SIAM J. OPTIM.
Vol. 13, No. 1, pp. 117–156


# GENERALIZED BUNDLE METHODS*

ANTONIO FRANGIONI†

**Abstract.** We study a class of generalized bundle methods for which the stabilizing term can be any closed convex function satisfying certain properties. This setting covers several algorithms from the literature that have been so far regarded as distinct. Under a different hypothesis on the stabilizing term and/or the function to be minimized, we prove finite termination, asymptotic convergence, and finite convergence to an optimal point, with or without limits on the number of serious steps and/or requiring the proximal parameter to go to infinity. The convergence proofs leave a high degree of freedom in the crucial implementative features of the algorithm, i.e., the management of the bundle of subgradients ($\beta$-strategy) and of the proximal parameter ($t$-strategy). We extensively exploit a dual view of bundle methods, which are shown to be a dual ascent approach to one nonlinear problem in an appropriate dual space, where nonlinear subproblems are approximately solved at each step with an inner linearization approach. This allows us to precisely characterize the changes in the subproblems during the serious steps, since the dual problem is not tied to the local concept of $\varepsilon$-subdifferential. For some of the proofs, a generalization of inf-compactness, called $^*$-compactness, is required; this concept is related to that of asymptotically well-behaved functions.

**Key words.** nondifferentiable optimization, bundle methods

**AMS subject classifications.** 90C25

**PII.** S1052623498342186

**Introduction.** We are concerned with the numerical solution of the *primal problem*

$$(\Pi) \qquad \inf_x \{f(x) : x \in X\}, \tag{0.1}$$

where $f : \Re^n \to \Re$ is finite-valued and convex (hence continuous) and $X \subseteq \Re^n$ is closed convex. Here $f$ is only known through an oracle ("black box") that, given any $x \in X$, returns the values $f(x)$ and $z \in \partial f(x)$. To simplify the treatment, we will assume $X = \Re^n$ until section 8, where the extension to the constrained case is studied.

We study a class of generalized bundle methods for the solution of (0.1), where a stabilizing term, which can be any closed convex function satisfying certain weak conditions, is added to (a model of) $f$. These methods sample $f$ in a sequence of *tentative points* $\{x_i\}$ to gather the $f$-values $\{f(x_i)\}$ and the *bundle* of first-order information $\beta = \{z_i \in \partial f(x_i)\}$. A distinguished vector $\bar{x}$ is taken as the *current point*, and $\beta$ is used to compute a *tentative descent direction* $d^*$ along which the next tentative point is generated. After a "successful" step, the current point can be updated; otherwise, the new information is used to enhance $\beta$, hopefully obtaining a better direction at the next iteration.

Several bundle methods proposed in the literature follow this pattern; some of them can be shown to actually belong to our class. Also, our generalized bundle methods provide implementable forms for some penalty-based algorithms for structured convex optimization. All of these algorithms have been analyzed either from the above primal viewpoint—the minimization of $f$—or from an application-specific

*Received by the editors July 20, 1998; accepted for publication (in revised form) December 6, 2001; published electronically, June 5, 2002.
http://www.siam.org/journals/siopt/13-1/34218.html
†Department of Computer Science, University of Pisa, Corso Italia 40, 56125 Pisa, Italy (frangio@di.unipi.it).

dual viewpoint, when $f$ itself is a dual function. A dual analysis of some general bundle methods exists—indeed, it motivated the development of the very first bundle methods—but it is related to the "local" concept of $\varepsilon$-subdifferential, and it does not easily extend to a wider class of methods. Instead, we extensively exploit a dual view of (0.1), where bundle methods are shown to be a penalty function approach to a "global" dual problem with approximate solution, via an inner linearization approach, of the penalized problem. The algorithms can be entirely described in terms of this dual problem; this is interesting for applications and helps in the convergence proofs.

We analyze in detail the features that are relevant for practical implementations, such as the management of the bundle ($\beta$-strategy) and of the proximal parameter ($t$-strategy). General rules are given which ensure convergence while leaving a large degree of freedom in practical implementations. For some variants of the algorithm, we require $f$ to be $^*$-*compact*, an assumption properly generalizing inf-compactness. $^*$-compact functions are *asymptotically well behaved* [Au97], but our definition seems to be better suited for the case of bundle methods.

The structure of this paper is the following: section 1 is devoted to the derivation of the dual viewpoint of generalized bundle methods. Some useful properties of pairs of primal and dual solutions of the stabilized master problems are proved in section 2. In section 3, the conditions on the stabilizing term are presented and discussed. Section 4 is devoted to the description of the algorithms and to the discussion of the rules for the $\beta$-strategy and the $t$-strategy. Convergence proofs of several variants of the algorithm are given next: section 5 is dedicated to convergence of the null step sequences, section 6 is dedicated to convergence of the serious step sequences, and section 7 is dedicated to the "third level" that is necessary for some classes of stabilizing terms. Some extensions of generalized bundle methods, e.g., to constrained optimization, are discussed in section 8, the relationships with other algorithms from the literature are analyzed in section 9, and conclusions are drawn in section 10.

Throughout the paper the following notation is used. The scalar product between two vectors $v$ and $w$ is denoted by $vw$. $\|v\|_p$ stands for the $L_p$ norm of the vector $v$, and the ball around 0 of radius $\delta$ in the $L_p$ norm will be denoted by $B_p(\delta)$. Given a set $X$, $I_X(x) = 0$ if $x \in X$ (and $+\infty$ otherwise) is its *indicator function*, $\sigma_X(z) = \sup_x\{zx : x \in X\}$ is its *support function*, and $d_X(y) = \inf_x\{\|x - y\| : x \in X\}$ is the *distance* from $y$ to $X$. Given a function $f$, $\partial_\varepsilon f(x)$ is its *$\varepsilon$-subdifferential* at $x$, $epi\, f = \{(v, x) : v \geq f(x)\}$ is its *epigraph*, $dom\ f = \{x : f(x) < \infty\}$ is its *domain*, and $S_\delta(f) = \{x : f(x) \leq \delta\}$ is its *level set* corresponding to the $f$-value $\delta$. Given a problem

$$\text{(P)} \qquad \inf[\sup]_x\{f(x) : x \in X\},$$

$v$(P) denotes the optimal value of $f$ over $X$; as usual, $X = \emptyset \Rightarrow v(\text{P}) = +\infty[-\infty]$.

**1. Duality for generalized bundle methods.** The dual description of generalized bundle methods relies on a well-established tool from convex analysis, the *conjugate* of $f$ (see [HL93b, Chapter X]):

$$f^*(z) = \sup_x\{zx - f(x)\}. \tag{1.1}$$

$f^*$ is a closed convex function and enjoys several properties; those useful in the paper are briefly recalled below.

(1.i) $(f^*)^* = f$ (duality of the conjugate operator),

(1.ii) $f_1 \le f_2 \Rightarrow f_1^* \ge f_2^*$ ("monotonicity" of the conjugate operator),
(1.iii) $(f(\cdot + x))^*(z) = f^*(z) - zx \quad \forall z, x$ (effect of a simple variable change),
(1.iv) $z \in \partial_\varepsilon f(x) \Leftrightarrow x \in \partial_\varepsilon f^*(z)$ (duality of the subdifferential mappings),
(1.v) $z \in \partial_\varepsilon f(x) \Leftrightarrow f(x) + f^*(z) \le zx + \varepsilon$ (characterization of the $\varepsilon$-subdifferentials),
(1.vi) $zx = f(x) + f^*(z) \Leftrightarrow z \in \partial f(x)$ (basic relation between the function values),
(1.vii) $zx \le f(x) + f^*(z) \quad \forall z, x$ (Fenchel's inequality).

A fundamental property of $f^*$ is that it characterizes all the affine functions supporting *epi* $f$ as

$$zx - \varepsilon \le f(x) \quad \forall x \Leftrightarrow \sup_x \{zx - f(x)\} = f^*(z) \le \varepsilon.$$

Note that, when the oracle is called at some point $x$ returning $f(x)$ and $z \in \partial f(x)$, $f^*(z)$ can be calculated via (1.vi); that is, the $f^*$-values are available if the $f$-values are, and vice-versa.

We remark that the above properties hold for any closed convex function; in the following, we will often take the conjugate of other functions apart from $f$, most notably of the "stabilizing term" to be introduced shortly.

**1.1. The dual problem.** Since $f^*$ is related with the minimization of $f$ by

$$v(\Pi) = \inf_x \{f(x)\} = -\sup_x \{0x - f(x)\} = -f^*(0),$$

we propose the following (apparently weird) *dual problem* as the dual of (0.1):

$$(\Delta) \qquad \inf_z \{f^*(z) : z = 0\}. \tag{1.2}$$

Problem (1.2) is a reasonable dual, since $v(\Pi) = -v(\Delta)$ and it deals with dual objects: every vector $z$ that is a subgradient of $f$ at some point belongs to *dom* $f^*$ (cf. (1.v)). Furthermore, consider the *Lagrangian relaxation* of (1.2) w.r.t. the constraints $z = 0$, using $\bar{x}$ as Lagrangian multipliers:

$$(\Delta_{\bar{x}}) \qquad \inf_z \{f^*(z) - z\bar{x}\}. \tag{1.3}$$

From (1.1) and (1.i), one has

$$-v(\Delta_{\bar{x}}) = \sup_z \{z\bar{x} - f^*(z)\} = (f^*)^*(\bar{x}) = f(\bar{x});$$

therefore, the *dual pricing problem* (1.3) can be seen as the problem that the oracle has to solve for computing $f(\bar{x})$. From the dual viewpoint, the oracle inputs $\bar{x}$ and returns a contact point $(f^*(z), z)$ between *epi* $f^*$ and the affine function with slope $(1, -\bar{x})$ that supports the set. This notation reveals that (0.1) itself is the Lagrangian dual of (1.2) w.r.t. the constraints $z = 0$.

**1.2. Approximations of $f$ and bundle algorithms.** Our aim is the construction of an algorithm that solves (0.1)—or, equivalently, (1.2)—given the oracle for $f$. A number of bundle algorithms have been proposed for this task, all based on the idea of using the bundle $\beta$ for constructing a *model* $f_\beta$ of the original function $f$. The

model is usually required to be a lower approximation of the function, i.e., $f_\beta \leq f$, so that the *primal master problem*

$$(\Pi_{\beta,\bar{x}}) \qquad \inf_d \{f_\beta(\bar{x}+d)\} \tag{1.4}$$

gives a lower bound on the primal problem (0.1). The optimal solution $d^*$ of (1.4) is then used as a (tentative) descent direction, analogously to what is done in Newton methods. From the dual viewpoint, $f^*_\beta \geq f^*$ (cf. (1.ii)) implies that the *dual master problem*

$$(\Delta_{\beta,\bar{x}}) \qquad \inf_z \{f^*_\beta(z) - z\bar{x} : z = 0\} \tag{1.5}$$

is an upper approximation of the dual problem (1.2).

The most popular model of $f$ is the *cutting plane model*

$$\hat{f}_\beta(x) = \max_z \{zx - f^*(z) : z \in \beta\}, \text{ for which} \tag{1.6}$$

$$\hat{f}^*_\beta(z) = \inf_\theta \left\{ \sum_{w\in\beta} f^*(w)\theta_w : \sum_{w\in\beta} w\theta_w = z, \quad \theta \in \Theta \right\}, \tag{1.7}$$

where $\Theta = \{\sum_{w\in\beta} \theta_w = 1, \theta \geq 0\}$ is the unitary simplex [HL93b, Proposition X.3.4.1]; note that $dom\ \hat{f}^*_\beta = conv(\beta)$. Using $\hat{f}_\beta$ in (1.4) gives the well-known cutting plane algorithm [HL93b, Algorithm XII.4.2.1], where the unknown $f$ is replaced with its known polyhedral outer approximation $\hat{f}_\beta$. In the corresponding (1.5), the unknown $f^*$ is replaced with its known polyhedral inner approximation $\hat{f}^*_\beta$ (a "pin-function").

**1.3. Stabilized master problems.** The cutting plane algorithm has some serious drawbacks, both in theory and in practice. First of all, the primal master problem (1.4) may be unbounded, that is, the dual master problem (1.5) may be infeasible; this is usually the case in the first iterations. Furthermore, two subsequent tentative points can be arbitrarily far apart; this is known as the "instability" of the cutting plane method. Most bundle methods try to alleviate this problem by introducing some "stabilizing device" into (1.4). Here, the *stabilizing term* $D_t$—a closed convex function—is added to $f_\beta$ to discourage points "far away" from $\bar{x}$, where $t > 0$ is the *proximal parameter* dictating the "strength" of $D_t$. That is, at each step the *stabilized primal master problem*

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \{f_\beta(\bar{x}+d) + D_t(d)\} \tag{1.8}$$

is solved, and its optimal solution $d^*$ is used as a (tentative) descent direction. By *Fenchel's duality* [HL93b, section XII.5.4], the dual of (1.8) is (using (1.iii)) the *stabilized dual master problem*

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_z \{f^*_\beta(z) - z\bar{x} + D^*_t(-z)\}. \tag{1.9}$$

Under proper assumptions (cf. Lemma 2.1 below), $v(\Delta_{\beta,\bar{x},t}) = -v(\Pi_{\beta,\bar{x},t})$. We see that the *primal stabilizing term* $D_t$ corresponds to a *dual penalty function* $D^*_t$ associated with the constraints $z = 0$; (1.9) is a (generalized) *augmented Lagrangian* of (1.5). The stabilizing term is a member of a family of functions parameterized in $t$;

in the bundle methods proposed so far, $t$ is either a factor, like in $D_t = \frac{1}{2t}\|\cdot\|_2^2$, or the radius of a ball, like in $D_t = I_{B_\infty(t)}$. In general, we will not require the function $t \to D_t(d)$ to have any specific form.

Note that $f$-values or $f^*$-values must be stored in memory together with the subgradients; due to (1.vi) the two choices are equivalent. In the standard notation of bundle methods, for $z \in \partial f(x)$ the *linearization error* (cf. [HL93b, Definition XI.4.2.3])

$$\alpha = f^*(z) - z\bar{x} + f(\bar{x}) = f(\bar{x}) - f(x) - z(\bar{x} - x) \geq 0 \tag{1.10}$$

of $z$ w.r.t. $\bar{x}$ is typically used in place of $f^*(z)$. This notation corresponds to defining the *translated function* $f_{\bar{x}}(d) = f(\bar{x} + d) - f(\bar{x})$ and its *translated model* $f_{\bar{x},\beta}$, and to considering a "local" form of (1.8) that uses $f_{\bar{x},\beta}$ [Fr98]. However, the corresponding dual problem is written in terms of $f^*_{\bar{x}}$, i.e., of a family of functions changing with $\bar{x}$, rather than in terms of the unique $f^*$. Furthermore, the notation based on linearization errors hides the dependency of some of the subproblem's data on the current point $\bar{x}$; that is why we use $f^*$-values.

**1.4. Stabilization in the original problems.** The above duality argument can also be applied to the original function $f$; the stabilized dual problem

$$(\Delta_{\bar{x},t}) \qquad \inf_z \{f^*(z) - z\bar{x} + D_t^*(-z)\} \tag{1.11}$$

is the (Fenchel) dual of the *stabilized primal problem*

$$(\Pi_{\bar{x},t}) \qquad \phi_t(\bar{x}) = \inf_d \{f(\bar{x} + d) + D_t(d)\}. \tag{1.12}$$

A primal analysis of generalized bundle methods would focus on (1.12), that is, the calculation of the *generalized Moreau–Yosida regularization* $\phi_t$ of $f$ in $\bar{x}$. With a proper $D_t$ [BPP91], $\phi_t$ has the same set of minima as $f$ but enjoys additional properties, e.g., smoothness; hence, minimizing $\phi_t$ could be an advantageous alternative to the minimization of $f$. Unfortunately, solving (1.12) with the sole help of the black box for $f$ is as difficult as solving (0.1); therefore, bundle methods resort to a two-level approach, repeatedly solving the approximation (1.8) until the accumulation of information in $\beta$ makes $f_\beta$ a "good enough" approximation of $f$, and only then changing $\bar{x}$. If $t$ is properly managed, the whole process eventually solves (0.1).

But a dual analysis of generalized bundle methods is also possible, which focuses instead on (1.2) and its *generalized augmented Lagrangian* (1.11), where the constraints $z = 0$ are replaced with the linear term $-\bar{x}z$ (with Lagrangian multipliers $\bar{x}$) and the nonlinear term $D_t^*(-z)$ in the objective function. A classical ascent method would require repeatedly solving (1.11) and updating $\bar{x}$ using the corresponding first-order information; unfortunately, solving (1.11)—which is equivalent to (1.12)—is difficult. On the contrary, (1.9) may be efficiently solvable; furthermore, the oracle for $f$ solves (1.3), and hence $v(\Delta_{\bar{x}+d})$ gives a lower bound on (1.11) if $-zd$ is a linear lower approximation of $D_t^*(-z)$. Hence, a viable approach is again a two-level one, where in the inner level a sequence of (1.9) and (1.3) is solved for fixed $\bar{x}$ in order to approximate (1.11), while in the outer level the Lagrangian multipliers $\bar{x}$ and the parameter $t$, dictating the "strength" of the penalty function, are updated.

This dual interpretation of bundle methods is related to—although independently obtained from—the general dual algorithmic scheme of [ACC93]; by taking their "perturbation function" $\varphi(x, \bar{x})$ as $f(x - \bar{x})$, the Lagrangian dual of (1.3), i.e., (1.2), is

obtained. However, in our case the relevant dual object is simply the conjugate $f^*$, and the whole process takes place in the graph space of $f^*$. This is confirmed by [Nu97], where a step in the same direction has been made using the graph of the $\varepsilon \to \partial_\varepsilon f(0)$ multifunction that is equivalent to *epi* $f^*$ (cf. section 9.1).

**2. Properties of subproblem solutions.** The following two lemmas will be useful in the analysis of the algorithm.

LEMMA 2.1. *Let $f_\beta$ and $D_t$ be two closed convex functions such that dom $f_\beta(\bar{x} + \cdot) \cap$ int dom $D_t \neq \emptyset$, and assume that* (1.8) *and of* (1.9) *have optimal solutions $d^*$ and $z^*$, respectively; then*

$$v(\Delta_{\beta,\bar{x},t}) = -v(\Pi_{\beta,\bar{x},t}), \tag{2.1}$$

$$-z^* \in \partial D_t(d^*) \qquad \text{and} \qquad d^* \in \partial D_t^*(-z^*), \tag{2.2}$$

$$z^* \in \partial f_\beta(\bar{x} + d^*) \qquad \text{and} \qquad \bar{x} + d^* \in \partial f_\beta^*(z^*), \tag{2.3}$$

$$f_\beta(\bar{x} + d^*) + f_\beta^*(z^*) = z^*(\bar{x} + d^*), \tag{2.4}$$

$$D_t(d^*) + D_t^*(-z^*) = -z^* d^*. \tag{2.5}$$

*Proof.* Equation (2.1) is [HL93b, (X.2.3.2)]. Apply [HL93b, Proposition XII.5.4.1] with the nonsymmetric assumption [HL93b, (X.2.3.Q.jj′)] to the pair (1.8)–(1.9) to show that any optimal solution $d^*$ of (1.8) belongs to $\partial[f_\beta(\bar{x} + \cdot)]^*(z^*) \cap \partial D_t^*(-z^*)$; this gives $d^* \in \partial D_t^*(-z^*)$ and, via (1.iii), $\bar{x} + d^* \in \partial f_\beta^*(z^*)$. For the rest, apply (1.iv) and (1.vi). ∎

We remark that Lemma 2.1 works for any closed convex function $f_\beta$, even if it is not a model of $f$. We will always keep the requirement on $f_\beta$ to the bare minimum, in the spirit of [CL93]; this will provide more general results, and it will be useful in section 8 where extensions of the method are discussed. Also, note that Lemma 2.1 with $f_\beta = f$ characterizes the properties of the solutions $d^*$ and $z^*$ of the primal and dual stabilized problems (1.12) and (1.11). When $f_\beta \leq f (\Rightarrow f_\beta^* \geq f^*$ by (1.ii)), the optimal solutions of the master problems allow us to derive information on those of the original problems.

LEMMA 2.2. *If $f_\beta \leq f$ and the hypothesis of Lemma* 2.1 *hold, then the optimal value of* (1.11) *can be bracketed using* (1.9) *and*

$$\Delta f = f(\bar{x} + d^*) - f_\beta(\bar{x} + d^*) \geq 0, \tag{2.6}$$

$$\textit{i.e.,} \qquad v(\Delta_{\bar{x},\beta,t}) - \Delta f \leq v(\Delta_{\bar{x},t}) \leq v(\Delta_{\bar{x},\beta,t}). \tag{2.7}$$

*Proof.* $v(\Delta_{\bar{x},t}) \leq v(\Delta_{\beta,\bar{x},t})$ comes from $f_\beta^* \geq f^*$. From (2.2),

$$D_t^*(-z) \geq D_t^*(-z^*) - d^*(z - z^*) \quad \forall z.$$

Add $f^*(z) - z\bar{x}$ to both sides, then add and remove $f_\beta^*(z^*) - z^*\bar{x}$ to the right-hand side to obtain

$$f^*(z) - z\bar{x} + D_t^*(-z) \geq v(\Delta_{\beta,\bar{x},t}) - [f_\beta^*(z^*) - f^*(z) + (\bar{x} + d^*)(z - z^*)] \quad \forall z.$$

Take the inf on $z$ on both sides and recognize the stabilized dual problem (1.11) on the left and the dual pricing problem (1.3) at $\bar{x} + d^*$ plus $f_\beta(\bar{x} + d^*)$ (via (2.4)) on the right. ∎

For future reference, let us record here the alternative formula

$$\Delta f = f_\beta^*(z^*) - f^*(z) + (\bar{x} + d^*)(z - z^*), \tag{2.8}$$

where $z \in \partial f(\bar{x} + d^*)$. ($z$ is an optimal solution of (1.3) at $\bar{x} + d^*$.)

Let us briefly comment on the above lemmas. Equation (2.3) shows that the dual optimal solution $z^*$ gives, in primal terms, a linear lower approximation of the model $f_\beta$ which, by (2.4), is tight in $\bar{x} + d^*$. Conversely, by (2.2) the primal direction $d^*$ gives, in dual terms, a subgradient of $D_t^*$ at $-z^*$. Lemma 2.2 shows that the gap between the model and the original function in $\bar{x}+d^*$ is a measure of the gap between (1.9) and (1.11); thus, if $\Delta f = 0$, then $z^*$ is optimal for (1.11) ($f_\beta^*(z^*) = f^*(z^*)$), and $d^*$ is optimal for (1.12).

If $f_\beta^* \geq f^*$, a useful object in the analysis of the algorithms is

$$\alpha^* = f_\beta^*(z^*) - z^*\bar{x} + f(\bar{x}) \geq 0 \tag{2.9}$$

(use (1.vii)); using (1.v) in (2.9), one obtains

$$z^* \in \partial_{(\alpha^*)} f(\bar{x}). \tag{2.10}$$

Note that all of the above relations are independent of the choice of $f_\beta$ and $D_t$; in the literature, analogous results have usually been obtained algebraically for specific choices, such as $D_t = \frac{1}{2t}\|\cdot\|_2^2$ and $f_\beta = \hat{f}_\beta$. However, not all the results for particular cases generalize; a relevant example is $d^* = -tz^*$, which is central in the analysis of proximal bundle methods but it is not true in general.

**3. Conditions on $D_t$.** Of course, the primal stabilizing term $D_t$ has to satisfy some conditions. First of all, in order to be able to apply the results of the previous paragraph, $D_t$ has to be a closed convex function $\forall t > 0$. Then, a set of weak properties that suffice for constructing a convergent algorithm is the following:

- (P1) $\forall t > 0$, $D_t(0) = 0$ and $0 \in \partial D_t(0)$ ($D_t$ is *nonnegative*).
- (P2) $\forall t > 0$ and $\varepsilon > 0$, $S_\varepsilon(D_t)$ is *compact* and $0 \in$ int $S_\varepsilon(D_t)$ ($S_\varepsilon(D_t)$ is *full-dimensional*).
- (P3) $\forall t > 0$, $\lim_{\|d\|\to\infty} D_t(d)/\|d\| = +\infty$ ($D_t$ is *strongly coercive*).
- (P4) $\forall t > 0$, $D_t \geq D_\tau$ for each $\tau \geq t$ ($D_t$ is *nonincreasing* in $t$).
- (P5) $\lim_{t\to\infty} D_t(d) = 0$ $\forall d$ ($\{D_t\}$ *converges pointwise* to the constant zero function).

We will show that the above conditions on $D_t$ are equivalent to the following conditions on $D_t^*$:

- (P*1) $\forall t > 0$, $D_t^*(0) = 0$ and $0 \in \partial D_t^*(0)$ ($D_t^*$ is *nonnegative*).
- (P*2) $\forall t > 0$ and $\varepsilon > 0$, $S_\varepsilon(D_t^*)$ is *compact* and $0 \in$ int $S_\varepsilon(D_t^*)$ ($S_\varepsilon(D_t^*)$ is *full-dimensional*).
- (P*3) $\forall t > 0$, $D_t^*$ is *finite everywhere*.
- (P*4) $\forall t > 0$, $D_t^* \leq D_\tau^*$ for each $\tau \geq t$ ($D_t^*$ is *nondecreasing* in $t$).
- (P*5) $\forall \varepsilon > 0$, $\lim_{t\to\infty} \inf_z\{D_t^*(z) : \|z\| \geq \varepsilon\} = +\infty$ ($\{D_t^*\}$ *converges* "*uniformly*" to $I_{\{0\}}$).

The following remarks about (P1)–(P5) are useful:

- Having a minimum in 0 where they evaluate to 0, both $D_t$ and $D_t^*$ are nonnegative functions $\forall t > 0$.
- As a consequence of (P1) and (P*1), $D_t$ and $D_t^*$ are *radially nondecreasing*, i.e.,

$$\forall \alpha \geq 1 \qquad D_t(\alpha d) \geq D_t(d) \quad \forall d \qquad \text{and} \qquad D_t^*(\alpha z) \geq D_t^*(z) \quad \forall z, \tag{3.1}$$

since, e.g., $d = (1/\alpha)\alpha d + (1-1/\alpha)0$ and, by convexity, $D_t(d) \le (1/\alpha)D_t(\alpha d) + (1-1/\alpha)D_t(0) \le D_t(\alpha d)$ as $\alpha \ge 1$ and $D_t(\alpha d) \ge 0$.

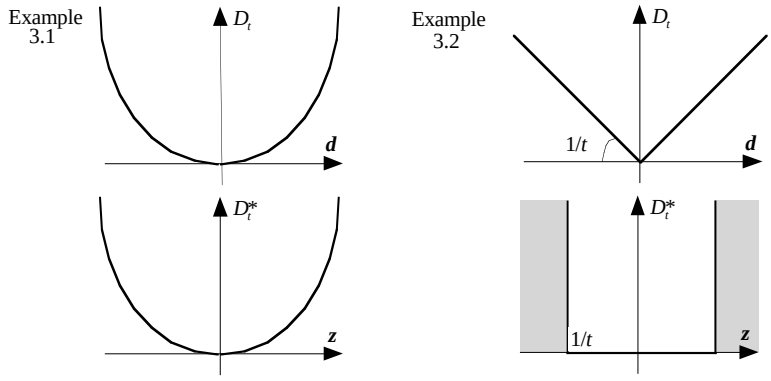
– Another consequence of (P1) and (1.v) is

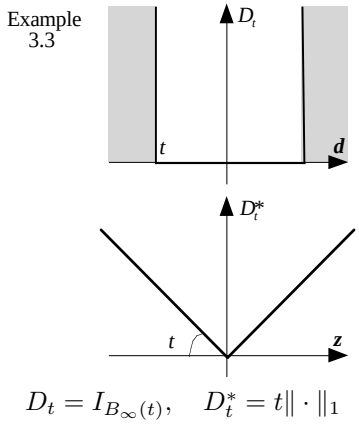
$$S_\varepsilon(D_t) = \partial_\varepsilon D_t^*(0) \qquad \text{and} \qquad S_\varepsilon(D_t^*) = \partial_\varepsilon D_t(0); \tag{3.2}$$

a rephrasing of (P2) is therefore that *both* the level sets of $D_t$ *and* its $\varepsilon$-subdifferentials at 0 must be compact, and the same holds for $D_t^*$.

– (P2) guarantees that the hypothesis of Lemma 2.1 is true, as $0 \in \text{int } dom\ D_t$ and $0 \in dom\ f_\beta(\bar{x} + \cdot)$. (This is true even in the constrained case, cf. section 8.1, assuming of course that $\bar{x} \in X$.)

– (P2) and (P*2) are stated for $\varepsilon > 0$: $S_0(D_t)$ and $S_0(D_t^*)$ may or may not be full-dimensional, as in the following examples.




– It is intuitive why (P2) and (P*2) are necessary. The noncompactness of $S_\varepsilon(D_t)$ for some $\varepsilon > 0$ means that $D_t$ is constantly 0 along some direction $d$ and therefore cannot "stabilize" $f$ along $d$. In fact, all the nonempty level sets of a closed convex function have the same *asymptotic cone* [HL93a, Proposition IV.3.2.5], so that $S_0(D_t)$ is also noncompact. On the other hand, if 0 belongs to the frontier of $dom\ D_t$, then some $d$ is a "forbidden" direction, i.e., $D_t(\alpha d) = +\infty\ \forall \alpha > 0$.

– Strongly coercive (or 1-coercive) functions increase faster than any linear

function at infinity; (P3) guarantees that (1.8) has a bounded nonempty set of optimal solutions.
- Concerning (P4) and (P*4), note the role of $t$ in Examples 3.1–3.3 above.
- The need for (P4) and (P5) is also intuitively clear: $t$ must make $D_t$ "weaker" as it grows, and it must be possible to make $D_t$ as weak as desired in order to avoid "blocking" promising directions. Dually, a penalty term must increase as the penalty parameter does (see (P*4)), and it must be equivalent to the constraints it replaced, at least in the limit (see (P*5)).

$D_t$ need not be "norm-like" [KCL95, Be96] or a *Bregman distance* [CT93]; in particular, it is not necessary that $D_t(0) = 0 \Leftrightarrow d = 0$ [IST94, Ki99]. Also, $t \to D_t(d)$ need not have the $1/t$ form.

THEOREM 3.1. (P1)–(P5) *are equivalent to* (P*1)–(P*5).

*Proof.* For the first four properties, the equivalence is pairwise.

1. The equivalence between (P1) and (P*1) is an easy consequence of (1.iv) and (1.vi).
2. The equivalence between (P2) and (P*2) can be obtained as a consequence of the following little-known result: for any proper convex function $D$, $\bar{d} \in \text{int } dom\ D \Leftrightarrow \bar{d} \in \text{int } S_\delta(D)\ \forall \delta > D(\bar{d})$. One of the implications is obvious; for the other, $\bar{d} \in \text{int } dom\ D$ means that there exists a ball $B(\bar{d}, \varepsilon)$ with $\varepsilon > 0$ such that $B(\bar{d}, \varepsilon) \subseteq \text{int } dom\ D$. By [HL93a, Theorem IV.3.1.2], $D$ is Lipschitz over the ball, i.e., $|D(d) - D(\bar{d})| \le L\|d - \bar{d}\|\ \forall d \in B(\bar{d}, \varepsilon)$ for some constant $L > 0$; hence, $S_\delta(D) \supseteq B(\bar{d}, \min\{\varepsilon, (\delta - D(\bar{d}))/L\})$ as desired.
   Using this result, [HL93b, Theorem XI.1.1.4], and (3.2), one has
   $$0 \in \text{int} S_\varepsilon(D_t) \Leftrightarrow 0 \in \text{int } dom\ D_t \Leftrightarrow \partial_\varepsilon D_t(0) \text{ compact} \Leftrightarrow S_\varepsilon(D_t^*) \text{ compact}.$$
   To complete the proof of the equivalence, simply exchange $D_t$ with $D_t^*$.
3. The equivalence between (P3) and (P*3) is [HL93b, Remark X.1.3.10].
4. The equivalence between (P4) and (P*4) is (1.ii).
5. For the last step, we will show that [(P1) + (P4) + (P5)] ⇒ (P*5) and [(P*1) + (P*4) + (P*5)] ⇒ (P5).

**[(P1) + (P4) + (P5)] ⇒ (P*5)**. Due to (3.2) (which requires (P1) ≡ (P*1)), (P*5) can be rewritten as

$$\forall \varepsilon > 0 \lim_{t \to \infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon\} = +\infty.$$

Now, assume by contradiction that $\varepsilon > 0$ exists such that the limit is not $+\infty$; since the feasible set is compact and $D_t^*$ is closed, for each $t$ there exists a $z_t$ achieving the inf, and we can write

$$\lim_{t \to \infty} D_t^*(z_t) \le M < +\infty.$$

From (1.vii) $\forall t\ \forall d\ \forall z,\ D_t(d) + D_t^*(z) \ge zd$; choosing $z = z_t$ and using $D_t^*(z_t) \le M$, one obtains

$$\forall t\ \forall d \qquad D_t(d) \ge z_t d - M.$$

But all the $z_t$ belong to a compact set, and therefore some cluster point $z^*$ exists with $\|z^*\| = \varepsilon$; plugging $d^* = (2M/\varepsilon^2)z^*$ into the above inequality and taking the limit for $t \to \infty$, one gets

$$\lim_{t \to \infty} D_t(d^*) \ge \lim_{t \to \infty} \left( \frac{2M}{\varepsilon^2} \right) z_t z^* - M = 2M - M > 0,$$

which contradicts (P5).

**[(P*1) + (P*4) + (P*5)] ⇒ (P5)**. As a preliminary, we must show that for every $d$ there exists a sufficiently large $\bar{t}$ such that $D_t(d) < +\infty$ for each $t \geq \bar{t}$; due to (P4) ≡ (P*4), it is only necessary to show that this happens for at least one $t$. Assume by contradiction that one $\bar{d}$ exists such that $\bar{d} \notin dom\ D_t\ \forall t$. Using [HL93a, Theorem V.2.2.2], one has

$$\forall t\ \exists z_t : \|z_t\| = 1 \quad \sup_d \{z_t d : d \in dom\ D_t\} \leq z_t \bar{d}.$$

Now

$$\begin{aligned} D_t^*(z_t) &= \sup_d\{z_t d - D_t(d)\} = \sup_d\{z_t d - D_t(d) : d \in dom\ D_t\} \\ &\leq \sup_d\{z_t d : d \in dom\ D_t\} \leq z_t \bar{d} \qquad (D_t \geq 0). \end{aligned}$$

Using $\|z_t\| = 1$, this finally gives $\forall t,\ D_t^*(z_t) \leq \|\bar{d}\|_2$, which contradicts (P*5). Hence, each $d$ is in $dom\ D_t$ for a sufficiently large $t$.

We now want to prove that (P5) holds, so assume by contradiction that one $\bar{d}$ exists such that $D_t(\bar{d}) \geq \varepsilon > 0\ \forall t > 0$. (It must be $\bar{d} \neq 0$ due to (P1) ≡ (P*1), and note that we are using (P4) ≡ (P*4).) Since $D_t(\bar{d}) > D_t(0) = 0$, $0 \notin \partial D_t(\bar{d})$. In fact, from the subgradient inequality

$$D_t(d) \geq D_t(\bar{d}) + z(d - \bar{d}) \qquad \forall d\ \forall z \in \partial D_t(\bar{d})$$

one gets for $d = 0$, using (P1),

$$z\bar{d} \geq D_t(\bar{d}) \geq \varepsilon \quad \forall z \in \partial D_t(\bar{d}) \Rightarrow \|z\| \geq \varepsilon' = \varepsilon/\|\bar{d}\| \quad \forall z \in \partial D_t(\bar{d}).$$

Now, for each $t$ choose any $z_t \in \partial D_t(\bar{d})$. Using (1.vi), $\|z_t\| \geq \varepsilon'$, and (P1), we obtain

$$\lim_{t\to\infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon'\} \leq \liminf_{t\to\infty} D_t^*(z_t) \leq \liminf_{t\to\infty} z_t\bar{d} - D_t(\bar{d}) \leq \liminf_{t\to\infty} z_t\bar{d}.$$

There exists a large enough $\bar{t}$ such that $2d \in dom\ D_{\bar{t}}$; hence, by (3.1), $d \in dom\ D_{\bar{t}}$ also. Again using the subgradient inequality, (P4) (which is implied by (P*4)), and (P1), we have

$$\forall t \geq \bar{t} \qquad D_{\bar{t}}(2\bar{d}) \geq D_t(2\bar{d}) \geq D_t(\bar{d}) + z_t(2\bar{d} - \bar{d}) \geq z_t\bar{d},$$

which finally gives

$$\lim_{t\to\infty} \inf_z \{D_t^*(z) : \|z\| = \varepsilon'\} \leq \liminf_{t\to\infty} z_t\bar{d} \leq D_{\bar{t}}(2\bar{d}) < \infty,$$

contradicting (P*5) and therefore finishing the proof of the theorem. □

Condition (P*5) may be a bit clumsy to check. The following result gives a handy sufficient condition that should work in most cases.

THEOREM 3.2. *If* (P*4) *holds,* $\{D_t^*\}$ *converges pointwise to* $I_{\{0\}}$, *i.e.,*

$$\lim_{t\to\infty} D_t^*(z) = +\infty \quad \forall z \neq 0,$$

*and for any two sequences* $\{t_i\} \to +\infty$ *and* $\{z_i\} \to \bar{z}$, *where* $z_i \in dom\ D_{t_i}^*$, *one has*

$$\liminf_{t\to\infty} D_{t_i}^*(z_i) \geq \lim_{i\to\infty} D_{t_i}^*(\bar{z}),$$

*then* (P*5) *holds.*

*Proof.* The thesis is obvious if for any fixed $\varepsilon$ there exists a $t$ such that *dom* $D_t^* \subseteq B_2(\varepsilon)$, since by (P*4) the domain of $D_t^*$ can only shrink as $t$ increases. Hence, we can assume that *dom* $D_t^* \backslash B_2(\varepsilon)$ is nonempty $\forall t$. Assume by contradiction that for some $\varepsilon > 0$ and $\{t_i\} \to +\infty$ there exist one $\delta < \infty$ and a sequence $\{z_i\}$ of points outside $B_2(\varepsilon)$ such that $D_{t_i}^*(z_i) \leq \delta \ \forall i$. Let $\bar{z}_i = (\varepsilon/\|z_i\|_2)z_i$ (the projection of $z_i$ on $B_2(\varepsilon)$). By (3.1), $D_{t_i}^*(\bar{z}_i) \leq D_{t_i}^*(z_i)$. Now, $B_2(\varepsilon)$ is a compact set; hence we can assume $\{\bar{z}_i\} \to \bar{z}$ with $\|\bar{z}\|_2 = \varepsilon > 0$. Using the hypothesis,

$$\infty = \lim_{i\to\infty} D_{t_i}^*(\bar{z}) \leq \liminf_{i\to\infty} D_{t_i}^*(\bar{z}_i) \leq \delta < \infty. \qquad \square$$

All the $D_t^*$ proposed so far satisfy (P*5); they are either continuous in both $z$ and $t$ (cf. Examples 3.1, 3.3) or indicator functions of balls shrinking as $t$ increases (cf. Example 3.2). It is clear from the proof of Theorem 3.2 that these two possibilities—which in our setting can be mixed—have two distinct ways of ensuring that (P*5) holds. Bundle methods using these two different types of stabilizing term, i.e., *penalty* and *trust-region*, have so far been viewed as distinct [HL93b, sections XV.2.1 and XV.2.2].

It is possible to avoid the strong coercivity assumption (P3) (cf. Example 3.2), provided that other assumptions guarantee that (1.8) is bounded below.

(P3′) $f$ is *bounded below*, a finite $f_*$ such that $f_* \leq v(\Pi)$ is *known* and $f_\beta \geq f_* \ \forall \beta$.

(P*3′) (1.2) is *nonempty*, a finite $f_*$ such that $f^*(0) \leq -f_*$ is *known* and $f_\beta^*(0) \leq -f_* \ \forall \beta$.

Note that there are three separate conditions in (P3′): a suitable $f_*$ must *exist*, must *be known*, and the corresponding "flat" subgradient must be *explicitly kept* in the bundle. From the dual viewpoint, (P*3′) guarantees that 0 is a feasible solution for (1.9). A more general condition would be requiring $f_\beta$ to be always bounded below; with such a model, the cutting plane algorithm could be directly applied without stabilization. However, the constant zero function is not a valid stabilizing term, even if (P3) is not enforced, due to the first part of (P2) (compactness).

Two other variants of the above properties allow us to obtain stronger convergence results:

(P3″) $\forall t$ $D_t$ is *strongly coercive* and *strictly convex.*

(P*3″) $\forall t$ $D_t^*$ is *finite everywhere* and *differentiable.*

(P5′) $\forall t$ $\partial D_t(0) = \{0\}$ ($D_t$ is *differentiable* in 0, i.e., $\nabla D_t(0) = 0$).

(P*5′) $\forall t$ $S_0(D_t^*) = \{0\}$ ($D_t^*$ is *strictly convex* in 0, i.e., 0 is the *unique minimum* of $D_t^*$).

(P3″) is a strengthening of (P3) that allows us to keep the size of $\beta$ bounded. The equivalence between (P3″) and (P*3″) is [HL93b, Theorem X.4.1.1]. Under (P5′), 0 is a stationary point of $f(\bar{x} + \cdot) + D_t$ if and only if $\bar{x}$ is a stationary point of $f$; with (P5′) replacing (P5), it is possible to prove convergence without requiring $t \to \infty$. The equivalence between (P5′) and (P*5′) is a consequence of (3.2). (P5′) implies the second part of (P2) (full dimensionality); this is easily seen in the dual, as (P*5′) implies the first part of (P*2) (compactness), since all the level sets of $D_t^*$ share the same asymptotic cone of $S_0(D_t^*) = \{0\}$.

So far, nothing has been required about the form of the $t \to D_t(d)$ functions; in this very general setting, $D_t$ and $D_{t'}$ for $t \neq t'$ may be two almost completely unrelated functions. In some cases, stronger results can be obtained under the following (pretty

```
⟨ let μ ≥ 1 and ε ≥ 0 be fixed; choose the initial x̄, t, and β ⟩        // initialization
do
    ⟨ solve (Π_{β,x̄,t}) and (Δ_{β,x̄,t}) for d* and z*, respectively ⟩;   // find a direction
    ⟨ move along d*, generating some new z and a trial point x ⟩;        // probe f along d*
    if (a large enough improvement has been obtained)                    // NS/SS decision
        then x̄ = x;                                                     // a serious step
    ⟨ add some new z to β, delete some old z from β ⟩;                   // the β-strategy
    ⟨ update t, depending on the previous history ⟩;                     // the t-strategy
while (α* + μ D*_t(−z*) > ε);                                           // stopping condition
```

FIG. 1. *The "two-level" bundle algorithm.*

```
⟨ choose the initial x̄, t̲ > 0, ε > 0, and β ⟩;
do forever
    ⟨ run the algorithm of Figure 1, ensuring that t ≥ t̲
      and using ε in the stopping criteria ⟩
    ⟨ increase t̲ and decrease ε ⟩;
enddo
```

FIG. 2. *The "three-level" bundle algorithm.*

reasonable) assumptions:

$$D_t = \frac{1}{t} D \Rightarrow D_t^* = \frac{1}{t} D^*(t\cdot), \tag{3.3}$$

where $D$ satisfies (P2) and (P2) and is finite everywhere ($\Rightarrow$ (P5));

$$D_t^* = tD^* \Rightarrow D_t = tD\left(\frac{1}{t}\cdot\right), \tag{3.4}$$

where $D^*$ satisfies (P*1) and (P*2) and is strictly convex in 0 ($\Rightarrow$ (P*5) + (P*5′)).

Of course, conditions equivalent to (P3)/(P*3) ($D$ strongly coercive/$D^*$ finite everywhere) or (P3′)/(P*3′) will also be required, whereas (P4)/(P*4) come directly from the nonnegativity of $D/D^*$.

Finally, let us record for future use two useful consequences of (P1)–(P5), the second being just that a penalty method using $D_t$ works.

LEMMA 3.3. *$\forall \varepsilon > 0\ \forall \delta > 0$ there exists a $\underline{t}$ such that $S_\varepsilon(D_t) \supseteq B_2(\delta) \forall t \geq \underline{t}$.*

*Proof.* $\{D_t\}$ converges *uniformly* to $0(\cdot)$ on every compact set $C$, i.e., $\forall \varepsilon > 0$ there exists a $\underline{t}$ such that $D_t(d) \leq \varepsilon\ \forall d \in C, \forall t \geq \underline{t}$: use (P5), [HL93a, Theorem IV.3.1.5], and the fact that *ri dom* $0(\cdot) = \Re^n$. The result follows, using $C = B_2(\delta)$, since, due to (P4), $s_\varepsilon(D_t)$ are nondecreasing in $t$. □

LEMMA 3.4. *For any fixed $\bar{x}$, $\lim_{t\to\infty} v(\Pi_{\bar{x},t}) = v(\Pi)$.*

*Proof.* Note that $v(\Pi_{\bar{x},t})$ is nonincreasing in $t$ by (P4). Assume by contradiction $\lim_{t\to\infty} v(\Pi_{\bar{x},t}) = \underline{v} > v(\Pi)$, i.e., one $\bar{d}$ exists such that $f(\bar{x} + \bar{d}) < \underline{v}$: using (P5), we get

$$\underline{v} = \lim_{t\to\infty} v(\Pi_{\bar{x},t}) \leq \lim_{t\to\infty} [f(\bar{x} + d) + D_t(\bar{d})] = f(\bar{x} + \bar{d}) < \underline{v}. \qquad \square$$

**4. The bundle algorithm.** We will analyze two main variants of the generalized bundle algorithm, described, respectively, in Figures 1 and 2.

The "two-level" bundle algorithm of Figure 1 implements the standard ideas of a bundle approach: the generalized Moreau–Yosida regularization $\phi_t$ of $f$ (cf. section 1.4) is minimized (2nd level), with sequences of consecutive null steps performing

the approximate computation of $\phi_t(\bar{x})$ (1st level). The algorithm of Figure 2 adds another level, where $t$ is forced to increase, possibly to $+\infty$; this is useful for those cases in which, due to properties of $D_t$, the standard two-level approach is not able to guarantee convergence unless $t$ is "large enough."

In order to obtain a convergent algorithm, assumptions are needed about the following:

- properties of the stabilizing term $D_t$,
- choice of the model $f_\beta$,
- properties of the function $f$,
- handling of the $t$ parameter ($t$-strategy) and the NS/SS decision,
- handling of the bundle ($\beta$-strategy).

The required properties for $D_t$ have been described in the previous section. We will always assume $f_\beta$ to be a closed convex function such that $f_\beta \le f$; for some results, $f_\beta$ will be required to be the cutting plane model $\hat{f}_\beta$ (1.6). The assumptions on the last three points will be discussed in the following.

**4.1. Assumptions on $f$.** For some variants of the algorithm, we will require $f$ to be a $^*$-*compact* function, i.e., such that

$$e(l,L) := \sup_x\{d_{S_l(f)}(x) : x \in S_L(f)\} < \infty \qquad \forall L \ge l > v(\Pi) \ge -\infty.$$

Here $f$ is $^*$-compact if the excess of any level set $S_L(f)$ over $S_l(f)$ is finite; that is, $f$ never becomes "infinitely flat."

Let us now briefly present some properties of $^*$-compact functions that are useful in our treatment. (The interested reader is referred to [Fr98] for a more detailed study.) Recall that a nonempty closed convex set $C$ is compact if and only if its *asymptotic cone*

$$C_\infty = \{d : x + \alpha d \in C \ \forall x \in C, \forall \alpha \ge 0\}$$

is the set $\{0\}$ (see [HL93a, Proposition III.2.2.3]); all the nonempty level sets of a closed convex function $f$ have the same asymptotic cone (see [HL93a, Proposition IV.3.2.5]), denoted by $f_\infty$.

THEOREM 4.1. *If $\forall L > v(\Pi)$ there exists a compact set $C_L$ such that $S_L(f) \subseteq C_L + f_\infty$, then $f$ is $^*$-compact.*

*Proof.* Select $L \ge l > v(\Pi)$ and choose any $x_l \in S_l(f)$ (there must be at least one) to be kept fixed. From the hypothesis, for any $\bar{x} \in S_L(f)$ there exists an $x_L \in C_L$ and a $d \in f_\infty$ such that $\bar{x} = x_L + d$. Since $x_l + d \in S_l(f)$, we obtain

$$\inf_x\{\|x - \bar{x}\| : x \in S_l(f)\} \le \|(x_l + d) - (x_L + d)\| = \|x_l - x_L\|.$$

Therefore,

$$\sup_x \left\{\inf_x\{\|x - \bar{x}\| : x \in S_l(f)\} : \bar{x} \in S_L(f)\right\}$$
$$\le \sup_x\{\|x_l - x\| : x \in C_L\} < \infty, \qquad \text{since } C_L \text{ is compact.} \qquad \square$$

Note that $C_L$ is not required to be convex, just compact.

COROLLARY 4.2. *All polyhedral functions are $^*$-compact.*

*Proof.* The level sets of polyhedral functions are obviously polyhedra. Any polyhedron has a minimal representation as the sum of a (compact) polytope and a polyhedral cone [Ro70, Theorem 19.1]. The cone appearing in the minimal representation of each level set can only be $f_\infty$. $\square$

Note that the hypothesis of Theorem 4.1 is obviously true if $f_\infty = \{0\}$, i.e., all inf-compact functions are $^*$-compact. The converse is not true, however, since by Corollary 4.2 there are $^*$-compact functions that are not inf-compact; $^*$-compactness properly generalizes inf-compactness. It is easy to prove that many other functions are $^*$-compact, such as the quadratic ones. $^*$-compactness is a powerful assumption, since it allows us to prove the following result.

LEMMA 4.3. *If $f$ is $^*$-compact, then for any $\infty > L \geq l > v(\Pi)$ and $\varepsilon > 0$ there exists a $\underline{t} > 0$ such that $v(\Pi_{\bar{x},t}) \leq l + \varepsilon \ \forall \bar{x} \in S_L(f)$ and $t \geq \underline{t}$.*

*Proof.* Given any $\bar{x} \in S_L(f)$, call $\hat{x}$ the projection of $\bar{x}$ over $S_l(f)$; since $f(\bar{x}) \leq L$, using $^*$-compactness, one has $\|\hat{x} - \bar{x}\| \leq e(l, L) = \delta < \infty$. By Lemma 3.3, there exists $\underline{t}$ such that $S_\varepsilon(D_t) \supseteq B_2(\delta) \ \forall t \geq \underline{t}$, and therefore $v(\Pi_{\bar{x},t}) \leq f(\hat{x}) + D_t(\hat{x} - \bar{x}) \leq l + \varepsilon \ \forall t \geq \underline{t}$. ∎

Using the above property, we can supplement Lemma 3.4, proving "convergence" for the optimal value of (1.12) for every "reasonable" choice of the sequences $\{\bar{x}_i\}$ and $\{t_i\}$.

LEMMA 4.4. *If $f$ is $^*$-compact, then for any sequence $\{\bar{x}_t\}$ such that $f(\bar{x}_t) \leq L < \infty$,*

$$\underline{v} := \liminf_{t \to \infty} v(\Pi_{\bar{x}_t,t}) = v(\Pi).$$

*Proof.* Assume by contradiction that $v(\Pi) < l = \underline{v} - 3\varepsilon$ for some $\varepsilon > 0$. Applying Lemma 4.3, we obtain that, for large enough $t$, $v(\Pi_{\bar{x}_t,t}) \leq \underline{v} - 3\varepsilon + \varepsilon$. Furthermore, from the definition of $\underline{v}$, there exists a large enough $t$ such that $\underline{v} \leq v(\Pi_{\bar{x}_t,t}) + \varepsilon$. Hence, for this (large enough) $t$,

$$\underline{v} \leq v(\Pi_{\bar{x}_t,t}) + \varepsilon \leq \underline{v} - 2\varepsilon + \varepsilon + \varepsilon < \underline{v}. \qquad ∎$$

A final observation has to be made about polyhedral functions. In order to prove finite convergence results, a natural (but in principle nontrivial) assumption about the black box is required: as $f$ is characterized by a finite set of vectors and their $f^*$-values, (cf. (1.6)), the black box has to return as subgradients only those "extreme" vectors characterizing $f$. More generally, one could require

(4.1)
only *finitely many different* pairs $(f^*(z), z)$ can be returned by the black box.

**4.2. Assumptions on the *t*-strategy and the NS/SS decision.** In order to leave a large degree of freedom in the implementation of the algorithm, we prove convergence under four general rules; several different $t$-strategies, with different performances in practice, can be designed following these guidelines [Fr97, Chapter I.5]. Since these rules measure improvements w.r.t. the current value $f(\bar{x})$, let us introduce the following notation:

$$\delta_{\bar{x}}(d) = f(\bar{x} + d) - f(\bar{x}) \qquad \text{is the } \textit{actual} \text{ improvement and} \tag{4.2}$$

$$\delta_{\beta,\bar{x}}(d) = f_\beta(\bar{x} + d) - f(\bar{x}) \qquad \text{is the } \textit{predicted} \text{ improvement} \tag{4.3}$$

for a step at $\bar{x}+d$. Note that $\delta_{\bar{x}}(d) - \delta_{\beta,\bar{x}}(d) = \Delta f$, and that $\delta_{\beta,\bar{x}}(d^*) \leq 0$. (Otherwise, $d = 0$ would be a better solution of (1.8) than $d^*$.) In the following, we will use "SS" as a shorthand for "serious step," i.e., an iteration of the algorithm where the current point $\bar{x}$ is changed. Analogously, "NS" will stand for "null step," i.e., an iteration of the algorithm where $\bar{x}$ is not changed.

(4.i) If an SS is performed, then

$$\delta_{\bar{x}}(d^*) \le m\delta_{\beta,\bar{x}}(d^*) \tag{4.4}$$

for a fixed $m \in (0,1)$; the converse is *not* required, i.e., an SS may not be done even if a "considerable" improvement has been obtained, except for what is required by (4.iii) below.

(4.ii) During a sequence of *consecutive* NS, $t$ can *increase* only *finitely many times*.

(4.iii) During a sequence of *consecutive* NS, (4.4) can happen only *finitely many times*; that is, after finitely many NS, *any* step such that

$$\delta_{\bar{x}}(d^*) > m\delta_{\beta,\bar{x}}(d^*) \tag{4.5}$$

must be accepted.

(4.iv) During a sequence of *consecutive* NS, at all iterations (but possibly a finite number) $f$ must be evaluated in $\bar{x}+d^*$, and the model $f_+$ of the following iteration must take into account the corresponding $z \in \partial f(\bar{x}+d^*)$, in the sense that $f_+^*(z) \le f^*(z)$.

Let us briefly discuss the above rules. By (4.i), an SS is performed only if a consistent improvement is obtained. Changing the current point is not mandatory if some alternative strategy—typically increasing $t$—appears to be preferable, but, by (4.iii), this must not happen forever. A reasonable answer to a "bad" step is to decrease $t$; increasing $t$ is also possible, but it must be properly limited, e.g., by (4.ii). Finally, inserting the newly obtained subgradient into $\beta$ is not mandatory if some alternative strategy—typically decreasing $t$—appears to be preferable, but, by (4.iv), this must not happen forever. Using $f_+^* \ge f^*$, (4.iv) is equivalent to $f_+^*(z) = f^*(z)$; from the primal viewpoint, it says that

$$f(\bar{x}+d^*) = f_+(\bar{x}+d^*) \quad \text{and} \quad z \in \partial f_+(\bar{x}+d^*). \tag{4.6}$$

In some cases, a strengthened form of rule (4.ii) is useful, as follows.

(4.ii′) During a sequence of *consecutive* NS, $t$ can *change* only *finitely many times*.

A consequence of rules (4.ii) (or (4.ii′)), (4.iii), and (4.iv) is that, for any sequence of consecutive NS, there exists an iteration index $h$ such that for all the subsequent iterations in the sequence, $t$ is nonincreasing (fixed), $\delta_{\bar{x}}(d^*) > m\delta_{\beta,\bar{x}}(d^*)$, and $z$ is added to $\beta$. In the following, we will often refer to this $h$.

Inhibiting serious steps allows us to drop the $^*$-compactness assumption in some variants of the algorithm; thus, we will sometimes use the following rule.

(4.iii′) Only *finitely many* SS are done; after the last one, the stopping condition becomes $\Delta f \le \varepsilon$.

This rule is rather abstract, but several practical implementations can be imagined. For instance, the current point can just be kept fixed. Alternatively, if $v(\Pi)$ is finite, one could choose some $\varepsilon > 0$ and inhibit SS if $\delta_{\beta,\bar{x}}(d^*) \ge -\underline{\varepsilon}$ (a "negligible" step), as long as the $t$-strategy is properly managed.

With the three-level algorithm of Figure 2, sometimes the following weakened form of (4.iii′), which allows any *total* number of SS to be performed, suffices.

(4.iii″) For *each run* of the two-level bundle algorithm, only finitely many SS are done; after the last one, the stopping condition becomes $\Delta f \le \varepsilon$.

At the end of this section, let us remark that the very concept of SS, although apparently primal in nature, has a noteworthy "dual interpretation." From the dual viewpoint, a bundle method is an approximated ascent approach to $\sup_x\{v(\Delta_{x,t})\}$,

where an ascent in the value of the stabilized dual problem (1.11), i.e., $v(\Delta_{\bar{x}+d^*,t}) \geq v(\Delta_{\bar{x},t})$, is desired. Unfortunately, the values of $v(\Delta_{\bar{x}+d^*,t})$ and $v(\Delta_{\bar{x},t})$ are unknown, and therefore the condition cannot be checked; however, they can be estimated, using the dual pricing problem (1.3) ($v(\Delta_x) = -f(x)$) and the stabilized dual master problem (1.9), as

$$v(\Delta_{\bar{x}+d^*,t}) \geq v(\Delta_{\bar{x}+d^*}) \quad \text{and} \quad v(\Delta_{\beta,\bar{x},t}) \geq v(\Delta_{\bar{x},t}) \geq v(\Delta_{\bar{x}}).$$

(Remember Lemma 2.1: $(\Delta_{\bar{x}+d^*})$ is a linearization of $(\Delta_{\bar{x}+d^*,t})$ in $-z^*$ using the subgradient $d^*$.) Now, (4.4) is equivalent, via (2.6), to

$$v(\Delta_{\bar{x}+d^*}) \geq mv(\Delta_{\beta,\bar{x},t}) + (1-m)v(\Delta_{\bar{x}}).$$

Therefore, $v(\Delta_{\bar{x}+d^*})$ and $mv(\Delta_{\beta,\bar{x},t})+(1-m)v(\Delta_{\bar{x}})$ are taken as estimates of $v(\Delta_{\bar{x}+d^*,t})$ and $v(\Delta_{\bar{x},t})$, respectively, and used to decide whether $\bar{x}+d^*$ are better multipliers than $\bar{x}$. Note that there is a safeguard against "wild" decisions: at least, $v(\Delta_{\bar{x}+d^*}) \geq v(\Delta_{\bar{x}})$. Hence, even if $v(\Delta_{\bar{x},t})$ does not actually improve moving to $\bar{x}+d^*$, at least its lower approximation $v(\Delta_{\bar{x}})$ does.

**4.3. Assumptions on the $\beta$-strategy.** An important detail of any implementable bundle method is the *$\beta$-strategy*, i.e., how the information in $\beta$ is managed to keep the computational cost of the solution of (1.8)/(1.9) reasonably low. Removing subgradients from $\beta$ is important in practice, but heedless removals can impair convergence of the algorithm. A "minimal" requirement for any $\beta$-strategy is the following.

DEFINITION 4.5. *A $\beta$-strategy is* weakly monotone *if, during a sequence of consecutive NS, for each $i \geq h$ the optimal value of* (1.9) *is monotonically nonincreasing or, equivalently, the optimal value of* (1.8) *is monotonically nondecreasing.*

The equivalence between the two conditions in Definition 4.5 is (2.1). A weakly monotone $\beta$-strategy ensures at least convergence (to some value) of the optimal value of (1.8)/(1.9) during a sequence of consecutive NS. The definition does not specify how that monotonicity is obtained; a pretty minimal assumption on $f_\beta$ is the following.

DEFINITION 4.6. *A $\beta$-strategy is* monotone *if, during a sequence of consecutive NS, for each $i \geq h$*

$$f^*_{\beta_{i+1}}(z^*_i) \leq f^*_{\beta_i}(z^*_i), \tag{4.7}$$

*or, equivalently,*

$$f_{\beta_{i+1}}(\bar{x}+d) \geq f_{\beta_i}(\bar{x}+d^*_i) + z^*_i(d-d^*_i) \quad \forall d. \tag{4.8}$$

The equivalence between (4.7) and (4.8) can be easily proved using (2.4) and (1.1). A monotone $\beta$-strategy is weakly monotone; since $t_{i+1} \leq t_i \Rightarrow D^*_{t_{i+1}} \leq D^*_{t_i}$ for $i \geq h, v(\Delta_{\bar{x},\beta_{i+1},t_{i+1}}) \leq f^*_{\beta_{i+1}}(z^*_i) - z^*_i\bar{x} + D^*_{t_{i+1}}(-z^*_i) \leq f^*_{\beta_i}(z^*_i) - z^*_i\bar{x} + D^*_{t_i}(-z^*_i) = v(\Delta_{\bar{x},\beta_i,t_i})$.

The practical implementation of a monotone $\beta$-strategy depends on the model. For the cutting plane model $\hat{f}_\beta$, at each iteration the following two *moves* can be considered:

- remove some $z$ from $\beta$ (removal),
- add $z^*$ to $\beta$ (*aggregation*), with $f^*$-value $\hat{f}^*_\beta(z^*)$.

Rewriting (1.9) with $f_\beta = \hat{f}_\beta$ in the following equivalent form (cf. (1.7))

$$\inf_\theta \{\Sigma_{z\in\beta}(f^*(z) - z\bar{x})\theta_z + D^*_t(-\Sigma_{z\in\beta}z\theta_z) : \Sigma_{z\in\beta}\theta_z = 1, \theta \geq 0\}, \tag{4.9}$$

it is clear that aggregation offers one way for implementing a monotone $\beta$-strategy. If—as often happens—(1.9) is actually solved via (4.9), an alternative is to just avoid discarding all the $z \in \beta$ whose corresponding optimal multiplier $\theta_z^*$ is strictly positive, as

$$\hat{f}_\beta^*(z^*) = \sum_{z \in \beta} f^*(z)\theta_z^* \quad \text{and} \quad z^* = \sum_{z \in \beta} z\theta_z^*.$$

In principle, no more than $n+1$ of the optimal multipliers need to be strictly positive, although in practice whether or not such a minimal solution is obtained depends on the actual solver; even for $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$, active-set algorithms [Ki89, Fr96] would guarantee it, while interior-point algorithms may not. The above discussion justifies the following result.

LEMMA 4.7. *If $f_\beta = \hat{f}_\beta$ and, during a sequence of consecutive NS, for each iteration after $h$ either all the $z$ such that $\theta_z^* > 0$ are kept in $\beta$ or $z^*$ is added to $\beta$ with $\hat{f}_\beta^*(z^*)$ as the corresponding $f^*$-value, then the $\beta$-strategy is monotone.*

A monotone $\beta$-strategy allows us to keep the size of $\beta$ bounded (down to 2); if (P3″) does not hold, however, it is not sufficient to guarantee convergence [Fr97, section I.4.2]. A stronger property has to be used, which essentially inhibits all removals at length.

DEFINITION 4.8. *A $\beta$-strategy is* strictly monotone *if it is monotone and, if some $z$ has been removed from $\beta$, no other removal is permitted until $v(\Pi_{\beta,\bar{x},t})$ increases by a fixed $\mu > 0$.*

A strictly monotone $\beta$-strategy guarantees convergence for every choice of $D_t$; although it does not give any finite bound on the size of $\beta$, it can still be practical. Furthermore, there is a trade-off between the size of $\beta$—hence the computational cost of (1.9)—and the speed of convergence of the overall process [HL93b, section XIV.4.5]; a small $\beta$ is a good choice only in some cases [CFG01].

Finally, if $f$ is a polyhedral function, *finite* termination to an optimal solution can be proved, provided that aggregation is properly limited.

DEFINITION 4.9. *A $\beta$-strategy is* safe *if only finitely many aggregations are done.*

**5. Convergence of NS sequences (1st level).** The convergence proof is divided into three parts. In this section we assume that no SS occurs, i.e., we examine infinite sequences of consecutive NS; we shall show that these sequences allow us to compute the generalized Moreau–Yosida regularization with any finite precision. Therefore, in the next section we will be allowed to disregard what happens between two consecutive SS, i.e., focus on the convergence of the minimization process of the generalized Moreau–Yosida regularization (2nd level). Finally, in section 7 we will discuss the convergence of the 3rd level.

In this section, the iteration index $i$ denotes the $i$th NS of the (only) infinite sequence of consecutive NS that the algorithm is supposed to perform, and therefore the current point $\bar{x}$ is fixed. The iteration index $h$ is the one implied by the rules (4.ii) (or (4.ii′)), (4.iii), and (4.iv). To simplify the notation, let $(\Delta_i)$ and $(\Pi_i)$ denote, respectively, the dual and primal stabilized master problems (1.9) and (1.8) solved in that iteration; $z_i^*$ and $d_i^*$ their solutions; $f_i(\hat{f}_i)$ the corresponding (cutting plane) model; $z_i$ the subgradient reported by the evaluation of $f(\bar{x} + d_i^*)$; $\delta_i$ the predicted improvement; and so on. Also, we will use the shorthand index "+" for "$i+1$."

In the following, we will always assume that (P1), (P2), (P4), and (P5) hold; additional assumptions will be explicitly listed. The first step in the convergence proof is to show that the algorithm is well defined, i.e., that the primal and dual

stabilized master problems have optimal solutions. This requires (P3) or (P3′), as well as minimal assumptions on $f_\beta$.

LEMMA 5.1. *Under the hypothesis of Lemma* 2.1*, if either* (P3′) *or* (P3) *hold, then* $(\Delta_i)$ *and* $(\Pi_i)$ *attain finite optimal solutions* $z_i^*$ *and* $d_i^*$*, respectively.*

*Proof.* If (P3′) holds, then from $f_i \geq f_*$ we have that $f_i(\bar{x}+d)+D_{t_i}(d) > f(\bar{x})\ \forall d \notin S_\delta(D_{t_i})$, where $\delta = f(\bar{x}) - f_*$. Since $S_\delta(D_{t_i})$ is compact by (P2), the infimum must be finitely attained. Otherwise (P3) holds, i.e., $D_t$ is strongly coercive; hence $f_\beta(\bar{x} + \cdot) + D_t(\cdot)$ is strongly coercive too. (Strongly coercive functions increase faster than any linear function at infinity, and any convex function is minorized by an affine function [HL93a, Proposition IV.1.2.1].) Therefore, $(\Pi_i)$ has a bounded nonempty set of optimal solutions [HL93a, Remark IV.3.2.8]. Finally, [HL93a, Theorem X.2.3.2] shows that an optimal $z_i^*$ exists for $(\Delta_i)$ whenever an optimal $d_i^*$ exists for $(\Pi_i)$. ∎

We will now focus on proving the boundedness of the sequences $\{d_i^*\}$ and/or $\{z_i^*\}$, under a set of different assumptions.

LEMMA 5.2. *Under the hypothesis of Lemma* 2.2*, if* (P3′) *holds, then* $\{d_i^*\}$ *and* $\{z_i^*\}$ *are bounded.*

*Proof.* Boundedness of $\{d_i^*\}$ was in fact established in Lemma 5.1, as $D_{t_i}(d_i^*) \leq \delta = f(\bar{x}) - f_*$ and, by (P4), $S_\delta(D_{t_i}) \subseteq S_\delta(D_{t_h})$; the latter is compact by (P2). By (2.3) and $f_i \leq f$, $z_i^*$ is an $\varepsilon_i$-subgradient of $f$ at $\bar{x}+d_i^*$ for $\varepsilon_i = \Delta f_i = f(\bar{x}+d_i^*) - f_i(\bar{x}+d_i^*) \geq 0$; since $f$ is finite everywhere, and therefore bounded over any compact set, and $f_i \geq f_*$, $\varepsilon_i \leq \bar{\varepsilon} < \infty$. Hence, $z_i^* \in \partial_{\bar{\varepsilon}} f(\bar{x} + d_i^*)\ \forall i \geq h$. The image of a compact set in int *dom* $f = \Re^n$ under the $\bar{\varepsilon}$-subdifferential mapping (see [HL93b, Proposition XI.4.1.2]) is compact. ∎

Thus, (P3′)/(P*3′) guarantee the boundedness of both solution sequences. In all the development, the first part of (P2) (compactness) is only used in Lemma 5.2, and therefore it could be dropped if (P3) holds; however, strong coercivity implies the boundedness of the level sets [HL93a, Proposition IV.3.2.5.(ii)], and hence there is no loss of generality—and a gain in symmetry—in requiring it to hold in general.

In Lemma 5.2, the boundedness of $\{z_i^*\}$ is obtained as a consequence of the boundedness of $\{d_i^*\}$, finiteness of $f$, and $f_i \geq f_*$; with basically the same argument, it is possible to prove the boundedness of $\{d_i^*\}$, *given* the boundedness of $\{z_i^*\}$ and (P3)/(P*3).

LEMMA 5.3. *Under the hypothesis of Lemma* 2.2*, if* (P*3) *holds and* $\{z_i^*\}$ *is bounded, then* $\{d_i^*\}$ *is bounded.*

*Proof.* By (2.2), $d_i^* \in \partial D_{t_i}^*(-z_i^*)$. From $t_i \leq t_h (\Rightarrow D_{t_h}^* \geq D_{t_i}^*$ by (P*4)), it is clear that $d_i^*$ must also be an $\varepsilon$-subgradient of $D_{t_h}^*$ for a proper $\varepsilon$; indeed, it is easy to check that $d_i^*$ is a $\varepsilon_i$-subgradient of $D_{t_i}^*$ at $-z_i^*$ for $\varepsilon_i = D_{t_h}^*(-z_i^*) \geq D_{t_h}^*(-z_i^*) - D_{t_i}^*(-z_i^*) \geq 0$. Since, by (P*3), $D_{t_h}^*$ is finite convex (hence continuous) and $\{z_i^*\}$ is bounded, $\varepsilon_i \leq \bar{\varepsilon} < \infty$. Hence, $d_i^* \in \partial_{\bar{\varepsilon}} D_{t_h}^*(-z_i^*)\ \forall i \geq h$; reasoning as in Lemma 5.2, we obtain that $\{d_i^*\}$ is bounded. ∎

Boundedness of $\{z_i^*\}$ under (P3)/(P*3) is not easy to establish in general; however, when obtained, it allows us to prove the boundedness of all the relevant sequences, as the same argument proves boundedness of $\{z_i\}$, given the boundedness of $\{d_i^*\}$.

LEMMA 5.4. *Under the hypothesis of Lemma* 2.1*, if* $\{d_i^*\}$ *is bounded, then the sequences* $\{f^*(z_i)\}$ *and* $\{z_i\}$ *obtained by evaluating* $f(\bar{x} + d_i^*)$ *are bounded.*

*Proof.* Since $f$ is finite everywhere and $z_i \in \partial f(\bar{x} + d_i^*)$, we can invoke [HL93a, Remark VI.6.2.3] to conclude that all the $z_i$ belong to a compact set. Also, $-f(0) \leq f^*(z_i) = z_i(\bar{x} + d_i^*) - f(\bar{x} + d_i^*)$, which is bounded above since both $\{d_i^*\}$ and $\{z_i\}$ are bounded and $f$ is bounded below over any compact set. ∎

Note that the above results do not depend on the $\beta$-strategy. Indeed, there are several situations in which the boundedness of $\{d_i^*\}$ is "free"; among them, let us mention the following:

- $dom(D_{t_h})$ is compact ($D_t$ is a "trust region," cf. Example 3.3), as $dom(D_{t_i}) \subseteq dom(D_{t_h})$ by (P4);
- (P3) holds and $dom(f^*)$ is compact ($f$ is globally Lipschitz, e.g., polyhedral), as $dom(f_i^*) \subseteq dom(f^*)$ and Lemma 5.3 gives boundedness of $\{d_i^*\}$.

Conversely, let us mention that the boundedness of $\{d_i^*\}$ implies the boundedness of $\{z_i^*\}$ whenever the cutting plane model $\hat{f}_\beta$ is used, as from (1.7) every $z_i^*$ belongs to the convex hull of $\{z_i\}$ and, from Lemma 5.4, the latter set is bounded whenever $\{d_i^*\}$ is.

**5.1. Results with a weakly monotone $\beta$-strategy.** We will now prove some results which only require a weakly monotone $\beta$-strategy and $f_\beta \le f$. The basic observation is that, with a weakly monotone $\beta$-strategy, by Definition 4.5 we have, $\forall i \ge h$,

(5.1)
$$D_{t_i}^*(-z_i^*) \le f_i^*(z_i^*) - z_i^*\bar{x} + f(\bar{x}) + D_{t_i}^*(-z_i^*) = v(\Delta_i) + f(\bar{x}) \le v(\Delta_h) + f(\bar{x}) < \infty$$

(use (1.vii) and $f_i^* \ge f^*$). In the proximal bundle case ($D_t^* = \frac{1}{2}t\|\cdot\|_2^2$), where $d_i^* = -t_i z_i^*$, (5.1) proves the boundedness of $\{d_i^*\}$; this is also true in the more general case, provided that $D_t$ has the form (3.3). The proof relies on the following "primal view" of (5.1),

$$D_{t_i}(0) - D_{t_i}(d_i^*) - (-z_i^*)(0 - d_i^*) \le v(\Delta_h) + f(\bar{x}) < \infty \tag{5.2}$$

(use (2.5) and (P1)); (5.2) can be expressed by saying that the *linearization error* (cf. (1.10)) in 0, made by approximating $D_{t_i}$ with its linearization in $d_i^*$ using slope $-z_i^*$, is bounded.

LEMMA 5.5. *Under the hypothesis of Lemma* 2.2, *if* (P3) *holds, a weakly monotone $\beta$-strategy is used, and $D_t$ has the form* (3.3), *then $\{d_i^*\}$ is bounded.*

*Proof.* From (2.2), $-z_i^* \in \partial D_{t_i}(d_i^*)$; since $D_t$ has the form (3.3), defining $\bar{z}_i^* := t_i z_i^*$, we have that $-\bar{z}_i^* \in \partial D(d_i^*)$. Hence, (5.2) can be written as

$$\frac{1}{t_i}D(0) - \frac{1}{t_i}D(d_i^*) - \frac{1}{t_i}(-\bar{z}_i^*)(0 - d_i^*) \le \varepsilon < \infty,$$

whence $D(0) - D(d_i^*) - (-\bar{z}_i^*)(0 - d_i^*) \le \varepsilon t_i \le \varepsilon t_h < \infty$.

Now, call $\bar{V}_\varepsilon(\bar{d})$ the set of all $d$ such that the linearization error in $\bar{d}$, made by approximating $D$ with its linearization in $d$ using any $z \in \partial D(d)$, is smaller than $\varepsilon$. Since $D$ is strongly coercive, $\bar{V}_\varepsilon(\bar{d})$ is compact for any $\bar{d}$ and any fixed $\varepsilon$ [HL93b, Proposition XI.4.2.6.(i)]. □

An alternative result does not require (3.3) but rather that $t_i$ remain bounded away from zero; actually, in this case the boundedness of $\{z_i^*\}$ is obtained, which, in view of Lemma 5.3, is a stronger result.

LEMMA 5.6. *Under the hypothesis of Lemma* 2.2, *if* (P3) *holds, a weakly monotone $\beta$-strategy is used, and $t_i \ge \underline{t} > 0$ ($t$ is bounded away from 0), then the sequences $\{f_i^*(z_i^*)\}$ and $\{z_i^*\}$ are bounded.*

*Proof.* From (5.1), (P*4), and $\underline{t} \le t_i$ we have $D_{\underline{t}}^*(-z_i^*) \le v(\Delta_h) + f(\bar{x}) < \infty$; the level sets of $D_{\underline{t}}^*$ are compact from (P*2), and hence $\{z_i^*\}$ is bounded. Looking again

at (5.1), we notice that $f_i^*(z_i^*)$ is bracketed between bounded quantities; hence it is also bounded. □

In practice, $t$ should not become too small anyway, so the condition in the above lemma is not really binding; yet, in many cases it can simply be dropped.

**5.2. Convergence with a monotone $\beta$-strategy.** The above boundedness results are instrumental for proving the actual convergence of a sequence of NS, that is, the fact that the stabilized master problems (1.8) and (1.9) can be used to approximate the stabilized problems (1.12) and (1.11) within any required degree of accuracy. Due to (2.7), it is only necessary to prove that $\{\Delta f_i\} \to 0$. With a monotone $\beta$-strategy, this requires (P*3″).

Consider the (convex) function

$$r_i(z) := f_i^*(z) - z\bar{x} + f(\bar{x}),$$

so that $r_i(z) + D_{t_i}^*(z)$ is, but for the constant $f(\bar{x})$, the objective function of $(\Delta_i)$; from $f_i^* \geq f^*$ and (1.vii), $r_i \geq 0$. Now define

$$\zeta_i := z_i - z_i^* \quad \text{and} \quad z_i(\gamma) := z_i^* + \gamma\zeta_i.$$

From Definition 4.6 ($f_+^*(z_i^*) \leq f_i^*(z_i^*)$) and (4.iv) ($f_+^*(z_i) \leq f^*(z_i)$) we have that, $\forall \gamma \in [0,1]$,

$$\begin{aligned} h_i(\gamma) &:= [f_i^*(z_i^*)(1-\gamma) + f^*(z_i)\gamma - z_i(\gamma)\bar{x} + f(\bar{x})] + D_{t_i}^*(-z_i(\gamma)) \\ &\geq [f_+^*(z_i^*) - z_i^*\bar{x} + f(\bar{x})](1-\gamma) + [f_+^*(z_i) - z_i\bar{x} + f(\bar{x})]\gamma + D_{t_i}^*(-z_i(\gamma)) \\ &\geq r_+(z_i^*)(1-\gamma) + r_+(z_i)\gamma + D_{t_+}^*(-z_i(\gamma)) \geq r_+(z_i(\gamma)) + D_{t_+}^*(-z_i(\gamma)). \end{aligned}$$

(We have also used $t_+ \leq t_i \Rightarrow D_{t_+}^* \leq D_{t_i}^*$ and the convexity of $r_+$.) Therefore, defining

$$(\vartheta_i) \qquad \min_\gamma \{h_i(\gamma) : \gamma \in [0,1]\},$$

$$v_i(z) := \begin{cases} r_i(z) + D_{t_i}^*(-z) & \text{if } z = z_i(\gamma) \text{ for some } \gamma \in [0,1], \\ +\infty & \text{otherwise,} \end{cases}$$

one clearly has

$$v(\vartheta_i) \geq \min_z \{v_+(z)\} \geq v(\Delta_+) + f(\bar{x}).$$

We will study the behavior of $v(\vartheta_i)$ during sequences of consecutive NS to estimate the convergence speed of $v(\Delta_i)$. In practice, this corresponds to the "aggressive" monotone $\beta$-strategy, where aggregation is performed at every step and all subgradients but $z_i^*$ and $z_i$ are discarded.

Due to (P*3″), $D_{t_i}^*$ is differentiable, and hence so is $h_i$; from (2.2), we have $d_i^* = \nabla D_{t_i}^*(-z_i(0))$; hence by (2.6)

$$h_i'(0) = f^*(z_i) - f_i^*(z_i^*) - (z_i - z_i^*)(\bar{x} + d_i^*) = -\Delta f_i. \tag{5.3}$$

Using (1.vi) and (2.4), the NS condition (4.5) can be written as

$$z_i(\bar{x} + d_i^*) - f^*(z_i) - f(\bar{x}) > m[z_i^*(\bar{x} + d_i^*) - f_i^*(z_i^*) - f(\bar{x})];$$

hence

$$h_i'(0) < (1-m)[z_i^*(\bar{x}+d_i^*) - f_i^*(z_i^*) - f(\bar{x})].$$

From (1.vi), $-z_i^* d_i^* = D_{t_i}(d_i^*) + D_{t_i}^*(-z_i^*) \geq D_{t_i}^*(-z_i^*)$; hence

$$h_i'(0) < (1-m)[-f_i^*(z_i^*) + z_i^*\bar{x} - D_{t_i}^*(-z_i^*) - f(\bar{x})] = -(1-m)h_i(0). \tag{5.4}$$

Using (5.4) it is possible to show, adapting standard results from smooth optimization [OR70], that $\{-h_i'(0) = \Delta f_i\} \to 0$ if $\{z_i^*\}$ is bounded and (P*3″) holds. In the general case, this requires some assumptions on the behavior of $t_i$, the simplest one being rule (4.ii′).

THEOREM 5.7. *Under the hypothesis of Lemma* 2.2, *if* (P*3″) *holds, rule* (4.ii′) *is in force, a monotone* $\beta$*-strategy is used, and* $\{z_i^*\}$ *is bounded, then* $\{\Delta f_i\} \to 0$.

*Proof.* Wait until the iteration $h$ implied by rules (4.ii′), (4.iii), and (4.iv): $t_i = t\ \forall i \geq h$. The boundedness of $\{z_i^*\}$, together with Lemma 5.3 ((P*3″) ⇒ (P*3)) and Lemma 5.4, implies that $\{z_i\}$ is also bounded; therefore, the set

$$Z := conv(\{z_i^*\} \cup \{z_i\})$$

is compact and contains all the segments $[z_i^*, z_i]$. From (P*3″), $D_t^*$ is differentiable and therefore *continuously differentiable* [HL93a, Remark VI.6.2.6] on $Z$; that is, $\nabla D_t^*$ is continuous, and therefore *uniformly continuous*, on the compact set $Z$. Note that if $\nabla D_t^*(z_i^*) = d_i^* = d_+^* = \nabla D_t^*(z_+^*)$, then $\Delta f_+ = 0$ as, from (4.6), $f(\bar{x}+d_i^*) = f_+(\bar{x}+d_i^*) = f_+(\bar{x}+d_+^*)$. Hence

$$\sup\{\|\nabla D_t^*(z') - \nabla D_t^*(z'')\| : z', z'' \in Z\} > 0.$$

The *reverse modulus of continuity* of $\nabla D_t^*$ over $Z$

$$\kappa(v) := \inf\{\|z'-z''\| : \|\nabla D_t^*(-z') - \nabla D_t^*(-z')\| \geq v, z', z'' \in Z\}$$

is an $F$-function, i.e., nondecreasing and such that $\kappa(0) = 0$ and $\kappa(v) > 0$ for $v > 0$ [OR70, Definition 14.2.6]. (Our definition of $\kappa$ is nonstandard, in that *dom* $\kappa$ may not be the whole $\Re_+$, but we will always evaluate $\kappa$ at points of its domain.)

We claim the existence of an $F$-function $\rho$ such that

$$v(\Delta_i) - v(\Delta_+) \geq h_i(0) - v(\vartheta_i) \geq \rho(-h_i'(0)) > \rho((1-m)h_i(0)),$$

which clearly implies that $\{-h_i'(0) = \Delta f_i\} \to 0$ and $\{h_i(0)\} \to 0$ and therefore proves the theorem. (Note that $\{v(\Delta_i)\}$ is bounded below, as $v(\Delta_i) \geq -f(\bar{x})$.) The function $\rho$ estimates how much of the decrease "promised" by $h_i'(0)$ is actually attained in the optimal solution of $(\vartheta_i)$.

A special case for which the estimate is easy is $z_i = z_i^*$, i.e., $\zeta_i = 0$: the corresponding $h_i(\gamma)$ is linear, the optimal solution of $(\vartheta_i)$ is $\gamma = 1$, and $h_i(1) = h_i(0) + h_i'(0)$; hence $\rho \equiv 1$.

Otherwise, for the reverse modulus of continuity of $h_i'$ over $[0,1]$ one has

$$\begin{aligned}
\sigma_i(v) &:= \inf\{|\gamma'-\gamma''| : |h_i'(\gamma') - h_i(\gamma'')| \geq v, \gamma', \gamma'' \in [0,1]\} \\
&\geq \frac{1}{\|\zeta_i\|} \inf\{\|(\gamma'-\gamma'')\zeta_i\| : \|(\nabla D_t^*(-z_i(\gamma')) - \nabla D_t^*(-z_i(\gamma')))\zeta_i\| \\
&\qquad\qquad \geq v, \gamma', \gamma'' \in [0,1]\} \\
&\geq \frac{1}{\|\zeta_i\|} \inf\left\{\|z'-z''\| : \|\nabla D_t^*(z') - \nabla D_t^*(z'')\| \geq \frac{v}{\|\zeta_i\|}, z', z'' \in Z\right\}
\end{aligned}$$

(note that $\zeta_i \neq 0$), and therefore

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right). \tag{5.5}$$

Now, define

$$\gamma^* := \inf_{\gamma}\left\{\gamma \geq 0 : h_i'(\gamma) \geq \frac{1}{2}h_i'(0)\right\} \geq \sigma_i\left(-\frac{1}{2}h_i'(0)\right). \tag{5.6}$$

(1/2 is arbitrary; any strictly positive number would do.) By (5.3), if $\gamma^* = 0$, then $\Delta f_i = 0$ and the theorem is proved. Otherwise, the following two cases may arise.

If $\gamma^* \geq 1$, then $\gamma = 1$ is the optimal solution of $(\vartheta_i)$ and $h_i'(\gamma) \leq \frac{1}{2}h_i'(0)\ \forall\gamma \in [0,1]$. ($h_i'$ is nondecreasing since $h_i$ is convex.) In particular, $h_i'(1) \leq \frac{1}{2}h_i'(0) < 0$, and therefore

$$h_i(1) + h_i'(1)(0-1) \leq h_i(0) \Rightarrow h_i(1) \leq h_i(0) + \frac{1}{2}h_i'(0).$$

If $\gamma^* < 1$, then by the mean-value theorem there exists some $\bar\gamma \in (0,\gamma^*)$ such that

$$h_i(\gamma^*) = h_i(0) + h_i'(\bar\gamma)\gamma^* \Rightarrow h_i(\gamma^*) \leq h_i(0) + \frac{1}{2}h_i'(0)\gamma^*.$$

Hence, using (5.5) and (5.6),

$$v(\Delta_i) - v(\Delta_+) \geq h_i(0) - h_i(\gamma^*) \geq -\frac{1}{2\|\zeta_i\|}h_i'(0)\kappa\left(-\frac{1}{2\|\zeta_i\|}h_i'(0)\right).$$

Thus, the claim is proved with $\rho(v) = \frac{v}{2}\min\left\{1, \frac{1}{diam(Z)}\kappa\left(\frac{v}{2diam(Z)}\right)\right\}$, where $diam(Z) < \infty$ is the maximum distance of any two points in $Z$. □

In the above proof, rule (4.ii′) is needed because $t \to D_t^*(z)$ may be almost any function; rule (4.ii) suffices, thereby allowing $\{t_i\} \to 0$, if this function is "simple."

THEOREM 5.8. *Under the hypothesis of Lemma* 2.2, *if* (P*3″) *holds, a monotone* $\beta$*-strategy is used,* $\{z_i^*\}$ *is bounded, and* $D_t^*$ *has either the form* (3.3) *or the form* (3.4), *then* $\{\Delta f_i\} \to 0$.

*Proof.* With the notations of Theorem 5.7, let $D_t^* = tD^*$ and call $\kappa_i$ and $\kappa$ the reverse modulus of continuity of $\nabla D_{t_i}^*$ and of $\nabla D^*$, respectively, on $Z$. It is easy to check that

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa_i\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{t_i\|\zeta_i\|}\right) \geq \frac{1}{\|\zeta_i\|}\kappa\left(\frac{v}{t_h\|\zeta_i\|}\right),$$

as $t_i \leq t_h$ and $\kappa$ is nondecreasing. If $D_t = \frac{1}{t}D$ instead, one has $D_t^*(z) = \frac{1}{t}D^*(tz)$ and therefore $\nabla D_t^*(z) = \nabla D^*(tz)$; simple calculations yield

$$\sigma_i(v) \geq \frac{1}{\|\zeta_i\|}\kappa_i\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{t_i\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right) \geq \frac{1}{t_h\|\zeta_i\|}\kappa\left(\frac{v}{\|\zeta_i\|}\right),$$

where $\kappa$ is the reverse modulus of continuity of $\nabla D^*$ on the (compact) set $\{tz : z \in Z, t \in [0,t_h]\}$. In both cases, the proof of Theorem 5.7 can be easily adapted by using the above functions in place of the reverse modulus of continuity of $D_t^*$ for the fixed $t$ provided by rule (4.ii′). □

The similar theorem [Au87, Theorem 2.3] (with a different proof) is proved for a *fixed* $t$ and $D_t$ of the form (3.3), differentiable and satisfying (P3″). Note that differentiability—which would at first appear to be a natural assumption—is necessary *in the dual* rather than in the primal, the critical property of $D_t$ being strict convexity. It is also interesting to note that in [Au87] a primal notation is used, but (1.9) is developed—only for the simple case $\beta_+ = \{z_i^*, z_i\}$—as a tool for proving [Au87, Theorem 2.3].

The above theorems rely on the compactness of $Z$, which is a consequence of the boundedness of $\{z_i^*\}$. The latter is, in several cases, either free or a consequence of the boundedness of $\{d_i^*\}$, which may not require $t_i$ bounded away from 0 (cf. Lemma 5.5). Thus, these results generalize those available for the proximal bundle method. Indeed, applying Theorem 5.7 to $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$, whose reverse modulus of continuity is $\kappa(v) = v/t$ (that does *not* depend on $Z$), one obtains an estimate that is only a 1/2 factor—due to the arbitrary 1/2 in the proof—away from the tightest possible one, if aggregation is allowed:

$$v(\Delta_i) - v(\Delta_+) \geq \frac{(1-m)(v(\Delta_i) + f(\bar{x}))}{2} \min\left\{1, \frac{(1-m)(v(\Delta_i) + f(\bar{x}))}{t_i \|z_i - z_i^*\|_2^2}\right\}.$$

The above estimate was obtained in [Fr97, Theorem I.2.2.2] (apart from a minor error) with basically the same arguments of Theorem 5.7, only using ad hoc relations.

Finally, note that all the results until now do not require $f_\beta$ to be the cutting plane model, and therefore they can be used in the analysis of "nonstandard" bundle methods [GM91].

**5.3. Convergence with a strictly monotone $\beta$-strategy.** When (P3′) does not hold, a monotone $\beta$-strategy does not guarantee convergence [Fr97, section I.4.2], and strict monotonicity is required. Furthermore, the following strengthened form of the rule in (4.iv)

$$\exists \text{ an index } h \text{ such that}, \forall i > j \geq h, f^*(z_j) = f_i^*(z_j) \tag{5.7}$$

is required; that is, at length the "accuracy" of $f_\beta^*$ as a model of $f^*$ cannot "deteriorate" once it has become "exact" in the dual points $z_i$.

THEOREM 5.9. *Under the hypothesis of Lemma* 2.2*, if* (5.7) *holds, a strictly monotone $\beta$-strategy is used, and $\{d_i^*\}$ is bounded, then $\{\Delta f_i\} \to 0$.*

*Proof.* By (2.3), $\bar{x} + d_i^* \in \partial f_i^*(z_i^*)$; hence, using (5.7),

$$f^*(z_j) - (\bar{x} + d_i^*)z_j \geq f_i^*(z_j) - (\bar{x} + d_i^*)z_j \geq f_i^*(z_i^*) - (\bar{x} + d_i^*)z_i^* \quad \forall i > j. \tag{5.8}$$

From $z_i \in \partial f(\bar{x} + d_i^*)$ and $z_i^* \in \partial f_i(\bar{x} + d_i^*)$ (cf. (2.3)), using (1.vi), one has

$$\Delta f_i = f(\bar{x} + d_i^*) - f_i(\bar{x} + d_i^*) = f_i^*(z_i^*) - f^*(z_i) + (\bar{x} + d_i^*)(z_i - z_i^*). \tag{5.9}$$

Using (5.8) in (5.9) to eliminate $z_i^*$, one obtains

$$\Delta f_i \leq \min_{i>j}\{f^*(z_j) - f^*(z_i) + (\bar{x} + d_i^*)(z_i - z_j)\}. \tag{5.10}$$

Sending $j \to \infty$, the min in (5.10) goes to 0 since $\{d_i^*\}$ is bounded, and hence, by Lemma 5.4, $\{z_i\}$ and $\{f^*(z_i)\}$ are also bounded. (Extract a subsubsequence such that both the $z$-values and the $f^*$-values converge to a cluster point.) □

The proof of Theorem 5.9 is essentially that of [HL93b, Theorem XII.4.2.3] for the cutting plane method, working *in the dual space* rather than in the primal space.

This shows the usefulness of our dual treatment, as the primal proofs of convergence of proximal and trust-region bundle methods are not easy to unify. Also, note that $t$ is not even mentioned in the proof, and hence nothing prevents $t_i \to 0$.

It is easy to verify that the cutting plane model $\hat{f}_\beta$ with a strictly monotone $\beta$-strategy guarantees (5.7). A strictly monotone $\beta$-strategy is weakly monotone, i.e., $v(\Pi_i)$ is nondecreasing; since it is also upper bounded by $f(\bar{x})$, it is clear that it can increase by any fixed quantity $\mu > 0$ only finitely many times. Hence, after some iteration $h$, no information is removed from $\beta$. Now, from (1.7) one has that $\hat{f}_i^*(z_j) \leq f^*(z_j)\ \forall i > j (z_j \in \beta_i)$, but $\hat{f}_i^* \geq f^*$.

Finally, note that this theorem requires the boundedness of $\{d_i^*\}$ (which implies that of $\{z_i\}$) but *not* of $\{z_i^*\}$. With $f_\beta = \hat{f}_\beta$, however, this is actually not an advantage, since, as we noted previously, in this case the boundedness of $\{d_i^*\}$ implies that of $\{z_i^*\}$.

**5.4. Overall NS convergence result.** We have shown that, under a number of different assumptions on the function, the model, and the stabilizing term, $\{\Delta f_i\} \to 0$ during infinite sequences of NS. In view of (2.7), this means that NS can be used to approximate (1.12) and (1.11) as closely as desired. This is more easily seen if $t$ is fixed at length (e.g., rule (4.ii′) is in effect); then, an infinite sequence of NS solves (1.12) and (1.11) for the current point $\bar{x}$ and the fixed $t$. Compactness of $\{d_i^*\}$ is typically required, and that of $\{z_i^*\}$ is usually available as well; hence, by the lower semicontinuity of the objective functions, subsequences of $\{d_i^*\}$ and $\{z_i^*\}$ can be extracted which converge to finite optimal solutions, respectively, for (1.12) and (1.11).

From the algorithmic viewpoint, $\{\Delta f_i\} \to 0$ implies the finite termination of the sequences of NS for $\varepsilon > 0$; this uses the following basic relation about the predicted improvement:

$$-\delta_i \geq -\delta_i - D_{t_i}(d_i^*) = -f_i(\bar{x} + d_i^*) + f(\bar{x}) - D_{t_i}(d_i^*) = \alpha_i^* + D_{t_i}^*(-z_i^*) = 0 \tag{5.11}$$

(use $D_{t_i} \geq 0$, (2.4), (2.5), and (2.9)).

THEOREM 5.10. *Assume that $\{\Delta f_i\} \to 0$; if $\varepsilon > 0$, then after finitely many consecutive NS either the stopping condition of the algorithm in Figure* 1 *holds or an SS is done; otherwise ($\varepsilon = 0$), $\{D_{t_i}^*(-z_i^*)\} \to 0$, and $\{\alpha_i^*\} \to 0$.*

*Proof.* Assume that infinitely many NS are done; (4.5) can be rewritten, using (4.2) and (4.3) first and then (5.11), as

$$\Delta f_i > -(1-m)\delta_i \geq (1-m)[\alpha_i^* + D_{t_i}^*(-z_i^*)]. \tag{5.12}$$

Since $\{\Delta f_i\} \to 0$, both $\{\alpha_i^*\}$ and $\{D_{t_i}^*(-z_i^*)\}$ must go to zero; if $\varepsilon > 0$, then the stopping condition of the algorithm in Figure 1 eventually holds. ∎

Note that, in general, $\{D_{t_i}^*(-z_i^*)\} \to 0$ does not imply $\{z_i^*\} \to 0$; consider the case where $D_t^*$ is a "trust region" (cf. Example 3.2) and/or $\{t_i\} \to 0$.

The above development can be extended to the case where $\delta_i$ in (4.4)/(4.5) is replaced with $\underline{\delta}_i = \delta_i + D_{t_i}(d_i^*)$; this corresponds to checking $f(\bar{x}+d_i^*)$ against $v(\Pi_i) = f_i(\bar{x} + d_i^*) + D_{t_i}(d_i^*)$ rather than against $f_i(\bar{x} + d_i^*)$. In fact, it is easy to check that (5.12) can be obtained as well from (5.11) and the modified from of (4.5) using $\underline{\delta}_i$. As observed in [HL93b, section XV.3], [Ki99, section 5], this descent test is weaker than (4.4) ($\underline{\delta}_i \geq \delta_i$), and therefore it may reduce the number of NS.

**5.5. The polyhedral case.** Finite termination of NS sequences requires $\varepsilon > 0$ and $m < 1$; in general, there is no chance of solving $(\Delta_{\bar{x},t})$ to optimality, i.e., of

obtaining $f_i(\bar{x}+d_i^*) = f(\bar{x}+d_i^*)$, unless $f$ is polyhedral. A finite convergence theorem for NS sequences can be proved for two different sets of assumptions, basically corresponding to those of Theorem 5.7 (with a *safe* $\beta$-strategy) and those of Theorem 5.9 (with $f_\beta = \hat{f}_\beta$).

THEOREM 5.11. *Assume that* $\{\Delta f_i\} \to 0$, *$f$ is polyhedral, and* (4.1) *is satisfied. If either* (P*3″) *holds, rule* (4.ii′) *is in effect, and a safe $\beta$-strategy is used, or $f_\beta = \hat{f}_\beta$ and a strictly monotone $\beta$-strategy is used, then after finitely many consecutive NS either the stopping condition of the algorithm in Figure* 1 *holds or an SS is done, even if $\varepsilon = 0$ and $m = 1$.*

*Proof.* Assume by contradiction that the stopping condition does not hold and that $\Delta f_i = \delta_{\bar{x}}(d_i^*) - \delta_i \geq \delta_{\bar{x}}(d_i^*) - m\delta_i > 0$ for infinitely many $i$.

If (P*3″) holds, (5.3) gives $h_i'(0) = -\Delta f_i < 0$, and therefore $v(\Delta_+) > v(\Delta_i)$; we can conclude that the set $\{v(\Delta_i)\}$ must be infinite. But the assumptions on $f$ and the safe $\beta$-strategy ensure that there are only finitely many different possible sets $\beta$; after the iteration $h$ implied by rule (4.ii′), $v(\Delta_i)$ can have only finitely many different values ($\bar{x}$ and $t$ are fixed).

For the other case ($f_\beta = \hat{f}_\beta$ and a strictly monotone $\beta$-strategy), note that if the pair $(f^*(z_i), z_i)$ already belongs to $\beta_i$, then $\Delta f_i = 0$. (Use (1.6) and $\hat{f}_i \leq f$.) Now, Definition 4.8 and $\{\Delta f_i\} \to 0$ ensure that, at length, removals are inhibited; by (4.1), only finitely many "new" pairs $(f^*(z_i), z_i)$ can ever be generated, which yields the contradiction. □

**6. Convergence of SS sequences (2nd level).** Having proved convergence of the NS sequences, in the following we disregard what happens between two consecutive SS. However, we are not allowed to entirely disregard NS; in fact, it may happen that only finitely many SS are done, so that a "tail" of (possibly infinitely many) consecutive NS is done after the last SS. In order to deal with the two different cases—finitely many and infinitely many serious steps—in a unified way, in this section we will use the following notation: the index $i$ denotes the $i$th serious step if at least $i$ SS are performed; otherwise it denotes the $(i-k)$th NS of the only infinite sequence of NS that starts right after that the last SS (the $k$th) is performed. With this notation, $\bar{x}_i, d_i^*, z_i^*, \delta_i, \ldots$ refer to the status of the algorithm just *before* the change of the current point occurring at step $i$, if any.

The standing assumption for all the results in this section is

conditions sufficient to guarantee $\{\Delta f_i\} \to 0$ during an infinite sequence of NS hold.

Several different such conditions exist, as we have shown in the previous sections. About the model, without further notice, we will require only $f_i \leq f$.

The first step for proving the convergence of the SS sequences consists in bounding the decrease that each step obtains. From (4.4), $f(\bar{x}_+) - f(\bar{x}_i) \leq m\delta_i$; hence this boils down to bounding the predicted improvement $\delta_i$, for which one can use (5.11) and the stopping condition:

$$-\delta_i \geq \frac{1}{\mu}\alpha_i^* + D_{t_i}^*(-z_i^*) = \frac{1}{\mu}[\alpha_i^* + \mu D_{t_i}^*(-z_i^*)] > \frac{\varepsilon}{\mu}$$

(use $\mu \geq 1$ and $\alpha_i^* \geq 0$). Hence

$$f(\bar{x}_i) \leq f(\bar{x}_0) + m\left(\sum_{j<i}\delta_j\right) \leq f(\bar{x}_0) - \frac{m}{\mu}\left(\sum_{j<i}\alpha_j^* + \mu D_{t_j}^*(-z_j^*)\right). \tag{6.1}$$

Note that (6.1) holds even if $\delta_i$ in (4.4) is replaced by $\underline{\delta}_i = \delta_i + D_{t_i}(d_i^*)$, as discussed in section 5.4. Finite termination, at least, is at hand whenever $\underline{f} := \lim_{i\to\infty} f(\bar{x}_i) > -\infty$.

LEMMA 6.1. *If $\underline{f} > -\infty$ and $\varepsilon > 0$, then only finitely many iterations can be done.*

*Proof.* By Theorem 5.10, only finitely many NS can be done between two consecutive SS; from (6.1), $-\infty < \underline{f} \le f(\bar{x}_0) - m\varepsilon i/\mu$, and therefore only finitely many SS can be done. ∎

Lemma 6.1 gives no information about how "good" the obtained solution is when the algorithm stops. Without qualification, nothing can be said; if $D_t$ is nonsmooth in 0—(P5′) does not hold—the fact that 0 is optimal for (1.12) does not imply that $\bar{x}$ is optimal for (0.1); i.e., a minimum of the generalized Moreau–Yosida regularization $\phi_t$ may not be a minimum of $f$.

**6.1. Convergence under (P5′)/(P*5′).** The immediate effect of assumption (P5′)/(P*5′) is to guarantee convergence of the dual iterates to 0, provided that $\underline{f} > -\infty$ and $t$ remains bounded away from zero.

THEOREM 6.2. *If $\underline{f} > -\infty$, (P*5′) holds, $t_i \ge \underline{t} > 0$, and $\varepsilon = 0$, then $\{\alpha_i^*\} \to 0$ and $\{z_i^*\} \to 0$.*

*Proof.* Since $\varepsilon = 0$, $\{D_{t_i}^*(-z_i^*)\} \to 0$ and $\{\alpha_i^*\} \to 0$. This is guaranteed by the stopping condition if the algorithm terminates finitely, by Theorem 5.10 if finitely many SS are done, and by (6.1) and $\underline{f} > -\infty$ if infinitely many SS are done.

Under (P5′), $\{D_{t_i}^*(-z_i^*)\} \to 0$ and $t_i \ge \underline{t}$ imply that $\{z_i^*\} \to 0$; in fact, from (P*4) $\{D_{\underline{t}}^*(-z_i^*)\} \to 0$, so that from (P*2) all the $z_i^*$ belong to a compact set (a proper level set of $D_{\underline{t}}^*$) and, extracting a subsequence if necessary, $\{z_i^*\} \to z^*$. $D_{\underline{t}}^*$ is lower semicontinuous; hence

$$0 = \liminf_{i\to\infty} D_{\underline{t}}^*(-z_i^*) \ge D_{\underline{t}}^*(z^*) \ge 0.$$

Due to (P*5′), $D_{\underline{t}}^*(z^*) = 0 \Leftrightarrow z^* = 0$. ∎

The requirement on $t$ can be weakened if $D_t^*$ has the special form (3.4) (which implies (P*5′)). In fact, by (6.1) and $\underline{f} > -\infty$,

$$\infty > \frac{m}{\mu}\left(\sum_{i\to\infty} \alpha_i^* + \mu D_{t_i}^*(-z_i^*)\right) \ge m\left(\sum_{i\to\infty} t_i D^*(-z_i^*)\right).$$

Hence, we can replace $t_i \ge \underline{t}$ with the milder condition

$$\text{if } \textit{infinitely many} \text{ SS are done, then } \sum_{i\to\infty} t_i = \infty$$

and still be guaranteed that $\{D^*(-z_i^*)\} \to 0$. In turn, this implies $\{z_i^*\} \to 0$, since all the $z_i^*$ belong to a proper level set of $D^*$, which is compact by (P*2), and $D^*$ is strictly convex in 0. Note that, by Theorem 5.8, under proper conditions (3.4) allows us to drop $t_i \ge \underline{t} > 0$ for sequences of NS also.

Therefore, in the following we will assume that

$$\begin{aligned} &\textit{either } t \text{ is bounded away from zero} \\ &\textit{or } D_t^* \text{ has the form (3.4) and } \sum_{i\to\infty} t_i = \infty. \end{aligned} \tag{6.2}$$

Yet, without qualification, convergence of the dual iterates does not imply convergence of the function values; a possibility is the usual "asymptotic complementary slackness" condition.

THEOREM 6.3. *If* (P*5′) *and* (6.2) *hold,* $\varepsilon = 0$, *and* $\liminf_{i\to\infty} z_i^* \bar{x}_i = 0$, *then* $\{f(\bar{x}_i)\} \to v(\Pi)$.

*Proof.* If $\underline{f} = -\infty$, then $\{\bar{x}_i\}$ is a minimizing sequence, so assume $\underline{f} > -\infty$; the hypotheses of Theorem 6.2 are satisfied. From (2.9) and $f_\beta^* \geq f^*$,

$$z_i^* \bar{x}_i = f_i^*(z_i^*) + f(\bar{x}_i) - \alpha_i^* \geq f^*(z_i^*) + f(\bar{x}_i) - \alpha_i^*.$$

Taking the lim inf on both sides and using the hypothesis, we obtain

$$\begin{aligned} 0 = \liminf_{i\to\infty} z_i^* \bar{x}_i &\geq \liminf_{i\to\infty} [f^*(z_i^*) + f(\bar{x}_i) - \alpha_i^*] \\ &\geq \liminf_{i\to\infty} f^*(z_i^*) + \liminf_{i\to\infty} f(\bar{x}_i) + \liminf_{i\to\infty} -\alpha_i^*. \end{aligned}$$

Now, use $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$ (by Theorem 6.2), $\{f(\bar{x}_i)\} \to \underline{f}$, and the lower semicontinuity of $f^*$ to obtain $0 \geq f^*(0) + \underline{f}$, i.e., $v(\Pi) = -f^*(0) \geq \underline{f}$; since $\underline{f} \geq v(\Pi)$, the thesis is proved. □

Under (P*5′), if $\{\bar{x}_i\}$ has a cluster point $x^*$—which happens, for instance, if only finitely many serious steps are done—then $x^*$ is optimal for (0.1); in fact, Theorem 6.2 applies, and therefore $\{z_i^* \bar{x}_i\} \to 0$ as $\{z_i^*\} \to 0$. This could have been directly proved in primal notation using (2.10), the fact that $\{\alpha_i^*\} \to 0$, and [HL93b, Proposition XI.4.1.1]. Hence, the bundle algorithm converges at least if $f$ is inf-compact; however, something better can be done.

THEOREM 6.4. *If* (P*5′) *and* (6.2) *hold,* $\varepsilon = 0$, *and* $f$ *is* *-*compact, then* $\{f(\bar{x}_i)\} \to v(\Pi)$.

*Proof.* Assume by contradiction that $v(\Pi) < l = \underline{f} - \lambda$ for $\lambda > 0$, and let $\hat{x}_i$ be the projection of $\bar{x}_i$ over $S_l(f)$. Since $f(\bar{x}_i)$ is nonincreasing, $f(\bar{x}_i) \leq f(\bar{x}_0) = L \ \forall i$; therefore, by *-compactness, $\|\hat{x}_i - \bar{x}_i\| \leq e(l, L) < \infty \ \forall i$. From (2.10), $f(\bar{x}_i) \geq \underline{f}$ and the $\varepsilon$-subgradient inequality

$$\underline{f} - \lambda = f(\hat{x}_i) \geq f(\bar{x}_i) + z_i^*(\hat{x}_i - \bar{x}_i) - \alpha_i^* \geq \underline{f} - \|z_i^*\| \cdot \|\hat{x}_i - \bar{x}_i\| - \alpha_i^*$$

that yield the desired contradiction since, from Theorem 6.2, $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$. □

Theorem 6.4 in fact proves that a *-compact $f$ is *asymptotically well-behaved* (a.w.b.) [Au97]. A sequence $\{x_i\}$ is a *stationary sequence* for the function $f$ if two sequences $\{z_i\} \to 0$ and $\{\varepsilon_i\} \to 0$ exist such that $z_i$ is an $\varepsilon_i$-subgradient of $f$ at $x_i$; $f$ is a.w.b. if every stationary sequence is a minimizing sequence. In [Au97] it is proved that $f$ is a.w.b. if and only if all the following three functions

$$\begin{aligned} r(l) &= \inf_x \left\{ \inf_z \{\|z\| : z \in \partial f(x)\} : f(x) = l \right\}, \\ k(l) &= \inf_x \left\{ \inf_z \left\{ f'\left(x, \frac{z}{\|z\|}\right) : z \in \partial f(x) \right\} : f(x) = l \right\}, \\ l(l) &= \inf_x \left\{ \frac{(f(x) - l)}{d_{S_l(f)}(x)} : f(x) > l \right\} \end{aligned}$$

are strictly positive for each $l > v(\Pi)$; by Theorem 6.4, *-compactness is another sufficient condition for "well-behavedness." A result quite similar to Lemma 4.4, in

a more general setting, can be found in [Au97], but it requires *weak coercivity* of $f(0 \in ri\ dom\ f^*)$. Clearly, $^*$-compact functions need not be weakly coercive (take an affine function). On the other hand, weak coercivity ensures convergence of the primal iterates as well as of the function values [Au97, Theorem 6], and therefore it can be a convenient alternative to $^*$-compactness when stronger convergence properties are required.

Note that, even when $\{\bar{x}_i\}$ is guaranteed to be a minimizing sequence, stopping as soon as $\bar{x}_i$ is $\varepsilon$-optimal for some fixed $\varepsilon > 0$ is not straightforward. Indeed, an estimate of the quality of $\bar{x}_i$ is available only if $z_i^* = 0$, since then $\bar{x}_i$ is $\alpha_i^*$-optimal (use (2.10)). In practice, the stopping condition has to require that $z_i^*$ is "small enough"; this is the meaning of the extra stopping parameter $\mu \geq 1$. For $D_t^*$ in the special form (3.4), for instance, $\mu$ makes the stopping condition be that of $t^* = t\mu \geq t$; in our experience, guessing a value of $\mu$ that produces a true $\varepsilon$-optimal solution is usually fairly easy.

**6.2. The polyhedral case.** If $f$ is polyhedral ($\Rightarrow^*$-compact), one can prove *finite* convergence for $\varepsilon = 0$; of course, this first requires finite convergence of the 1st level. The basic result is that, at length, the primal stabilized master problem (1.8) is equivalent to its nonstabilized version; this follows from the next technical lemma.

LEMMA 6.5. *Assume that $f$ is polyhedral,* (4.1) *is satisfied, $f_\beta = \hat{f}_\beta$, and a safe $\beta$-strategy is used; for any function $h^*$ satisfying* (P$^*$1) *and* (P$^*$5$'$) *there exists a constant $\Psi_f > 0$ such that, however fixed $\beta$, if a $z \in \partial \hat{f}_\beta(x)$ exists such that $h^*(z) < \psi_f$, then $0 \in \partial \hat{f}_\beta(x)$.*

*Proof.* From (4.1) and the safe $\beta$-strategy, there is only a *finite* number of different possible $\beta$. Since each $\hat{f}_\beta$ has only a *finite* set of possible different subdifferentials [HL93a, Corollary VI.4.3.2], there is a *finite set* $\Gamma_f$ containing all possible subdifferentials of some $\hat{f}_\beta$ at some point $x$. Let $\psi(Z) = \inf_{z \in Z}\{h^*(z)\}$ ($\geq 0$ due to (P$^*$1)) and $\psi_f = \min\{\psi(Z) : Z \in \Gamma_f, \psi(Z) > 0\} > 0$; $\psi(Z) = 0$ for any Z such that $h^*(z) < \psi_f$ for some $z \in Z$. Closedness of the subdifferentials and $h^*(z) = 0 \Leftrightarrow z = 0$ (via (P$^*$5$'$)) do the rest. ∎

Note that, when $f$ itself is polyhedral, there exists one finite $\beta$ such that $f = \hat{f}_\beta$; hence, a fortiori for each $h^*$ there exists a $\psi_f > 0$ such that $z \in \partial f(x)$ and $h^*(z) < \psi_f \Rightarrow x$ is optimal for ($\Pi$).

THEOREM 6.6. *Under the hypotheses of Theorem* 5.11 *and Lemma* 6.5*, if $f$ is bounded below, $t_i = \underline{t} > 0$,* (P$^*$5$'$) *holds, $\varepsilon = 0$, and $m = 1$, then the two-level bundle algorithm finitely solves* ($\Pi$).

*Proof.* Setting $m = 1$ and $\varepsilon = 0$ is allowed by Theorem 5.11; after finitely many consecutive NS, either the algorithm stops or $\hat{f}_i(\bar{x}_i + d_i^*) = f(\bar{x}_i + d_i^*)$ and an SS is done. If the algorithm stops, by $\varepsilon = 0$ one has $\alpha_i^* = 0$ and, from (P5$'$), $z_i^* = 0$; therefore, $\bar{x}_i$ is optimal (cf. (2.10)). Hence, assume by contradiction that infinitely many SS are done; by (6.1) and the boundedness of $f$, as in the proof of Theorem 6.2, we get $\{D_{\underline{t}}^*(-z_i^*)\} \to 0$. Since $z_i^* \in \partial \hat{f}_i(\bar{x}_i + d_i^*)$, applying Lemma 6.5 with $h^* = D_{\underline{t}}^*$ shows that, for large enough $i$, $0 \in \partial \hat{f}_i(\bar{x}_i + d_i^*)$; i.e., $\bar{x}_i + d_i^*$ is a minimum of $\hat{f}_i$. Hence, at length every $f(\bar{x}_i)$ is a minimum of some $\hat{f}_\beta$; but from the hypotheses there are only finitely many different sets $\beta$, which contradicts $f(\bar{x}_+) > f(\bar{x}_i)$. ∎

Note that, as for Theorem 6.2, the requirement over $t$ can be weakened if $D_{\underline{t}}^*$ has the form (3.4).

Let us mention that setting $m = 1$ all along is only the simplest possibility; what is really required is that only finitely many "inexact" SS (with $\Delta f > 0$) be performed between two "exact" SS (with $\Delta f = 0$). Hence, $m$ can be reset to any value $< 1$ after

each exact SS, provided that it is set to 1 after finitely many consecutive (inexact) SS.

**7. Convergence of the 3rd level.** If (P*5′) does not hold, convergence requires $t \to \infty$ and therefore the "three-level" bundle algorithm of Figure 2. Hence, let us once again change our notation: from now on, the index $i$ refers to the end of the $i$th call to the algorithm of Figure 1, with $\underline{t} = \underline{t}_i$ and $\varepsilon = \varepsilon_i > 0$, from within the cycle of the "three-level" bundle algorithm. Therefore, the standing assumption is now

> conditions sufficient to guarantee finite termination
>
> of the two-level bundle algorithm hold.

We also assume that $\{\underline{t}_i\} \to \infty$ and $\{\varepsilon_i\} \to 0$.

**7.1. Primal convergence.** It is instructive to compare Lemma 5.4 with Theorem 6.4. In the former—where $\bar{x}$ is fixed—the optimal values of (1.12) converge to that of (0.1) without the *-compactness assumption, while in the latter—where SS are allowed—it is required. The same happens with the bundle algorithm.

THEOREM 7.1. *If $f$ is *-compact, then* $\lim_{i\to\infty} f(\bar{x}_i) = v(\Pi)$.

*Proof.* Since $t_i \geq \underline{t}_i$, $\{t_i\} \to \infty$. The stopping condition implies $v(\Delta_i) \leq \varepsilon_i - f(\bar{x}_i)$, i.e., $v(\Pi_i) + \varepsilon_i \geq f(\bar{x}_i)$, and since $v(\Pi_{\bar{x}_i,t_i}) \geq v(\Pi_i)$, we obtain $v(\Pi) \leq f(\bar{x}_i) \leq v(\Pi_{\bar{x}_i,t_i}) + \varepsilon_i$; now apply Lemma 4.4. □

Note that *-compactness is used in Lemma 4.4 ⇒ Theorem 7.1 without any reference to a stationary sequence; hence, unlike Theorem 6.4, a.w.b.-ness could not be used here. Furthermore, $\{t_i\} \to \infty$ is required in order to solve (Π) with "infinite accuracy"; a suitably large $t$ suffices for obtaining any finite accuracy (of course, $f$ must be bounded below). In fact, using Lemma 4.3, it is easy to show that, if $f$ is bounded below, then for any starting point $\bar{x}_0$ and any fixed $\varepsilon > 0$ there exists a $\bar{t}$ such that $v(\Pi_{\bar{x}_i,\bar{t}}) \leq v(\Pi) + \varepsilon$ (use $f(\bar{x}_i) \leq f(\bar{x}_0)$). Given a suitable estimate of $\bar{t}$, the two-level bundle algorithm can directly solve (Π) with any finite accuracy.

Eliminating the *-compactness assumption is possible, at the cost of inhibiting—at length—the serious steps, i.e., using rule (4.iii′). In this case, $t$ needs to go all the way up to $\infty$.

THEOREM 7.2. *With rule* (4.iii′) *in force,* $\liminf_{i\to\infty} f(\bar{x}_i + d_i^*) = v(\Pi)$.

*Proof.* Wait for the last SS to be performed, and call $\bar{x}(= \bar{x}_i)$ the fixed current point. Assume by contradiction that $\liminf_{i\to\infty} f(\bar{x} + d_i^*) - 2\delta > v(\Pi)$ for some $\delta > 0$; hence, there exists a $\bar{d}$ such that $f(\bar{x} + \bar{d}) \leq f(\bar{x} + d_i^*) - 2\delta \ \forall i$. Due to (P5) and $\{t_i\} \to \infty$, $D_{t_i}(\bar{d}) \leq \delta$ for a large enough $i$; therefore

$$v(\Pi_{\bar{x},t_i}) \leq f(\bar{x} + \bar{d}) + D_{t_i}(\bar{d}) \leq f(\bar{x} + d_i^*) - \delta \leq f(\bar{x} + d_i^*) + D_{t_i}(d_i^*) - \delta.$$

When the inner loop terminates, $\Delta f_i \leq \varepsilon_i$; hence, using (2.6) and $v(\Pi_{\bar{x},t_i}) \geq v(\Pi_i)$,

$$\varepsilon_i + v(\Pi_{\bar{x},t_i}) \geq \varepsilon_i + v(\Pi_i) \geq \Delta f_i + v(\Pi_i) = f(\bar{x} + d_i^*) + D_{t_i}(d_i^*),$$

which leads to $\varepsilon_i \geq f(\bar{x} + d_i^*) + D_{t_i}(d_i^*) - v(\Pi_{\bar{x},t_i}) \geq \delta$, contradicting $\{\varepsilon_i\} \to 0$. □

**7.2. Dual convergence.** From the dual viewpoint, $\{\bar{x}_i + d_i^*\}$ is a maximizing sequence for the Lagrangian dual of (1.2) w.r.t. the constraints $z = 0$ (cf. section 1), and $\{z_i\}$ are the optimal solutions of the corresponding dual pricing problems (1.3), with $\bar{x} = \bar{x}_i + d_i^*$. Further, from $f^* \leq f_i^*$ and (2.8), the alternative stopping condition of (4.iii′) ($\Delta f_i \leq \varepsilon_i$) gives

$$f^*(z_i^*) - z_i^*(\bar{x}_i + d_i^*) \leq f_i^*(z_i^*) - z_i^*(\bar{x}_i + d_i^*) \leq f^*(z_i) - z_i(\bar{x}_i + d_i^*) + \varepsilon_i,$$

i.e., $z_i^*$ is an $\varepsilon_i$-optimal solution for (1.3) with $\bar{x} = \bar{x}_i + d_i^*$. Using (1.v) in the above relation gives

$$z_i^* \in \partial_{\varepsilon_i} f(\bar{x} + d_i^*),$$

i.e., $\varepsilon_i$-optimal solutions for (1.3) are $\varepsilon_i$-*subgradients* of $f$; this is of particular interest when $f$ itself is a dual function (cf. section 9). Thus, if $\{\bar{x}_i + d_i^*\} \to x^*$ and $\{z_i^*\} \to z^*$, then $z^* \in \partial f(x^*)$ [HL93b, Proposition XI.4.1.1]; one would like to show that $\{z_i^*\} \to 0$ whenever $f$ is bounded. This is possible, and it does not require $^*$-compactness.

THEOREM 7.3. *If $f$ is bounded below and rule* (4.iii′) *is in force, then* $\{z_i^*\} \to 0$.

*Proof.* Using (2.1) and $D_{t_i} \geq 0$, we obtain

$$-v(\Delta_i) = v(\Pi_i) = f_i(\bar{x}_i + d_i^*) + D_{t_i}(-z_i^*) \geq f_i(\bar{x}_i + d_i^*).$$

Using the previous relation with the stopping condition of (4.iii′), $\Delta f_i = f(\bar{x}_i + d_i^*) - f_i(\bar{x}_i + d_i^*) \leq \varepsilon_i$, gives, together with boundedness of $f$ and monotonicity of $\{\varepsilon_i\}$,

$$v(\Delta_i) \leq -f_i(\bar{x}_i + d_i^*) \leq \varepsilon_i - f(\bar{x}_i + d_i^*) \leq \varepsilon_0 - v(\Pi) < \infty.$$

Now, using (2.9) and $f_i^* \geq f^*$, one obtains

$$\infty > v(\Delta_i) = f_i^*(z_i^*) - z_i^* \bar{x}_i + D_{t_i}^*(-z_i^*) \geq D_{t_i}^*(-z_i^*) - f(\bar{x}_i).$$

By rule (4.iii′), only finitely many serious steps are done, hence at length, $f(\bar{x}_i) = f(\bar{x})$ for a fixed $\bar{x}$; by (P*5), $\|z_i^*\|_2 \geq \varepsilon > 0$ for infinitely many $i$ and $\{t_i\} \to \infty$ imply $\{D_{t_i}^*(-z_i^*)\} \to \infty$. □

Note that, if $f$ is bounded below, a dual proof of Theorem 7.1 exists, using Theorem 7.3 ($\{z_i^*\} \to 0$) and the fact that $\{\bar{x}_i\}$ is a stationary sequence; however, the case of $f$ unbounded below would need a separate treatment (a.w.b.-ness is tailored over bounded functions with unbounded level sets).

**7.3. The polyhedral case.** The three-level bundle method allows us to drop assumption (P5′) from the finite termination proofs in the polyhedral ($\Rightarrow^*$-compact) case. Indeed, for bounded polyhedral functions one can prove the following strengthened form of Lemma 4.4, where $z_t$ and $d_t$ denote the optimal solutions of $(\Delta_{\bar{x},t})$ and $(\Pi_{\bar{x},t})$, respectively.

LEMMA 7.4. *If $f$ is polyhedral and bounded below, then for each $L < \infty$ there exists a $\underline{t} > 0$ such that $\bar{x} + d_t$ is an optimal solution of* (Π) $\forall t > \underline{t}$ *and $\bar{x}$ such that* $f(\bar{x}) \leq L$.

*Proof.* Fix any $\bar{x}$ such that $f(\bar{x}) \leq L$; it is easy to show, mirroring Theorem 7.3, that $\{z_t\} \to 0$ for $t \to \infty$ (use $D_t^*(-z_t) - L \leq D_t^*(-z_t) - f(\bar{x}) \leq f^*(z_t) - z_t\bar{x} + D_t^*(-z_t) = v(\Delta_{\bar{x},t}) \leq -v(\Pi) < \infty$ and (P*5)). Then, using $z_t \in \partial f(\bar{x} + d_t)$ (cf. (2.3)) and Lemma 6.5 with $h^* = \|\cdot\|$, we obtain that, for large enough $t$, $0 \in \partial f(\bar{x} + d_t)$; i.e., $\bar{x} + d_t$ is a minimum of $f$. □

This result allows us to derive a finite convergence proof; since $f$ is polyhedral, we can directly fix $\varepsilon_i = 0$ and use rule (4.iii″).

THEOREM 7.5. *Under the hypotheses of Theorem* 5.11 *and Lemma* 6.5, *if $\varepsilon_i = 0$ $\forall i$ and rule* (4.iii″) *is in force, then the three-level bundle algorithm finitely solves* (Π).

*Proof.* From Theorem 5.11, we know that only finitely many consecutive NS can be done: either the normal stopping rule fires or an SS is performed. However, from rule (4.iii″), only finitely many SS can be done; hence, either the stopping rule fires, or a sequence of consecutive NS is started. Theorem 5.11 tells us that such a sequence

finitely produces $\Delta f = 0$; hence the two-level bundle algorithm finitely terminates with either $\alpha^* + D_t^*(-z^*) = 0$ or $\Delta f = 0$.

If $\alpha_i^* + D_{t_i}^*(-z_i^*) = 0$ happens infinitely many times, $\alpha_i^* = 0$ and (2.10) tell us that $z_i^* \in \partial f(\bar{x}_i)$. Theorem 7.3 shows that $\|z_i^*\| \to 0$ as $\{t_i\} \to \infty$; hence, applying Lemma 6.5 with $h^* = \|\cdot\|$ shows that, for large enough $i$, $0 \in \partial f(\bar{x}_i)$, i.e., $\bar{x}_i$ is optimal for ($\Pi$). If $\Delta f_i = 0$ happens infinitely many times, recall from (2.7) that this means that $d_i^*$ is optimal for $(\Pi_{\bar{x}_i,t_i})$ and use Lemma 7.4. □

Let us remark that the three-level bundle algorithm applied to a polyhedral $f$ lacks a convenient stopping criterion; either $\bar{x}_i$ or $\bar{x}_i + d_i^*$ at some point becomes optimal, but there is no easy way to tell when this happens. In order to be able to stop, either the solver of $(\Delta_{\beta,\bar{x},t})$ should always return $z^* = 0$ whenever it can, or an estimate of $\underline{t}$ of Lemma 7.4 should be available.

**8. Extensions.** The generalized bundle algorithm presented in the previous paragraphs can incorporate a number of important algorithmic variants. For instance, (4.iv) allows us to seamlessly add a line search on $d^*$, provided only that, at length, the unit step is always probed. (4.i)–(4.iii) allow us to adapt the curved search approach of [SZ92] to our more general setting; other $t$-strategies, originally devised for $D_t = \frac{1}{2t}\|\cdot\|_2^2$, can be adapted as well [Fr97, section I.5]. Multiple [$\varepsilon$-]subgradients can be added to $\beta$ at each call of the oracle if the latter is—as happens in some applications—capable of providing them. Finally, it should not be hard to extend the proofs of convergence to the case in which $f$ is not computed exactly, following what is done in [GV97, Ki99]. More complex extensions are discussed in the following section.

**8.1. The constrained case.** Generalized bundle methods can cope with constraints $x \in X$ if $X$ is a closed convex set. Basically, all that is needed is to insert full knowledge about $X$ into (1.8), i.e., to solve at each iteration

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \{f_\beta(\bar{x}+d) + D_t(d) : (\bar{x}+d) \in X\}. \tag{8.1}$$

Problem (8.1) can be viewed as (1.8) using the *restricted model* $f_{X,\beta} = f_\beta + I_X$, which is a model of the actual function to be minimized, the *restricted function* $f_X = f + I_X$. Under the natural assumption that $\bar{x} \in dom\ f_\beta \cap X$, the dual of (8.1) is just (1.9) with $f_\beta^*$ replaced by

$$f_{X,\beta}^*(z) = (f_\beta + I_X)^*(z) = \inf_w \{f_\beta^*(z-w) + \sigma_X(w)\} \tag{8.2}$$

(see [HL93b, Theorem X.2.3.1]) as $(I_X)^* = \sigma_X$. The problem can be written in a "direct" form, avoiding the complicated-looking infimal convolution (8.2), by means of the simple variable change $z = \bar{z} + w$:

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_{\bar{z},w} \{f_\beta^*(\bar{z}) + \sigma_X(w) - \bar{x}(\bar{z}+w) + D_t^*(-\bar{z}-w)\}. \tag{8.3}$$

The extension of the theory is not completely straightforward: $f_X$ is *not* finite everywhere, and $f_{X,\beta}$ is a model of $f_X$ rather than of $f$. Hence, (2.3)/(2.4) are valid with $f_{X,\beta}$ replacing $f_\beta$; in particular, we have that $z^* \in \partial f_{X,\beta}(\bar{x}+d^*)$. On the other hand, the black box produces subgradients of $f$ rather than of $f_X$, i.e., $z \in \partial f(\bar{x}+d^*)(\bar{x}+d^* \in X)$; there is an "asymmetry" that has to be taken into account.

The main observation is that some of the properties of $z^*$ have now to be referred to $\bar{z}^*$. In fact, from $f^*_{X,\beta}(z^*) = f^*_\beta(\bar{z}^*) + \sigma_X(w^*)$ and $z^* \in \partial f_{X,\beta}(\bar{x}+d^*)$, using (1.vi), we obtain

$$[f^*_\beta(\bar{z}^*) - \bar{z}^*(\bar{x}+d^*) + f_\beta(\bar{x}+d^*)] + [\sigma_X(w^*) - w^*(\bar{x}+d^*) + I_X(\bar{x}+d^*)] = 0.$$

By (1.vii) both quantities in square brackets are nonnegative, and therefore both must be zero; hence, by (1.vi) we get $\bar{z}^* \in \partial f_\beta(\bar{x}+d^*)$ (and $w^* \in \partial I_X(\bar{x}+d^*)$). Thus, in the constrained case one has to carefully distinguish $\bar{z}^*$ from $z^*$. For instance, when aggregation is done, it is $\bar{z}^*$, together with its $f^*$-value $f^*_i(\bar{z}^*_i)$, that is added to $\beta$ instead of $z^*$; the inequality in Definition 5.6 becomes

$$f^*_i(\bar{z}^*_i) \geq f^*_+(\bar{z}^*_i). \tag{8.4}$$

In this setting, Lemma 5.2 proves that $\{\bar{z}^*_i\}$, rather than $\{z^*_i\}$, is bounded; however, Lemma 5.1 and Lemmas 5.3–5.6 do not change. The boundedness of $\{\bar{z}^*_i\}$ also implies that of $\{z^*_i\}$ under certain assumptions, as the following lemma shows.

LEMMA 8.1. *If* (P*3) *holds,* $D^*_t$ *has the form* (3.4)*, and* $\{\bar{z}^*_i\}$ *is bounded, then* $\{z^*_i\}$ *is bounded.*

*Proof.* Since $(\bar{z}^*_i, w^*_i)$ is the optimal solution of (8.3) and $\sigma_X(w^*_i) - \bar{x}w^*_i \geq 0$ as $\bar{x} \in X$, we have

$$\begin{aligned} &f^*_i(\bar{z}^*_i) - \bar{x}\bar{z}^*_i + D^*_i(-\bar{z}^*_i - w^*_i) \\ &\quad \leq f^*_i(\bar{z}^*_i) + \sigma_X(w^*_i) - \bar{x}(\bar{z}^*_i + w^*_i) + D^*_i(-\bar{z}^*_i - w^*_i) \leq f^*_i(\bar{z}^*_i) - \bar{x}\bar{z}^*_i + D^*_i(-\bar{z}^*_i), \end{aligned}$$

and therefore

$$D^*_i(-\bar{z}^*_i - w^*_i) \leq D^*_i(-\bar{z}^*_i). \tag{8.5}$$

Since $D^*_t$ has the form (3.4), we can divide both sides of (8.5) by $t_i$ to obtain

$$D^*(-\bar{z}^*_i - w^*_i) \leq D^*(-\bar{z}^*_i).$$

Since, by (P*3), $D^*$ is finite everywhere and $\{\bar{z}^*_i\}$ is bounded, the left-hand side is finite; therefore, all $z^*_i = -\bar{z}^*_i - w^*_i$ belong to a level set of $D^*$, which is compact by (P*2). □

In order to extend the proof of Theorem 5.7, "asymmetric" definitions of $h_i$ and $r_i$,

$$\begin{aligned} h_i(\gamma) &:= [f^*_{X,i}(z^*_i)(1-\gamma) + f^*(z_i)\gamma - z_i(\gamma)\bar{x} + f(\bar{x})] + D^*_{t_i}(-z_i(\gamma)), \\ r_i(z) &:= f^*_{X,i}(z) - z\bar{x} + f(\bar{x}) = f^*_{X,i}(z) - z\bar{x} + f_X(\bar{x}) \geq 0, \end{aligned}$$

are required. Using (8.4), one obtains

$$f^*_{X,i}(z^*_i) = f^*_i(\bar{z}^*_i) + \sigma_X(w^*_i) \geq f^*_+(\bar{z}^*_i) + \sigma_X(w^*_i) \geq \inf_w \{f^*_+(z^*_i - w) + \sigma_X(w)\} = f^*_{X,+}(z^*_i),$$

while from (4.iv) and $\sigma_X(0) = 0$,

$$f^*(z_i) \geq f^*_+(z_i) = f^*_+(z_i) + \sigma_X(0) \geq \inf_w \{f^*_+(z_i - w) + \sigma_X(w)\} = f^*_{X,+}(z_i);$$

now, proceeding as in section 5.2 $v(\vartheta_i) = v(\Delta_+) + f(\bar{x})$ is readily obtained. Furthermore, (2.6)/(2.8) can be written (in an asymmetric fashion) as

$$\Delta f = f(\bar{x}+d^*) - f_{X,\beta}(\bar{x}+d^*) = f^*_{X,\beta}(z^*) - f^*(z) + (\bar{x}+d^*)(z-z^*), \tag{8.6}$$

which easily gives the equivalent to (5.4),

$$h_i'(0) = -\Delta f_i < -(1-m)[f_{X,i}^*(z_i^*) - z_i^*\bar{x} + D_{t_i}^*(-z_i^*) + f(\bar{x})] = -(1-m)h_i(0),$$

which allows us immediately to extend the proofs of Theorems 5.7 and 5.8 to the constrained case. Note that $D_t^* = \frac{1}{2}t\|\cdot\|_2^2$ has *both* the forms (3.3) and (3.4); therefore, exploiting Lemma 8.1, our convergence results for $f_\beta = \hat{f}_\beta$ generalize the best ones known for the proximal bundle case.

The only difficulty in extending the proof of Theorem 5.9 comes from the fact that (5.7) does *not* guarantee $f_{X,i}^*(z_h) = f_X^*(z_h)\ \forall i \geq h$. However, $f_{X,i} \geq f_i$ and (5.7) give $f_{X,i}^*(z_h) \leq f_i^*(z_h) = f^*(z_h)$; thus, operating as in Theorem 5.9, one obtains the equivalent to (5.8):

$$f^*(z_h) - (\bar{x} + d_i^*)z_h \geq f_{X,i}^*(z_i^*) - (\bar{x} + d_i^*)z_i^* \qquad \forall i > h.$$

Combined with the "asymmetric" definition (8.6) of $\Delta f_i$ (with $z = z_i$), this gives (5.10). All the other results in section 5 plainly extend to the constrained case.

It is then easy to check that almost all other results in section 6 and section 7 remain valid, with the only provision being that we look at $f_X$, rather than at $f$, as the actual function to be minimized. In particular, note that, by (8.1), $\bar{x} + d^*$ is always feasible and rule (4.iv) can be satisfied. The only exceptions are the results about polyhedral functions, which also require $X$ to be a *polyhedral* set. In fact, it is easy to prove that Lemma 6.5 fails if $X$ is not polyhedral, as $f_X$ may have infinitely many different subdifferentials (take $f$ affine and $X = B_2(\delta)$). However, if $f$ satisfies condition (4.1) and $X$ is polyhedral, then $f_X$ has finitely many different subdifferentials; this allows us to extend Lemma 6.5 and all the subsequent results.

Finally, let us mention that, when $X$ is a polyhedron $Hx \leq h$, (8.3) boils down to

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_{z,\omega}\{f_\beta^*(z) + \omega h - \bar{x}(z + \omega H) + D_t^*(-z - \omega H) : \omega \geq 0\}$$

($\omega$ being the "dual" variables). In this case, it is not even required that all the defining inequalities of $X$ be known in advance; when an unfeasible $x$ is probed, the black box should just return $+\infty$ and some "extremal" violated inequality. (Assumption (4.1) on the black box must be satisfied.) Clearly, only finitely many steps are required to eventually acquire a complete description of $X$.

**8.2. Decomposable functions.** Another important extension is a different treatment of *decomposable* functions,

$$f(x) = \sum_{h \in K} f^h(x),$$

where $1 < |K| = k < \infty$; examples are cost-decomposition approaches to block-structured convex problems [PZ92, GK95, CFG01]. Here, the computation of each $f^h(\bar{x})$ gives a $z^h \in \partial f^h(\bar{x})$; rather than aggregating this information into the unique $z = \sum_{h \in K} z^h$, one may keep it in a disaggregated form [Ki95, GV97], where $\beta$ is partitioned into $k$ disjoint subsets $\beta^h$ and there is one model $f_\beta^h$ for each $f^h$. The

disaggregated subproblems

$$(\Pi_{\beta,\bar{x},t}) \qquad \inf_d \left\{ \sum_{h\in K} f^h_\beta(\bar{x}+d) + D_t(d) \right\},$$

$$(\Delta_{\beta,\bar{x},t}) \qquad \inf_z \left\{ \sum_{h\in K} (f^h_\beta)^*(z^h) - \left( \sum_{h\in K} z^h \right) \bar{x} + D^*_t \left( - \sum_{h\in K} z^h \right) \right\}$$

are then solved instead of the aggregated versions. Using the disaggregated model $f^K_\beta = \sum_{h\in K} f^h_\beta$ is well known to be potentially beneficial: in the polyhedral case, for instance, $\hat{f}^K_\beta$ is a (much) better description of $f$ than the ordinary aggregate model $\hat{f}_\beta$.

It is easy to show that the "critical" properties are inherited by the disaggregated model $f^K_\beta$ if they hold for all the $f^h_\beta$ individually. For (4.iv), for instance, one has that $(f^h_+)^*(z^h) \leq (f^h)^*(z^h)\ \forall h$ implies

$$f^*(z) = \sum_{h\in K} (f^h)^*(z^h) \geq \sum_{h\in K} (f^h_+)^*(z^h) \geq \inf_{\bar{z}} \left\{ \sum_{h\in K} (f^h_+)^*(\bar{z}^h) : \sum_{h\in K} \bar{z}^h = z \right\} = (f^K_+)^*(z).$$

Analogously, it is possible to show that if (4.7)/(5.7) hold for all the $f^h_\beta$, then they hold for $f^K_\beta$. Thus, the analysis of the previous paragraphs immediately extends to the "disaggregated" variant of generalized bundle methods, independently on the stabilizing term $D_t$. Of course, these results can be used together with those of section 8.1 to construct a disaggregated constrained generalized bundle method.

**9. Comparisons.** A number of algorithms that have been proposed in the literature can be shown to be special cases of, or closely related to, the generalized bundle algorithm.

**9.1. Other bundle approaches.** The algorithm in Figure 1 covers the proximal bundle method [HL93b, Algorithm XV.3.3.4], where $D_t = \frac{1}{2t}\|\cdot\|^2_2$ and $D^*_t = \frac{1}{2}t\|\cdot\|^2_2$. A dual interpretation of this method is well known [HL93b, section XV.2.4]: (1.9) is a Lagrangian relaxation of the problem of finding the *steepest $\varepsilon$-descent direction* for $\hat{f}_\beta$ in $\bar{x}$. Historically, this dual interpretation motivated the development of the first bundle methods; however, it has drawbacks in that (1.9) (resp., (1.11)) is described in terms of a "local" object, the $\varepsilon$-subdifferential of $f_\beta$ (resp., $f$) in $\bar{x}$, so that it is difficult to relate two problems corresponding to different current points. Conceptual descent methods have been proposed, based on this dual interpretation, where the $L_2$-norm in the dual is replaced with any norm $|||\cdot|||$ (see [HL93a, Algorithm VIII.2.1.5]); however, this does not readily extend to other forms of bundle methods, where $D_t$ is

- $\frac{1}{t}\|\cdot\|_p$ for $p \geq 1$ (in practice, the $L_1$- and $L_\infty$-norms) [KCL95];
- $\frac{1}{t}h(|||d|||)$, where $|||\cdot|||$ is any norm and $h$ is a convex continuous and differentiable function with invertible derivative such that $h(0) = h'(0) = 0$ [Be96];
- $\frac{1}{t}D(d)$ for $D$ strictly convex, strongly coercive, differentiable, and finite everywhere [Au87];
- the *indicator function* of the ball of radius $t$ under some norm $|||\cdot|||$; this amounts to restricting the next trial point inside a *trust region* [HL93b, Algorithm XV.2.1.1].

It is easy to see that conditions (P1)–(P5) are less restrictive than all those above. Remarkably, the convergence proofs for the first three cases, where $D_t(0) = 0 \Leftrightarrow d = 0$, are quite different from those used in the fourth case, where $D_t(d) = 0$ in some ball around the origin. Our analysis is the first that covers both situations in a uniform way. Furthermore, our analysis is the first that fully exploits duality. In [Be96] it was noted that using a norm $|||\cdot|||$ in the primal leads to some dual problem involving the *conjugate norm* $|||\cdot|||^*$, much in the spirit of [HL93a, Algorithm VIII.2.1.5], but this was not extended to a dual interpretation of the algorithm. In [Au87], (1.9) is only used to prove [Au87, Theorem 2.3]. In other cases duality was completely overlooked, even when linear duality could have been used [KCL95]. A first step towards this development was done in [Fr97], where $D_t^* = \frac{1}{t}\|\cdot\|_p$ with $p \in \{1, 2, \infty\}$ was studied; due to the interpretation of (1.9) in terms of $\varepsilon$-subgradients, those bundle variants had an interest on their own, as a bundle algorithm with a dual trust region was one of the open questions in [HL93b, Remark XV.2.5.1].

Other approaches directly related to generalized bundle methods are proximal-type algorithms; there, the stabilized problem (1.12) is solved with a "nonuniform" stabilizing term, which depends on $\bar{x}$ as well as on $t$. This is used to incorporate constraints in the stabilizing term, which also serves as a barrier function to keep the iterates feasible. Stabilizing terms studied in the literature are either *D-functions* [Ec93, CT93],

$$D_{\bar{x},t}(d) = \frac{1}{t}(\psi(\bar{x}+d) - \psi(\bar{x}) - \nabla\psi(\bar{x})d),$$

where $\psi$ is a fixed strictly convex and differentiable function such that the level sets of $D_{\bar{x},t}$ are compact, or *$\varphi$-divergences* [IST94, IT95, Te97],

$$D_{\bar{x},t}(d) = \frac{1}{t}\sum_{i=1,\dots,n} \bar{x}_i\varphi\left(\frac{\bar{x}_i + d_i}{\bar{x}_i}\right),$$

where $\varphi$ is a fixed univariate function that is (among other things) continuously differentiable, strictly convex, and such that $\varphi(1) = \varphi'(1) = 0$. These stabilizing terms satisfy (P1), (P4), and (P5), and they have bounded level sets [IST94] which contain 0 in the interior if $\bar{x}$ lies in the zone of $D_{\bar{x},t}$ (int *dom* $\psi$ in the first case and $\Re^n_{++}$ in the second), where proximal-type algorithms work. Conditions parallel to (P3) and (P3′) are also required: boundedness of $f$, that corresponds to (P3′), is widely used, but in [CT93] the requirement is rather *im* $\nabla\psi = \Re^n$, i.e., *dom* $\psi^* = \Re^n$, i.e., (P*3) as

$$D^*_{\bar{x},t}(z) = \frac{1}{t}\psi^*(tz + \nabla\psi(\bar{x})) - \bar{x}(tz + \nabla\psi(\bar{x})) + \psi(\bar{x}).$$

In both cases, $D_{\bar{x},t}$ is differentiable and $D_{\bar{x},t}(d) = 0 \Leftrightarrow d = 0$; this is not required in our approach, even though both differentiability (in 0) and strict convexity help to enhance (different parts of) the convergence proofs. Also, all of the above methods require the exact solution of (1.12), which is a rather strong condition. Finally, our dual viewpoint extends the one that has been developed for proximal-type algorithms, which is limited to the case in which (0.1) is itself a Lagrangian dual (cf. section 9.2).

The differentiability of $D_{\bar{x},t}$, but not strict convexity, is dropped in [Ki98], where *B-functions* are introduced; there, the compactness requirement is also different. (There is no need for "local" compactness, as the solution of (1.12) is assumed to be given.) An implementable version of the proximal method using B-functions, the

*bundle Bregman proximal method*, is then proposed in [Ki99]. The analysis provides strong convergence results, for instance allowing inexact solution of the stabilized master problem and avoiding the $^*$-compactness assumption. However, it does not subsume the results of the present article, which do not require the stabilizing term to be a B-function. Furthermore, our "more technical" (cf. [CL93, Remark 4.6]) dual proof of Theorem 5.7 provides estimates on the rate of convergence during NS sequences, and we don't require $f_\beta$ to be the cutting plane model, thereby allowing easy extensions, e.g., to the disaggregated case (cf. section 8.2).

Finally, a related but different approach can be found in [Nu97]. There, the dual object is the graph of the $\varepsilon \to \partial_\varepsilon f(0)$ mapping, which is equivalent (modulo a rotation) to *epi* $f^*$. The approach in [Nu97] can be summarized, in our notation, as follows: at each step $i$, find a separating hyperplane between *epi* $\hat{f}_i^*$ and the point $(-\underline{f}_i, 0)$, where $\underline{f}_i$ is the best $f$-value found so far. The hyperplane must be nonvertical, i.e., in the form $(1, -x_i)$; it is easy to check that $(1, -x_i)$ is a separating hyperplane if and only if $\hat{f}_i(x_i) \leq \underline{f}_i$. Condition (P$^*$3$'$) is required in order to ensure that $\hat{f}_i^*(0) < \infty$. Not all choices of separating hyperplanes give a convergent algorithm; in [Nu97], an abstract rule is given, and an implementation is proposed under the form of the min-problem

$$\inf_{\sigma,z}\{|||(-\underline{f}_i, 0) - (\sigma, z)||| : (\sigma, z) \in \mathit{epi}\ \hat{f}_i^*\}, \tag{9.1}$$

where $|||\cdot|||$ is any norm whose dual optimal solution provides $x_i$. Problem (9.1) is clearly related to (1.9) (cf. [Fr98]), but with a decidedly different flavor. On one hand, in (9.1) the cost function of the $\hat{f}^*$-values need not be linear, but, on the other hand, $D_t^*$ in (1.9) need not be norm-like. Furthermore, the treatment in [Nu97] ignores the concept of current point and the updating of the proximal parameter $t$.

To conclude this section, let us mention that there are important classes of bundle methods that are *not* covered by our analysis: such are *proximal level* methods [LNN95], [HL93b, Algorithm XV.2.3.1], *analytic center cutting plane* methods [Ne95, GV97], *dual $\varepsilon$-descent algorithms* [HL93b, Algorithm XIV.3.4.2], algorithms based on a biobjective view of the direction finding problem [Fu98], and Newton-type bundle methods [LS98, LV98, MSQ98]. The extension of our theory to some of the above algorithms might be possible and is currently under research.

**9.2. Algorithms for structured convex problems.** It is well known that, under proper assumptions [HL93b, Chap. XII], the convex problem

$$(\mathrm{P}) \qquad \sup_u\{c(u) : h(u) = 0, u \in U\} \tag{9.2}$$

is equivalent to its Lagrangian dual (0.1), where

$$(\mathrm{D}_{\bar{x}}) \qquad f(\bar{x}) = \sup_u\{c(u) + \bar{x}h(u) : u \in U\}. \tag{9.3}$$

Here,

$$f^*(z) = \sup_x \left\{ -\sup_u\{c(u) + x(h(u) - z) : u \in U\} \right\} \tag{9.4}$$

$$= -\sup_u\{c(u) : h(u) = z, u \in U\}. \tag{9.5}$$

((9.4) is the Lagrangian dual of (9.5), whence the identity.) Thus, $-f^*$ is the *value function* of (9.2) w.r.t. the constraints $h(u)$; plugging (9.5) into (1.11), one obtains

$$(\mathrm{D}_{\bar{x},t}) \qquad \sup_u\{c(u) + \bar{x}h(u) - D_t^*(-h(u)) : u \in U\}. \tag{9.6}$$

Hence, generalized bundle methods applied to a Lagrangian dual are approximated generalized augmented Lagrangian approaches to the solution of (9.2). If $c$ and $h$ are affine, and the cutting plane model $\hat{f}_\beta$ is used, in view of (1.7) the stabilized dual master problem (1.9) becomes

(9.7)

$$(\mathrm{D}_{\beta,\bar{x},t}) \qquad \inf_z \left\{ \inf_\theta \left\{ \sum_{u\in\beta} -c(u)\theta_u : \sum_{u\in\beta} h(u)\theta_u = z, \theta \in \Theta \right\} - z\bar{x} + D_t^*(-z) \right\}$$
$$= \sup_u \{ c(u) + \bar{x}h(u) - D_t^*(-h(u)) : u \in Conv(\beta) = U_\beta \},$$

where $\beta$ is now considered a set of optimal solutions $u_i \in U$ of the dual pricing problem (1.3) such that $z_i = h(u_i)$. Thus, the generalized bundle method uses an *inner linearization* approach, where $U$ is substituted with its inner linearization $U_\beta$, to approximately solve $(\mathrm{D}_{\bar{x},t})$. In fact, let $u^*$ be the optimal solution of (9.7); from (4.7), the sequence $\{z_i^* = h(u_i^*)\}$ of optimal solutions of (1.9) corresponds to a sequence $\{u_i^*\}$ of $\alpha_i^*$-*optimal* solutions for (9.3) (cf. section 6.1). If $\{z_i^*\} \to 0$ and $\{\alpha_i^*\} \to 0$, any cluster point of $\{u_i^*\}$ is optimal for (9.2). Similar results hold for inequality constraints $h(u) \le 0$.

Hence, generalized bundle methods are related to nonquadratic penalty methods. For instance, in [PZ92, PZ94], (9.6) is considered with $\bar{x} = 0$ and $D_t^*(z) = t\sum_i \Phi_\varepsilon^*(z_i) \Rightarrow D_t(z) = t\sum_i \Phi_\varepsilon(\frac{1}{t}d_i)$ for some $\varepsilon > 0$ and

$$\Phi_\varepsilon^*(z_i) = \begin{cases} \dfrac{z_i^2}{2\varepsilon} & \text{if } -\varepsilon \le z_i \le \varepsilon, \\ |z_i| - \dfrac{\varepsilon}{2} & \text{otherwise,} \end{cases} \qquad \Phi_\varepsilon(d_i) = \begin{cases} \dfrac{\varepsilon}{2}d_i^2 & \text{if } -1 \le d_i \le 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Here $\Phi_\varepsilon$ is a smooth approximation of the nonsmooth exact penalty function $t\|z\|_1$. The algorithm of [PZ94] requires us to compute an exact optimal solution $u^*$ of (9.6) for given $t$ and $\varepsilon$, and then either increases $t$ if $\|h(u^*)\|_\infty > \varepsilon$ ($u^*$ is not $\varepsilon$-feasible), or decreases $\varepsilon$ otherwise. The suggested procedure for solving (9.6), used in [PZ92], is *simplicial decomposition*, i.e., inner linearization. Hence, in the affine case the algorithm in [PZ94] is very similar to a three-level bundle algorithm that never performs SS. The only difference is that $\varepsilon$ is not decreased to improve the approximation of (9.6) (which is assumed to be exactly solved, although this may not be practical) but rather to force $D_t$ to behave more and more like $t\|\cdot\|_1$; however, this is permitted by our theory. Thus, a generalized bundle method with the above $D_t^*$ offers an alternative to the algorithm of [PZ94], which may be more efficient because (9.6) is only approximately solved and changes of $\bar{x}$ are allowed. Furthermore, we remark that, although $\Phi_\varepsilon$ is mentioned in [PZ94], the corresponding stabilized primal master problem is not described there; however, the corresponding (1.9) is a box-constrained quadratic problem that could be solved with specialized codes (see [Ki89, Fr96]) more efficiently than the nonlinear problem (9.7).

A similar idea has been used to develop $\varepsilon$-approximation algorithms for (block-)structured convex problems [GK95]. In order to solve (9.2) (with $h(u) \le 0$), (9.6) (with $\bar{x} = 0$) is considered, where

$$D_t^*(z) = \ln \sum_i \mathrm{e}^{tz_i}. \tag{9.8}$$

This $D_t^*$ is a smooth approximation of $t\|z\|_\infty$, i.e., $(\ln n) + t\|z\|_\infty \geq D_t^*(z) \geq t\|z\|_\infty$. Problem (9.6) is then approximately solved with an inner linearization approach, i.e., solving (9.7) and using the gradient of $D_t^*$ in $z^*$ (resp., $u^*$) to generate a new point $z$ (resp., $u$). At each step, only the "minimal" bundle $\{u^*, u\}$ is kept. ($D_t^*$ satisfies (P$^*$3$''$)). This approach is not exactly a generalized bundle method, as $D_t^*$ is not zero in the feasible region. However, generalized bundle methods could use slightly modified forms of the above exponential penalty function while allowing changes in $\bar{x}$.

**10. Conclusions.** We have proved convergence of several variants of generalized bundle methods; different convergence properties can be obtained according to the characteristics of the function to be minimized and of the stabilizing term employed. The statements of the properties needed for convergence allow great flexibility in the implementation of the algorithm; several different $t$-strategies and $\beta$-strategies, which are well known to be crucial in practice, can be fitted within this framework.

Our conditions on $D_t$ are less restrictive than those in [Au87, KCL95, Be96], are different from those in [Ki99], and allow a unified treatment of "penalty-like" and "trust-region-like" stabilizing terms [HL93b, sections XV.2.1 and XV.2.2], which have so far been considered as distinct. Very little regularity is required for $D_t(d)$ as a function of $t$. Weak requirements on $f$, such as $^*$-compactness, avoid stronger requirements on $D_t$. A distinguishing feature of our analysis is the extensive exploitation of a new dual viewpoint of bundle methods. Some algorithms that have been proposed outside the bundle framework [PZ94, GK95] can be shown to be closely related to our class.

Our results suggest that practical implementations of generalized bundle algorithms are possible with several different nonquadratic stabilizing terms; examples are primal and/or dual trust regions based on "linear" ($L_1$- or $L_\infty$-)norms, which require the solution of just a linear program at each step. Preliminary computational experiences [Be96] seem to confirm the effectiveness of these approaches. Other stabilizing terms, e.g., exponential or linear-quadratic, may exhibit better convergence in practice than the $L_2$-norm, and thus compensate for the more difficult subproblem to be solved.

Finally, it may be possible to extend these results to an even larger class of bundle algorithms.

**Acknowledgments.** I'm deeply indebted to Claude Lemaréchal for his precious advice, which considerably improved both the presentation and the contents of this paper; in particular, his contribution was fundamental in correcting an error in a previous version of Theorem 3.1 and its consequences. I'm also grateful to a referee for his numerous and detailed comments and for pointing out several glitches, among which was an error in the proof of Theorem 7.3.

REFERENCES

[Au87] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Program. Study, 30 (1987), pp. 102–126.

[Au97] A. Auslender, *How to deal with the unbounded in optimization: Theory and algorithms*, Math. Programming, 79 (1997), pp. 3–18.

[ACC93] A. Auslender, R. Cominetti, and J.-P. Crouzeix, *Convex functions with unbounded level sets and applications to duality theory*, SIAM J. Optim., 3 (1993), pp. 669–687.

[Be96] C. Berger, *Contribution à l'Optimisation Non-Différentiable et à la Décomposition en Programmation Mathématique*, Ph.D. Thesis, Département de Mathématiques et de

Génie Industriel, École Polytechnique de Montréal, Montreal, QC, Canada, 1996.

[BPP91] M. Bougeard, J.-P. Penot, and A. Pommelet, *Towards minimal assumptions for the infimal convolution regularization*, J. Approx. Theory, 64 (1991), pp. 245–270.

[CL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[CT93] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[CFG01] T. G. Crainic, A. Frangioni, and B. Gendron, *Bundle-based relaxation methods for multicommodity capacitated fixed charge network design problems*, Discrete Appl. Math., 112 (2001), pp. 73–99.

[Ec93] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 292–226.

[Fr96] A. Frangioni, *Solving semidefinite quadratic problems within nonsmooth optimization algorithms*, Comput. Oper. Res., 23 (1996), pp. 1099–1118.

[Fr97] A. Frangioni, *Dual Ascent Methods and Multicommodity Flow Problems*, Ph.D. Dissertation, TD 5/97, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 1997.

[Fr98] A. Frangioni, *Generalized Bundle Methods*, Technical report TR 04/98, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 1998.

[Fu98] A. Fuduli, *Metodi Numerici per la Minimizzazione di Funzioni Convesse NonDifferenziabili*, Ph.D. Thesis, DEIS, Università della Calabria, Calabria, Italy, 1998.

[GK95] M. D. Grigoriadis and L. G. Kahchiyan, *An exponential-function reduction method for block-angular convex programs*, Networks, 26 (1995), pp. 59–68.

[GM91] M. Gaudioso and M. F. Monaco, *Quadratic approximations in convex nondifferentiable optimization*, SIAM J. Control Optim., 29 (1991), pp. 58–70.

[GV97] J. Gondzio and J.-P. Vial, *Warm Start and $\varepsilon$-Subgradients in Cutting Plane Scheme for Block-angular Linear Programs*, Logilab Technical report, 1997.1, Paris, France, 1997.

[HL93a] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I—*Fundamentals*, Grundlehren Math. Wiss. 305, Springer-Verlag, New York, 1993.

[HL93b] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II—*Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.

[IST94] A. N. Iusem, B. F. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[IT95] A. N. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal methods for convex and linear programming*, Math. Oper. Res., 20 (1994), pp. 657–677.

[Ki89] K. C. Kiwiel, *A dual method for certain positive semidefinite quadratic programming problems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 175–186.

[Ki95] K. C. Kiwiel, *Approximations in proximal bundle methods and decomposition of convex programs*, J. Optim. Theory Appl., 84 (1997), pp. 529–548.

[Ki98] K. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), pp. 1142–1168.

[Ki99] K. Kiwiel, *A bundle Bregman proximal method for convex nondifferentiable optimization*, Math. Program., 85 (1999), pp. 241–258.

[KCL95] S. Kim, K. N. Chang, and J. Y. Lee, *A descent method with linear programming subproblems for nondifferentiable convex optimization*, Math. Programming, 71 (1995), pp. 17–28.

[LNN95] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111–147.

[LS98] C. Lemaréchal and C. Sagastizàbal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Programming, 16 (1997), pp. 393–410.

[LV98] L. Luksan and J. Vlcek, *A bundle-Newton method for nonsmooth unconstrained optimization*, Math. Programming, 83 (1998), pp. 373–391.

[MSQ98] R. Mifflin, D. Sun, and L. Qi, *Quasi-Newton bundle-type methods for nondifferentiable convex optimization*, SIAM J. Optim., 8 (1998), pp. 583–603.

[Ne95] Y. Nesterov, *Complexity estimates of some cutting plane methods based on the analytic barrier*, Math. Programming, 69 (1995), pp. 149–176.

[Nu97] E. A. Nurminski, *Separating plane algorithms for convex optimization*, Math. Programming, 76 (1997), pp. 373–391.

[OR70] J. M. Ortega and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in*

*Several Variables*, Academic Press, New York, 1970.

[PZ92] M. C. Pinar and S. A. Zenios, *Parallel decomposition of multicommodity network flows using a linear-quadratic penalty algorithm*, ORSA J. Comput., 4 (1992), pp. 235–248.

[PZ94] M. C. Pinar and S. A. Zenios, *On smoothing exact penalty functions for convex constrained optimization*, SIAM J. Optim., 4 (1994), pp. 486–511.

[Ro70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[SZ92] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim., 2 (1992), pp. 121–152.

[Te97] M. Teboulle, *Convergence of proximal-like algorithms*, SIAM J. Optim., 7 (1997), pp. 1069–1083.