Math. Program., Ser. A (2013) 141:319–348
DOI 10.1007/s10107-012-0528-9

FULL LENGTH PAPER

A practical relative error criterion for augmented Lagrangians

Jonathan Eckstein · Paulo J. S. Silva

Received: 17 August 2010 / Accepted: 27 February 2012 / Published online: 24 March 2012

Abstract This paper develops a new error criterion for the approximate minimization of augmented Lagrangian subproblems. This criterion is practical since it is readily testable given only a gradient (or subgradient) of the augmented Lagrangian. It is also "relative" in the sense of relative error criteria for proximal point algorithms: in particular, it uses a single relative tolerance parameter, rather than a summable parameter sequence. Our analysis first describes an abstract version of the criterion within Rockafellar's general parametric convex duality framework, and proves a global convergence result for the resulting algorithm. Specializing this algorithm to a standard formulation of convex programming produces a version of the classical augmented Lagrangian method with a novel inexact solution condition for the subproblems. Finally, we present computational results drawn from the CUTE test set—including many nonconvex problems—indicating that the approach works well in practice.

Mathematics Subject Classification 90C25 · 90C30

This material is based in part upon work supported by the National Science Foundation under Grant CCF-1115638. Jonathan Eckstein was also 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.

J. Eckstein (✉)
Department of Management Science and Information Systems and RUTCOR,
Rutgers University, 640 Bartholomew Road, Busch Campus, Piscataway, NJ 08854, USA
e-mail: jeckstei@rci.rutgers.edu

P. J. S. Silva
Department of Computer Science, University of São Paulo, Rua do Matão,
1010, São Paulo, SP, CEP 05508-090, Brazil
e-mail: pjssilva@ime.usp.br

1 Introduction, motivation, and summary

This paper presents a new rule for approximately solving the subproblems of augmented Lagrangian methods. Consider the following convex optimization problem:

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & h(x) = 0 \\ & g(x) \le 0, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $h : \mathbb{R}^n \to \mathbb{R}^{m_1}$ is affine, and $g(x) = \big(g_1(x), \ldots, g_{m_2}(x)\big)$, where $g_1, \ldots, g_{m_2} : \mathbb{R}^n \to \mathbb{R}$ are convex. While it is not essential to our analysis, we assume for simplicity of presentation that f and g are differentiable. Augmented Lagrangian methods for such convex optimization problems are intimately connected with the *proximal point algorithm* (PPA) for set-valued monotone operators [25,26]: here, given a maximal monotone operator $T : \mathbb{R}^\ell \rightrightarrows \mathbb{R}^\ell$, the canonical problem is to locate a *root* of T, that is, a point z such that $0 \in T(z)$. The PPA refers to computing a sequence $\{z^k\}$ obeying the recursion

$$z^k = (I + c_k T)^{-1}(z^{k-1}), \quad \text{or equivalently} \quad z^k + c_k T(z^k) \ni z^{k-1}, \tag{2}$$

where $\{c_k\}$ is a sequence of scalars with $\inf_k\{c_k\} > 0$. Any sequence $\{z^k\}$ conforming to (2) converges to a root of T, if one exists. Applying the PPA to an operator derived from the dual of the problem (1), one obtains as described in [25] the *method of multipliers*

$$x^k \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^n} \left\{ f(x) + \langle \lambda^{k-1}, h(x) \rangle + \frac{c_k}{2}\|h(x)\|^2 + \frac{1}{2c_k}\big\| \max\big\{0, \mu^{k-1} + c_k g(x)\big\}\big\|^2 \right\} \tag{3}$$

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

$$\mu^k = \max\big\{0, \mu^{k-1} + c_k g(x^k)\big\}, \tag{5}$$

where $\{c_k\}$ is as above and the "max" operations are interpreted componentwise. In the augmented Lagrangian minimization step (3), $\lambda^{k-1} \in \mathbb{R}^{m_1}$ and $\mu^{k-1} \in \mathbb{R}^{m_2}_+$ are the previous iteration's estimates of the Lagrange multipliers for the equality and inequality constraints in (1), respectively.

If one applies the proximal point algorithm to a monotone operator derived from a primal-dual formulation of (1), one instead obtains—again, see [25]—the *proximal* method of multipliers, in which the subproblem minimand contains an additional augmenting term:

$$
\begin{aligned}
x^k &= \arg\min_{x\in\mathbb{R}^n}\Big\{ f(x) + \langle\lambda^{k-1}, h(x)\rangle + \frac{c_k}{2}\|h(x)\|^2 \\
&\qquad + \frac{1}{2c_k}\left\| \max\{0, \mu^{k-1} + c_k g(x)\}\right\|^2 + \frac{1}{2c_k}\|x - x^{k-1}\|^2\Big\} &(6)\\
\lambda^k &= \lambda^{k-1} + c_k h(x^k) &(7)\\
\mu^k &= \max\{0, \mu^{k-1} + c_k g(x^k)\}. &(8)
\end{aligned}
$$

In practice, one would prefer not to solve all the subproblems of the form (3) or (6) to high precision, but to instead solve them inexactly using tolerances that tighten as the algorithm proceeds. The idea is to avoid expending excessive effort on computing exact minimizers x^k of the augmented Lagrangian in early iterations when the estimates $(\lambda^{k-1}, \mu^{k-1})$ of the Lagrange multipliers may be poor. To prove convergence of methods of this kind, it is natural to appeal to the theory of approximation criteria for the proximal point algorithm, that is, for approximating the recursion (2) while still maintaining convergence of $\{z^k\}$ to a root of T. The oldest approximation conditions for (2), which we call *absolute summable* error criteria, involve a theoretically infinite sequence of error tolerance parameters $\{\epsilon_k\} \subset [0, \infty)$, but provide no direct guidance as to how to select it, except for requiring that $\sum_{k=1}^{\infty} \epsilon_k < \infty$. For instance, one of the original approximation criteria proposed in [26] is

$$\left\| z^{k+1} - (I + c_k T)^{-1}(z^k)\right\| \le \epsilon_k \quad \forall k, \quad \text{where} \quad \sum_{k=1}^{\infty} \epsilon_k < \infty. \tag{9}$$

The past 10–15 years, starting with [27–29], have seen the development of a new family of *relative* error criteria for approximating (2); such criteria have only a single scalar parameter and are based on the ratio of the error in evaluating the proximal operator $(I + c_k T)^{-1}$ to some other quantity maintained by the algorithm, such as the tentative step length. For example, the method of [27] involves finding $\tilde{z}^k$ and $v^k \in T(\tilde{z}^k)$ such that

$$\|\tilde{z}^k + c_k v^k - z^k\| \le \sigma \|\tilde{z}^k - z^k\|, \tag{10}$$

where $\sigma \in [0, 1)$ is a scalar parameter; note that if $\sigma = 0$, then $\tilde{z}^k$ is simply the next exact iterate specified by (2). In [27], the next iterate is not $\tilde{z}^k$, but $z^{k+1} = z^k - c_k v^k$ (although these two vectors are identical if $\sigma = 0$); [28] gives a different, projective "corrector" formula. In general, the advantage of such 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 do not require selection of an infinite sequence of parameters.

Approximation criteria like (9) or (10) for the generic proximal point recursion (2) tend to translate straightforwardly into implementable criteria for approximately performing the minimization (6) in the *proximal* method of multipliers, but regrettably this is not the case for the far more commonly used augmented Lagrangian minimization (3), for which generic proximal-point approximation criteria reduce to conditions that are not readily testable. This situation is unfortunate, since methods such as (3)–(5)

tend to be faster than their regularized cousins like (6)–(8) and are much more often used in practice; for a recent example of empirical results to this effect, see [13].

This paper develops new relative error criteria for approximating "pure dual" augmented Lagrangian calculations like (3). Besides being relative rather than absolute summable, the other critical feature of the criteria proposed here is that they are *practical*, that is, they are readily testable given knowledge of the gradient, or a subgradient, of the augmented Lagrangian. Direct application of approximation analyses such as those in [27–29] or even [26] to dual formulations of problems like (1) does not yield algorithms that are practical in this sense. Other approaches to approximating (3) while retaining a global proof of convergence in the convex case, for example techniques based on ϵ-subgradients, have similar difficulties except in special cases. The one exception we are aware of is the analysis in [12]: our analytical approach to deriving practical approximation criteria draws on [12], and makes similar use of Rockafellar's general parametric convex duality framework; see [22, Chaps. 29–30] and [24]. The analysis in [12] 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, [12] developed only an absolute summable error criterion with an infinite sequence $\{\epsilon_k\}$ of error parameters, and, as in [26], provides no direct guidance as to how to select this sequence. By contrast, the analysis here develops true relative error criteria with only one or two scalar parameters.

Some of our recent experiments in [13] used a heuristic approximation rule for dual methods like (3)–(5), obtained by taking a provably convergent rule for a proximal, primal-dual method of multipliers of the form (6)–(8), and simply deleting the terms relating to primal regularization. Our computational results for this criterion, given as (59) and discussed in Sect. 6 below, were very promising, but a corresponding formal convergence proof is not known, even in the convex case.

Here, we derive an outwardly similar but somewhat more complicated approximation criterion and prove that it is globally convergent for convex problems. We also present empirical computational results conducted in a similar environment to [13], using an advanced nonlinear conjugate gradient method [17,18] for the subproblems and standard nonconvex test problems from the CUTE test set [7]. These results show that our new algorithm's performance is very close to the heuristic rule of [13] and, in the context in which we performed our tests, superior to approximation rules adapted from those used by the popular augmented Lagrangian solvers Algencan [1,2] and Lancelot-B [9,10]. These rules are in turn faster than absolute summable criteria based on [12], or solving each subproblem essentially exactly.

As a preview and summary of the main results, specializing the general approximation framework we will propose to the problem (1) produces the following set of recursive conditions, with arbitrary starting values $\lambda^0 \in \mathbb{R}^{m_1}$, $\mu^0 \in \mathbb{R}^{m_2}_+$, and $w^0 \in \mathbb{R}^n$:

$$y^k = \nabla_x \Big[f(x) + \langle \lambda^{k-1}, h(x) \rangle + \frac{c_k}{2} \|h(x)\|^2 + \big\| \max\{0, \mu^{k-1} + c_k g(x)\} \big\|^2 \Big] \tag{11}$$

$$\frac{2}{c_k} \left| \langle w^{k-1} - x^k, y^k \rangle \right| + \|y^k\|^2 \leq \sigma \left(\|h(x^k)\|^2 + \left\| \min\left\{ \frac{1}{c_k} \mu^{k-1}, -g(x^k) \right\} \right\|^2 \right) \tag{12}$$

$$\lambda^k = \lambda^{k-1} + c_k h(x^k) \tag{13}$$

$$\mu^k = \max\{0, \mu^{k-1} + c_k g(x^k)\}. \tag{14}$$

$$w^k = w^{k-1} - c_k y^k. \tag{15}$$

Here, (11)–(12) replace the exact augmented Lagrangian minimization (3); the "min" in (12), like the "max" in (11), is interpreted componentwise. To execute the first step of the classical augmented Lagrangian method (3)–(5), one would in general apply some iterative method to the problem (3) until it finds point x where the augmented Lagrangian gradient y is nearly 0; to satisfy (11)–(12), we similarly apply an iterative method to (3), but truncate it as soon as the trial iterate x has a corresponding augmented Lagrangian gradient y small enough to satisfy (12). Next, (13) and (14) are just the usual augmented Lagrangian multiplier updates, but (15) and the way the sequence $\{w^k\}$ appears in (12) are novel features. Relative-error variants of the proximal point algorithm typically involve some kind of "corrector" to the basic proximal step, either a projection as in [28], or an extragradient step as in [27,29]. 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 in (11) or the multiplier updates (13)–(14). Instead, it only appears in the approximation condition (12) and the extragradient-like auxiliary update (15). Note that if we were 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 accumulated "error drift" in the sequence of calculations, something novel in augmented Lagrangian algorithms.

The quantity $\nu^k = \|h(x^k)\|^2 + \|\min\{(1/c_k)\mu^{k-1}, -g(x^k)\}\|^2$ on the right-hand side of (12) measures how much the trial primal-dual solution $\left(x^k, (\lambda^k, \mu^k)\right)$ satisfies the feasibility and complementarity conditions for optimality of (1), whereas $\|y^k\|$ measures how much $\left(x^k, (\lambda^k, \mu^k)\right)$ violates the remaining condition for optimality, stationarity of the Lagrangian. In the exact augmented Lagrangian method (3)–(5), one could in principle use $\nu^k \approx 0$ as a termination condition, since exact optimization of the augmented Lagrangian guarantees $y^k = 0$. Because its left-hand side is closely related to $\|y^k\|$, one can interpret the approximation criterion (12) as requiring that stationarity violation be bounded by a quantity proportional to the feasibility/complementarity violation.

There is another line of research into augmented Lagrangians algorithms that directly addresses differentiable nonconvex problems; see [1–5,8–10,15]. Among these references, the most general global convergence results are given in [4]; there, as in [1–3], the approximation criterion is simply $\|y^k\| \leq \epsilon_k$, where $\epsilon_k \to 0$ is a sequence of positive scalars. In this case, under a suitable weak constraint qualification, it is possible to show that the limit points of the primal sequence $\{x^k\}$ are either KKT points—that is, points x such that there exists a multiplier vector $p = (\lambda, \mu)$ with (x, p) satisfying the KKT conditions—or stationary for a natural infeasibility measure. However, such results do not show that the dual sequence $\{p^k\}$ converges in the convex case; actually, it is in general not possible to show that the dual sequence is bounded. To avoid unbounded multipliers, the methods described in [1–4] impose artificial bounds on them, but update the penalty parameter sequence $\{c_k\}$ in such a way that if the artificial multiplier bounds are binding, one has $c_k \to \infty$, essentially converting the algorithms into a pure penalty methods.

Thus, an important topic in this line of research concerns finding conditions that ensure that the penalty parameters remain bounded. To obtain such boundedness, relatively stringent assumptions on the problem are required, usually the strong second-order sufficiency condition and regularity [1], or, more recently, the strict Mangasarian-Fromowitz constraint qualification [6]. The analysis of [1], for example, guarantees boundedness of the penalty parameters if the approximation criterion is strengthened to

$$\|y^k\| \le \min\left\{\epsilon_k, \eta_k \left\| \begin{array}{c} h(x^k) \\ \min\{\frac{1}{c_k} p^{k-1}, -g(x^k)\} \end{array} \right\| \right\},$$

where both $\{\epsilon_k\}$ and $\{\eta_k\}$ are positive sequences converging to zero. The form of this last condition suggests an interesting possible connection with the criterion proposed in this paper; see (12) and also (59) below. But observe that, once again, the theory does not provide any direct guidance on how to choose the sequence of parameters $\{\epsilon_k\}$; the same comment also applies here to $\{\eta_k\}$. Finally, an interesting recent result is able to avoid the requirement of a constraint qualification and still ensure the boundedness of the penalty parameters whenever the initial primal-dual pair is close enough to a KKT pair that conforms to the second-order sufficient condition and the initial penalty parameter is large enough [14].

We organize the rest of this paper as follows: Sect. 2 below will briefly review the general parametric duality framework of [22,24], apparently required to derive error criteria that are practical in the sense meant here. Next, Sect. 3 will develop a much more general version of the framework (11)–(15), and show that it reduces to (11)–(15) in the case of problem (1). Section 4 will establish the convergence properties of the generalized framework, which carry over immediately to the special case (11)–(15). The fundamental convergence result is slightly weaker than traditionally obtained for methods like (3)–(5), in that we do not show convergence of the dual sequence $\{p^k\} = \{(\lambda^k, \mu^k)\}$ to a unique limit; however, we do show that $\{p^k\}$ is bounded, with all its limit points being dual optimal solutions, while the sequence $\{x^k\}$ is asymptotically optimal, with all its limit points being primal optimal solutions. We will also show that a stronger result, asserting 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 (12). Section 5 discusses applying our framework to a more practical formulation which enhances (1) with "box" constraints, Sect. 6 presents the computational results, and Sect. 7 presents some concluding remarks.

The results of this paper should extend to inner products more general than the canonical $\langle u, v\rangle = u^\top v$, so long as the corresponding norm and adjoint operator are used in place of "$\|\cdot\|$" and "$^\top$". Such techniques could also be used to extend the results to infinite dimension. For simplicity, we will not address such generalizations further here.

2 General parametric convex duality framework

We now briefly review the general parametric convex duality framework from [22, Chaps. 29–30] and [24]. We suppose that we have a closed (lower semicontinuous) proper convex function $F : \mathbb{R}^n \times \mathbb{R}^m \to (\infty, +\infty]$, and we wish to solve the *primal* problem

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

The second argument to F represents some kind of perturbation to the primal problem (16). The customary choice for modeling problem (1) is to set $m = m_1 + m_2$, partition $u = (r, s)$ with $r \in \mathbb{R}^{m_1}$ and $s \in \mathbb{R}^{m_2}$, and define

$$F\big(x, (r, s)\big) = \begin{cases} f(x), & \text{if } h(x) + r = 0 \quad \text{and} \quad g(x) + s \leq 0 \\ +\infty, & \text{otherwise.} \end{cases} \tag{17}$$

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

$$Q(y, p) = \inf_{\substack{x \in \mathbb{R}^n \\ u \in \mathbb{R}^m}} \big\{F(x, u) - \langle x, y\rangle - \langle u, p\rangle\big\}, \tag{18}$$

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

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

A simple application of Fenchel's inequality—see for example [22, 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 $\partial Q : \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m$ denote its subgradient map (the negative of its supergradient map), that is,

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

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

$$L(x, p) = \inf_{u \in \mathbb{R}^m} \big\{F(x, u) - \langle u, p\rangle\big\}. \tag{21}$$

If we compute L for the choice of F given in (17), with p partitioned as $p = (\lambda, \mu)$ for $\lambda \in \mathbb{R}^{m_1}$ and $\mu \in \mathbb{R}^{m_2}$, we obtain

$$L\big(x, (\lambda, \mu)\big) = \begin{cases} f(x) + \langle \lambda, h(x)\rangle + \langle \mu, g(x)\rangle, & \mu \geq 0, \\ -\infty, & \text{otherwise,} \end{cases} \tag{22}$$

which is the Lagrangian ordinarily associated with problem (1), along with the requirement that the inequality-constraint Lagrange multipliers μ 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,

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

We point out that the point-to-set maps ∂F, ∂Q, and ∂L are all maximal monotone operators, and

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

that is, ∂F and ∂Q are inverses of one another, and ∂L is a *partial inverse* [30] 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 *saddle point* (of the Lagrangian L). If such a saddle point 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 (11)–(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

$$\left(y^k, \tfrac{1}{c_k}(p^{k-1} - p^k)\right) \in \partial L(x^k, p^k) \tag{24}$$

$$2c_k \left|\langle w^{k-1} - x^k, y^k\rangle\right| + c_k^2 \|y^k\|^2 \leq \sigma \|p^{k-1} - p^k\|^2 \tag{25}$$

$$w^k = w^{k-1} - c_k y^k. \tag{26}$$

At this point, (25) and (26) may seem unmotivated; we will attempt to provide insight into these choices later, as we proceed with the convergence proof. First, however, we will show that, with F defined as in (17), the conditions (24)–(26) reduce exactly to our proposed algorithm (11)–(15).

Consider F as defined in (17), with its perturbation argument partitioned $u = (r, s)$. Similarly partitioning $p = (\lambda, \mu)$, we obtain when $\mu \geq 0$ that

$$\begin{aligned} \partial L\big(x, (\lambda, \mu)\big) &= \{\nabla f(x) + \nabla h(x)^\top \lambda + \nabla g(x)^\top \mu\} \times \{-h(x)\} \times \big(-g(x) + N_{\mathbb{R}^{m_2}_+}(\mu)\big) \\ &= \{\nabla f(x) + \nabla h(x)^\top \lambda + \nabla g(x)^\top \mu\} \times \{-h(x)\} \\ &\quad \times \{-g(x) + q \mid q \leq 0, \langle \mu, q\rangle = 0\}, \end{aligned}$$

where $N_{\mathbb{R}^{\ell}_{+}}$ denotes the normal cone mapping of the nonnegative orthant in $\mathbb{R}^{\ell}$; note that if $\mu \not\geq 0$, then $\partial L\big(x, (\lambda, \mu)\big) = \emptyset$. Inserting this form of L into (24) and partitioning p^k as $p^k = (\lambda^k, \mu^k)$, we obtain that (24) is equivalent to the conditions

$$y^k = \nabla f(x^k) + \nabla h(x^k)\lambda^k + \nabla g(x^k)\mu^k \tag{27}$$

$$\tfrac{1}{c_k}(\lambda^{k-1} - \lambda^k) = -h(x^k) \tag{28}$$

$$\tfrac{1}{c_k}(\mu^{k-1} - \mu^k) \in -g(x^k) + N_{\mathbb{R}^m_+}(\mu^k). \tag{29}$$

Rearranging (28), we obtain $\lambda^k = \lambda^{k-1} + c_k h(x^k)$, exactly as in (4) and (13). Rearranging (29) in a similar manner yields

$$\big(\mu^{k-1} + c_k g(x^k)\big) - \mu^k \in N_{\mathbb{R}^m_+}(\mu^k),$$

which is equivalent to μ^k being the unique projection of $\mu^{k-1} + c_k g(x^k)$ onto $\mathbb{R}^{m_2}_+$, that is,

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

Thus, we obtain exactly the classical inequality multiplier update in (5) and (14). Substituting this expression and $\lambda^k = \lambda^{k-1} + c_k h(x^k)$ into (27), we obtain

$$y^k = \nabla f(x^k) + \nabla h(x^k)^{\top}\big(\lambda^{k-1} + c_k h(x^k)\big) + \nabla g(x^k)^{\top} \max\Big\{0, \mu^{k-1} + c_k g(x^k)\Big\},$$

which is equivalent to (11). Thus, condition (24) is simply equivalent to y^k being the gradient of the usual augmented Lagrangian of (3) at x^k, with (λ^k, μ^k) being obtained by the usual multiplier updates (13) and (14).

Next, we turn our attention to the approximation condition (25). Using the multiplier update formulas to substitute into the expression on its right-hand side, we note that

$$\begin{aligned} p^{k-1} - p^k = \begin{bmatrix} \lambda^{k-1} - \lambda^k \\ \mu^{k-1} - \mu^k \end{bmatrix} &= \begin{bmatrix} \lambda^{k-1} - (\lambda^{k-1} + c_k h(x^k)) \\ \mu^{k-1} - \max\big\{0, \mu^{k-1} + c_k g(x^k)\big\} \end{bmatrix} \\ &= \begin{bmatrix} -c_k h(x^k) \\ \min\{\mu^{k-1}, -c_k g(x^k)\} \end{bmatrix}. \end{aligned}$$

Substituting this expression into (25) produces

$$2c_k \Big|\langle w^{k-1} - x^k, y^k\rangle\Big| + c_k^2\|y^k\|^2 \leq \sigma\Big(c_k^2\|h(x^k)\|^2 + \big\|\min\{\mu^{k-1}, -c_k g(x^k)\}\big\|^2\Big).$$

Dividing this relation by c_k^2 yields precisely (12). Since (26) and (15) are identical, it follows that when F is defined as in (17), the abstract recursions (24)–(26) reduce exactly to the approximate multiplier method (11)–(15) presented in Sect. 1.

The same basic mode of analysis may be used to specialize (24)–(26) to many other kinds of convex problems, such as those involving general conic constraints (for example, for cones of semidefinite matrices), and nonsmooth functions. We next prove the convergence of the abstract method, which immediately yields convergence of specializations such as (11)–(15).

4 Convergence proof for the abstract method

Proposition 1 *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to (\infty, +\infty]$ be closed proper convex, with Q and L defined as in (18) and (21), respectively. Let $\sigma \in [0, 1)$ and let $\{c_k\}_{k=1}^{\infty} \subset \mathbb{R}$ be such that $\inf_{k \geq 1}\{c_k\} > 0$. Suppose that $\{x^k\}_{k=1}^{\infty}, \{y^k\}_{k=1}^{\infty}, \{w^k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ and $\{p^k\}_{k=0}^{\infty} \subset \mathbb{R}^m$ obey for all $k \geq 1$ the recursions (24)–(26). Define, for all $k \geq 1$,*

$$u^k = \tfrac{1}{c_k}(p^{k-1} - p^k). \tag{30}$$

If there exists any saddle point of L, that is, any $(x^*, p^*) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $0 \in \partial L(x^*, p^*)$, then the following hold:

- *The sequences $\{p^k\}$ and $\{w^k\}$ are bounded.*
- *$u^k \to 0$ and $y^k \to 0$.*
- *$F(x^k, u^k)$ 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 saddle point exists, then at least one of the sequences $\{p^k\}$ or $\{x^k\}$ is unbounded.

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

$$\begin{aligned}\|p^{k-1} - p^*\|^2 &= \|p^{k-1} - p^k + p^k - p^*\|^2 \\ &= \|p^{k-1} - p^k\|^2 + 2\langle p^{k-1} - p^k, p^k - p^*\rangle + \|p^k - p^*\|^2.\end{aligned} \tag{31}$$

Using the definition of u^k, we have $p^{k-1} - p^k = c_k u^k$, which we may substitute into (31) 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

$$\|p^k - p^*\|^2 = \|p^{k-1} - p^*\|^2 - 2c_k\langle u^k, p^k - p^*\rangle - \|p^{k-1} - p^k\|^2. \tag{32}$$

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

$$\begin{aligned}\|w^k - x^*\|^2 &= \|w^{k-1} - c_k y^k - x^*\|^2\\ &= \|w^{k-1} - x^*\|^2 - 2\langle w^{k-1} - x^*, c_k y^k\rangle + c_k^2\|y^k\|^2\\ &= \|w^{k-1} - x^*\|^2 - 2c_k\langle w^{k-1} - x^k, y^k\rangle - 2c_k\langle x^k - x^*, y^k\rangle + c_k^2\|y^k\|^2.\end{aligned} \tag{33}$$

Next, we add (32) and (33) to obtain

$$\begin{aligned}\|p^k - p^*\|^2 + \|w^k - x^*\|^2 = \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2&\\ -2c_k\left[\langle x^k - x^*, y^k\rangle + \langle p^k - p^*, u^k\rangle\right]&\\ -2c_k\langle w^{k-1} - x^k, y^k\rangle + c_k^2\|y^k\|^2&\\ -\|p^{k-1} - p^k\|^2.&\end{aligned} \tag{34}$$

Next, we use the monotonicity of ∂L to eliminate the expression on the second line of (34). Specifically, from (24) and (30) we have that $(y^k, u^k) \in \partial L(x^k, p^k)$, and since (x^*, p^*) is a saddle point, we also have $(0, 0) \in \partial L(x^*, p^*)$. Thus, the monotonicity of ∂L yields

$$\langle x^k - x^*, y^k - 0\rangle + \langle p^k - p^*, u^k - 0\rangle = \langle x^k - x^*, y^k\rangle + \langle p^k - p^*, u^k\rangle \geq 0. \tag{35}$$

Combining this inequality with (34) yields

$$\begin{aligned}\|p^k - p^*\|^2 + \|w^k - x^*\|^2 \leq \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2&\\ - 2c_k\langle w^{k-1} - x^k, y^k\rangle + c_k^2\|y^k\|^2&\\ - \|p^{k-1} - p^k\|^2.&\end{aligned} \tag{36}$$

The error criterion (25) is designed so that the terms on the second line of (36) may be “buried” in the last term. Specifically, (25) implies

$$\begin{aligned}- 2c_k\langle w^{k-1} - x^k, y^k\rangle + c_k^2\|y^k\|^2 &\leq 2c_k\left|\langle w^{k-1} - x^k, y^k\rangle\right| + c_k^2\|y^k\|^2\\ &\leq \sigma\|p^{k-1} - p^k\|^2.\end{aligned} \tag{37}$$

Substituting this inequality into (36), we obtain an inequality that is the key to the convergence analysis:

$$\|p^k - p^*\|^2 + \|w^k - x^*\|^2 \leq \|p^{k-1} - p^*\|^2 + \|w^{k-1} - x^*\|^2 - (1-\sigma)\|p^{k-1} - p^k\|^2. \tag{38}$$

Since (38) holds for all $k \geq 1$, a cascade of deductions follows:

- $\left\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\right\}$ is a nonincreasing sequence, so the sequences $\{p^k\}$ and $\{w^k\}$ are bounded. Since $\left\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\right\}$ is bounded below by 0, it must be convergent. Since (x^*, p^*) was an arbitrary saddle point, we further conclude that $\{(x^k, p^k)\}$ is Fejér monotone to the set of saddle points.
- Summing (38) over k and using that $\sigma < 1$, we conclude that $\{\|p^{k-1} - p^k\|^2\}$ is a summable sequence and hence that $p^k - p^{k-1} \to 0$.
- Since c_k is bounded away from 0 and $u^k = (1/c_k)(p^k - p^{k-1})$, it follows that $u^k \to 0$ and $\{\|u^k\|^2\}$ is summable.
- Referring to (25) or (37), and using the just-established properties of the sequence $\{p^k - p^{k-1}\}$, it then also follows that the two sequences $\{c_k|\langle w^{k-1} - x^k, y^k\rangle|\}$ and $\{c_k^2\|y^k\|^2\}$ are both summable, and thus converge to 0. Immediately, we have $c_k\langle w^{k-1} - x^k, y^k\rangle \to 0$.
- Since it is absolutely summable, $\{c_k\langle w^{k-1} - x^k, y^k\rangle\}$ is also summable.
- Again using that c_k is bounded away from 0, the last two sets of observations imply that $\{|\langle w^{k-1} - x^k, y^k\rangle|\}$, $\{\langle w^{k-1} - x^k, y^k\rangle\}$, and $\{\|y^k\|^2\}$ are all summable and convergent to 0. and in particular we have $y^k \to 0$.
- 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 $y^k \to 0$, we have $\langle w^{k-1}, y^k\rangle \to 0$. Since we have already established that $\langle w^{k-1} - x^k, y^k\rangle \to 0$, it follows that $\langle x^k, y^k\rangle \to 0$.

Next, applying (23) to $(y^k, u^k) \in \partial L(x^k, p^k)$ gives that $(x^k, u^k) \in \partial Q(y^k, p^k)$. Thus, combining 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

$$Q(y^k, p^k) \geq Q(0, p^*) - \langle x^k, y^k\rangle + \langle u^k, p^* - p^k\rangle. \tag{39}$$

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

$$\liminf_{k\to\infty} Q(y^k, p^k) \geq Q(0, p^*). \tag{40}$$

We now consider $\limsup_{k\to\infty} Q(y^k, p^k)$, which must by (40) be at least $Q(0, p^*)$; however, we have not yet excluded the possibility that it may be larger, perhaps $+\infty$. Let $\mathcal{K}$ be a subsequence such that $Q(y^k, p^k) \to_{\mathcal{K}} \limsup_{k\to\infty} Q(y^k, p^k)$. 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

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

Combining this result with (40), we conclude that

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

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

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

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

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 $y^k \to 0$, analogous reasoning implies that all limit points of $\{p^k\}$ are dual solutions. The proof for the case that at least one saddle point exists is now complete.

It remains only to consider the case that no saddle point exists. In this case, we use a variant of the analysis originally given in [26] for the behavior of the proximal point algorithm for operators with no roots. The proof proceeds by contradiction: suppose that no saddle point 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 [23], T is maximal monotone, since T and ∂L are both maximal monotone and $\operatorname{dom} \partial L$ and $\operatorname{int} \operatorname{dom} N_B$ have nonempty intersection. Furthermore, since $\operatorname{dom} T$ lies within B and is therefore a bounded set, it follows from [21, Proposition 2] that there exists at least one point $(x^*, p^*) \in \mathbb{R}^n \times \mathbb{R}^m$ for which $(0, 0) \in T(x^*, p^*)$.

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 (35) still holds, but with (x^*, p^*) assumed to be any root of T. Following the logic above for the case in which at least one saddle point exists, all the results above from (36) through $\langle x^k, y^k \rangle \to 0$ 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 saddle point, which contradicts the hypothesis. Therefore, the assumption above that one can simultaneously have no saddle points with both $\{p^k\}$ and $\{x^k\}$ bounded cannot hold. □

The properties of $\{w^k\}$ are unusual. Although $\{(w^k, p^k)\}$ is Fejér monotone to the set of saddle points, 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 intepretation of the role of $\{w^k\}$ is that it accumulates, through (26), the total "error drift" of the algorithm. If a large amount of drift accumulates in the sense that $w^{k-1} - x^k$ becomes large, the subproblem optimality tolerance may be effectively tightened, in that the component of y^k parallel to $w^{k-1} - x^k$ will have to be small.

The conclusions of Proposition 1 are somewhat weaker than are typically obtained for multiplier methods, either in their exact form or with the exact summable error criterion of [12]. 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 saddle points exist, but the set of primal solutions is unbounded; however, in practical implementations, such behavior of $\{x^k\}$ is generally not of concern. When saddle points 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 saddle point exists. If for some scalar $\zeta \geq 0$ it is also true for all $k \geq 1$ that*

$$c_k\|y^k\| \le \zeta\|p^{k-1} - p^k\|^2, \tag{41}$$

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 saddle point exists must hold. Again letting (x^*, p^*) denote an arbitrary saddle point and rearranging (34), we obtain

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

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

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

Since $\left\{\|p^k - p^*\|^2 + \|w^k - x^*\|^2\right\}$ is convergent, 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

$$\left\{c_k\langle x^k - x^*, y^k\rangle + c_k\langle p^k - p^*, u^k\rangle\right\} \tag{42}$$

is summable. Next, we will use the additional hypothesis (41) to show that the first term $\{c_k\langle x^k - x^*, y^k\rangle\}$ above is summable, with the consequence 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 (41) 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 (42), 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 its right-hand side above are summable, summing the above equation leads to the conclusion that $\{\|p^k - p^*\|^2\}$ converges.

Next, consider any limit point p^∞ of the bounded sequence $\{p^k\}$, which we know from Proposition 1 must be a dual solution. Since the set of primal-dual solutions form a Cartesian product [22, Corollary 30.5.1], we know that (x^*, p^∞) is also a saddle point, 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 $p^k \to p^\infty$. □

Although the constant ζ may be arbitrarily large, the additional approximation criterion (41) 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 Including variable bounds

Although in principle it is no more general, we now consider a version of (1) including explicit bounds on the variables, namely

$$\begin{array}{ll}\min_{x\in\mathbb{R}^n} & f(x) \\ \text{s.t.} & g(x) \le 0 \\ & h(x) = 0 \\ & a \le x \le b,\end{array} \tag{43}$$

where $a \in [-\infty, \infty)^n$, $b \in (-\infty, \infty]^n$, and $a \le b$.

Due to their simple structure, we will directly enforce the constraints $a \le x \le b$ in the subproblems, rather than attaching Lagrange multipliers to them. Let

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

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 $z \in \mathbb{R}^n$,

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

Partitioning the second argument to F as before into $u = (r, s)$, we choose F as follows:

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

Again partitioning $p = (\lambda, \mu)$, the corresponding Lagrangian takes the form

$$L\big(x, (\lambda, \mu)\big) = \begin{cases} f(x) + \langle \lambda, h(x)\rangle + \langle \mu, g(x)\rangle, & \text{if } a \leq x \leq b \quad \text{and} \quad \mu \geq 0 \\ -\infty, & \text{if } a \leq x \leq b \quad \text{and} \quad \mu \ngeq 0 \\ +\infty, & \text{if } a \nleq x \text{ or } x \nleq b. \end{cases}$$

Note that the form of L effectively enforces the dual constraint $\mu \geq 0$. The corresponding extended dual function is $Q\big(y, (\lambda, \mu)\big) = \inf_{x \in \mathbb{R}^n} \big\{L\big(x, (\lambda, \mu)\big) - \langle x, y\rangle\big\}$. Applying the proximal point algorithm to the map $\partial Q\big(0, (\cdot\,, \cdot)\big)$ produces an augmented Lagrangian method identical to (3)–(5), except that (3) is replaced by the box-constrained augmented Lagrangian minimization

$$x^k \in \operatorname*{Arg\,min}_{x \in B(a,b)} \Big\{ f(x) + \langle \lambda^{k-1}, h(x)\rangle + \tfrac{c_k}{2}\|h(x)\|^2 + \tfrac{1}{2c_k}\big\| \max\{0, \mu^{k-1} + c_k g(x)\}\big\|^2 \Big\}. \tag{44}$$

Applying the recursions (24)–(26) to this form of L reduces, after defining

$$L_k(x) = f(x) + \langle \lambda^{k-1}, h(x)\rangle + \frac{c_k}{2}\|h(x)\|^2 + \frac{1}{2c_k}\big\| \max\{0, \mu^{k-1} + c_k g(x)\}\big\|^2, \tag{45}$$

to the approximate augmented Lagrangian recursions

$$y^k \in \nabla L_k(x^k) + N_{B(a,b)}(x^k) \tag{46}$$

$$\frac{2}{c_k}\Big|\langle w^{k-1} - x^k, y^k\rangle\Big| + \|y^k\|^2 \leq \sigma\left(\|h(x^k)\|^2 + \left\|\min\left\{\frac{1}{c_k}\mu^{k-1}, -g(x^k)\right\}\right\|^2\right) \tag{47}$$

$$\lambda^k = \lambda^{k-1} + c_k h(x^k) \tag{48}$$

$$\mu^k = \max\{0, \mu^{k-1} + c_k g(x^k)\} \tag{49}$$

$$w^k = w^{k-1} - c_k y^k. \tag{50}$$

The only difference from (11)–(15) is the presence of $N_{B(a,b)}(x^k)$ in (46), due to the box constraints $a \leq x \leq b$. Note that (46) 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 (44).

To implement (46)–(50) computationally, one would apply some iterative bound-constrained solver to the problem (44), 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 (47). 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 $x^k \leq b$ is binding. Observing from the Cauchy-Schwarz inequality that $|\langle w^{k-1} - x^k, y^k \rangle| \leq \|w^{k-1} - x^k\| \|y^k\|$, one simple strategy for trying to satisfy (47) with as few iterations as possible of the bound-constrained subproblem solver is to 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 [13]. To compute the vector y with the minimum norm in the set $t + N_{B(a,b)}$, for any $t \in \mathbb{R}^n$, one may use the following simple calculation:

$$y_i = \begin{cases} \min\{t_i, 0\}, & \forall i : \ x_i = a_i \\ t_i, & \forall i : \ a_i < x_i < b_i \\ \max\{t_i, 0\}, & \forall i : \ x_i = b_i . \end{cases} \tag{51}$$

We close this section by observing that Proposition 1 implies that, whenever there exists at least one saddle point $(x^*, p^*) = \big(x^*, (\lambda^*, \mu^*)\big)$, we have that $u^k \to 0$ and $F(x^k, u^k) \to F(x^*, 0)$, where $\{u^k\}$ is defined by (30). For the current choice of F, these conditions reduce to

$$\limsup_{k\to\infty} g_i(x) \leq 0, \quad i = 1, \ldots, m_1 \qquad h(x^k) \to 0 \qquad f(x^k) \to f(x^*).$$

Such behavior of the primal sequence $\{x^k\}$ is often referred to as *asymptotic optimality*; see for example [12]. In particular, all accumulation points of $\{x^k\}$ are primal solutions.

6 Computational testing

We now describe some preliminary computational testing of the algorithm (46)–(50), using a subset of problems from the CUTE test set [7]. 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 $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^{m_1}$, and $h : \mathbb{R}^n \to \mathbb{R}^{m_2}$ 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 [13]. As in [13], we implemented our main algorithm (the "outer loop") in Python [31], using SciPy [19], an open-source software environment with capabilities similar to MATLAB; in fact, the implementation is a minor enhancement to the existing "pyauglag" prototype code already developed in [13].

The "inner loop" of the implementation consists of the procedure necessary to identify some pair (x^k, y^k) jointly satisfying (46) and (47). To this end, much as in [13], we used a slightly modified version of the ASA bound-constrained conjugate gradient code of Hager and Zhang [18], which is based on the advanced unconstrained conjugate gradient algorithm described in [17]. Our only significant modification to the base ASA code [16] is the ability to use a user-specified termination criterion. Specifically, we ran ASA on the bound-constrained problem (44), 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 (51) with $t = \nabla L_k(x)$.
2. Determine whether

$$\frac{2}{c_k}\left|\langle w^{k-1} - x, y\rangle\right| + \|y\|^2 \ \leq\ \sigma\left(\|h(x)\|^2 + \left\|\min\left\{\frac{1}{c_k}\mu^{k-1}, -g(x)\right\}\right\|^2\right) \tag{52}$$

or

$$\|y\|_\infty \leq \delta, \tag{53}$$

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 (52) or (53) 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 [7] 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 [13], to facilitate direct comparison with our recent computational work there. The exact selection of problems is described in [13]; in brief, some of the larger problems were excluded due to prototype nature of our implementation, which contains extensive intepreted Python code.

To terminate the outer loop, we use a condition similar to that of [13] and also inspired by Algencan [1,2], a well established general-purpose augmented Lagrangian solver. Given a positive penalty parameter $c > 0$, we define

$$\gamma(x, p, q) = \min\left\{\|y\|_\infty \ \middle|\ y \in \nabla_x L\big(x, (p, q)\big) + N_{B(a,b)}(x)\right\} \tag{54}$$

$$\phi_c(x) = \max\{\|h(x)\|_\infty, \|\max\{-p/c, g(x)\}\|_\infty\}. \tag{55}$$

Here, $\gamma(x, p, q)$ measures how close x comes to being the minimizer of $L\big(\cdot, (p, q)\big)$ over $B(a, b)$, while $\phi_c(x)$ simultaneously measures the feasibility of x and the violation of complementary slackness for the inequality constraints $g(x) \leq 0$. If $\gamma(x, p, q) = 0$ and $\phi_c(x) = 0$, then (x, p, q) satisfies the KKT conditions for (43). 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

$$\gamma(x^k, p^k, q^k) = \|y^k\|_\infty < \epsilon \quad \text{and} \quad \phi_{c_k}(x^k) < \epsilon. \tag{56}$$

We also experimented with using a fixed c in place of c_k in (56), without any significant change in the computational results. For easier comparison with the results presented in [13], 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 (54)–(55). We defined the constant δ appearing in (53) to be $\epsilon/10$. Note that we allow (53) to serve as an alternative subproblem termination condition to (52) because checking only (52) might require more exact solution of subproblems than warranted by the overall tolerance ϵ; if (52) is more stringent than (53), one may consider the relative error criterion to have already "served its purpose", at least for the time being, and instead proceed as one would in classical augmented Lagrangian implementation.

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

- We still have not satisfied the approximate KKT conditions (56) after 200 outer iterations ($k \geq 200$)
- The ASA subproblem solver fails more than 5 times
- There are more than 1 million function evaluations
- The total CPU time exceeds 1 h (on a single core of a 2.83 GHz Intel Core 2 Quad Q9550 processor with 800 MHz memory).

We use the following strategy to adjust the penalty parameter c_k, once again based on the technique used in the Algencan [1,2]. At the end of iteration k, we test whether

$$\phi_{c_k}(x^k) < \epsilon \quad \text{or} \quad \phi_{c_k}(x^k) \leq 0.5\,\phi(x^{k-1}). \tag{57}$$

If (57) holds, 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, with 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 future, a more sophisticated approach might base c_0 on a scaling analysis of the problem and starting point).

Below, we experimentally compare the new approximation criterion (52)–(53) with a number of alternative criteria described in existing literature, as well as to solving each subproblem to the fixed "exact" precision δ, that is, using (53) only. We tested all the alternatives within the same prototype solver implementation described above and in [13]. The first alternative approximation criterion is a special case of the absolute summable error criterion from [12, formula (17)]. Specifically, we tested an approximation criterion of the form

$$\|y^k\| \leq \frac{\epsilon_k}{c_k \gamma_k}, \quad \text{with} \quad \gamma_k = \begin{cases} 1, & \|x^k\| \leq \beta \\ \|x^k\|/\beta, & \|x^k\| > \beta, \end{cases} \tag{58}$$

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 (58) is a special case of the criterion specified in [12]. As for the choice of the summable sequence $\{\epsilon_k\}$, we experimented with various sequences of the form $\epsilon_k = \eta/k^\zeta$, where η is a positive constant and $\zeta > 1$. After some numerical experimentation, we settled on $\eta = 0.1$ and $\zeta = 2$, which seemed to perform the best on our experimental test set.

Next, we compared the new approach with adaptations of the error criteria used by the existing augmented Lagrangian codes Algencan [1,2] and Lancelot-B [9,10]. The Algencan criterion involves a simple two-stage process: recall that δ is the threshold that determines when the subproblems are solved "almost exactly"; see (53) and the following discussion. The Algencan criterion starts by requiring that each subproblems be solved with fixed precision $\sqrt{\delta}$, that is, the subproblem solver is stopped once $\|y^k\|_\infty \le \sqrt{\delta}$. As soon as the stopping conditions for the outer loop (56) hold with precision $\sqrt{\epsilon}$, the precision required in the subproblems is tightened to δ, and remains fixed thereafter.

The criterion used by Lancelot is more complicated, and can be adapted to our setting as follows: let ϵ_k be the precision required in the k^{th} iteration, that is, the subproblem solution process halts once $\|y^k\|_\infty \le \epsilon_k$. After iterations where the algorithm has made good progress toward feasibility and complementarity slackness, as defined in (57), set $\epsilon_{k+1} = \epsilon_k/c_{k+1}$, otherwise set $\epsilon_{k+1} = \epsilon_0/c_{k+1}$. Note that, even though ϵ_{k+1} may be larger than ϵ_k in the former case, lack of good progress also triggers an increase in c_k, which will accelerate the rate of tolerance tightening in future iterations. We set the initial precision ϵ_0 to $0.1/c_0$, as Lancelot does.

Finally, we also compared the new approach with the heuristic criterion suggested in [13], 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

$$\|y^k\|^2 \le \sigma\left(\|h(x^k)\|^2 + \left\|\min\left\{\frac{1}{c_k}\mu^{k-1}, -g(x^k)\right\}\right\|^2\right), \tag{59}$$

that is, (47) with the term involving w^{k-1} deleted. This criterion was proposed in [13], where it was developed by analogy with a similar but rigorously derived criterion for primal-dual methods like (6)–(8). In order to ensure theoretical convex-case convergence, the use of (59) in [13] included supplementary safeguards based on the summable criterion (58). 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 (52) to the best-performing heuristic criterion available, we simply used (59) without any safeguards.

Note the augmented Lagrangian tests based on the various acceptance criteria described above all use the same conditions to declare success and failure of the overall algorithm; see (56) and the following discussion. Moreover, all the algorithm variations use the same strategy to update the penalty parameters.

Both criteria (47) and (59) 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 $\sigma = 0.99$.

In the new error criterion, the choice of the initial reference vector 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 it is possible that $|\langle w^{k-1} - x, y\rangle|$ will be large in (52), making it much stricter than the heuristic criterion (59), 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 $(\lambda^{k-1}, \mu^{k-1})$ for which the inner loop is having trouble solving the subproblem (44) 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 [13]. 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 (47), effectively reducing it to (59). 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 (47) 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 $|\langle w^{k-1} - x, y\rangle|$ becoming large. Thus, at each trial point computed by the ASA solver, we check whether $\|w^k - x^k\| > 100c_k\|y^k\|$. 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 (45)–(50) and our convergence theory applies, at least in the convex case.

Figure 1 shows two performance profiles [11]. The upper profile compares the summable criterion (58), the Algencan criterion, the Lancelot criterion, and "exact" solution of all subproblems using only (53). We measure performance by counting the number of gradient evaluations; results for the number of function evaluations are similar. The raw data used to generate all the profiles are displayed in Appendix A. It is clear that the inexact approaches are clear improvements over exact subproblem solution, requiring less computational effort without any compromise in robustness. Moreover, the criteria from Algencan and Lancelot are faster than the summable criterion and very similar to one another. The lower profile in Fig. 1 compares the three previous inexact criteria 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 better sense the rate of approximation tightening appropriate to each problem instance.

Figure 2 displays a performance profile comparing the new relative error criterion with the heuristic relative error criterion (59), which was the best-performing approach

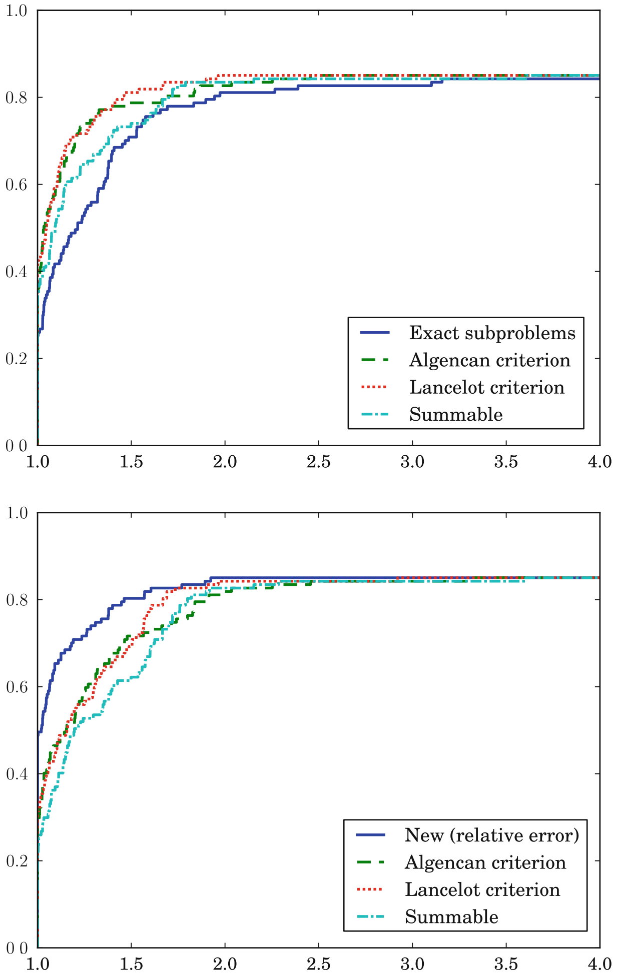

Fig. 1 Performance profiles, in terms of the number of gradient evaluations: the top profile compares exact subproblem minimization ("Exact subproblems"), the Algencan approximation criterion, the Lancelot approximation criterion, and using a predetermined summable error sequence $\{\epsilon_k\}$ ("Summable"), as in (58). The lower profile compares the three approximation criteria from the upper profile with our new relative error criterion (52)–(53) ("New (relative error)")

in [13]. Here, we see that the extra term involving w^k on the left-hand side of (47) has very little practical impact, slowing down convergence only slightly. On the other hand, it appears to have a small (but probably 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 currently 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.

In conclusion, our computational experiments suggest that, at least in the setting of our prototype implementation using a modern conjugate gradient method to solve the subproblems, the new approximation criterion improves on the procedures used in

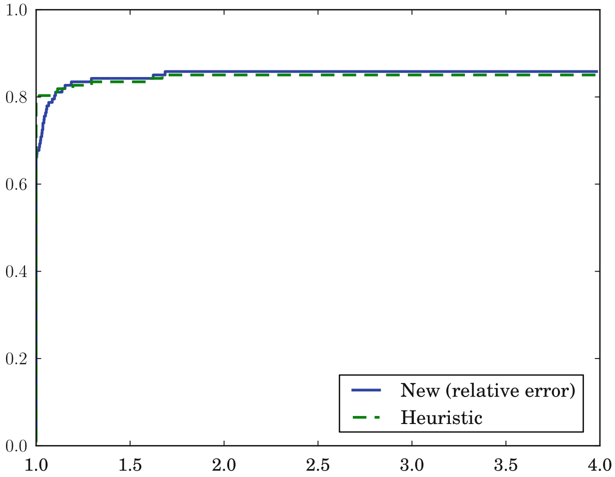

Fig. 2 Performance profile comparing the proposed algorithm ("New (relative error)") with the best algorithm in [13] ("Heuristic"), in terms of the number of gradient evaluations

established, popular augmented Lagrangian nonlinear optimization solvers, and thus has significant practical value. That the heuristic criterion (59), without the sequence $\{w^k\}$, seems to work about as well in practice as our proposed method, or perhaps slightly better, 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

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 (6), and yet requires only readily available information, namely the gradient (or a subgradient) of the augmented Lagrangian. It is also a true relative error criterion, in that it sets the precision requirement for each subproblem proportionally to the current violation of feasibility and complementarity—as, for example, on the right-hand side of (47). Thus, it is an advance over the results of [12], which require a summable sequence of error parameters $\{\epsilon_k\}$, and provide no direct guidance how to select it. Furthermore, in our computational tests using a conjugate-gradient "inner loop", the performance of our new, provably convergent method is very close to the best heuristic approach with which we have experimented so far, and appears to improve on approximation rules adapted from currently popular augmented Lagrangian solvers.

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 [12] 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 [12] into one involving a relative error criterion, because the resulting criteria always seem to require knowledge of the saddle point (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 saddle points. In some sense, one can view the method as a proximal algorithm in the dual variables p—or (λ, μ) for the problem (1) and (43)—and a kind of extragradient algorithm [20] 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.

Although the $\sigma = 0$ case of our method is exactly the pure-dual augmented Lagrangian method, our proposed error criterion (25) and the more concrete criteria we have derived from it are "primal-dual" in a similar sense to the absolute-error criterion proposed in [12], in that the new primal iterate x^k appears not only implicitly through the multiplier change $\|p^{k-1} - p^k\|$ appearing on the right of (25), but also in the expression involving the gradient norm $\|y^k\|$ on the left. In particular, if x^k becomes large relative to w^k, then the acceptance criterion may be effectively tightened by the presence of the term $|\langle w^{k-1} - x^k, y^k\rangle|$ on the left of (25). The criterion in [12], although it uses a given parameter sequence $\{\epsilon_k\}$, has a similar property. For example, in (58), if $\|x^k\|$ becomes large relative to β, then the error tolerance on $\|y^k\|$ becomes smaller than ϵ_k/c_k. In both cases, if $\{x^k\}$ appears to be "blowing up", the error criterion may be significantly tightened, which one may interpret as making sure that the behavior of $\{x^k\}$ is "legitimate", and not the result of excessively accumulation of error in the augmented Lagrangian minimizations.

In our planned continued work on this topic, it is of obvious interest to analyze the behavior of the method for nonconvex problems (to the extent possible), and further improve its practical reliability. Rate of convergence is another topic deserving of study. 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. Parallelizing the algorithm will require parallelization of the ASA subproblem solver, but this task seems likely to be feasible due to the separable structure of the box constraints and the relatively simple linear algebra required by the underlying conjugate gradient method of [17]. The outer loop of our method also seems readily parallelizable due to the the simple form of the updates to the multipliers and $\{w^k\}$.

We will also continue trying to prove the convergence of the simpler and more intuitive relative error criterion (59), which has similar practical performance to the method proposed here, but has so far resisted analysis.

Appendix A: Performance data table

We present below the raw data used to generate the perfomace profiles in Sect. 6. Each line of Table 1 displays the total number of evaluations of the problem Jacobian under each possible approximation criterion. Asterisks indicate cases in which the corresponding solver failed.

Table 1

Problem	Relative error	Heuristic	Algencan	Lancelot	Summable	Exact
airport	1,122	1,022	1,347	1,754	1,748	1,785
avgasa	101	93	125	122	117	145
avgasb	74	70	123	116	96	150
batch	38,885*	158,131*	105,764*	45,489*	39,396*	40,469*
cb2	27	27	28	25	43	33
cb3	27	27	29	24	40	33
chaconn1	27	27	28	25	43	33
chaconn2	27	27	29	24	40	33
congigmz	406	342	400	595	303	724
core1	440,077*	526,500*	602,649*	602,649*	602,649*	602,649*
core2	635,900*	548,055*	751,808*	767,343*	596,229*	596,229*
coshfun	4,735	2,808	6,044	2,461	3,688	7,772
cresc100	89,718*	89,718*	122,640*	76,069*	89,718*	89,718*
cresc4	2,870	2,826	2,969	3,750	2,880	3,033
cresc50	78,990*	98,399*	81,250*	77,144*	81,625*	75,818*
demymalo	156	156	149	155	160	146
dipigri	280	280	299	325	318	457
disc2	74,956*	4,561*	12,950*	2,488*	4,967*	8,071*
discs	1,061	1,017	841	923	844	808
dualc1	254	254	263	238	238	238
dualc2	69	69	83	83	81	81
dualc8	293	293	285	328	306	336
eg3	3,492	3,428	403	564	685	667
expfita	5,663	3,489	3,912	6,531	14,100	4,095
expfitb	2,115	2,115	3,942	2,985	1,937	3,278
fletcher	71	71	115	120	112	140
gigomez1	130	130	125	133	132	120
goffin	8,251	8,251	8,249	8,253	8,253	8,249
gpp	19,570	23,397	28,646	29,869	31,326	32,656
hadamard	16	16	16	16	16	18
haifam	528,240*	528,240*	528,240*	528,240*	528,240*	528,240*

Table 1 continued

Problem	Relative error	Heuristic	Algencan	Lancelot	Summable	Exact
haifas	138	138	82	78	109	148
haldmads	1,550	2,495	5,094	4,528	2,726	4,296
hanging	1,067	967	1,525	1,697	1,730	2,134
himmelbi	1	1	1	1	1	1
himmelp3	5	5	5	5	5	5
himmelp4	5	5	5	5	5	5
himmelp5	14	14	14	14	14	14
himmelp6	2	2	2	2	2	2
hs015	59	59	59	59	59	59
hs016	24	24	24	24	24	24
hs017	62	62	81	79	74	93
hs018	240	240	240	253	227	246
hs019	152	152	291	180	175	175
hs020	33	33	31	31	33	33
hs022	42	40	43	43	49	48
hs023	61	61	73	72	73	89
hs024	50	48	43	42	56	49
hs033	32	32	32	34	35	33
hs034	284	283	272	267	283	285
hs043	100	100	125	125	107	157
hs044	32	32	37	37	52	37
hs059	13	13	13	19	28	13
hs066	54	54	94	71	76	96
hs072	115	111	159	173	181	193
hs076	28	28	41	38	43	41
hs085	520,539*	520,539*	472,359*	532,854*	517,589*	518,435*
hs086	129	125	169	192	151	231
hs093	206*	206*	206*	206*	206*	206*
hs095	154	154	154	150	154	154
hs096	154	154	154	150	154	154
hs097	40	40	29	29	40	40
hs098	40	40	29	29	40	40
hs100	280	280	299	325	318	457
hs100mod	2,602	2,602	5,155	4,513	4,516	5,132
hs101	3,236	3,286	5,549	5,121	5,504	6,150
hs102	5,414	5,282	7,507	8,486	8,349	10,315
hs103	5,967	6,648	8,578	10,081	10,866	13,865
hs104	555	481	511	661	628	1,158
hs106	843,447*	771,795*	722,088*	640,932*	391,617*	391,617*

Table 1 continued

Problem	Relative error	Heuristic	Algencan	Lancelot	Summable	Exact
hs108	83	83	90	84	92	76
hs109	94,415*	94,415*	94,415*	94,415*	94,415*	94,415*
hs113	356	356	345	288	331	560
hs114	76,479*	76,479*	85,905*	88,428*	88,428*	88,428*
hs116	598,890*	289,163*	604,430*	1,124,882*	867,720*	1,236,783*
hs117	4,082	3,872	4,309	3,575	3,477	5,050
hs268	1	1	1	1	1	1
hs44new	19	19	19	19	34	19
kiwcresc	14	14	16	14	14	17
loadbal	351	351	469	267	255	379
lootsma	6	6	11	6	6	11
madsen	37	37	47	48	46	46
madsschj	1,728	1,728	1,709	1,703	1,690	1,709
makela1	32	32	34	33	39	36
makela2	23	23	23	24	31	19
makela3	417	417	431	422	425	436
makela4	632	632	632	630	634	632
matrix2	30	30	33	27	29	42
mifflin1	113	113	108	114	115	111
mifflin2	205	204	179	182	185	185
minmaxbd	404	404	397	370	429	516
minmaxrb	326	326	276	308	280	286
mistake	66	66	87	73	73	75
model	622,687	775,610*	860,808	679,486	840,008	648,121
optmass	643,890*	643,890*	643,890*	528,286*	528,286*	643,890*
optprloc	91,053	91,053	92,034	91,027	91,045	91,053
pentagon	105	105	201	169	240	147
polak1	5	5	5	5	5	5
polak3	305	268	442	401	299	590
polak4	1,893	1,893	1,205	2,278*	1,897	3,737
polak5	23	23	36	36	32	34
polak6	543	543	531	527	530	571
prodpl0	359	351	487	540	598	679
prodpl1	307	298	561	512	583	721
qpcstair	125,083	125,083	95,574	132,981	129,524	129,524
qpnstair	164,354*	147,700*	133,590*	135,286*	133,786*	134,960*
rosenmmx	309	309	304	294	327	331
s365mod	920*	920*	405*	390*	1,178*	416*
simpllpa	13	13	13	13	13	13

Table 1 continued

Problem	Relative error	Heuristic	Algencan	Lancelot	Summable	Exact
simpllpb	9	9	9	9	9	9
snake	24,072*	24,072*	23,122*	23,122*	23,122*	23,122*
spiral	657	657	671	720	731	663
sseblin	2,186	2,049	1,375	1,413	1,362	1,362
ssebnln	2,708	2,950	1,808	1,904	1,725	1,725
stancmin	71	71	73	77	73	85
swopf	248,526	321,896	558,715*	293,960	569,905*	569,905*
trimloss	1,195,701*	1,132,121*	2,483,114*	2,483,114*	2,483,114*	2,483,114*
twirism1	12,026	9,295	9,517	18,081	10,869	202,490*
twobars	43	43	54	56	49	63
vanderm1	54,434	90,951	83,864	229,563	37,233	49,192
vanderm2	217	217	195	169	271	180
vanderm3	205	205	377	327	393	225
vanderm4	142	142	75	100	107	78
womflet	42	42	30	42	48	38
zecevic2	24	24	23	23	23	23
zecevic3	29	28	38	38	51	44
zecevic4	15	15	27	22	27	30

References

1. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: On augmented Lagrangian methods with general lower-level constraints. SIAM J. Optim. **18**(4), 1286–1309 (2007)
2. Andreani, R., Birgin, E.G., Martínez, J.M., Schuverdt, M.L.: Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. Math. Program. **111**(1–2), 5–32 (2008)
3. Andreani, R., Haeser, G., Schuverdt, M., Silva, P.J.S.: A relaxed constant positive linear dependence constraint qualification and applications. Math. Program. Published electronically. doi:10.1007/s10107-011-0456-0 (2011)
4. Andreani, R., Haeser, G., Schuverdt, M., Silva, P.J.S.: Two new weak constraint qualifications and applications. Available at Optimization Online: http://www.optimization-online.org/DB_HTML/2011/07/3105.html (2011)
5. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York, NY (1982)
6. Birgin, E.G., Fernández, D., Martínez, J.M.: On the boundedness of penalty parameters in an augmented Lagrangian method with constrained subproblems. Optim. Meth. Softw. (2012, in press)
7. Bongartz, I., Conn, A.R., Gould, N., Toint, P.L.: CUTE: constrained and unconstrained testing environment. ACM Trans. Math. Softw. **21**(1), 123–160 (1995)
8. Conn, A.R., Gould, N., Sartenaer, A., Toint, P.L.: 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)
9. Conn, A.R., Gould, N.I.M., Toint, P.L.: A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. SIAM J. Numer. Anal. **28**(2), 545–572 (1991)
10. Conn, A.R., Gould, N.I.M., Toint, P.L.: LANCELOT: A Fortran package for Large-Scale Nonlinear Optimization (Release A). Springer, Berlin (1992)
11. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**(2), 201–213 (2002)

12. Eckstein, J.: A practical general approximation criterion for methods of multipliers based on Bregman distances. Math. Program. **96**(1), 61–86 (2003)
13. Eckstein, J., Silva, P.J.S.: Proximal methods for nonlinear programming: double regularization and inexact subproblems. Comput. Optim. Appl. **46**(2), 279–304 (2010)
14. Fernández, D., Solodov, M.V.: Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. Technical Report A677, Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro (2011)
15. Friedlander, M.P., Saunders, M.A.: A globally convergent linearly constrained Lagrangian method for nonlinear optimization. SIAM J. Optim. **15**(3), 863–897 (2005)
16. Hager, W.W., Zhang, H.: ASA-CG source code. http://www.math.ufl.edu/~hager/papers/CG/
17. Hager, W.W., Zhang, H.: A new conjugate gradient method with guaranteed descent and an efficient line search. SIAM J. Optim. **16**(1), 170–192 (2005)
18. Hager, W.W., Zhang, H.: A new active set algorithm for box constrained optimization. SIAM J. Optim. **17**(2), 526–557 (2006)
19. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python. http://www.scipy.org/ (2001)
20. Korpelevich, G.M.: Extrapolation gradient methods and their relation to modified Lagrange functions. Èkonom. i Mat. Metody **19**(4), 694–703 (1983)
21. Rockafellar, R.T.: Local boundedness of nonlinear, monotone operators. Michigan Math. J. **16**, 397–407 (1969)
22. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ (1970)
23. Rockafellar, R.T.: On the maximality of sums of nonlinear monotone operators. Trans. Am. Math. Soc. **149**, 75–88 (1970)
24. Rockafellar, R.T.: Conjugate Duality and Optimization. SIAM, Philadelphia (1974)
25. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
26. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
27. Solodov, M.V., Svaiter, B.F.: A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. Set-Valued Anal. **7**(4), 323–345 (1999)
28. Solodov, M.V., Svaiter, B.F.: A hybrid projection-proximal point algorithm. J. Convex Anal. **6**(1), 59–70 (1999)
29. Solodov, M.V., Svaiter, B.F.: 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)
30. Spingarn, J.E.: Partial inverse of a monotone operator. Appl. Math. Optim. **10**(3), 247–265 (1983)
31. van Rossum, G., et al.: Python language website. http://www.python.org/