

R U T C O R
R E S E A R C H
R E P O R T

A PRACTICAL
RELATIVE ERROR CRITERION
FOR AUGMENTED LAGRANGIANS

Jonathan Eckstein ^a Paulo J. S. Silva ^b

RRR 11-2010, JULY 19, 2010

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
<http://rutcor.rutgers.edu/~rrr>

^aDepartment of Management Science and Information Systems and RUTCOR, 640 Bartholomew Road, Busch Campus, Rutgers University, Piscataway NJ 08854 USA, jeckstei@rci.rutgers.edu.

^bDepartment of Computer Science, Rua do Matão, 1010, University of São Paulo, CEP: 05508-090, São Paulo, SP, Brazil. pjssilva@ime.usp.br

RUTCOR RESEARCH REPORT

RRR 11-2010, JULY 19, 2010

A PRACTICAL RELATIVE ERROR CRITERION FOR AUGMENTED LAGRANGIANS

Jonathan Eckstein

Paulo J. S. Silva

Abstract. This paper develops a new error criterion for the approximate minimization of augmented Lagrangian subproblems. This criterion is practical in the sense that it requires only information that is ordinarily readily available, such as the gradient (or a subgradient) of the augmented Lagrangian. It is also “relative” in the sense of relative error criteria for proximal point algorithms, in that it is based on the relative magnitude of two quantities and requires only a single parameter, not the choice of an theoretically infinite sequences of parameters. It involves a novel auxiliary sequence that appears only in the approximation criterion, and not in the augmented Lagrangian minimand, nor in the multiplier update. We give a proof of the global convergence of our method in the setting of the abstract convex duality framework of Rockafellar, along with some more concrete applications. The dual convergence result is slightly weaker than usually obtained for multiplier methods, but may be strengthened by enforcing an additional condition in the algorithm. We give some computational results drawn from the CUTE test set, indicating that our approach works well in practice.

Acknowledgements: Jonathan Eckstein’s work was partially supported by Rutgers Business School Research Resources Committee grants. Paulo J. S. Silva was partially supported by CNPq (grants 303030/2007-0 and 474138/2008-9) and PRONEX–Optimization.

1 Introduction, motivation, and summary

The past 10-15 years, starting with [22, 23, 24], have seen the development of a number of relative error criteria for approximately solving the subproblems arising in proximal point algorithms. The advantage of these criteria is that they set the exactness tolerance for each subproblem in a manner sensitive to the algorithm's convergence on each particular problem instance, and involve only a single scalar parameter; by contrast, more traditional approximation criteria like those originally proposed in [21] involve a theoretically infinite sequence of error tolerance parameters $\{\epsilon_k\} \subset [0, \infty)$, and provide no direct guidance as to how to select it, except for requiring that $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Unfortunately, however, neither the traditional nor the existing relative error criteria apply readily to one of the most important application of proximal point algorithms, augmented Lagrangian algorithms, also called *methods of multipliers*. Consider the following two simple convex optimization problems:

$$\begin{array}{ll} \min & f(x) \\ \text{ST} & h(x) = 0 \end{array} \quad (1)$$

and

$$\begin{array}{ll} \min & f(x) \\ \text{ST} & g(x) \leq 0, \end{array} \quad (2)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is affine, and $g(x) = (g_1(x), \dots, g_m(x))$, where $g_1, \dots, g_m : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex. Let us assume for the purposes of these simple examples that f and g are continuously differentiable. Applying the proximal point algorithm to the duals of these problems, we obtain as described in [20] the respective methods of multipliers

$$x^k \in \text{Arg min}_{x \in \mathbb{R}^n} \left\{ f(x) + \langle p^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 \right\} \quad (3)$$

$$p^k = p^{k-1} + c_k h(x^k) \quad (4)$$

and

$$x^k \in \text{Arg min}_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max \{0, p^{k-1} + c_k g_i(x)\}^2 \right\} \quad (5)$$

$$p^k = \max \{0, p^{k-1} + c_k g(x^k)\}, \quad (6)$$

where $\{c_k\} \subset (0, \infty)$ is a sequence with $\inf_k \{c_k\} > 0$, and the “max” operation in (6) is interpreted componentwise. If one applies the proximal point algorithm to primal-dual formulations of (1) and (2), one instead obtains — again, see [20] — the respective *proximal* methods of multipliers

$$x^k \in \text{Arg min}_{x \in \mathbb{R}^n} \left\{ f(x) + \langle p^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 + \frac{1}{2c_k} \|x - x^{k-1}\|^2 \right\} \quad (7)$$

$$p^k = p^{k-1} + c_k h(x^k) \quad (8)$$

and

$$x^k \in \operatorname{Arg} \min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max \{0, p^{k+1} + c_k g_i(x)\}^2 + \frac{1}{2c_k} \|x - x^{k-1}\|^2 \right\} \quad (9)$$

$$p^k = \max \{0, p^{k-1} + c_k g(x^k)\}. \quad (10)$$

That is, we obtain similar methods, but with an additional primal regularizing term of the form $(1/2c_k)\|x - x^{k-1}\|^2$ in the minimand of each subproblem. Approximation criteria for abstract proximal point algorithms tend to translate straightforwardly into implementable criteria for approximately performing the minimizations (7) and (9), but this is not the case for the far more commonly used augmented Lagrangian minimizations (3) or (5). This situation is unfortunate, since methods such as (3)-(4) and (5)-(6) tend to be faster than their regularized cousins (7)-(8) and (9)-(10), and are much more often used in practice; for a recent example of empirical results to this effect, see [9].

This paper develops implementable criteria for approximating “pure dual” augmented Lagrangian calculations like (3) and (5). These criteria are also “relative” in a sense similar to the abstract proximal point algorithm criteria of [22, 23, 24]: they involve the choice of one or two scalar parameters, rather than an indefinite sequence of parameters, and allow the problem instance to “guide” the gradual tightening of the approximation tolerance as the algorithm proceeds. Our recent computational results in [9], using a heuristic approximation rule of this type, were very promising; however, a formal convergence proof was lacking, even in the convex case. Here, we will develop a related but different algorithm and prove that it is globally convergent in the convex case. We will conclude by presenting some empirical computational results on standard nonconvex test problems from the CUTE test set [4], showing that our algorithm’s performance is essentially identical to the heuristic rule of [9].

Our analytical approach draws on [8], and makes similar use of Rockafellar’s general convex duality framework; see [17, Chapters 29-30] and [19]. The analysis in [8] derives an augmented Lagrangian error criterion that is practical in the same sense we use here, in that it requires only readily available information such as the gradient of the augmented Lagrangian; however, [8] developed only an absolute error criterion with a summable sequence $\{\epsilon_k\}$ of error parameters, and does not provide direct guidance as to how to select this sequence. By contrast, the analysis here will develop true relative error criteria with only one or two scalar parameters.

Section 2 will briefly review the the generalized duality framework of [17, 19], apparently required to derive error criteria requiring only knowledge of the augmented Lagrangian gradient, showing how to apply the general framework to several specific problem formulations. To give a preview of the main results, consider for example the simple inequality-constrained problem (2). In this case, the framework we will develop reduces to the following algorithmic structure: define y^k to be the gradient of the augmented Lagrangian with primal variables x^k and multipliers p^{k-1} :

$$y^k = \nabla_x \left[f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max \{0, p^{k-1} + c_k g_i(x)\}^2 \right]_{x=x^k}.$$

Rather than having to exactly minimize the augmented Lagrangian, which would be equivalent to computing x^k such that $y^k = 0$, we will instead only require that

$$\frac{2}{c_k} \|w^{k-1} - x^k\| \|y^k\| + \|y^k\|^2 \leq \sigma \left\| \min \left\{ \frac{1}{c_k} p^{k-1}, -g(x^k) \right\} \right\|^2, \quad (11)$$

where w^{k-1} is a certain auxiliary vector, and $\sigma \in [0, 1)$ is a scalar parameter. The quantity on the left of (11) clearly vanishes as x^k approaches the set of minimizers of the augmented Lagrangian and y^k thus approaches 0, while the quantity $\left\| \min \left\{ (1/c_k) p^{k-1}, -g(x^k) \right\} \right\|^2$ on the right-hand side is a measure of the degree to which feasibility and complementary slackness are currently violated. Thus, rather than having to drive the augmented Lagrangian gradient to zero in order to update the multipliers, we need only reduce it within a tolerance dictated by how close we are to satisfying the KKT conditions. In the case of problem (2), the full algorithmic framework we will develop below reduces to the following set of recursive conditions, with arbitrary starting values $p^0 \in \mathbb{R}^m$ and $w^0 \in \mathbb{R}^n$:

$$y^k = \nabla_x \left[f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max \{0, p^{k-1} + c_k g_i(x)\}^2 \right]_{x=x^k} \quad (12)$$

$$\frac{2}{c_k} \|w^{k-1} - x^k\| \|y^k\| + \|y^k\|^2 \leq \sigma \left\| \min \left\{ \frac{1}{c_k} p^{k-1}, -g(x^k) \right\} \right\|^2 \quad (13)$$

$$p^k = \max \{0, p^{k-1} + c_k g(x^k)\} \quad (14)$$

$$w^k = w^{k-1} - c_k y^k. \quad (15)$$

Here, (12)-(13) just express our specific form of approximate minimization of the augmented Lagrangian and (14) is just the usual multiplier update for the class of problems (2), but (15), along with the way the sequence $\{w^k\}$ appears in (13), is a novel feature. Relative-error variants of the proximal point algorithm typically involve some kind of ‘‘corrector’’ to the basic proximal step, either a projection as in [23], or an extragradient step as in [22, 24]. Here, (15) appears to fulfill this role, but in an unusual manner, since w^k plays no direct role in either the subproblem objective function of (12) or the multiplier update (14). Instead, it only appears in the approximation condition (13) and the extragradient-like auxiliary update (15). Note that if we are able to minimize all the augmented Lagrangians exactly, and thus obtain $y^k = 0$ for all k , then $\{w^k\}$ would simply be a constant sequence. The sequence $\{w^k\}$ appears to play the role of tracking the total ‘‘drift’’ in the sequence of calculations, something novel in augmented Lagrangian algorithms.

Section 3 will develop a much more general version of the framework (12)-(15), and Section 4 will establish its convergence. The fundamental convergence result is slightly weaker than traditionally obtained for methods like (5)-(6), in that we do not show convergence of $\{p^k\}$ to a unique limit; however, we do show that the dual sequence $\{p^k\}$ is bounded and all its limit points are dual solutions, the sequence $\{x^k\}$ is asymptotically optimal, and all limit points of $\{x^k\}$ are primal solutions. We will also show that a stronger result, asserting full convergence of $\{p^k\}$ and akin to those typically obtained for multiplier methods,

may be obtained by enforcing a second approximation condition in addition to a condition generalizing (13).

Finally, Section 5 gives an application of our framework to a more complicated, realistic formulation than (1) or (2), Section 6 gives computational results, and Section 7 presents some concluding remarks.

It also bears mention that there is an alternate thread of research involving approximate augmented Lagrangian minimization, as exemplified by analyses such as in [2, 3, 5, 6, 10], which directly analyzes convergence on nonconvex problems, but with compensating restrictive assumptions. It would also be interesting to examine our new algorithm in a similar light, but that is outside the scope of this paper.

2 General convex duality framework

We now briefly review the general convex duality framework from [17, Chapters 29-30] and [19]. We suppose that we have a closed (lower semicontinuous) convex function $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow (\infty, +\infty]$, and we wish to solve the *primal* problem

$$\min_{x \in \mathbb{R}^n} F(x, 0). \quad (16)$$

The second argument to F represents some kind of perturbation to the primal problem (16). The customary choice for modeling the equality-constrained problem (1) is

$$F(x, u) = \begin{cases} f(x), & \text{if } h(x) + u = 0 \\ +\infty, & \text{otherwise,} \end{cases} \quad (17)$$

while for the inequality-constrained problem (2) one typically makes the similar choice

$$F(x, u) = \begin{cases} f(x), & \text{if } g(x) + u \leq 0 \\ +\infty, & \text{otherwise.} \end{cases} \quad (18)$$

Further, $\partial F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ denotes the subgradient mapping of F . We define Q to be the concave conjugate of F , that is

$$Q(y, p) = \inf_{\substack{x \in \mathbb{R}^n \\ u \in \mathbb{R}^m}} \{F(x, u) - \langle x, y \rangle - \langle u, p \rangle\},$$

and the *dual* problem to (16) to be

$$\max_{p \in \mathbb{R}^m} Q(0, p). \quad (19)$$

A simple application of Fenchel's inequality — see for example [17, Theorem 23.5] — shows that weak duality holds, that is, $Q(0, p) \leq F(x, 0)$ for all $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^m$. Q is a closed

(upper semicontinuous) concave function, and we let $\partial Q : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ denote its subgradient map (the negative of its supergradient map), that is,

$$(x, u) \in \partial Q(y, p) \Leftrightarrow Q(y', p') \leq Q(y, p) - \langle x, y' - y \rangle - \langle u, p' - p \rangle \quad \forall y' \in \mathbb{R}^n, p' \in \mathbb{R}^m. \quad (20)$$

We define $L : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [-\infty, \infty]$ to be the function obtained by taking the concave conjugate of F with respect to only its second argument, that is,

$$L(x, p) = \inf_{u \in \mathbb{R}^m} \{F(x, u) - \langle u, p \rangle\}.$$

If we compute L for the choice of F given in (17), we obtain

$$L(x, p) = f(x) + \langle p, h(x) \rangle, \quad (21)$$

which is the Lagrangian ordinarily associated with problem (1). For the choice of F in (2), we obtain

$$L(x, p) = \begin{cases} f(x) + \langle p, g(x) \rangle, & p \geq 0, \\ -\infty, & \text{otherwise,} \end{cases} \quad (22)$$

which expresses the usual Lagrangian for the inequality-constrained problem (2), along with the requirement that the Lagrange multipliers p be nonnegative. By analogy, one in general calls L the *Lagrangian* corresponding to (16). L is convex in its first argument, and concave in the second, and we let ∂L denote its subgradient map, that is,

$$(y, u) \in \partial L(x, p) \Leftrightarrow \begin{cases} L(x', p) \geq L(x, p) + \langle y, x' - x \rangle & \forall x' \in \mathbb{R}^n \\ L(x, p') \leq L(x, p) - \langle u, p' - p \rangle & \forall p' \in \mathbb{R}^m. \end{cases}$$

The point-to-set maps ∂F , ∂Q , and ∂L are all maximal monotone operators, and

$$(y, p) \in \partial F(x, u) \Leftrightarrow (y, u) \in \partial L(x, p) \Leftrightarrow (x, u) \in \partial Q(y, p), \quad (23)$$

that is, ∂F and ∂Q are inverses of one another, and ∂L is a *partial inverse* [25] of both ∂F and ∂Q . If $(x^*, p^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is such that $(0, 0) \in \partial L(x^*, p^*)$, then x^* solves the primal problem (16) and p^* solves the dual problem (19). In this case, we say that (x^*, p^*) is a *KKT pair*. If a KKT pair exists, then strong duality holds, that is, $F(x^*, 0) = Q(0, p^*)$ and thus the optimal values of the primal and dual problems (16) and (19) exist and are equal.

3 An abstract approximate method of multipliers

We now formulate a set of recursions analogous to (12)-(15), but in the much more general setting of the abstract problem (16). Specifically, for some $\sigma \in [0, 1)$, we suppose we have sequences $\{x^k\}_{k=1}^\infty, \{y^k\}_{k=1}^\infty, \{w^k\}_{k=0}^\infty \subset \mathbb{R}^n$, and $\{p^k\}_{k=0}^\infty \subset \mathbb{R}^m$ satisfying for all $k \geq 1$ the conditions

$$(y^k, \frac{1}{c_k}(p^{k-1} - p^k)) \in \partial L(x^k, p^k) \quad (24)$$

$$2c_k \|w^{k-1} - x^k\| \|y^k\| + c_k^2 \|y^k\|^2 \leq \sigma \|p^{k-1} - p^k\|^2 \quad (25)$$

$$w^k = w^{k-1} - c_k y^k. \quad (26)$$

At this point, (25) and (26) may seem somewhat unmotivated; we will attempt to provide insight into these choices later, as we proceed with the analysis. For the moment, we focus on the abstract condition (24). Consider first the simple case of the equality constrained problem (1), with F chosen as in (17): in this case, using (21), we obtain that $\partial L(x, p)$ is the singleton set given by

$$\partial L(x, p) = \left\{ \begin{bmatrix} \nabla f(x) + \nabla h(x)^\top p \\ -h(x) \end{bmatrix} \right\},$$

and so (24) reduces to

$$y^k = \nabla f(x^k) + \nabla h(x^k)^\top p^k \tag{27}$$

$$\frac{1}{c_k}(p^{k-1} - p^k) = -h(x^k). \tag{28}$$

Solving (28) for p^k yields $p^k = p^{k-1} + c_k h(x^k)$, that is, precisely the usual multiplier update (4) for equality constraints. Substituting this value of p^k into (27), we obtain

$$\begin{aligned} y^k &= \nabla f(x^k) + \nabla h(x^k)^\top (p^{k-1} + c_k h(x^k)) \\ \Leftrightarrow y^k &= \nabla f(x^k) + \nabla h(x^k)^\top p^{k-1} + c_k \nabla h(x^k)^\top h(x^k) \\ \Leftrightarrow y^k &= \nabla f(x^k) + \nabla_x [\langle p^{k-1}, h(x) \rangle]_{x=x^k} + \nabla_x \left[\frac{c_k}{2} \|h(x)\|^2 \right]_{x=x^k} \\ \Leftrightarrow y^k &= \nabla_x \left[f(x) + \langle p^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 \right]_{x=x^k}. \end{aligned}$$

Thus, we have in this case that the abstract condition (24) is simply equivalent to y^k being the gradient of the usual augmented Lagrangian of (3) at x^k , with p^k being obtained by the usual multiplier update (4). Substituting $p^k = p^{k-1} + c_k h(x^k)$ into (25) and dividing by c_k^2 , we obtain that for the simple equality constrained problem (1) and the corresponding customary choice (17) that the abstract recursions (24)-(26) reduce to

$$y^k = \nabla_x \left[f(x) + \langle p^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 \right]_{x=x^k} \tag{29}$$

$$\frac{2}{c_k} \|w^{k-1} - x^k\| \|y^k\| + \|y^k\|^2 \leq \sigma \|h(x^k)\|^2 \tag{30}$$

$$p^k = p^{k-1} + c_k h(x^k) \tag{31}$$

$$w^k = w^{k-1} - c_k y^k. \tag{32}$$

Comparing these recursions to the exact multiplier method (3)-(4), we do not reduce the gradient of that augmented Lagrangian to 0, as in (3), but only sufficiently that the left-hand side of (30) drops below a threshold proportional to the current constraint violation. We then update the multipliers in the usual manner, and perform the (somewhat curious) update $w^k = w^{k-1} - c_k y^k$. Note that if we set $\sigma = 0$, we would force exact minimization of every augmented Lagrangian, and obtain exactly the same sequence of iterates $\{x^k\}$ and $\{p^k\}$

as (3)-(4), with $\{w^k\}$ being a constant sequence. So, (29)-(32) is a generalization of (3)-(4) with relative-error approximate minimization of the augmented Lagrangian.

We now repeat the above exercise, but for the inequality-constrained problem (2) and the corresponding customary choice (18). In this case, we obtain

$$\begin{aligned}\partial L(x, p) &= \{\nabla f(x) + \nabla g(x)^\top p\} \times (-g(x) + N_{\mathbb{R}_+^m}(p)) \\ &= \{\nabla f(x) + \nabla g(x)^\top p\} \times \{-g(x) + q \mid q \leq 0, \langle p, q \rangle = 0\},\end{aligned}$$

where $N_{\mathbb{R}_+^m}$ denotes the normal cone mapping of the nonnegative orthant in \mathbb{R}^m . Using this definition in (24), we obtain

$$y^k = \nabla f(x^k) + \nabla g(x^k)^\top (p^k) \quad (33)$$

$$\frac{1}{c_k}(p^{k-1} - p^k) \in -g(x^k) + N_{\mathbb{R}_+^m}(p^k). \quad (34)$$

Manipulating (34), we obtain the condition

$$(p^{k-1} + c_k g(x^k)) - p^k \in N_{\mathbb{R}_+^m}(p^k),$$

which is equivalent to p^k being the unique projection of $p^{k-1} + c_k g(x^k)$ onto \mathbb{R}_+^m , that is,

$$p^k = \max \{0, p^{k-1} + c_k g(x^k)\}.$$

Thus, we have obtained exactly the usual inequality multiplier update (6). Substituting this expression for p^k into (33) and manipulating the result in much the same manner as in the previous example yields

$$y^k = \nabla_x \left[f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max \{0, p^{k-1} + c_k g_i(x)\}^2 \right]_{x=x^k},$$

which is exactly the same as (12). Finally, substituting the right-hand side of (6) into (25) and dividing by c_k^2 produces

$$\frac{2}{c_k} \|w^{k-1} - x^k\| \|y^k\| + \|y^k\|^2 \leq \sigma \left\| \min \left\{ \frac{1}{c_k} p^{k-1}, -g(x^k) \right\} \right\|^2,$$

which is identical to (13). Thus, we conclude that in the case of the inequality-constrained problem (2) and its corresponding customary choice (18), the abstract approximate multiplier method (24)-(26) reduces exactly to the example method (12)-(15) described in Section 1.

The same basic mode of analysis may be used to specialize (12)-(15) to many other kinds of convex problems, including those involving mixtures of equality and inequality constraints (see Section 5), general conic constraints (for example, for cones of semidefinite matrices), and nonsmooth functions.

4 Convergence proof for the abstract method

Proposition 1 *With F , Q , and L defined as in Section 2, suppose for some $\sigma \in [0, 1)$ and $\{c_k\}_{k=1}^\infty$ with $\inf_{k \geq 1} \{c_k\} > 0$ that the sequences $\{x^k\}_{k=1}^\infty, \{y^k\}_{k=1}^\infty, \{w^k\}_{k=0}^\infty \subset \mathbb{R}^n$ and $\{p^k\}_{k=0}^\infty \subset \mathbb{R}^m$ obey for all $k \geq 1$ the recursions (24)-(26). Define, for all $k \geq 1$,*

$$u^k = \frac{1}{c_k}(p^{k-1} - p^k). \quad (35)$$

Then, if a KKT pair exists, the following hold:

- *The sequences $\{p^k\}$ and $\{w^k\}$ must be bounded.*
- *$u^k \rightarrow 0$ and $y^k \rightarrow 0$.*
- *$F(x^k, u^k)$ and $Q(y^k, p^k)$ both converge to the common optimal value of the primal and dual problems (16) and (19).*
- *All accumulation points of $\{x^k\}$ are solutions to the primal problem (16) and all accumulation points of $\{p^k\}$ are solutions to the dual problem (19).*

If no KKT pair exists, then at least one of the sequences $\{p^k\}$ or $\{x^k\}$ must be unbounded.

Proof. First, we consider the case that some KKT pair exists, and let (x^*, p^*) be any such pair. For any $k \geq 1$,

$$\begin{aligned} \|p^{k-1} - p^*\|^2 &= \|p^{k-1} - p^k + p^k - p^*\|^2 \\ &= \|p^{k-1} - p^k\|^2 + 2\langle p^{k-1} - p^k, p^k - p^* \rangle + \|p^k - p^*\|^2. \end{aligned} \quad (36)$$

From the definition of u^k , we have $p^{k-1} - p^k = c_k u^k$, which we may substitute into (36) to obtain

$$\|p^{k-1} - p^*\|^2 = \|p^{k-1} - p^k\|^2 + 2c_k \langle u^k, p^k - p^* \rangle + \|p^k - p^*\|^2,$$

which may be rearranged into

$$\|p^k - p^*\|^2 = \|p^{k-1} - p^*\|^2 - 2c_k \langle u^k, p^k - p^* \rangle - \|p^{k-1} - p^k\|^2 \quad (37)$$

Next, using that $w^k = w^{k-1} - c_k y^k$, we perform a similar expansion of $\|w^k - x^*\|^2$:

$$\begin{aligned} \|w^k - x^*\|^2 &= \|w^{k-1} - c_k y^k - x^*\|^2 \\ &= \|w^{k-1} - x^*\|^2 - 2\langle w^{k-1} - x^*, c_k y^k \rangle + c_k^2 \|y^k\|^2 \\ &= \|w^{k-1} - x^*\|^2 - 2c_k \langle w^{k-1} - x^k + x^k - x^*, y^k \rangle + c_k^2 \|y^k\|^2 \\ &= \|w^{k-1} - x^*\|^2 - 2c_k \langle w^{k-1} - x^k, y^k \rangle - 2c_k \langle x^k - x^*, y^k \rangle + c_k^2 \|y^k\|^2. \end{aligned} \quad (38)$$

Next, we add (37) and (38) to obtain

$$\begin{aligned} \|p^k - p^*\|^2 + \|w^k - x^*\|^2 &= \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2 \\ &\quad - 2c_k [\langle x^k - x^*, y^k \rangle + \langle p^k - p^*, u^k \rangle] \\ &\quad - 2c_k \langle w^{k-1} - x^k, y^k \rangle + c_k^2 \|y^k\|^2 \\ &\quad - \|p^{k-1} - p^k\|^2. \end{aligned} \quad (39)$$

Next, we use the monotonicity of ∂L to eliminate the expression on the second line of (39). Specifically, from (24) and (35) we have that $(y^k, u^k) \in \partial L(x^k, p^k)$, and since (x^*, p^*) is a KKT pair, we also have $(0, 0) \in \partial L(x^*, p^*)$. Thus, the monotonicity of ∂L yields

$$\langle x^k - x^*, y^k - 0 \rangle + \langle p^k - p^*, u^k - 0 \rangle = \langle x^k - x^*, y^k \rangle + \langle p^k - p^*, u^k \rangle \geq 0. \quad (40)$$

Combining this inequality with (39) yields

$$\begin{aligned} \|p^k - p^*\|^2 + \|w^k - x^*\|^2 &\leq \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2 \\ &\quad - 2c_k \langle w^{k-1} - x^k, y^k \rangle + c_k^2 \|y^k\|^2 \\ &\quad - \|p^{k-1} - p^k\|^2. \end{aligned} \quad (41)$$

The error criterion (25) is designed so that the terms on the second line of (41) may be “buried” in the last term. Specifically, using the Cauchy-Schwarz inequality in conjunction with (25), we have

$$-2c_k \langle w^{k-1} - x^k, y^k \rangle + c_k^2 \|y^k\|^2 \leq 2c_k \|w^{k-1} - x^k\| \|y^k\| + c_k^2 \|y^k\|^2 \leq \sigma \|p^{k-1} - p^k\|^2. \quad (42)$$

Substituting this inequality into (41), we obtain an inequality that is the key to the convergence analysis:

$$\|p^k - p^*\|^2 + \|w^k - x^*\|^2 \leq \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2 - (1 - \sigma) \|p^{k-1} - p^k\|^2. \quad (43)$$

Since (43) holds for all $k \geq 1$, a cascade of deductions follows:

- $\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\}$ is a nonincreasing sequence, so in particular the sequences $\{p^k\}$ and $\{w^k\}$ must be bounded. Since $\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\}$ is bounded below by 0, it must be convergent. Since (x^*, p^*) was an arbitrary KKT pair, we further conclude that $\{(x^k, p^k)\}$ is Fejér monotone to the set of KKT pairs.
- Summing (43) over k and using that $\sigma < 1$, we conclude that $\{\|p^{k-1} - p^k\|^2\}$ is a summable sequence and hence that $p^k - p^{k-1} \rightarrow 0$.
- Since c_k is bounded away from 0 and $u^k = (1/c_k)(p^k - p^{k-1})$, it follows that $u^k \rightarrow 0$ and $\{\|u^k\|^2\}$ is summable.

- Referring to (25) or (42), and using the just-established properties of the sequence $\{p^k - p^{k-1}\}$, it then also follows that the sequences $\{c_k \|w^{k-1} - x^k\| \|y^k\|\}$ and $\{c_k^2 \|y^k\|^2\}$ are both summable, and thus converge to 0.
- By the Cauchy-Schwarz inequality, it follows immediately that $\{c_k \langle w^{k-1} - x^k, y^k \rangle\}$ is summable and thus converges to 0.
- Again using that c_k is bounded away from 0, the last two sets of observations imply that $\{\|w^{k-1} - x^k\| \|y^k\|\}$, $\{\langle w^{k-1} - x^k, y^k \rangle\}$, and $\{\|y^k\|\}$ are all summable, and thus converge to 0, and in particular we have $y^k \rightarrow 0$.
- Writing

$$\langle x^k, y^k \rangle = \langle w^{k-1}, y^k \rangle - \langle w^{k-1} - x^k, y^k \rangle,$$

we note that since $\{w^k\}$ is bounded and $y^k \rightarrow 0$, we have $\langle w^{k-1}, y^k \rangle \rightarrow 0$. Since we have already established that $\langle w^{k-1} - x^k, y^k \rangle \rightarrow 0$, it follows that $\langle x^k, y^k \rangle \rightarrow 0$.

Next, applying (23) to the relation $(y^k, u^k) \in \partial L(x^k, p^k)$ gives that $(x^k, u^k) \in \partial Q(y^k, p^k)$. Thus, the subgradient inequality (20) with $y' = 0$ and $p' = p^*$ yields

$$Q(0, p^*) \leq Q(y^k, p^k) - \langle x^k, 0 - y^k \rangle - \langle u^k, p^* - p^k \rangle,$$

which we may rearrange into

$$Q(y^k, p^k) \geq Q(0, p^*) - \langle x^k, y^k \rangle + \langle u^k, p^* - p^k \rangle. \tag{44}$$

Passing to the limit and using that $\langle x^k, y^k \rangle \rightarrow 0$, $u^k \rightarrow 0$, and $\{p^k\}$ is bounded, we obtain

$$\liminf_{k \rightarrow \infty} Q(y^k, p^k) \geq Q(0, p^*). \tag{45}$$

We now consider $\limsup_{k \rightarrow \infty} Q(y^k, p^k)$, which must by (45) be at least $Q(0, p^*)$; however, at present we have not excluded the possibility that it may be larger, or even $+\infty$. Let \mathcal{K} be a subsequence such that $Q(y^k, p^k) \rightarrow_{\mathcal{K}} \limsup_{k \rightarrow \infty} Q(y^k, p^k)$. By the boundedness of $\{p^k\}$, there exists a subsequence $\mathcal{K}' \subseteq \mathcal{K}$ such that $\{p^k\}_{k \in \mathcal{K}'}$ converges to some limit p^∞ . We then observe that

$$\begin{aligned} Q(0, p^*) &\geq Q(0, p^\infty) && \text{[Since } p^* \text{ is optimal for the dual]} \\ &= Q\left(\lim_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}'}} y^k, \lim_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}'}} p^k\right) && \text{[Since } y^k \rightarrow 0, p^k \rightarrow_{\mathcal{K}'} p^\infty\text{]} \\ &\geq \limsup_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}'}} Q(y^k, p^k) && \text{[Since } Q \text{ is upper semicontinuous]} \\ &= \limsup_{k \rightarrow \infty} Q(y^k, p^k). && \text{[By the choice of } \mathcal{K} \supseteq \mathcal{K}'\text{]} \end{aligned}$$

Combining this result with (45), we conclude that

$$\liminf_{k \rightarrow \infty} Q(y^k, p^k) \geq Q(0, p^*) \geq \limsup_{k \rightarrow \infty} Q(y^k, p^k) \quad \Rightarrow \quad \lim_{k \rightarrow \infty} Q(y^k, p^k) = Q(0, p^*),$$

which is the common optimal value of the primal and dual problems.

We now consider the sequence $\{F(x^k, u^k)\}$. Since F and $-Q$ are convex conjugates and $(x^k, u^k) \in \partial Q(y^k, p^k)$, the Fenchel equality — see for example [17, Theorem 23.5] — implies

$$F(x^k, u^k) = Q(y^k, p^k) + \langle y^k, x^k \rangle + \langle p^k, u^k \rangle.$$

Since we already know that $Q(y^k, p^k) \rightarrow Q(0, p^*)$, and we have that $\langle y^k, x^k \rangle \rightarrow 0$, $\{p^k\}$ is bounded, and $u^k \rightarrow 0$, it follows that $F(x^k, u^k) \rightarrow Q(0, p^*) = F(x^*, 0)$.

We now show, using the lower semicontinuity of F and that $u^k \rightarrow 0$, that all limit points of $\{x^k\}$ must be solutions to the primal problem (16). Specifically, considering any such limit point x^∞ and corresponding subsequence \mathcal{K} , we have

$$F(x^\infty, 0) = F\left(\lim_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}}} x^k, \lim_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}}} u^k\right) \leq \liminf_{\substack{k \rightarrow \infty \\ k \in \mathcal{K}}} F(x^k, u^k) = F(x^*, 0),$$

but since $F(x^*, 0)$ is the minimum possible value of $F(\cdot, 0)$, we must have $F(x^\infty, 0) = F(x^*, 0)$. Using the upper semicontinuity of Q and that $y^k \rightarrow 0$, analogous reasoning implies that all limit points of $\{p^k\}$ are dual solutions. The proof for the case that at least one KKT pair exists is now complete.

It remains only to consider the case that no KKT pair exists. In this case, we use a variant of the analysis originally given in [21] for the behavior of the proximal point algorithm for operators with no roots. The proof proceeds by contradiction: suppose that no KKT pair exists, but the conclusion of the proposition does not hold, so that $\{p^k\}$ and $\{x^k\}$ are both bounded. In this case, there exists some scalar $R \in (0, \infty)$ such that $\sup_{k \geq 1} \{\|(x^k, p^k)\|\} < R$. Let B denote the closed ball of radius R around the origin in $\mathbb{R}^n \times \mathbb{R}^m$, and consider the point-to-set operator $T = \partial L + N_B$, where N_B is the normal cone mapping of B . From the results of [18], T is maximal monotone, since T and ∂L are both maximal monotone and $\text{dom } \partial L$ and $\text{int dom } N_B$ have nonempty intersection. Furthermore, since $\text{dom } T$ lies within B and is thus a bounded set, it follows from [16, Proposition 2] that there exists at least one point $(x^*, p^*) \in \mathbb{R}^n \times \mathbb{R}^m$ for which $(0, 0) \in T(x^*, p^*)$, that is, (x^*, p^*) is a *root* of T .

Since the entire sequence $\{(x^k, p^k)\}$ lies in the interior of B , we have for all k that

$$N_B(x^k, p^k) = \{0\} \quad \Rightarrow \quad T(x^k, p^k) = \partial L(x^k, p^k) \quad \Rightarrow \quad (y^k, u^k) \in T(x^k, p^k).$$

Using the monotonicity of T , we may conclude that (40) still holds, but with (x^*, p^*) assumed to be any root of T . Continuing in this way, all the conclusions above for the case in which at least one KKT pair exists, through the conclusion that $\langle x^k, y^k \rangle \rightarrow 0$, continue to hold. From the boundedness of $\{(x^k, p^k)\}$, it must have at least one limit point (x^∞, p^∞) , whose norm must be less than R . Taking limits over an appropriate subsequence in the relation $(y^k, u^k) \in T(x^k, p^k)$, and using the maximality of T , we conclude that $(0, 0) \in T(x^\infty, p^\infty)$. But since

$\|(x^\infty, p^\infty)\| < R$, it lies in the interior of B and hence $\partial L(x^\infty, p^\infty) = T(x^\infty, p^\infty)$. Thus, $(0, 0) \in \partial L(x^\infty, p^\infty)$, and (x^∞, p^∞) must be a KKT pair, which contradicts the hypothesis. Therefore, the assumption above that one can simultaneously have no KKT pairs with both $\{p^k\}$ and $\{x^k\}$ bounded cannot hold. \square

The properties of $\{w^k\}$ are unusual. Although $\{(w^k, p^k)\}$ is Fejér monotone to the set of KKT pairs, and all limit points of $\{p^k\}$ are dual solutions, $\{w^k\}$ need not approach the set of primal solutions, and may behave very differently from $\{x^k\}$. Indeed, if we were able to solve the augmented Lagrangian subproblems exactly and achieve $y^k = 0$, then $\{w^k\}$ would simply be a constant sequence. One possible interpretation of the role of $\{w^k\}$ is that it accumulates, through (26), the total “error drift” of the algorithm. Then, for example, if a large amount of drift accumulates, in the sense that the magnitude of w^{k-1} becomes large relative to $\|x^k\|$, the subproblem optimality tolerance will be effectively tightened, because $\|y^k\|$ will have to be small in order to satisfy (25).

The conclusions of Proposition 1 are somewhat weaker than are typically obtained for multiplier methods, either in their exact form or using the absolute error criterion of [8]. In particular, in the case in which no solution exists, one typically obtains that the dual sequence $\{p^k\}$ is unbounded, but here we obtain only that either $\{p^k\}$ or $\{x^k\}$ is unbounded. The latter can in theory happen even if KKT pairs exist, but the set of primal solutions is unbounded; however, in practical implementations, such behavior of $\{x^k\}$ is generally not of concern. When KKT pairs exist, Proposition 1 is also weaker than results normally obtained for multiplier methods, in that full convergence of $\{p^k\}$ is not guaranteed. If the optimal solution of the dual problem is unique, then the results of Proposition 1 — that $\{p^k\}$ is bounded and all its limit points are solutions — are equivalent to convergence. If the optimal dual solution is nonunique, the results are somewhat weaker, but the differences seem unlikely to be of practical concern. We now show how the dual convergence results may be strengthened to full convergence by imposing a second approximation criterion in addition to (25); however, it is doubtful such a criterion would be needed in practice.

Proposition 2 *Suppose all the hypotheses of Proposition 1 hold, in the case that at least one KKT pair exists. If for some scalar $\zeta \geq 0$ it is also true for all $k \geq 1$ that*

$$c_k \|y^k\| \leq \zeta \|p^{k-1} - p^k\|^2, \tag{46}$$

then $\{p^k\}$ must converge to a unique limit, which is necessarily a dual solution.

Proof. By hypothesis, all the conclusions and intermediate results of Proposition 1 for the case that at least one KKT pair exists must hold. Again letting (x^*, p^*) denote an arbitrary KKT pair and rearranging (39), we obtain

$$\begin{aligned} 2c_k [\langle x^k - x^*, y^k \rangle + \langle p^k - p^*, u^k \rangle] \\ = \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2 - [\|p^k - p^*\|^2 + \|w^k - x^*\|^2] \\ - 2c_k \langle w^{k-1} - x^k, y^k \rangle + c_k^2 \|y^k\|^2 - \|p^{k-1} - p^k\|^2. \end{aligned}$$

Summing this equation for $k = 1, \dots, K$, we obtain

$$\begin{aligned} & 2 \sum_{k=1}^K (c_k \langle x^k - x^*, y^k \rangle + c_k \langle p^k - p^*, u^k \rangle) \\ &= \|p^0 - p^*\|^2 + \|w^0 - x^*\|^2 - \left[\|p^K - p^*\|^2 + \|w^K - x^*\|^2 \right] \\ &\quad - 2 \sum_{k=1}^K c_k \langle w^{k-1} - x^k, y^k \rangle + \sum_{k=1}^K c_k^2 \|y^k\|^2 - \sum_{k=1}^K \|p^{k-1} - p^k\|^2. \end{aligned}$$

Since $\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\}$ is a convergent sequence, and $\{c_k \langle w^{k-1} - x^k, y^k \rangle\}$, $\{c_k^2 \|y^k\|^2\}$, and $\{\|p^{k-1} - p^k\|^2\}$ are all summable, we conclude that the sequence

$$\{c_k \langle x^k - x^*, y^k \rangle + c_k \langle p^k - p^*, u^k \rangle\} \quad (47)$$

is summable. Next, we will use the additional hypothesis (46) to show that the first term $\{c_k \langle x^k - x^*, y^k \rangle\}$ above is summable, meaning that the second term $\{c_k \langle p^k - p^*, u^k \rangle\}$ must also be summable. To this end, we write

$$c_k \langle x^k - x^*, y^k \rangle = c_k \langle x^k - w^{k-1}, y^k \rangle + \langle w^{k-1} - x^*, c_k y^k \rangle.$$

Now, the first term on the right-hand side above, $c_k \langle x^k - w^{k-1}, y^k \rangle$ was already been shown to be summable in the proof of Proposition 1. As for the second, we note that from the extra condition (46) and the summability of $\{\|p^{k-1} - p^k\|^2\}$, we have that $\{\|c_k y^k\|\}$ (without the norm being squared) is summable. Now, since $\{(w^k, p^k)\}$ is bounded, we have that $\{w^{k-1} - x^*\}$ is bounded, and since $\{\|c_k y^k\|\}$ is summable, it follows that $\{\langle w^{k-1} - x^*, c_k y^k \rangle\}$ and therefore $\{c_k \langle x^k - x^*, y^k \rangle\}$ are summable. From the summability of (47), it follows that $\{c_k \langle p^k - p^*, u^k \rangle\}$ is also summable.

Next, we perform the expansion

$$\begin{aligned} \|p^{k-1} - p^*\|^2 &= \|p^k - p^* + (p^{k-1} - p^k)\|^2 \\ &= \|p^k - p^*\|^2 + 2\langle p^k - p^*, p^{k-1} - p^k \rangle + \|p^k - p^{k-1}\|^2 \\ &= \|p^k - p^*\|^2 + 2c_k \langle p^k - p^*, u^k \rangle + \|p^k - p^{k-1}\|^2. \end{aligned}$$

where the last equality follows from the definition of u^k . Rearranging the resulting equation, we have

$$\|p^k - p^*\|^2 - \|p^{k-1} - p^*\|^2 = -2c_k \langle p^k - p^*, u^k \rangle - \|p^k - p^{k-1}\|^2.$$

Because the two terms on the right-hand side above are summable, we may therefore conclude that $\{\|p^k - p^*\|^2\}$ and consequently $\{\|p^k - p^*\|\}$ must converge to a finite limit.

Next, consider any limit point p^∞ of the bounded sequence $\{p^k\}$, which we know from Proposition 1 must be a dual solution. From [17, Corollary 30.5.1], we know that (x^*, p^∞) is also a KKT pair, and therefore we may set $p^* = p^\infty$ to conclude that $\{\|p^k - p^\infty\|\}$ converges. But since p^∞ is a limit point of $\{p^k\}$, $\{\|p^k - p^\infty\|\}$ must have a subsequence converging to 0, and thus the entire sequence converges to 0 and we must have $p^k \rightarrow p^\infty$. \square

Although the constant ζ may be arbitrarily large, the additional approximation criterion (46) is potentially stringent in the limit, since the norm on its right-hand side is squared, but the norm on its left is not. Again, it seems doubtful that this extra criterion would be needed in practice.

5 An application with inequality constraints, equality constraints, and variable bounds

We now give a somewhat more complicated application of (24)-(26) than was discussed in sections 1 and 3, and more reflective of the kind of formulations that would be encountered by a practical nonlinear optimization solver. Although in theory it is no more general than (2), we will consider a problem of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{ST} \quad & g(x) \leq 0 \\ & h(x) = 0 \\ & a \leq x \leq b, \end{aligned} \tag{48}$$

where

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and continuously differentiable, as in Section 1
- $g : \mathbb{R}^n \rightarrow \mathbb{R}^{m_1}$ is continuously differentiable and componentwise convex (similar to Section 1, but with range dimension m_1 instead of m)
- $h : \mathbb{R}^n \rightarrow \mathbb{R}^{m_2}$ is affine (as in Section 1, except with range dimension m_2 instead of m)
- $a \in [-\infty, \infty]^n$, $b \in (-\infty, \infty]^n$, and $a \leq b$.

Due to their simple structure, we will enforce the bound constraints $a \leq x \leq b$ in the subproblems, and will not attempt to attach Lagrange multipliers to them. Let

$$B(a, b) = \{x \in \mathbb{R}^n \mid a \leq x \leq b\}$$

denote the “box” set defined by these constraints, and let $N_{B(a,b)}$ denote its normal cone map. In particular, $N_{B(a,b)}(x) = \emptyset$ if $x \notin B(a, b)$, and otherwise, for $z \in \mathbb{R}^n$,

$$z \in N_{B(a,b)}(x) \quad \Leftrightarrow \quad \begin{cases} z_i \leq 0 & \forall i : x_i < b_i \\ z_i \geq 0 & \forall i : x_i > a_i. \end{cases}$$

Writing the second argument to F (the perturbation vector) as (u, v) , where $u \in \mathbb{R}^{m_1}$ and $v \in \mathbb{R}^{m_2}$, we choose F as follows:

$$F(x, (u, v)) = \begin{cases} f(x), & \text{if } a \leq x \leq b \text{ and } g(x) + u \leq 0 \text{ and } h(x) + v = 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

The corresponding Lagrangian takes the form

$$L(x, (p, q)) = \begin{cases} f(x) + \langle p, g(x) \rangle + \langle q, h(x) \rangle, & \text{if } a \leq x \leq b \text{ and } p \geq 0, \\ -\infty, & \text{if } a \leq x \leq b \text{ and } p \not\geq 0 \\ +\infty, & \text{if } a \not\leq x \text{ or } x \not\leq b, \end{cases}$$

where the multiplier vector is written as $(p, q) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$; note that the form of L effectively enforces the dual constraint $p \geq 0$. The corresponding extended dual function is $Q(y, (p, q)) = \inf_{x \in \mathbb{R}^n} \{L(x, (p, q)) - \langle x, y \rangle\}$. Applying the proximal point algorithm to the map $\partial Q(0, (\cdot, \cdot))$ produces the following augmented Lagrangian method for problem (48):

$$x^k \in \text{Arg min}_{x \in B(a,b)} \left\{ f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max\{0, p^{k-1} + c_k g_i(x)\}^2 + \langle q^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 \right\} \quad (49)$$

$$p^k = \max\{0, p^{k-1} + c_k g(x^k)\} \quad (50)$$

$$q^k = q^{k-1} + c_k h(x^k). \quad (51)$$

Note that the augmented Lagrangian is minimized over the explicit box constraint set $B(a, b)$ given by the constraints $a \leq x \leq b$.

Through manipulations resembling those of Section 3, but somewhat more complicated, the recursions (24)-(26) can be shown equivalent, when applied to this form of L , to the following approximate version of (49)-(51):

$$\text{Define } L_k(x) = f(x) + \frac{1}{2c_k} \sum_{i=1}^m \max\{0, p^{k-1} + c_k g_i(x)\}^2 + \langle q^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 \quad (52)$$

$$y^k \in \nabla L_k(x^k) + N_{B(a,b)}(x^k) \quad (53)$$

$$\frac{2}{c_k} \|w^{k-1} - x^k\| \|y^k\| + \|y^k\|^2 \leq \sigma \left(\min\left\{ \frac{1}{c_k} p^{k-1}, -g(x^k) \right\}^2 + \|h(x^k)\|^2 \right) \quad (54)$$

$$p^k = \max\{0, p^{k-1} + c_k g(x^k)\} \quad (55)$$

$$q^k = q^{k-1} + c_k h(x^k) \quad (56)$$

$$w^k = w^{k-1} - c_k y^k. \quad (57)$$

The approximation condition (53) is equivalent to y^k being a subgradient of the function $L_k + \delta_{B(a,b)}$, where $\delta_{B(a,b)}(x) = 0$ for $x \in B(a, b)$, and $\delta_{B(a,b)}(x) = +\infty$ if $x \notin B(a, b)$; note that $L_k + \delta_{B(a,b)}$ is effectively the function being minimized in (49). To implement (52)-(57) computationally, one would apply some iterative bound-constrained solver to the problem (49), but truncate its calculations as soon as it finds a vector x^k such that there exists a $y^k \in \nabla L_k(x^k) + N_{B(a,b)}(x^k)$ satisfying (54). For a given trial value of x^k , the possible corresponding choices of y^k will be nonunique if any component of the constraints $a \leq x^k$ or $x^k \leq b$ are binding. To maximize the chance of satisfying (54) with as few iterations as possible of the bound-constrained subproblem solver, one should choose y^k to have the minimum possible norm among all vectors in the set $\nabla L_k(x^k) + N_{B(a,b)}(x^k)$; a similar strategy is

used in the computational tests of [9]. To compute the vector y with the minimum possible norm in the set $t + N_{B(a,b)}$, for any $t \in \mathbb{R}^n$, one may use the following simple calculation:

$$y_i = \begin{cases} \min\{t_i, 0\}, & \forall i : x_i = a_i \\ t_i, & \forall i : a_i < x_i < b_i \\ \max\{t_i, 0\}, & \forall i : x_i = b_i. \end{cases} \quad (58)$$

We close this section by observing that Proposition 1 states that, whenever there exists at least one KKT pair (x^*, p^*) , we have that $u^k \rightarrow 0$ and $F(x^k, u^k) \rightarrow F(x^*, 0)$, where $\{u^k\}$ is defined by (35). For the current choice of F , this means that $\limsup_{k \rightarrow \infty} g_i(x) \leq 0$ for $i = 1, \dots, m_1$, $h(x^k) \rightarrow 0$, and $f(x^k) \rightarrow f(x^*)$. Such behavior of the primal sequence $\{x^k\}$ is often referred to as *asymptotic optimality*; see for example [8]. In particular, all accumulation points of $\{x^k\}$ are primal solutions.

6 Computational testing

We now describe some preliminary computational testing of the algorithm (53)-(57), using a subset of problems from the CUTE test set [4]. For our tests, we did not require f or the component functions of g to be convex, nor did we require h to be affine; we merely assumed $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^{m_1}$, and $h : \mathbb{R}^n \rightarrow \mathbb{R}^{m_2}$ to be once continuously differentiable. Our current convergence theory does not cover such potentially nonconvex problems, and analyzing our algorithm’s behavior in the nonconvex case is a topic for future research. Nevertheless, to assess our approach’s computational promise, it seemed best to test it on a standard, realistic, demanding test set, even if the majority of its problems are nonconvex.

We based our testing closely on our recent work in [9]. As in [9], we implemented our main algorithm (the “outer loop”) in Python [26], using SciPy [14], an open-source environment with capabilities similar to MATLAB; in fact, the implementation is a minor enhancement to the existing `pyauglag` prototype code already developed in [9].

The “inner loop” of the implementation consists of the procedure necessary to identify some pair (x^k, y^k) jointly satisfying (53) and (54). To this end, much as in [9], we used a slightly modified version of the ASA bound-constrained conjugate gradient code of Hager and Zhang [13], which is based on the advanced unconstrained conjugate algorithm described in [12]. Our only significant modification to the base ASA code [11] is the ability to use a user-specified termination criterion. Specifically, we ran ASA on the bound-constrained problem (49), starting from the previous primal iterate x^{k-1} , and checking the trial solution x produced at each ASA iteration as follows:

1. Calculate the minimum-norm member y of the set $\nabla L_k(x) + N_{B(a,b)}(x)$ by using (58) with $t = \nabla L_k(x)$.
2. Determine whether

$$\frac{2}{c_k} \|w^{k-1} - x\| \|y\| + \|y\|^2 \leq \sigma \left(\min\left\{ \frac{1}{c_k} p^{k-1}, -g(x) \right\}^2 + \|h(x)\|^2 \right) \quad (59)$$

or

$$\|y\|_\infty \leq \delta, \quad (60)$$

where δ is a fixed constant that is small enough to assert that the subproblem was solved “exactly”. This parameter depends on the termination criterion for the outer loop and will be defined below.

If either (59) or (60) holds, set $x^k = x$ and $y^k = y$ and exit the ASA subroutine; if not, continue to the next ASA iteration.

The remainder of the algorithm comprises the updates of the multiplier estimates and w^k , which consist of simple vector calculations implemented in SciPy.

In our computational tests, we used a subset of 127 of the AMPL versions of the CUTE [4] test problems made available by Hande Benson at <http://orfe.princeton.edu/~rvdb/ampl/nlmodels/cute/>. We used exactly the same subset of these problems as in [9], to facilitate direct comparison with our recent computational work there. The exact selection of problems is described in [9]; in brief, some of the larger problems were excluded due to prototype nature of our implementation, which contains extensive interpreted Python code.

We compared our new approach to an implementation based on the summable error criterion from [8, formula (17)]. Specifically, we tested an approximation criterion of the form

$$\|y^k\| \leq \frac{\epsilon_k}{c_k \gamma_k}, \quad \text{with} \quad \gamma_k = \begin{cases} 1, & \|x^k\| \leq \beta \\ \|x^k\|/\beta, & \|x^k\| > \beta. \end{cases} \quad (61)$$

where ϵ_k is some summable sequence and $\beta > 0$ is a given constant; in our experiments we used $\beta = 10^4 \sqrt{n}$. Using that the penalty parameter sequence $\{c_k\} \subset (0, \infty)$ is assumed to be bounded away from 0, it can easily be shown that (61) is a special case of the criterion specified in [8]. As for the choice of the summable sequence $\{\epsilon_k\}$, we experimented with various sequences of the form

$$\epsilon_k = \eta/k^\zeta, \quad (62)$$

where η is a positive constant and $\zeta > 1$. After some numerical experimentation, we settled on the values $\eta = 0.1$ and $\zeta = 2$, which seemed to perform the best on our experimental test set.

We also compared the new approach with the heuristic criterion suggested in [9], which used a form of the relative error criterion for which there is (at least at present) no global convergence proof even in the convex case. In our current notation, this heuristic uses the error criterion

$$\|y^k\|^2 \leq \sigma \left(\left\| \min \left\{ \frac{1}{c_k} p^{k-1}, -g(x^k) \right\} \right\|^2 + \|h(x^k)\|^2 \right), \quad (63)$$

that is, (54) with the term involving w^{k-1} deleted. In order to ensure theoretical convex-case convergence, the use of (63) in [9] included supplementary safeguards based on the summable criterion (61). However, at least for our current set of test problems, these safeguards are not required in practice, and only slow down the method. So, in order to compare (59) to the best-performing heuristic criterion available, we simply used (63) without any safeguards.

Both criteria (54) and (63) require us to select the parameter $\sigma \in [0, 1)$. In both cases, based on some numerical experimentation, we used the following “adaptive” method to control σ : at the outset, we set $\sigma = 0.99$; however, if at iteration k the starting point $x = x^{k-1}$ for the ASA algorithm already satisfies the error criterion, we decrease σ by setting $\sigma \leftarrow \sigma/10$. Conversely, if the ASA inner loop fails to find a solution of the subproblem within the required precision, we set $\sigma \leftarrow \min\{0.99, 10\sigma\}$. Note that since $\sigma \leq 0.99$ at all iterations, our procedure fulfills the assumptions of the convergence proof with $\sigma = 0.99$.

To terminate the outer loop, we use the same rules as in [9]. We define

$$\gamma(x, p, q) = \min \{ \|y\|_\infty \mid y \in \nabla_x L(x, (p, q)) + N_{B(a,b)}(x) \} \quad (64)$$

$$\phi(x) = \max \{ \|h(x)\|_\infty, \|\max\{0, g(x)\}\|_\infty \} \quad (65)$$

$$\kappa_\epsilon(x, p) = \max \{ |p_i| \mid i : g_i(x) \leq -\epsilon \}. \quad (66)$$

Here, $\gamma(x, p, q)$ measures how close x comes to being the minimizer over $B(a, b)$ of $L(\cdot, (p, q))$, while $\phi(x)$ measures how close x is to being feasible, and $\kappa_\epsilon(x, p)$ measures the violation of complementary slackness for the inequality constraints $g(x) \leq 0$. If $\gamma(x, p, q) = 0$, $\phi(x) = 0$, and $\kappa_0(x, p) = 0$, then (x, p, q) satisfies the KKT conditions for (48). Note that under the recursions of our proposed algorithm, we have $\gamma(x^k, p^k, q^k) = \|y^k\|_\infty$.

Given a parameter $\epsilon > 0$, we terminate, declaring success, whenever

$$\gamma(x^k, p^k, q^k) = \|y^k\| < \epsilon \quad \phi(x^k) < \epsilon \quad \kappa_\epsilon(x^k, p^k) < \epsilon. \quad (67)$$

For direct comparison with the results presented in [9], we set $\epsilon = 10^{-4}$; in future, we plan to use tighter tolerances, but scale them in some relation to the problem data, or apply coordinatewise scaling in the definitions of the convergence metrics (64)-(66). We defined the constant δ appearing in (60) to be $\epsilon/10$.

We considered a method to have failed if any of the following occur:

- We still have not satisfied the approximate KKT conditions (67) after 200 outer iterations ($k \geq 200$)
- The ASA subproblem solver declares failure 5 or more times
- There are more than 1 million function evaluations
- The total CPU time exceeds one hour (on a single core of a 2.83GHz Intel Core 2 Quad Q9550 processor with 800 MHz memory).

For all algorithms, we use the same strategy to adjust the penalty parameter c_k , based on the technique used in the Algencan augmented Lagrangian solver of Andreani *et al.* [1, 2]. At the end of iteration k , we test whether

$$\phi(x^k) < \epsilon \quad \text{or} \quad \phi(x^k) \leq 0.5 \phi(x^{k-1}) \quad (68)$$

$$\kappa_\epsilon(x^k, y^k) < \epsilon \quad \text{or} \quad \kappa_\epsilon(x^k, p^k, q^k) \leq 0.5 \kappa_\epsilon(x^{k-1}, p^{k-1}, q^{k-1}). \quad (69)$$

If both (68) and (69) hold, we consider our method to be making “good progress” towards feasibility and complementarity, and keep the penalty parameter c_k unchanged. Otherwise, the penalty parameter increases by a factor of 5; note that Algencan, which uses a Newton subproblem solver instead of the conjugate gradient approach employed here, uses a larger increase factor of 10. We set the initial penalty parameter c_0 to 5.

In the new error criterion, the choice of the initial reference point w^0 is arbitrary, but can have great bearing on the strictness of the error criterion. If at some point the current trial solution x of the ASA algorithm is far from w^{k-1} , then $\|w^{k-1} - x\|$ will be large in (59), making it much stricter than the heuristic criterion (63), and thus requiring a much smaller subgradient y . A small value of y^k , once an acceptable pair (x^k, y^k) has been identified, means that the update $w^k = w^{k-1} - c_k y^k$ will leave w^k close to w^{k-1} , and the error criterion for the next iteration will be similarly strict if x remains in the same region. If this phenomenon occurs for values of the multiplier estimates p^{k-1} and q^{k-1} for which the inner loop is having trouble solving the subproblem (49) accurately, it has the potential to “jam” the progress of the overall algorithm. Initially, we observed this pattern occurring for a few of the test problems, causing a minor loss of robustness in comparison to the heuristic method of [9]. To ameliorate such behavior, we make a “smart” initial choice of w^k , and allow a finite number of “resets” to the $\{w^k\}$ sequence. Specifically, for the first three iterations, we ignore the w^{k-1} term in the error criterion (54), effectively reducing it to (63). Then, we initialize $w^4 = x^3$, the idea being that henceforth w^k is likely to be roughly equal to x^k , and the criterion (54) will not be overly stringent. However, we have occasionally observed (especially for nonconvex problems) that x^k can shift substantially later in the algorithm, again raising the possibility of $\|w^{k-1} - x\|$ becoming large. Thus, at each trial point computed by the ASA solver, we check whether

$$\|w^k - x^k\| > 100c_k \|y^k\|,$$

that is, whether the new error criterion (54) is more than two orders of magnitude stricter than the heuristic criterion (63). If so, we reset $w^k \leftarrow x^k$. However, we allow at most 5 resets of this kind, so that in the limit we are using algorithm (52)-(57) and our convergence theory applies, at least in the convex case.

Figure 1 shows two performance profiles [7]. The left-hand profile compares the summable criterion (61), using the best-performing parameters we could identify, with a classical augmented Lagrangian approach in which all subproblems are solved essentially exactly, that is, using only condition (60). We measure performance by counting the number of gradient evaluations; results for the number of function evaluations are similar. It is clear that the summable inexact variation is a clear improvement, requiring less computational effort without any compromise in robustness. The right-hand profile in Figure 1 compares the summable error with the new, relative error criterion introduced in this paper. Once again, we see a clear improvement, with no robustness sacrifice. We interpret this improvement as resulting from the relative criterion’s ability to sense the rate of approximation tightening appropriate to each problem instance, giving it an advantage over criteria like (61) which employ a predetermined “one-size-fits-all” summable sequence $\{\epsilon_k\}$, no matter how carefully it is chosen.

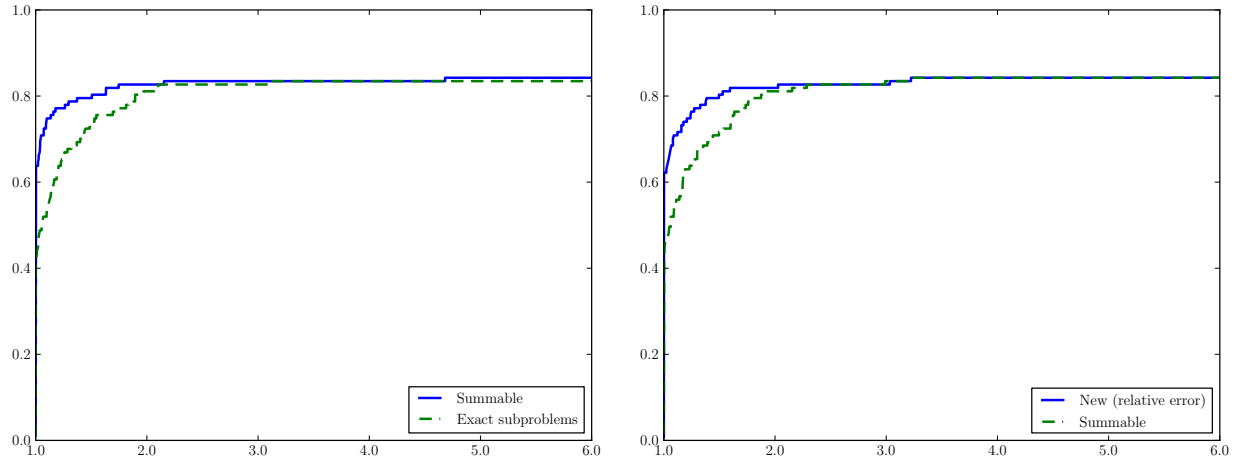


Figure 1: Performance profiles comparing the criterion (61), based on a summable sequence $\{\epsilon_k\}$ (“Summable”) with computing high-precision solutions to all subproblems (“Exact”) and the new error criterion (54) (“New (relative error)”), in terms of gradient evaluations.

Figure 2 displays a performance profile comparing the new relative error criterion with the heuristic relative error criterion (63), which was the best-performing approach in [9]. Here, we see that the extra term involving w^k in the left-hand side of (54) does not seem to have significant practical impact, although it does seem to slow down convergence very slightly. On the other hand, it appears to have a small (but possibly statistically insignificant) benefit in terms of robustness. The main difference between the two methods is that the new criterion has a convergence proof for the convex case, while the heuristic does not. Thus, our new method works about as well as the best heuristic method we are aware of, but has the advantage of a global convergence proof.

That the heuristic criterion (63), without the sequence $\{w^k\}$, seems to work about as well in practice as our proposed method suggests that it may be worthwhile trying to prove its convergence. Such a proof appears difficult, and we do not know if one is possible, but we plan to further investigate this topic in the future.

7 Concluding discussion

In conclusion, we have developed a promising new error criterion for approximate minimization of augmented Lagrangian subproblems. It does not require a primal regularization term as in (7) and (9), and yet requires only readily available information, namely the gradient (or a subgradient) of the augmented Lagrangian. Furthermore, it is a relative error criterion, in that it uses a quantity proportional the current violation of the KKT conditions — as for example in the right-hand side of (54) — to guide the degree of precision needed in each subproblem. Thus, it is an advance over the results of [8], which require a summable sequence of error parameters $\{\epsilon_k\}$, and provide no direct guidance how to select it. Furthermore, the performance of our new, provably convergent method is nearly identical to the best heuristic

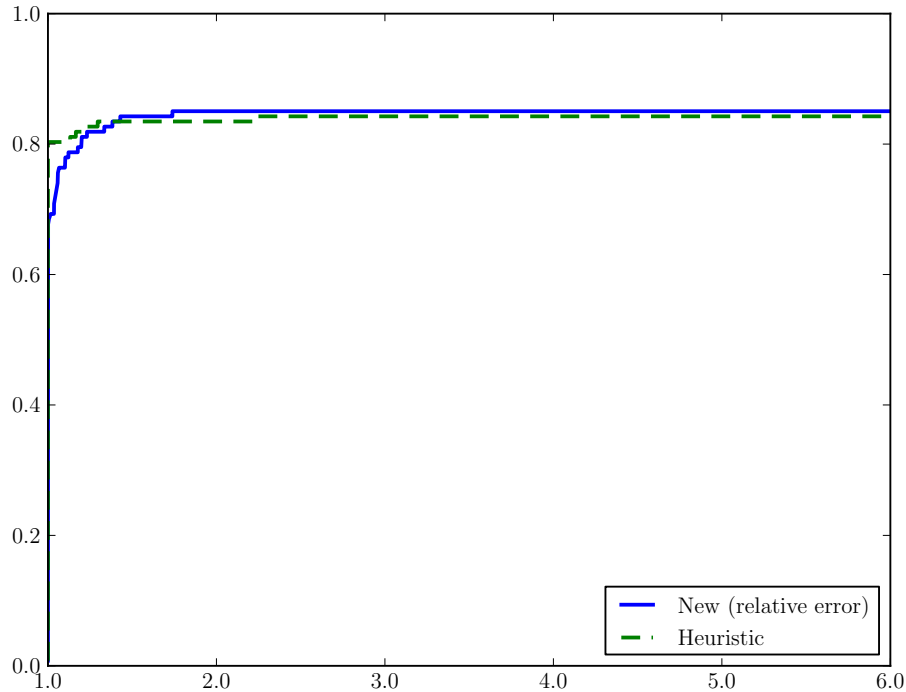


Figure 2: Performance profile comparing the proposed algorithm (“New (relative error)”) with the best algorithm in [9] (“Heuristic”), in terms of the number of gradient evaluations.

approach we have experimented with so far.

The new error criterion involves an unusual auxiliary sequence $\{w^k\}$; referring to the proof of Proposition 1, we now make some additional comments about the role of this sequence, and why it appears to be necessary to obtain a global convergence result in the convex case. To have a practical condition requiring only knowledge of augmented Lagrangian gradient, as opposed to more problematic conditions involving ϵ -optimality or ϵ -subgradients, it appears that any convex-case global convergence proof must be based on the monotonicity of the Lagrangian subgradient operator $\partial L(\cdot, \cdot)$ and upper semicontinuity of the extended dual function $Q(\cdot, \cdot)$, rather than working with the lower-dimensional dual functional $Q(0, \cdot)$ and its subgradient map, as is traditionally the case for augmented Lagrangian methods. The earlier analysis in [8] is based on this same idea, but establishes simple Fejér monotonicity of $\{p^k\}$ to the dual solution set and does not require an auxiliary sequence like $\{w^k\}$. However, we have as yet been unsuccessful in directly modifying the analysis of [8] into one involving a relative error criterion, because the resulting criteria always seem to require knowledge of the KKT pair (x^*, p^*) , which (even though it might be assumed to exist) is necessarily unknown if one is trying to solve the corresponding optimization problem. The approach presented here manages to sidestep this difficulty by instead establishing Fejér monotonicity of (w^k, p^k) to the set of KKT pairs. In some sense, one can view the method as a proximal algorithm in the dual variables p — or (p, q) for the problem (48) — and a kind of extragradient algorithm [15] in the primal variables: the extragradient-like step (26) provides the necessary

Fejér monotonicity in the primal variables. However, that $\{w^k\}$ is distinct from the ordinary primal iterates and need not, and in general does not, approach the set of primal solutions, is a curious new feature of the algorithm.

In our planned continued work on this topic, it is of obvious interest is to analyze the behavior of the method for nonconvex problems (to the extent possible), and further improve its practical dependability. However, the present results are promising enough that we plan to embark on more sophisticated, fully compiled implementations aimed at larger-scale problems and parallel computing architectures. If the basic workings of the ASA algorithm, or something similar, can be implemented in parallel, then the entire algorithm should be quite straightforwardly parallelizable, due to the simple form of the updates to the multipliers and w^k . Because of the separable structure of the box constraints and the relatively simple linear algebra required by the underlying conjugate gradient method of [12], parallelization of ASA seems likely to be feasible. We will also continue trying to prove the convergence of the simpler and more intuitive relative error criterion (63), which has similar practical performance to the method proposed here, but has so far resisted analysis.

References

- [1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM J. Optim.*, 18(4):1286–1309, 2007.
- [2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. *Math. Program.*, 111(1-2):5–32, 2008.
- [3] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, New York, 1982.
- [4] I. Bongartz, A. R. Conn, N. Gould, and P. L. Toint. CUTE: constrained and unconstrained testing environment. *ACM Trans. Math. Softw.*, 21(1):123–160, 1995.
- [5] A. R. Conn, N. Gould, A. Sartenaer, and P. L. Toint. Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints. *SIAM J. Optim.*, 6(3):674–703, 1996.
- [6] A. R. Conn, N. I. M. Gould, and P. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.*, 28(2):545–572, 1991.
- [7] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2):201–213, 2002.

- [8] J. Eckstein. A practical general approximation criterion for methods of multipliers based on Bregman distances. *Math. Program.*, 96(1):61–86, 2003.
- [9] J. Eckstein and P. J. S. Silva. Proximal methods for nonlinear programming: double regularization and inexact subproblems. *Comput. Optim. Applic.*, 46(2):279–304, 2010.
- [10] M. P. Friedlander and M. A. Saunders. A globally convergent linearly constrained Lagrangian method for nonlinear optimization. *SIAM J. Optim.*, 15(3):863–897, 2005.
- [11] W. W. Hager and H. Zhang. ASA-CG source code. <http://www.math.ufl.edu/~hager/papers/CG/>.
- [12] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optim.*, 16(1):170–192, 2005.
- [13] W. W. Hager and H. Zhang. A new active set algorithm for box constrained optimization. *SIAM J. Optim.*, 17(2):526–557, 2006.
- [14] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001. <http://www.scipy.org/>.
- [15] G. M. Korpelevich. Extrapolation gradient methods and their relation to modified Lagrange functions. *Èkonom. i Mat. Metody*, 19(4):694–703, 1983.
- [16] R. T. Rockafellar. Local boundedness of nonlinear, monotone operators. *Michigan Math. J.*, 16:397–407, 1969.
- [17] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
- [18] R. T. Rockafellar. On the maximality of sums of nonlinear monotone operators. *Trans. Amer. Math. Soc.*, 149:75–88, 1970.
- [19] R. T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, Philadelphia, 1974.
- [20] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.
- [21] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.
- [22] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal.*, 7(4):323–345, 1999.
- [23] M. V. Solodov and B. F. Svaiter. A hybrid projection-proximal point algorithm. *J. Convex Anal.*, 6(1):59–70, 1999.

- [24] M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Math. Oper. Res.*, 25(2):214–230, 2000.
- [25] J. E. Spingarn. Partial inverse of a monotone operator. *Appl. Math. Optim.*, 10(3):247–265, 1983.
- [26] G. van Rossum et al. Python language website. <http://www.python.org/>.