Contents

Chapter 4

Model Augmented Lagrangian Algorithm

We will consider the optimization problem defined by

$$\begin{array}{ll} \text{Minimize} & f(x) \\ \text{subject to} & h(x) = 0, \\ & g(x) \leq 0, \\ & x \in \Omega, \end{array} \tag{4.1}$$

where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^p$, $f : \mathbb{R}^n \to \mathbb{R}$ are continuous and $\Omega \subseteq \mathbb{R}^n$ is closed (not necessarily convex!). In this chapter we do not require the existence of derivatives.

The Lagrangian function $\mathscr{L}$ will be defined by

$$\mathscr{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x) \tag{4.2}$$

for all $x \in \Omega$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^p_+$, whereas the Augmented Lagrangian [144, 217, 228] will be given by

$$L_\rho(x, \lambda, \mu) = f(x) + \frac{\rho}{2} \left\{ \sum_{i=1}^{m} \left[h_i(x) + \frac{\lambda_i}{\rho} \right]^2 + \sum_{i=1}^{p} \left[\max\left(0, g_i(x) + \frac{\mu_i}{\rho}\right) \right]^2 \right\} \tag{4.3}$$

for all $\rho > 0$, $x \in \Omega$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^p_+$.

In order to understand the meaning of the definition (4.3), first consider the case in which $\lambda = 0$ and $\mu = 0$. Then,

$$L_\rho(x, 0, 0) = f(x) + \frac{\rho}{2} \left(\|h(x)\|_2^2 + \|g(x)_+\|_2^2 \right). \tag{4.4}$$

Therefore, $L_\rho(x, 0, 0)$ is the "external penalty function" that coincides with $f(x)$ within the feasible set and penalizes the lack of feasibility by means of the term

$$\frac{\rho}{2} \left(\|h(x)\|_2^2 + \|g(x)_+\|_2^2 \right).$$

For big values of the penalty parameter ρ, the penalty term "dominates" $f(x)$ and the level sets of $L_\rho(x, 0, 0)$ tend to be those of $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$ in the infeasible region.

Consider the problem

$$\text{Minimize } f(x) \text{ subject to } h(x) = 0, \tag{4.5}$$

where
$$f(x) = (x_1 - 6)^2 + x_2^2$$
and
$$h(x) = (x_2 - (x_1/4)^2)^2 + (x_1/4 - 1)^2 - 1,$$
illustrated in Figure 4.1. The level sets of $h(x)^2$ are shown in Figure 4.2(a), while Figures 4.2(b)–4.2(d) show the level sets of $L_\rho(x, 0, 0)$ for $\rho \in \{1, 100, 1{,}000\}$. It is easy to see that, with $\rho = 1{,}000$, the penalty term $\rho h(x)^2$ dominates $f(x)$ in $L_\rho(x, 0, 0)$ and, therefore, the level sets depicted in Figures 4.2(a) and 4.2(d) are very similar (in fact, they are indistinguishable).

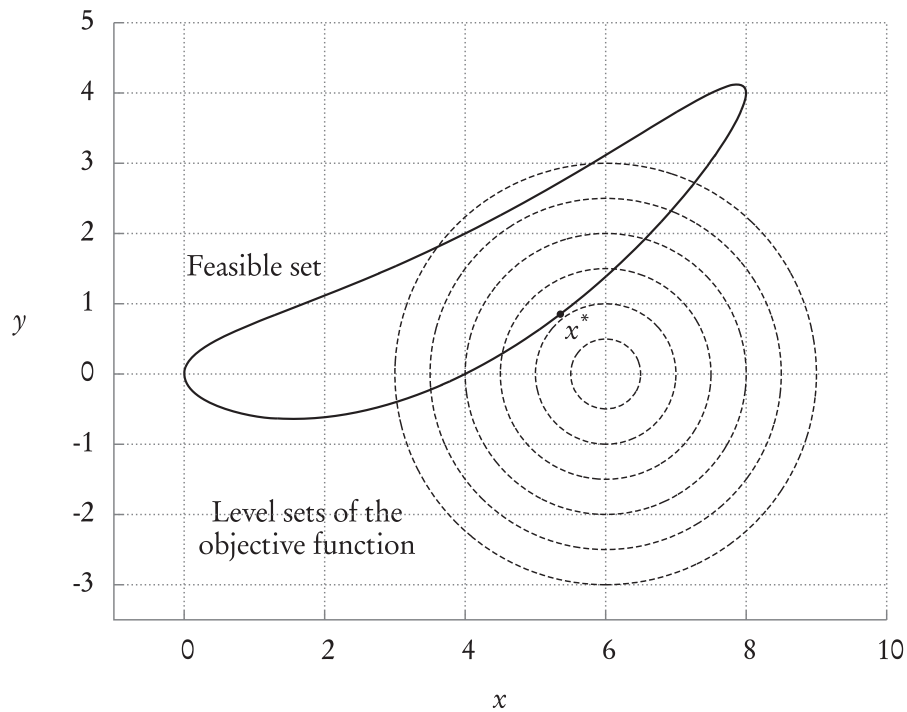

Figure 4.1. *Feasible set and level sets of the objective function of problem (4.5). Solution is given by* $x^* \approx (5.3541, 0.8507)^T$.

This means that, for x infeasible, the external penalty function (4.4) gives little information about the objective function if ρ is very large. The Augmented Lagrangian (4.3) may be seen as the penalized function in which "punishment" of infeasibility occurs, not with respect to the true constraints $h(x) = 0$ and $g(x) \le 0$ but with respect to the shifted constraints $h(x) + \lambda/\rho = 0$ and $g(x) + \mu/\rho \le 0$. The reasons for employing shifted constraints, instead of the original ones, will be explained after the definition of the following model algorithm.

4.1 ▪ Main model algorithm and shifts

The following algorithm is a basic Augmented Lagrangian algorithm for solving (4.1). The algorithm proceeds by minimizing the Augmented Lagrangian function at each iteration and updating Lagrange multipliers and penalty parameters between iterations. Its generality will allow us to analyze several particular cases that address the problem (4.1) under different conditions.

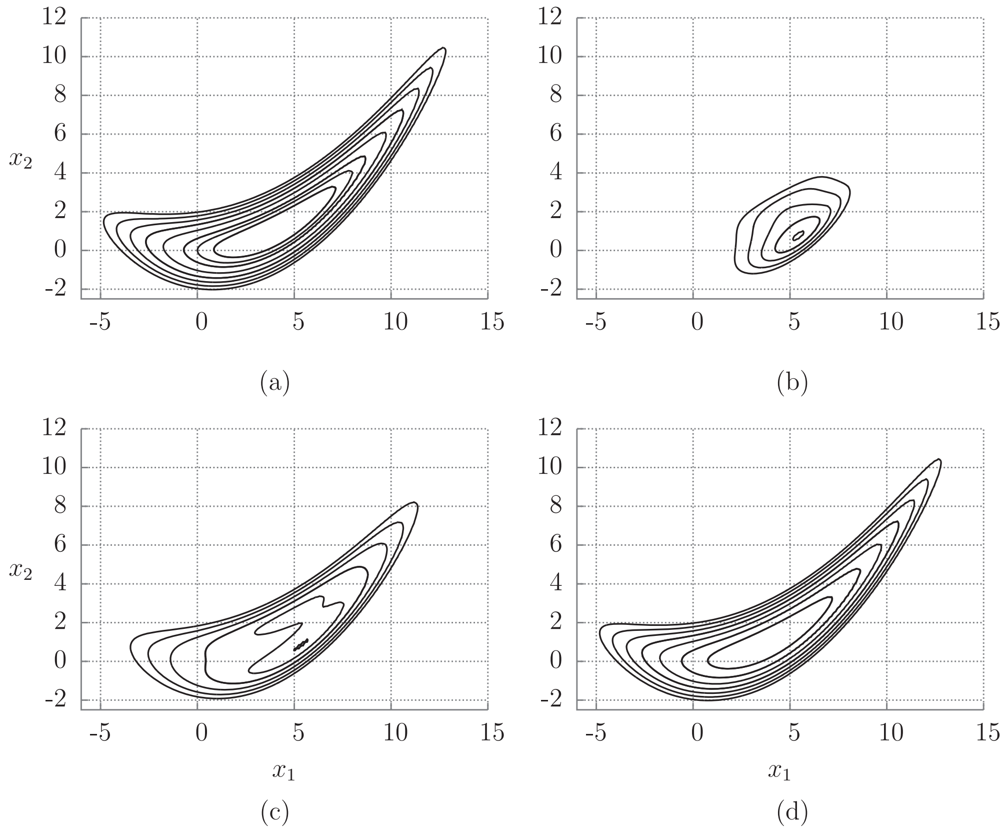

Figure 4.2. *Level sets of (a) $h(x)^2$ and (b)–(d) level sets of $L_\rho(x,0,0)$ for $\rho \in \{1, 100, 1{,}000\}$, respectively.*

Algorithm 4.1.

Let $\lambda_{\min} < \lambda_{\max}$, $\mu_{\max} > 0$, $\gamma > 1$, and $0 < \tau < 1$. Let $\bar{\lambda}^1 \in [\lambda_{\min}, \lambda_{\max}]^m$, $\bar{\mu}^1 \in [0, \mu_{\max}]^p$, and $\rho_1 > 0$. Initialize $k \leftarrow 1$.

Step 1. Find $x^k \in \mathbb{R}^n$ as an approximate solution of

$$\text{Minimize } L_{\rho_k}(x, \bar{\lambda}^k, \bar{\mu}^k) \text{ subject to } x \in \Omega. \tag{4.6}$$

Step 2. Compute new approximations of the Lagrange multipliers

$$\lambda^{k+1} = \bar{\lambda}^k + \rho_k h(x^k) \tag{4.7}$$

and

$$\mu^{k+1} = \left(\bar{\mu}^k + \rho_k g(x^k)\right)_+. \tag{4.8}$$

Step 3. Define

$$V_i^k = \min\left\{-g_i(x^k), \bar{\mu}_i^k / \rho_k\right\} \text{ for } i = 1, \dots, p.$$

If $k = 1$ or

$$\max\left\{\|h(x^k)\|, \|V^k\|\right\} \le \tau \max\left\{\|h(x^{k-1})\|, \|V^{k-1}\|\right\}, \tag{4.9}$$

choose $\rho_{k+1} \ge \rho_k$. Otherwise, define $\rho_{k+1} = \gamma \rho_k$.

Step 4. Compute $\bar{\lambda}^{k+1} \in [\lambda_{\min}, \lambda_{\max}]^m$ and $\bar{\mu}_i^{k+1} \in [0, \mu_{\max}]^p$.

Step 5. Set $k \leftarrow k+1$ and go to Step 1.

Note that λ^{k+1} and μ^{k+1} are not used in this algorithm but will be used in more specific versions to define $\bar{\lambda}^{k+1}$ and $\bar{\mu}^{k+1}$, respectively.

Algorithm 4.1 is said to be conceptual because Step 1, which establishes that the iterate x^k should be an "approximate minimizer" of the subproblem (4.6), is defined in a deliberately ambiguous way. (Strictly speaking, any point $x^k \in \mathbb{R}^n$ could be accepted as an "approximate minimizer"!) Constraints defined by $x \in \Omega$ will be said to be "hard," "simple," or "nonrelaxable," following the terminology of Audet and Dennis [24]. Other authors prefer the terms "lower-level constraints" [8] or "subproblem constraints" [48]. The term "hard" only appears contradictory to "simple." These constraints are generally simple in the sense that it is not difficult to satisfy them. However, they are hard (or nonrelaxable) in the sense that it is not admissible not to satisfy them.

The quantities ρ_k are said to be "penalty parameters." Ideally, at each "outer iteration" one solves the subproblem (4.6) and, if enough progress has been obtained in terms of improvement of feasibility and complementarity, the same penalty parameter *may be* used at the next outer iteration ($\rho_{k+1} \geq \rho_k$), while the penalty parameter *must be* increased if progress is not satisfactory.

The quantities $\bar{\lambda}_i^k/\rho_k \in \mathbb{R}$ and $\bar{\mu}_i^k/\rho_k \in \mathbb{R}_+$ can be interpreted as "shifts." In (4.6), one penalizes not merely the violation of infeasibilities (this would be the case if the shifts were null) but the infeasibilities modified by the shifts $\bar{\lambda}_i^k/\rho_k$ and $\bar{\mu}_i^k/\rho_k$. The idea is that, even with a penalty parameter of moderate value, a judicious choice of the shifts makes possible the coincidence, or near coincidence, of the solution to subproblem (4.6) with the desired minimizer of (4.1). See Figure 4.3, where we compare the solution for the problem

$$\text{Minimize } x \text{ subject to } -x \leq 0$$

to the solution for a subproblem with $\rho_k = 1$ and null shift ($\bar{\mu}^k = 0$) and to the solution for a subproblem with $\rho_k = 1$ and the "correct" shift ($\bar{\mu}^k = 1$).

The shifts are corrected according to formulae (4.7) and (4.8). The idea is as follows. Assume that x^k came from solving (4.6) with shifts $\bar{\lambda}_k/\rho_k$ and $\bar{\mu}_k/\rho_k$. After this process, if one verifies that the feasibility of an inequality has been violated by a quantity $g_i(x^k) > 0$, it is reasonable to think that the shift should be increased by a quantity equal to this violation. This leads to the formula $\mu^{k+1}/\rho_k = \bar{\mu}^k/\rho_k + g_i(x^k)$ or, equivalently, $\mu^{k+1} = \bar{\mu}^k + \rho_k g(x^k)$. Moreover, if the infeasibility has not been violated and $-\mu^k/\rho_k < g_i(x^k) \leq 0$, the suggestion is that the shift has been excessively big, and a reduction is necessary to make it possible that $g_i(x) = 0$ at the new iteration, with a possible improvement of the objective function. Again, this suggests the updating rule $\mu^{k+1}/\rho_k = \bar{\mu}^k/\rho_k + g_i(x^k)$. Finally, if $g_i(x^k) < -\mu^k/\rho_k$, we guess that the shift was not necessary and should be null from now on. Similar reasoning may be done with respect to formula (4.7), which updates the shifts corresponding to the equality constraints.

Algorithms are conceivable in which the penalty parameters become fixed and only shift modifications are made. These algorithms can be interpreted, under convexity assumptions, as "proximal point" methods in the dual problem of (4.1), and their properties have been exhaustively studied in textbooks and survey papers (see [39, 150, 228], among others).

In Algorithm 4.1, not only are the shifts updated, but so are the penalty parameters, according to the test (4.9). In (4.9), the progress of two different quantities is considered.

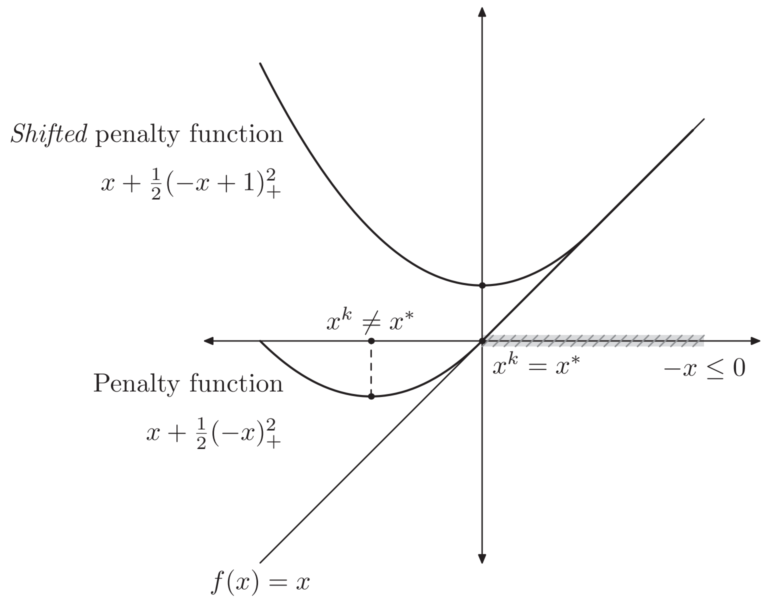

Figure 4.3. *Comparison between the solution to the problem Minimize x subject to $-x \leq 0$ and (a) the solution to a subproblem with $\rho_k = 1$ and no shift ($\bar{\mu}^k = 0$) and (b) the solution to a subproblem with $\rho_k = 1$ and the "correct" shift ($\bar{\mu}^k = 1$).*

On the one hand, one needs progress in terms of the feasibility of equality constraints, measured by $||h(x^k)||$. In the absence of improvement of this infeasibility measure, one decides to increase the penalty parameter. On the other hand, through the test (4.9), one also requires reduction of the quantities $\min\{-g_i(x^k), \bar{\mu}_i^k/\rho_k\}$. Note that $\bar{\mu}_i^k/\rho_k$, the shift already employed in (4.6), is nonnegative. Therefore, through consideration of $\min\{-g_i(x^k), \bar{\mu}_i^k/\rho_k\}$, we are implicitly testing the progress in terms of fulfillment of the inequality constraint $g_i(x) \leq 0$. In fact, if $g_i(x^k)$ tends to zero, improvement of $\min\{-g_i(x^k), \bar{\mu}_i^k/\rho_k\}$ very likely occurs, independently of the shifts $\bar{\mu}_i^k/\rho_k$. The interesting question is why we require this improvement even in the case that $g_i(x^k) \ll 0$ and x^k probably converges to a point at which $g_i(x)$ is inactive.

The answer is the following. If $g_i(x^k)$ is "very feasible" and the shift $\bar{\mu}_i^k/\rho_k$ is big, very likely it was the shift that forced $g_i(x^k)$ to be very negative at the solution to (4.6), since the subproblem penalizes deviations of the constraint from the shift, instead of mere infeasibility. However, although we are getting a feasible point and we may get a feasible point in the limit, since the feasible set is unnecessarily being reduced in this case, it is unlikely that an optimum could be obtained in this way. Therefore, we need to decrease the shift by increasing the penalty parameter.

According to the arguments above, it is sensible to choose the new Lagrange multipliers $\bar{\lambda}^{k+1}$ and $\bar{\mu}^{k+1}$ as λ^{k+1} and μ^{k+1}, respectively. In general, this is what is done in practice, but safeguards that guarantee the boundedness of $\{\bar{\lambda}^k\}$ and $\{\bar{\mu}^k\}$ are necessary. Safeguarded boundedness guarantees a crucial commonsense property: The shifts should tend to zero when the penalty parameter tends to infinity. Clearly, if we are led to penalize violations of the constraints with a very large ρ_k, it does not make sense to use shifts bounded away from zero since, in this case, we would be punishing hardly suitable feasible points. So, when ρ_k tends to infinity, common sense dictates that the shifts should tend to zero, and the most straightforward way to guarantee this is to impose bounds on the multipliers. Therefore, we may think of $\bar{\mu}^k$ and $\bar{\lambda}^k$ as being "safeguarded multipliers."

In the Augmented Lagrangian context, some authors prefer, at each outer iteration, to update either the multipliers or the penalty parameters, but not both. Our formulation in Algorithm 4.1 allows this possibility, although in practice we prefer to update penalty parameters and multipliers simultaneously.

The reader should observe that, for the motivation arguments given above, differentiability of the objective function or the constraints has not been invoked. Penalty and Augmented Lagrangian ideas are independent of the degree of smoothness of the functions that define the problem. This characteristic makes possible the application of the Augmented Lagrangian techniques to many nonstandard optimization problems.

4.2 ▪ Multipliers and inactive constraints

Despite the generality of Algorithm 4.1, it is possible to prove a useful property: Inequality multipliers corresponding to constraints that are inactive in the limit are asymptotically null, independent of the feasibility of the limit point. Note that, in the statement of Theorem 4.1, the existence of a limit point x^* is assumed. The existence of limit points in this and several other results should be confirmed by employing, in general, boundedness arguments concerning the feasible set.

Theorem 4.1. *Assume that the sequence $\{x^k\}$ is generated by Algorithm 4.1 and $K \underset{\infty}{\subset} \mathbb{N}$ is such that $\lim_{k\in K} x^k = x^*$. Then, for $k \in K$ large enough,*

$$\mu_i^{k+1} = 0 \text{ for all } i = 1,\dots,p \text{ such that } g_i(x^*) < 0. \tag{4.10}$$

Proof. By (4.8), $\mu^{k+1} \in \mathbb{R}^p_+$ for all $k \in \mathbb{N}$.

Assume that $g_i(x^*) < 0$ and let $k_1 \in \mathbb{N}$ and $c < 0$ be such that

$$g_i(x^k) < c < 0 \text{ for all } k \in K, k \geq k_1.$$

We consider two cases:

1. The sequence $\{\rho_k\}$ tends to infinity.

2. The sequence $\{\rho_k\}$ is bounded.

In the first case, since $\{\bar{\mu}_i^k\}$ is bounded, there exists $k_2 \geq k_1$ such that, for all $k \in K$, $k \geq k_2$,

$$\bar{\mu}_i^k + \rho_k g_i(x^k) < 0.$$

By (4.8), this implies that

$$\mu_i^{k+1} = 0 \text{ for all } k \in K, k \geq k_2.$$

Consider now the case in which $\{\rho_k\}$ is bounded. Then, (4.9) holds for all k large enough and, consequently,

$$\lim_{k\to\infty} V_i^k = 0.$$

Thus,

$$\lim_{k\to\infty} \left| \min\{-g_i(x^k), \bar{\mu}_i^k/\rho_k\} \right| = 0.$$

Since $g_i(x^k) < c < 0$ for $k \in K$ large enough, we have that

$$\lim_{k\in K} \bar{\mu}_i^k / \rho_k = 0.$$

Thus, since the sequence $\{\rho_k\}$ is bounded,

$$\lim_{k\in K} \bar{\mu}_i^k = 0.$$

Therefore, for $k \in K$ large enough,

$$\bar{\mu}_i^k + \rho_k g_i(x^k) < 0.$$

By the definition (4.8) of μ^{k+1}, this implies that $\mu_i^{k+1} = 0$ for $k \in K$ large enough, as we wanted to prove. □

4.3 ▪ Review and summary

The basic Augmented Lagrangian algorithm is composed of outer iterations, and at each, one minimizes the objective function plus a term that penalizes shifted constraints. The use of shifts has the appeal of avoiding the necessity of increasing the penalty parameter up to values at which the objective function becomes numerically neglected. Both the penalty parameter and the shifts are updated after each outer iteration. In particular, shifts are updated according to commonsense rules whose plausibility does not depend on the differentiability of the problem.

4.4 ▪ Further reading

In the case in which h is a linear mapping, there are no inequality constraints, and Ω represents an n-dimensional box, subproblems generated by an Augmented Lagrangian method based on the Powell–Hestenes–Rockafeller (PHR) [144, 217, 228] Augmented Lagrangian function (4.3) are box-constrained quadratic optimization problems. This fact was exhaustively exploited by Dostál [102] in order to define "optimal" quadratic programming methods. The use of different penalty functions (instead of the quadratic loss) and the generalization of the shifting ideas give rise to many alternative Augmented Lagrangian algorithms [5, 25, 32, 33, 77, 128, 149, 150, 164, 165, 195, 209, 210, 227, 249, 257, 260]. Most of these algorithms can be recommended for particular structures, but, for general problems reported in popular collections, the classical PHR approach seems to be more efficient and robust than nonclassical approaches [44, 104]. In this book, all the theory is dedicated to properties on general (not necessarily convex) problems. When convexity is assumed for the objective function and constraints, profound results can be obtained using the dual equivalence with the so-called proximal point methods. Iusem's survey [150] offers a good overview of this subject. An alternative Augmented Lagrangian approach that deals with nonlinear semidefinite programming was proposed in [166] and gave rise to the PENNON software package [166, 167, 168].

Sometimes, the penalty parameter is increased at the first iterations of the Augmented Lagrangian method, but, at later iterations, smaller penalty parameters are admissible. An extension of the basic Augmented Lagrangian method in which a nonmonotone strategy for penalty parameters is employed may be found in [56]. In cases in which the objective function takes very low values (perhaps going to $-\infty$) at infeasible points, penalty and

Augmented Lagrangian algorithms may be attracted by those points at early iterations and practical convergence could be discouraged. This phenomenon is called "greediness" in [43] and [75], where theoretically justified remedies are suggested. An application of the Augmented Lagrangian philosophy to a relevant family of nonsmooth problems may be found in [42].

4.5 ▪ Problems

4.1 Describe Algorithm 4.1 in terms of "shifts" instead of multipliers. Write explicitly the updating rules for shifts. Replace the boundedness condition on the multipliers with some condition on the shifts guaranteeing that shifts go to zero if multipliers go to infinity.

4.2 Justify the updating formula for the Lagrange multipliers corresponding to equality constraints using commonsense criteria, as we did in the case of inequalities in Section 4.1.

4.3 Analyze Algorithm 4.1 in the case that $\bar{\lambda}^k = 0$ and $\bar{\mu}^k = 0$ for all k.

4.4 Analyze Algorithm 4.1 in the case that at the resolution of the subproblem, one defines x^k to be an arbitrary, perhaps random, point of $\mathbb{R}^n$.

4.5 Analyze Algorithm 4.1 in the case that there are no constraints at all ($m = p = 0$) besides those corresponding to $x \in \Omega$.

4.6 Compare Algorithm 4.1 in the following two situations: when constraints $x \in \Omega$ remain in the lower level and when they are incorporated into the relaxable set $h(x) = 0$ and $g(x) \leq 0$ (if possible).

4.7 Give examples in which constraints $x \in \Omega$ cannot be expressed as systems of equalities and inequalities.

4.8 Assume that you are convinced that (4.7) is the reasonable way to update the equality Lagrange multipliers, but you are not convinced about the plausibility of (4.8). Replace each constraint $g_i(x) \leq 0$ in (4.1) with $g_i(x)+z_i^2 = 0$, where z_i is a slack variable, and reformulate Algorithm 4.1 for the new problem, now without inequality constraints. At the solution x^k of the new formulation of (4.6), observe that it is sensible (why?) to define

$$(z_i^k)^2 = -g_i(x^k) - \bar{\mu}_i^k/\rho_k \text{ if } g_i(x^k) + \bar{\mu}_i^k/\rho_k < 0$$

and

$$(z_i^k)^2 = 0 \text{ if } g_i(x^k) + \bar{\mu}_i^k/\rho_k \geq 0.$$

Deduce that the infeasibility for the constraint $g_i(x) + z_i^2 = 0$, or, equivalently, $g_i(x) \leq 0$, may be defined by

$$\left|\min\left\{-g_i(x^k), \bar{\mu}_i^k/\rho_k\right\}\right|$$

and that, according to (4.7), the formula for the new multiplier μ^{k+1} should be given by (4.8). In other words, the scheme defined in Algorithm 4.1 with inequality constraints can be deduced from the scheme defined for problems with equality constraints only. See [39].

4.9 Consider the following alternative definition for V^k in Algorithm 4.1:

$$V_i^k = \min\{-g_i(x^k), \mu_i^{k+1}\} \text{ for } i = 1,\ldots,p.$$

Discuss the adequacy of this definition and prove Theorem 4.1 for the modified algorithm.

4.10 Consider the following alternative definition for V^k in Algorithm 4.1:

$$V_i^k = |\mu_i^{k+1} g_i(x^k)| \text{ for } i = 1,\ldots,p, \tag{4.11}$$

and replace the test (4.9) with

$$\max\{\|h(x^k)\|, \|g(x^k)_+\|, \|V^k\|\} \le \tau \max\{\|h(x^{k-1})\|, \|g(x^{k-1})_+\|, \|V^{k-1}\|\}.$$

Discuss the adequacy of these alternatives and prove Theorem 4.1 for the modified algorithm.

4.11 Define V^k as in (4.11) and

$$W_i^k = |\lambda_i^{k+1} h_i(x^k)| \text{ for } i = 1,\ldots,m.$$

Replace the test (4.9) with

$$\max\{\|h(x^k)\|, \|g(x^k)_+\|, \|V^k\|, \|W^k\|\}$$
$$\le \tau \max\{\|h(x^{k-1})\|, \|g(x^{k-1})_+\|, \|V^{k-1}\|, \|W^{k-1}\|\}.$$

Discuss the adequacy of these alternatives and prove Theorem 4.1 for the modified algorithm.

4.12 Assume that there exists $c \in \mathbb{R}$ such that at Step 1 of Algorithm 4.1, we have that

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

for all $k \in \mathbb{N}$. Prove that any limit point x^* of $\{x^k\}$ verifies $h(x^*) = 0$ and $g(x^*) \le 0$.

4.13 Consider constrained optimization problems that include at least one "semidefinite constraint," which says that a set of variables defines a symmetric positive semidefinite matrix. Consider the possibility of coding that constraint as a set of upper-level constraints of the form $X = MM^T$, where M is an auxiliary matrix. Consider a different possibility: setting positive semidefinitness as a lower-level constraint on which we know how to project (how?). Analyze advantages and disadvantages and repeat a copy of this problem in all chapters of the book. (Specific Augmented Lagrangian methods for this case may be found in Kocvara and Stingl [162] and Stingl [241].)

4.14 Make your choices: Discuss reasonable values for the algorithmic parameters $\lambda_{\min}$, $\lambda_{\max}$, $\mu_{\max}$, ρ_1, γ, and τ. Implement Algorithm 4.1 (in your favorite language) in the case that Ω is a box and using some simple trial and error strategy for the approximate minimization of the Augmented Lagrangian subproblem (4.6). Run your code using simple examples and draw conclusions.

4.15 Try to exploit algorithmically the consequences of these facts:

(a) The original problem that you want to solve is equivalent to the problem in which you add a fixed penalization to the objective function.

(b) In your specific problem a feasible initial point is easily available.

Second Proofs

Chapter 5

Global Minimization Approach

In this chapter, the subproblems (4.6) at Step 1 of Algorithm 4.1 will be interpreted in terms of global optimization. Namely, at each outer iteration, we will assume that x^k is an approximate global minimizer of the Augmented Lagrangian on Ω. In principle, the global minimization of the Augmented Lagrangian on Ω could be as difficult as the original problem, since we make no assumptions on the geometry of this set. However, in practice, the set Ω is, in general, simple enough to make global minimization on Ω much easier than on the feasible set of problem (4.1).

The global minimization method for solving (4.1) considered in this chapter will be Algorithm 4.1 with the following algorithmic assumption.

Assumption 5.1. *For all $k \in \mathbb{N}$, we obtain $x^k \in \Omega$ such that*

$$L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) \leq L_{\rho_k}(x, \bar{\lambda}^k, \bar{\mu}^k) + \varepsilon_k \text{ for all } x \in \Omega,$$

where the sequence of tolerances $\{\varepsilon_k\} \subseteq \mathbb{R}_+$ is bounded.

As in Chapter 4, in this chapter we only assume continuity of the objective function and the functions that define the constraints.

5.1 ▪ Feasibility result

Assumption 5.1 says that, at each outer iteration, one finds an approximate global minimizer of the subproblem. In principle, the tolerances ε_k do not need to be small at all. In the following theorem, we prove that, even using possibly big tolerances, we obtain, in the limit, a global minimizer of the infeasibility measure.

Theorem 5.1. *Assume that $\{x^k\}$ is a sequence generated by Algorithm 4.1 under Assumption 5.1. Let x^* be a limit point of $\{x^k\}$. Then, for all $x \in \Omega$, we have that*

$$\|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2 \leq \|h(x)\|_2^2 + \|g(x)_+\|_2^2.$$

Proof. Since Ω is closed and $x^k \in \Omega$, we have that $x^* \in \Omega$. We consider two cases: $\{\rho_k\}$ bounded and $\rho_k \to \infty$.

If $\{\rho_k\}$ is bounded, there exists k_0 such that $\rho_k = \rho_{k_0}$ for all $k \geq k_0$. Therefore, for all $k \geq k_0$, (4.9) holds. This implies that $\|h(x^k)\| \to 0$ and $\|V^k\| \to 0$, so $g_i(x^k)_+ \to 0$ for all $i = 1, \ldots, p$. Thus, the limit point is feasible.

Now, assume that $\rho_k \to \infty$. Let $K \underset{\infty}{\subset} \mathbb{N}$ be such that

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

Assume by contradiction that there exists $x \in \Omega$ such that

$$||h(x^*)||_2^2 + ||g(x^*)_+||_2^2 > ||h(x)||_2^2 + ||g(x)_+||_2^2.$$

By the continuity of h and g, the boundedness of $\{\bar{\lambda}^k\}$ and $\{\bar{\mu}^k\}$, and the fact that ρ_k tends to infinity, there exist $c > 0$ and $k_0 \in \mathbb{N}$ such that for all $k \in K, k \geq k_0$,

$$\left\| h(x^k) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 > \left\| h(x) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 + c.$$

Therefore, for all $k \in K, k \geq k_0$,

$$f(x^k) + \frac{\rho_k}{2} \left[\left\| h(x^k) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 \right]$$
$$> f(x) + \frac{\rho_k}{2} \left[\left\| h(x) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 \right] + \frac{\rho_k c}{2} + f(x^k) - f(x).$$

Since $\lim_{k \in K} x^k = x^*$, f is continuous, and $\{\varepsilon_k\}$ is bounded, there exists $k_1 \geq k_0$ such that, for $k \in K$, $k \geq k_1$,

$$\frac{\rho_k c}{2} + f(x^k) - f(x) > \varepsilon_k.$$

Therefore,

$$f(x^k) + \frac{\rho_k}{2} \left[\left\| h(x^k) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 \right]$$
$$> f(x) + \frac{\rho_k}{2} \left[\left\| h(x) + \frac{\bar{\lambda}^k}{\rho_k} \right\|_2^2 + \left\| \left(g(x) + \frac{\bar{\mu}^k}{\rho_k} \right)_+ \right\|_2^2 \right] + \varepsilon_k$$

for $k \in K, k \geq k_1$. This contradicts Assumption 5.1. □

5.2 ▪ Optimality result

Theorem 5.1 says that Algorithm 4.1, with the iterates defined by Assumption 5.1, finds minimizers of the infeasibility. Therefore, if the original optimization problem is feasible, every limit point of a sequence generated by the algorithm is feasible. Note that we only used boundedness of the sequence of tolerances $\{\varepsilon_k\}$ in the proof of Theorem 5.1. Now, we will see that, assuming that ε_k tends to zero, it is possible to prove that, in the feasible case, the algorithm asymptotically finds global minimizers of (4.1).

Theorem 5.2. *Assume that $\{x^k\}$ is a sequence generated by Algorithm 4.1 under Assumption 5.1 and $\lim_{k \to \infty} \varepsilon_k = 0$. Moreover, assume that, in the case that (4.9) holds, we always choose $\rho_{k+1} = \rho_k$. Let x^* be a limit point of $\{x^k\}$. Suppose that problem (4.1) is feasible. Then, x^* is a global minimizer of (4.1).*

Proof. Let $K \underset{\infty}{\subset} \mathbb{N}$ be such that $\lim_{k\in K} x^k = x^*$. By Theorem 5.1, since the problem is feasible, we have that x^* is feasible. Let $x \in \Omega$ be such that $h(x) = 0$ and $g(x) \le 0$.

We consider two cases: $\rho_k \to \infty$ and $\{\rho_k\}$ bounded.

Case 1 ($\rho_k \to \infty$). By the definition of the algorithm, we have that

$$
\begin{aligned}
&f(x^k) + \frac{\rho_k}{2}\left[\left\| h(x^k) + \frac{\bar\lambda^k}{\rho_k}\right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar\mu^k}{\rho_k}\right)_+ \right\|_2^2\right] \\
&\le f(x) + \frac{\rho_k}{2}\left[\left\| h(x) + \frac{\bar\lambda^k}{\rho_k}\right\|_2^2 + \left\| \left(g(x) + \frac{\bar\mu^k}{\rho_k}\right)_+ \right\|_2^2\right] + \varepsilon_k
\end{aligned} \tag{5.1}
$$

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

Since $h(x) = 0$ and $g(x) \le 0$, we have that

$$
\left\| h(x) + \frac{\bar\lambda^k}{\rho_k}\right\|_2^2 = \left\| \frac{\bar\lambda^k}{\rho_k}\right\|_2^2 \text{ and } \left\| \left(g(x) + \frac{\bar\mu^k}{\rho_k}\right)_+ \right\|_2^2 \le \left\| \frac{\bar\mu^k}{\rho_k}\right\|_2^2 .
$$

Therefore, by (5.1),

$$
f(x^k) \le f(x^k) + \frac{\rho_k}{2}\left[\left\| h(x^k) + \frac{\bar\lambda^k}{\rho_k}\right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar\mu^k}{\rho_k}\right)_+ \right\|_2^2\right] \le f(x) + \frac{\|\bar\lambda^k\|_2^2}{2\rho_k} + \frac{\|\bar\mu^k\|_2^2}{2\rho_k} + \varepsilon_k .
$$

Taking limits for $k \in K$ and using that $\lim_{k\in K} \|\bar\lambda^k\|/\rho_k = \lim_{k\in K}\|\bar\mu^k\|/\rho_k = 0$ and $\lim_{k\in K} \varepsilon_k = 0$, by the continuity of f and the convergence of x^k, we get

$$
f(x^*) \le f(x).
$$

Since x is an arbitrary feasible point, it turns out that x^* is a global minimizer, as we wanted to prove.

Case 2 ($\{\rho_k\}$ bounded). In this case, there exists $k_0 \in \mathbb{N}$ such that $\rho_k = \rho_{k_0}$ for all $k \ge k_0$. Therefore, by Assumption 5.1,

$$
\begin{aligned}
&f(x^k) + \frac{\rho_{k_0}}{2}\left[\left\| h(x^k) + \frac{\bar\lambda^k}{\rho_{k_0}}\right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar\mu^k}{\rho_{k_0}}\right)_+ \right\|_2^2\right] \\
&\le f(x) + \frac{\rho_{k_0}}{2}\left[\left\| h(x) + \frac{\bar\lambda^k}{\rho_{k_0}}\right\|_2^2 + \left\| \left(g(x) + \frac{\bar\mu^k}{\rho_{k_0}}\right)_+ \right\|_2^2\right] + \varepsilon_k
\end{aligned}
$$

for all $k \ge k_0$. Since $g(x) \le 0$ and $\bar\mu^k/\rho_{k_0} \ge 0$,

$$
\left\| \left(g(x) + \frac{\bar\mu^k}{\rho_{k_0}}\right)_+ \right\|_2^2 \le \left\| \frac{\bar\mu^k}{\rho_{k_0}}\right\|_2^2 .
$$

Thus, since $h(x) = 0$,

$$
f(x^k) + \frac{\rho_{k_0}}{2}\left[\left\| h(x^k) + \frac{\bar\lambda^k}{\rho_{k_0}}\right\|_2^2 + \left\| \left(g(x^k) + \frac{\bar\mu^k}{\rho_{k_0}}\right)_+ \right\|_2^2\right] \le f(x) + \frac{\rho_{k_0}}{2}\left[\left\| \frac{\bar\lambda^k}{\rho_{k_0}}\right\|_2^2 + \left\| \frac{\bar\mu^k}{\rho_{k_0}}\right\|_2^2\right] + \varepsilon_k
$$

for all $k \ge k_0$. Let $K_1 \underset{\infty}{\subset} K$, $\lambda^* \in R^m$, and $\mu^* \in \mathbb{R}^p$ be such that

$$
\lim_{k\in K_1} \bar\lambda^k = \lambda^* \text{ and } \lim_{k\in K_1} \bar\mu^k = \mu^* .
$$

By the feasibility of x^*, taking limits in the inequality above for $k \in K_1$, we get

$$f(x^*) + \frac{\rho_{k_0}}{2}\left[\left\|\frac{\bar{\lambda}^*}{\rho_{k_0}}\right\|_2^2 + \left\|\left(g(x^*) + \frac{\bar{\mu}^*}{\rho_{k_0}}\right)_+\right\|_2^2\right] \le f(x) + \frac{\rho_{k_0}}{2}\left[\left\|\frac{\bar{\lambda}^*}{\rho_{k_0}}\right\|_2^2 + \left\|\frac{\bar{\mu}^*}{\rho_{k_0}}\right\|_2^2\right].$$

Therefore,

$$f(x^*) + \frac{\rho_{k_0}}{2}\left\|\left(g(x^*) + \frac{\bar{\mu}^*}{\rho_{k_0}}\right)_+\right\|_2^2 \le f(x) + \frac{\rho_{k_0}}{2}\left\|\frac{\bar{\mu}^*}{\rho_{k_0}}\right\|_2^2.$$

Thus,

$$f(x^*) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{p}\left(g_i(x^*) + \frac{\bar{\mu}_i^*}{\rho_{k_0}}\right)_+^2 \le f(x) + \frac{\rho_{k_0}}{2}\sum_{i=1}^{p}\left(\frac{\bar{\mu}_i^*}{\rho_{k_0}}\right)^2. \tag{5.2}$$

Now, if $g_i(x^*) = 0$, since $\bar{\mu}_i^*/\rho_{k_0} \ge 0$, we have that

$$\left(g_i(x^*) + \frac{\bar{\mu}_i^*}{\rho_{k_0}}\right)_+ = \frac{\bar{\mu}_i^*}{\rho_{k_0}}.$$

Therefore, by (5.2),

$$f(x^*) + \frac{\rho_{k_0}}{2}\sum_{g_i(x^*)<0}\left(g_i(x^*) + \frac{\bar{\mu}_i^*}{\rho_{k_0}}\right)_+^2 \le f(x) + \frac{\rho_{k_0}}{2}\sum_{g_i(x^*)<0}\left(\frac{\bar{\mu}_i^*}{\rho_{k_0}}\right)^2. \tag{5.3}$$

But, by (4.9), $\lim_{k\to\infty} \max\{g_i(x^k), -\bar{\mu}_i^k/\rho_{k_0}\} = 0$. Therefore, if $g_i(x^*) < 0$, we necessarily have that $\bar{\mu}_i^* = 0$. Therefore, (5.3) implies that $f(x^*) \le f(x)$. Since x was an arbitrary feasible point, the proof is complete. $\square$

5.3 ▪ Optimality subject to minimal infeasibility

Under an additional assumption, this time on the rule for updating multipliers, a stronger result can be proved that concerns the behavior of the algorithm for infeasible problems. We have already seen that, in this case, the algorithm considered in this chapter converges to global minimizers of the infeasibility, measured by the sum of squares $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$. Under Assumption 5.2 below, we will show that limit points minimize the objective function subject to minimal infeasibility.

Assumption 5.2. *For all $k \in \mathbb{N}$, if $\lambda^{k+1} \notin [\lambda_{\min}, \lambda_{\max}]^m$ or $\mu^{k+1} \notin [0, \mu_{\max}]^p$, we choose $\bar{\lambda}^{k+1} = 0$ and $\bar{\mu}^{k+1} = 0$.*

The case in which $\|\lambda^{k+1}\| + \|\mu^{k+1}\|$ is big, contemplated in Assumption 5.2, generally corresponds to situations in which ρ_k is big too. As mentioned in Chapter 4, when infeasibility is severely penalized ($\rho_k \gg 1$), it makes no sense to employ shifts at all, because one could be adding a heavy penalization to the objective function even at reasonably feasible points. This argument, which supports the use of safeguards for the multipliers, can also be used to support the sensibility of the decision made in Assumption 5.2 (see [59]).

Theorem 5.3. *Assume that $\{x^k\}$ is a sequence generated by Algorithm 4.1 under Assumptions 5.1 and 5.2 and that $\lim_{k\to\infty} \varepsilon_k = 0$. Moreover, assume that, in the case in which (4.9) holds, we always choose $\rho_{k+1} = \rho_k$. Let x^* be a limit point of $\{x^k\}$. Then,*

$$\|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2 \le \|h(x)\|_2^2 + \|g(x)_+\|_2^2 \textit{ for all } x \in \Omega$$

and

$$f(x^*) \le f(x) \text{ for all } x \in \Omega \text{ such that } \|h(x)\|_2^2 + \|g(x)_+\|_2^2 = \|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2.$$

Proof. If $h(x^*) = 0$ and $g(x^*) \le 0$, the first part of the thesis follows immediately and the second part of the thesis follows from Theorem 5.2.

Let us assume from now on that $\|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2 = c > 0$. This implies by Step 3 that $\lim_{k\to\infty} \rho_k = \infty$. Since by Theorem 5.1 x^* is a global minimizer of $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$, it turns out that, for all $k \in \mathbb{N}$, $\|h(x^k)\|_2^2 + \|g(x^k)_+\|_2^2 \ge c$. By (4.7), (4.8), the boundedness of $\{\bar{\lambda}^k\}$ and $\{\bar{\mu}^k\}$, and the fact that ρ_k tends to infinity, we have that, for all k large enough, either $\lambda^{k+1} \notin [\lambda_{\min}, \lambda_{\max}]^m$ or $\mu^{k+1} \notin [0, \mu_{\max}]^p$. Therefore, by Assumption 5.2, there exists $k_0 \in \mathbb{N}$ such that for all $k \ge k_0$ we have that $\bar{\lambda}^k = 0$ and $\bar{\mu}^k = 0$.

Let $K \underset{\infty}{\subset} \{k_0, k_0+1, k_0+2, \dots\}$ be such that $\lim_{k\in K} x^k = x^*$.

By Assumption 5.1 and the fact that $\|\bar{\lambda}^k\| = \|\bar{\mu}^k\| = 0$, we have that, for all $x \in \Omega$,

$$f(x^k) + \frac{\rho_k}{2}[\|h(x^k)\|_2^2 + \|g(x^k)_+\|_2^2] \le f(x) + \frac{\rho_k}{2}[\|h(x)\|_2^2 + \|g(x)_+\|_2^2] + \varepsilon_k \tag{5.4}$$

for all $k \in K$. In particular, if $x \in \Omega$ is such that $\|h(x)\|_2^2 + \|g(x)_+\|_2^2 = \|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2$, we have that x is a global minimizer of the infeasibility on Ω. Thus,

$$\frac{\rho_k}{2}[\|h(x^k)\|_2^2 + \|g(x^k)_+\|_2^2] \ge \frac{\rho_k}{2}[\|h(x)\|_2^2 + \|g(x)_+\|_2^2].$$

Therefore, by (5.4) and Assumption 5.1,

$$f(x^k) \le f(x) + \varepsilon_k \quad \text{for all} \quad k \in K.$$

By the continuity of f, taking limits on both sides of this inequality, we obtain the desired result. □

5.4 ▪ Review and summary

The Augmented Lagrangian paradigm can be used for solving global optimization problems. The only requirement is that we need to use a global optimization procedure for solving the subproblems. Theorem 5.1 indicates that we can always expect to find feasible points (if they exist) if we globally solve the subproblems, even with a loose tolerance. If the original problem is feasible, the global form of the Augmented Lagrangian method finds global solutions. Moreover, with a special safeguard of the Lagrange multipliers, the method finds global minimizers subject to minimal infeasibility.

5.5 ▪ Further reading

Global optimization has many applications in all branches of engineering, sciences, and production. Several textbooks addressing different aspects of global optimization theory and applications are available [28, 117, 123, 148, 239, 246, 250, 258]. Useful review papers have also appeared [118, 211]. In [49], global minimizers of linearly constrained subproblems are computed α-BB method [3, 4]. In [226], Augmented Lagrangian box-constrained subproblems are solved employing a stochastic population–based strategy that aims to guarantee global convergence. A variation of the algorithm introduced in this chapter, with finite termination and finite detection of possible infeasibility, was introduced by Birgin, Martínez, and Prudente [58]. Employing duality arguments, some authors (see, for example, Burachik and Kaya [72] and [124]) transform the original constrained optimization problem into a simpler problem whose variables are Lagrange multipliers and penalty parameters. Applying subgradient techniques in the dual, convergence to global solutions is obtained.

5.6 ▪ Problems

5.1 If the second derivatives of an unconstrained optimization problem evaluated at a global minimizer are very big, a slight perturbation of the global minimizer could represent a large increase in the objective function. In this sense, perhaps, local minimizers with small second derivatives should be preferred over global minimizers with big second derivatives. Reformulate unconstrained minimization problems, taking into account this robustness issue.

5.2 Discuss the following apparent paradox: You need to solve the subproblems only very loosely if you want only to minimize the infeasibility (Theorem 5.1). In particular, it seems that, in order to minimize $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$, you do not need to employ global optimization procedures at all for solving the subproblems. Does this mean that global minimization of $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$ can be achieved without using global optimization? Does this contradict the fact that for obtaining global minimizers without further information one needs to evaluate the function on a dense set?

5.3 Prove Theorem 5.2 without the assumption that $\rho_{k+1} = \rho_k$ when (4.9) holds.

5.4 Note that Assumption 5.2 says that, under some circumstances, it is sensible to eliminate shifts and reduce the algorithm to the penalty method. Suggest alternative tests that could be employed to decide to annihilate $\bar{\lambda}^k$ and $\bar{\mu}^k$.

5.5 Make your choices: Implement Algorithm 4.1 with the assumptions made in this chapter. Consider the possibility of using some heuristic for obtaining an approximate global minimizer of the Augmented Lagrangian at Step 1. Run simple examples and draw conclusions.

5.6 Employ your code to solve problems in which the feasible region is empty. Observe whether it behaves as described by the theorems presented in this chapter.

5.7 Discuss the application of the algorithm and theory presented in this chapter to the capacity expansion planning problem [176] presented in Chapter 2. Note that the binary-variable constraints may be modeled as nonlinear constraints, but this does not prevent the use of mixed-integer techniques in the solution process.

5.8 In terms of detecting infeasibility, the pure penalty method seems to have better convergence properties than the Augmented Lagrangian algorithm (why?). Suggest a safeguarded updating procedure for the multipliers taking advantage of this property.

Chapter 6

General Affordable Algorithms

In the global optimization literature, algorithms that are designed to converge not to global minimizers but to mere stationary points (in fact, not necessarily local minimizers) are known as local algorithms. This denomination could be adopted with a warning that local algorithms are generally guaranteed to converge in some sense to stationary points of the optimization problem, independently of the initial approximation. In this sense, they are said to be globally convergent. Roughly speaking, "local algorithm" is synonymous with the "affordable algorithm" of Chapter 3. In general, global optimization algorithms are not reliable for solving large-scale problems, and, for small to medium-scale problems, they are much slower than local algorithms. On the other hand, global optimization software makes use of local algorithms when associated with branch-and-bound procedures by means of which the search space for a global minimizer is reduced.

The denomination local algorithm does not allude to the concept of local convergence, which is related to the convergence of the whole sequence if one starts close enough to a solution. In fact, local algorithms are usually globally convergent in the sense of stationarity of limit points but are not necessarily locally convergent as they may generate sequences that accumulate in more than one cluster point.

In this chapter, the description of a local algorithm based on the Augmented Lagrangian corresponds to Algorithm 4.1 with a precise interpretation of Step 1, which says that the subproblem solution x^k is approximately a KKT point of the subproblem.

First, we consider lower-level constraints of the following form:

$$\Omega = \{x \in \mathbb{R}^n \mid \underline{h}(x) = 0,\ \underline{g}(x) \le 0\},$$

where $\underline{h} : \mathbb{R}^n \to \mathbb{R}^{\underline{m}}$, $\underline{g} : \mathbb{R}^n \to \mathbb{R}^{\underline{p}}$, and f, h, g, $\underline{h}$, and $\underline{g}$ admit continuous first derivatives on $\mathbb{R}^n$. Consequently, the constrained optimization problem that we wish to solve is

$$\text{Minimize } f(x) \text{ subject to } h(x) = 0,\ g(x) \le 0,\ x \in \Omega. \tag{6.1}$$

Note that the simple box constraints $\ell \le x \le u$ can be expressed trivially in the form $\underline{g}(x) \le 0$.

Assumption 6.1 below defines the sense in which the approximate minimization at Step 1 of Algorithm 4.1 should be interpreted in the local minimization context.

Assumption 6.1. *At Step 1 of Algorithm 4.1, we obtain $x^k \in \mathbb{R}^n$ such that there exist $v^k \in \mathbb{R}^{\underline{m}}$ and $w^k \in \mathbb{R}_+^{\underline{p}}$ satisfying*

$$\|\nabla L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) + \nabla \underline{h}(x^k) v^k + \nabla \underline{g}(x^k) w^k\| \leq \varepsilon_k, \tag{6.2}$$

$$\|\underline{h}(x^k)\| \leq \varepsilon'_k, \text{ and } \|\min\{-\underline{g}(x^k), w^k\}\| \leq \varepsilon'_k, \tag{6.3}$$

where the sequence $\{\varepsilon_k\}$ is bounded and the sequence $\{\varepsilon'_k\}$ tends to zero.

Assumption 6.1 establishes a criterion to measure the degree of feasibility and optimality at the approximate solution of the subproblem

$$\text{Minimize } L_{\rho_k}(x, \bar{\lambda}^k, \bar{\mu}^k) \text{ subject to } x \in \Omega. \tag{6.4}$$

Sometimes it is useful to replace Assumption 6.1 with a condition that involves the projection of the gradient $\nabla L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)$ onto the tangent approximation of the lower feasible set Ω. Namely, the requirement for being an approximate solution of the subproblem is given in that case by

$$\|P_k(x^k - \nabla L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)) - x^k\| \leq \varepsilon_k, \tag{6.5}$$

$$\|\underline{h}(x^k)\| \leq \varepsilon'_k, \text{ and } \|\underline{g}(x^k)_+\| \leq \varepsilon'_k, \tag{6.6}$$

where the sequence $\{\varepsilon_k\}$ is bounded, the sequence $\{\varepsilon'_k\}$ tends to zero, and P_k represents the Euclidean projection operator onto T_k,

$$T_k = \{x \in \mathbb{R}^n \mid \nabla \underline{h}(x^k)^T (x - x^k) = 0 \text{ and } \underline{g}(x^k)_- + \nabla \underline{g}(x^k)^T (x - x^k) \leq 0\}. \tag{6.7}$$

The case in which $\lim_{k\to\infty} \varepsilon_k = 0$ corresponds to the AGP condition introduced by Martínez and Svaiter [191]. The geometrical interpretation for (6.5) is given in Figure 6.1. Note that, in the case that $x^k \in \Omega$ and $\underline{h}$ and $\underline{g}$ are affine functions, the subproblem constraints define a polytope that coincides with T_k. In particular, this is the case when Ω is a box (see Figure 6.2). It can be proved that (6.5), (6.6) imply (6.2), (6.3) (see problem 6.1).

It is interesting to interpret conditions (6.2) and (6.3) in the case in which the constraints of the subproblem define a box, i.e., in the case in which we have $\underline{m} = 0$, $\underline{p} = 2n$,

$$\underline{g}_i(x) = \ell_i - x_i \text{ and } \underline{g}_{n+i}(x) = x_i - u_i \text{ for all } i = 1, \ldots, n.$$

Methods that solve the subproblem (6.4) when the constraints define a box usually preserve feasibility of all the iterates. Therefore, we will have $\underline{g}(x^k) \leq 0$ for all k. Now, let us define, for $i = 1, \ldots, n$,

$$w_i^k = \max\left\{0, \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\} \text{ and } w_{n+i}^k = \max\left\{0, -\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\}.$$

With these definitions, condition (6.2) is trivially satisfied (even for $\varepsilon_k = 0$). Now, let us define, for $i = 1, \ldots, 2n$,

$$z_i^k = \min\{-\underline{g}_i(x^k), w_i^k\}.$$

By definition, if $\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) \geq 0$, we have that

$$w_i^k = \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k), w_{n+i}^k = 0, z_i^k = \min\left\{x_i^k - \ell_i, \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\},$$
$$\text{and } z_{n+i}^k = 0,$$

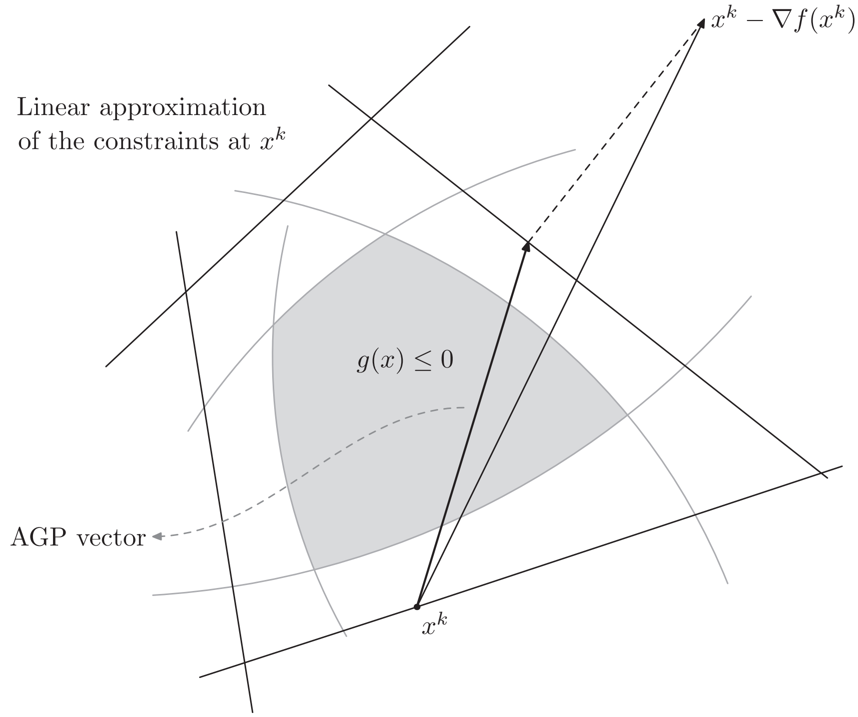

Figure 6.1. *The AGP vector tends to zero if x^k tends to a local minimizer.*

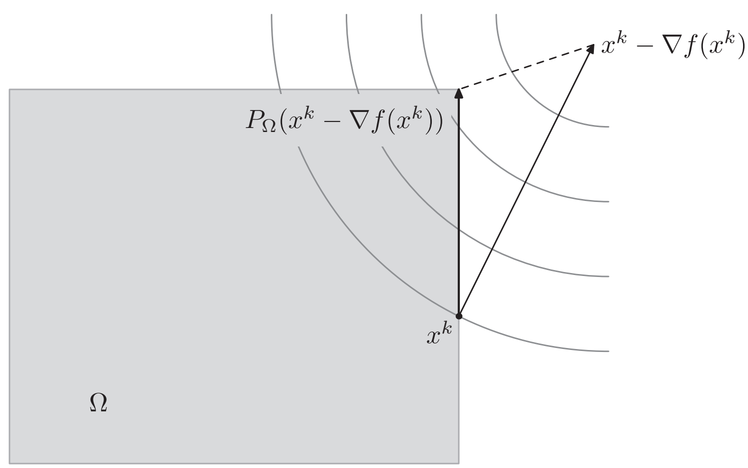

Figure 6.2. *Gradient projection onto a box.*

and, if $\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) < 0$, we have that

$$w_i^k = 0, w_{n+i}^k = -\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k), z_i^k = 0,$$

$$\text{and } z_{n+i}^k = \min\left\{u_i - x_i^k, -\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\}$$

for $i = 1, \ldots, n$.

Therefore, condition (6.3) imposes that $\|\hat{z}^k\| \le \varepsilon'_k$, where $\hat{z}^k \in \mathbb{R}^n$ is defined by

$$\hat{z}_i^k = \begin{cases} \min\left\{x_i^k - \ell_i, \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\} & \text{if } \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) \ge 0, \\ \min\left\{u_i - x_i^k, -\frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)\right\} & \text{if } \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) < 0 \end{cases}$$

for $i = 1, \dots, n$. Thus,

$$\hat{z}_i^k = \begin{cases} \left|\max\{x_i^k - \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k), \ell_i\} - x_i^k\right| & \text{if } \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) \ge 0, \\ \left|\min\{x_i^k - \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k), u_i\} - x_i^k\right| & \text{if } \frac{\partial}{\partial x_i} L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) < 0 \end{cases}$$

for $i = 1, \dots, n$. Therefore, $\|\hat{z}^k\|$ is the norm of $P_\Omega(x^k - \nabla L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k)) - x^k$, where P_Ω represents the projection onto the box Ω. This means that criterion (6.3) coincides with the one ordinarily used for box-constrained minimization, based on projected gradients.

The following theorem plays the role of Theorem 4.1 with respect to the multipliers associated with the subproblem inequality constraints $\underline{g}(x) \le 0$.

Theorem 6.1. *Assume that the sequence $\{x^k\}$ is generated by Algorithm 4.1 under Assumption 6.1, $\lim_{k\to\infty} \varepsilon_k = 0$, and $K \underset{\infty}{\subset} \mathbb{N}$ is such that $\lim_{k\in K} x^k = x^*$. Then, for $k \in K$ large enough there exists $\tilde{w}_k \in \mathbb{R}_+^{\underline{p}}$ such that*

$$\lim_{k\in K} \left\| \nabla L_{\rho_k}(x^k, \bar{\lambda}^k, \bar{\mu}^k) + \nabla \underline{h}(x^k) v^k + \nabla \underline{g}(x^k) \tilde{w}^k \right\| = 0 \tag{6.8}$$

and

$$\tilde{w}_i^k = 0 \text{ for all } i \in \{1, \dots, \underline{p}\} \text{ such that } g_i(x^*) < 0. \tag{6.9}$$

Moreover, $\tilde{w}_i^k = w_i^k$ for all $i \in \{1, \dots, \underline{p}\}$ such that $\underline{g}_i(x^*) \ge 0$.

Proof. Assume that $\underline{g}_i(x^*) < 0$. By (6.3), since $\min\{-\underline{g}_i(x^k), w_i^k\}$ tends to zero, we have that w_i^k tends to zero.

By the continuity of $\nabla \underline{g}_i$, this implies that

$$\lim_{k\in K} w_i^k \nabla \underline{g}_i(x^k) = 0$$

for all $i \in \{1, \dots, \underline{p}\}$ such that $\underline{g}_i(x^*) < 0$. Therefore, by (6.2) and $\varepsilon_k \to 0$, we have that (6.8) holds by taking

$$\tilde{w}_i^k = 0 \text{ if } \underline{g}_i(x^*) < 0 \text{ and } \tilde{w}_i^k = w_i^k \text{ if } \underline{g}_i(x^*) \ge 0$$

for $i = 1, \dots, \underline{p}$. This completes the proof. $\square$

In the next theorem, we prove that, when the algorithm analyzed in this chapter admits a feasible limit point, this point satisfies the optimality AKKT condition. In this case, the AKKT condition makes reference to all the constraints of the problem, not only those given by $h(x) = 0$ and $g(x) \le 0$. More precisely, in the case of the problem of minimizing $f(x)$ subject to $h(x) = 0$, $g(x) \le 0$, $\underline{h}(x) = 0$, and $\underline{g}(x) \le 0$, according to

Definition 3.1, we say that x^* satisfies the AKKT condition when there exist sequences $\{x^k\} \subseteq \mathbb{R}^n$, $\{\lambda^k\} \subseteq \mathbb{R}^m$, $\{\mu^k\} \subseteq \mathbb{R}^p_+$, $\{v^k\} \in \mathbb{R}^{\underline{m}}$, and $\{w^k\} \in \mathbb{R}^{\underline{p}}_+$ such that

$$\lim_{k\to\infty} x^k = x^*, \tag{6.10}$$

$$\lim_{k\to\infty} \left\| \nabla f(x^k) + \nabla h(x^k)\lambda^{k+1} + \nabla g(x^k)\mu^{k+1} + \nabla \underline{h}(x^k)v^k + \nabla \underline{g}(x^k)w^k \right\| = 0, \tag{6.11}$$

$$\lim_{k\to\infty} \|\min\{-g(x^k), \mu^{k+1}\}\| = 0, \tag{6.12}$$

and

$$\lim_{k\to\infty} \|\min\{-\underline{g}(x^k), w^k\}\| = 0. \tag{6.13}$$

Of course, we can formally state conditions (6.11) and (6.12) writing λ^k and μ^k instead of λ^{k+1} and μ^{k+1}, respectively. We prefer to write λ^{k+1} and μ^{k+1} here in order to stress the relation with the notation adopted in (4.7) and (4.8).

Theorem 6.2. *Assume that the sequence $\{x^k\}$ is generated by Algorithm 4.1 with Assumption 6.1 for the minimization of $f(x)$ subject to $h(x) = 0, g(x) \le 0, \underline{h}(x) = 0$, and $\underline{g}(x) \le 0$ and $K \underset{\infty}{\subset} \mathbb{N}$ is such that $\lim_{k\in K} x^k = x^*$ and x^* is feasible. Moreover, assume that the bounded sequence $\{\varepsilon_k\}$ in Assumption 6.1 is such that $\lim_{k\in K} \varepsilon_k = 0$. Then, x^* satisfies the AKKT conditions for the optimization problem.*

Proof. By (6.2) and straightforward calculations using the definitions (4.7) and (4.8) of λ^{k+1} and μ^{k+1}, we have that

$$\lim_{k\in K} \left\| \nabla f(x^k) + \nabla h(x^k)\lambda^{k+1} + \nabla g(x^k)\mu^{k+1} + \nabla \underline{h}(x^k)v^k + \nabla \underline{g}(x^k)w^k \right\| = 0. \tag{6.14}$$

Moreover, by Theorem 4.1, we have that $\lim_{k\in K} \|\min\{-g(x^k), \mu^{k+1}\}\| = 0$. Therefore, by the feasibility of x^* and (6.3), it turns out that x^* is an AKKT point, as we wanted to prove. □

Theorem 6.2 induces a natural stopping criterion for Algorithm 4.1 under Assumption 6.1. Given $\varepsilon > 0$, it is sensible to stop (declaring success) when

$$\left\| \nabla f(x^k) + \nabla h(x^k)\lambda^{k+1} + \nabla g(x^k)\mu^{k+1} + \nabla \underline{h}(x^k)v^k + \nabla \underline{g}(x^k)w^k \right\| \le \varepsilon, \tag{6.15}$$

$$\|h(x^k)\| \le \varepsilon, \|\min\{-g(x^{k+1}), \mu^{k+1}\}\| \le \varepsilon, \tag{6.16}$$

$$\|\underline{h}(x^k)\| \le \varepsilon, \text{ and } \|\min\{-\underline{g}(x^{k+1}), w^k\}\| \le \varepsilon. \tag{6.17}$$

Of course, different tolerances may be used in (6.15), (6.16), and (6.17).

Corollary 6.1. *Under the assumptions of Theorem 6.2, if x^* is a feasible limit point of a sequence generated by Algorithm 4.1, and x^* fulfills the CPLD constraint qualification, then x^* is a KKT point of the problem.*

Proof. The proof is a consequence of Theorems 3.6 and 6.2. □

Constrained optimization algorithms have two goals: finding feasible points and minimizing the objective function subject to feasibility. The behavior of algorithms with respect to feasibility thus demands independent study. Employing global optimization techniques, we saw in Chapter 5 that one necessarily finds global minimizers of the infeasibility, a property that cannot be guaranteed using affordable local optimization procedures. In the next theorem, we prove that, by means of Algorithm 4.1 under Assumption 6.1, we necessarily find stationary points of the sum of squares of infeasibilities. The reader will observe that we do not need $\varepsilon_k \to 0$ for proving this important property.

Theorem 6.3. *Assume that the sequence $\{x^k\}$ is obtained by Algorithm 4.1 under Assumption 6.1. Let x^* be a limit point of $\{x^k\}$. Then, x^* satisfies the AKKT condition of the problem*

$$\text{Minimize } \|h(x)\|_2^2 + \|g(x)_+\|_2^2 \text{ subject to } \underline{h}(x) = 0, \underline{g}(x) \le 0. \tag{6.18}$$

Proof. Since $\underline{h}$ and $\underline{g}$ are continuous and, by Assumption 6.1, $\lim_{k\to\infty} \varepsilon'_k = 0$, we have that $\underline{h}(x^*) = 0$ and $\underline{g}(x^*) \le 0$.

If the sequence $\{\rho_k\}$ is bounded, we have by (4.9), that $\lim_{k\to\infty} \|h(x^k)\| = \lim_{k\to\infty} \|g(x^k)_+\| = 0$. Thus, the gradient of the objective function of (6.18) vanishes. This implies that KKT (and, hence, AKKT) holds with null Lagrange multipliers corresponding to the constraints.

Let us consider the case in which ρ_k tends to infinity. Defining

$$\begin{aligned}\delta^k = \nabla f(x^k) &+ \sum_{i=1}^m (\bar\lambda_i^k + \rho_k h_i(x^k)) \nabla h_i(x^k) + \sum_{i=1}^p \max\{0, \bar\mu_i^k + \rho_k g_i(x^k)\} \nabla g_i(x^k) \\ &+ \sum_{i=1}^m v_i^k \nabla \underline{h}_i(x^k) + \sum_{i=1}^p w_i^k \nabla \underline{g}_i(x^k)\end{aligned} \tag{6.19}$$

by (6.2) and the fact that $\{\varepsilon_k\}$ is bounded, we have that $\{\|\delta^k\|\}$ is bounded too.

Let $K \underset{\infty}{\subset} \mathbb{N}$ be such that $\lim_{k\in K} x^k = x^*$. By Theorem 6.1, we may assume, without loss of generality, that $w_i^k = 0$ for all $i \in \{1,\dots,p\}$ such that $\underline{g}_i(x^*) < 0$. Therefore, for all $k \in K$, we have that

$$\begin{aligned}\delta^k = \nabla f(x^k) &+ \sum_{i=1}^m (\bar\lambda_i^k + \rho_k h_i(x^k)) \nabla h_i(x^k) + \sum_{i=1}^p \max\{0, \bar\mu_i^k + \rho_k g_i(x^k)\} \nabla g_i(x^k) \\ &+ \sum_{i=1}^{\underline{m}} v_i^k \nabla \underline{h}_i(x^k) + \sum_{\underline{g}_i(x^*)=0} w_i^k \nabla \underline{g}_i(x^k).\end{aligned}$$

Dividing by ρ_k, we obtain

$$\begin{aligned}\frac{\delta^k}{\rho_k} = \frac{1}{\rho_k}\nabla f(x^k) &+ \sum_{i=1}^m \left(\frac{\bar\lambda_i^k}{\rho_k} + h_i(x^k)\right) \nabla h_i(x^k) + \sum_{i=1}^p \max\left\{0, \frac{\bar\mu_i^k}{\rho_k} + g_i(x^k)\right\} \nabla g_i(x^k) \\ &+ \sum_{i=1}^{\underline{m}} \frac{v_i^k}{\rho_k} \nabla \underline{h}_i(x^k) + \sum_{\underline{g}_i(x^*)=0} \frac{w_i^k}{\rho_k} \nabla \underline{g}_i(x^k)\end{aligned}$$

and, since $\{\|\delta^k\|\}$ is bounded and ρ_k tends to infinity, we have that $\delta_k/\rho_k \to 0$.

If $g_i(x^*) < 0$, since $\{\bar\mu^k\}$ is bounded and ρ_k tends to infinity, we have that $\max\{0, \bar\mu_i^k/\rho_k + g_i(x^k)\} = 0$ for $k \in K$ large enough. Therefore, by the boundedness of $\{\nabla f(x^k)\}$ and

$\{\bar{\lambda}^k\}$, $\delta_k/\rho_k \to 0$ implies that

$$\lim_{k\in K}\left\|\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)\geq 0} \max\left\{0, \frac{\bar{\mu}_i^k}{\rho_k} + g_i(x^k)\right\}\nabla g_i(x^k) \right.$$
$$\left. + \sum_{i=1}^{m} \frac{v_i^k}{\rho_k}\nabla \underline{h}_i(x^k) + \sum_{\underline{g}_j(x^*)=0} \frac{w_j^k}{\rho_k}\nabla \underline{g}_j(x^k)\right\| = 0. \tag{6.20}$$

If $g_i(x^*) = 0$, we clearly have that, by the boundedness of $\{\bar{\mu}^k\}$ and $\{\nabla g(x^k)\}$,

$$\lim_{k\in K} \max\left\{0, \frac{\bar{\mu}_i^k}{\rho_k} + g_i(x^k)\right\}\nabla g_i(x^k) = 0.$$

Then, by (6.20), we have that

$$\lim_{k\in K}\left\|\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)>0} \max\left\{0, \frac{\bar{\mu}_i^k}{\rho_k} + g_i(x^k)\right\}\nabla g_i(x^k) \right.$$
$$\left. + \sum_{i=1}^{m} \frac{v_i^k}{\rho_k}\nabla \underline{h}_i(x^k) + \sum_{\underline{g}_j(x^*)=0} \frac{w_j^k}{\rho_k}\nabla \underline{g}_j(x^k)\right\| = 0.$$

But, if $g_i(x^*) > 0$ and $k \in K$ is large enough, we have that

$$\max\left\{0, \frac{\bar{\mu}_i^k}{\rho_k} + g_i(x^k)\right\} = g_i(x^k) + \frac{\bar{\mu}_i^k}{\rho_k};$$

thus

$$\lim_{k\in K}\left\|\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)>0} \left(\frac{\bar{\mu}_i^k}{\rho_k} + g_i(x^k)\right)\nabla g_i(x^k) \right.$$
$$\left. + \sum_{i=1}^{m} \frac{v_i^k}{\rho_k}\nabla \underline{h}_i(x^k) + \sum_{\underline{g}_j(x^*)=0} \frac{w_j^k}{\rho_k}\nabla \underline{g}_j(x^k)\right\| = 0.$$

Therefore, since $\bar{\mu}_i^k/\rho_k \to 0$, we have

$$\lim_{k\in K}\left\|\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)>0} g_i(x^k)\nabla g_i(x^k) \right.$$
$$\left. + \sum_{i=1}^{m} \frac{v_i^k}{\rho_k}\nabla \underline{h}_i(x^k) + \sum_{\underline{g}_j(x^*)=0} \frac{w_j^k}{\rho_k}\nabla \underline{g}_j(x^k)\right\| = 0.$$

This obviously implies that

$$\lim_{k\in K}\left\|\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)\geq 0} g_i(x^k)\nabla g_i(x^k) \right.$$
$$\left. + \sum_{i=1}^{m} \frac{v_i^k}{\rho_k}\nabla \underline{h}_i(x^k) + \sum_{\underline{g}_j(x^*)=0} \frac{w_j^k}{\rho_k}\nabla \underline{g}_j(x^k)\right\| = 0. \tag{6.21}$$

But

$$\nabla[\|h(x^k)\|_2^2 + \|g(x^k)_+\|_2^2] = 2\left[\sum_{i=1}^{m} h_i(x^k)\nabla h_i(x^k) + \sum_{g_i(x^*)\geq 0} g_i(x^k)\nabla g_i(x^k)\right].$$

Therefore, by (6.21), the limit point x^* satisfies the AKKT condition of (6.18). $\square$

Corollary 6.2. *Every limit point generated by Algorithm 4.1 under Assumption 6.1 at which the constraints $\underline{h}(x) = 0$, $\underline{g}(x) \leq 0$ satisfy the CPLD constraint qualification is a KKT point of the problem of minimizing the infeasibility $\|h(x)\|_2^2 + \|g(x)_+\|_2^2$ subject to $\underline{h}(x) = 0$ and $\underline{g}(x) \leq 0$.*

6.1 ▪ The algorithm with abstract constraints

We will finish this chapter by considering the case in which the lower-level set Ω, instead of being defined by $\underline{h}(x) = 0$ and $\underline{g}(x) \leq 0$, is an arbitrary closed and convex set, possibly without an obvious representation in terms of equalities and inequalities. In this case, it may be convenient to define Step 1 of Algorithm 4.1 in a different way than that presented in Assumption 6.1. Assumption 6.2 below gives the appropriate definition in this case.

Assumption 6.2. *At Step 1 of Algorithm 4.1, we obtain $x^k \in \Omega$ such that*

$$\left\| P_\Omega(x^k - \nabla L_{\rho_k}(x^k, \bar\lambda^k, \bar\mu^k)) - x^k \right\| \leq \varepsilon_k, \tag{6.22}$$

where the sequence $\{\varepsilon_k\}$ tends to zero.

Theorem 6.4. *Assume that the sequence $\{x^k\}$ is generated by Algorithm 4.1 under Assumption 6.2 for the minimization of $f(x)$ subject to $h(x) = 0$, $g(x) \leq 0$, and $x \in \Omega$, with Ω closed and convex. Assume that $K \underset{\infty}{\subset} \mathbb{N}$ is such that $\lim_{k\in K} x^k = x^*$ and x^* is feasible. Then,*

$$\lim_{k\in K} \left\| P_\Omega\left[x^k - \left(\nabla f(x^k) + \nabla h(x^k)\lambda^{k+1} + \nabla g(x^k)\mu^{k+1}\right)\right] - x^k \right\| = 0 \tag{6.23}$$

and

$$\lim_{k\in K} \min\{-g_i(x^k), \mu_{i+1}^k\} = 0 \text{ for all } i = 1, \ldots, p. \tag{6.24}$$

Proof. By (6.22) and straightforward calculations using the definitions of λ^{k+1} and μ^{k+1}, we have that (6.23) holds. Moreover, by Theorem 4.1, we have that

$$\lim_{k\in K} \|\min\{-g(x^k), \mu^{k+1}\}\| = 0.$$

This completes the proof. $\square$

Theorem 6.4 provides another useful stopping criterion. Given $\varepsilon > 0$, we stop declaring success when

$$\left\| P_\Omega\left[x^k - \left(\nabla f(x^k) + \nabla h(x^k)\lambda^{k+1} + \nabla g(x^k)\mu^{k+1}\right)\right] - x^k \right\| \leq \varepsilon, \tag{6.25}$$

$$\|\min\{-g(x^k), \mu^{k+1}\}\| \leq \varepsilon, \tag{6.26}$$

and

$$\|h(x^k)\| \le \varepsilon. \tag{6.27}$$

This is the criterion usually employed in practice when the lower-level constraints have the box form $\ell \le x \le u$.

It remains to prove the feasibility result that corresponds to Theorem 6.3 in the case that Ω is an "abstract" closed and convex set. In analogy to Theorem 6.3, Theorem 6.5 below shows that the algorithm makes the best possible work in the process of trying to find feasible points.

Theorem 6.5. *Assume that the sequence $\{x^k\}$ is obtained by Algorithm 4.1 under Assumption 6.2. Assume that x^* is a limit point of $\{x^k\}$. Then, x^* is a stationary point (in the sense of (3.31)) of*

$$\text{Minimize } \|h(x)\|_2^2 + \|g(x)_+\|_2^2 \text{ subject to } x \in \Omega. \tag{6.28}$$

Proof. If the sequence $\{\rho_k\}$ is bounded, the desired result follows as in Theorem 6.3.

Assume that $\rho_k \to \infty$ and let $K \underset{\infty}{\subset} \mathbb{N}$ be such that $\lim_{k\in K} x^k = x^*$. By (6.22), for all $k \in K$, we have that

$$\begin{aligned} \Bigg\| P_\Omega\Bigg(x^k - \Bigg(\nabla f(x^k) + \rho_k \Bigg[\nabla h(x^k)\left(h(x^k) + \frac{\bar\lambda^k}{\rho_k}\right) \\ + \nabla g(x^k)\left(g(x^k) + \frac{\bar\mu^k}{\rho_k}\right)_+ \Bigg]\Bigg)\Bigg) - x^k \Bigg\| \le \varepsilon_k. \end{aligned} \tag{6.29}$$

Since $\rho_k \to \infty$, we have that $1/\rho_k < 1$ for k large enough. Therefore, by (3.29), (6.29) implies

$$\begin{aligned} \Bigg\| P_\Omega\Bigg(x^k - \Bigg(\nabla f(x^k)/\rho_k + \nabla h(x^k)\left(h(x^k) + \frac{\bar\lambda^k}{\rho_k}\right) \\ + \nabla g(x^k)\left(g(x^k) + \frac{\bar\mu^k}{\rho_k}\right)_+ \Bigg)\Bigg) - x^k \Bigg\| \le \varepsilon_k \end{aligned} \tag{6.30}$$

for $k \in K$ large enough. Since $\rho_k \to \infty$ and $\{\bar\mu^k\}$ is bounded, we have that for $k \in K$ large enough, $(g_i(x^k) + \bar\mu_i^k/\rho_k)_+ = 0$ whenever $g_i(x^*) < 0$. Therefore, by (6.30),

$$\begin{aligned} \Bigg\| P_\Omega\Bigg(x^k - \Bigg(\nabla f(x^k)/\rho_k + \nabla h(x^k)\left(h(x^k) + \frac{\bar\lambda^k}{\rho_k}\right) \\ + \sum_{g_i(x^*)\ge 0} \nabla g_i(x^k)\left(g_i(x^k) + \frac{\bar\mu_i^k}{\rho_k}\right)_+ \Bigg)\Bigg) - x^k \Bigg\| \le \varepsilon_k \end{aligned} \tag{6.31}$$

for $k \in K$ large enough. By the boundedness of $\{\bar\lambda^k\}$, the continuity of ∇f, ∇h, ∇g, and P_Ω, and, consequently, the uniform continuity of these functions on a compact set that contains the points x^k for all $k \in K$, since $\{\varepsilon_k\}$ tends to zero, (6.31) implies

$$\lim_{k\in K} \Bigg\| P_\Omega\Bigg(x^k - \Bigg[\nabla h(x^k) h(x^k) + \sum_{g_i(x^*)\ge 0} \nabla g_i(x^k) g_i(x^k)_+ \Bigg]\Bigg) - x^k \Bigg\| = 0. \tag{6.32}$$

This implies the thesis of the theorem. ∎

6.2 ▪ Review and summary

In this chapter, we considered the model Algorithm 4.1, where the approximate solution of the subproblem is interpreted as the approximate fulfillment of its KKT condition. In this way, we may use a standard affordable solver for solving the subproblems. According to the theoretical results, if the sequence generated by the algorithm converges to a feasible point, this point satisfies the AKKT condition. Moreover, some iterate satisfies, up to any arbitrarily given precision, the KKT condition of the original problem. However, since the original problem may be infeasible, it is useful to show that the limit points of sequences generated by Algorithm 4.1 with the assumptions of this chapter are stationary points of the infeasibility measure (probably local or even global minimizers of infeasibility). Assumptions 6.1 and 6.2 represent different instances of the main algorithm, corresponding to different definitions of the subproblem constraints.

6.3 ▪ Further reading

Using an additional smoothness (generalized Kurdyka–Lojasiewicz) condition, the fulfillment of a stronger sequential optimality condition by the Augmented Lagrangian method was proved by Andreani, Martínez, and Svaiter [18]. The CAKKT defined in [18] states, in addition to the usual AKKT requirements, that the products between multipliers and constraint values tend to zero. CAKKT is strictly stronger than AKKT. Since stopping criteria based on sequential optimality conditions are natural for every constrained optimization algorithm, the question arises of whether other optimization algorithms generate sequences that satisfy AKKT. Counter-examples and results in [15] indicate that for algorithms based on sequential quadratic programming the answer is negative. The behavior of optimization algorithms in situations where Lagrange multipliers do not exist at all is a subject of current research (see [15]).

6.4 ▪ Problems

6.1 Prove that (6.5), (6.6) implies (6.2), (6.3).

6.2 Work on the calculations to prove (6.14) using (6.2) and the definitions (4.7) and (4.8) of λ^{k+1} and μ^{k+1}.

6.3 Formulate Algorithm 4.1 with the assumptions of this chapter for the case in which Ω is a box. Suggest a projected gradient criterion for deciding to stop the iterative subproblem solver.

6.4 Observing that, for proving Theorem 6.3, it is not necessary to assume that $\varepsilon_k \to 0$, define a version of Algorithm 4.1 in which ε_k is a function of the infeasibility measure $\|h(x^k)\| + \|g(x^k)_+\|$ (see Martínez and Prudente [188]).

6.5 Formulate Algorithm 4.1 with the assumptions of this chapter for the case in which $p = 0$ and $\Omega = \mathbb{R}^n$. Observe that, if $\varepsilon_k = 0$, the stopping condition for the subproblem is a nonlinear system of equations. Formulate Newton's method for this system and identify causes for ill-conditioning of the Newtonian linear system when ρ_k is large. Decompose the system in order to eliminate the ill-conditioning.

6.6 Suppose that your constrained optimization problem is feasible and that the only stationary points of the infeasibility are the feasible points. Assume that the sequence $\{x^k\}$ is obtained by Algorithm 4.1 under Assumption 6.1. Assume that $K \underset{\infty}{\subset} \mathbb{N}$ is such that $\lim_{k\in K} x^k = x^*$. Discuss the following arguments:

(a) By Theorem 6.3, x^* feasible.

(b) By (a) and Theorem 4.1, $\lim_{k\in K} \max\{\|h(x^k)\|, \|V_k\|\} = 0$.

(c) As a consequence of (a), (b), and (4.9), $\{\rho_k\}$ remains bounded.

(d) Hence, unboundedness of the penalty parameters occurs only in the case of infeasibility.

Transform these (wrong) arguments in a topic of research.

6.7 Prove (6.23).

6.8 Complete the details of the proof of Theorem 6.5.

6.9 In many nonlinear optimization problems, restoring feasibility is easy because there exists a problem-oriented procedure for finding feasible points efficiently. This possibility can be exploited within the Augmented Lagrangian framework in the choice of the initial point for solving the subproblems. Namely, one can choose that initial approximation as the result of approximately restoring feasibility starting from the Augmented Lagrangian iterate x^{k-1}. In order to take advantage of this procedure, the penalty parameter should be chosen in such a way that the Augmented Lagrangian function decreases at the restored point with respect to its value at x^{k-1}. Define carefully this algorithm and check the convergence theory. (This idea approximates the Augmented Lagrangian framework to the so-called inexact restoration methods and other feasible methods for constrained optimization [1, 53, 111, 129, 159, 172, 185, 187, 197, 198, 199, 229, 230, 231].)

6.10 It is numerically more attractive to solve (4.6) dividing the Augmented Lagrangian by ρ_k, at least when this penalty parameter is large. Why? Formulate the main algorithms in this form and modify the stopping criterion of the subproblems consequently.

6.11 The process of solving (4.6) with a very small tolerance for convergence can be painful. Suggest alternative practical stopping criteria for the subproblems in accordance (or not) with the theory. (Hints: Relative difference between consecutive iterates and lack of progress during some iterations.) Discuss the theoretical and practical impact of the suggested modifications.

6.12 In the context of the problem above, suggest a stopping criterion for the subproblems that depends on the best feasibility-complementarity achieved at previous outer iterations. Note that it may not be worthwhile to solve subproblems with great precision if we are far from optimality.

6.13 In the case of convergence to a nonfeasible point, the penalty parameter goes to infinity and, consequently, it is harder and harder to solve the subproblems (4.6) with the given stopping criterion. Why? However, probable infeasibility can be detected evaluating optimality conditions of the sum of squares of infeasibities subject to the lower-level constraints. Add this test to Algorithm 4.1. Discuss the possible effect of this modification in numerical tests. Note that you need a test of stationarity for the infeasibility measure with relative big value of the sum of squares (two tolerances are involved).

6.14 Study classical acceleration procedures for sequences in $\mathbb{R}^n$ and apply these procedures to the choice of initial points for solving the Augmented Lagrangian subproblems. Hint: See Brezinski and Zaglia [70] and the DIIS method of Pulay [218].

6.15 Study different ways of exploiting parallelization in Augmented Lagrangian algorithms, for example, using different initial points in parallel at the solution of the subproblems or solving subproblems with different penalty parameters simultaneously.