Chapter 8

Solving Unconstrained Subproblems

At each iteration of the Augmented Lagrangian method, we need to minimize the function

$$L_{\rho_k}(x, \bar{\lambda}^k, \bar{\mu}^k)$$

with respect to x on a generally simple set that we call Ω. In this chapter, we consider the less complicated case, in which $\Omega = \mathbb{R}^n$. This means that, at each outer iteration, we need to solve an unconstrained minimization problem. For simplicity, we denote $\rho = \rho_k$, $\lambda = \bar{\lambda}^k$, $\mu = \bar{\mu}^k$, and

$$F(x) = L_\rho(x, \lambda, \mu) \tag{8.1}$$

throughout this chapter.

In principle, we will assume that F has continuous first derivatives for all $x \in \mathbb{R}^n$ without mentioning second derivatives at all. This omission is convenient at the beginning since the Augmented Lagrangian function has second derivative discontinuities when the original optimization problem has inequality constraints $g(x) \leq 0$, independently of the smoothness of $g(x)$.

8.1 ▪ General algorithm

We will define a general algorithm for unconstrained minimization based on line searches. Many effective algorithms for unconstrained minimization have the general form of the algorithm described here. For simplicity (with some abuse of notation) we denote by $\{x^k\}$ the sequence of iterates generated by this and other subproblems' solvers. They must not be confused with the iterates $\{x^k\}$ of the main Augmented Lagrangian algorithm.

Algorithm 8.1.
Let $\theta \in (0,1)$, $\alpha \in (0,1/2)$, $\bar{M} \geq 2$, and $\beta > 0$ be algorithmic parameters. Let $x^0 \in \mathbb{R}^n$ be the initial approximation. Given $x^k \in \mathbb{R}^n$, the steps for computing x^{k+1} are the following:

Step 1. If $\|\nabla F(x^k)\| = 0$, finish the execution of the algorithm.

Step 2. Compute $d^k \in \mathbb{R}^n$ such that

$$\nabla F(x^k)^T d^k \leq -\theta \|d^k\|_2 \|\nabla F(x^k)\|_2 \text{ and } \|d^k\| \geq \beta \|\nabla F(x^k)\|. \tag{8.2}$$

Step 3. Compute $t_k > 0$ and $x^{k+1} \in \mathbb{R}^n$ such that

$$F(x^{k+1}) \leq F(x^k + t_k d^k),$$

$$F(x^k + t_k d^k) \leq F(x^k) + \alpha t_k \nabla F(x^k)^T d^k, \tag{8.3}$$

and

$$\left[t_k \geq 1\right] \text{ or } \left[F(x^k + \bar{t}_k d^k) > F(x^k) + \alpha \bar{t}_k \nabla F(x^k)^T d^k \text{ for some } \bar{t}_k \in [t_k, \bar{M} t_k]\right].$$

Let us explain the reasons that support each step of this algorithm:

1. At Step 1, we establish that, if the gradient at x^k is null, it is not worthwhile to continue the execution of the algorithm. Thus, we accept points where the gradient vanishes and we do not intend to go further in this case. This is not because we are happy with stationary points, but because we do not know how to proceed from that kind of point without using potentially expensive second-order information.

2. If the gradient does not vanish at x^k, we seek a *search direction* d^k for which two conditions are required. The first is that d^k should be a *first-order descent direction*. This means that the directional derivative $\nabla F(x^k)^T d^k$ should be negative. More precisely, the angle between the direction and $-\nabla F(x^k)$ should be smaller than or equal to a fixed angle smaller than $\pi/2$, whose cosine is defined by the algorithmic parameter θ. If $\theta = 1$, one forces the direction to be a multiple of the negative gradient. In general, we are far less exacting, and $\theta = 10^{-6}$ is a traditionally recommended tolerance. The second condition is that the size of d^k should be at least a fixed multiple of $||\nabla F(x^k)||$. The constant of proportionality is called β. The reason for this requirement is that we want to accept small directions only if the gradient is small.

3. At Step 3, we require that the final point of our line search, $x^k + t_k d^k$, satisfy the *Armijo condition* (8.3). If we define $\varphi(t) = F(x^k + t d^k)$, the Armijo condition is equivalent to
$$\varphi(t_k) \leq \varphi(0) + \alpha t_k \varphi'(0).$$
In other words, with this condition, we require that $\varphi(t_k)$ stay below the line that passes through $(0, \varphi(0))$ whose slope is $\alpha\varphi'(0)$. In this sense, $F(x^k + t_k d^k)$ should be sufficiently smaller than $F(x^k)$. For this reason, (8.3) is frequently known as a *sufficient descent* condition.

   However, satisfying (8.3) is not enough. We need to guarantee that we are not taking artificially small steps. (A step should be small only if it cannot be much larger.) Consequently, we impose that either $t_k \geq 1$ (a constant different from 1 having the same effect) or a frustrated step $\bar{t}_k$ exists, not much bigger than t_k, for which the Armijo condition did not hold.

4. Finally, we leave the door open to take a point x^{k+1} even better than $x^k + t_k d^k$. For this reason, we impose for x^{k+1} the only requirement that its functional value should not be greater than $F(x^k + t_k d^k)$. Obviously, it is admissible to choose $x^{k+1} = x^k + t_k d^k$.

Step 3 of Algorithm 8.1 may be implemented in many different ways. The most elementary strategy consists of choosing t_k as the first element of the sequence $\{1, 1/2, 1/4, \ldots\}$ satisfying the Armijo condition. It is easy to see that, in this way, we obtain a step as required in the algorithm with $\bar{M} = 2$. The direction d^k also admits many different definitions. The most obvious one is to choose $d^k = -\nabla F(x^k)$ (which satisfies (8.2) for any choice of $\theta \in (0,1)$ and $0 < \beta \leq 1$). With these choices, we obtain a version of the *steepest descent* method, one of the most popular procedures for unconstrained minimization.

Algorithm 8.1 requires four algorithmic parameters: θ, α, $\bar{M}$, and β. The first three are dimensionless, that is, their values do not depend on the unities in which the problem magnitudes are measured. For example, θ is the cosine of an angle and α is a pure fraction whose traditional value is 10^{-4}. It makes sense to specify recommended values for dimensionless parameters, since the effect of them should not be affected by the scaling of the problem.

The value of $\bar{M}$ is related to the strategy used in the line search to backtrack, when sufficient descent is not verified for some trial step $\bar{t}$. When this happens, we wish to choose a new trial t in the interval $(0, \bar{t})$. However, t should not be excessively close to 0, because, in that case, the evaluation of F at $x^k + td^k$ would add too little information to the knowledge of F. Therefore, line-search methods usually employ safeguards that impose $t \geq (1/\bar{M})\bar{t}$ with $1/\bar{M} \in (0,1)$. Consequently, the third condition of Step 3 is satisfied.

Usual algorithms for computing t_k at Step 3 of Algorithm 8.1 obey the following steps:

Step LS1. Start setting $t_{\rm trial} \leftarrow 1$.

Step LS2. If $t_{\rm trial}$ satisfies the Armijo condition, find $t_{\rm new} \in (t_{\rm trial}, \bar{M} t_{\rm trial}]$. If $t_{\rm new}$ also satisfies the Armijo condition, set $t_{\rm trial} \leftarrow t_{\rm new}$ and repeat Step LS2. Otherwise, define $t_k = t_{\rm trial}$ and finish the line search.

Step LS3. Find $t_{\rm new} \in [t_{\rm trial}/\bar{M}, t_{\rm trial}/2]$ and set $t_{\rm trial} \leftarrow t_{\rm new}$. If $t_{\rm trial}$ satisfies the Armijo condition, define $t_k = t_{\rm trial}$ and finish the line search. Otherwise, repeat Step LS3.

The value of $t_{\rm new}$ at Step LS3 may be computed as the safeguarded minimizer of a univariate quadratic or cubic interpolating function. At Step LS2, the computation of $t_{\rm new}$ may contemplate extrapolating techniques.

The choice of the parameter β is more tricky because, unlike the others, this parameter is a dimensional parameter and compares magnitudes of different types. For example, suppose that, instead of minimizing the function $F(x)$, we need to minimize the function $\bar{F}(x) = 10F(x)$. Both problems are completely equivalent, and therefore we would like to observe the same behavior of the algorithm in both cases. However, the conditions $\|d^k\| \geq \beta\|\nabla F(x^k)\|$ and $\|d^k\| \geq \beta\|\nabla \bar{F}(x^k)\|$ are not equivalent. In other words, we should use different values of β in those problems. A more careful analysis would reveal that β should be proportional to the norm of the inverse of the Hessian at x^k. Fortunately, the usual procedures used to compute d^k frequently satisfy the condition $\|d^k\| \geq \beta\|\nabla F(x^k)\|$ automatically for an unknown value of β. For this reason, we generally assign a small value to this parameter, trying to accept d^k as frequently as possible.

8.2 ▪ Magic steps and nonmonotone strategies

At Step 3 of Algorithm 8.1, we require that

$$F(x^k + t_k d^k) \leq F(x^k) + \alpha t_k \nabla F(x^k)^T d^k \tag{8.4}$$

and

$$F(x^{k+1}) \leq F(x^k + t_k d^k). \tag{8.5}$$

Clearly, the choice $x^{k+1} = x^k + t_k d^k$ already satisfies (8.5), but several reasons motivate one to try something better. On the one hand, after computing $x^k + t_k d^k$, it could be interesting to test whether some extrapolation of type $x^k + td^k$ with $t > t_k$ could cause some improvement. Moreover, improvements could come from persevering not only along the computed direction but also along other directions motivated by the specific problem we are trying to solve. Heuristic choices of x^{k+1} satisfying (8.5) are generally called "magic steps" (Conn, Gould, and Toint [83]) and could be crucial for the good practical behavior of the algorithm.

Magic steps may be computed by evoking nonmonotone strategies, watchdog techniques [79], and the spacer step theorem of [181, p. 255]. Sometimes, a sequence of iterates is deemed to converge to the solution of a problem (even very fast!), although without satisfying a monotone decrease of the objective function. This is typical of Newton-like methods for solving subproblems with extreme penalty parameters (and it is related to the phenomenon called the Maratos effect in the optimization literature). In these cases, it is sensible to tolerate some increase in the objective function during some iterations, before returning to rigorous line searches. These ideas may be formalized in the following algorithm.

Algorithm 8.2. Nonmonotone pseudomagic procedure
Given integers $M \geq 0$ and $L \geq 0$ (both around 10), in Algorithm 8.1, after the computation of x^k and before the computation of d^k, set $y^0 = x^k$, $j \leftarrow 0$,

$$F_{\max} = \max\{F(x^k), \dots, F(x^{\max\{k-M,0\}})\},$$

and execute Steps 1–5 below.
Step 1. If $j < L$, compute y^{j+1} and test whether

$$F(y^{j+1}) \leq F_{\max}. \tag{8.6}$$

If y^{j+1} was computed and (8.6) was satisfied, set $j \leftarrow j+1$ and repeat Step 1.
Step 2. Let $y \in \{y^0, \dots, y^j\}$ be such that

$$F(y) = \min\{F(y^0), \dots, F(y^j)\}. \tag{8.7}$$

If $F(y) \geq F(x^k)$, discard y, finish the execution of the present algorithm, and return to the computation of d^k, satisfying (8.2), at Step 2 of Algorithm 8.1.
Step 4. Define $d^k = y - x^k$.
Step 5. If d^k satisfies (8.2) and, in addition,

$$F(y) \leq F(x^k) + \alpha \nabla F(x^k)^T d^k, \tag{8.8}$$

define $x^{k+1} = x^k + d^k = y$ and consider that the kth iteration of Algorithm 8.1 is finished.
Step 6. If d^k does not satisfy (8.2) or does not fulfill (8.8), replace x^k with y and proceed to the computation of d^k at Step 2 of Algorithm 8.1.

Observe that with the inclusion of the nonmonotone magic procedure, the unconstrained Algorithm 8.1 preserves its basic form. The procedure may be considered as an auxiliary device for computing the search direction. If our heuristic for choosing the magic points y^j is good, it could be unnecessary to perform line searches. Note that, after the execution of Algorithm 8.2, we may obtain the next iterate x^{k+1} or we may improve (redefining it) the iterate x^k. In the second case, we should formally "forget" the existence of the previously computed x^k, and we should consider that the new x^k is a result of further improvement in the sense of (8.5) at iteration $k-1$. In the first case, the magic procedure computed, perhaps by chance, a direction d^k that satisfies the basic requirements of Algorithm 8.1. Apparently, there is not a big difference between the possibilities, since in both cases we come up with a better point y. The difference is that, when (8.2) and (8.8) hold, we avoid the necessity of computing a new direction d^k and we open the possibility that all the iterates could be computed as magic steps. In other words, in both cases the magic step was successful but the status given to the step at Step 5 is more relevant than the one given at Step 6. In any case, Algorithm 8.2 specifies aspects of the implementation of Algorithm 8.1 but does not alter its basic structure.

8.3 ▪ Well-definiteness and global convergence

We are going to show first that Algorithm 8.1 is well defined. This means that if the algorithm does not stop at x^k, it is possible to obtain x^{k+1} in finite time. This is the main theoretical and practical condition that an implementable algorithm must satisfy. Well-defined algorithms that do not satisfy additional global convergence properties may be effective in practice [90, 91, 193, 194] but the well-definiteness property is mandatory.

Theorem 8.1. *Algorithm 8.1 is well defined and stops at x^k if and only if $\nabla F(x^k) = 0$.*

Proof. Assume that $\nabla F(x^k) \neq 0$. By the conditions imposed at Step 2 of the algorithm and the differentiability of F,

$$\lim_{t\to 0} \frac{F(x^k + t d^k) - F(x^k)}{t} = \nabla F(x^k)^T d^k < 0.$$

Thus,

$$\lim_{t\to 0} \frac{F(x^k + t d^k) - F(x^k)}{t \nabla F(x^k)^T d^k} = 1.$$

Since $\alpha < 1$, for t small enough, we have that

$$\frac{F(x^k + t d^k) - F(x^k)}{t \nabla F(x^k)^T d^k} \geq \alpha.$$

Now, as $\nabla F(x^k)^T d^k < 0$, we deduce that

$$F(x^k + t d^k) \leq F(x^k) + \alpha t \nabla F(x^k)^T d^k$$

for $t > 0$ small enough. Thus, choosing t_k as the first element of the sequence $\{\bar{M}^{-\ell}\}_{\ell \in \mathbb{N}_0}$ satisfying the condition above, we have that the requirements of Step 3 of Algorithm 8.1 are fulfilled. □

The following theorem is said to be a global convergence theorem. It establishes that, *independently of the initial point*, the gradient must vanish at every limit point. Global convergence in this sense should not be confused with "convergence to global minimizers." Note that the existence of limit points is assumed, and not guaranteed, at this theorem. A sufficient condition for the existence of the limit points is the boundedness of the generated sequence, which, in turn, holds whenever the level set defined by $F(x^0)$ is bounded.

Theorem 8.2. *If x^* is a limit point of a sequence generated by Algorithm 8.1, we have that $\nabla F(x^*) = 0$.*

Proof. Let $K = \{k_0, k_1, k_2, k_3, \dots\} \underset{\infty}{\subset} \mathbb{N}$ be such that

$$\lim_{k \in K} x^k = x^*.$$

By the continuity of F,

$$\lim_{k\in K} F(x^k) = F(x^*).$$

By the Armijo condition, since $k_{j+1} \geq k_j + 1$, we have that

$$F(x^{k_{j+1}}) \leq F(x^{k_j+1}) \leq F(x^{k_j}) + \alpha t_{k_j} \nabla F(x^{k_j})^T d^{k_j} < F(x^{k_j})$$

for all $j \in \mathbb{N}$. Then,

$$\lim_{j\to\infty} t_{k_j} \nabla F(x^{k_j})^T d^{k_j} = 0.$$

Therefore, by (8.2),

$$\lim_{j\to\infty} t_{k_j} \|\nabla F(x^{k_j})\|_2 \|d^{k_j}\|_2 = 0$$

and, by the equivalence of norms in $\mathbb{R}^n$,

$$\lim_{j\to\infty} t_{k_j} \|\nabla F(x^{k_j})\| \|d^{k_j}\| = 0. \tag{8.9}$$

Thus, there exists $K_1 \underset{\infty}{\subset} K$ such that at least one of the two following possibilities is fulfilled:

(a)
$$\lim_{k\in K_1} \|\nabla F(x^k)\| = 0,$$

(b)
$$\lim_{k\in K_1} t_k \|d^k\| = 0. \tag{8.10}$$

In case (a), we deduce that $\nabla F(x^*) = 0$ and the thesis is proved.

In case (b), there exists $K_2 \underset{\infty}{\subset} K_1$ such that at least one of the two following possibilities holds:

(c)
$$\lim_{k\in K_2} \|d_k\| = 0,$$

(d)
$$\lim_{k\in K_2} t_k = 0.$$

In case (c), the conditions (8.2) also imply that $\|\nabla F(x^k)\| \to 0$ for $k \in K_2$ and, therefore, $\nabla F(x^*) = 0$.

Let us consider case (d). Without loss of generality, assume that $t_k < 1$ for all $k \in K_2$. Therefore, by Step 3 of the algorithm, for all $k \in K_2$ there exists $\bar{t}_k > 0$ such that

$$F(x^k + \bar{t}^k d^k) > F(x^k) + \alpha \bar{t}_k \nabla F(x^k)^T d^k. \tag{8.11}$$

Moreover, since $\bar{t}_k \leq \bar{M} t_k$, by (8.10), we have that

$$\lim_{k\in K_2} \bar{t}_k \|d_k\| = 0.$$

Thus, defining $s^k = \bar{t}_k d^k$ for all $k \in K_2$, we have that

$$\lim_{k\in K_2} \|s_k\| = 0. \tag{8.12}$$

By (8.11) and the mean value theorem, for all $k \in K_2$ there exists $\xi_k \in [0,1]$ such that

$$\nabla F(x^k + \xi_k s^k)^T s^k = F(x^k + s^k) - F(x^k) > \alpha \nabla F(x^k)^T s^k. \tag{8.13}$$

Let $K_3 \underset{\infty}{\subset} K_2$ and $s \in \mathbb{R}^n$ be such that $\lim_{k \in K_3} s^k / \|s^k\| = s$. By (8.12), dividing both sides of the inequality (8.13) by $\|s^k\|$ and taking limits for $k \in K_3$, we obtain

$$\nabla F(x^*)^T s \geq \alpha \nabla F(x^*)^T s.$$

Since $\alpha < 1$ and $\nabla F(x^k)^T s^k < 0$ for all k, this implies that $\nabla F(x^*)^T s = 0$. Since by (8.2),

$$-\frac{\nabla F(x^k)^T s^k}{\|s^k\|_2} \geq \theta \|\nabla F(x^k)\|_2 \tag{8.14}$$

for all $k \in K_2$, taking limits in (8.14) for $k \in K_3$, we obtain that $\nabla F(x^*) = 0$. □

The alert reader should observe that the proof of Theorem 8.2 has two essential arguments. The "argument of success" applies to the case in which the step is bounded away from zero. In this case, the size of the direction should go to zero; otherwise the functional value would tend to minus-infinity. The "argument of failure" applies to the case in which the step tends to zero. In this case, there is a different step that also tends to zero along which the function increased or did not decrease enough. This is possible only if the gradient in the limit is zero. These two arguments appear persistently in every global convergence theorem for continuous optimization algorithms. The reader is invited to find them in other line-search procedures, in trust-region methods, in nonsmooth optimization [69], and in derivative-free minimization algorithms [84].

8.4 ▪ Local convergence

Global convergence in the sense used in the former section is a very welcome property for practical unconstrained minimization algorithms, because it guarantees that convergence to points where the gradient does not vanish cannot occur. Nevertheless, the efficiency of algorithms is also linked to theoretical properties of *local convergence*. These properties say that, when the sequence generated by the algorithm passes close to a minimizer, such proximity is recognized and the sequence converges quickly to the solution. To obtain that property, the distance between x^{k+1} and x^k needs to be small when the gradient $\nabla F(x^k)$ is small. We will formalize this requirement in the following assumption.

Assumption 8.1. *There exists $b > 0$ such that consecutive iterates x^k and x^{k+1} in Algorithm 8.1 satisfy*

$$\|x^{k+1} - x^k\| \leq b \|\nabla F(x^k)\| \tag{8.15}$$

for all $k \in \mathbb{N}$.

This assumption is compatible with line searches that obey the requirements of Step 3 of Algorithm 8.1.

Our strategy to prove local superlinear convergence has three parts. In Theorem 8.3 below, we prove that, if x^* is an isolated limit point, the whole sequence converges to x^*. In Theorem 8.4 below, we prove that, if the initial point is close to a strict local minimizer x^*, the whole sequence converges to x^*. Although analogous, these two theorems are independent (none of them is deduced from the other). However, both show that the convergence of the whole sequence to a single point may be expected in many cases. In other words, convergence of the whole sequence is a *plausible* assumption. Using it, and assuming further that the directions d^k are obtained using a Newtonian philosophy, we will prove superlinear convergence.

A point x^* is said to be *isolated* if there exists $\epsilon > 0$ such that $\nabla F(x) \neq 0$ for all $x \in B(x^*, \epsilon)$ such that $x \neq x^*$.

Theorem 8.3. *Assume that x^* is isolated, the sequence $\{x^k\}$ is generated by Algorithm 8.1 with Assumption 8.1, and $\lim_{k \in K} x^k = x^*$ for some subsequence $K \underset{\infty}{\subset} \mathbb{N}$. Then, $\nabla F(x^*) = 0$ and $\lim_{k \to \infty} x^k = x^*$.*

Proof. The fact that $\nabla F(x^*) = 0$ is a consequence of Theorem 8.2. Since x^* is isolated, there exists $\epsilon > 0$ such that $\nabla F(x) \neq 0$ for all $x \in B(x^*, \epsilon) \setminus \{x^*\}$.

Let us define

$$C = \{x \in \mathbb{R}^n \mid \epsilon/2 \le \|x - x^*\| \le \epsilon\}.$$

Clearly, C is compact and does not contain points where the gradient vanishes. Therefore, by Theorem 8.2, C does not contain infinitely many iterates. Let $k_1 \in \mathbb{N}$ be such that $x^k \notin C$ for all $k \ge k_1$.

Define

$$K_1 = \{k \ge k_1 \mid \|x^k - x^*\| \le \epsilon/2\} \subseteq \mathbb{N}.$$

Note that K_1 is nonempty since, by hypothesis, $\lim_{k \in K} x^k = x^*$.

Since x^* is the unique stationary point in the ball with radius $\epsilon/2$, by Theorem 8.2, we have that $\lim_{k \in K_1} x^k = x^*$ and, by the continuity of ∇F, that $\lim_{k \in K_1} \|\nabla F(x^k)\| = 0$. Then, by Assumption 8.1, $\lim_{k \in K_1} \|x^{k+1} - x^k\| = 0$. This implies that, for all $k \in K_1$ large enough, $\|x^{k+1} - x^k\| < \epsilon/2$. Then, since, by the definition of K_1, we have $\|x^k - x^*\| \le \epsilon/2$, the triangle inequality implies that $\|x^{k+1} - x^*\| \le \epsilon$. However, since $x^{k+1} \notin C$, we have that $\|x^{k+1} - x^*\| \le \epsilon/2$. Therefore, we proved that, for all $k \in K_1$ large enough, we have that $k+1$ also belongs to K_1. This implies that $\|x^k - x^*\| \le \epsilon/2$ for all k large enough. Invoking again Theorem 8.2 and the isolation of x^*, it follows that

$$\lim_{k \to \infty} x^k = x^*,$$

as we wanted to prove. □

Theorem 8.4. *Assume that the isolated point x^* is a strict local minimizer and the sequence $\{x^k\}$ is generated by Algorithm 8.1 with Assumption 8.1. Then, there exists $\delta_1 > 0$ such that, if $\|x^{k_0} - x^*\| \le \delta_1$ for some k_0, we have that $\lim_{k \to \infty} x^k = x^*$.*

Proof. Let $\epsilon > 0$ be such that x^* is a strict global minimizer of f on the ball $B(x^*, \epsilon)$ and assume that this ball does not contain any other point in which the gradient vanishes. By the continuity of ∇F and Assumption 8.1, there exists $\delta \in (0, \epsilon/2)$ such that

$$\|x^k - x^*\| \le \delta \Rightarrow \|x^{k+1} - x^k\| \le \epsilon/2. \tag{8.16}$$

Let c be the minimum value of $F(x)$ on the set $\{x \in \mathbb{R}^n \mid \delta \le \|x - x^*\| \le \epsilon\}$. Let $\delta_1 \in (0, \delta)$ be such that $\|x - x^*\| \le \delta_1 \Rightarrow F(x) < c$.

Let us prove by induction that, if there exists k_0 such that $\|x^{k_0} - x^*\| \le \delta_1$, one obtains $\|x^k - x^*\| \le \delta$ and $F(x^k) < c$ for all $k \ge k_0$. By the definition of δ_1, this is trivially true if $k = k_0$. Now, let us assume it is valid for k and prove it for $k+1$. Observe that, by (8.16), $\|x^k - x^*\| \le \delta$ implies that $\|x^{k+1} - x^*\| \le \epsilon$. Moreover, $F(x^k) < c$ and the fact that $F(x^{k+1}) < F(x^k)$ imply $F(x^{k+1}) < c$, and this implies that $\|x^{k+1} - x^*\| < \delta$.

Therefore, since $\delta < \varepsilon/2$ for k large enough, all the elements of the sequence are contained in $B(x^*, \epsilon/2)$. Since x^* is the unique point where the gradient vanishes in this ball, Theorem 8.3 implies that the whole sequence converges to x^* as we wanted to prove. □

8.4.1 ▪ Convergence under Newton-like choices of the search directions

The next algorithm is a particular case of Algorithm 8.1 and defines a general (Newton-like) form in which the direction d^k may be computed. This direction will be the approximate solution of a linear system of the form $B_k d = -\nabla F(x^k)$. The idea is that the gradient $\nabla F(x)$ is well approximated by the linear function $B_k(x - x^k) + \nabla F(x^k)$ in a neighborhood of x^k or, equivalently, that the objective function $F(x)$ is well approximated by the quadratic $(1/2)(x-x^k)^T B_k(x-x^k) + \nabla F(x^k)^T(x-x^k) + F(x^k)$ in such a way that a solution of $B_k(x-x^k) + \nabla F(x^k) = 0$ could be a good approximation to the solution of the problem. In fact, this corresponds to the "Newtonian paradigm" for solving many nonlinear mathematical problems: Approximate the original problem by an easy (in some sense, linear) problem using information at x^k and use the solution of the "subproblem" to continue the process. The Newtonian linear system may be solved exactly (discarding rounding errors) and using B_k as the true Hessian at x^k (classical Newton). If the subproblem is solved approximately employing an iterative linear solver but still using the true Hessian, we say we are using the inexact Newton or truncated Newton approach. When B_k is only an approximation of the Hessian at x^k, we talk about quasi-Newton methods (inexact quasi-Newton in the case that the quasi-Newton system is solved only approximately).

In this context, step $t_k = 1$ should be preferred in some sense because it thoroughly corresponds to the Newtonian paradigm. When this step satisfies the Armijo condition, we decide to stop the line search. Global convergence will require, on the other hand, that the matrices B_k be positive definite and that their inverses be bounded.

Algorithm 8.3. This algorithm corresponds to an implementation of Algorithm 8.1 in which the following hold:

(a) The direction d^k is such that

$$\|B_k d^k + \nabla F(x^k)\| \le \eta_k \|\nabla F(x^k)\|, \tag{8.17}$$

where $B_k \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and $\eta_k \in [0,1)$.

(b) If $F(x^k + d^k) \le F(x^k) + \alpha \nabla F(x^k)^T d^k$, we choose $t_k = 1$ and $x^{k+1} = x^k + d^k$.

In order to say that Algorithm 8.3 is an implementation of Algorithm 8.1, we must show that the direction d^k computed in **(a)** satisfies (8.2). A sufficient condition for the fulfillment of (8.2) is the boundedness of $\|B_k\|$ and $\|B_k^{-1}\|$, stated in Assumption 8.2. This claim is proved in Theorem 8.5 below.

Assumption 8.2. *The sets* $\{\|B_k\|, k \in \mathbb{N}\}$ *and* $\{\|B_k^{-1}\|, k \in \mathbb{N}\}$ *are bounded.*

Theorem 8.5. *Assume that the sequence* $\{x^k\}$ *is generated by Algorithm 8.3 and that Assumption 8.2 holds. Then, there exist* $\eta_{\max} \in (0,1)$, $\theta \in (0,1)$, *and* $\beta > 0$ *such that, for all* $k \in \mathbb{N}$, *if* $\eta_k \le \eta_{\max}$, *then (8.2) is satisfied.*

Proof. Define $r^k = B_k d^k + \nabla F(x^k)$. Then,

$$\begin{aligned}\|\nabla F(x^k)\| &= \|\nabla F(x^k) - r^k + r^k\| \le \|\nabla F(x^k) - r^k\| + \|r^k\| \\ &= \|B^k d^k\| + \|r^k\| \le \|B^k\|\|d^k\| + \|r^k\|.\end{aligned}$$

Therefore,

$$\|\nabla F(x^k)\| - \|r^k\| \le \|B^k\|\|d^k\|.$$

Thus, since by (8.17) $||r^k|| \le \eta_k ||\nabla F(x^k)||$, we deduce that

$$(1-\eta_k)||\nabla F(x^k)|| \le ||B_k|| ||d^k||.$$

Consequently,

$$||d^k|| \ge \frac{1-\eta_k}{||B_k||} ||\nabla F(x^k)||.$$

Taking $\eta_k \le 1/2$, we have that

$$||d^k|| \ge \frac{1}{2||B_k||} ||\nabla F(x^k)||.$$

Assuming that $||B_k|| \le c$ for all $k \in \mathbb{N}$, the second requirement of (8.2) is satisfied with $\beta = \frac{1}{2c}$.

Let us prove the fulfillment of the angle condition in (8.2). Since $-\nabla F(x^k) = B_k d^k - r^k$, premultiplying by $(d^k)^T$ yields

$$-(d^k)^T \nabla F(x^k) = (d^k)^T B_k d^k - (d^k)^T r^k.$$

By the spectral decomposition [127] of B_k, we have that

$$\lambda_{\min}(B_k) \le \frac{(d^k)^T B_k d^k}{||d^k||_2^2} \le \lambda_{\max}(B_k),$$

where $\lambda_{\min}(B_k)$ and $\lambda_{\max}(B_k)$ represent the smallest and the largest eigenvalues of B_k. Since $||B_k^{-1}||_2 = \lambda_{\max}(B_k^{-1}) = 1/\lambda_{\min}(B_k)$, we have that

$$(d^k)^T B_k d^k \ge \frac{||d^k||_2^2}{||B_k^{-1}||_2}.$$

Therefore,

$$-(d^k)^T \nabla F(x^k) \ge \frac{||d^k||_2^2}{||B_k^{-1}||_2} - (d^k)^T r^k.$$

Thus, by the first part of the proof and the equivalence of norms in $\mathbb{R}^n$, there exists $\beta_2 > 0$ such that

$$-(d^k)^T \nabla F(x^k) \ge \beta_2 \frac{||d^k||_2 ||\nabla F(x^k)||_2}{||B_k^{-1}||_2} - (d^k)^T r^k$$

for all k such that $\eta_k \le 1/2$. Assuming that $||B_k^{-1}||_2 \le c_2$ for all $k \in \mathbb{N}$, this implies that

$$(d^k)^T \nabla F(x^k) \le -\beta_2 \frac{||d^k||_2 ||\nabla F(x^k)||_2}{c_2} + (d^k)^T r^k.$$

Now, since $||r^k|| \le \eta_k ||\nabla F(x^k)||$, by the equivalence of norms in $\mathbb{R}^n$, there exists a constant $c_{\text{norm}} > 0$ such that

$$||r^k||_2 \le c_{\text{norm}} \eta_k ||\nabla F(x^k)||_2.$$

Hence,

$$(d^k)^T r^k \le ||d^k||_2 ||r^k||_2 \le c_{\text{norm}} \eta_k ||d^k||_2 ||\nabla F(x^k)||_2$$

and, therefore,

$$\begin{aligned}(d^k)^T\nabla F(x^k) &\le -(\beta_2/c_2)\|d^k\|_2\|\nabla F(x^k)\|_2 + c_{\text{norm}}\eta_k\|d^k\|_2\|\nabla F(x^k)\|_2\\ &= -(\beta_2/c_2 - c_{\text{norm}}\eta_k)\|d^k\|_2\|\nabla F(x^k)\|_2.\end{aligned}$$

Thus if, say, $\eta_k \le \min\{\frac{1}{2}, \frac{1}{2}(\beta_2/(c_2 c_{\text{norm}}))\}$, we have that $\bar\theta = \beta_2/c_2 - c_{\text{norm}}\eta_k > 0$ and that

$$(d^k)^T\nabla F(x^k) \le -\bar\theta\|d^k\|_2\|\nabla F(x^k)\|_2.$$

This means that the angle condition of (8.2) is satisfied with $\theta = \bar\theta$. □

The following theorem completes the basic convergence theory of Algorithm 8.1. We will show that under Assumption 8.2, if the sequence generated by Algorithm 8.3 converges to a local minimizer x^* where the Hessian $\nabla^2 F(x^*)$ is positive definite and the matrices B_k are approximations of the Hessians $\nabla^2 F(x^k)$ in the sense of Dennis and Moré [97], the convergence is superlinear and, for k large enough, we have that $t_k = 1$. In other words, for k large enough, we will need only one function evaluation per iteration.

Theorem 8.6. *Assume that the sequence $\{x^k\}$ is generated by Algorithm 8.3 with Assumption 8.2 and $x^k \ne x^*$ for all $k \in \mathbb{N}$, $\lim_{k\to\infty} x^k = x^*$, F admits continuous third derivatives in a neighborhood of x^*, $\nabla^2 F(x^*)$ is positive definite, and the Dennis–Moré condition*

$$\lim_{k\to\infty} \frac{\left\|\left[B_k - \nabla^2 F(x^k)\right]d^k\right\|}{\|d^k\|} = 0$$

and the inexact Newton condition

$$\lim_{k\to\infty} \eta_k = 0$$

are verified. Then, there exists $k_0 \in \mathbb{N}$ such that $t_k = 1$ for all $k \ge k_0$ and the sequence $\{x^k\}$ converges superlinearly to x^*. Moreover, if $B_k = \nabla^2 F(x^k)$ and $\eta_k = 0$ for all $k \in \mathbb{N}$, the convergence is quadratic.

Proof. By Taylor's formula, we have that

$$\begin{aligned}&F(x^k + d^k) - F(x^k) - \alpha\nabla F(x^k)^T d^k\\ &\quad = (1-\alpha)\nabla F(x^k)^T d^k + \tfrac{1}{2}(d^k)^T\nabla^2 F(x^k)d^k + o(\|d^k\|^2)\\ &\quad = (1-\alpha)(d^k)^T\left[\nabla F(x^k) + \nabla^2 F(x^k)d^k\right] + \left(\alpha - \tfrac{1}{2}\right)(d^k)^T\nabla^2 F(x^k)d^k + o(\|d^k\|^2).\end{aligned}$$

Defining $r^k = B_k d^k + \nabla F(x^k)$, by Step 2 of the algorithm and $\eta_k \to 0$, we have that $\|r^k\| = o(\|\nabla F(x^k)\|) = o(\|d^k\|)$. Therefore,

$$\begin{aligned}F(x^k + d^k) - F(x^k) - \alpha\nabla F(x^k)^T d^k &= (1-\alpha)(d^k)^T r^k + (1-\alpha)(d^k)^T\left[\nabla^2 F(x^k) - B_k\right]d^k\\ + \left(\alpha - \tfrac{1}{2}\right)(d^k)^T\nabla^2 F(x^k)d^k + o(\|d^k\|^2) &= (1-\alpha)(d^k)^T\left[\nabla^2 F(x^k) - B_k\right]d^k\\ &\quad + \left(\alpha - \tfrac{1}{2}\right)(d^k)^T\nabla^2 F(x^k)d^k + o(\|d^k\|^2).\end{aligned}$$

Now, by the Dennis–Moré condition, we have that

$$(1-\alpha)(d^k)^T\left[\nabla^2 F(x^k) - B_k\right]d^k = o(\|d^k\|^2),$$

and, therefore,

$$F(x^k + d^k) - F(x^k) - \alpha(d^k)^T\nabla F(x^k) = \left(\alpha - \frac{1}{2}\right)(d^k)^T\nabla^2 F(x^k)d^k + o(\|d^k\|^2). \tag{8.18}$$

Let $\mu > 0$ be a lower bound for the eigenvalues of $\nabla^2 F(x^*)$. Then, there exists k_1 such that $\mu/2$ is a lower bound for the eigenvalues of $\nabla^2 F(x^k)$ for all $k \geq k_1$. Thus, for all $k \geq k_1$, we have that

$$\frac{(d^k)^T \nabla^2 F(x^k) d^k}{||d^k||^2} \geq \mu/2.$$

Since $\alpha < 1/2$, by (8.18), we have that

$$\frac{F(x^k + d^k) - F(x^k) - \alpha (d^k)^T \nabla F(x^k)}{||d^k||^2} \leq \left(\alpha - \frac{1}{2}\right)\frac{\mu}{2} + \frac{o(||d^k||^2)}{||d^k||^2} \qquad (8.19)$$

for $k \geq k_1$. But, since $\{||B_k^{-1}||, k \in \mathbb{N}\}$ is bounded, $\nabla F(x^k) \to 0$ (by Theorem 8.2), and $\eta_k \to 0$, $||r_k|| \leq \eta_k ||F(x^k)||$ implies that $||d^k|| \to 0$. Therefore, taking limits in (8.19) for $k \to \infty$, we obtain

$$F(x^k + d^k) - F(x^k) - \alpha \nabla F(x^k)^T d^k \leq 0$$

for k large enough. Then, by the definition of the algorithm, there exists $k_0 \in \mathbb{N}$ such that $t_k = 1$ for all $k \geq k_0$. Thus, the first part of the thesis is proved.

By the first part of the thesis and the definition of Algorithm 8.3, we have that

$$x^{k+1} - x^k = d^k \text{ for all } k \geq k_0. \qquad (8.20)$$

Then, by Taylor's formula,

$$\begin{aligned} \nabla F(x^{k+1}) &= \nabla F(x^k) + \nabla^2 F(x^k) d^k + O(||d^k||^2) \\ &= B_k d^k + \nabla F(x^k) + [\nabla^2 F(x^k) - B_k] d^k + O(||d^k||^2) \qquad (8.21) \\ &= r^k + [\nabla^2 F(x^k) - B_k] d^k + O(||d^k||^2). \end{aligned}$$

As in the first part of the proof we have that $||r^k|| = o(||d^k||)$. Therefore,

$$\nabla F(x^{k+1}) = [\nabla^2 F(x^k) - B_k] d^k + o(||d^k||).$$

Then, by the Dennis–Moré condition and (8.20),

$$\lim_{k\to\infty} \frac{||\nabla F(x^{k+1})||}{||x^{k+1} - x^k||} = 0.$$

Since, by the mean value theorem of integral calculus, we have that

$$\nabla F(x^{k+1}) - \nabla F(x^*) = \left[\int_0^1 \nabla^2 F(x^* + t(x^{k+1} - x^*) dt\right](x^{k+1} - x^*),$$

then, by the continuity and nonsingularity of the Hessian at x^*, we deduce that

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^{k+1} - x^k||} = 0.$$

Therefore,

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^{k+1} - x^*|| + ||x^k - x^*||} = 0.$$

Thus,

$$\lim_{k\to\infty} 1+\frac{||x^k-x^*||}{||x^{k+1}-x^*||}=\infty$$

and, consequently,

$$\lim_{k\to\infty} \frac{||x^{k+1}-x^*||}{||x^k-x^*||}=0. \tag{8.22}$$

Then, the convergence is superlinear, as we wanted to prove.

Finally, let us prove that the convergence is quadratic when $B_k=\nabla^2 F(x^k)$ and $\eta_k=0$ for all $k\in\mathbb{N}$. In this case, by (8.21) we have that

$$||\nabla F(x^{k+1})||=O(||d^k||^2).$$

So, since $t_k=1$ for k large enough, there exists $c>0$ such that

$$||\nabla F(x^{k+1})||\le c||x^{k+1}-x^k||^2 \tag{8.23}$$

for k large enough. By the mean value theorem of integral calculus and the continuity and nonsingularity of the Hessian at x^*, (8.23) implies that there exists $c_1>0$ such that

$$||x^{k+1}-x^*||\le c_1||x^{k+1}-x^k||^2$$

for k large enough. Then,

$$\frac{||x^{k+1}-x^*||}{||x^{k+1}-x^*||+||x^k-x^*||}\le c_1||x^{k+1}-x^k||.$$

Therefore,

$$\frac{||x^k-x^*||}{||x^{k+1}-x^*||}\ge\frac{1}{c_1||x^{k+1}-x^k||}-1=\frac{1-c_1||x^{k+1}-x^k||}{c_1||x^{k+1}-x^k||}.$$

So,

$$\frac{||x^{k+1}-x^*||}{||x^k-x^*||}\le\frac{c_1||x^{k+1}-x^k||}{1-c_1||x^{k+1}-x^k||}.$$

Taking k large enough, since $||x^{k+1}-x^k||\to 0$, we have that

$$\frac{||x^{k+1}-x^*||}{||x^k-x^*||}\le 2c_1||x^{k+1}-x^k||\le 2c_1(||x^{k+1}-x^*||+||x^k-x^*||).$$

But, by (8.22), $||x^{k+1}-x^*||\le||x^k-x^*||$ for k large enough; thus

$$\frac{||x^{k+1}-x^*||}{||x^k-x^*||}\le 4c_1||x^k-x^*||,$$

and the quadratic convergence follows from this inequality. □

Theorem 8.6 assumes the existence and continuity of the second derivatives at x^*. However, we know that the Augmented Lagrangian function $L_\rho(x,\lambda,\mu)$ does not admit second derivatives with respect to x at the points in which $g_i(x)+\mu_i/\rho=0$ for some i. This is not a serious drawback and does not eliminate the explicative power of the theorem because of three main reasons. On the one hand, in many constrained optimization problems, the *strict complementarity* property holds, meaning that $\mu_i^* > 0$ whenever $g_i(x^*) = 0$. Coming back to the chapter about boundedness of the penalty parameters, we see that, with additional conditions, this property is inherited by the approximate Lagrange multipliers and by their safeguarded approximations $\bar{\mu}_i$. Since, in addition, the penalty parameter will be bounded, we probably have continuous second derivatives in a neighborhood of the solution. On the other hand, even if the second derivatives are discontinuous at the solution, the gradient of the Augmented Lagrangian is *semismooth* in the sense of Qi and Sun [219]. Roughly speaking, this means that Newton's method and inexact Newton methods (Martínez and Qi [189]) can be defined and have good local convergence properties. Finally, constraints $g_i(x) \le 0$ may be replaced by $g_i(x) + z_i = 0, z_i \ge 0$, where z_i is a slack variable. With such reformulation, no second derivative discontinuity occurs.

It is important to interpret correctly the results of this section in order to understand the computational behavior in practical situations. The global convergence Theorem 8.2 provides a general frame for the behavior of the algorithms, but the user of computational optimization methods is not expected in practice to observe iterates jumping between different accumulation points of the sequence $\{x^k\}$. The reason is given by the "capture" Theorems 8.3 and 8.4. These theorems say that isolated stationary points are powerful attractors for the algorithms considered here. This means that, in practice, the algorithm produces a sequence $\{x^k\}$ such that $\|x^k\| \to \infty$ without limit points or such that it converges to a single point x^*. The possibility $\|x^k\| \to \infty$ is discarded if the level sets are bounded; therefore the assumption $x^k \to x^*$ of Theorem 8.6 is not arbitrary. However, this theorem includes additional hypotheses that deserve to be discussed.

The Dennis–Moré condition is one of these hypotheses. It states that

$$\left\| \left[B_k - \nabla^2 F(x^k) \right] d^k \right\| = o(\|d^k\|).$$

This condition is obviously satisfied if we choose $B_k = \nabla^2 F(x^k)$ and tends to be fulfilled if B_k is close to the true Hessian, in particular if $\|B_k - \nabla^2 F(x^k)\| \to 0$. (Consider, for example, the case in which the Hessian is computed using finite differences with discretization steps that tend to zero.) However, the Dennis–Moré condition tends to be fulfilled under much weaker assumptions. Observe first that, by Taylor,

$$\left\| \nabla^2 F(x^k) d^k - \left[\nabla F(x^k + d^k) - \nabla F(x^k) \right] \right\| = o(\|d^k\|).$$

Therefore, the Dennis–Moré condition will hold whenever

$$\left\| B_k d^k - \left[\nabla F(x^k + d^k) - \nabla F(x^k) \right] \right\| = o(\|d^k\|).$$

Now, assume that we choose the successive matrices B_k in order to satisfy the *secant equation* (Dennis and Schnabel [98]) given by

$$B_{k+1} d^k = \nabla F(x^k + d^k) - \nabla F(x^k).$$

Under the hypotheses of Theorem 8.6, this condition is asymptotically equivalent to

$$B_{k+1}(x^{k+1} - x^k) = \nabla F(x^{k+1}) - \nabla F(x^k). \tag{8.24}$$

In this case, the Dennis–Moré condition will be satisfied whenever $\lim_{k\to\infty} \|B_{k+1} - B_k\| = 0$. Over several decades, many efforts in numerical optimization have been made to define methods that enjoy compatibility between the secant equation (8.24) and a low-variation requirement. Those methods, generally called *secant methods*, avoid the computation of second derivatives in the context of unconstrained optimization. The Dennis–Moré condition is the key tool for their convergence analysis.

The second important hypothesis of Theorem 8.6 is $\lim_{k\to\infty} \eta_k = 0$. The strict interpretation of this hypothesis would require the a priori definition of a sequence that converges to zero, as $\{1/k\}$ or $\{1/k^2\}$. A less trivial interpretation comes from considering that η_k is the tolerance for the error in the solution of the linear system $B_k d = -\nabla F(x^k)$, measured by the comparison of the residuals at $d = 0$ and at the computed approximate solution. The smaller this tolerance, the more accurate the linear system solution, and we could expect something close to superlinear convergence. In general, we have two possibilities: (i) we solve the linear system "exactly" or (ii) we solve it only approximately using some iterative method such as conjugate gradients. In the first case, η_k is very small but not exactly zero, because in the computer we work with high precision but not infinite precision. The exact solution of the system is frequently associated with the observance of superlinear convergence. When we solve the system using an iterative method, it is usual to fix a unique value for η_k (small), so that the convergence, although not superlinear, is reasonably fast.

Now, how is superlinear convergence observed in practice? Before answering this question, let us formulate another: Is it observable that, for k large enough, $t_k = 1$, that is, that each iteration asymptotically involves a single function evaluation? The answer to the second question is *yes*. In well-behaved problems, in which the sequence generated by the algorithm converges to a point x^* with positive definite Hessian, we really observe that, near x^*, the first tentative $t = 1$ produces enough decrease and, consequently, $x^{k+1} = x^k + d^k$ if B_k is similar to a true Hessian $\nabla^2 F(x^k)$ and the solution of the linear system is accurate. This behavior is not observed only if the Hessian is almost singular or ill-conditioned or if the third derivatives are dominant, so the Lipschitz constant of the Hessians is very big and the basin of convergence of Newton's method is small. (Unfortunately this may be the situation when F has the form (8.1) and the penalty parameter is big.)

Finally, superlinear convergence means that $||x^{k+1} - x^*||/||x^k - x^*|| \to 0$, a property that, assuming that $\nabla^2 F(x^*)$ is nonsingular and continuous in a neighborhood of x^* (by the mean value theorem of integral calculus), is equivalent to

$$\lim_{k\to\infty} \frac{||\nabla F(x^{k+1})||}{||\nabla F(x^k)||} = 0. \tag{8.25}$$

Usually, in well-behaved problems, one observes that $||\nabla F(x^{k+1})||$ decreases significantly with respect to $||\nabla F(x^k)||$, but the user should not expect an academic immaculate convergence to zero of the quotient. If we observe that, at some iteration, the norm of the gradient is, say, one half the norm of the previous one, with some tendency to (usually nonmonotone) decrease, this does not mean that superlinear convergence is being violated. Ultimately, (8.25) is an asymptotic property.

8.5 ▪ Computing search directions

8.5.1 ▪ Newton and stabilized Newton approaches

Newton's method may be applied to minimize the function $F(x)$ given by (8.1). Denoting

$$H(x) = h(x) + \lambda/\rho \text{ and } G(x) = g(x) + \mu/\rho,$$

we have that

$$F(x)=f(x)+\frac{\rho}{2}\left(\|H(x)\|_2^2+\|G(x)_+\|_2^2\right). \tag{8.26}$$

The function F has continuous first derivatives. The second derivatives are discontinuous, but $\nabla F(x)=0$ is a semismooth [219] system of equations so that the Newton's approach makes sense and its unitary-step version converges quadratically under local nonsingularity conditions. Inexact Newton methods can also be applied for solving that system [189].

The iterates generated by the algorithm for minimizing $F(x)$ will be denoted, as usual, by x^k. Without loss of generality, let us assume that, given a generic iterate x^k, we have that $G_i(x^k)\geq 0$ for all $i=1,\dots,q$ and $G_i(x^k)<0$ for $i=q+1,\dots,p$. Consequently, we define

$$\underline{G}(x)=(G_1(x),\dots,G_q(x))^T.$$

Clearly, Newton's iteration for the minimization of $F(x)$ using the current point x^k coincides with Newton's iteration for the minimization of $f(x)+\frac{\rho}{2}(\|H(x)\|_2^2+\|\underline{G}(x)\|_2^2)$. With abuse of notation, let us redefine

$$F(x)=f(x)+\frac{\rho}{2}\left(\|H(x)\|_2^2+\|\underline{G}(x)\|_2^2\right).$$

Therefore,

$$\nabla F(x)=\nabla f(x)+\rho\left(\nabla H(x)H(x)+\nabla\underline{G}(x)\underline{G}(x)\right)$$

and

$$\nabla^2 F(x)=\nabla^2 f(x)+\rho\left(H'(x)^T H'(x)+\underline{G}'(x)^T\underline{G}'(x)+\sum_{i=1}^{m}H_i(x)\nabla^2 H_i(x)\right.$$
$$\left.+\sum_{i=1}^{q}\underline{G}_i(x)\nabla^2\underline{G}_i(x)\right).$$

In principle, Newton's iteration requires the solution of the linear system

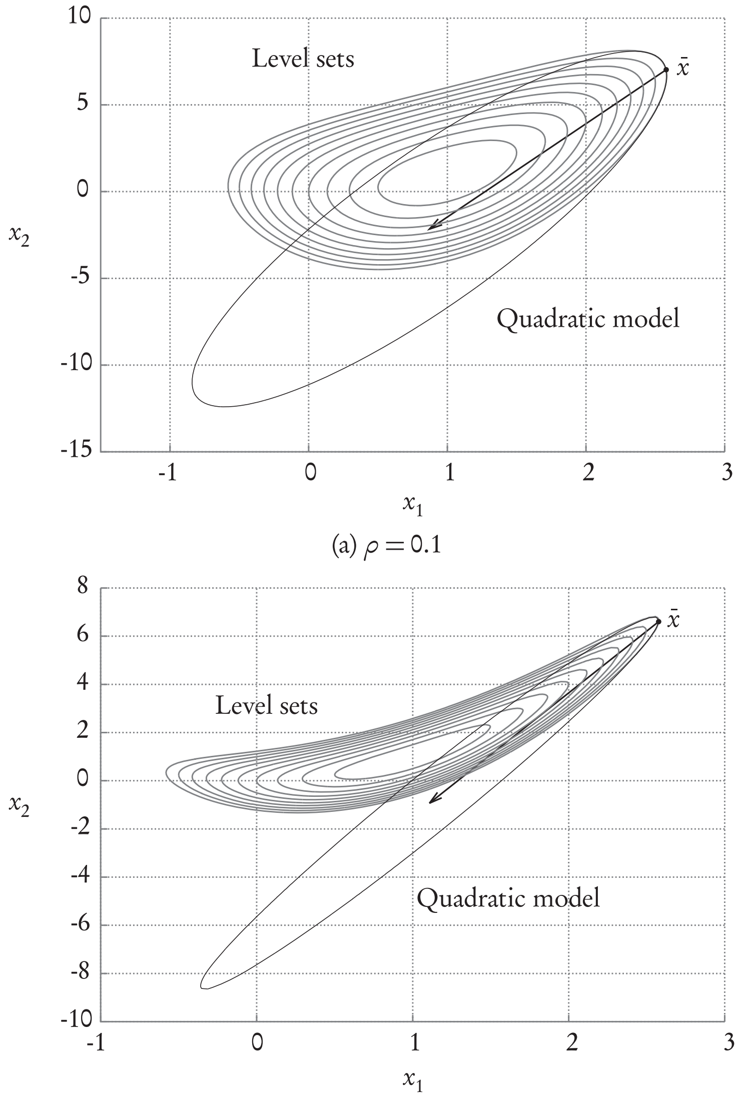
$$\nabla^2 F(x^k)(x-x^k)=-\nabla F(x^k). \tag{8.27}$$

The value of ρ may be large because it needed to be increased many times after the test (4.9) or, more frequently, because one deliberately decides to start with a big penalty parameter with the aim of getting a solution very fast. This "shortcut" strategy (Fletcher [113]) may be useful when we have a guaranteed good starting point, in addition to a good approximation of the Lagrange multipliers, and we do not want to lose feasibility at the first Augmented Lagrangian iterations. (This is exactly what we wish when dealing with parametric optimization [134, 156].) However, in this case, the naive application of Newton's method may lead to poor results due to the following reasons:

1. Although Newton's direction is generally a descent direction for $F(x)$ (at least after some possible correction on the matrix of the system), the unitary step, which should generate quadratic convergence, may not be accepted by monotone line search procedures unless the current point is very close to the solution.

2. The Newtonian linear system is generally very ill-conditioned if ρ is large.

Both phenomena are associated with the size of the penalty parameter but they are not the same phenomenon. The second one may be overcome by means of a decomposition of the Newtonian linear system. This fact leads some people to argue that there is no real problem with big penalty parameters. However, the first phenomenon persists even if we solve the linear system by means of decomposition techniques, because it is intrinsic



(a) $\rho = 0.1$

(b) $\rho = 1$

Figure 8.1. *Level sets and Newton's direction for $F(x) = f(x) + (\rho/2)||H(x)||_2^2$, where $f(x) = (1 - x_1)^2$ and $H(x) = 2(x_2 - x_1^2)$. (Note that $F(x)$ with $\rho = 100$ is the Rosenbrock's function.) In (a), $\rho = 0.1$, while in (b) $\rho = 1$. In both cases, $\bar{x}$ is approximately at the same distance to the solution $x^* = (1,1)^T$. In the first case, the unitary step along the Newton's direction produces a decrease in $F(x)$, while in the second case it does not.*

to the fact that for big values of ρ, functional values of $F(x)$ are dominated by those of $(\rho/2)[||H(x)||_2^2 + ||G(x)_+||_2^2]$. If ρ is large, the level sets of $(\rho/2)[||H(x)||_2^2 + ||G(x)_+||_2^2]$ tend to be parallel surfaces to the feasible set and $F(x)$ tends to produce flat curved valleys as level sets, more or less following the shape of the boundary of the feasible region. In these conditions, the minimizer of the quadratic approximation tends to be outside the level set, as shown in Figure 8.1. Completely overcoming this inconvenience is impossible, but nonmonotone techniques, which tolerate an occasional increase of the functional value, tend to alleviate it.

In any case, it is always convenient to decompose the system in such a way that the computation of Newtonian directions becomes well-conditioned.

The Newtonian linear system may be written as

$$\begin{aligned}&\left[B(x^k)+\rho\left(H'(x^k)^T H'(x^k)+\underline{G}'(x^k)^T\underline{G}'(x^k)\right)\right](x-x^k)\\&=-\left(\nabla f(x^k)+\rho\left(\nabla H(x^k)H(x^k)+\nabla\underline{G}(x^k)\underline{G}(x^k)\right)\right),\end{aligned} \tag{8.28}$$

where

$$B(x^k)=\nabla^2 f(x^k)+\rho\left(\sum_{i=1}^{m}H_i(x^k)\nabla^2 H_i(x^k)+\sum_{i=1}^{q}\underline{G}_i(x^k)\nabla^2\underline{G}_i(x^k)\right). \tag{8.29}$$

Note that the matrix of (8.28) should be positive definite. (In particular, its diagonal elements should be positive, a fact that is easy to verify and correct even in the decomposition context that follows.)

The system (8.28) is equivalent to

$$\begin{aligned}&B(x^k)(x-x^k)+\rho\nabla H(x^k)\left(H(x^k)+H'(x^k)(x-x^k)\right)\\&\quad+\rho\nabla\underline{G}(x^k)\left(\underline{G}(x^k)+\underline{G}'(x^k)(x-x^k)\right)=-\nabla f(x^k).\end{aligned}$$

Defining

$$\lambda_{\text{new}}=\rho\left(H(x^k)+H'(x^k)(x-x^k)\right)\ \text{ and }\ \mu_{\text{new}}=\rho\left(\underline{G}(x^k)+\underline{G}'(x^k)(x-x^k)\right),$$

the system becomes

$$B(x^k)(x-x^k)+\nabla H(x^k)\lambda_{\text{new}}+\nabla\underline{G}(x^k)\mu_{\text{new}}=-\nabla f(x^k).$$

In other words, the Newtonian direction $x-x^k$ comes from solving the linear equations

$$\begin{aligned}B(x^k)(x-x^k)+\nabla H(x^k)\lambda_{\text{new}}+\nabla\underline{G}(x^k)\mu_{\text{new}}&=-\nabla f(x^k),\\H'(x^k)(x-x^k)-\lambda_{\text{new}}/\rho&=-H(x^k),\\\underline{G}'(x^k)(x-x^k)-\mu_{\text{new}}/\rho&=-\underline{G}(x^k),\end{aligned}$$

which, in matricial terms, can be written as

$$\begin{pmatrix}B(x^k)&H'(x^k)^T&\underline{G}'(x^k)^T\\H'(x^k)&-I/\rho&0\\\underline{G}'(x^k)&0&-I/\rho\end{pmatrix}\begin{pmatrix}x-x^k\\\lambda_{\text{new}}\\\mu_{\text{new}}\end{pmatrix}=\begin{pmatrix}-\nabla f(x^k)\\-H(x^k)\\-\underline{G}(x^k)\end{pmatrix}. \tag{8.30}$$

Finally, note that $\rho H_i(x^k)=\lambda_i+\rho h_i(x^k)$ and $\rho\underline{G}_i(x^k)=\rho\max\{0,g_i(x^k)+\mu_i/\rho\}=\max\{0,\mu_i+\rho g_i(x^k)\}$. In the case that ρ is big and $h_i(x^k)$ or $g_i(x^k)$ is big too, the corresponding term may be inconveniently large, and so it is recommendable to redefine (8.29) as

$$B(x^k)=\nabla^2 f(x^k)+\sum_{i=1}^{m}\lambda_i^k\nabla^2 H_i(x^k)+\sum_{i=1}^{q}\mu_i^k\nabla^2\underline{G}_i(x^k), \tag{8.31}$$

where λ^k and μ^k may be given by the vectors λ_{new} and μ_{new} obtained in the previous Newton's iteration. By (4.7), (4.8), and the results related to the boundedness of the penalty parameter given in Chapter 7, $\rho H(x^k)$ and $\rho\underline{G}(x^k)$ are estimates of the Lagrange multipliers at the solution of the subproblem (4.6). On the other hand, if ρ is very big, and one replaces the matrix $B(x^k)$ given by (8.29) with the one given by (8.31) (with $\lambda^k=\lambda_{\text{new}}$ and $\mu^k=\mu_{\text{new}}$) in (8.30), the linear system (8.30) becomes a stabilized (with ρ) Newtonian linear system for the KKT condition of the original problem (4.1).

The linear system (8.30) (defined with the matrix $B(x^k)$ given by (8.29) or (8.31)) is a linear system with $n+m+q$ equations and unknowns and with no apparent ill-conditioning problems. Note that, when $\rho \to \infty$, the condition number of the matrix tends to the condition number of a matrix where ρ does not appear at all.

It is interesting to analyze the case in which, in the solution of (8.30), we have $(x - x^k, \lambda_{\text{new}}, \mu_{\text{new}}) = (0, \lambda, \mu)$. By the second and third blocks of (8.30), we have, in that case, that $h(x^k) = 0$ and $g_i(x^k) = 0$ for all i such that $g_i(x^k) \geq -\mu_i/\rho$. Therefore, x^k is a feasible point of the original problem (4.1). Moreover, the first block of (8.30) states that the KKT condition of the original problem holds. This means that the distance between the solution of (8.30) and (x^k, λ, μ) provides a sensible measure of optimality.

The system (8.30) needs a possible correction in the case in which the generated direction $d^k \equiv x - x^k$ is not a descent direction for $F(x)$ at x^k. Note that taking a sufficiently big correcting parameter c_k and replacing $B(x^k)$ with $B(x^k) + c_k I$ in (8.30), the angle between d^k and $-\nabla F(x^k)$ can be made as small as desired, so that the condition $\nabla F(x^k)^T d^k \leq -\theta \|\nabla F(x^k)\|_2 \|d^k\|_2$, required at Step 2 of Algorithm 8.1, is satisfied. Moreover, multiplying d^k by a suitable scalar, the condition $\|d^k\| \geq \beta \|\nabla F(x^k)\|$ will also be satisfied. After the computation of the (perhaps corrected) direction d^k, the next iterate is computed by means of a line search following the general Algorithm 8.1. The dominance phenomenon of the penalized term with respect to the function $f(x)$ for large values of ρ makes it desirable to employ a nonmonotone decreasing strategy.

Linear algebra

If ρ is not very large, we may try to obtain the Newton direction by directly solving (8.27). Since $\nabla^2 F(x^k)$ is symmetric, we may try to compute its Cholesky factorization. If this factorization can be completed, obtaining $\nabla^2 F(x^k) = LL^T$ with L lower-triangular and nonsingular, the direction obtained solving $LL^T d^k = -\nabla F(x^k)$ is a descent direction along which classical line searches produce sufficient descent. However, directions whose angle with $-\nabla F(x^k)$ is very close to $\pi/2$ should be replaced with more conservative directions that may be obtained by increasing the diagonal of $\nabla^2 F(x^k)$ by a tiny positive scalar. If the Cholesky factorization of the Hessian cannot be completed, several possibilities arise. Using a good estimation of the lowest eigenvalue we may add a suitable positive diagonal matrix to the Hessian by means of which a descent direction is guaranteed after solving the linear system that replaces (8.27). The direction d^k obtained in this way minimizes the quadratic approximation of $F(x)$ on the region defined by $\|d\|_2 \leq \|d^k\|_2$, a property that approximates this strategy to the well-known field of trust-region methods [83].

In the case of a large penalty parameter ρ, the system (8.30) may be solved by means of variations of a factorization proposed by Bunch and Parlett [71] for symmetric matrices. Improvements have been obtained in [23] and any of the successors of the classical Harwell subroutine MA27 (based on [103]), such as MA57, MA86, or MA97, is the rule of choice in this case. The interpretation of diagonal modifications of B_k is the same as above and also connects the strategy to the trust-region framework. Iterative linear solvers may also be used to solve (8.30).

8.5.2 ▪ Truncated Newton approach

In the truncated Newton approach, we consider the quadratic problem

$$\text{Minimize } \frac{1}{2} d^T \nabla^2 F(x^k) d + \nabla F(x^k)^T d \tag{8.32}$$

and we try to find an approximate solution d^k employing the classical conjugate gradient (CG) method of Hestenes and Stiefel [146]. To achieve this goal, we start with the iterate $d = 0$, and, at each CG iteration, we test the conditions (8.2), taking care that the final approximate solution of (8.32) satisfies them. This is not difficult since the first iterate of the CG method is in general a multiple of $-\nabla F(x^k)$, thus satisfying trivially (8.2). As far as (8.2) holds at the CG iterate, we continue until minimizing (8.32) with some required precision.

When the Hessian $\nabla^2 F(x^k)$ is not positive definite, the problem (8.32) may have no solution. In this case, CG is interrupted but the possibility of computing an adequate direction from the point of view of (8.2) is not affected. (In the worst case we may be forced to compute the approximate solution of (8.32) as a multiple of $-\nabla F(x^k)$.)

At each step of the CG method, we need to compute a Hessian-vector product. Sometimes, we do not wish to provide the true Hessian because we cannot compute it or because it is computationally expensive. In this case, each matrix-vector product may be replaced by an incremental quotient, using the approximation

$$\nabla^2 F(x^k) d \approx \frac{1}{t}\left(\nabla F(x^k + td) - \nabla F(x^k)\right)$$

with a small value of t. In this case, each CG iteration involves an additional evaluation of ∇F.

8.5.3 ▪ Quasi Newton and preconditioners

We aim to create an adequate quasi-Newton formula for approximating the Hessian matrix $\nabla^2 F(x)$ in the large-scale case. Recall that this Hessian is well defined whenever $\mu_i + \rho g_i(x) \neq 0$ for all $i = 1, \dots, p$. By direct calculations, we have that

$$\nabla^2 F(x) = \nabla^2 f(x) + A(x) + C(x),$$

where

$$A(x) = \rho\left(\sum_{i=1}^{m} \nabla h_i(x) \nabla h_i(x)^T + \sum_{i \in I(x,\mu)} \nabla g_i(x) \nabla g_i(x)^T\right),$$

$$C(x) = \sum_{i=1}^{m} [\lambda_i + \rho h_i(x)] \nabla^2 h_i(x) + \sum_{i \in I(x,\mu)} [\mu_i + \rho g_i(x)] \nabla^2 g_i(x),$$

and

$$I(x,\mu) = \{i \in \{1, \dots, p\} \mid \mu_i + \rho g_i(x) > 0\}.$$

Assume that x^c is the current iterate at Algorithm 8.1 and x^p is the previous one. We want to construct a reasonable and cheap approximation of $\nabla^2 F(x^c)$. Let us define $s = x^c - x^p$ and $y = \nabla F(x^c) - \nabla F(x^p)$. In order to obtain the Hessian approximation, the matrix $A(x^c)$ will be corrected twice. At the first correction, we add a diagonal matrix σI with the objective of guaranteeing nonsingularity. (Note that $A(x)$ is always positive semidefinite.) Following the "spectral approach" [30, 223, 224, 60], we define

$$\sigma_{\text{spec}} = \operatorname{argmin} \|(A(x) + \sigma I)s - y\|^2.$$

This implies that

$$\sigma_{\text{spec}} = \frac{(y - A(x^c)s)^T s}{s^T s}.$$

Using safeguards, we correct σ_{spec} taking

$$\sigma = \max\{\sigma_{\min}, \min\{\sigma_{\max}, \sigma_{\text{spec}}\}\},$$

where $0 < \sigma_{\min} < \sigma_{\max} < +\infty$ are given parameters, and

$$A_+ = A(x^c) + \sigma I.$$

If $s^T y \leq 10^{-8}||s||||y||$, we define $H = A_+$. Otherwise, we correct A_+ in order to satisfy the secant equation, maintaining its positive definiteness. Since A_+ is positive definite, it is natural to correct this matrix using the famous BFGS formula [98]. So,

$$H = A_+ + \frac{yy^T}{s^T y} - \frac{A_+ s s^T A_+}{s^T A_+ s}. \tag{8.33}$$

In this way, the Hessian approximation H satisfies the secant equation and remains positive definite.

The search direction will be obtained by solving a linear system, whose matrix is H, employing the CG method. A suitable preconditioner for the application of CG may be obtained as follows:

1. Define $D = \text{diag}(A(x^c))$.

2. Compute its associated spectral coefficient

$$\sigma_P = \max\left\{\sigma_{\min}, \min\left\{\sigma_{\max}, \frac{(y - Ds)^T s}{s^T s}\right\}\right\}.$$

3. Compute $D_+ = D + \sigma_P I$.

4. As in the computation of H, if $s^T y \leq 10^{-8}||s||||y||$, define $H_P = D_+$. Otherwise, define

$$H_P = D_+ + \frac{yy^T}{s^T y} - \frac{D_+ s s^T D_+}{s^T D_+ s}. \tag{8.34}$$

(Note that H_P does not need to be computed explicitly.)

In this process, H_P is the BFGS correction of a positive definite matrix, and, thus, it is positive definite too [98]. The inverse of H_P is given by

$$H_P^{-1} = D_+^{-1} + \frac{(s - D_+^{-1} y) s^T + s(s - D_+^{-1} y)^T}{s^T y} - \frac{(s - D_+^{-1} y)^T y s s^T}{(s^T y)^2}. \tag{8.35}$$

The explicit form of H_P^{-1} shows that H_P may be used as preconditioner for the CG method. Moreover, this preconditioner may be used too for the CG iterations in the truncated Newton scheme.

8.6 ▪ Review and summary

In this chapter, we addressed the unconstrained optimization problem with a focus on the case in which the objective function is the Augmented Lagrangian. We concentrated on line-search methods, although a similar chapter could be written along similar lines using the trust-region framework. For a general line-search algorithm that admits intermediate "magic" steps, we proved global and local convergence results.

8.7 ▪ Further reading

The present chapter may be read as a (biased) short course on unconstrained optimization. The classical bibliography on this subject complements this study. The Dennis–Schnabel book [98] covers the essence of Newtonian and quasi-Newton methods and the classical book by Ortega and Rheinboldt [214] is a mandatory reference for local and global properties of Newton and related methods. Trust-region methods, not addressed in this chapter, were exhaustively surveyed in [83]. Generalizations of the Hestenes–Stiefel CG method to the nonquadratic case began with the classical Fletcher–Reeves and Polak–Ribière papers [116, 216] and were followed by many variations. The main drawback of these CG generalizations is that they require rather strict line searches due to their connection with the Hestenes–Stiefel method, in which exact one-dimensional minimization is essential. Modern CG methods try to alleviate this inconvenience as much as possible. The unconstrained CG methods with the best performance according to the available literature are due to Dai and Yuan [93] and Hager and Zhang [138, 139, 140].

If one applies Newton's method to an unconstrained minimization problem and the Hessian is positive definite at every iteration, a reasonable conjecture is that every limit point is stationary. Surprisingly, this conjecture is not true (Mascarenhas [194]), showing that safeguards in terms of the angle between the Newton direction and the gradient are necessary. Related results, concerning the nonconvergence of the BFGS method, were given by Dai [90, 91] and Mascarenhas [193, 194]. For many years, it was believed that limit points generated by the BFGS method, with standard line-search procedures, should be stationary, but the counterexamples exhibited in [90, 91, 193, 194] show that this is not true.

8.8 ▪ Problems

8.1 Prove that, if the search direction is given by $d^k = -H_k \nabla F(x^k)$ and the matrix H_k is symmetric and positive definite, one has that $\nabla F(x^k)^T d^k < 0$ and, consequently, the basic unconstrained optimization algorithm is well defined.

8.2 Find an explicit formula for the global minimizer of a quadratic along a given direction, when such a formula exists.

8.3 Prove that, if the set $\{\|B_k\|, k \in \mathbb{N}\}$ is bounded, the condition $\|d^k\| \geq \beta \|\nabla F(x^k)\|$ is satisfied, as required by Algorithm 8.1. Moreover, prove that if $\eta_k = 0$, then the condition number $\|B_k\|_2 \|B_k^{-1}\|_2$ is smaller than or equal to $\frac{1}{\theta}$ and the angle condition defined in Step 2 also holds.

8.4 Consider $F(x_1, x_2) = x_1^2 + x_2$. Take $x^0 = (-\pi, 0)^T$ as the initial point. Define the search direction d^k as being $d^k = (1,0)^T$ if $x_1^k < 0$ and $d^k = (-1,0)^T$ if $x_1^k > 0$. Assume that t_k is the largest value in $\{1, 1/2, 1/4, \dots\}$ such that $x^k + t_k d^k$ satisfies the Armijo criterion (8.3) and that $x^{k+1} = x^k + t_k d^k$. Show that the sequence $\{x^k\}$ converges to $x^* = (0,0)^T$ and that $\nabla F(x^*) \neq 0$. Moreover, show that, although the search directions d^k are descent directions, for any $\theta \in (0,1)$ there exists k_0 such that the angle condition in (8.2) does not hold for every $k \geq k_0$.

8.5 Replace condition (8.3) with

$$F(x^{k+1}) \leq F(x^k + t_k d^k) \leq F(x^k) + \alpha t_k \max\{-c, \nabla F(x^k)^T d^k\},$$

where $c > 0$ is a given constant. Interpret geometrically and prove Theorems 8.1 and 8.2 with this modification.

8.6 Prove that, when the condition number of the positive definite Hessian is uniformly bounded by $c > 0$, the cosine of the angle between the negative gradient and the Newton direction is at least $1/c$.

8.7 Prove that, when the objective function is quadratic and one minimizes along the search direction, the corresponding minimizer satisfies the Armijo condition for any $\alpha \leq 1/2$. (This is a strong motivation for using $\alpha < 1/2$. Why?)

8.8 Consider the function $F(x_1, x_2) = (x_1 - 1)^2 + \frac{\rho}{2}(x_2 - x_1^2)^2$. Find $\delta(\rho)$ such that, if $\|(x_1, x_2)^T - (1,1)^T\| \leq \delta(\rho)$, the unitary step along the Newton direction provides descent. Note that $\delta(\rho) \to 0$ when $\rho \to \infty$.

8.9 Prove that the direction d^k obtained solving (8.27) minimizes the quadratic approximation of $F(x)$ on the region defined by $\|d\|_2 \leq \|d^k\|_2$, a property that approximates this strategy to the well-known field of trust-region methods (see the Conn, Gould, and Toint book [83]).

8.10 Consider a matrix B_0 that is symmetric and positive definite. Assume that B_{k+1} must satisfy the secant equation (8.24) and that the rank of $B_{k+1} - B_k$ must be 2. Discover the BFGS formula.

8.11 Consider a matrix H_0 that is symmetric and positive definite. Assume that H_{k+1} must satisfy the secant equation $H_{k+1}(\nabla F(x^{k+1}) - \nabla F(x^k)) = (x^{k+1} - x^k)$ and that the rank of $H_{k+1} - H_k$ must be 2. Discover the so-called DFP (Davidon–Fletcher–Powell) formula.

8.12 Show that H obtained by (8.33) is positive definite.

8.13 Prove that H_P^{-1} defined by (8.35) is the inverse of H_P defined by (8.34).

8.14 The spectral gradient approach comes from approximating the Hessian by a diagonal matrix with identical diagonal elements. The idea of employing more general diagonal approximations has been suggested many times. For general test functions it does not seem to be a good idea, perhaps because the spectral properties of the equally diagonal approach are lost. However, in real-life applications, the presence of separable or almost-separable objective functions is not rare. Practical optimization models should deal with noncorrelated variables as far as possible. The lack of correlation between variables is naturally linked to almost-diagonal Hessians. In this context, the diagonal updating idea may be useful. Define an optimization method based on diagonal updatings of a diagonal approximate Hessian and analyze convergence supported by the theory presented in this chapter.

8.15 Assume that $\nabla F(x^*) = 0$, $\nabla^2 F(x^*)$ is nonsingular, and $\{x^k\}$ converges superlinearly (respectively, quadratically) to x^*. Prove that $F(x^k)$ converges superlinearly (respectively, quadratically) to $F(x^*)$. Show that this result is not true if we replace "superlinearly" with "linearly." Derive consequences for numerical minimization algorithms from this property. (How reliable is to consider that $F(x^k) - F(x^*)$ measures the distance to the solution?)

8.16 Assume that $F(x^k)$ converges to $F(x^*)$ linearly, superlinearly, or quadratically. Imagine that you plot $F(x^k) - F(x^*)$ as a function of k. How does this graphic look in each case? What about the graphic of $\log(F(x^k) - F(x^*)$? Repeat this exercise with $\|\nabla F(x^k)\|$ instead of $F(x^k) - F(x^*)$.

Chapter 9

Solving Constrained Subproblems

Recall that, at each outer iteration of the Augmented Lagrangian method, we need to minimize the function $L_{\rho_k}(x, \bar{\lambda}^k, \bar{\mu}^k)$ on a lower-level set Ω. The case in which $\Omega = \mathbb{R}^n$ was studied in Chapter 8. Here, we will consider the case in which Ω is closed and convex and the more particular case in which Ω is a box. As in Chapter 8, we will denote $\rho = \rho_k$, $\lambda = \bar{\lambda}^k$, $\mu = \bar{\mu}^k$, and

$$F(x) = L_\rho(x, \lambda, \mu). \tag{9.1}$$

Continuous first derivatives of f, h, and g (and, consequently, of F) will be assumed to exist whenever necessary.

9.1 ▪ Spectral projected gradient

In this section, we consider the case in which the subproblem (4.6) takes the form

$$\text{Minimize } F(x) \text{ subject to } x \in \Omega \tag{9.2}$$

and $\Omega \subseteq \mathbb{R}^n$ is a closed and convex set, which perhaps is not described by a finite number of equalities or inequalities. However, we will consider that computing $P_\Omega(x)$, the Euclidean projection of an arbitrary $x \in \mathbb{R}^n$ onto Ω, is affordable. If in addition n is large, the spectral projected gradient method (SPG) [60, 61, 62, 63] may be the best alternative for solving subproblem (9.2).

The iterates of algorithms for solving (9.2) will be called (with some abuse of notation) x^k for $k = 0, 1, 2, \ldots$. We warn that these iterates should not be confused with the ones that define the Augmented Lagrangian iterations.

Given $x^0 \in \Omega$, $\alpha \in (0, 1/2)$, and $0 < \sigma_{\min} \ll \sigma_{\max}$, the SPG method computes

$$d^k = P_\Omega\left(x^k - \frac{1}{\sigma_k^{SPG}} \nabla F(x^k)\right) - x^k \tag{9.3}$$

and x^{k+1} such that $F(x^{k+1}) \le F(x^k + t_k d^k)$, where $\sigma_k^{SPG} \in [\sigma_{\min}, \sigma_{\max}]$ and t_k is obtained by means of a backtracking procedure that guarantees the sufficient descent condition

$$F(x^k + t_k d^k) \le F_k^{\text{ref}} + \alpha t_k \nabla F(x^k)^T d^k. \tag{9.4}$$

The reference value F_k^{ref} is generally chosen as

$$F_k^{\text{ref}} = \max\{f(x^k), f(x^{k-1}), \ldots, f(x^{\max\{0, k-M+1\}})\}$$

(Grippo, Lampariello, and Lucidi [132]) with M around 10. The SPG parameter σ_k^{SPG} is usually defined by

$$\sigma_k^{SPG} = \begin{cases} 1 & \text{if } k = 0, \\ \max\left\{\sigma_{\min}, \min\left\{\dfrac{(s^k)^T y^k}{(s^k)^T s^k}, \sigma_{\max}\right\}\right\} & \text{otherwise,} \end{cases} \tag{9.5}$$

where $s^k = x^k - x^{k-1}$ and $y^k = \nabla F(x^k) - \nabla F(x^{k-1})$. Alternatively, σ_0^{SPG} may be defined as

$$\sigma_0^{SPG} = \max\left\{\sigma_{\min}, \min\left\{\frac{\bar{s}^T \bar{y}}{\bar{s}^T \bar{s}}, \sigma_{\max}\right\}\right\}, \tag{9.6}$$

where $\bar{s} = x^0 - \bar{x}$ and $\bar{y} = \nabla F(x^0) - \nabla F(\bar{x})$, $\bar{x} = x^0 - t_{\text{small}} \nabla F(x^0)$, and t_{small} is a small positive number.

The backtracking procedure for computing t_k is as follows. It begins with the trial $t = 1$ and testing the fulfillment of (9.4) (for $t_k = t$). If (9.4) holds, the backtracking finishes. Otherwise, a new t in the interval $[0.1t, 0.5t]$ is chosen using safeguarded quadratic interpolation and the process is repeated. Since the direction d^k in (9.3) is a descent direction, this loop necessarily finishes with the fulfillment of (9.4) for a sufficiently small t_k.

The SPG method has three main ingredients: projected gradient ideas [38, 126, 173]; the choice of the steplength, motivated by Barzilai and Borwein [30] and elucidated by Raydan [223, 224] for unconstrained problems; and nonmonotone line searches [132]. Because of the simplicity of this method and its capacity to deal with huge problems, it has been used in multiple applications since its introduction by Birgin, Martínez, and Raydan [60, 61]. It can be proved (see [62]) that *every* limit point of a sequence generated by SPG satisfies the optimality condition (3.31), given by $P_\Omega(x^* - \nabla F(x^*)) = x^*$. Here, we will give a simplified and less ambitious proof, where we only show that, if the sequence $\{x^k\}$ is bounded, *at least one* limit point is stationary in the sense of (3.31). In this proof, we will follow the approach of Birgin, Martínez, and Raydan [62].

9.1.1 ▪ Convergence of SPG

Throughout this section, we will denote

$$Q_k(d) = \frac{1}{2}\sigma_k^{SPG} \|d\|_2^2 + \nabla F(x^k)^T d.$$

The global minimizer of $Q_k(d)$ subject to $x^k + d \in \Omega$ is given by (9.3).

Algorithm 9.1. SPG.
Let $\alpha \in (0,1)$, $0 < \sigma_{\min} < \sigma_{\max}$, and M be a positive integer. Let $x^0 \in \Omega$ be an arbitrary initial point. Given $x^k \in \Omega$, the steps to compute x^{k+1} are as follows:

Step 1. Compute $\sigma_k^{SPG} \in [\sigma_{\min}, \sigma_{\max}]$ and d^k as in (9.5) and (9.3), respectively. If $d^k = 0$, stop the execution of the algorithm declaring that x^k is a stationary point.

Step 2. Set $t \leftarrow 1$ and $F_k^{\text{ref}} = \max\{F(x^{k-j+1}) \mid 1 \le j \le \min\{k+1, M\}\}$.
If

$$F(x^k + t d^k) \le F_k^{\text{ref}} + t\alpha \nabla F(x^k)^T d^k, \tag{9.7}$$

set $t_k = t$, choose $x^{k+1} \in \Omega$ such that

$$F(x^{k+1}) \le F(x^k + t_k d^k), \tag{9.8}$$

and finish the iteration. Otherwise, choose $t_{\text{new}} \in [0.1t, 0.5t]$, set $t \leftarrow t_{\text{new}}$ and repeat test (9.7).

The lemma below shows that Algorithm 9.1 is well defined. In particular, it shows that the direction d^k computed as in (9.3) is a descent direction. Then, for completeness, it shows (as already shown in Theorem 8.1) that, if d^k is a descent direction, a steplength t_k that satisfies the Armijo condition can be computed in a finite number of steps.

Lemma 9.1. *Algorithm 9.1 is well defined.*

Proof. If $d^k = 0$ the algorithm stops. Otherwise, we have that $Q_k(d^k) \le Q_k(0) = 0$, i.e., $\nabla F(x^k)^T d^k \le -\frac{1}{2}\sigma_k^{SPG} \|d^k\|_2^2 < 0$, since $d^k \neq 0$ and, by (9.5), $0 < \sigma_{\min} \le \sigma_k^{SPG}$. Now,

$$\lim_{t \to 0} \frac{F(x^k + t d^k) - F(x^k)}{t} = \nabla F(x^k)^T d^k < 0.$$

Therefore,

$$\lim_{t \to 0} \frac{F(x^k + t d^k) - F(x^k)}{t \nabla F(x^k)^T d^k} = 1$$

and

$$\frac{F(x^k + t d^k) - F(x^k)}{t \nabla F(x^k)^T d^k} > \alpha$$

if t is small enough. Thus, for $t > 0$ small enough,

$$F(x^k + t d^k) < F(x^k) + t\alpha \nabla F(x^k)^T d^k \le F_k^{\text{ref}} + t\alpha \nabla F(x^k)^T d^k.$$

This completes the proof. □

The lemma below shows that, if Algorithm 9.1 does not generate an infinite sequence of iterates, it stops at a stationary point.

Lemma 9.2. *Assume that the sequence generated by Algorithm 9.1 stops at x^k. Then, x^k is stationary.*

Proof. The proof follows from the characterization given in Lemma 3.2. □

For the remaining results of this section, we assume that the algorithm does not stop. So, infinitely many iterates $\{x^k\}_{k \in \mathbb{N}}$ are generated, and, by (9.7), we have that $F(x^k) \le F(x^0)$ for all $k \in \mathbb{N}$. In order to ensure the existence of limit points, we state the following assumption.

Assumption 9.1. *The level set $\{x \in \Omega \mid F(x) \le F(x^0)\}$ is bounded.*

Assumption 9.1 will be supposed to be true all along the present section. Note that Assumption 9.1 holds if Ω is bounded.

The five lemmas below will be used to prove a final theorem that says that the sequence generated by Algorithm 9.1 has at least one limit point that is stationary.

Lemma 9.3. *Assume that $\{x^k\}_{k \in \mathbb{N}}$ is a sequence generated by Algorithm 9.1. Define, for all $j = 1, 2, 3, \dots$,*

$$V_j = \max\{F(x^{jM-M+1}), F(x^{jM-M+2}), \dots, F(x^{jM})\}$$

and $\nu(j) \in \{jM - M + 1, jM - M + 2, \dots, jM\}$ such that $F(x^{\nu(j)}) = V_j$. Then,

$$V_{j+1} \le V_j + t_{\nu(j+1)-1} \alpha \nabla F(x^{\nu(j+1)-1})^T d^{\nu(j+1)-1} \tag{9.9}$$

for all $j = 1, 2, 3, \dots$.

Proof. We will prove by induction on ℓ that, for all $\ell = 1,2,\dots,M$ and for all $j = 1,2,3,\dots,$

$$F(x^{jM+\ell}) \le V_j + t_{jM+\ell-1}\alpha \nabla F(x^{jM+\ell-1})^T d^{jM+\ell-1} < V_j. \tag{9.10}$$

By (9.7) and (9.8), we have that, for all $j \in \mathbb{N}$,

$$F(x^{jM+1}) \le V_j + t_{jM}\alpha \nabla F(x^{jM})^T d^{jM} < V_j,$$

so (9.10) holds for $\ell = 1$.

Assume, as the inductive hypothesis, that

$$F(x^{jM+\ell'}) \le V_j + t_{jM+\ell'-1}\alpha \nabla F(x^{jM+\ell'-1})^T d^{jM+\ell'-1} < V_j \tag{9.11}$$

for $\ell' = 1,\dots,\ell$. Now, by (9.7) and (9.8) and the definition of V_j, we have that

$$\begin{aligned} F(x^{jM+\ell+1}) &\le \max_{1\le r\le M}\{F(x^{jM+\ell+1-r}\} + t_{jM+\ell}\alpha \nabla F(x^{jM+\ell})^T d^{jM+\ell} \\ &= \max\{F(x^{(j-1)M+\ell+1}),\dots,F(x^{jM+\ell})\} + t_{jM+\ell}\alpha \nabla F(x^{jM+\ell})^T d^{jM+\ell} \\ &\le \max\{V_j, F(x^{jM+1}),\dots,F(x^{jM+\ell})\} + t_{jM+\ell}\alpha \nabla F(x^{jM+\ell})^T d^{jM+\ell}. \end{aligned}$$

But, by the inductive hypothesis,

$$\max\{F(x^{jM+1}),\dots,F(x^{jM+\ell})\} < V_j,$$

so

$$F(x^{jM+\ell+1}) \le V_j + t_{jM+\ell}\alpha \nabla F(x^{jM+\ell})^T d^{jM+\ell} < V_j.$$

Therefore, the inductive proof is complete, and hence (9.10) is proved. Since $\nu(j+1) = jM+\ell$ for some $\ell \in \{1,\dots,M\}$, this implies the desired result. □

From now on, we define

$$K = \{\nu(1)-1, \nu(2)-1, \nu(3)-1, \dots\}, \tag{9.12}$$

where $\{\nu(j)\}$ is the sequence of indices defined in Lemma 9.3. Note that, when $M = 1$, we have that $K = \{0,1,2,\dots\}$. Clearly,

$$\nu(j) < \nu(j+1) \le \nu(j) + 2M - 1$$

for all $j = 1,2,3,\dots$.

Lemma 9.4. *Let K be given by (9.12). Then,* $\lim_{k\in K} t_k Q_k(d^k) = 0$.

Proof. By (9.9), since F is continuous and bounded below,

$$\lim_{k\in K} t_k \nabla F(x^k)^T d^k = 0. \tag{9.13}$$

But

$$0 > Q_k(d^k) = \frac{1}{2}\sigma_k^{SPG}\|d^k\|_2^2 + \nabla F(x^k)^T d^k \ge \nabla F(x^k)^T d^k \text{ for all } k \in \mathbb{N}.$$

Hence, by (9.13), the desired result is proved. □

Lemma 9.5. *Assume that $K_1 \underset{\infty}{\subset} \mathbb{N}$ is a sequence of indices such that*

$$\lim_{k\in K_1} x^k = x^* \in \Omega \text{ and } \lim_{k\in K_1} Q_k(d^k) = 0.$$

Then, x^* is stationary.

Proof. Let $K_2 \underset{\infty}{\subset} K_1$ be such that

$$\lim_{k\in K_2} \sigma_k^{SPG} = \sigma > 0.$$

We define

$$Q(d) = \frac{1}{2}\sigma \|d\|_2^2 + \nabla F(x^*)^T d \text{ for all } d \in \mathbb{R}^n.$$

Suppose that there exists $\hat{d} \in \mathbb{R}^n$ such that $x^* + \hat{d} \in \Omega$ and

$$Q(\hat{d}) < 0. \tag{9.14}$$

Define

$$\hat{d}^k = x^* + \hat{d} - x^k \text{ for all } k \in K_2.$$

Clearly, $x^k + \hat{d}^k \in \Omega$ for all $k \in K_2$. By continuity, since $\lim_{k\in K_2} x^k = x^*$, we have that

$$\lim_{k\in K_2} Q_k(\hat{d}^k) = Q(\hat{d}) < 0. \tag{9.15}$$

But, by the definition of d^k, we have that $Q_k(d^k) \le Q_k(\hat{d}^k)$. Therefore, by (9.15), $Q_k(d^k) \le Q(\hat{d})/2 < 0$ for $k \in K_2$ large enough. This contradicts the fact that $\lim_{k\in K_2} Q_k(d^k) = 0$. The contradiction came from the assumption that $\hat{d}$ with the property (9.14) exists. Therefore, $Q(d) \ge 0$ for all $d \in \mathbb{R}^n$ such that $x^* + d \in \Omega$. Thus, $\nabla F(x^*)^T d \ge 0$ for all $d \in \mathbb{R}^n$ such that $x^* + d \in \Omega$. So, x^* is stationary. □

Lemma 9.6. *$\{d^k\}_{k\in\mathbb{N}}$ is bounded.*

Proof. For all $k \in \mathbb{N}$,

$$Q_k(d^k) = \frac{1}{2}\sigma_k^{SPG}\|d^k\|_2^2 + \nabla F(x^k)^T d^k < 0.$$

Therefore,

$$\|d^k\|_2^2 < -\frac{2}{\sigma_k^{SPG}} \nabla F(x^k)^T d^k \le \frac{2}{\sigma_{\min}} \|\nabla F(x^k)\|_2 \|d^k\|_2.$$

Thus,

$$\|d^k\|_2 \le \frac{2}{\sigma_{\min}} \|\nabla F(x^k)\|_2.$$

Since $\{x^k\}_{k\in\mathbb{N}}$ is bounded and F has continuous derivatives, $\{\nabla F(x^k)\}_{k\in\mathbb{N}}$ is bounded. Therefore, the set $\{d^k\}_{k\in\mathbb{N}}$ is bounded. □

Lemma 9.7. *Assume that $K_3 \underset{\infty}{\subset} \mathbb{N}$ is a sequence of indices such that*

$$\lim_{k\in K_3} x^k = x^* \in \Omega \text{ and } \lim_{k\in K_3} t_k = 0.$$

Then,

$$\lim_{k\in K_3} Q_k(d^k) = 0, \tag{9.16}$$

and hence x^* is stationary.

Proof. Suppose that (9.16) is not true. Then, for some infinite set of indices $K_4 \underset{\infty}{\subset} K_3$, $Q_k(d^k)$ is bounded away from zero.

On the other hand, since $t_k \to 0$, by the definition of Algorithm 9.1, for $k \in K_4$ large enough, there exists $t'_k \geq t_k$ such that $\lim_{k\in K_4} t'_k = 0$, and (9.7) does not hold for $t = t'_k$. So,

$$F(x^k + t'_k d^k) > \max\{F(x^{k-j+1}) \mid 1 \leq j \leq \min\{k+1, M\}\} + t'_k \alpha \nabla F(x^k)^T d^k,$$

and hence

$$F(x^k + t'_k d^k) > F(x^k) + \alpha t'_k \nabla F(x^k)^T d^k$$

for all $k \in K_4$. Therefore,

$$\frac{F(x^k + t'_k d^k) - F(x^k)}{t'_k} > \alpha \nabla F(x^k)^T d^k$$

for all $k \in K_4$. By the mean-value theorem, there exists $\xi_k \in [0,1]$ such that

$$\nabla F(x^k + \xi_k t'_k d^k)^T d^k > \alpha \nabla F(x^k)^T d_k \tag{9.17}$$

for all $k \in K_4$. Since, by Lemma 9.6, the set $\{d^k\}_{k\in K_4}$ is bounded, there exists a sequence of indices $K_5 \underset{\infty}{\subset} K_4$ such that $\lim_{k\in K_5} d^k = d$ and $\lim_{k\in K_5} \sigma_k^{SPG} = \sigma$ for some $d \in \mathbb{R}^n$ and some $\sigma > 0$. Taking limits for $k \in K_5$ in both sides of (9.17), we obtain $\nabla F(x^*)^T d \geq \alpha \nabla F(x^*)^T d$. This implies that $\nabla F(x^*)^T d \geq 0$. So,

$$\frac{1}{2}\sigma \|d\|_2^2 + \nabla F(x^*)^T d \geq 0.$$

Therefore, since $Q_k(d^k) < 0$ for all k,

$$\lim_{k\in K_5} \frac{1}{2}\sigma_k^{SPG}\|d^k\|_2^2 + \nabla F(x^k)^T d^k = 0.$$

Thus, $\lim_{k\in K_5} Q_k(d^k) = 0$. This contradicts the assumption that $Q_k(d^k)$ is bounded away from zero for $k \in K_4$. Therefore, (9.16) is true. Thus, the hypothesis of Lemma 9.5 holds, with K_3 replacing K_1, and, therefore, by Lemma 9.5, x^* is stationary. □

Theorem 9.1. *Every limit point of $\{x^k\}_{k\in K}$ is stationary.*

Proof. Let $K_6 \underset{\infty}{\subset} K$ be such that $\lim_{k\in K_6} x^k = x^*$. By Lemma 9.4, we have that

$$\lim_{k\in K_6} t_k Q_k(d^k) = 0.$$

Now we have two possibilities: (a) $\lim_{k\in K_6} Q_k(d^k) = 0$ or (b) there exists $K_7 \underset{\infty}{\subset} K_6$ such that $Q_k(d^k) \le c < 0$ for all $k \in K_7$. In case (a), by Lemma 9.5, x^* is stationary. In case (b), $t_k \to 0$ for $k \in K_7$. By Lemma 9.7, this implies that x^* is stationary. □

Theorem 9.1 shows that every limit point of $\{x^k\}_{k\in K}$ is stationary. If $M = 1$, since in this case we have that $K = \mathbb{N}$, this means that every limit point is stationary. Even when $M > 1$, it is also true that every limit point is stationary. See [62] for details.

9.1.2 ▪ SPG and magic steps

As discussed in the case of unconstrained subproblems, the requirement (9.8) allows one to employ a big variety of heuristic procedures, accelerations, and "magic" ideas that may improve the behavior of SPG. In many cases, the efficiency of an algorithm depends on such heuristics, although its global convergence is guaranteed by an underlying algorithm like SPG. Assuming that one believes in a standard heuristic procedure, we may formalize its interlacing with SPG in the following way.

Algorithm 9.2. SPG with magic steps.
Let $\theta_{\text{progress}} \in [0,1)$ and the parameters $\alpha \in (0,1)$ and $0 < \sigma_{\min} < \sigma_{\max}$ that are necessary to execute SPG. Let $x^0 \in \Omega$ be an arbitrary initial point. Set $\text{ItType}_0 = \text{MAGIC}$. Given $x^k \in \Omega$ and the iteration type ItType_k, in order to compute x^{k+1}, proceed as follows:

Step 1. If $\text{ItType}_k = \text{MAGIC}$, execute Steps 1.1–1.3 below. Otherwise, go to Step 2.

Step 1.1. Compute $y \in \Omega$ by means of a heuristic procedure.

Step 1.2. If $F(y) > F(x^k)$, discard y and go to Step 2.

Step 1.3. Set $x^{k+1} = y$. If

$$\|P_\Omega(x^{k+1} - \nabla F(x^{k+1})) - x^{k+1}\| \le \theta_{\text{progress}} \|P_\Omega(x^k - \nabla F(x^k)) - x^k\|, \tag{9.18}$$

set $\text{ItType}_{k+1} = \text{MAGIC}$. Otherwise, set $\text{ItType}_{k+1} = \text{SPG}$. In any case, finish the kth iteration.

Step 2. Compute x^{k+1} using SPG (with $M = 1$) and set $\text{ItType}_{k+1} = \text{MAGIC}$.

If in employing Algorithm 9.2, infinitely many SPG iterations are performed, the SPG convergence theory guarantees that every limit point of the corresponding subsequence is stationary. The other possibility is that, for all k large enough, the inequality (9.18) holds. In this case, by the continuity of the projection, every limit point is stationary. Algorithm 9.2 also defines a watchdog strategy in the sense of [79] and can be employed as an alternative to the Grippo, Lampariello, and Lucidi [132] nonmonotone strategy in the implementation of SPG. For example, SPG may be executed using $M = 1$ but establishing that the heuristic y is also obtained by means of some SPG iterations without requiring descent.

9.2 ▪ Active set methods

In this section, we consider the case in which the lower-set Ω is a box, given by

$$\Omega = \{x \in \mathbb{R}^n \mid \ell \le x \le u\}.$$

The vectors ℓ and u in $\mathbb{R}^n$, with $\ell < u$, are the lower and upper bounds of Ω, respectively.

It will be useful to consider that Ω is the union of disjoint *faces*. Given a set of indices $I \subseteq \{1,\ldots,2n\}$, we denote by $\mathscr{F}_I$ the set of points $x \in \Omega$ such that

1. $x_i = \ell_i$ if $i \in I$,
2. $x_i = u_i$ if $n+i \in I$,
3. $\ell_i < x_i < u_i$ if $i \notin I$ and $n+i \notin I$.

For example, $\mathscr{F}_\emptyset$ is the topological interior of Ω, $\mathscr{F}_{\{1,\ldots,n\}}$ is the set whose unique element is $(\ell_1,\ldots,\ell_n)$, and so on. Clearly, $\cup \mathscr{F}_I = \Omega$ and, if $I \neq J$, we have that $\mathscr{F}_I \cap \mathscr{F}_J = \emptyset$. A face is also characterized by its *free* variables and *fixed* variables. The free variables at the face $\mathscr{F}_I$ are those variables x_i such that $i \notin I$ and $n+i \notin I$. The variables x_i such that $i \in I$ are said to be *fixed at the lower bound* and the variables x_i such that $n+i \in I$ are said to be *fixed at the upper bound*. Given $\bar{x} \in \Omega$, we also denote by $\mathscr{F}_{I(\bar{x})}$ the face to which $\bar{x}$ belongs, i.e., $I(\bar{x}) = \{i \mid \bar{x}_i = \ell_i\} \cup \{n+i \mid \bar{x}_i = u_i\}$. We denote by $\bar{\mathscr{F}}_I$ the closure of $\mathscr{F}_I$. For example, $\bar{\mathscr{F}}_\emptyset = \Omega$. Since boxes are simple sets, reasonable algorithms for minimizing onto boxes generate points x^k that belong to the box for all k. Therefore, each iterate x^k belongs to one (and only one) face of Ω. Many algorithms are based on the *principles of active constraints* that we state below.

9.2.1 ▪ Strong principle of active constraints

Assume that our goal is to find a global minimizer of the continuous function $F(x)$ subject to $x \in \Omega$. Let $x^0 \in \Omega$ and assume that the sequence $\{x^k\}$, generated starting from x^0, obeys the following axioms:

A1. If x^k is a global minimizer of F on Ω, the sequence *stops* at x^k. Otherwise, the point x^{k+1} satisfies $F(x^{k+1}) < F(x^k)$.

A2. If x^k is not a global minimizer of F on $\mathscr{F}_{I(x^k)}$, the point x^{k+1} is a global minimizer of F on $\mathscr{F}_{I(x^k)}$ or belongs to the boundary of $\mathscr{F}_{I(x^k)}$, i.e., belongs to $\bar{\mathscr{F}}_{I(x^k)} \setminus \mathscr{F}_{I(x^k)}$.

We say that an algorithm whose iterates x^k obey the axioms A1 and A2 follows the *strong principle of active constraints*. The philosophy behind this principle is that a face must be explored until a global minimizer on that face is found or until a point on its boundary is reached (and, in consequence, x^{k+1} belongs to a "boundary" face).

The iterations of an algorithm of this type may be of three types:

Internal iterations: These are the iterations in which x^k is not a global minimizer of F on $\mathscr{F}_{I(x^k)}$ and $x^{k+1} \in \mathscr{F}_{I(x^k)}$ is a global minimizer of F on $\mathscr{F}_{I(x^k)}$.

Boundary iterations: These are the iterations in which x^k is not a global minimizer of F on $\mathscr{F}_{I(x^k)}$ and x^{k+1} belongs to the boundary of $\mathscr{F}_{I(x^k)}$.

Leaving-face iterations: These are the iterations in which x^k *is* a global minimizer of F on $\mathscr{F}_{I(x^k)}$ (and hence on its closure) but, because it is not a global minimizer on Ω, we have that $F(x^{k+1}) < F(x^k)$ and $x^{k+1} \notin \bar{\mathscr{F}}_{I(x^k)}$.

In the following theorem, we prove that, if an algorithm is able to obey the strong principle of active constraints, it finds a global minimizer in a finite number of steps.

Theorem 9.2. *Assume that the sequence $\{x^k\}$ was generated by an algorithm that obeys the strong principle of active constraints. Then, the sequence is finite and there exists k such that x^k is a global minimizer of F onto Ω.*

Proof. Assume that the sequence $\{x^k\}$ has infinitely many elements. Since the number of faces is finite and $F(x^{k+1}) < F(x^k)$ whenever x^k is not a global solution of the problem, the number of internal iterations and the number of leaving-face iterations are finite. This means that there exists k_0 such that, for all $k \geq k_0$, all the iterations are of boundary type. But, at each boundary iteration, the number of free variables strictly decreases. Therefore, there cannot be infinitely many boundary iterations either. Thus, the sequence necessarily stops at some x^k, which must be a global minimizer of F onto Ω. □

9.2.2 ▪ Practical principle of active constraints

The strong principle of active constraints indicates an algorithmic direction but does not generate affordable methods, at least for problems with a large number of variables. The reason is that we are almost never able to meet, in finite time, a global minimizer of F on the face to which x^k belongs.

Nevertheless, the principle of staying in a face while good progress is obtained at every internal iteration is valid. The sensible way to stay in the face to which the iterate x^k belongs comes from evoking the unconstrained problem whose variables are the free variables at the face and to apply an unconstrained minimization iteration starting from x^k. The unconstrained algorithm could converge to a stationary point with respect to the free variables (a point where the derivatives with respect to the free variables vanish) or could hit the boundary of the face, stopping at a face of lower dimension. On the other hand, at each iteration, it will be necessary to decide whether it is worthwhile to continue in the same face (perhaps hitting the boundary) or if it is convenient to abandon it, in order to explore a face with additional free variables. We now give a reasonable criterion for such a decision.

Decision based on the continuous projected gradient

The decision of persisting on a face or changing it may be taken using the components of the *continuous projected gradient*.

Given $x^k \in \Omega$, consider the problem

$$\text{Minimize } \nabla F(x^k)^T(x - x^k) + \frac{1}{2}\|x - x^k\|_2^2 \text{ subject to } x \in \Omega. \qquad (9.19)$$

It is easy to see that this problem is equivalent to

$$\text{Minimize } \|x^k - \nabla F(x^k) - x\|_2^2 \text{ subject to } x \in \Omega. \qquad (9.20)$$

The solution $\bar{x}$ of (9.20) is, by definition, the projection of $x^k - \nabla F(x^k)$ onto Ω. By direct inspection, we see that this solution is given by

$$\bar{x}_i = \max\left\{\ell_i, \min\left\{x_i^k - \frac{\partial F}{\partial x_i}(x^k), u_i\right\}\right\}, i = 1, \dots, n.$$

We denote $\bar{x} = P_\Omega(x^k - \nabla F(x^k))$. It is easy to see that $\bar{x}$ is the unique stationary point of (9.19) and that $\bar{x}$ depends continuously on x^k.

On the other hand, if x^k were a KKT point of the problem of minimizing $F(x)$ onto Ω, it would also be a KKT point of (9.19), and vice versa. This fact would be equivalent to saying that $\|\bar{x}-x^k\|=0$. These considerations lead us to define the continuous projected gradient $g_P(x^k)$ as

$$g_P(x^k)=P_\Omega(x^k-\nabla F(x^k))-x^k$$

and to define the degree of stationarity of x^k as the norm of this vector.

Moreover, $g_P(x^k)$ may be decomposed in a unique way as

$$g_P(x^k)=g_I(x^k)+\left[g_P(x^k)-g_I(x^k)\right],$$

where $g_I(x^k)$ belongs to the subspace associated with $\mathscr{F}_{I(x^k)}$ and $g_P(x^k)-g_I(x^k)$ belongs to its orthogonal complement. In other words, $g_I(x^k)$ is identical to $g_P(x^k)$ for the free variables while the remaining coordinates are null. Of course $g_I(x^k)$, which will be called the *internal gradient*, also depends continuously on x^k.

If $g_I(x^k)=0$, we have that x^k is a stationary point of $F(x)$ restricted to $x\in\mathscr{F}_{I(x^k)}$. In this case, nothing else can be expected from an unconstrained algorithm that uses gradients, and so the sensible recommendation is to abandon the face (unless, of course, $g_P(x^k)=0$, in which case the problem of minimizing F on the box should be considered solved). The same recommendation should be made if the norm of $g_I(x^k)$ is small with respect to the norm of $g_P(x^k)$. Summing up, in a practical algorithm, the face $\mathscr{F}_I(x^k)$ should be abandoned when

$$\|g_I(x^k)\|\le\eta\|g_P(x^k)\|,$$

where $\eta\in(0,1)$ is an algorithmic parameter.

9.2.3 ▪ Practical scheme for the active set strategy

In [8] and [51, 52, 19], box-constrained minimization problems are solved by using unconstrained methods within the faces and abandoning the faces, when necessary, using monotone SPG iterations.

Monotone SPG iterations

As described in Section 9.1, an SPG iteration, which requires constants $\alpha\in(0,1/2)$ and $0<\sigma_{\min}\ll\sigma_{\max}$ defined independently of k, computes the (scaled) projected gradient direction d^k and the SPG parameter σ_k^{SPG} given by (9.3) and (9.5), respectively. Given the direction d^k, a monotone SPG iteration requires a step t_k satisfying the Armijo criterion

$$F(x^k+t_kd^k)\le F(x^k)+\alpha t_k\nabla F(x^k)^Td^k. \tag{9.21}$$

Note that it corresponds to satisfying (9.4) with $F_k^{\text{ref}}\equiv F(x^k)$, i.e., with $M=1$. As already described, the backtracking procedure for computing t_k begins with $t=1$ and computes a new $t\in[0.1t,0.5t]$ while (9.21) is not satisfied. When a value of t that satisfies (9.21) is found, we define $t_k=t$ and x^{k+1} such that $F(x^{k+1})\le F(x^k+t_kd^k)$.

Lemma 9.8. *The monotone SPG iteration is well defined.*

Proof. See Lemma 9.1. □

Lemma 9.9. *Assume that there exists an infinite sequence of indices $K=\{k_1,k_2,\dots\}$ such that, for all $k\in K$, x^{k+1} is obtained by means of a monotone SPG iteration. Then, every limit point of the sequence $\{x^k\}_{k\in K}$ is stationary for the box-constrained optimization problem.*

Proof. This lemma is a particular case of Theorem 9.1 with $M=1$. (Note that iterates x^{k_j} with $j>1$ in the sequence $\{x^k\}_{k\in K}$ can be seen as the result of some magic steps computed after the iterate $x^{k_{j-1}+1}$ that was computed by means of a monotone SPG iteration.) □

Internal iterations

Following the active set strategy, we minimize functions on a box by combining iterations that do not modify the fixed variables of the current iterate (perhaps hitting the boundary of the current face) with monotone SPG iterations that abandon the current face when the corresponding test indicates this decision.

The internal iterations only modify the free variables, in such a way that they can be considered as unconstrained iterations with respect to those variables. The objective function remains to be F but modifying the fixed variables is impossible. Infinitely many consecutive iterations of such an algorithm should lead to a point at which the internal gradient vanishes. This property is formalized in the following assumption.

Assumption 9.2. *If $x^k, x^{k+1}, \dots \in \mathcal{F}_I$ is a sequence of infinitely many iterations obtained by an unconstrained internal algorithm, then*

$$\lim_{k\to\infty} \frac{\partial F}{\partial x_j}(x^k) = 0$$

for all the free variables x_j. (This implies that $g_I(x^{k+\ell}) \to 0$ when $\ell \to \infty$.)

The results of Chapter 8 indicate that Assumption 9.2 is easy to verify if we use a reasonable unconstrained algorithm within the faces. There is subtlety in this statement due to the existence of bounds for the free variables. In fact, the unconstrained algorithm may indicate as a new iterate a point that does not fulfill the constraints for the free variables. In this case, one should be able to reject that iterate returning to the interior of the face or be able to hit the boundary providing a decrease of the objective function.

Convergence of the active set strategy

The following algorithm condenses the main characteristics of a suitable active set strategy for the box-constrained minimization problems that must be solved in the Augmented Lagrangian context.

Algorithm 9.3. Active set strategy.

Let $\eta \in (0,1)$ be the algorithmic parameter that provides the criterion for abandoning the faces. (Typically one takes $\eta = 0.1$ in practical calculations.) Let $x^0 \in \Omega$ be the initial point. If $x^k \in \Omega$ is a typical iterate, the steps for obtaining $x^{k+1} \in \Omega$ or interrupting the execution of the algorithm are the following.

Step 1. If $g_P(x^k) = 0$, stop. (The point x^k is stationary for minimizing F on Ω.)

Step 2. If $\|g_I(x^k)\| \le \eta \|g_P(x^k)\|$, obtain x^{k+1} using a monotone SPG iteration.

Step 3. If $\|g_I(x^k)\| > \eta \|g_P(x^k)\|$, obtain $x^{k+1} \in \mathcal{F}_{I(x^k)}$ such that $F(x^{k+1}) < F(x^k)$ using an unconstrained internal algorithm or obtain x^{k+1} in the boundary of $\mathcal{F}_{I(x^k)}$ such that $F(x^{k+1}) < F(x^k)$.

Theorem 9.3. *Let $\{x^k\}$ be generated by Algorithm 9.3. Then, this sequence stops at a stationary point x^k or admits a stationary limit point.*

Proof. Assume that the sequence does not stop and thus generates infinitely many iterations. If infinitely many iterations are of SPG type, the thesis follows from Lemma 9.9.

Assume that only a finite number of iterations are of type SPG. Therefore, for k large enough, x^{k+1} belongs to the same face as x^k or belongs to a face of lower dimension.

Then, there exist k_0 and a face $\mathcal{F}_{I(x^{k_0})}$ such that for all $k \geq k_0$ all the iterates belong to $\mathcal{F}_{I(x^{k_0})}$. By Assumption 9.2, this implies that $g_I(x^k) \to 0$. But, by Step 2, we have that $\|g_I(x^k)\| > \eta \|g_P(x^k)\|$ for all $k \geq k_0$. Therefore, $g_P(x^k) \to 0$. By the continuity of g_P, this implies that $g_P(x)$ vanishes at every limit point. □

Theorem 9.3 has the merit of showing that, ultimately, the active set algorithm finds stationary points. However, it is rather disappointing that the gradient SPG iterations are of overwhelming importance in the proof. It would be better if we were able to show that only a finite number of SPG iterations would be needed at each execution of the algorithm, so that convergence should rest on the properties of the internal algorithm, which, as we showed in Chapter 8, may be quite strong. In Theorem 9.4 below, we will give a sufficient condition to ensure that, from some iteration on, all the iterations are internal.

We say that a stationary point is *dual-degenerate* if there exists $i \in \{1, \dots, n\}$ such that $i \in I(x^*)$ or $n + i \in I(x^*)$, but

$$\frac{\partial F}{\partial x_i}(x^*) = 0.$$

A *dual-nondegenerate* point is a point that is not dual-degenerate. In other words, at dual-nondegenerate points, only derivatives with respect to free variables can vanish.

Theorem 9.4. *Assume that all the stationary points of the box-constrained problem are dual-nondegenerate. Then, the number of SPG iterations in Algorithm 9.3 is finite.*

Proof. Let K be the set of indices such that x^{k+1} is obtained by means of monotone SPG iterations for all $k \in K$. Suppose that K has infinitely many terms. By Lemma 9.9, there exists $K_1 \underset{\infty}{\subset} K$ such that

$$\lim_{k \in K_1} x^k = x^*$$

and x^* is stationary. Without loss of generality, let us assume, by contradiction, that there exist $K_2 \underset{\infty}{\subset} K_1$ and $i \in \{1, \dots, n\}$ such that $x_i^k = \ell_i$ and $x_i^{k+1} > \ell_i$ for all $k \in K_2$. Hence, we have that $x_i^* = \ell_i$. However, since x^* is dual-nondegenerate,

$$\frac{\partial F}{\partial x_i}(x^*) > 0.$$

By the continuity of ∇F, we deduce that

$$\frac{\partial F}{\partial x_i}(x^k) > 0$$

for all $k \in K_2$ large enough. Therefore, for those indices k,

$$[g_P(x^k)]_i = 0.$$

This implies that the constraint $x_i = \ell_i$ could not be abandoned at iteration k. □

9.2.4 ▪ Active set stabilized Newton

Consider again the case in which $\Omega = \{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}$. The active set philosophy described in the previous section allows one to use any unconstrained minimization method for minimizing $L_\rho(x,\lambda,\mu)$ within a particular face. Newton's method is one of these possibilities. Writing, as before, $F(x) = L_\rho(x,\lambda,\mu)$, and, in addition,

$$H(x) = h(x) + \lambda/\rho \text{ and } G(x) = g(x) + \mu/\rho,$$

we have that $F(x)$ has the form (8.26). Moreover, the function that we need to minimize within the current face also has the form (8.26) with different values for the number of (free) variables. If we consider, with some abuse of notation, that $F : \mathbb{R}^n \to \mathbb{R}$, $H : \mathbb{R}^n \to \mathbb{R}^m$, and $G : \mathbb{R}^n \to \mathbb{R}^p$ (even in the case in which x is constrained to some face and hence the number of variables is less than n), then the stabilizing techniques explained in Section 8.5.1 may be applied.

9.3 ▪ Interior stabilized Newton

Unfortunately, there is no unanimous opinion with respect to the best strategy for solving box-constrained subproblems. Interior-point strategies are the main competitors of active set ones. Assume, as before, that we wish to minimize $F(x)$ subject to $\ell \leq x \leq u$, where F is given by (8.26). The interior-point (or barrier) idea consists of considering the barrier function

$$F_\nu(x) = F(x) - \nu\left(\sum_{i=1}^{n} \log(x_i - \ell_i) + \sum_{i=1}^{n} \log(u_i - x_i)\right) \tag{9.22}$$

for a small $\nu > 0$ that tends to zero. The application of Newton's method to the minimization of $F_\nu(x)$ should take into account possible instabilities due not only to big values of ρ but also to small values of ν. As a consequence, the basic Newtonian linear systems give rise to stabilized decoupled systems as in (8.30).

9.4 ▪ Review and summary

The reason the Augmented Lagrangian method is useful for solving large-scale optimization problems is that well-established large-scale methods for solving the subproblems are available. The most usual case is when the nonrelaxable constraints define a box. In this chapter, we described, in a self-contained way, an active set framework based on unconstrained methods and projected gradients for solving the Augmented Lagrangian subproblems. Chapters 8 and 9 may be used as an independent (biased) short course on unconstrained, bound-constrained, and convex-constrained optimization.

9.5 ▪ Further reading

In 1988, Barzilai and Borwein published a new gradient algorithm (the BB method) for minimizing convex quadratics. Given $q : \mathbb{R}^n \to \mathbb{R}$ defined by $q(x) = \frac{1}{2}x^T Ax + b^T x$, where A is symmetric and positive definite, the BB method computes

$$x^{k+1} = x^k - \alpha_k \nabla q(x^k), \tag{9.23}$$

where $\alpha_0 > 0$ is arbitrary and, for all $k = 0, 1, 2, \ldots,$

$$\alpha_{k+1} = \frac{\nabla q(x^k)^T \nabla q(x^k)}{\nabla q(x^k)^T A \nabla q(x^k)}. \tag{9.24}$$

Formula (9.24) used in the BB method for defining the step at iteration $k+1$ was used in the classical Cauchy steepest descent method for defining the step at iteration k [78] and could also be derived from the scaling strategy of Oren [213]. Raydan [223] proved the convergence of the BB method for general strictly convex quadratic functions. The possibility of obtaining superlinear convergence for arbitrary n was discarded by Fletcher [114]. However, the conditions were given for the implementation of the BB method for general unconstrained minimization with the help of a nonmonotone procedure. Raydan [224] defined this method in 1997 using the Grippo–Lampariello–Lucidi strategy [132]. He proved global convergence and exhibited numerical experiments that showed that the method was more efficient than classical CG methods for minimizing general functions. These nice comparative numerical results were possible because although the CG method of Hestenes and Stiefel continued to be the rule of choice for solving many convex quadratic problems, its efficiency was hardly inherited by generalizations for minimizing general functions. Therefore, a wide space existed for variations of the Barzilai–Borwein idea [64]. The SPG method of Birgin, Martínez, and Raydan [60, 61, 62] combines Barzilai–Borwein (spectral) nonmonotone ideas with classical projected gradient strategies [40, 126, 173]. SPG is applicable to convex-constrained problems in which the projection onto the feasible set is easy to compute. Since its appearance, the method has been intensively used in applications, including optics [7, 26, 45, 46, 80, 89, 207, 208, 221, 251, 252, 253], support vector machines [86, 92, 237], optimal control [47], topology optimization [245], compressive sensing [36, 37, 110, 178], geophysics [31, 41, 87, 95, 259], image restoration [35, 68, 135], and atmospheric sciences [155, 206]. Moreover, it has been the object of several spectral-parameter modifications, alternative nonmonotone strategies have been suggested, convergence and stability properties have been elucidated, and it has been combined with other algorithms for different optimization problems.

Fletcher [115] introduced a limited memory steepest descent method that generalizes the spectral gradient method by using a few additional vectors of storage and obviously can be extended to convex-constrained minimization. On the CG box-constrained side, the CG active set method of Hager and Zhang [139] seems to be the best known alternative. Spectral residual methods that extend spectral gradient ideas to nonlinear systems of equations were introduced by La Cruz, Martínez, and Raydan [170, 171].

9.6 ▪ Problems

9.1 In the discussion of SPG, we observed that a theorem exists that says every limit point satisfies the optimality condition. For simplicity, we included a shorter theorem that guarantees that if limit points exist, at least one of them satisfies the optimality condition. Do you think that the results are equally relevant from the practical point of view? Give arguments supporting positive and negative answers to this question.

9.2 Define variations of the algorithms presented in this chapter in which the initial step size at the current iteration depends on the successful step size at the previous one. Furthermore, try to exploit "regularities" observed at different iterations regarding step size acceptance.

9.3 Consider the following alternative to Algorithm 9.3. At each ordinary iteration, we define as free variables those variables that verify $\ell_i < x_i^k < u_i$ (as always) plus those that verify $x_i = \ell_i$ with $\partial F/\partial x_i < 0$ and those that verify $x_i = u_i$ with $\partial F/\partial x_i > 0$. We compute a descent direction d^k for F in the face defined by these free variables. For all i such that $x_i^k = \ell_i$ with $d_i^k < 0$ and for all i such that $x_i^k = u_i$ with $d_i^k > 0$, we redefine $d_i^k = 0$. Prove that d^k is a descent direction. We redefine d^k again, taking the maximal step such that $x^k + t d^k$ is feasible. Prove that $t > 0$. If the descent direction computed so far is smaller than a fixed small multiple of the projected gradient, we discard this direction and we proceed to a projected gradient iteration. Otherwise, we execute a sufficient descent line search. Prove convergence, suggest different choices for d^k, and write a code.

9.4 In the case of minimization with box constraints, consider the stopping criterion in which one finishes the execution when no progress is obtained along coordinate variations with small relative tolerances. Analyze this stopping criterion in connection with subproblems of Augmented Lagrangian methods.

9.5 Generalize the active set strategies to the case of general linear (equality and inequality) constraints. Moreover, extend the generalization to general nonlinear constraints, pointing out the difficulties of implementation.

9.6 Try different strategies for the choice of steplength in projected gradient methods, including random choices. Compare with the one-dimensional exact minimization strategy in the case of convex quadratics. Develop an “artificial intelligence” strategy that chooses the steplength according to the performance of different strategies at previous iterations.

9.7 Analyze the strategy that consists of leaving the current face following gradient components orthogonal to the current face (called “chopped gradient” by Friedlander and Martínez [121] and “proportional gradient” by Dostál [101]). Show that, in the case of a convex quadratic objective function, it is possible to define a leaving criterion that guarantees returning to the current face a bounded number of times, which leads to complexity conclusions.

9.8 Write subroutines implementing the active set stabilized Newton method and the interior stabilized Newton method and compare.

9.9 Box-constrained optimization problems can be reformulated as unconstrained optimization problems by means of a nonlinear change of variables. To fix ideas, consider the change of variables $x_i = y_i^2$ to eliminate a constraint of the form $x_i \geq 0$. The inconvenience of this approach is that derivatives with respect to null variables are null, and so algorithms tend to stay artificially close to the boundary. A possible remedy for this drawback is to consider second derivatives information [9]. Discuss.