The difference of the terms inside the norms on the left hand sides in (5.33a) and (5.33c) is $[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})$. Its norm can be estimated as follows:

$$\begin{aligned} &\|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad \leq \|F'(x^{(k)}) - F'(x^*)\|_{M^{-1} \leftarrow M} \|x^{(k+1)} - x^{(k)}\|_M \\ &\quad \leq \varepsilon \|x^{(k+1)} - x^{(k)}\|_M \quad \text{by the continuity of } F' \end{aligned}$$

for sufficiently large k. The triangle inequality, either in the form

$$\begin{aligned} &\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad \leq \|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad \leq \|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \varepsilon \|x^{(k+1)} - x^{(k)}\|_M \end{aligned}$$

or in the form

$$\begin{aligned} &\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad \leq \|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad \leq \|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \varepsilon \|x^{(k+1)} - x^{(k)}\|_M, \end{aligned}$$

each for sufficiently large k, now shows the equivalence of Statement (*ii*) and Statement (*iv*). □

We now apply this lemma to Algorithm 5.34, where the sequence of iterates $x^{(k)}$ is generated via (5.32). The residual these iterates leave in the true Newton systems can be expressed as

$$\begin{aligned} &F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)}) \\ &\quad = F(x^{(k)}) + F'(x^{(k)})\,d^{(k)} \overbrace{\qquad\qquad}^{=0} \\ &\quad = F(x^{(k)}) + F'(x^{(k)})\,d^{(k)} \underbrace{- F(x^{(k)}) - H^{(k)}\,d^{(k)} + \zeta^{(k)}}_{} \\ &\quad = \left[F'(x^{(k)}) - H^{(k)}\right] d^{(k)} + \zeta^{(k)}. \end{aligned}$$

We thus obtain from Lemma 5.36 the following corollary.

Corollary 5.37 (Characterization of fast local convergence). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^{(k)}$ is a sequence generated by (5.32) that converges to x^* with non-singular Jacobian $F'(x^*)$. Then the following are equivalent:*

(*i*) *$x^{(k)}$ converges Q-superlinearly w.r.t. the M-norm, and we have $F(x^*) = 0$.*

(*ii*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\left\| \left[F'(x^{(k)}) - H^{(k)} \right] d^{(k)} + \zeta^{(k)} \right\|_{M^{-1}} \leq \varepsilon \, \|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \geq k_0. \tag{5.34a}$$

(*iii*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\left\| \left[F'(x^{(k)}) - H^{(k)} \right] d^{(k)} + \zeta^{(k)} \right\|_{M^{-1}} \leq \varepsilon \, \|x^{(k)} - x^*\|_M \quad \text{for all } k \geq k_0. \tag{5.34b}$$

(*iv*) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\left\| \left[F'(x^*) - H^{(k)} \right] d^{(k)} + \zeta^{(k)} \right\|_{M^{-1}} \leq \varepsilon \, \|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \geq k_0. \tag{5.34c}$$

This set of equivalent of conditions that ensure the local Q-superlinear convergence are known as **Dennis-Moré conditions**, introduced in Dennis, Moré, 1974. They exhibit that two requisites are sufficient to ensure fast convergence:

(1) The residual in the linear system, $\|\zeta^{(k)}\|_{M^{-1}}$, goes to zero faster than $\|x^{(k+1)} - x^{(k)}\|_M$.

(2) The difference between the Jacobian $F'(x^{(k)})$ and the model Jacobian $H^{(k)}$, evaluated in the direction of $d^{(k)}$, goes to zero faster than $\|x^{(k+1)} - x^{(k)}\|_M$. **Note:** It is not necessary for $H^{(k)}$ to approximate the Jacobian $F'(x^{(k)})$ in its entirety!

We will discuss in the following two classes of methods that are specializations of Algorithm 5.34. The first class of methods are inexact Newton methods (§ 5.6), which use $H^{(k)} = F'(x^{(k)})$. The second class of methods are quasi-Newton algorithms (**??**), which feature $\zeta^{(k)} = 0$.

§ 5.6 Inexact Newton Methods

Inexact Newton methods use the true Jacobian $H^{(k)} = F'(x^{(k)})$ in the linear systems (5.31), but they solve them only inexactly, leaving a residual $\zeta^{(k)}$:

$$F'(x^{(k)})\, d^{(k)} = -F(x^{(k)}) + \zeta^{(k)} \tag{5.35}$$

We measure the norm of the residual in the linear system (5.35) relative to the norm of the outer residual $F(x^{(k)})$ associated with the current iterate $x^{(k)}$. We require

$$\|\zeta^{(k)}\|_{M^{-1}} = \|F'(x^{(k)})\, d^{(k)} + F(x^{(k)})\|_{M^{-1}} \leq \eta^{(k)} \, \|F(x^{(k)})\|_{M^{-1}} \tag{5.36}$$

with some $\eta^{(k)} \in (0, 1)$. The sequence $\left(\eta^{(k)}\right)$ is known as a **forcing sequence**.

Note that $F(x^{(k)})$ is the residual associated with the zero vector, and hence

$$\frac{\|\text{residual associated with } d^{(k)}\|_{M^{-1}}}{\|\text{residual associated with } 0\|_{M^{-1}}} = \frac{\|\zeta^{(k)}\|_{M^{-1}}}{\|F(x^{(k)})\|_{M^{-1}}} \leq \eta^{(k)}. \tag{5.37}$$

Thus we can interpret the forcing sequence as the relative reduction of the residual required in the linear system $F'(x^{(k)})\, d^{(k)} = -F(x^{(k)})$, compared to a zero initial guess. It is evident that we should demand $\eta^{(k)} < 1$. Otherwise, $d^{(k)} = 0$ would constitute a sufficiently accurate solution.

We refer to Algorithm 5.34 as an **inexact local Newton's method** in case $H^{(k)} = F'(x^{(k)})$. For completeness, we state the algorithm as

Algorithm 5.38 (lokal inexact Newton's method for $F(x) = 0$)**.**
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$
Input: *routine to evaluate* F *and* F'
Input: *spd Matrix* M *(oder Matrix-Vektor-Produkte mit* M^{-1}*)*
Input: *routine to determine the* forcing sequence $\eta^{(k)}$
Output: *approximate zero of* F

1: *Set* $k := 0$
2: ***while*** *stopping criterion not met* ***do***
3: *Determine a search direction* $d^{(k)}$ *by (inexactly) solving* $H^{(k)}\, d^{(k)} = -F(x^{(k)})$ *so that the residual* $\zeta^{(k)} := F'(x^{(k)})\, d^{(k)} + F(x^{(k)})$ *satisfies the condition*

$$\|\zeta^{(k)}\|_{M^{-1}} \leq \eta^{(k)}\, \|F(x^{(k)})\|_{M^{-1}} \tag{5.36}$$

4: *Set* $x^{(k+1)} := x^{(k)} + d^{(k)}$
5: *Set* $k := k + 1$
6: ***end while***
7: ***return*** $x^{(k)}$

Note: With $\eta^{(k)} \equiv 0$, we obtain again the exact local Newton's method.

We can now specify a local convergence theorem for Algorithm 5.38.

Theorem 5.39 (Convergence of Algorithm 5.38)**.** *Suppose that* $F\colon \mathbb{R}^n \to \mathbb{R}^n$ *is a* C^1 *function and that* $x^* \in \mathbb{R}^n$ *is a point where* $F(x^*) = 0$ *and* $F'(x^*)$ *is non-singular. Suppose that* $x^{(k)}$ *is a sequence generated by Algorithm 5.38, where the elements of the forcing sequence satisfy* $\eta^{(k)} \leq \overline{\eta} < 1$ *for all* $k \in \mathbb{N}_0$*. Then there exists a neighborhood* $B_\delta^M(x^*)$ *such that*

(*i*) x^* *is the unique zero of* F *in* $B_\delta^M(x^*)$.

(*ii*) *For any initial guess* $x^{(0)} \in B_\delta^M(x^*)$*, the local inexact Newton's method is well-defined, and it generates a sequence* $x^{(k)}$ *which converges to* x^*.

(*iii*) $\left(x^{(k)}\right)$ *converges to* x^* *Q-linearly w.r.t. the* M*-norm.*

(*iv*) *If, in addition,* $\eta^{(k)} \searrow 0$ *holds, then the convergence is Q-superlinear.*

(*v*) *If* F' *is Lipschitz continuous in* $B_\delta^M(x^*)$*, and if, in addition to* $\eta^{(k)} \searrow 0$*, we even have* $\eta^{(k)} \le C \|F(x^{(k)})\|_{M^{-1}}$ *with some constant* $C > 0$*, then this convergence is even Q-quadratic.*

Proof. We only give a sketch of the proof. Statement (*i*) can be shown as Theorem 5.27. A guide to proving Statement (*ii*) and **??** can be found in Geiger, Kanzow, 1999, Satz 10.3. For Statement (*iv*), we use the characterization of Q-superlinear convergence by Corollary 5.37. We have

$$\begin{aligned}\|(\underbrace{f''(x^{(k)}) - H^{(k)}}_{=0})\, d^{(k)} + \zeta^{(k)}\|_{M^{-1}} &= \|\zeta^{(k)}\|_{M^{-1}} \\ &\le \eta^{(k)} \|\nabla f(x^{(k)})\|_{M^{-1}} \quad \text{by (5.36).}\end{aligned}$$

As in the proof of Lemma 5.36, see (∗), we have $\|\nabla f(x^{(k)})\|_{M^{-1}} \le C \|x^{(k)} - x^*\|_M$ for sufficiently large indices k and thus

$$\|(f''(x^{(k)}) - H^{(k)})\, d^{(k)} + \zeta^{(k)}\|_{M^{-1}} \le \eta^{(k)}\, C\, \|x^{(k)} - x^*\|_M.$$

Since $\eta^{(k)} \searrow 0$, we satisfy (5.34b).

Statement (*v*) follows similarly as in Theorem 5.27. □

A possible rule for the choice of the forcing sequence $\eta^{(k)}$ that guarantees the local Q-superlinear convergence is

$$\eta^{(k)} := \min\{\overline{\eta},\ \|\nabla f(x^{(k)})\|_{M^{-1}}^{\theta}\} \tag{5.38}$$

with some $\overline{\eta} < 1$ and $\theta \in (0, 1]$, for instance $\overline{\eta} = 1/2$ and $\theta = 0.5$.[28]

In the remainder of § 5.6 we consider a practical approach to the inexact solution of the Newton systems, while simultaneously globalizing the inexact local Newton's method (Algorithm 5.38). Since in this class we are discussing globalization for Newton-like methods only in the context of optimization (and not for general root-finding), we switch back to the optimization context now. That is, we have $F(x) = \nabla f(x)$ and $F'(x) = f''(x)$.

We need to take into account the following:

(1) The Newton system $f''(x^{(k)})\, d = -\nabla f(x^{(k)})$ is to be solved *iteratively*[29]. In this way, we can take advantage of the fact that an inexact solution is sufficient and we can stop once the residual norm for the linear system falls below the threshold dictated by the forcing sequence; see (5.36). We refer to the inexact Newton direction as $d_N^{(k)}$.

(2) The inexact Newton direction $d_N^{(k)}$ is required to be, at the very least, a descent direction for the objective f at the current outer iterate $x^{(k)}$.

[28] More precisely, we even obtain the Q-superlinear convergence with rate $1 + \theta$ with this choice.
[29] rather than using a direct solver such as Gaussian elimination

(3) As we did in the globalized exact Newton's method (Algorithm 5.30), we need to verify whether the inexact Newton direction $d_N^{(k)}$ offers sufficient descent. If not, then we fall back to taking a step in the steepest descent direction.

It turns out that we can reach the first and the second goal simultaneously by a clever use of the conjugate gradient method (Algorithm 4.17), applied to the symmetric linear system $A\,d = b$, where

$$A = f''(x^{(k)}) \quad \text{and} \quad b = -\nabla f(x^{(k)}).$$

As a stopping criterion, we employ the relative criterion (4.14a) with $\varepsilon_{\text{rel}} = \eta^{(k)}$, and the zero vector serves as initial guess. In case the CG algorithm finishes "without an incidence", then — due to (5.37) — the solution returned is an inexact solution of the Newton system with sufficiently small residual norm in the sense of (5.36).

Remark 5.40 (inner and outer iterations)**.** *In what follows we will sometimes use the terms inner iterations and outer iterations. The* ***outer iterations*** *of those of the outer optimization method, which is the inexact Newton's method in this subsection. The quantities used in the outer iterations are the iterates $x^{(k)}$, search directions $d^{(k)}$, step sizes $\alpha^{(k)}$, etc.*

On the other hand, every search direction $d^{(k)}$ will now be found in an iterative way, which refer to as ***inner iterations****. In order to help avoid confusion, we will denote the inner iteration index by ℓ. Also, the iterates of the inner solver for the linear system $A\,d = b$ will be termed $d^{(\ell)}$ instead of $x^{(\ell)}$. The search directions in the inner solver will be $p^{(\ell)}$ instead of $d^{(\ell)}$. The residuals in the inner solver will be $\zeta^{(\ell)}$ instead of $r^{(\ell)}$.*

What could be the incidences that might occur in the CG algorithm in the present context? On the one hand, we might reach the maximum number of iterations before reaching the relative tolerance. On the other hand, the symmetric matrix A might not be positive definite. This means that the function

$$\phi(z) := \frac{1}{2} z^{\mathsf{T}} A\, z - b^{\mathsf{T}} z$$

has a least one direction $p \in \mathbb{R}^n$, $p \neq 0$, of non-positive curvature; i. e., $p^{\mathsf{T}} A\, p \leq 0$ holds. A lack of positive definiteness does not mean that a search direction of non-positive curvature will actually be encountered during the inner iterations. On the one hand, the required tolerance may be reached beforehand. But even for exact solutions ($\varepsilon_{\text{rel}} = 0$), not all right hand sides b actually invoke directions of non-positive curvature.

In any case, if a direction $p^{(\ell)}$ with $\theta^{(\ell)} := (p^{(\ell)})^{\mathsf{T}} A\, p^{(\ell)} \leq 0$ is encountered, a reaction is required since otherwise,

- in case $\theta^{(\ell)} = 0$, a division by zero would occur in Line 8 of the CG algorithm (Algorithm 4.17),
- in case $\theta^{(\ell)} < 0$, the CG algorithm could be continued; however, we might lose the property that the iterates $d^{(\ell)}$ are descent directions for f at $x^{(k)}$. This can be confirmed by examples. As long as all search directions $p^{(\ell)}$ are directions of positive curvature ($\theta^{(\ell)} > 0$), the descent property remains intact; see Lemma 5.42.

For the reasons above, it is customary to employ a variant of the CG method known as **truncated conjugate gradient method** (**truncated CG method**) as inner solver in a globalized inexact Newton method. Starting from a zero initial guess, iterate until either the relative stopping criterion (4.14a) is verified, or a search direction of non-positive curvature is encountered. In that case, the most recent iterate $d^{(\ell)}$ is returned as inexact solution.

For completeness, we state below the truncated CG algorithm. Notice that we chose the specific stopping criterion (5.37) instead of a general criterion.

Algorithm 5.41 (Truncated conjugate gradient method for symmetric systems $A\,d = b$ w.r.t. the M-inner product; compare Algorithm 4.17)**.**

Input: *right-hand side* $b \in \mathbb{R}^n$
Input: *symmetric matrix* A *(or matrix-vector products with* A*)*
Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*
Input: *relative residual* $\varepsilon_{\mathrm{rel}}$
Output: *approximate solution of* $A\,d = b$

Set $\ell := 0$
Set $d^{(0)} := 0$ // *zero initial guess*
Set $\zeta^{(0)} := -b$ // *evaluate the initial residual*
Set $p^{(0)} := -M^{-1}\zeta^{(0)}$
Set $\delta^{(0)} := -(\zeta^{(0)})^{\mathsf{T}} p^{(0)}$ // $\delta^{(0)} = \|\zeta^{(0)}\|^2_{M^{-1}}$
while $\delta^{(\ell)} \geq \varepsilon_{\mathrm{rel}}^2 \delta^{(0)}$ ***do*** // *check stopping criterion* (5.37)
    Set $q^{(\ell)} := A\,p^{(\ell)}$
    Set $\theta^{(\ell)} := (q^{(\ell)})^{\mathsf{T}} p^{(\ell)}$
    if $\theta^{(\ell)} > 0$ ***then***
        Set $\alpha^{(\ell)} := \delta^{(\ell)} / \theta^{(\ell)}$
        Set $d^{(\ell+1)} := d^{(\ell)} + \alpha^{(\ell)} p^{(\ell)}$
        Set $\zeta^{(\ell+1)} := \zeta^{(\ell)} + \alpha^{(\ell)} q^{(\ell)}$
        Set $p^{(\ell+1)} := -M^{-1}\zeta^{(\ell+1)}$
        Set $\delta^{(\ell+1)} := -(\zeta^{(\ell+1)})^{\mathsf{T}} p^{(\ell+1)}$ // $\delta^{(\ell+1)} = \|\zeta^{(\ell+1)}\|^2_{M^{-1}}$
        Set $\beta^{(\ell+1)} := \delta^{(\ell+1)} / \delta^{(\ell)}$
        Set $p^{(\ell+1)} := p^{(\ell+1)} + \beta^{(\ell+1)}\, p^{(\ell)}$
        Set $\ell := \ell + 1$
    else
        Abort the ***while*** *loop*
    end if
end while
return $d^{(\ell)}$

It still remains to be proved that the approximate solution $d^{(\ell)}$ that the truncated CG method generates and that is to be used as inexact Newton direction $d_N^{(k)}$, is indeed a descent direction for the objective f at the current outer iterate $x^{(k)}$. This means that we need to show $f'(x^{(k)})\, d^{(\ell)} < 0$ or equivalently, $b^{\mathsf{T}} d^{(\ell)} > 0$.

Lemma 5.42 (The truncated CG method generates descent directions)**.** *Suppose that* $b \neq 0$ *and that*

$d^{(0)}, \dots, d^{(\ell)}$ have been generated by Algorithm 5.41 for some $\ell \geq 1$. Then the following holds.

(*i*) $b^\mathsf{T} M^{-1} \zeta^{(j)} = 0$ *for* $j = 1, \dots, \ell$.

(*ii*) $b^\mathsf{T} p^{(j)} = \|\zeta^{(j)}\|^2_{M^{-1}}$ *for* $j = 0, \dots, \ell$.

(*iii*) $b^\mathsf{T} d^{(\ell)} = \sum_{j=0}^{\ell-1} \alpha^{(j)} \, \|\zeta^{(j)}\|^2_{M^{-1}}$ *ist positive and strictly monotonically increasing in* ℓ.

Proof. Statement (*i*): Since we use the zero vector as initial guess, we have $\zeta^{(0)} = A\,0 - b = -b$ for the initial residual. Therefore,

$$b^\mathsf{T} M^{-1} \zeta^{(j)} = -(\zeta^{(0)})^\mathsf{T} M^{-1} \zeta^{(j)} = 0 \quad \text{for } j \geq 1$$

according to (4.28).

Statement (*ii*): The initial search direction is $p^{(0)} = -M^{-1}\zeta^{(0)}$, and hence we have

$$b^\mathsf{T} p^{(0)} = (\zeta^{(0)})^\mathsf{T} M^{-1} \zeta^{(0)} = \|\zeta^{(0)}\|^2_{M^{-1}}.$$

By induction, we find for $j \geq 0$:

$$\begin{aligned} b^\mathsf{T} p^{(j+1)} &= b^\mathsf{T}\big(-M^{-1}\zeta^{(j+1)} + \beta^{(j+1)}\, p^{(j)}\big) \\ &= 0 + \beta^{(j+1)}\, b^\mathsf{T} p^{(j)} && \text{by Statement (i)} \\ &= \frac{\|\zeta^{(j+1)}\|^2_{M^{-1}}}{\|\zeta^{(j)}\|^2_{M^{-1}}}\, b^\mathsf{T} p^{(j)} && \text{by (4.24')} \\ &= \|\zeta^{(j+1)}\|^2_{M^{-1}} && \text{by the induction hypothesis.} \end{aligned}$$

Statement (*iii*): Since Algorithm 5.41 generated the iterates $d^{(0)}, \dots, d^{(\ell)}$, the numbers $\theta^{(0)}, \dots, \theta^{(\ell-1)}$ are all strictly positive. Consequently, $\alpha^{(j)} = \delta^{(j)}/\theta^{(j)} > 0$ is also positive for $j = 0, \dots, \ell - 1$. We consider the expression

$$b^\mathsf{T} d^{(\ell)} = b^\mathsf{T} \sum_{j=0}^{\ell-1} \alpha^{(j)}\, p^{(j)} = \sum_{j=0}^{\ell-1} \alpha^{(j)}\, \|\zeta^{(j)}\|^2_{M^{-1}}$$

with the last equality due to Statement (*ii*). The residuals $\zeta^{(0)}, \dots, \zeta^{(\ell-1)}$ are all $\neq 0$, otherwise the stopping criterion in Algorithm 5.41 would have been triggered. Therefore, the above expression is strictly increasing w.r.t. ℓ. □

Remark 5.43 (on Algorithm 5.41)**.**

(*i*) *The first search direction is* $p^{(0)} = M^{-1}b$*, which is equal to the steepest descent direction* $-M^{-1}\nabla f(x^{(k)})$ *in the optimization context. When* $p^{(0)}$ *is a direction of positive curvature (if* $\theta^{(0)} > 0$*), then* $d^{(1)}$ *is the same as though we had applied the steepest descent method with Cauchy step size (Algorithm 4.6).*

(*ii*) *By contrast, when $p^{(0)}$ is a direction of non-positive curvature, Algorithm 5.41 stops and returns $d^{(0)} = 0$. This is, of course, not a useful descent direction for the outer, inexact Newton method, as will be detected by a quality test for the inexact Newton direction, and a fallback to a gradient step will be the consequence.*

(*iii*) *The strictly increasing monotonicity of $b^\mathsf{T} d^{(\ell)} = -f'(x^{(k)})\, d^{(\ell)}$ w.r.t. the iteration counter ℓ means that the descent properties of the iterates $d^{(\ell)}$ progressively improve, as long as the search directions $p^{(\ell)}$ remain directions of positive curvature for A. Therefore, it is reasonable to continue performing CG iterations until either the desired tolerance is reached, or a direction of non-positive curvature is encountered. This is the strategy Algorithm 5.41 is following.*

As we already mentioned, the globalization of the inexact Newton method can be done along the same lines as in Algorithm 5.30. This leads to the following algorithm.

Algorithm 5.44 (Globalized inexact Newton method for (**UP**); compare Algorithm 5.30)**.**
Input: *initial guess $x^{(0)} \in \mathbb{R}^n$*
Input: *routine to evaluate f and f' (or ∇f)*
Input: *routine to evaluate f'' (or matrix-vector products with f'')*
Input: *s. p. d. matrix M (or matrix-vector products with M^{-1})*
Input: *routine to determine the* forcing sequence *$\eta^{(k)}$*
Input: *globalization parameters $\eta \in (0,1)$, $\rho > 0$ and exponent $p > 0$*
Input: *Armijo parameter $\sigma \in (0, 1/2)$ // to be passed through to the Armijo backtracking line search*
Input: *backtracking parameter $\beta \in (0,1)$ // to be passed through to the Armijo backtracking line search*
Output: *approximate stationary point of* (**UP**)

1: *Set $k := 0$*
2: *Set $f^{(0)} := f(x^{(0)})$* // *evaluate the initial objective value*
3: *Set $r^{(0)} := f'(x^{(0)})^\mathsf{T} = \nabla f(x^{(0)})$* // *evaluate the initial residual*
4: *Set $d_G^{(0)} := -M^{-1} r^{(0)}$* // *evaluate the negative M-gradient*
5: *Set $\delta^{(0)} := -(r^{(0)})^\mathsf{T} d_G^{(0)}$* // *$\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$*
6: ***while*** *stopping criterion not met* ***do***
7: *Determine the inexact Newton direction $d_N^{(k)}$ using Algorithm 5.41 with $A = f''(x^{(k)})$, $b = -r^{(k)}$, preconditioner M and relative residual $\varepsilon_{\mathrm{rel}} = \eta^{(k)}$*
8: *Evaluate the generalized angle condition for the Newton direction*

$$f'(x^{(k)})\, d_N^{(k)} \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d_N^{(k)}\|_M \tag{5.27}$$

9: ***if*** *true* ***then***
10: *Set $d^{(k)} := d_N^{(k)}$* // *use the inexact Newton direction*
11: ***else***
12: *Set $d^{(k)} := d_G^{(k)}$* // *use the steepest descent direction as fallback*
13: ***end if***
14: *Determine a step size $\alpha^{(k)} > 0$ from an Armijo backtracking line search procedure (Algorithm 5.11), applied to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$, with initial trial step size $\alpha^{(k,0)} = 1$, Armijo parameter σ and backtracking parameter β* // *$\varphi(0) = f^{(k)}$ and $\varphi'(0) = -\delta^{(k)}$ in case of $d^{(k)} = d_G^{(k)}$, or $\varphi'(0) = f'(x^{(k)})\, d_N^{(k)}$ in case of $d^{(k)} = d_N^{(k)}$, are already known*

15: Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
16: Set $f^{(k+1)} := f(x^{(k+1)})$ // *can be returned by the Armijo backtracking line search routine*
17: Set $r^{(k+1)} := f'(x^{(k+1)})^\mathsf{T} = \nabla f(x^{(k+1)})$
18: Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$ // *evaluate the negative M-gradient*
19: Set $\delta^{(k+1)} := -(r^{(k+1)})^\mathsf{T} d_G^{(k+1)}$ // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
20: Set $k := k + 1$
21: ***end while***
22: ***return*** $x^{(k)}$

Remark 5.45 (on Algorithm 5.44)**.**

(*i*) *See Remark 5.31 on choosing the globalization parameters ρ and p.*

(*ii*) *The quantity $\|d_N^{(k)}\|_M$ required to evaluate the generalized angle condition (5.27) can be returned at negligible additional cost by the truncated CG algorithm (Algorithm 5.41), as described in (4.33)–(4.34).*

The global convergence of Algorithm 5.44 can be verified very similarly as in Theorem 5.32. In fact, **Step** (1) in the proof (admissibility of search directions) remains exactly the same since the generalized angle condition and the fallback to steepest descent directions remains the same as in Algorithm 5.30. In **Step** (2) (admissibility of step sizes), we need to take into account the fact that the inexact Newton direction satisfies the Newton system only with a residual. We end up replacing the estimate (5.28) by

$$\|d^{(k)}\|_M \geq \min\left\{\frac{1-\eta^{(k)}}{C}, 1\right\} \|\nabla_M f(x^{(k)})\|_M \geq \min\left\{\frac{1-\eta^{(k)}}{C}, 1\right\} \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \tag{5.39}$$

for all $k \in K$, and we have to modify the function ψ accordingly. (**Quiz 5.6:** Can you fill in the details?)

The transition to fast local convergence can be shown similarly as in Theorem 5.33. We can verify again that

$$d^{(k)} = d_N^{(k)} \quad \text{and} \quad \alpha^{(k)} = 1 \tag{5.29}$$

holds for sufficiently large indices k. Consequently, the convergence mode (Q-linear, Q-superlinear or even Q-quadratic) follows depending on the choice of forcing sequence, using Theorem 5.39; see also Geiger, Kanzow, 1999, Satz 10.8.

The combination of the inexact Newton method as outer algorithm with the truncated CG algorithm as inner solver is often referred to as **truncated Newton CG method**. Since we do not necessarily need to set up the full Hessian matrix $f''(x^{(k)})$, but matrix-vector products with $f''(x^{(k)})$ are sufficient, one also speaks of a **Hessian-free optimization**. Matrix-vector products with $f''(x^{(k)})$ can be realized, e. g., using algorithmic diffentiation techniques (Chapter 4).

End of Week 5

Chapter 2 Theory for Constrained Optimization Problems

Chapter 3 Numerical Techniques for Constrained Optimization Problems

Chapter 4 Differentiation Techniques

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.

Alt, W. (2002). *Nichtlineare Optimierung*. Vieweg Studium: Aufbaukurs Mathematik. Eine Einführung in Theorie, Verfahren und Anwendungen. [An introduction to theory, procedures and applications]. Friedrich Vieweg & Sohn, Braunschweig. DOI: 10.1007/978-3-322-84904-5.

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. (1967). *Calcul Différentiel*. Paris: Hermann.

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.

Dennis Jr., J. E.; J. J. Moré (1974). "A characterization of superlinear convergence and its application to quasi-Newton methods". *Mathematics of Computation* 28, pp. 549–560. DOI: 10.1090/s0025-5718-1974-0343581-1.

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.

Ulbrich, M.; S. Ulbrich (2012). *Nichtlineare Optimierung*. New York: Springer. DOI: 10.1007/978-3-0346-0654-7.