Chapter 4

Unconstrained optimization

An **unconstrained optimization problem** takes the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{4.1}$$

for a **target functional** (also called **objective function**) $f : \mathbb{R}^n \to \mathbb{R}$. In this chapter and throughout most of our discussion on optimization, we will assume that f is sufficiently smooth, that is, at least continuously differentiable.

In most applications of optimization problem, one is usually interested in a **global minimizer** $\mathbf{x}^*$, which satisfies $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all x in $\mathbb{R}^n$ (or at least for all x in the domain of interest). Unless f is particularly nice, optimization algorithms are often not guaranteed to yield global minima but only yield local minima. A point $\mathbf{x}^*$ is called a **local minimizer** if there is a neighborhood $\mathcal{N}$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{N}$. Similarly, $\mathbf{x}^*$ is called a **strict local minimizer** $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{N}$ with $\mathbf{x} \neq \mathbf{x}^*$.

4.1 Fundamentals

Sufficient and necessary conditions for local minimizers can be developed from the Taylor expansion of f. Let us recall Example 2.3: If f is two times continuously differentiable then

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{h} + \frac{1}{2}\mathbf{h}^\mathsf{T} H(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|^3), \tag{4.2}$$

where ∇f is the gradient and $H = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j=1}^m$ is the Hessian *[matrice Hessienne]* of f.

Theorem 4.1 (First-order necessary condition) *If $\mathbf{x}^*$ is a local minimizer and f is continuously differentiable in an open neighborhood of $\mathbf{x}^*$ then $\nabla f(\mathbf{x}^*) = 0$.*

Proof. By local optimality $[f(\mathbf{x}^* + t\mathbf{d}) - f(\mathbf{x}^*)]/t$ is nonnegative for sufficiently small $t > 0$ and for arbitrary $\mathbf{d}$. It follows that

$$\lim_{t \to 0^+} \frac{1}{t}[f(\mathbf{x}^* + t\mathbf{d}) - f(\mathbf{x}^*)] = \nabla f(\mathbf{x}^*)^T \mathbf{d} \geq 0.$$

Choosing $\mathbf{d} = -\nabla f(\mathbf{x}^*)$ implies $\nabla f(\mathbf{x}^*) = 0$. ∎

A $\mathbf{x}^*$ satisfying $\nabla f(\mathbf{x}^*) = 0$ is called **stationary point**. A **saddle point** is a stationary point that is neither a local minimizer nor a local maximizer. More can be said if the Hessian of f is available.

Theorem 4.2 (Second-order necessary condition) *If $\mathbf{x}^*$ is a local minimizer and f is two times continuously differentiable in an open neighborhood of $\mathbf{x}^*$ then $\nabla f(\mathbf{x}^*) = 0$ and the Hessian $H(\mathbf{x}^*)$ is positive semidefinite.*

Proof. Theorem 4.1 already yields $\nabla f(\mathbf{x}^*) = 0$, so it remains to prove the positive semidefiniteness of $H(\mathbf{x}^*)$. For sufficiently small t and arbitrary $\mathbf{d}$, we have

$$\begin{aligned} 0 \leq f(\mathbf{x}^* + t\mathbf{d}) - f(\mathbf{x}^*) &= t\nabla f(\mathbf{x}^*)^T \mathbf{d} + \frac{t^2}{2}\mathbf{d}^\mathsf{T} H(\mathbf{x}^*)\mathbf{d} + O(t^3) \\ &= \frac{t^2}{2}\mathbf{d}^\mathsf{T} H(\mathbf{x}^*)\mathbf{d} + O(t^3), \end{aligned}$$

where we used the Taylor expansion (4.2). Hence, $\mathbf{d}^\mathsf{T} H(\mathbf{x}^*)\mathbf{d} \geq O(t)$ and the result follows by taking the limit $t \to 0$. ∎

Theorem 4.3 (Second-order sufficient condition) *Suppose that f is two times continuously differentiable in an open neighborhood of $\mathbf{x}^*$ and that $\nabla f(\mathbf{x}^*) = 0$. Moreover, suppose that the Hessian $H(\mathbf{x}^*)$ is positive definite. Then $\mathbf{x}^*$ is a strict local minimizer.*

Proof. Since $H(\mathbf{x}^*)$ is positive definite, there is a constant μ such that

$$\mathbf{d}^T H(\mathbf{x}^*)\mathbf{d} \geq \mu \|\mathbf{d}\|_2^2$$

for all $\mathbf{d} \in \mathbb{R}^n$. Using the Taylor expansion and $\nabla f(\mathbf{x}^*) = 0$, we have

$$f(\mathbf{x}^* + \mathbf{d}) - f(\mathbf{x}^*) = \frac{1}{2}\mathbf{d}^\mathsf{T} H(\mathbf{x}^*)\mathbf{d} + O(\|\mathbf{d}\|^3) \geq \frac{\mu}{2}\|\mathbf{d}\|_2^2 + O(\|\mathbf{d}\|^3) \geq \frac{\mu}{4}\|\mathbf{d}\|_2^2 > 0$$

for all $\mathbf{d} \neq 0$ of sufficiently small norm. Hence, $\mathbf{x}^*$ is a strict local minimizer. ∎

4.2 Line search methods

General line search methods for solving the optimization problem (4.1) take the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \tag{4.3}$$

where $\alpha_k > 0$ is called the **step length** and $\mathbf{p}_k$ is called the **search direction**.

As we will see, there are many choices for α and $\mathbf{p}$. A natural requirement is that $\mathbf{p}$ should be chosen such that the slope of f in the direction $\mathbf{p}$ is negative. Because of

$$\lim_{t \to 0^+} \frac{f(\mathbf{x} + t\mathbf{p}) - f(\mathbf{x})}{\|t\mathbf{p}\|_2} = \nabla f(\mathbf{x})^T \mathbf{p},$$

this motivates the following definition.

Definition 4.4 *A vector $\mathbf{p} \neq 0$ is called* **descent direction** *of a continuously differentiable function f at a point $\mathbf{x}$ if $\nabla f(\mathbf{x})^T \mathbf{p} < 0$.*

4.2.1 Method of steepest descent

It makes sense to choose $\mathbf{p}$ such that the slope of f in the direction $\mathbf{p}$ is as small as possible. This leads to the minimization problem

$$\min_{\|\mathbf{p}\|_2 = 1} \nabla f(\mathbf{x})^T \mathbf{p}, \tag{4.4}$$

which can be easily solved.

Lemma 4.5 *If $\nabla f(\mathbf{x}) \neq 0$ then (4.4) has the unique solution*

$$\mathbf{p} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2}.$$

Proof. By the Cauchy-Schwarz inequality we have

$$\nabla f(\mathbf{x})^T \mathbf{p} \geq -\|\nabla f(\mathbf{x})\|_2 \|\mathbf{p}\|_2 = -\|\nabla f(\mathbf{x})\|_2,$$

with equality if and only if $\mathbf{p}$ takes the form (4.4). ∎

Any vector of the form

$$\mathbf{p} = -\alpha \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2}, \quad \alpha > 0,$$

is called **direction of steepest descent**. It remains to choose the step length α_k.

The **Armijo rule** applies to a general line search method (4.3) and proceeds as follows: Let $\beta \in]0,1[$ (typically $\beta = 1/2$) and $c_1 \in]0,1[$ (for example $c_1 = 10^{-4}$) be fixed parameters.

Armijo rule:
Determine the largest number $\alpha_k \in \{1, \beta, \beta^2, \beta^3, \ldots\}$ such that

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - f(\mathbf{x}_k) \leq c_1 \alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{p}_k \tag{4.5}$$

holds.

In words, the condition (4.5) ensures that the reduction in f is proportional to the step length and the directional derivative. The following lemma guarantees that (4.5) can always be satisfied provided that $\mathbf{p}_k$ is a descent direction.

Lemma 4.6 *Let $c_1 \in]0,1[$ and let $f : U \to \mathbb{R}$ be continuously differentiable in an open set $U \subset \mathbb{R}^n$. If $\mathbf{x} \in U$ and if $\mathbf{p}$ is a descent direction of f at $\mathbf{x}$ then there is $\overline{\alpha} > 0$ such that*

$$f(\mathbf{x} + \alpha \mathbf{p}) - f(\mathbf{x}) \leq c_1 \alpha \nabla f(\mathbf{x})^T \mathbf{p} \quad \forall \alpha \in [0, \overline{\alpha}].$$

Proof. The inequality trivially holds for $\alpha = 0$. Now, let $\alpha > 0$. Then

$$\frac{f(\mathbf{x} + \alpha \mathbf{p}) - f(\mathbf{x})}{\alpha} - c_1 \nabla f(\mathbf{x})^T \mathbf{p} \overset{\alpha \to 0^+}{\longrightarrow} \nabla f(\mathbf{x})^T \mathbf{p} - c_1 \nabla f(\mathbf{x})^T \mathbf{p} = (1 - c_1) \nabla f(\mathbf{x})^T \mathbf{p} < 0.$$

Hence, by choosing $\overline{\alpha}$ sufficiently small, we have

$$\frac{f(\mathbf{x} + \alpha \mathbf{p}) - f(\mathbf{x})}{\alpha} - c_1 \nabla f(\mathbf{x})^T \mathbf{p} \leq 0 \quad \forall \alpha \in [0, \overline{\alpha}].$$

This shows the result. □

In summary, we obtain Algorithm 4.7.

Algorithm 4.7 Steepest descent with Armijo line search
Input: *Function f, starting vector $\mathbf{x}_0$ and parameters $\beta > 0, c_1 > 0$. Tolerance* tol.
Output: *Vector $\mathbf{x}_k$ approximating stationary point.*

1: ***for*** $k = 0, 1, 2, \ldots$ ***do***
2: *Set* $\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
3: *Stop if* $\|\mathbf{p}_k\| \leq$ tol.
4: *Determine step length α_k according to the Armijo rule* (4.5).
5: *Set* $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$.
6: ***end for***

4.2.2 Convergence of general line search methods

In the following, we will analyse the convergence of general line search method (4.3). Of course, we cannot choose arbitrary $\alpha_k, \mathbf{p}_k$ and still expect convergence. First of all, we would like to maintain the **Armijo condition** of Lemma 4.6:

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - f(\mathbf{x}_k) \leq c_1 \alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{p}_k \tag{4.6}$$

for some $c_1 \in]0,1[$. This ensures a sufficient decrease in the objective function.

The condition (4.6) is not enough by itself to guarantee reasonable progress of the linear search method. Lemma 4.6 shows that it is always satisfied for sufficiently small α_k, provided that $\mathbf{p}_k$ is a descent direction, but we have to make sure that α_k does not become unacceptably small. This is the purpose of the so called **curvature condition**

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^T \mathbf{p}_k \geq c_2 \nabla f(\mathbf{x}_k)^T \mathbf{p}_k, \tag{4.7}$$

which should hold for some $c_2 \in]c_1, 1[$. Typical values for c_2 are 0.9 (when $\mathbf{p}_k$ is chosen by a Newton or quasi-Newton method) or 0.1 (when $\mathbf{p}_k$ is chosen by the nonlinear conjugate gradient method).

The two conditions (4.6)–(4.7) taken together are called *Wolfe conditions*. The following lemma shows that there always exist step lengths satisfying the Wolfe conditions under reasonable assumptions on f.

Lemma 4.8 *Let f be continuously differentiable and assume that is bounded from below along the ray $\{\mathbf{x}_k + \alpha \mathbf{p}_k \mid \alpha > 0\}$ for some descent direction $\mathbf{p}_k$. Then there exist intervals of step lengths satisfying the Wolfe conditions* (4.6)–(4.7)*, provided that $0 < c_1 < c_2 < 1$.*

Proof. Let us consider the function

$$\psi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k) - f(\mathbf{x}_k) - c_1 \alpha \nabla f(\mathbf{x}_k)^T \mathbf{p}_k.$$

Since $\mathbf{p}_k$ is a descent direction and $c_1 < 1$, it follows that $\psi'(0) = (1-c_1)\nabla f(\mathbf{x}_k)^T \mathbf{p}_k < 0$. Because ψ' is continuous and $f(\mathbf{x}_k + \alpha \mathbf{p}_k)$ is bounded from below there exists a smallest $\alpha^* > 0$ such that $\psi(\alpha^*) = 0$. It follows that $\psi(\alpha) \leq 0$ and hence (4.6) holds for all $\alpha \in]0, \alpha^*]$.

By the mean value theorem, there is $\alpha^{**} \in]0, \alpha^*[$ such that

$$f(\mathbf{x}_k + \alpha^* \mathbf{p}_k) - f(\mathbf{x}_k) = \alpha^* \nabla f(\mathbf{x}_k + \alpha^{**} \mathbf{p}_k)^T \mathbf{p}_k.$$

Combined with $\psi(\alpha^*) = 0$, this implies

$$\nabla f(\mathbf{x}_k + \alpha^{**} \mathbf{p}_k)^T \mathbf{p}_k = c_1 \nabla f(\mathbf{x}_k)^T \mathbf{p}_k > c_2 \nabla f(\mathbf{x}_k)^T \mathbf{p}_k.$$

Therefore, there is α^{**} satisfying the Wolfe conditions (4.6)–(4.7). By the continuous differentiability of f, they also hold for a (sufficiently small) interval around α^{**}. □

One of the great advantages of the Wolfe conditions is that they allow to prove convergence of the line search method (4.3) under fairly general assumptions.

Theorem 4.9 *Consider a line search method* (4.3)*, where $\mathbf{p}_k$ is a descent direction and α_k satisfies the the Wolfe conditions* (4.6)–(4.7) *in each iteration k. Suppose that f is bounded from below in $\mathbb{R}^n$ and continuously differentiable*

in an open set $\mathcal{U} \subset \mathbb{R}^n$ *with* $\{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \subset \mathcal{U}$. *Moreover,* ∇f *is assumed to be Lipschitz continuous on* $\mathcal{U}$. *Then*

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \cdot \|\nabla f(\mathbf{x}_k)\|_2^2 < \infty, \quad \textit{where} \quad \cos\theta_k := \frac{-\nabla f(\mathbf{x}_k)^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|_2 \|\mathbf{p}_k\|_2}. \tag{4.8}$$

Proof. The curvature condition (4.7) implies

$$\big(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big)^T \mathbf{p}_k \geq (c_2 - 1) \nabla f(\mathbf{x}_k)^T \mathbf{p}_k,$$

On the other hand, the Lipschitz condition implies $\|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\|_2 \leq L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2$ and hence

$$\big(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\big)^T \mathbf{p}_k \leq \alpha_k L \|\mathbf{p}_k\|_2^2.$$

By combining both inequalities, we obtain

$$\alpha_k \geq \frac{c_2 - 1}{L} \frac{\nabla f(\mathbf{x}_k)^T \mathbf{p}_k}{\|\mathbf{p}_k\|_2^2}.$$

Using the Armijo condition (4.6) then yields

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + c_1 \alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{p}_k \\ &\leq f(\mathbf{x}_k) - c_1 \frac{1 - c_2}{L} \frac{\big(\nabla f(\mathbf{x}_k)^T \mathbf{p}_k\big)^2}{\|\mathbf{p}_k\|_2^2} \\ &= f(\mathbf{x}_k) - c \cos^2 \theta_k \cdot \|\nabla f(\mathbf{x}_k)\|_2^2, \end{aligned}$$

with $c = c_1 \frac{1-c_2}{L}$. Recursively inserting this relation gives

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_0) - c \sum_{j=0}^{k} \cos^2 \theta_k \cdot \|\nabla f(\mathbf{x}_k)\|_2^2.$$

Since $f(\mathbf{x}_0) - f(\mathbf{x}_{k+1})$ is bounded, the statement of the theorem follows by taking $k \to \infty$. □

Let us discuss the implications of Theorem 4.9. First of all, (4.8) implies

$$\cos^2 \theta_k \cdot \|\nabla f(\mathbf{x}_k)\|_2^2 \overset{k\to\infty}{\longrightarrow} 0. \tag{4.9}$$

(In fact, one can say a little more, e.g., the sequence must converge faster than $1/k$ to zero.) It is quite natural to assume that θ_k is bounded away from 90 degrees, that is, there is $\delta > 0$ such that

$$\cos\theta_k \geq \delta > 0, \quad \forall k.$$

For example, this is clearly the case for steepest descent, with $\delta = 1$. Under this condition, it immediately follows from (4.9) that

$$\|\nabla f(\mathbf{x}_k)\|_2^2 \overset{k\to\infty}{\longrightarrow} 0.$$

In other words, the line search method converges (globally) to a stationary point. Note that we cannot conclude that the method converges to a local minimizer. Making such a statement requires to inject additional information about the Hessian; this will lead to the Newton methods discussed in Section 4.2.4.

4.2.3 Rate of convergence for steepest descent

In the following, we aim at quantifying the (local) convergence speed for the steepest descent method. Let us first perform this analysis for a quadratic objective function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{4.10}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $\mathbf{b} \in \mathbb{R}^n$. This is about the best objective function one can dream up; it is strictly convex and $\mathbf{x}^* = A^{-1}\mathbf{b}$ is the global minimizer. Note that $\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$.

We consider an **exact line search** strategy, that is, α_k is chosen to minimize $f(\mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k))$. This can be easily implemented for (4.10) as

$$\begin{aligned}\frac{\partial}{\partial\alpha} f(\mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)) &= -\nabla f(\mathbf{x}_k)^T A\mathbf{x}_k + \alpha\nabla f(\mathbf{x}_k)^T A\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)^T\mathbf{b}\\ &= -\nabla f(\mathbf{x}_k)^T A\mathbf{x}_k + \alpha\nabla f(\mathbf{x}_k)^T A\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)^T\mathbf{b}\\ &= -\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_k) + \alpha\nabla f(\mathbf{x}_k)^T A\nabla f(\mathbf{x}_k)\end{aligned}$$

is zero for

$$\alpha_k = \frac{\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_k)}{\nabla f(\mathbf{x}_k)^T A\nabla f(\mathbf{x}_k)}. \tag{4.11}$$

Hence, the steepest descent method for (4.10) with exact linear search takes the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\nabla f(\mathbf{x}_k)^T\nabla f(\mathbf{x}_k)}{\nabla f(\mathbf{x}_k)^T A\nabla f(\mathbf{x}_k)}\nabla f(\mathbf{x}_k). \tag{4.12}$$

To quantify the convergence of (4.12), it will be convenient to measure the error in the norm induced by A: $\|\mathbf{y}\|_A^2 := \mathbf{y}^T A\mathbf{y}$.

Theorem 4.10 *The steepest descent method with exact line search applied to* (4.10) *satisfies*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A \le \frac{\kappa(A) - 1}{\kappa(A) + 1} \cdot \|\mathbf{x}_k - \mathbf{x}^*\|_A \tag{4.13}$$

where $\kappa(A) = \|A\|_2\|A\|_2^{-1} = \lambda_{\max}(A)/\lambda_{\min}(A)$ *denotes the condition number of* A.

Proof. Subtracting $\mathbf{x}^*$ on both sides of (4.11) gives

$$\mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \frac{\nabla f(\mathbf{x}_k)^T \nabla f(\mathbf{x}_k)}{\nabla f(\mathbf{x}_k)^T A \nabla f(\mathbf{x}_k)} \nabla f(\mathbf{x}_k)$$

Letting $\mathbf{v} := \nabla f(\mathbf{x}_k) = A(\mathbf{x}_k - \mathbf{x}^*)$, we obtain

$$\rho := \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_A^2}{\|\mathbf{x}_k - \mathbf{x}^*\|_A^2} = \left(1 - \frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_A^2 \|\mathbf{x}_k - \mathbf{x}^*\|_A^2}\right)^2 = \left(1 - \frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_A^2 \|\mathbf{v}\|_{A^{-1}}^2}\right)^2 \quad (4.14)$$

To proceed further, we need the so called Kantorovich inequality,

$$\frac{\|\mathbf{v}\|_2^4}{\|\mathbf{v}\|_A^2 \|\mathbf{v}\|_{A^{-1}}^2} \geq \frac{4\lambda_{\min}(A)\lambda_{\max}(A)}{\big(\lambda_{\min}(A) + \lambda_{\max}(A)\big)^2} = \frac{4\kappa(A)}{(1+\kappa(A))^2}, \quad (4.15)$$

which can be shown by noting that can restrict ourselves to vectors of the form $\mathbf{v} = \alpha \mathbf{v}_{\min} + \beta \mathbf{v}_{\max}$, where $\mathbf{v}_{\min}, \mathbf{v}_{\max}$ are eigenvectors belonging to $\lambda_{\min}(A), \lambda_{\max}(A)$. Combining (4.14) and (4.15) yields

$$\rho \leq \left(\frac{\kappa(A) - 1}{\kappa(A) + 1}\right)^2,$$

which concludes the proof. □

Let us now consider the case of general smooth f. By Taylor expansion of f around a strict local minimizer x^*, we have

$$f(\mathbf{x}_k) = f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}_k - \mathbf{x}^*)^\mathsf{T} A (\mathbf{x}_k - \mathbf{x}^*) + O(\|\mathbf{x}_k - \mathbf{x}^*\|^3),$$

where $A = H(\mathbf{x}^*)$ is the Hessian at $\mathbf{x}^*$. Hence,

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \approx \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}^*\|_A^2.$$

Moreover, one step of the steepest descent method will – in first order – produce nearly the same next iterate if we replace f by the quadratic model

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x}_k - \mathbf{x}^*)^\mathsf{T} A (\mathbf{x}_k - \mathbf{x}^*), \quad (4.16)$$

provided that $\|\mathbf{x}_k - \mathbf{x}^*\|$ is sufficiently small. These considerations allow us to generalize Theorem 4.10.

Theorem 4.11 *Suppose that the steepest descent method with exact line search applied to a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ converges to*

a stationary point x^* with symmetric positive definite Hessian $H(\mathbf{x}^*)$. Then – for sufficiently large k – we have

$$f(\mathbf{x}_{k+1}) \leq \rho^2\big(f(\mathbf{x}_k)\big),$$

for any

$$\rho = \frac{\kappa(H(\mathbf{x}^*)) - 1}{\kappa(H(\mathbf{x}^*)) + 1} + \epsilon < 1$$

with $\epsilon > 0$.

4.2.4 The Newton method

From the discussion in Section 4.2.2, one may be tempted to conclude that choosing $\cos\theta_k$ large is advisable and hence steepest descent will produce fastest convergence. Nothing could be more misleading! In this section, we discuss the (locally) much faster Newton method for minimizing $f(\mathbf{x})$, which amounts to choosing the search direction

$$\mathbf{p}_k = -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k). \tag{4.17}$$

There are many different ways of motivating this choice for $\alpha_k = 1$. On the one hand, this amounts to the standard Newton method for solving the nonlinear equation $\nabla f(\mathbf{x}) = 0$. On the other hand, this minimizes the quadratic model (4.16) exactly. Both motivations indicate that the Newton method converges locally quadratically.

Theorem 4.12 *Consider a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ for which the Hessian is symmetric positive definite at a stationary point $\mathbf{x}^*$ and Lipschitz continuous in a neighborhood of $\mathbf{x}^*$. Then the following statements hold for the iteration $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ with the Newton direction (4.17):*

1. *$\mathbf{x}_k \overset{k\to\infty}{\longrightarrow} \mathbf{x}^*$, provided that $\mathbf{x}_0$ is sufficiently close to x^*;*
2. *the sequence $\{\mathbf{x}_k\}$ converges locally quadratically;*
3. *the sequence $\{\|\nabla f(\mathbf{x}_k)\|\}$ converges locally quadratically to zero.*

Proof. By the definition of the Newton method, we have

$$\begin{aligned}\mathbf{x}_{k+1} - \mathbf{x}^* &= \mathbf{x}_k - \mathbf{x}^* + \mathbf{p}_k = \mathbf{x}_k - \mathbf{x}^* - H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k) \\ &= H(\mathbf{x}_k)^{-1}\left[H(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\right]. \end{aligned} \tag{4.18}$$

From the Taylor expansion and the Lipschitz continuity of H, we obtain a constant $L > 0$ such that

$$\|H(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\|_2 \leq L\|\mathbf{x}_k - \mathbf{x}^*\|_2^2$$

holds for all $\mathbf{x}_k$ sufficiently close to $\mathbf{x}^*$. Combined with (4.18), this gives

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq L\|H(\mathbf{x}_k)^{-1}\|_2\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \leq \underbrace{2L\|H(\mathbf{x}^*)^{-1}\|_2}_{=:\tilde{L}}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2, \qquad (4.19)$$

where used the fact that $\|H(\mathbf{x}_k)^{-1}\|_2 \leq 2\|H(\mathbf{x}^*)^{-1}\|_2$ for $\mathbf{x}_k$ sufficiently close to $\mathbf{x}^*$. The inequality (4.19) shows the local quadratic convergence of $\mathbf{x}_k$.

It remains to prove the local quadratic convergence of $\{\|\nabla f(\mathbf{x}_k)\|\}$. This is shown by the same arguments as above:

$$\begin{aligned}\|\nabla f(\mathbf{x}_{k+1})\|_2 &= \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) - H(\mathbf{x}_k)\mathbf{p}_k\|_2 \\ &\leq L\|\mathbf{p}_k\|^2 \leq L\|H(\mathbf{x}_k)^{-1}\|_2^2\|\nabla f(\mathbf{x}_k)\|_2^2 \\ &\leq 2L\|H(\mathbf{x}^*)^{-1}\|_2^2\|\nabla f(\mathbf{x}_k)\|_2^2\end{aligned}$$

where we used $\nabla f(\mathbf{x}_k) + H(\mathbf{x}_k)\mathbf{p}_k = 0$. □

As shown by Theorem 4.12, choosing the step length $\alpha_k = 1$ yields *local* quadratic convergence. In general, $\alpha_k = 1$ is not a good choice in the beginning of the iteration. In practice, the Newton method should therefore be combined with the Armijo or the Wolfe conditions. The expectation is that, initially, α_k is less than 1. Once the region of local quadratic convergence for the Newton method is reached, the conditions allow for choosing $\alpha_k = 1$. This indicates that local quadratic convergence will also be attained in such a setting, see [UU] for the precise statement.

4.2.5 Quasi-Newton methods

The computation of the Newton direction $\mathbf{p}_k = -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ is often too expensive, due to the need for determining the Hessian and solving a linear system. The general idea of **quasi-Newton methods** is to approximate $H(\mathbf{x}_k)$ by a symmetric positive matrix B_k, leading to a search direction of the form

$$\mathbf{p}_k = -B_k^{-1}\nabla f(\mathbf{x}_k). \qquad (4.20)$$

It is important to quantify the extent to which B_k shall approximate $H(\mathbf{x}_k)$ to obtain good convergence, that is, faster convergence than the steepest descent method. As we will see below, it is sufficient to require that B_k provides an increasingly accurate approximation of $H(\mathbf{x}_k)$ *along the search direction* $\mathbf{p}_k$:

$$\lim_{k\to\infty} \frac{\|(B_k - H(\mathbf{x}_k))\mathbf{p}_k\|_2}{\|\mathbf{p}_k\|_2} = 0. \qquad (4.21)$$

Theorem 4.13 *Consider a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ for which the Hessian is symmetric positive definite at a stationary point $\mathbf{x}^*$ and Lipschitz continuous in a neighborhood of $\mathbf{x}^*$. Suppose that the iteration $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ with the quasi-Newton direction (4.20) converges to $\mathbf{x}^*$. Then $\{\mathbf{x}_k\}$ converges superlinearly if and only if (4.21) holds.*

Proof. The key idea of the proof is to relate the quasi-Newton direction to the Newton direction $\mathbf{p}_k^N := -H(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$. Assuming that (4.21) holds, we have

$$\begin{aligned}\|\mathbf{p}_k - \mathbf{p}_k^N\|_2 &= \big\|H(\mathbf{x}_k)^{-1}\big(H(\mathbf{x}_k)\mathbf{p}_k + \nabla f(\mathbf{x}_k)\big)\big\|_2 \\ &\leq \|H(\mathbf{x}_k)^{-1}\|_2\|(H(\mathbf{x}_k) - B_k)\mathbf{p}_k\|_2 = o(\|\mathbf{p}_k\|_2).\end{aligned}$$

On the other hand, $\|\mathbf{p}_k - \mathbf{p}_k^N\|_2 = o(\|\mathbf{p}_k\|_2)$ immediately implies (4.21). Hence, both conditions are equivalent.

By using the result of Theorem 4.12, we thus obtain

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 = \|\mathbf{x}_k + \mathbf{p}_k - \mathbf{x}^*\|_2 &\leq \|\mathbf{x}_k + \mathbf{p}_k^N - \mathbf{x}^*\|_2 + \|\mathbf{p}_k - \mathbf{p}_k^N\|_2 \\ &\leq O(\|\mathbf{x}_k - \mathbf{x}^*\|^2) + o(\|\mathbf{p}_k\|_2).\end{aligned}$$

This proves superlinear convergence if (4.21) holds. The other direction of the statement is an immediate consequence of the fact that superlinear convergence is only possible if $\|\mathbf{p}_k - \mathbf{p}_k^N\|_2 = o(\|\mathbf{p}_k\|_2)$. □

There is a lot of freedom in choosing B_k. Quasi-Newton methods choose a sequence $B_0, B_1, B_2, \ldots$ satisfying the condition

$$B_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) \tag{4.22}$$

starting from an inital symmetric positive definite matrix B_0 (which preferably is an approximation of $H(\mathbf{x}_0)$). Note that (4.22) mimicks the approximation of a tangent vector by the secant vector.

Even when imposing (4.22), there remains a lot of freedom. One usually restricts the freedom further by requiring the update $B_{k+1} - B_k$ to be a low-rank matrix, which allows for the efficient inversion of B_{k+1} using the inverse of B_k. When requiring the update to be symmetric and of rank 1, the choice of B_{k+1} becomes unique:

$$B_{k+1} = B_k + \frac{(\mathbf{y}_k - B_k\mathbf{s}_k)(\mathbf{y}_k - B_k\mathbf{s}_k)^T}{(\mathbf{y}_k - B_k\mathbf{s}_k)^T\mathbf{s}_k}, \tag{4.23}$$

where

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k\mathbf{p}_k, \quad \mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k).$$

The quasi-Newton method resulting from (4.23) is called **SR1** (symmetric rank-1). Some care needs to be applied when using SR1; the denominator $(\mathbf{y}_k - B_k\mathbf{s}_k)^T\mathbf{s}_k$ may become negative (or even zero!), potentially destroying positive definitness.

By far, the most popular quasi-Newton method is **BFGS** (Broyden-Fletcher-Goldfarb-Shanno):

$$B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} - \frac{(B_k \mathbf{s}_k)(B_k \mathbf{s}_k)^T}{\mathbf{s}_k^T B_k \mathbf{s}_k}. \tag{4.24}$$

It can be easily seen that this update satisfies (4.22). Much can (and should) be said about the properties of BFGS. However, the analysis of BFGS is significantly more complicated than the analysis of the Newton method, due to the evolution of B_k. Under suitable conditions, it can be shown that BFGS satisfies (4.21) and hence converges superlinearly.

4.3 The nonlinear conjugate gradient method

4.3.1 The linear conjugate gradient method

Let us recall the (linear) conjugate gradient method for the objective function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A \mathbf{x} - \mathbf{b}^T \mathbf{x},$$

with a symmetric positive definite matrix A. Recall that the gradient of f at $\mathbf{x}_k$ is given by

$$\mathbf{r}_k = A\mathbf{x}_k - \mathbf{b}_k.$$

In Section 4.2.3 we have already seen and analysed the method of steepest descent for this problem. We also seen that it exhibits a 'zigzag' behavior for ill-conditioned matrices, resulting in slow convergence. This 'zigzag' behavior can be avoided by choosing the search directions $\{\mathbf{p}_0, \mathbf{p}_1, \ldots\}$ orthogonal to each other, in the inner product induced by A:

$$\mathbf{p}_i^T A \mathbf{p}_j = 0 \qquad \forall i \neq j. \tag{4.25}$$

Then we generate a sequence $\{\mathbf{x}_k\}$ by setting

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k,$$

with parameter α_k obtained from exact line search:

$$\alpha_k = -\frac{\mathbf{r}_k^T \mathbf{p}_k}{\mathbf{p}_k^T A \mathbf{p}_k}.$$

Because of (4.25), this is sometimes called **conjugate directions method**.

In the **conjugate gradient method**, the directions are chosen such that

$$\operatorname{span}\{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_k\} = \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k\} \tag{4.26}$$

holds, that is, the search directions span the same space as the gradients. To generate such directions, let us suppose we can do this via a recursion of the form

$$\mathbf{p}_k = -\mathbf{r}_k + \beta_k \mathbf{p}_{k-1}.$$

The condition $\mathbf{p}_{k-1}^T A\mathbf{p}_k = 0$ implies

$$\beta_k = \frac{\mathbf{r}_k^T A\mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T A\mathbf{p}_{k-1}}$$

It then follows that (4.25) and (4.26) hold (which is by no means trivial to show). Moreover, it can be shown that

$$\alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A\mathbf{p}_k}, \qquad \beta_{k+1} = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k},$$

which avoids unnecessary multiplications with A in the computation of these scalars. Putting everything together yields Algorithm 4.14.

Algorithm 4.14 CG method
Input: Symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$. Starting vector $\mathbf{x}_0 \in \mathbb{R}^n$. $k_{\max} \in \mathbb{N}$.
Output: Approximate solution $\mathbf{x}_k$ to $A\mathbf{x} = \mathbf{b}$.

$\mathbf{r}_0 \leftarrow \mathbf{b} - A\mathbf{x}_0$, $\mathbf{p}_0 \leftarrow -\mathbf{r}_0$
for $k = 0, 1, \ldots, k_{\max}$ **do**
  $\alpha_k \leftarrow \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T A\mathbf{p}_k}$
  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \alpha_k \mathbf{p}_k$
  $\mathbf{r}_{k+1} \leftarrow \mathbf{r}_k + \alpha_k A\mathbf{p}_k$
  $\beta_{k+1} \leftarrow \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k}$
  $\mathbf{p}_{k+1} \leftarrow -\mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k$
end for

It is informative to compare the following convergence result with Theorem 4.13.

Theorem 4.15 *Let $\mathbf{x}_k$ denote the approximate solution obtained after applying k steps of CG with starting vector $\mathbf{x}_0$. Then*

$$\frac{\|\mathbf{x} - \mathbf{x}_k\|_A}{\|\mathbf{x} - \mathbf{x}_0\|_A} \leq 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k.$$

4.3.2 The Fletcher-Reeves method

Algorithm 4.14 can be applied to a general nonlinear optimization problem by simply replacing $\mathbf{r}_k$ with a gradient. Only one additional change is necessary, as it is in general not possible to obtain the step length α_k by exact line search. This can be replaced by, e.g., the Armijo rule. As a result, we obtain Algorithm 4.16.

Algorithm 4.16 Fletcher-Reeves method
Input: Objective function f. Starting vector $\mathbf{x}_0 \in \mathbb{R}^n$. $k_{\max} \in \mathbb{N}$.
Output: Approximate minimizer $\mathbf{x}_k$ of f.

Evaluate $\nabla_0 = \nabla f(\mathbf{x}_0)$ and set $\mathbf{p}_0 \leftarrow -\nabla_0$
for $k = 0, 1, \ldots, k_{\max}$ **do**
  Compute α_k by line search and set $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \alpha_k \mathbf{p}_k$
  Evaluate $\nabla_{k+1} = \nabla f(\mathbf{x}_{k+1})$
  $\beta_{k+1}^{\mathrm{FR}} \leftarrow \frac{\nabla_{k+1}^T \nabla_{k+1}}{\nabla_k^T \nabla_k}$
  $\mathbf{p}_{k+1} \leftarrow -\nabla_{k+1} + \beta_{k+1}^{\mathrm{FR}} \mathbf{p}_k$
end for

It would be comforting to know that the directions produced by Algorithm 4.16 are indeed descent directions. To check this, we compute

$$\nabla f(\mathbf{x}_k)^T \mathbf{p}_k = -\|\nabla f(\mathbf{x}_k)\|^2 + \beta_k^{\mathrm{FR}} \nabla f(\mathbf{x}_k)^T \mathbf{p}_{k-1}. \tag{4.27}$$

If the line search is exact, we have $\nabla f(\mathbf{x}_k)^T \mathbf{p}_{k-1} = 0$ and hence (4.27) is always negative. For inexact line search this is not clear at all. Fortunately, it will turn out to be the case if we impose the **strong Wolfe conditions**:

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - f(\mathbf{x}_k) \leq c_1 \alpha_k \nabla f(\mathbf{x}_k)^T \mathbf{p}_k, \tag{4.28}$$

$$|\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)^T \mathbf{p}_k| \leq -c_2 \nabla f(\mathbf{x}_k)^T \mathbf{p}_k.. \tag{4.29}$$

While (4.6) and (4.28) are identical, the absolute value in (4.29) is not present in (4.7). Moreover, we impose $0 < c_1 < c_2 < \frac{1}{2}$ (instead of $0 < c_1 < c_2 < 1$).

Theorem 4.17 *Suppose that Algorithm 4.16 makes use of a step length α_k satisfying* (4.29) *with $c_2 < 1/2$. Then the method generates descent directions that satisfy*

$$-\frac{1}{1-c_2} \leq \frac{\nabla f(\mathbf{x}_k)^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|_2^2} \leq \frac{2c_2 - 1}{1 - c_2}. \tag{4.30}$$

Proof. By elementary considerations,

$$-1 < \frac{2c_2 - 1}{1 - c_2} < 0, \tag{4.31}$$

and we therefore obtain the descent property once (4.30) is established.

The proof of (4.30) is by induction in k. The case $k = 0$ follows from (4.31). Assume now that (4.30) holds for some k. Then, by Algorithm 4.16,

$$\frac{\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_{k+1}}{\|\nabla f(\mathbf{x}_{k+1})\|_2^2} = -1 + \beta_{k+1}^{\mathrm{FR}} \frac{\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_{k+1})\|_2^2} = -1 + \beta_{k+1}^{\mathrm{FR}} \frac{\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|_2^2}.$$

Combining this with the curvature condition (4.29),

$$|\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_k| \leq -c_2 \nabla f(\mathbf{x}_k)^T \mathbf{p}_k,$$

we obtain

$$-1 + c_2 \frac{\nabla f(\mathbf{x}_k)^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|_2^2} \leq \frac{\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_{k+1}}{\|\nabla f(\mathbf{x}_{k+1})\|_2^2} \leq -1 - c_2 \frac{\nabla f(\mathbf{x}_k)^T \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|_2^2}.$$

Using the induction hypothesis yields

$$-1 - \frac{c_2}{1-c_2} \leq \frac{\nabla f(\mathbf{x}_{k+1})^T \mathbf{p}_{k+1}}{\|\nabla f(\mathbf{x}_{k+1})\|_2^2} \leq -1 + \frac{c_2}{1-c_2},$$

which completes the proof. □

Note that Theorem 4.17 does not make use of the Armijo condition (4.28). This condition is still needed, to ensure global convergence. However, in contrast to 4.15, the global convergence statements for nonlinear CG methods are much weaker; see Chapter 5 in [NW].

4.3.3 The Polak-Ribière method

If it happens that $\mathbf{p}_k$ is a very poor search direction, that is, it is nearly orthogonal to $\mathbf{p}_k$ then Algorithm 4.16 makes very little progress in one step and hence $\mathbf{x}_{k+1} \approx \mathbf{x}_k$, $\nabla_{k+1} \approx \nabla_k$. So, we have

$$\beta_{k+1}^{\mathrm{FR}} = \frac{\nabla_{k+1}^T \nabla_{k+1}}{\nabla_k^T \nabla_k} \approx 1.$$

Moreover, it can be shown that, in such a situation, $\|\nabla_k\|$ (and therefore also $\|\nabla_{k+1}\|$) needs to be tiny for (4.29) to be satisfied. Consequently,

$$\mathbf{p}_{k+1} \approx \mathbf{p}_k$$

and Algorithm 4.16 will also make very little progress in the next step. In other words, it gets stuck.

The **Polak-Ribière method** aims to avoid the described situation by replacing $\beta_{k+1}^{\mathrm{FR}}$ with

$$\beta_{k+1}^{\mathrm{PR}} = \frac{\nabla_{k+1}^T (\nabla_{k+1} - \nabla_k)}{\nabla_k^T \nabla_k}.$$

The Fletcher-Reeves and Polak-Ribière methods with *exact* line search are identical for strongly convex functions. In all other situations, they can differ, sometimes to a large extent. Often, the Polak-Ribière method yields more robust and faster convergence.

There are several other strategies for choosing β_{k+1}, see [NW].