TOP manuscript No.
(will be inserted by the editor)

Critical Lagrange multipliers: what we currently know about them, how they spoil our lives, and what we can do about it

A. F. Izmailov · M. V. Solodov

Received: date / Accepted: date

Abstract We discuss a certain special subset of Lagrange multipliers, called critical, which usually exist when multipliers associated to a given solution are not unique. This kind of multipliers appear to be important for a number of reasons, some understood better, some (currently) not fully. What is clear, is that Newton and Newton-related methods have an amazingly strong tendency to generate sequences with dual components converging to critical multipliers. This is quite striking because, typically, the set of critical multipliers is "thin" (the set of noncritical ones is relatively open and dense, meaning that its closure is the whole set). Apart from mathematical curiosity to understand the phenomenon for something as classical as the Newton method, the attraction to critical multipliers is relevant computationally. This is because convergence to such multipliers is the reason for slow convergence of the Newton method in degenerate cases, as convergence to noncritical limits (if it were to happen) would have given the superlinear rate. Moreover, the attraction phenomenon shows up not only for the basic Newton method, but also for other related techniques (for example, quasi-Newton, and the linearly-constrained augmented Lagrangian method). In spite of clear computational evidence, proving that convergence to a critical limit must occur appears to be a challenge, at least

Research of the first author is supported by the Russian Foundation for Basic Research Grant 14-01-00113 and by CNPq Grant PVE 401119/2014-9 (Brazil). The second author is supported in part by CNPq Grant 302637/2011-7, by PRONEX–Optimization, and by FAPERJ.

A. F. Izmailov
Operations Research Department, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (MSU), Uchebniy Korpus 2, Leninskiye Gory, 119991 Moscow, Russia
E-mail: izmaf@ccas.ru

M. V. Solodov
IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil
E-mail: solodov@impa.br

for general problems. We outline the partial results obtained up to now. We also discuss the important role that noncritical multipliers play for stability, sensitivity, and error bounds. Finally, an important issue is dual stabilization, i.e., techniques to avoid moving along the multiplier set towards a critical one (since it leads to slow convergence). We discuss the algorithms that do the job locally, i.e., when initialized close enough to a noncritical multiplier, their dual behavior is as desired. These include the stabilized sequential quadratic programming method and the augmented Lagrangian algorithm. However, when the starting point is far, even those algorithms do not appear to provide fully satisfactory remedies. We discuss the challenges with constructing good algorithms for the degenerate case, which have to incorporate dual stabilization for fast local convergence, at an acceptable computational cost, and also be globally efficient.

Keywords Critical Lagrange multipliers · Second-order sufficiency · Newton-type methods · Sequential quadratic programming · Newton-Lagrange method · Superlinear convergence

Mathematics Subject Classification (2010) 90C30 · 90C33 · 65K05

1 Introduction

In this exposition, we shall restrict our attention to the equality-constrained optimization problem

$$\text{minimize } f(x) \text{ subject to } h(x) = 0, \tag{1}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraints mapping $h : \mathbb{R}^n \to \mathbb{R}^l$ are at least twice differentiable, with their second derivatives being continuous at the point of interest. The issues discussed below extend (at least partially, sometimes completely) in various directions: to optimization problems involving inequality constraints; to more general variational problems of which optimization is a special case; to problems with weaker smoothness requirements. Also, while we shall refer for illustration to the most basic form of the Newton method for (1), it is important to mention that the conceptual conclusions apply also to its various important modifications [32], and sometimes even to methods which do not look clearly "Newtonian" [35]. We keep the focus here on sufficiently smooth equality-constrained optimization and the basic Newton method, in order to avoid technical details and branching of the exposition, and to make sure we transmit the main ideas with reasonable brevity.

We start by recalling some classical terminology. Let $L : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ be the usual Lagrangian of problem (1), i.e.,

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product (the space would always be clear from the context). Then stationary points and associated Lagrange multipliers of problem (1) are characterized by the Lagrange optimality system

$$\frac{\partial L}{\partial x}(x,\, \lambda) = 0, \quad h(x) = 0, \tag{2}$$

with respect to $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^l$. Let $\mathcal{M}(\bar{x})$ stand for the set of Lagrange multipliers associated to a stationary point $\bar{x}$ of problem (1), i.e.,

$$\mathcal{M}(\bar{x}) = \left\{ \lambda \in \mathbb{R}^l \;\middle|\; \frac{\partial L}{\partial x}(\bar{x},\, \lambda) = 0 \right\}.$$

As is well known, if $\bar{x}$ is a local solution of problem (1), and the constraints regularity condition

$$\operatorname{rank} h'(\bar{x}) = l \tag{3}$$

holds, then $\mathcal{M}(\bar{x})$ is a singleton. When the multiplier set $\mathcal{M}(\bar{x})$ is nonempty, but (3) is violated, $\mathcal{M}(\bar{x})$ is an affine manifold of some positive dimension. For such (degenerate, difficult) situations of nonisolated solutions, the goal of this article is to discuss a certain special subclass of Lagrange multipliers with very interesting properties. The multipliers in question are called "critical"; see the next section for the definition. The relevance of this subcluss is (at least) two-fold. First, Lagrange multipliers of this kind tend to attract dual sequences of a good number of important optimization algorithms, and this can be seen to be the reason for their slow convergence in the degenerate cases. Second, multipliers that *do not* belong to this special subclass play a key role in various theoretical developments; in particular, concerned with stability and sensitivity issues, and with error bounds. Some facts of this kind will be mentioned below, but we shall mostly concentrate here on the (negative) influence of critical multipliers on computational methods. We shall discuss our current understanding of the nature of the destructive phenomenon of attraction to critical multipliers, outline the possible ways out of this unpleasant situation, and indicate what still needs to be done to construct more satisfactory algorithms capable to handle degenerate problems.

In what follows, I stands for the identity matrix of an appropriate size. All vector norms are Euclidean; all matrix norms are spectral (i.e., induced by Euclidean). By $\mathbb{S}^n$ we denote the space of $n \times n$ symmetric matrices.

2 Critical and noncritical Lagrange multipliers: definitions and some basic properties

A Lagrange multiplier $\bar{\lambda} \in \mathbb{R}^l$ associated to a stationary point $\bar{x}$ is called *critical* if there

$$\text{exists } \xi \in \ker h'(\bar{x}) \setminus \{0\} \text{ such that } \frac{\partial^2 L}{\partial x^2}(\bar{x},\, \bar{\lambda})\xi \in \operatorname{im}(h'(\bar{x}))^{\mathrm{T}},$$

and *noncritical* otherwise. In other words, $\bar{\lambda}$ is critical if the corresponding reduced Hessian of the Lagrangian (i.e., the symmetric matrix $H(\bar{\lambda}) = H(\bar{x}, \bar{\lambda})$ of the quadratic form $\xi \to \langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi\rangle : \ker h'(\bar{x}) \to \mathbb{R}$) is singular. This notion was introduced in [23]; its theoretical and computational implications are further studied in [30,31,10,32,33,28,39]; see also [34].

Observing that $\mathrm{im}(h'(\bar{x}))^{\mathrm{T}} = (\ker h'(\bar{x}))^{\perp}$, it is immediate that the multiplier is always noncritical if it satisfies the second-order sufficient optimality condition (SOSC)

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi \right\rangle > 0 \quad \forall\, \xi \in \ker h'(\bar{x}) \setminus \{0\}, \tag{4}$$

or the corresponding condition for maximizers (i.e., with the reverse sign). If $\bar{\lambda}$ satisfies the condition

$$\left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda})\xi, \xi \right\rangle \geq 0 \quad \forall\, \xi \in \ker h'(\bar{x}) \setminus \{0\} \tag{5}$$

then it is noncritical if and only if it satisfies the SOSC (4). It is important to emphasize, however, that (5) is *not necessary* for local optimality of $\bar{x}$ in the absence of the regularity assumption (3). Thus, in the degenerate context, noncriticality and (4) are *not* equivalent even for minimizers. In fact, the common situation is that there are many more noncritical multipliers than those that satisfy the SOSC (4), even counting also the ones with the reverse sign if such exist, and that critical multipliers are "very few" within the set $\mathcal{M}(\bar{x})$. This will be discussed again and illustrated by examples further below.

Even though of interest in this discussion are degenerate problems when $\mathcal{M}(\bar{x})$ is not a singleton, let us start with the following well-understood property clarifying the role of noncritical multipliers in the regular/nondegenerate case.

Proposition 1 *For any stationary point $\bar{x}$ of problem* (1) *and any associated Lagrange multiplier $\bar{\lambda}$, the Jacobian of the Lagrange system* (2) *(or in other words, the full Hessian of the Lagrangian) is nonsingular at $(\bar{x}, \bar{\lambda})$ if and only if the regularity condition* (3) *holds (implying that $\bar{\lambda}$ is unique) and $\bar{\lambda}$ is noncritical.*

Consider now the Newton–Lagrange method (NLM), that is, the Newton method applied to the Lagrange system (2). Specifically, for a current primal-dual iterate $(x^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$, the next iterate (x^{k+1}, λ^{k+1}) is defined by the linear system

$$\begin{gathered} \frac{\partial^2 L}{\partial x^2}(x^k, \lambda^k)(x - x^k) + (h'(x^k))^{\mathrm{T}}(\lambda - \lambda^k) = -\frac{\partial L}{\partial x}(x^k, \lambda^k), \\ h'(x^k)(x - x^k) = -h(x^k). \end{gathered} \tag{6}$$

As is well known, this is also a way to write the iteration of the sequential quadratic programming method (SQP) for the optimization problem (1); see,

e.g., [34, Chapter 4]. An iteration of the latter consists in solving the subproblem

$$\begin{array}{l}\text{minimize} \quad \langle f'(x^k),\, x-x^k\rangle + \dfrac{1}{2}\left\langle \dfrac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)(x-x^k),\, x-x^k\right\rangle \\ \text{subject to } h(x^k) + h'(x^k)(x-x^k) = 0.\end{array}$$

As is easily seen, stationary points of this subproblem and the associated Lagrange multipliers are characterized precisely by the linear system (6).

According to Proposition 1, local superlinear convergence of NLM/SQP to a solution $(\bar{x},\, \bar{\lambda})$ of system (2) is guaranteed if the regularity condition (3) holds and the multiplier $\bar{\lambda}$ is noncritical. Moreover, when the regularity condition (3) holds, the unique $\bar{\lambda} \in \mathcal{M}(\bar{x})$ is typically noncritical (at least, there is no particular reason why this should not be so).

At this point, we discard the regularity condition (3) and, thus, enter the territory where the multiplier set $\mathcal{M}(\bar{x})$ is a nontrivial affine manifold.

The first (obvious) issue to comment is that now an algorithm (any primal-dual algorithm) applied to solve (1) has (infinitely) many "correct dual targets". Even if the algorithm appears to be successfully solving the problem, i.e., the primal sequence converges to some degenerate solution $\bar{x}$ and the dual sequence approaches $\mathcal{M}(\bar{x})$ which is not a singleton, how the dual sequence behaves is clearly of potential influence. It turns out that the latter is actually crucial, and directly affects the speed of convergence, including that of the primal sequence.

The second issue to mention is that $\mathcal{M}(\bar{x})$ may now very well consist of "various kinds" of multipliers; in particular, it may contain critical multipliers, and in a stable way: they usually do not disappear after small perturbations preserving the deficient rank of the constraints' Jacobian. At the same time, it is important to emphasize that the set of critical multipliers is usually "thin" within $\mathcal{M}(\bar{x})$. This follows from the fact that the set of critical multipliers is characterized by the algebraic equation

$$\det H(\lambda) = 0 \tag{7}$$

over the affine set $\mathcal{M}(\bar{x})$. Thus, if there exists some $\lambda \in \mathcal{M}(\bar{x})$ violating this equation (i.e., there exists a noncritical multiplier), then this is the case for all λ in some relatively open and dense subset of $\mathcal{M}(\bar{x})$. In other words (and somewhat informally), typically "almost all" the multipliers are noncritical, while critical multipliers are "few". This will be seen from the examples discussed below again and again.

One characteristic property of a noncritical multiplier is the primal-dual Lipschitzian error bound estimating the distance to the solution set of the Lagrange system (2), or equivalently, the upper-Lipschitzian behavior of this solution set under canonical perturbations. The related results for much more general problem settings can be found in [13,21,25,28]; see also [34, Chapter 1.3].

Proposition 2 *For any stationary point $\bar{x}$ of problem* (1) *and any associated Lagrange multiplier $\bar{\lambda}$, the following three properties are equivalent:*

(a) The multiplier $\bar{\lambda}$ is noncritical.
(b) The error bound

$$\|x-\bar{x}\|+\operatorname{dist}(\lambda,\,\mathcal{M}(\bar{x}))=O\left(\left\|\left(\frac{\partial L}{\partial x}(x,\,\lambda),\,h(x)\right)\right\|\right) \tag{8}$$

holds as $(x,\,\lambda)\in\mathbb{R}^n\times\mathbb{R}^l$ tends to $(\bar{x},\,\bar{\lambda})$.
(c) For every $\sigma=(a,\,b)\in\mathbb{R}^n\times\mathbb{R}^l$ close enough to $(0,\,0)$, any solution $(x(\sigma),\,\lambda(\sigma))$ of the canonically perturbed Lagrange system

$$\frac{\partial L}{\partial x}(x,\,\lambda)=a,\quad h(x)=b, \tag{9}$$

which is close enough to $(\bar{x},\,\bar{\lambda})$, satisfies the estimate

$$\|x(\sigma)-\bar{x}\|+\operatorname{dist}(\lambda(\sigma),\,\mathcal{M}(\bar{x}))=O(\|\sigma\|) \tag{10}$$

as $\sigma\to 0$.

In particular, neither the Lipschitzian error bound (8), nor the upper-Lipschitzian behavior of the solution set, expressed by (10), can be expected to hold near a critical $\bar{\lambda}$. Hence, the consequences. And this is how the injurious nature of critical multipliers starts to show up!

One interesting fact implied by Proposition 2 is the following: if $\bar{x}$ is a nonisolated stationary point of problem (1), and $\bar{\lambda}$ is the limit of a sequence of Lagrange multipliers associated with stationary points forming a sequence convergent to $\bar{x}$, then $\bar{\lambda}$ is *necessarily* a critical Lagrange multiplier associated with $\bar{x}$. In particular, if $\bar{x}$ is a nonisolated stationary point satisfying the regularity condition (3), then the unique Lagrange multiplier associated with $\bar{x}$ is necessarily critical.

Finally, the next result from [23] demonstrates that when (3) is violated, noncritical multipliers can be stable only subject to very special perturbations of problem (1). The proof of this result employs the estimate (10) from Proposition 2.

Proposition 3 *Let $\bar{x}$ be a stationary point of problem* (1)*, and let $\bar{\lambda}$ be an associated noncritical Lagrange multiplier.*

Then for any $d\in\mathbb{R}^l$, if there exist sequences $\{t_k\}$ of positive reals, $\{x^k\}\subset\mathbb{R}^n$, and $\{\lambda^k\}\subset\mathbb{R}^l$, such that $t_k\to 0$, $\{x^k\}\to\bar{x}$, $\{\lambda^k\}\to\bar{\lambda}$, and for each k the point $(x^k,\,\lambda^k)$ is a solution of the system (9) *with $a=O(t_k)$ and $b=t_kd+o(t_k)$, then the sequence $\{(x^k-\bar{x})/t_k\}$ has a limit point ξ, and any such limit point satisfies the equality*

$$h'(\bar{x})\xi=d. \tag{11}$$

The equality (11) implies that

$$d \in \operatorname{im} h'(\bar{x}),$$

and when (3) is violated, the right-hand side of the previous inclusion is a proper subspace in $\mathbb{R}^l$. Therefore, the situation specified in Proposition 3 may only happen for very special directions d of constraints' perturbations. By contrast, critical multipliers are stable under very reasonable assumptions, and for "generic" perturbations (loosely speaking); see [23]. The next example is a good illustration.

Example 1 Consider the problem

$$\text{minimize } x_1^2 \text{ subject to } x_1^2 - x_2^2 = b,$$

where $b \in \mathbb{R}$ is a parameter perturbing the right-hand side of the constraint. For $b = 0$, this problem has the unique solution $\bar{x} = 0$, with $h'(\bar{x}) = 0$ (hence, the regularity condition (3) is violated) and the multiplier set is $\mathcal{M}(\bar{x}) = \mathbb{R}$. It holds that $\det H(\lambda) = -4(1 + \lambda)\lambda$, and there are two critical multipliers: $\bar{\lambda}^1 = -1$ and $\bar{\lambda}^2 = 0$.

For $a = 0$ the perturbed Lagrange system (9) always has the unique solution, given by

$$(x(b),\, \lambda(b)) = \begin{cases} ((\pm\sqrt{b},\, 0),\, -1) \text{ if } b \geq 0, \\ ((0,\, \pm\sqrt{-b}),\, 0) \text{ if } b < 0. \end{cases}$$

This means that the critical multiplier $\bar{\lambda}^1$ is stable when b moves from zero to the right, while the other critical multiplier $\bar{\lambda}^2$ is stable when b moves from zero to the left. Noncritical multipliers are never stable: they disappear subject to perturbations (in the sense that around each of them there is a neighborhood such that the perturbed problem does not have Lagrange multipliers in this neighborhood for any $b \neq 0$).

3 Attraction of algorithms to critical multipliers and its consequences

Passing from bad to worse. By now, there exists convincing theoretical and numerical evidence of the following striking phenomenon. Despite critical multipliers being typically very "few" (for example, just one point when the set of all Lagrange multipliers is a line, as in Example 4 below), dual sequences generated by NLM (6), and by other Newton-type methods for problem (1), have a remarkably strong tendency to converge to critical multipliers when they exist. This is the case not only for generic mathematical formulations of algorithms, but also for professional implementations of influential software; see [32] for experiments with quasi-Newton SQP and its SNOPT implementation [15], and linearly-constrained Lagrangian methods [48,45,14] and the MINOS package [46]. Moreover, convergence to critical multipliers appears to be the

precise reason for the lack of superlinear convergence rate, which is typical for problems with degenerate constraints. See [30–32] for numerous examples illustrating this behavior. These references also give some theoretical results of a "negative" nature, showing that convergence to a noncritical multiplier is highly unlikely in some sense. Here, we start with a couple of examples from [30], which demonstrate well the attraction phenomenon.

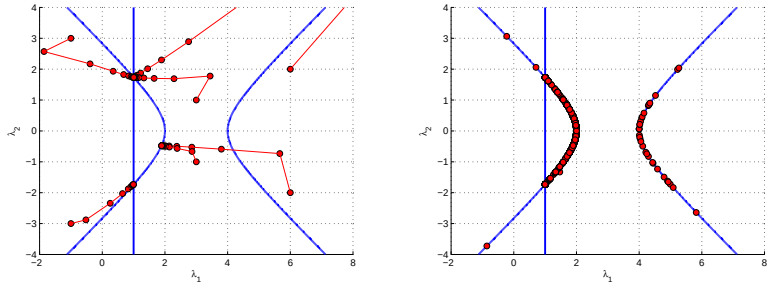
Example 2 The problem with quadratic data

$$\text{minimize } x_1^2 - x_2^2 + 2x_3^2 \text{ subject to } -\frac{1}{2}x_1^2 + x_2^2 - \frac{1}{2}x_3^2 = 0,\ x_1x_3 = 0$$

has the unique solution $\bar{x} = 0$, with $h'(\bar{x}) = 0$ (hence, the regularity condition (3) is violated) and the multiplier set is $\mathcal{M}(\bar{x}) = \mathbb{R}^2$. Furthermore, $\det H(\lambda) = 2(1-\lambda_1)((2-\lambda_1)(4-\lambda_1) - \lambda_2^2)$, and hence, critical multipliers are those $\lambda \in \mathbb{R}^2$ satisfying $\lambda_1 = 1$ or $(\lambda_1 - 3)^2 - \lambda_2^2 = 1$ (vertical line and two branches of hyperbola in Figure 1).

Some NLM dual sequences corresponding to the primal starting point $x^0 = (1, 2, 3)$ are shown in Figure 1a. Figure 1b shows the distribution of dual iterates at the time of termination of the method according to the stopping criterion, for dual trajectories generated starting from the points on the grid in the domain $[-2, 8] \times [-2, 4]$ (step of the grid is $1/4$).

Convergence of the dual sequences is inevitably to critical multipliers (and convergence of both primal and dual sequences is slow).



(a) Dual sequences.

(b) Distribution of dual iterates at the time of termination.

Fig. 1: NLM for Example 2; $x^0 = (1, 2, 3)$.

In the previous example, all the functions are quadratic and the problem is fully degenerate (the multiplier set is the whole dual space). In the example that follows, some functions are not quadratic, and the multiplier set is a proper subspace of the dual space.

Example 3 The problem with non-quadratic data

$$\begin{array}{l}\text{minimize} \quad x_1^2 + x_2^2 + x_3^2 \\ \text{subject to } \sin x_1 + \sin x_2 + \sin x_3 = 0,\ x_1 + x_2 + x_3 + \sin x_1 x_3 = 0\end{array}$$

has the unique solution $\bar{x} = 0$, and the regularity condition (3) is violated (even though $h'(\bar{x}) \neq 0$). The multiplier set is $\mathcal{M}(\bar{x}) = \{\lambda \in \mathbb{R}^2 \mid \lambda_1 + \lambda_2 = 0\}$ (thick straight line in Figure 2), and it can be seen that the only critical multipliers are $(-2,\, 2)$ and $(6, -6)$.

Some NLM dual sequences corresponding to the primal starting point $x^0 = (0.1,\, 0.2,\, 0.3)$ are shown in Figure 2.

Again, the dual sequences always converge to one of the two critical multipliers.

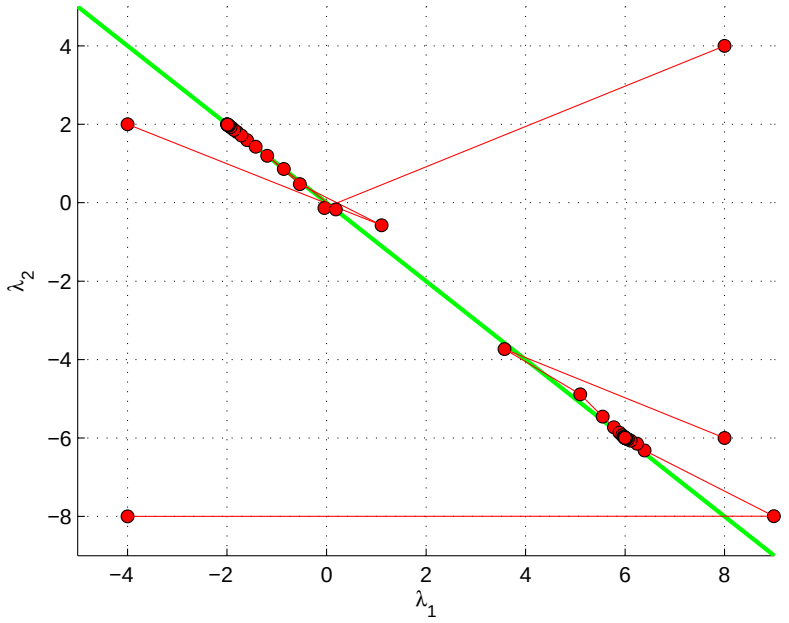


Fig. 2: Dual sequences of NLM for Example 3; $x^0 = (0.1,\, 0.2,\, 0.3)$.

Our next example, although very simple, allows us to demonstrate the phenomenon in question analytically.

Example 4 Consider the following problem with quadratic data:

$$\text{minimize } ax^2 \text{ subject to } bx^2 = 0,$$

where a and b are nonzero real parameters. The unique feasible point (hence, the unique solution of this problem) is $\bar{x} = 0$, and since $h'(\bar{x}) = 0$, it holds that $\mathcal{M}(\bar{x}) = \mathbb{R}$ and

$$H(\lambda) = \frac{\partial^2 L}{\partial x^2}(\bar{x}, \lambda) = 2(a + \lambda b).$$

Therefore, the unique critical multiplier is $\bar{\lambda} = -a/b$.

For each k, the equalities in (6) take the form

$$(a + \lambda^k b)(x - x^k) + bx^k(\lambda - \lambda^k) = -(a + \lambda^k b)x^k, \quad 2bx^k(x - x^k) = -b(x^k)^2.$$

Assuming that $x^k \neq 0$, we obtain from the second equality that $x^{k+1} = x^k/2$, and then the first equality gives $\lambda^{k+1} = (\lambda^k - a/b)/2$, which can be further written as

$$\lambda^{k+1} + a/b = \frac{1}{2}(\lambda^k + a/b).$$

Therefore, if $x^0 \neq 0$, then $x^k \neq 0$ for all k, and the sequence $\{(x^k, \lambda^k)\}$ is well defined by (6) and converges linearly to $(0, \bar{\lambda})$. Moreover, if $\lambda^0 \neq \bar{\lambda}$, then the convergence rates of both $\{x^k\}$ and $\{\lambda^k\}$ are linear.

Some primal-dual sequences of NLM for this example with $a = b = 1$ are shown in Figure 3.

Given the illustrations above, as well as many more examples in [30–32], it is natural to try to *prove* that NLM always converges to a critical multiplier when one exists. Despite all the evidence, this turned out to be very hard. No proof is known at this time for the general problem setting of (1). We proceed to discuss what is currently known for some special cases.

The behavior observed in Example 4 is fully explained by the following result for a one-dimensional problem with a single constraint, obtained in [41]. Even in this seemingly simple case, the proof is not at all simple!

Proposition 4 *Let $n = l = 1$ and assume that $f'(\bar{x}) = h(\bar{x}) = h'(\bar{x}) = 0$, $h''(\bar{x}) \neq 0$.*

Then for any $x^0 \in \mathbb{R} \setminus \{\bar{x}\}$ close enough to $\bar{x}$, and any $\lambda^0 \in \mathbb{R}$, there exists the unique sequence $\{(x^k, \lambda^k)\} \subset \mathbb{R} \times \mathbb{R}$ such that (x^{k+1}, λ^{k+1}) satisfies (6) for all k; this sequence converges to $(\bar{x}, \bar{\lambda})$, where $\bar{\lambda} = -f''(\bar{x})/h''(\bar{x})$; $x^k \neq \bar{x}$ holds for all k, and

$$\lim_{k\to\infty} \frac{x^{k+1} - \bar{x}}{x^k - \bar{x}} = \frac{1}{2}.$$

Note that the assumption $f'(\bar{x}) = h(\bar{x}) = h'(\bar{x}) = 0$ implies that $\bar{x}$ is a stationary point of problem (1), with $\mathcal{M}(\bar{x}) = \mathbb{R}$, while $h''(\bar{x}) \neq 0$ implies that $\bar{\lambda} = -f''(\bar{x})/h''(\bar{x})$ is the unique critical Lagrange multiplier associated with $\bar{x}$. Linear convergence rate of $\{\lambda^k\}$ can also be established under some additional assumptions; see [41].

Proposition 4 is nice in the sense of its clear theoretical characterization of convergence of the dual sequence (from any dual starting point!) to the

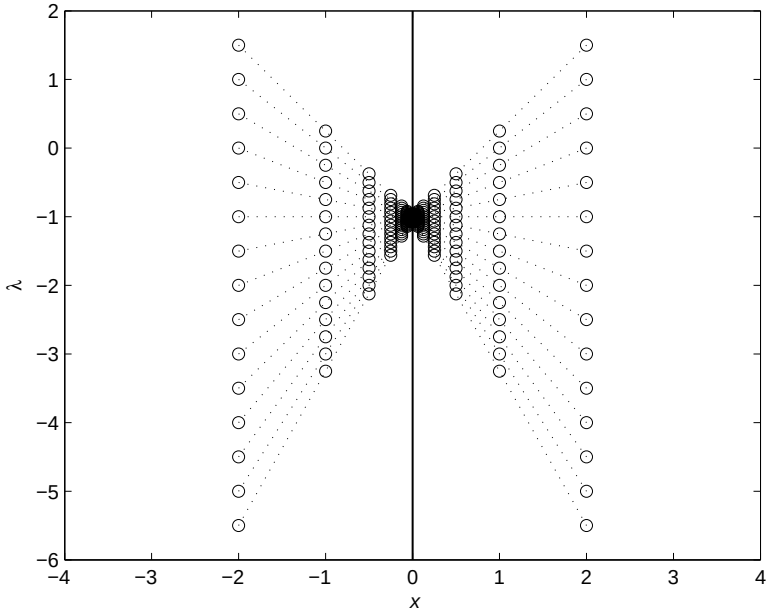


Fig. 3: Primal-dual sequences of NLM for Example 4.

(unique!) critical multiplier, and the associated slow primal convergence. But it is also somewhat discouraging, as it suggests that there may be no chances to escape convergence to a critical multiplier, and thus slow convergence. As already mentioned, unfortunately (or fortunately), currently there exist no full extensions of this nice (but discouraging) result to higher dimensions and with f and h still general enough (moreover, such full extension could hardly be possible: some further assumptions, and restrictions on the starting point would have to come into play). We proceed with discussing a more special case, namely, that of all the problem data being quadratic. Note, however, that this model setting captures the most intrinsic consequences of constraints degeneracy.

Consider the problem

$$\text{minimize } \frac{1}{2}\langle Ax,\, x\rangle \text{ subject to } \frac{1}{2}B[x,\, x] = 0, \tag{12}$$

where A is a symmetric $n \times n$ matrix, and $B : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^l$ is a symmetric bilinear mapping, that is, $B[x,\, x] = (\langle B_1 x,\, x\rangle,\, \ldots,\, \langle B_l x,\, x\rangle)$, where $B_1,\, \ldots,\, B_l$ are symmetric $n \times n$ matrices. Observe that $\bar{x} = 0$ is always a stationary point of problem (12), $\mathcal{M}(\bar{x}) = \mathbb{R}^l$, and $H(\lambda) = A + \lambda B$, where $\lambda B = \sum_{i=1}^{l} \lambda_i B_i$.

For problem (12), the Lagrange system (2) takes the form

$$H(\lambda)x = 0, \quad \frac{1}{2}B[x,\, x] = 0,$$

and the iteration subproblem (6) of NLM can be written as

$$\begin{pmatrix} H(\lambda^k) & (B[x^k])^{\mathrm{T}} \\ B[x^k] & 0 \end{pmatrix} \begin{pmatrix} x - x^k \\ \lambda - \lambda^k \end{pmatrix} = - \begin{pmatrix} H(\lambda^k)x^k \\ \dfrac{1}{2}B[x^k,\, x^k] \end{pmatrix}, \tag{13}$$

where for a given $\xi \in \mathbb{R}^n$, the matrix $B[\xi] \in \mathbb{R}^{l \times n}$ is defined by $B[\xi]x = B[\xi,\, x]$.

Extension of the results presented below to more general settings (with linear terms and/or higher-order terms) might be possible using, in particular, the Lyapunov–Schmidt procedure (e.g., [19, Ch. VII]). This is a subject for future research. Such an extension will certainly involve technical complications, as even in the purely quadratic setting the required analysis is already quite involved [39].

Consider first an even more special case when $l = 1$, i.e., when (12) is a problem with a single constraint:

$$\text{minimize } \frac{1}{2}\langle Ax,\, x\rangle \text{ subject to } \frac{1}{2}\langle Bx,\, x\rangle = 0, \tag{14}$$

where B is a symmetric $n \times n$ matrix. Assume that B is nonsingular. The latter is, of course, a generic property, thus not restrictive. But in fact, this assumption can be dropped, as is demonstrated below by the analysis for the general quadratic problem (12). We impose this condition only in order to present our assumptions for the special case of $l = 1$ in standard algebraic terms. Specifically, under this assumption, $\bar{\lambda}$ satisfies (7) (and hence, is a critical multiplier associated to $\bar{x}$) if and only if $-\bar{\lambda}$ is an eigenvalue of the matrix $B^{-1}A$.

We shall further assume that the eigenvalue $-\bar{\lambda}$ has the algebraic multiplicity 1, which also holds generically, and which essentially means that the critical multiplier $\bar{\lambda}$ is not "too critical" (one can further study critical multipliers "of order 2", i.e., those for which the corresponding eigenvalue has multiplicity 2, etc.). This property can be equivalently expressed in the form

$$\Delta'(\bar{\lambda}) \neq 0, \tag{15}$$

where the function $\Delta : \mathbb{R}^l \to \mathbb{R}$ is given by $\Delta(\lambda) = \det H(\lambda)$. The geometric multiplicity of an eigenvalue is never higher than its algebraic multiplicity, and hence, under our assumption,

$$\dim \ker H(\bar{\lambda}) = 1. \tag{16}$$

Moreover, it can be seen that our assumption is in fact equivalent to the combination of (16) and the condition

$$\langle B\bar{\xi}, \bar{\xi}\rangle \neq 0 \tag{17}$$

for any $\bar{\xi} \in \mathbb{R}^n$ spanning $\ker H(\bar{\lambda})$. These properties are used in order to establish the following result [50].

Proposition 5 *Let B be a nonsingular matrix, and assume that $-\bar{\lambda}$ is an eigenvalue of the algebraic multiplicity 1 of the matrix $B^{-1}A$.*

Then there exists $\delta > 0$ such that for any $x^0 \in \mathbb{R}^n \setminus \{0\}$ satisfying

$$\operatorname{dist}(x^0/\|x^0\|, \ker H(\bar{\lambda})) < \delta, \tag{18}$$

and any $\lambda^0 \in (\bar{\lambda} - \delta, \bar{\lambda}) \cup (\bar{\lambda}, \bar{\lambda} + \delta)$, there exists the unique sequence $\{(x^k, \lambda^k)\} \subset \mathbb{R} \times \mathbb{R}$ such that (x^{k+1}, λ^{k+1}) satisfies (13) for all k; this sequence converges to $(0, \bar{\lambda})$; $x^k \neq 0$ and $\lambda^k \neq \bar{\lambda}$ for all k;

$$\lim_{k\to\infty} \frac{\|x^{k+1}\|}{\|x^k\|} = \frac{1}{2}, \quad \lim_{k\to\infty} \frac{\lambda^{k+1} - \bar{\lambda}}{\lambda^k - \bar{\lambda}} = \frac{1}{2},$$

and the sequence $\{x^k/\|x^k\|\}$ converges to an element of $\ker H(\bar{\lambda})$.

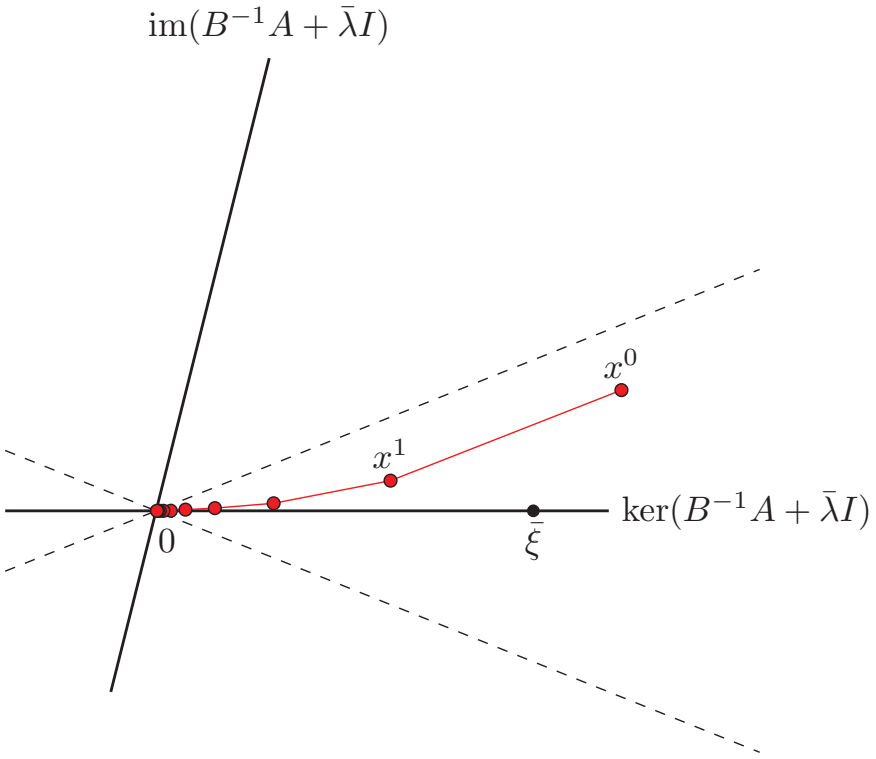


Fig. 4: Primal behavior of NLM.

The assertion of Proposition 5 is illustrated in Figure 4. The main idea of the proof is as follows. In the case of problem (14), the NLM iteration system

(13) can be seen to yield

$$x^{k+1} = \frac{1}{2}\,\frac{\langle B\xi^k,\,\xi^k\rangle}{\langle B\xi^k,\,P_k\xi^k\rangle}P_k x^k, \quad \lambda^{k+1} - \bar{\lambda} = \left(1 - \frac{1}{2}\,\frac{\langle B\xi^k,\,\xi^k\rangle}{\langle B\xi^k,\,P_k\xi^k\rangle}\right)(\lambda^k - \bar{\lambda})$$

(assuming that $\langle B\xi^k,\,P_k\xi^k\rangle \neq 0$), where $\xi^k = x^k/\|x^k\|$, and $P_k = P(\lambda^k,\,\bar{\lambda})$ is an $n \times n$ matrix such that $P(\lambda^k,\,\bar{\lambda})$ tends to the projector onto the one-dimensional subspace $\ker(A + \bar{\lambda}B) = \ker(B^{-1}A + \bar{\lambda}I)$ along the subspace $\mathrm{im}(B^{-1}A + \bar{\lambda}I)$, as $\lambda^k \to \bar{\lambda}$. Therefore, if λ^k is close to $\bar{\lambda}$, and ξ^k is close to $\ker(A + \bar{\lambda}B)$, then $P_k\xi^k$ is close to ξ^k, and hence, x^{k+1} is close to $x^k/2$, and $\lambda^{k+1} - \bar{\lambda}$ is close to $(\lambda^k - \bar{\lambda})/2$.

One may wonder whether the restriction (18) on the domain of attraction is essential in Proposition 5. The answer is yes, as demonstrated by the following example from [50].

Example 5 For the problem

$$\text{minimize } 2x_1^2 + 2x_2^2 + 5x_2x_3 \text{ subject to } x_1x_2 + x_1x_3 + x_2x_3 = 0,$$

we have that $\det H(\lambda) = 2(\lambda - 5)(\lambda^2 + 6\lambda + 10)$, implying that the unique critical multiplier is $\bar{\lambda} = 5$, and the algebraic multiplicity of the corresponding eigenvalue is 1.

However, if the iteration sequence $\{(x^k,\,\lambda^k)\}$ of NLM is well defined, and if for some k it turns out that x^k belongs to the linear subspace $\mathcal{L}$ spanned by $(1,\,-1,\,3)$ and $(3,\,2,\,2)$, then all the subsequent primal iterates remain in $\mathcal{L}$, and the sequence $\{\lambda^k\}$ cannot converge (and in particular, cannot converge to the critical multiplier). Apparently, this sequence usually has noncritical accumulation points, though it is not clear whether the latter can be proven formally. This behavior is rather persistent: even if $x^0 \notin \mathcal{L}$, the directions $x^k/\|x^k\|$ quite often tend to the subspace $\mathcal{L}$. One run of this kind is shown in Figure 5, where $x^0 = (1,\,-1,\,2)$, $\lambda^0 = 1$. Moreover, small perturbations of the starting point do not destroy this behavior.

What happens in this example is that the sequence $\{x^k/\|x^k\|\}$ stays separated from $\ker H(\bar{\lambda})$, preventing the dual sequence initialized arbitrarily close to the unique critical multiplier to converge to this multiplier, and this happens when $x^0/\|x^0\|$ is not close enough to $\ker H(\bar{\lambda})$.

Convergence to the critical multiplier is also a common behavior in Example 5, especially if λ^0 is taken close to $\bar{\lambda}$ (of course, in this case $x^k \notin \mathcal{L}$ for all k). However, the above considerations show, in particular, that even for problem (14), and even when a critical multiplier is unique, one cannot expect that convergence to it can be proven for *all* starting points $(x^0,\,\lambda^0)$ with λ^0 close enough to that critical multiplier.

Define the set of critical multipliers:

$$\mathcal{M}_0 = \{\lambda \in \mathbb{R}^l \mid \det H(\lambda) = 0\}.$$

The case when $l > 1$ is qualitatively more involved than $l = 1$ because in the former case points in $\mathcal{M}_0$ are usually non-isolated (since this set is given by

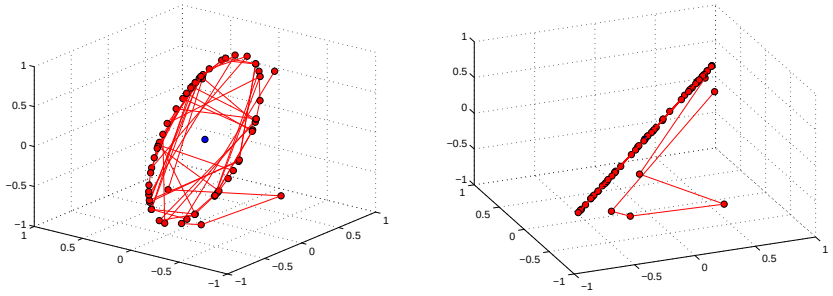


Fig. 5: Sequence $\{x^k/\|x^k\|\}$ for NLM for Example 5; $x^0 = (1, -1, 2)$, $\lambda^0 = 1$.

a single algebraic equation in $l > 1$ variables). This means, in particular, that one cannot expect to prove convergence to a *given* point of this set: limits of dual sequences would naturally depend on starting points.

Nevertheless, let us consider the dual behavior near a given $\bar{\lambda} \in \mathcal{M}_0$. Our first key assumption is (15), which certainly makes sense when $l > 1$ as well. (We cannot be talking about any eigenvalues though, since $\bar{\lambda}$ is not a scalar anymore). It can be seen that (15) is still equivalent to (16) combined with the natural counterpart of (17):

$$B[\bar{\xi}, \bar{\xi}] \neq 0. \tag{19}$$

Critical multipliers satisfying (15) can still be considered as critical "of order 1", though now we do not refer to any multiplicities.

We need to state one more assumption, which is rather technical. Let $\mathcal{J}_l$ be the set of all subsets of $\{1, \ldots, n\}$ whose cardinality is equal to l, and for any $J \in \mathcal{J}_l$, any $l \times n$ matrix M, and any $n \times n$ matrix H, define the $n \times n$ matrix $S(M, H; J)$ with the rows

$$(S(M, H; J))_j = \begin{cases} (M^{\mathrm{T}}M)_j & \text{if } j \in J, \\ H_j & \text{if } j \in \{1, \ldots, n\} \setminus J. \end{cases}$$

The assumption in question has the form

$$\sum_{J \in \mathcal{J}_l} \det S(B[\bar{\xi}], H(\bar{\lambda}); J) \neq 0. \tag{20}$$

It can be easily seen that this property implies, in particular, that

$$\operatorname{rank} B[\bar{\xi}] = l. \tag{21}$$

Moreover, we are currently not aware of any examples satisfying (15) (or equivalently, (16) and (19)) and (21), but violating (20). Therefore, the conjecture is that under (15), condition (20) is actually equivalent to (21). At the same time, it can be seen by the transversality theorem that if $n + 2 > 2l$, then for

all $(A,\, B)$ in some massive (hence, dense) set in $\mathbb{S}^n \times \bigotimes_{i=1}^{l} \mathbb{S}^n$, for any $\bar{\lambda} \in \mathcal{M}_0$ condition (21) holds for all nonzero $\bar{\xi} \in \ker H(\bar{\lambda})$.

The following result was established in [39].

Theorem 1 *Assume that* (15) *and* (20) *hold for some* $\bar{\lambda} \in \mathcal{M}_0$ *and for* $\bar{\xi}$ *spanning* $\ker H(\bar{\lambda})$.

Then for any $\ell > 0$ *and any* $\rho > 0$ *there exists* $\delta \in (0,\, \rho)$ *such that for any* $x^0 \in \mathbb{R}^n \setminus \{0\}$ *and any* $\lambda^0 \in \mathcal{B}(\bar{\lambda},\, \delta) \setminus \mathcal{M}_0$ *satisfying* (18) *and*

$$\frac{\| \left(B[x^0](H(\lambda^0))^{-1}(B[x^0])^{\mathrm{T}}\right)^{-1} B[x^0,\, x^0]\|}{|\det H(\lambda^0)|} \leq \ell \tag{22}$$

with the well-defined left-hand side, there exists the unique sequence $\{(x^k,\, \lambda^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l$ *such that* $(x^{k+1},\, \lambda^{k+1})$ *satisfies* (13) *for all* k*; this sequence converges to* $(0,\, \lambda^*)$ *for some* $\lambda^* \in \mathcal{M}_0$*;* $x^k \neq 0$ *and* $\lambda^k \in \mathcal{B}(\bar{\lambda},\, \rho) \setminus \mathcal{M}_0$ *for all* k*;*

$$\lim_{k\to\infty} \frac{\|x^{k+1}\|}{\|x^k\|} = \frac{1}{2}, \quad \lim_{k\to\infty} \frac{\lambda^{k+1} - \lambda^*}{\lambda^k - \lambda^*} = \frac{1}{2};$$

the sequence $\{x^k/\|x^k\|\}$ *converges to* ξ^* *spanning* $\ker H(\lambda^*)$*; the sequence* $\{(\lambda^k - \lambda^*)/\|\lambda^k - \lambda^*\|\}$ *converges to some* $d \in \mathbb{R}^l$ *which is transverse (not tangent) to* $\mathcal{M}_0$ *at* λ^*.

Example 5 demonstrates that it is essential to take x^0 satisfying (18). However, the behavior in that example is extremely atypical: usually NLM eventually enters the attraction domain of a critical multiplier (when it exists), even when the behavior is "chaotic" in the beginning. For the problem from Example 2, dual sequence for $x^0 = (3,\, 2,\, 1)$, $\lambda^0 = (3.75,\, -0.25)$ is shown in Figure 6. The initial part of this run also shows that $x^0/\|x^0\|$ must be close enough to $\ker H(\bar{\lambda})$ for the assertion of Theorem 1 to take effect.

Generally, in our experience, closeness of $x^0/\|x^0\|$ to $\ker H(\bar{\lambda})$ plays a crucial role for convergence of the dual sequence to a critical multiplier close to $\bar{\lambda}$.

The additional assumption (20) and the additional restriction (22) on the domain of attraction in Theorem 1 can be dropped when $l = 1$, but not when $l > 1$ (see [39] for counterexamples).

By the parametric transversality theorem it can be derived that for all $(A,\, B)$ in some massive set in $\mathbb{S}^n \times \bigotimes_{i=1}^{l} \mathbb{S}^n$, for any $\bar{\lambda} \in \mathcal{M}_0$ it holds that

$$\dim \ker H(\bar{\lambda})(\dim \ker H(\bar{\lambda}) + 1) \leq 2l,$$

and if (16) is valid, then (15) holds as well. In particular, if $l \leq 2$, then (15) holds generically for all critical multipliers. In Example 2, which is a fully quadratic problem with two constraints, critical multipliers violating (15) correspond to the points where the vertical line intersects the hyperbola; see Figure 7. The thin line demonstrates the typical structure of the set of critical multipliers of the slightly perturbed (though still fully quadratic) problem: the points of intersection disappear, and this set is now a smooth manifold.

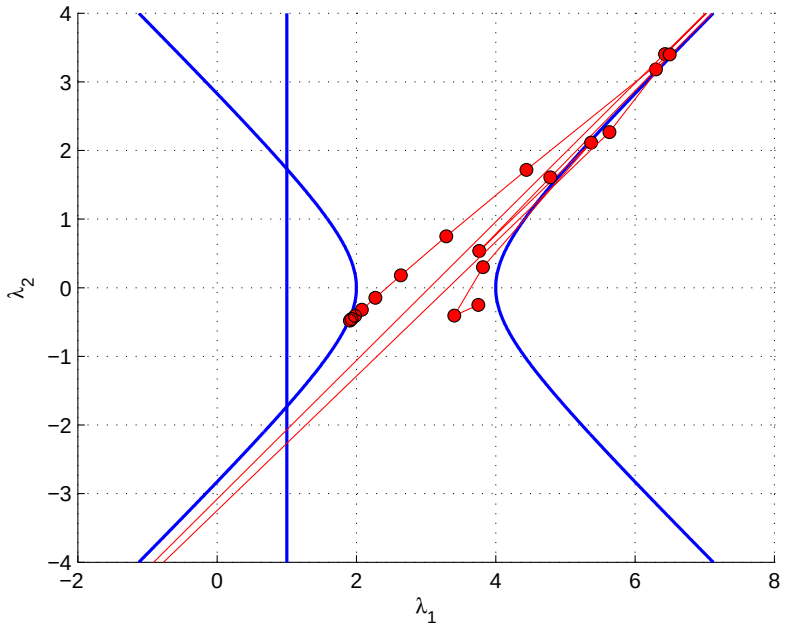


Fig. 6: Dual sequence of the NLM for Example 2.

If $l \geq 3$, then critical multipliers violating (15) (hence, violating at least one of the conditions (16) or (19)) may exist in a stable way, but usually (15) holds for "almost all" critical multipliers.

Currently, there exists no theory of attraction to multipliers which are critical "of order higher than 1" (that is, violate (15)), but problems like the one in Example 2 suggest that such multipliers are even more attractive for the Newton-type methods than critical multipliers "of order 1". There are also no formal proofs of attraction for problems with non-quadratic data; however, problems like in Example 3 show that the phenomenon is there for general functions as well. Advancing the theory in those directions constitutes some open questions.

4 Dual stabilization techniques

We hope the exposition above (and supporting references) convinced the reader that the effect of attraction to critical multipliers is real, very persistent, and it slows down convergence of NLM. While we focus the discussion on NLM, it is important to stress again that the same happens for various related algorithms; for example, quasi-Newton SQP and linearly constrained Lagrangian methods.

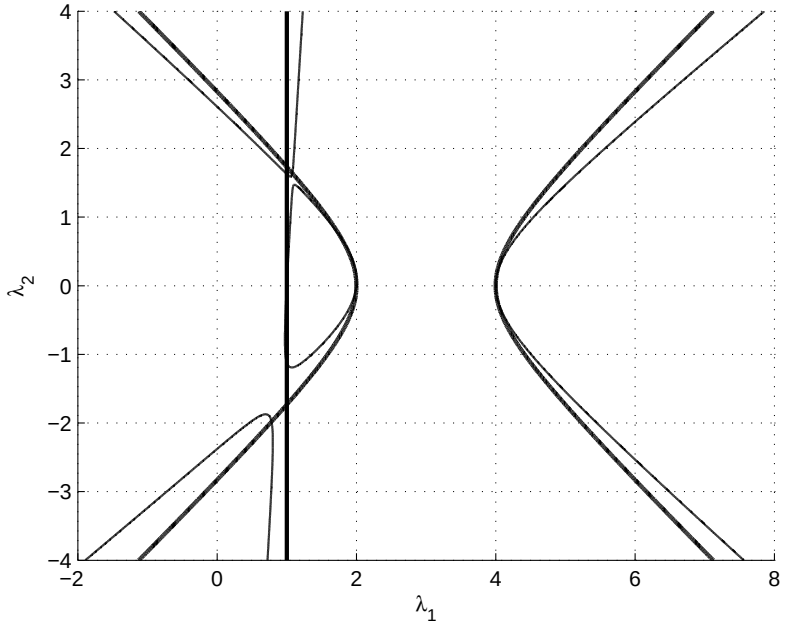


Fig. 7: Critical multipliers in Example 2 subject to perturbations.

Moreover, this is observed in professional software such as MINOS [46] and SNOPT [15]; see [32].

If we do not want to resign to this unfortunate state of affairs (and we do not!), we need to develop tools for avoiding (or at least significantly alleviating) the exhibited undesirable dual behavior. Such tools must aim to prevent the dual sequence from moving *along* the set of Lagrange multipliers towards critical ones; i.e., they should force the dual sequence to approach the multipliers set *transversely*. The natural name for such techniques is *dual stabilization*: the shorter the way of dual iterates to their eventual limit, the better.

Fortunately, there are dual stabilization tools that do the job at least *locally*, in the sense that if the dual starting point is close enough to a noncritical multiplier (or to a multiplier satisfying SOSC) then convergence to a critical multiplier does not occur, and thus convergence to a solution is also fast. Such local stabilization techniques are the subject of this section. In fact, somewhat surprisingly, it turns out that one classical algorithm, known and widely used for decades, possesses the needed dual stabilization property. However, as will be discussed in the next section, the difficulties are still not resolved completely, in the *global* sense. In particular, when the dual starting point is not close enough to a noncritical multiplier (or to a multiplier satisfying SOSC), attraction to critical multipliers is still observed with a certain frequency, with all the negative consequences.

We shall discuss two methods that have the local stabilization property. The first one has the word "stabilized" in its name; it is the stabilized sequential quadratic programming (sSQP) method or, in our context, the stabilized Newton–Lagrange method (sNLM). For a current primal-dual iterate $(x^k,\, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^l$, it generates the next iterate $(x^{k+1},\, \lambda^{k+1})$ by solving the linear system

$$\begin{gathered}\frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)(x - x^k) + (h'(x^k))^{\mathrm{T}}(\lambda - \lambda^k) = -\frac{\partial L}{\partial x}(x^k,\, \lambda^k), \\ h'(x^k)(x - x^k) - \sigma_k(\lambda - \lambda^k) = -h(x^k),\end{gathered} \tag{23}$$

where $\sigma_k \geq 0$ is the dual stabilization parameter. When $\sigma_k = 0$, the sNLM iteration system (23) coincides with the NLM iteration system (6). The stabilization effect is achieved by positive σ_k, controlled appropriately.

The sSQP idea was introduced in [51] for inequality-constrained problems, and in the form of solving certain min-max subproblems. Later, it was recognized that the min-max subproblem is actually equivalent to a quadratic programming problem in the primal-dual space [42], and that the method is applicable to equality-constrained problems as well. In the latter case, which is the setting of our discussion, the sSQP subproblem takes the following form:

$$\begin{aligned}&\text{minimize} \quad \langle f'(x^k),\, x - x^k\rangle + \frac{1}{2}\left\langle \frac{\partial^2 L}{\partial x^2}(x^k,\, \lambda^k)(x - x^k),\, x - x^k \right\rangle + \frac{\sigma_k}{2}\|\lambda\|^2 \\ &\text{subject to } h(x^k) + h'(x^k)(x - x^k) - \sigma_k(\lambda - \lambda^k) = 0.\end{aligned}$$

The Lagrange optimality system for this problem is precisely (23) above. The idea is that the quadratic term $\|\lambda\|^2$ puts an "anchor" on the dual sequence, preventing it from drifting along the multiplier set, towards critical multipliers. Note also that there is the so-called "ellastic mode" feature in the constraints, which makes the subproblems feasible without any constraint qualifications or other assumptions. For the analysis of sSQP, see [51,20,13,11,33] and [34, Chapter 7].

The (currently) sharpest local convergence result for sNLM was obtained in [33]. It reveals precisely the dual stabilization property discussed above. Here, we give a slightly strengthened version of the result in [33], with weaker requirements on stabilization parameter values.

Theorem 2 *Let $\bar\lambda$ be a noncritical Lagrange multiplier associated to a stationary point $\bar x$ of problem (1). Let $\sigma : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ be any function satisfying with some $C > 0$ the following requirements: σ is continuous at every point of $\{\bar x\} \times \mathcal{M}(\bar x)$ close enough to $(\bar x,\, \bar\lambda)$, $\sigma(\bar x,\, \lambda) = 0$ for all $\lambda \in \mathcal{M}(\bar x)$ close enough to $\bar\lambda$, and for all $(x,\, \lambda) \in (\mathbb{R}^n \times \mathbb{R}^l) \setminus (\{\bar x\} \times \mathcal{M}(\bar x))$ close enough to $(\bar x,\, \bar\lambda)$ it holds that*

$$\sigma(x,\, \lambda) \neq 0, \quad \|x - \bar x\| \leq C|\sigma(x,\, \lambda)|.$$

Then for any $(x^0,\, \lambda^0) \in \mathbb{R}^n \times \mathbb{R}^l$ close enough to $(\bar x,\, \bar\lambda)$ there exists the unique sequence $\{(x^k,\, \lambda^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l$ such that $(x^{k+1},\, \lambda^{k+1})$ satisfies, with $\sigma_k = \sigma(x^k,\, \lambda^k)$ and for each k, the system (23); this sequence converges superlinearly to $(\bar x,\, \lambda^*)$, where $\lambda^* \in \mathcal{M}(\bar x)$ is such that $\lambda^* \to \bar\lambda$ as $(x^0,\, \lambda^0) \to (\bar x,\, \bar\lambda)$.

Since $\bar{\lambda}$ is assumed noncritical, appropriate practical choices of the function σ in Theorem 2 are suggested by the error bound property (8) in Proposition 2: for any fixed $\tau > 0$ and $\theta \in (0, 1]$, one can take

$$\sigma(x, \lambda) = \tau \left\| \left(\frac{\partial L}{\partial x}(x, \lambda), h(x) \right) \right\|^{\theta}, \tag{24}$$

which is readily computable.

Some primal-dual sequences of sNLM for Example 4 are shown by dashed lines in Figure 8, and the dual stabilization effect is evident (just compare the outcomes to the dotted lines of SNM, all leading to the critical multiplier). At the same time, those dual sequences which are initialized close to the critical multiplier are still attracted to it. The attraction domain of the critical multiplier in Example 4 for sNLM is shown in Figure 9. Observe that there is no contradiction with Theorem 2: from some neighborhood of each noncritical multiplier, convergence to the critical one does not occur. However, the size of such neighborhood tends to zero (at superlinear rate) as the noncritical multiplier in question approaches the critical one.

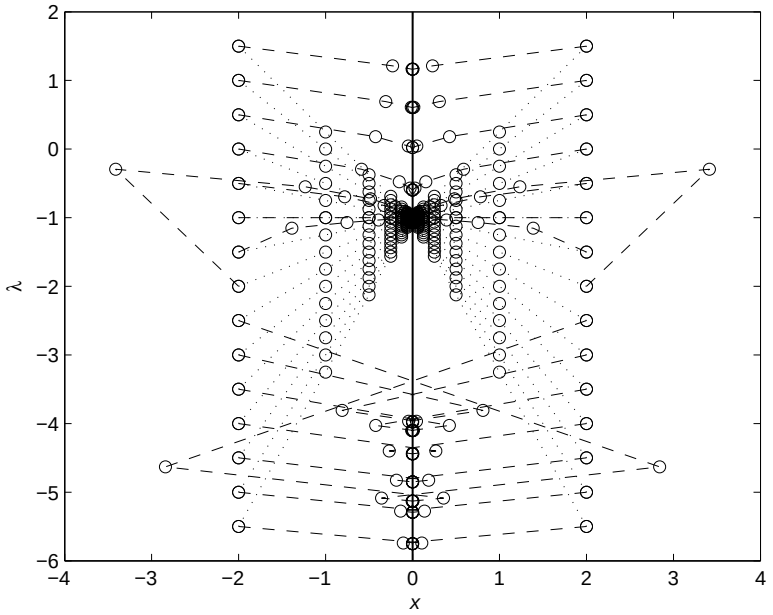


Fig. 8: Primal-dual sequences of NLM and SNLM for Example 4.

Another method which turns out to have the dual stabilization property is the classical augmented Lagrangian algorithm (called also the method of multipliers), which dates back to [22] and [47]. It remains an important technique

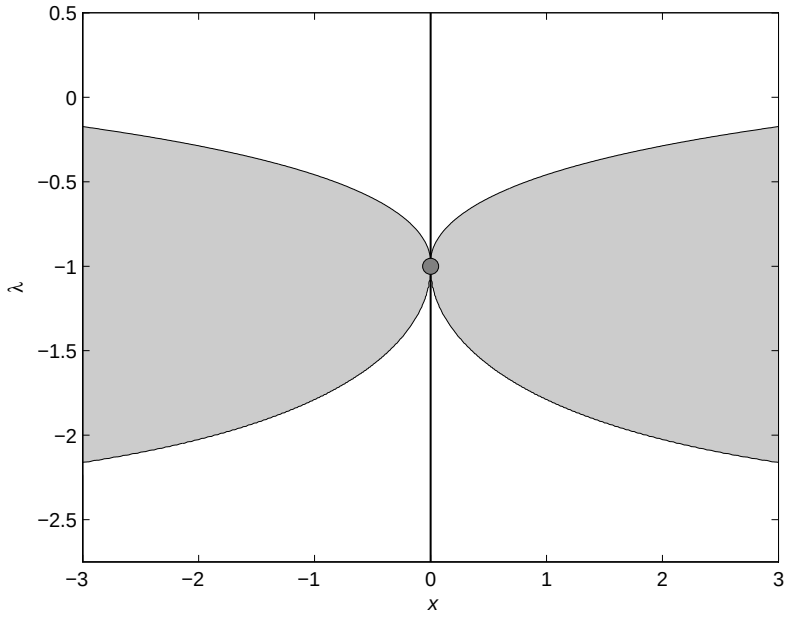


Fig. 9: Primal-dual attraction domain of the critical multiplier for sNLM in Example 4.

still attracting interest, both theoretical and practical; see [2–4,12,27,36] for some recent developments, and also [7].

The augmented Lagrangian $L_\sigma : \mathbb{R}^n \times \mathbb{R}^l \mapsto \mathbb{R}$ for problem (1) is given by

$$L_\sigma(x,\, \lambda) = L(x,\, \lambda) + \frac{1}{2\sigma}\|h(x)\|^2,$$

where $\sigma > 0$ is the (inverse of) penalty parameter. Given the current estimate $\lambda^k \in \mathbb{R}^l$ of a Lagrange multiplier and $\sigma_k > 0$, an iteration of the augmented Lagrangian method consists of computing the primal iterate x^{k+1} by solving (usually approximately) the unconstrained subproblem

$$\text{minimize } L_{\sigma_k}(x,\, \lambda^k) \text{ subject to } x \in \mathbb{R}^n, \tag{25}$$

and then updating the multipliers by the explicit formula

$$\lambda^{k+1} = \lambda^k + \frac{1}{\sigma_k} h(x^{k+1}). \tag{26}$$

The following local convergence result, exposing the dual stabilization property of the augmented Lagrangian method, was recently obtained in [29]. Again, we present here a slightly strengthened version of the result in [29].

Theorem 3 *Let $\bar\lambda$ be a noncritical Lagrange multiplier associated to a stationary point $\bar x$ of problem* (1)*. Let $\sigma : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$ be any function satisfying with some $C > 0$ the following requirements: $\sigma(x,\lambda) \to 0$ as $(x,\lambda) \to (\bar x,\bar\lambda)$, and for all $(x,\lambda) \in (\mathbb{R}^n \times \mathbb{R}^l) \setminus (\{\bar x\} \times \mathcal{M}(\bar x))$ close enough to $(\bar x,\bar\lambda)$ it holds that*

$$\|x-\bar x\| + \operatorname{dist}(\lambda, \mathcal{M}(\bar x)) \le C|\sigma(x,\lambda)|.$$

Then for any $c > 3$, for any $(x^0,\lambda^0) \in \mathbb{R}^n \times \mathbb{R}^l$ close enough to $(\bar x,\bar\lambda)$ there exists a sequence $\{(x^k,\lambda^k)\} \subset \mathbb{R}^n \times \mathbb{R}^l$ such that x^{k+1} is a stationary point of problem (25) *with $\sigma_k = \sigma(x^k,\lambda^k)$, λ^{k+1} computed according to* (26)*, and*

$$\|(x^{k+1}-x^k,\lambda^{k+1}-\lambda^k)\| \le c\operatorname{dist}((x^k,\lambda^k),\bar U)$$

for all k, where $\bar U$ is the solution set of the Lagrange optimality system (2)*; any such sequence converges superlinearly to some $(\bar x,\lambda^*)$ with $\lambda^* \in \mathcal{M}(\bar x)$ such that $\lambda^* \to \bar\lambda$ as $(x^0,\lambda^0) \to (\bar x,\bar\lambda)$.*

Again, appropriate choices of σ in Theorem 3 are given by (24).

One may say that the augmented Lagrangian method is not of Newton type, and therefore, should not be considered as a dual stabilization procedure in our context. However, writing the optimality conditions for the subproblem (25), and combining it with the updating rule for dual variables (26), we obtain that

$$\begin{aligned} 0 &= \frac{\partial L_{\sigma_k}}{\partial x}(x^{k+1},\lambda^k) \\ &= f'(x^{k+1}) + (h'(x^{k+1}))^{\mathrm{T}}\lambda^k + (h'(x^{k+1}))^{\mathrm{T}}h(x^{k+1}) \\ &= \frac{\partial L}{\partial x}(x^{k+1},\lambda^{k+1}). \end{aligned} \tag{27}$$

Therefore, the next primal-dual iterate (x^{k+1},λ^{k+1}) of the augmented Lagrangian method is, in fact, defined by the system of equations

$$\frac{\partial L}{\partial x}(x,\lambda) = 0, \quad h(x) - \sigma_k(\lambda-\lambda^k) = 0. \tag{28}$$

Linearizing this system at (x^k,λ^k) yields precisely the iteration system (23) of sNLM. This gives the relation between sNLM and the augmented Lagrangian method: the former is a "linearization" of the latter, or in other words, the sNLM step is just the Newton step for the primal-dual iteration system (28) of the augmented Lagrangian method. This explains the similarities in local convergence properties of the two methods, and also similar techniques needed for their state-of-the-art local convergence analysis [11,12]; see also [35] for a discussion of a broader view on Newtonian methods. That said, there are also some subtle but remarkable differences. For instance, there exist examples [29] demonstrating that stabilization function $\sigma(x,\lambda)$ employed in sNLM cannot be allowed to tend to zero arbitrarily fast as $(x,\lambda) \to (\bar x,\bar\lambda)$, as this may result in unsolvable iteration systems (23). At the moment, such examples are not known for the augmented Lagrangian method. In other words, the

augmented Lagrangian subproblem may possess some "good" solutions whose "counterparts" are missing for the sNLM subproblem. On the other hand, the augmented Lagrangian subproblems (which are unconstrained optimization problems) are of course more difficult to solve than the sNLM subproblems (which are linear systems).

In summary, there exist good local dual stabilization procedures. But this is not the happy end of the story, precisely because they are local. For those procedures to make the desired effect, they must be initialized close enough to a noncritical multiplier. The role of critical multipliers in the context of global convergence of algorithms is discussed next.

5 Critical multipliers and global behavior of algorithms

Augmented Lagrangian methods are very appealing in the context of degenerate problems because they combine (local) dual stabilization discussed above with strong global convergence properties; see [2–4,36]. Note that, from (27), it readily follows that every accumulation point $(\bar{x}, \bar{\lambda})$ of any primal-dual iterative sequence $\{(x^k, \lambda^k)\}$ generated by such methods satisfies the first equation in the Lagrange optimality system (2). Therefore, if $\bar{x}$ is also feasible, it is automatically a stationary point of problem (1), and $\bar{\lambda}$ is an associated Lagrange multiplier. This observation does not rely on any constraint qualification-type conditions [49]. However, those conditions do come into play when one wants to show the existence of accumulation points of dual sequences. In addition, the penalty parameter has to be controlled appropriately, in order to reduce the chances of convergence to infeasible points. In any case, theoretical and numerical results in [36] put in evidence that augmented Lagrangian methods are a good choice for solving (potentially) degenerate problems, especially when the primary concern is not speed but rather "quality of the outcome" (that is, the tendency of the algorithm to successfully terminate at a good approximate solution, rather than fail or terminate close to a nonoptimal stationary point). But things are not as good if one is concerned with speed. Superlinear convergence of the augmented Lagrangian method can be expected only when the inverse penalty parameter is asymptotically driven to zero. The latter leads to ill-conditioned subproblems which are difficult to solve. In addition, there exists some theoretical and numerical evidence of attraction to critical multipliers for the augmented Lagrangian method as well [24,40], when the dual starting point is arbitrary. Although, the consequences of this for the convergence rate are not so persistently bad as in the case of Newton-type methods. Perhaps, the reason is the exact penalization property of the augmented Lagrangian when the dual variable coincides with a true Lagrange multiplier. However, in any case, superlinear convergence rate of major iterations of the augmented Lagrangian method usually does not compensate for high computational costs of those iterations. One possible solution might be switching to some cheaper Newton-type iterations at the final stage of the process, and in fact, this strategy is currently implemented in the ALGENCAN solver [1,

7], where the method for the final stage is precisely NLM [6]. For degenerate problems, one might consider replacing NLM by sNLM, but usually this does not seem to give any serious benefits [40]. The reason is precisely that the final phase sNLM algorithm is usually triggered "too late", near a critical multiplier to which the augmented Lagrangian dual iterates have already been attracted in the previous "global" phase. This nullifies potential fast convergence of the final sNLM phase. Thus, critical multipliers have negative influence in this context as well, not allowing to combine in a practically efficient scheme good local convergence properties of sNLM with good global convergence properties of the augmented Lagrangian method. The situation is similar for the augmented Lagrangian method itself: while close to noncritical multipliers convergence is fast, the problem is that often enough (when from arbitrary dual starting points) dual iterates are still steered towards critical multipliers.

In view of the preceding discussion, it is natural to look for other globalization strategies for sNLM. In fact, this has been a matter of interest for several research groups. One possibility is to combine sNLM with the augmented Lagrangian method in a more sophisticated way, trying to employ stabilized steps not only at the final phase but as often as possible. Recall that sNLM is a "linearization" of the augmented Lagrangian method. Thus, one can try to use sNLM directions when solving the augmented Lagrangian subproblem (25), motivated by the understanding that the two methods are doing "more-or-less the same thing" at each step, at least when the linearization (23) approximates well the primal-dual iteration system (28) of the augmented Lagrangian method. The idea of combining some stabilized Newton-type steps for the Lagrange optimality system with augmented Lagrangian methods dates back at least to [18] (see also [5, p. 240]). This idea has found its further development in the very recent works [16,17], employing the so-called primal-dual augmented Lagrangian. The usual augmented Lagrangian is used in [38].

Furthermore, combining with augmented Lagrangian methods is, of course, not the only possibility to globalize sNLM. Hybrid globalizations employ the standard globalized NLM as an outer-phase algorithm, trying to switch to the full-step sNLM when convergence to a degenerate solution is detected. But, in our experience, the positive effect of such switches is usually again nullified by the attraction of the outer-phase algorithm to critical multipliers [26]. Another attempt to globalize sSQP was suggested in [9], where it was combined with the inexact restoration method [43,44]. However, as a globalization of sSQP, this strategy is somewhat indirect; it can be more naturally viewed as using sSQP to solve the subproblems of the inexact restoration method. In particular, it is not known whether close to a solution one sSQP step per iteration is enough or not (and thus, it is not known whether the method reduces to the usual sSQP with fast convergence).

Yet another approach would be to find or construct merit functions for which the sNLM direction is a direction of descent. This, however, proved difficult. It turns out that penalty functions employed for globalization of the usual SQP (such as the l_1-penalty, for example) are not suitable. A promising

recent proposal is based on a certain smooth two-parameter primal-dual exact penalty function from [8], see [37]. This development is currently for the equality-constrained case only.

Much work remains to be done. At this time, we are not aware of numerical results convincingly showing that any of the globalized versions of sNLM/sSQP steadily and reliably outperforms the usual NLM/SQP globalized in standard ways, on degenerate problems. (The experience in [37] shows that this globalization is reasonable, and preferable for certain types of degenerate problems. It is, however, for equality constraints only.) As discussed above, NLM/SQP is almost always attracted by critical multipliers when they exist (which is the typical situation), thus losing superlinear convergence. While for the already available globalizations of sNLM/sSQP the effect of attraction is actually not so persistent, and superlinear convergence does in fact show up, in many cases it does not compensate for extra computational costs needed to achieve it, and/or for the still quite frequent cases of convergence to critical multipliers. Morerover, in the latter cases sNLM may converge even slower than NLM. Thus, developing really satisfactory globalization techniques for sSQP, or more generally, algorithms efficient for the degenerate case, continues to be an important challenge, with many open questions. The effect of attraction to critical Lagrange multipliers is the key difficulty which, depending on the algorithm, can manifest itself locally, "semi-locally", or globally. The goal should be either to develop some practical (i.e., not prohibitively expensive) global dual stabilization techniques allowing to avoid the attraction phenomenon completely (or at least typically), or to improve efficiency in the cases of convergence to critical multipliers.

References

1. http://www.ime.usp.br/~egbirgin/tango/.
2. Andreani R, Birgin EG, Martínez JM, Schuverdt ML (2007) On augmented Lagrangian methods with general lower-level constrints. SIAM J Optim 18:1286–1309
3. Andreani R, Birgin EG, Martínez JM, Schuverdt ML (2008) Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. Math Program 111:5–32
4. Andreani R, Haeser G, Schuverdt ML, Silva PJS (2012) A relaxed constant positive linear dependence constraint qualification and applications. Math Program 135:255–273
5. Bertsekas DP (1982) Constrained optimization and Lagrange multiplier methods. Academic Press, New York
6. Birgin EG, Martínez JM (2008) Improving ultimate convergence of an augmented Lagrangian method. Optim Meth Software 23:177–195
7. Birgin EG, Martínez JM (2014) Practical augmented Lagrangian methods for constrained optimization. SIAM, Philadelphia
8. Di Pillo G, Grippo L (1979) A new class of augmented Lagrangians in nonlinear programming. SIAM J Control Optim 17:618–628
9. Fernández D, Pilotta EA, Torres GA (2013) An inexact restoration strategy for the globalization of the sSQP method. Comput Optim Appl 54:595–617
10. Fernández D, Izmailov AF, Solodov MV (2010) Sharp primal superlinear convergence results for some Newtonian methods for constrained optimization. SIAM J Optim 20:3312–3334

11. Fernández D, Solodov M (2010) Stabilized sequential quadratic programming for optimization and a stabilized Newton-type method for variational problems. Math Program 125:47–73
12. Fernández D, Solodov M (2012) Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. SIAM J Optim 22:384–407
13. Fischer A (2002) Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math Program 94:91–124
14. Friedlander MP, Saunders MA (2005) A globally convergent linearly constrained Lagrangian method for nonlinear optimization. SIAM J Optim 15:863–897
15. Gill PE, Murray W, Saunders MA (2002) SNOPT: An SQP algorithm for large-scale constrained optimization. SIAM J Optim 12:979–1006
16. Gill PE, Robinson DP (2012) A primal-dual augmented Lagrangian. Comput Optim Appl 51:1–25
17. Gill PE, Robinson DP (2013) A globally convergent stabilized SQP method. SIAM J Optim 23:1983–2010
18. Glad ST (1979) Properties of updating methods for the multipliers in augmented Lagrangian. J Optim Theory Appl 28:135–156
19. Golubitsky M, Schaeffer DG (1985) Singularities and groups in bifurcation theory. Vol. 1. Springer-Verlag, New York, Berlin, Heidelberg
20. Hager WW (1999) Stabilized sequential quadratic programming. Comput Optim Appl 12:253–273
21. Hager WW, Gowda MS (1999) Stability in the presence of degeneracy and error estimation. Math Program 85:181–192
22. Hestenes MR (1969) Multiplier and gradient methods. J Optim Theory Appl 4:303–320
23. Izmailov AF (2005) On the analytical and numerical stability of critical Lagrange multipliers. Comput Math Math Phys 45:930–946
24. Izmailov AF (2011) On the limiting properties of dual trajectories in the Lagrange multipliers method. Comput Math Math Phys 51:3–23
25. Izmailov AF (2010) Solution sensitivity for Karush–Kuhn–Tucker systems with nonunique Lagrange multipliers. Optimization 59:747–775
26. Izmailov AF, Krylova AM, Uskov EI (2011) Hybrid globalization of stabilized sequential quadratic programming method. In Russian. In: Bereznyov VA (ed) Theoretical and applied problems of nonlinear analysis. Computing Center RAS, Moscow, pp 47–66
27. Izmailov AF, Kurennoy AS (2013) Abstract Newtonian frameworks and their applications. SIAM J Optim 23:2369–2396
28. Izmailov AF, Kurennoy AS, Solodov MV (2013) A note on upper Lipschitz stability, error bounds, and critical multipliers for Lipschitz-continuous KKT systems. Math Program 142:591–604
29. Izmailov AF, Kurennoy AS, Solodov MV (2013) Local convergence of the method of multipliers for variational and optimization problems under the noncriticality assumption. Comput Optim Appl 60:111–140
30. Izmailov AF, Solodov MV (2009) Examples of dual behaviour of Newton-type methods on optimization problems with degenerate constraints. Comput Optim Appl 42:231–264
31. Izmailov AF, Solodov MV (2009) On attraction of Newton-type iterates to multipliers violating second-order sufficiency conditions. Math Program 117:271–304
32. Izmailov AF, Solodov MV (2011) On attraction of linearly constrained Lagrangian methods and of stabilized and quasi-Newton SQP methods to critical multipliers. Math Program 126:231–257
33. Izmailov AF, Solodov MV (2012) Stabilized SQP revisited. Math Program 122:93–120
34. Izmailov AF, Solodov MV (2014) Newton-type methods for optimization and variational problems. Springer Series in Operations Research and Financial Engineering
35. Izmailov AF, Solodov MV (2015) Newton-type methods: a broader view. J Optim Theory Appl 164:577–620
36. Izmailov AF, Solodov MV, Uskov E.I (2012) Global convergence of augmented Lagrangian methods applied to optimization problems with degenerate constraints, including problems with complementarity constraints. SIAM J Optim 22:1579–1606
37. Izmailov AF, Solodov MV, Uskov E.I (2014) Globalizing stabilized SQP by smooth primal-dual exact penalty function. IMPA preprint A752/2014, Rio de Janero

38. Izmailov AF, Solodov MV, Uskov E.I (2014) Combining stabilized SQP with the augmented Lagrangian algorithm. IMPA preprint A754/2014, Rio de Janero
39. Izmailov AF, Uskov E.I (2014) Attraction of Newton method to critical Lagrange multipliers: fully quadratic case. Math Program. doi:10.1007/s10107-014-0777-x
40. Izmailov AF, Uskov E.I (2012) On the influence of the critical Lagrange multipliers on the convergence rate of the multiplier method. Comput Math Math Phys 52:1504–1519
41. Izmailov AF, Uskov E.I (2012) The effect of attraction of the Newton–Lagrange method to critical Lagrange multipliers: full analysis in the one-dimensional case. In Russian. In: Bereznyov VA (ed) Theoretical and applied problems of nonlinear analysis. Computing Center RAS, Moscow, pp 53–71
42. Li D-H, Qi L (2000) Stabilized SQP method via linear equations. Applied Mathematics Technical Reptort AMR00/5, University of New South Wales, Sydney
43. Martínez JM, Pilotta EA (2000) Inexact restoration algorithms for constrained optimization. J Optim Theory Appl 104:135–163
44. Martínez JM, Pilotta EA (2005) Inexact restoration methods for nonlinear programming: advances and perspectives. In: Optimization and control with applications. Springer, New York, pp 271–292
45. Murtagh BA, Saunders MA (1982) A projected Lagrangian algorithm and its implementation for sparse nonlinear constraints. Math Program Study 16:84–117
46. Murtagh BA, Saunders MA (1983) MINOS 5.0 user's guide. Technical Report SOL 83.20, Stanford University
47. Powell MJD (1969) A method for nonlinear constraints in minimization problems. In: Fletcher R (ed) Optimization. Academic Press, New York, pp 283–298
48. Robinson SM (1972) A quadratically convergent algorithm for general nonlinear programming problems. Math Program 3:145–156
49. Solodov MV (2010) Constraint qualifications. In: Cochran JJ (ed) Wiley encyclopedia of operations research and management science. John Wiley & Sons, Inc
50. Uskov EI (2013) On the attraction of Newton method to critical Lagrange multipliers. Comp Math Math Phys 53:1099–1112
51. Wright SJ (1998) Superlinear convergence of a stabilized SQP method to a degenerate solution. Comput Optim Appl 11:253–275