SIAM J. OPTIM.
Vol. 23, No. 3, pp. 1784–1809

A NONMONOTONE PROXIMAL BUNDLE METHOD WITH (POTENTIALLY) CONTINUOUS STEP DECISIONS*

A. ASTORINO†, A. FRANGIONI‡, A. FUDULI§, AND E. GORGONE¶

Abstract. We present a convex nondifferentiable minimization algorithm of proximal bundle type that does not rely on measuring descent of the objective function to declare the so-called serious steps; rather, a merit function is defined which is decreased at each iteration, leading to a (potentially) continuous choice of the stepsize between zero (the null step) and one (the serious step). By avoiding the discrete choice the convergence analysis is simplified, and we can more easily obtain efficiency estimates for the method. Some choices for the step selection actually reproduce the dichotomic behavior of standard proximal bundle methods but shed new light on the rationale behind the process, and ultimately with different rules; furthermore, using nonlinear upper models of the function in the step selection process can lead to actual fractional steps.

Key words. nonsmooth optimization, bundle methods, nonmonotone algorithm

AMS subject classifications. 90C26, 65K05

DOI. 10.1137/120888867

1. Introduction.

We are concerned with the numerical solution of the problem

$$f^* = \inf\{f(x) : x \in X\}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a finite-valued proper convex possibly nondifferentiable function and $X \subseteq \mathbb{R}^n$ is closed convex; for notational simplicity we will initially assume $X = \mathbb{R}^n$ with extension to the constrained case discussed later on. *Proximal bundle methods* are known to be among the most efficient implementable algorithms for solving (1.1) when f is known only through an oracle ("black box") that, given any $x \in X$, returns $f(x)$ and one subgradient $z \in \mathbb{R}^n$, i.e., such that $f(y) \geq f(x) + z(y - x)$ for all y; the set of all subgradients is denoted by $\partial f(x)$ (the subdifferential). Thus, as a sequence of *tentative points* $\{x_i\}$ is generated, the information $f_i = f(x_i)$ and $z_i \in \partial f(x_i)$ provided by the oracle is gathered into the *bundle* $\mathcal{B} = \{(x_i, f_i, z_i)\}$ that is used to construct a *model* $f_{\mathcal{B}}$ of the function f, usually the *cutting plane* one,

$$\hat{f}_{\mathcal{B}}(x) = \max\{f_i + z_i(x - x_i) : (x_i, f_i, z_i) \in \mathcal{B}\}, \tag{1.2}$$

that in turn drives the selection of the next tentative point. This is done by taking a distinguished vector $\bar{x}$ as the *current point* and solving a (primal) *master problem*

$$\phi_{\mathcal{B},t}(\bar{x}) = \inf\left\{f_{\mathcal{B}}(\bar{x} + d) + \frac{1}{2t}||d||^2\right\}, \tag{1.3}$$

whose optimal solution d^* provides a *tentative descent direction* along which the next iterate is generated. In (1.3) one minimizes the model $f_{\mathcal{B}}$ plus the *stabilizing term*

*Received by the editors August 22, 2012; accepted for publication (in revised form) June 19, 2013; published electronically September 12, 2013.
http://www.siam.org/journals/siopt/23-3/88886.html

†Istituto di Calcolo e Reti ad Alte Prestazioni, C.N.R., 87036 Rende (CS), Italy (astorino@icar.cnr.it).
‡Dipartimento di Informatica, Università di Pisa, 56127 Pisa, Italy (frangio@di.unipi.it).
§Dipartimento di Matematica, Università della Calabria, 87036 Rende (CS), Italy (antonio.fuduli@unical.it).
¶DEIS, Università della Calabria, 87036 Rende (CS), Italy (egorgone@deis.unical.it).

$\frac{1}{2t}||d||^2$ that discourages points "far away" from $\bar{x}$, where $f_{\mathcal{B}}$ is presumably a "bad model" of f; the *proximal parameter* $t > 0$ controls the "strength" of the stabilization.

In all proximal bundle methods (just "bundle methods" in the following unless otherwise stated) known so far, a *binary* decision is taken depending on the quality of the best point generated along d^*. Typically, only the unitary step is probed; that is, one compares $f(\bar{x})$ with $f(x)$, where $x = \bar{x} + d^*$. If $f(x)$ is "significantly smaller" than $f(\bar{x})$, then $\bar{x}$ is moved to x; this is called a *serious step* (SS). Otherwise, $(x, f(x), z \in \partial f(x))$ is added to $\mathcal{B}$ in order to obtain a (hopefully) better direction at the next iteration; this is called a *null step* (NS). With proper rules for the NS/SS decision, and by appropriate handling of $\mathcal{B}$ and t, these approaches can be shown to minimize f; however, the convergence analysis is somewhat complicated because two basically distinct processes are going on.

In fact, sequences of consecutive NS aim at solving the *stabilized primal problem*

$$\phi_t(\bar{x}) = \inf \left\{ f(\bar{x} + d) + \frac{1}{2t}||d||^2 \right\}, \tag{1.4}$$

i.e., computing the *Moreau–Yosida regularization* ϕ_t of f in $\bar{x}$, which has the same set of minima as f (cf. section 3) but is smooth. The optimal solution of (1.4) (if nonzero) is guaranteed to be a descent direction at $\bar{x}$, but finding it with the sole help of the black box for f is, in principle, as difficult as solving (1.1); thus, bundle methods solve (1.4) *approximately and iteratively* by a standard *cutting-plane-type approach*. Indeed, an NS happens when $f_{\mathcal{B}}(x) \ll f(\bar{x})$ but $f(x) \not\ll f(\bar{x})$; in this case, the new $(x, f(x), z)$ will significantly enrich $\mathcal{B}$, eventually leading to a better approximation to the solution of (1.4). However, after an SS is declared the algorithm can basically be restarted from scratch (apart from keeping $\bar{x}$): the convergence analysis allows one to completely change $\mathcal{B}$ then. In other words, bundle methods are usually analyzed as if being made by two loosely related processes: sequences of consecutive NS, where an approximate solution to (1.4) is computed for the given $\bar{x}$ and t, (possibly) interrupted by the sought-after SS, with each sequence of NS basically seen as being almost completely unrelated to all the others.

It can be argued that this may lead to substantially underestimating the rate of convergence of these methods in practice. Indeed, the (few) available theoretical estimates [18] depict an almost hopelessly slow method, even worse than what methods of this kind necessarily are [21]. While convergence can be slow, and tailing-off often rears its ugly head [5], this is not always the case: when the model $f_{\mathcal{B}}$ "rapidly becomes a good approximation to f," convergence can actually be pretty quick [13, 14]. Thus, it appears that (as one would intuitively expect) the accumulation of information in $\mathcal{B}$ can make a substantial difference in the convergence of the approach; yet, this phenomenon is completely lost by the available theoretical analysis.

The *dichotomic* NS/SS decision to be made at each step is clearly at the basis of the characterization of bundle methods as being made of two loosely related processes. We aim at developing a bundle method where this distinction is removed; that is, convergence is proven by monitoring *one single* merit function and showing that it is improved at all steps, until eventually optimality is reached. This requires doing away with monotonicity in the (anyway, somewhat awkward) form that is usual for bundle methods, i.e., that of the objective function value between two consecutive SS. Our algorithm is nonmonotone in that particular sense, in a different way from previous proposals [8, 15] that are based on the well-known trick of setting a fixed k and requiring that the SS improves over the worst among the last k values of the

current point. Also, while the nonmonotone behavior of the filter-bundle approach of [16] for constrainted optimization is dictated by the need to balance function value improvements with constraint satisfaction improvements, our approach is nonmonotone even in the unconstrained case. The *level bundle* methods [21] are nonmonotone since they always move the current point to the last iterate, making up for convergence with a clever update of the *stabilizing device*; however they require some compactness assumptions that may not be satisfied in applications, and they can be more costly in practice since they can require the solution of two problems at each iteration. Our proposal will instead keep the basic structure of proximal bundle methods, with its advantages: the trait that clearly distinguishes our approach from all other bundle methods so far is that the NS/SS decision is not eliminated, as in level methods, but rather made (ideally) *continuous*. Remarkably, the most natural choices for the step selection process come back to dichotomic decisions; therefore, the simplest implementations of the proposed algorithm still perform NS/SS, although possibly nonmonotone (in terms of f-value) ones. This sheds new light on the rationale of the process and may suggest further developments of classical methods.

The structure of the paper is as follows. In section 2 we introduce the necessary notation and motivate our approach, starting with a naïve initial version of our main idea that does not work but that leads to the definition of the merit function we employ, which is analyzed in section 3. Then, in section 4 we introduce the algorithm, discussing in detail the crucial step of finding the optimal stepsize, and in section 5 we analyze its convergence, propose rules for the online management of the proximal parameter (section 5.1), provide speed-of-convergence estimates (section 5.2), and briefly discuss the impact of employing nonlinear *upper* models of f (section 5.3). Finally, in section 6 we report some preliminary computational results aimed at giving a first estimate of the actual convergence behavior of the new algorithm in relation to that of the standard proximal bundle method, and in section 7 we draw some conclusions.

2. Motivation. To motivate the key ideas in our development, we need to introduce some notation. The *lower model* $f_{\mathcal{B}}$ has to be intended as (1.2), which clearly satisfies $\hat{f}_{\mathcal{B}} \leq f$. Apart from easing the solution of the master problem, as $\hat{f}_{\mathcal{B}}$ is a polyhedral function that can be represented by linear constraints, this has the extra advantage that $\mathcal{B}$ can be considered as a set of *pairs* $\{(\alpha_i^{\bar{x}}, z_i)\}$, where

$$\alpha_i(\bar{x}) = f(\bar{x}) - [f_i + z_i(\bar{x} - x_i)] \tag{2.1}$$

is the *linearization error* of the subgradient z_i obtained at x_i w.r.t. $\bar{x}$. In fact

$$z_i \in \partial_{\alpha_i(\bar{x})} f(\bar{x}), \tag{2.2}$$

where the ε-subdifferential $\partial_\varepsilon f(\bar{x})$ contains all ε-subgradients $z \in \mathbb{R}^n$ such that $f(x) \geq f(\bar{x}) + z(x - \bar{x}) - \varepsilon$ for all $x \in \mathbb{R}^n$. Thus, unlike with other models (e.g., [1]) one does not need to keep track of the iterates x_i in $\mathcal{B}$, since the linearization error can be easily updated using the *information transport property* when $\bar{x}$ is moved to any x,

$$\alpha_i(x) = z_i\big(\bar{x} - x\big) + \alpha_i(\bar{x}) + \big(f(x) - f(\bar{x})\big). \tag{2.3}$$

(Just write (2.1) for x and $\bar{x}$ and simplify out common terms.) Usually $\bar{x}$ is regarded as being fixed, and thus there is no need to stress the fact that the linearization errors depend on it; in our case this is sometimes necessary, but for the sake of notational

simplicity we will still use α_i as much as possible when $\bar{x}$ is clear from the context. Since we will move to points of the form $x(\lambda) = \bar{x} + \lambda d^*$, for which (2.3) gives

$$\alpha_i(x(\lambda)) = \alpha_i(\bar{x}) + f(x(\lambda)) - f(\bar{x}) - \lambda z_i d^*, \tag{2.4}$$

to simplify the notation we will denote $\alpha_i(x(\lambda))$ simply as α_i^λ; note, however, that $f(x(\lambda))$ in (2.4) is usually known only for $\lambda \in \{0, 1\}$. To further ease notation, we will often use the shorthand "$i \in \mathcal{B}$" for "$(\alpha_i, z_i) \in \mathcal{B}$"; let us also remark here that while the index i can upon a first reading be considered that of the iteration, in general the $\mathcal{B}$ evolves in rather different ways. Yet, all this allows us to rewrite (1.3) as

$$\min \left\{ v + \frac{1}{2t}||d||^2 : v \geq z_i d - \alpha_i \quad i \in \mathcal{B} \right\} \quad [+f(\bar{x})], \tag{2.5}$$

where the constant term $+f(\bar{x})$ is most often disregarded, but it is crucial in our development, as we shall see. Solving (2.5) is equivalent to solving its dual

$$\min \left\{ \frac{1}{2} t \left\| \sum_{i \in \mathcal{B}} z_i \theta_i \right\|^2 + \sum_{i \in \mathcal{B}} \alpha_i \theta_i : \theta \in \Theta \right\} \quad [-f(\bar{x})], \tag{2.6}$$

where $\Theta = \{\sum_{i \in \mathcal{B}} \theta_i = 1, \theta_i \geq 0 \quad i \in \mathcal{B}\}$ is the unitary simplex of appropriate dimension, in that $\nu(2.5) = -\nu(2.6)$ (where $\nu(\cdot)$ denotes the optimal value of an optimization problem), and the dual optimal solution θ^* of (2.6), appropriately translated in the (z, α)-space by

$$z^* = \sum_{i \in \mathcal{B}} z_i \theta_i^* \quad \alpha^* = \sum_{i \in \mathcal{B}} \alpha_i \theta_i^*, \tag{2.7}$$

gives the primal optimal solution (v^*, d^*) of (2.5) as

$$d^* = -tz^* \quad v^* = -t||z^*||^2 - \alpha^*. \tag{2.8}$$

The dual form of the master problem not only is useful for algorithmic purposes [9], but it is also closely related to the stopping criterion of the method. In fact, from

$$z^* \in \partial_{\alpha^*} f(\bar{x}) \tag{2.9}$$

(cf. (2.2)) it follows that $z^* = 0$ and $\alpha^* = 0$ imply that $\bar{x}$ is optimal. In practice one would therefore stop when $\|z^*\|$ and α^* are "small," e.g., as in

$$s^* = t^* \|z^*\|^2 + \alpha^* \leq \varepsilon, \tag{2.10}$$

where t^* is an appropriately chosen scaling factor and ε is the final (absolute) accuracy required. In general t must be tuned online to reflect the quality of $f_{\mathcal{B}}$ in the neighborhood of $\bar{x}$, while t^* can stay fixed to a large enough value to ensure that $\|z^*\|^2$ actually is small enough when the algorithm terminates; however, at first reading one may take $t^* = t$, with further discussion provided later on.

In the standard approach, after (2.5) (or, equivalently, (2.6)) are solved one sets $\bar{x} = x$ if $f(x) \ll f(\bar{x})$ (with $x = \bar{x} + d^*$); while this sounds quite natural (after all, we are minimizing f), it is not necessarily the best choice. Indeed, what one would really want is that s^* *decreases as fast as possible*, so that (2.10) is attained as early as possible. But an SS ($\lambda = 1$) does not necessarily decrease it; indeed, any decrease

in s^* is at best indirectly caused by the fact that the α_i's change and in particular that the new pair (z, α) obtained evaluating $f(x)$ has $\alpha(x) = \alpha(x(1)) = \alpha^1 = 0$. Therefore, one may wonder whether, even if $f(x) \ll f(\bar{x})$, a different move than an SS could be better in terms of (2.10). A simple possibility is to consider the line segment $\mathcal{L} = conv(\{\bar{x}, x\}) = \{x(\lambda) : \lambda \in [0, 1]\}$ and *determine the optimal value* λ^* *of* λ *that minimizes* s^* *at the next iteration*, then set the new current point $\bar{x}$ to $x(\lambda^*)$.

One initial issue with this idea is that in order to set $x(\lambda)$ as the current point one needs to know $f(x(\lambda))$ to compute the linearization errors (cf. (2.4)). However, one can alternatively develop an *upper model* $f^{\mathcal{B}}$ of f that is correct at least on $\mathcal{L}$, i.e., such that $f^{\mathcal{B}}(\lambda) = f^{\mathcal{B}}(x(\lambda)) \geq f(x(\lambda))$ for all $x(\lambda) \in \mathcal{L}$, and use it to (conservatively) estimate the true function value and therefore the linearization errors:

$$\bar{\alpha}_i^\lambda = f^{\mathcal{B}}(\lambda) - [f_i + z_i(x(\lambda) - x_i)] \geq f(x(\lambda)) - [f_i + z_i(x(\lambda) - x_i)] = \alpha_i^\lambda. \tag{2.11}$$

(Note that f_i is known exactly, although it should be easy to extend the approach to *approximate* bundle methods where the oracle itself has errors [20].) In particular, the straightforward *worst case upper model* is

$$f^{\mathcal{B}}(\lambda) = (1 - \lambda)f(\bar{x}) + \lambda f(x) = f(\bar{x}) + \lambda \Delta f, \tag{2.12}$$

where $\Delta f = f(x) - f(\bar{x})$; we will mainly work with (2.12) since any more accurate $f^{\mathcal{B}}$ can only improve (decrease) the $\bar{\alpha}_i^\lambda$'s and thus result in faster convergence. Note that $f^{\mathcal{B}}$ depends on the next iterate x, so one cannot use the upper model to drive its selection; indeed, for this task we still use the lower model $f_{\mathcal{B}}$, restricting the use of $f^{\mathcal{B}}$ only to the selection of λ, i.e., of the next current point $x(\lambda) \in \mathcal{L}$. However, (2.12) is still not directly useable because $\bar{x}$ itself may be the $x(\lambda)$ of the previous iteration, hence $f(\bar{x})$ may actually be unknown; for the approach to work in general one has to employ the recursive definition

$$\bar{f}^{\mathcal{B}}(\lambda) = (1 - \lambda)\bar{f}^{\mathcal{B}}(\bar{x}) + \lambda f(x) = \bar{f}^{\mathcal{B}}(\bar{x}) + \lambda \Delta \bar{f}, \tag{2.13}$$

where $\Delta \bar{f} = f(x) - \bar{f}^{\mathcal{B}}(\bar{x})$, which still gives

$$\bar{\alpha}_i^\lambda = \bar{f}(\lambda) - [f_i + z_i(x(\lambda) - x_i)] = \bar{\alpha}_i(\bar{x}) + \lambda \Delta \bar{f} - \lambda z_i d^* \geq \alpha_i^\lambda \tag{2.14}$$

(cf. (2.4)). In the following we will assume knowledge of an upper model like (2.13), which we will denote simply as $\bar{f}$, where possibly some of the data actually comes from an upper estimate of f at previous iterations; this allows us to define the *family of QP dual pairs, parametric over* λ,

$$-\delta(\lambda) = \min \left\{ v + \frac{1}{2t}||d||^2 : v \geq z_i d - \bar{\alpha}_i^\lambda \quad i \in \mathcal{B} \right\} [+f(x(\lambda))], \tag{2.15}$$

$$\delta(\lambda) = \min \left\{ \frac{1}{2} t \left\| \sum_{i \in \mathcal{B}} z_i \theta_i \right\|^2 + \sum_{i \in \mathcal{B}} \bar{\alpha}_i^\lambda \theta_i : \theta \in \Theta \right\} [-f(x(\lambda))]. \tag{2.16}$$

The optimal solution $\theta^*(\lambda)$ of (2.16), or better its representative in the (z, α)-space

$$z(\lambda) = \sum_{i \in \mathcal{B}} z_i \theta_i^*(\lambda), \quad \alpha(\lambda) = \sum_{i \in \mathcal{B}} \bar{\alpha}_i^\lambda \theta_i^*(\lambda), \tag{2.17}$$

immediately gives (cf. (2.9) and (2.11))

$$z_i \in \partial_{\bar{\alpha}_i^\lambda} f(x(\lambda)) \Longrightarrow z(\lambda) \in \partial_{\alpha(\lambda)} f(x(\lambda)). \tag{2.18}$$

Note that (2.18) refers to the "true" $f(x(\lambda))$ rather than its approximation $\bar{f}(\lambda)$, since the estimate of linearization errors is "kept in check" by the fact that at each application of (2.14) we do use the true value of $f(x)$ to compute $\Delta\bar{f}$. Said otherwise, applying (2.14) twice cancels out the term "$\bar{f}^{\mathcal{B}}(\bar{x})$" for the middle point together with any error, so that the error in the estimate of the $\bar{\alpha}_i^{\lambda}$ depends only on that of the initial $f(\bar{x})$ and on that of the final $f(x(\lambda))$; in particular, for an SS $\lambda^* = 1 \Longrightarrow x(\lambda) = x \Longrightarrow f(\lambda) = f(x)$ gives $\bar{\alpha}_i = \alpha_i$. Also, we remark for future reference that $\delta(\lambda) = \nu(2.16) = -\nu(2.15)$ is a *concave* function. Indeed, λ appears linearly in the right-hand side of the constraints in (2.15); therefore, $-\delta(\lambda)$ is the *value function* of a convex problem and hence convex.

Due to (2.18), the stopping condition (2.10)—with $z(\lambda)$ and $\alpha(\lambda)$ replacing z^* and α^*, respectively—can still be applied. One could then seek the value of λ such that $z(\lambda)$ and $\alpha(\lambda)$ are the best possible for (2.10), i.e., the optimal solution λ^* of

$$\min\left\{s^*(\lambda) = t^*\|z(\lambda)\|^2 + \alpha(\lambda) : \lambda \in [0,1]\right\}, \tag{2.19}$$

and set $\bar{x} = x(\lambda^*)$, in order to (hopefully) obtain (2.10) faster. However, it is easy to prove that this approach does not work in general: for the linear $f(x) = rx$ ($r \in \mathbb{R}^n$ fixed) (1.1) is unbounded below, and a bundle algorithm would prove it by making infinitely many consecutive SS along the direction $-r$. It is also easy to see that $\mathcal{B}$ contains all and only identical copies $(r, 0)$; hence, $(z(\lambda), \alpha(\lambda)) = (r, 0)$ whatever the value of λ, i.e., $s^*(\lambda)$ actually does not depend on λ. Therefore *any* $\lambda \in [0,1]$ *is optimal* to (2.19), and nothing can prevent the approach to always select $\lambda^* = 0$, thereby dramatically failing to solve the problem. The example shows that $s^*(\lambda)$ of (2.19) is not an appropriate merit function for our problem; in the next section we will therefore propose a modification of the approach that solves this issue. As we shall see, the basic idea is simply that *the constant* $f(x(\lambda))$ *in* (2.15)–(2.16) (or, better, its readily available approximation $\bar{f}(\lambda)$) *needs be taken into account as well.*

3. The merit function. An appropriate merit function can be devised exploiting the Moreau–Yosida regularization ϕ_t (cf. (1.4)) of f, for which

$$\phi_t \le f, \tag{3.1}$$

$$\phi_t(x) = f(x) \iff x \text{ optimal for (1.1)}. \tag{3.2}$$

Indeed, $d = 0$ is a feasible solution to (1.4), which gives (3.1). Furthermore, the optimality conditions of (1.4),

$$0 \in \partial[f(x + \cdot) + ||\cdot||^2/(2t)](d^*) \iff -d^*/t \in \partial f(x + d^*),$$

clearly imply that $d^* = 0 \iff x$ is optimal for (1.1), but $d^* = 0 \iff \phi_t(x) = f(x)$, whence (3.2). As remarked, ϕ_t is only a *conceptual* object because it is difficult to compute (the oracle being for f); however, owing to $f_{\mathcal{B}} \le f$, for the *readily available* $\phi_{\mathcal{B},t}$ (cf. (1.3)) one clearly has

$$\phi_{\mathcal{B},t} \le \phi_t [\le f], \tag{3.3}$$

$$\phi_{\mathcal{B},t} = f(x) \Longrightarrow x \text{ optimal for (1.1)}. \tag{3.4}$$

We are now ready to propose the merit function, in both the conceptual form,

$$\zeta_t(x) = 2f(x) - \phi_t(x),$$

and the implementable form,

$$\zeta_{\mathcal{B},t}(x) = 2f(x) - \phi_{\mathcal{B},t}(x) \ [\geq \zeta_t(x)]. \tag{3.5}$$

Another way to look at the definition is to write it as $\zeta_t(x) = f(x) + (f(x) - \phi_t(x))$, thereby revealing the *gap function* associated with ζ_t (and with ϕ_t)

$$\delta_t(x) = \zeta_t(x) - f(x) = f(x) - \phi_t(x) \ [\geq 0], \tag{3.6}$$

which gives $\zeta_t(x) = f(x) \iff f(x) = \phi_t(x) \iff \delta_t(x) = 0$, and therefore

$$\zeta_t \geq f, \tag{3.7}$$

$$\zeta_t(x) = f(x) \iff x \text{ optimal for (1.1)} \tag{3.8}$$

via (3.1)–(3.2). This and (3.5) then immediately give

$$\zeta_{\mathcal{B},t} \geq \zeta_t [\geq f], \tag{3.9}$$

$$\zeta_{\mathcal{B},t} = f(x) \Longrightarrow x \text{ optimal for (1.1)}, \tag{3.10}$$

which is equivalently rewritten in terms of

$$\delta_{\mathcal{B},t} = \zeta_{\mathcal{B},t} - f = f - \phi_{\mathcal{B},t} \geq \delta_t \geq 0.$$

The link with the master problems (2.15)–(2.16) is

$$\begin{aligned} \zeta_{\mathcal{B},t}(x(\lambda)) &= 2f(x(\lambda)) - \phi_{\mathcal{B},t}(x(\lambda)) = 2f(x(\lambda)) - \nu \ (2.15) \\ &= f(x(\lambda)) + t\|z(\lambda)\|^2/2 + \alpha(\lambda) = f(x(\lambda)) + \delta_{\mathcal{B},t}(x(\lambda)) \end{aligned} \tag{3.11}$$

due to $\nu(2.15) = -\nu(2.16)$ and the fact that the constant term $-f(x(\lambda))$ in (2.16) cancels out with the factor of 2.

Hence, both ζ_t and ϕ_t coincide with f only at optimality, but the former is an upper approximation, whereas the latter is a lower approximation; even more relevant is that while ϕ_t is convex, ζ_t in general is not (although clearly it is a Difference of Convex (DC) function, as is δ_t), as shown in Figure 3.1 for the simple piecewise-linear function $f(x) = \max\{-3x + 8, -x + 4, 1, x - 3, 2x - 9\}$. Thus, one could hardly use $\zeta_{\mathcal{B},t}$ instead of $\phi_{\mathcal{B},t}$ in (1.3), but still ζ_t is an appropriate merit function, at least on converging (sub) sequences.

LEMMA 3.1. *Assume that a sequence* $\{\bar{x}_i, \mathcal{B}_i, t_i\}$ *is given such that* $t_i \geq \underline{t} > 0$, $\{\bar{x}_i\} \to \bar{x}$ *and* $\liminf_{i\to\infty}(\delta_i = \delta_{\mathcal{B}_i,t_i}(\bar{x}_i)) = 0$*; then,* $\bar{x}$ *is optimal for* (1.1).

Proof. Since $\delta_{\mathcal{B},t} \geq \delta_t$, $t_i \geq \underline{t}$, ζ_t is nondecreasing in t (hence δ_t is), and $\delta_t \geq 0$ is lower semicontinuous,

$$\liminf_{i\to\infty} \delta_{\underline{t}}(\bar{x}_i) = 0 \Longrightarrow \delta_{\underline{t}}(\bar{x}) = 0 \Longrightarrow \zeta_{\underline{t}}(\bar{x}) = f(\bar{x}),$$

and the result follows from (3.10). ∎

Nota that Lemma 3.1 requires nothing specific about how $\mathcal{B}_i$ is handled, provided of course that one succeeds in sending $\delta_{\mathcal{B}_i,t_i}(\bar{x}_i)$ to zero; however, $t_i \geq \underline{t} > 0$ is crucial. In fact, there is an easy but fictitious way to ensure $\zeta_t(\bar{x}) = f(\bar{x})$: just take $t = 0$ (or, in a sequence, have $t_i \to 0$ fast). This does not harm much ϕ_t, except of course killing any regularization effect, $\phi_0 \equiv f$, which still means that all minima of ϕ_0 are minima of f. Conversely, it is disastrous for ζ_t: $\zeta_0 - f = \delta_0 \equiv 0$, hence the merit function is of

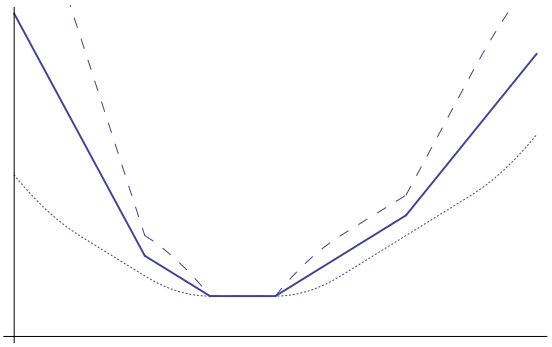

FIG. 3.1. *ϕ_t (dotted line) and ζ_t (dashed line) for a simple f (thick solid line).*

no use whatsoever for detecting minima of f. Requiring t to be bounded away from zero is less than ideal: many algorithms converge even if $t_i \to 0$, provided this happens slowly enough [6]. Furthermore, Lemma 3.1 requires a converging sequence to start with, which may not be trivial to attain in the context of a numerical algorithm. It is possible to improve on both aspects if f is regular enough; one useful concept, already used, e.g., in [1, 10], is the following.

DEFINITION 3.2. *Let $S_\delta(f) = \{x : f(x) \le \delta\}$ be the level set corresponding to the f-value δ: a function f is *-compact if for all $\bar{l} \ge \underline{l} > f^* = \nu(1.1) \ge -\infty$*

$$e(\underline{l}, \bar{l}) = \sup_x \{ dist(x, S_{\underline{l}}(f)) : x \in S_{\bar{l}}(f) \} < \infty. \tag{3.12}$$

In a *-compact function the excess of any two level sets is bounded; many functions are *-compact, e.g., all the inf-compact ones, as discussed in more detail in [10], where the class was introduced to study the convergence of bundle methods with nonquadratic stabilizing terms. Interestingly, the same concept can be used to extend Lemma 3.1 to nonconverging sequences, using the following technical lemma.

LEMMA 3.3. *For each x and $\tilde{x}$ such that $f(\tilde{x}) < f(x)$ and $g \in \partial_\varepsilon f(x)$, it holds that*

$$\delta_t(x) \ge \min \left\{ \frac{(f(x) - f(\tilde{x}))^2 t}{2||\tilde{x} - x||^2}, \frac{(f(x) - f(\tilde{x}) - \varepsilon)^2}{2t||g||^2} \right\}. \tag{3.13}$$

Proof. By the very definition (1.4) we have

$$\delta_t(x) = f(x) - \phi_t(x) = f(x) - \inf \left\{ f(x+d) + \frac{1}{2t}||d||^2 \right\}.$$

To construct a lower estimate for δ_t, pick arbitrarily any $\tilde{x}$ and consider the function

$$\bar{f}(y) = \begin{cases} \beta f(\tilde{x}) + (1-\beta) f(x) & \text{if } y = \beta\tilde{x} + (1-\beta)x \text{ and } \beta \in [0,1], \\ +\infty & \text{otherwise.} \end{cases}$$

By convexity of f it clearly holds $\bar{f} \ge f$, and therefore

$$\begin{aligned} \delta_t(x) &\ge f(x) - \inf \left\{ \bar{f}(x+d) + \frac{1}{2t}||d||^2 \right\} \\ &= -\inf \left\{ \beta(f(\tilde{x}) - f(x)) + \frac{1}{2t}||\beta(\tilde{x} - x)||^2 : \beta \in [0,1] \right\} \\ &= \max \left\{ \psi(\beta) = \beta \Delta f - \frac{1}{2t}\beta^2 \Delta x : \beta \in [0,1] \right\}, \end{aligned}$$

where $\Delta f = f(x) - f(\tilde{x}) > 0$ and $\Delta x = ||\tilde{x} - x||^2 > 0$. The quadratic function $\psi(\beta)$ has maximum $\tilde{\beta} = t\Delta f/\Delta x > 0$ with $\psi(\tilde{\beta}) = t(\Delta f)^2/(2\Delta x)$; therefore, the optimal solution to the maximization problem is

$$\beta^* = \begin{cases} \tilde{\beta} = t\Delta f/\Delta x & \text{if } \tilde{\beta} \le 1 \iff \Delta f \le \Delta x/t, \\ 1 & \text{if } \tilde{\beta} \ge 1 \iff \Delta f \ge \Delta x/t, \end{cases}$$

which immediately gives that its optimal value is

$$\psi(\beta^*) = \begin{cases} t(\Delta f)^2/(2\Delta x) & \text{if } \Delta f \le \Delta x/t, \\ \Delta f - \Delta x/(2t) & \text{if } \Delta f \ge \Delta x/t. \end{cases}$$

For the case $\Delta f \ge \Delta x/t$, we have that $\psi(\beta^*) = \Delta f - \Delta x/(2t) \ge \Delta x/(2t)$ and we need to bound this in terms of Δf. To do that, pick any $g \in \partial_\varepsilon f(x)$ and write the ε-subgradient inequality:

$$f(\tilde{x}) \ge f(x) + g(\tilde{x} - x) - \varepsilon \Longrightarrow ||g|| \cdot ||x - \tilde{x}|| \ge \Delta f - \varepsilon.$$

This gives $\Delta x \ge (\Delta f - \varepsilon)^2/||g||^2$ and hence the desired result. ∎

LEMMA 3.4. *Assume that f is *-compact: any sequence $\{\bar{x}_i, \mathcal{B}_i, t_i\}$ such that $\zeta_i = \zeta_{\mathcal{B}_i,t_i}(\bar{x}_i)$ is monotone nonincreasing, $0 < t_i \le \bar{t} < \infty$, and*

$$\liminf_{i\to\infty} \delta_{\mathcal{B}_i,t_i}(\bar{x}_i)/t_i = 0 \tag{3.14}$$

is an optimizing sequence, i.e., $f_\infty = \liminf_{i\to\infty}(\bar{f}_i = f(\bar{x}_i)) = f^*$.

Proof. Because $\zeta_i \ge \bar{f}_i$ and ζ_i is nonincreasing, the sequence $\{\bar{f}_i\}$ is bounded above: $\bar{f}_i \le \bar{l} < \infty$. If $f_\infty = -\infty$, then the thesis is proved: clearly, $f^* = -\infty$ and $\{\bar{x}_i\}$ is an optimizing sequence. Otherwise, assume by contradiction that $\bar{f}_i - f^* > \gamma > 0$ and set $-\infty < \underline{l} = \inf\{\bar{f}_i\} - \gamma \le \bar{l} - \gamma$. Since $(\delta_i = \delta_{\mathcal{B}_i,t_i}(\bar{x}_i))/t_i = ||z_i^*||^2/2 + \alpha_i^*/t_i$, (3.14) implies that, at least on one subsequence, $||z_i^*|| \to 0$ and $\alpha_i^* \to 0$. Hence we can apply (on the subsequence) Lemma 3.3 with $\tilde{x}$ the projection of $\bar{x}_i$ over the level set $S_{\underline{l}}(f)$ (so that $\bar{f}_i - f(\tilde{x}_i) \ge \gamma$), $g = z_i^*$, and $\varepsilon = \alpha_i^*$: using (3.12) in (3.13) gives

$$\delta_{t_i}(\bar{x}_i) \ge \min\left\{ \frac{\gamma^2 t_i}{2e(\underline{l}, \bar{l})^2}, \frac{(\gamma - \alpha_i^*)^2}{2t_i||z_i^*||^2} \right\}.$$

Since $\alpha_i^* \to 0$, $\gamma - \alpha_i^*$ is bounded away from zero; furthermore in the denominator of the rightmost term t_i is bounded above and $||z_i^*||$ goes to zero, so the whole term eventually becomes large. Finally, divide by t_i to get $\delta_i = \delta_{t_i}(\bar{x}_i)/t_i \ge \gamma^2/(2e(\underline{l}, \bar{l})^2)$, which contradicts (3.14) and hence proves the result. ∎

Under mild conditions on f, any sequence $\{\bar{x}_i\}$ that nullifies the gap function is minimizing; t_i can even go to zero, provided this happens "slowly enough w.r.t. δ_i" so that δ_i/t_i still goes to zero. In fact, a familiar way to obtain (3.14) is

$$\sum_{i=1}^{\infty} t_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \delta_i < \infty; \tag{3.15}$$

if δ_i goes to zero fast enough, so that the series converges, then t_i can even be allowed to go to zero slowly enough, so that the series diverges. Condition (3.15) implies (3.14) since $\liminf_{i\to\infty} \delta_i/t_i > 0$ implies that for some $\delta > 0$, a large enough h, and all $i \ge h$ one has $\delta_i \ge \delta t_i$, hence $\sum_{i=h}^{\infty} \delta_i \ge \delta \sum_{i=h}^{\infty} t_i$, which contradicts (3.15). Alternatively, if

t_i is bounded away from zero, then (3.14) just requires that δ_i vanishes. Indeed, if f is inf-compact (hence clearly *-compact), then we can extract a converging subsequence out of $\{\bar{x}_i\} \subset S_L(f)$ (that is compact) and therefore recover Lemma 3.1.

Thus, under appropriate conditions $\zeta_{\mathcal{B},t}$ can be used to construct a nonmonotone bundle method along the lines of section 2; at least it solves the $f(x) = rx$ counterexample. This first requires working with the approximate version of (3.11) employing $\bar{f}$,

$$\bar{\zeta}_{\mathcal{B},t}(x(\lambda)) = \bar{f}(\lambda) + t\|z(\lambda)\|^2/2 + \alpha(\lambda) \geq \zeta_{\mathcal{B},t}(x(\lambda)) \tag{3.16}$$

(as $\bar{f}(\lambda) \geq f(x(\lambda))$ and $\bar{\alpha}^\lambda \geq \alpha^\lambda$). However, in the linear case $\bar{f} = f \Longrightarrow \bar{\alpha} = \alpha$; furthermore, $z(\lambda)$ and $\alpha(\lambda)$ are independent on λ. Yet, due to the extra term $\bar{f}(\lambda)$ in (3.16) w.r.t. (2.19), it is easy to see that the merit function approach correctly assesses that $\lambda^* = 1$, which could therefore lead to an algorithm capable of solving this (very easy) problem. This algorithm is described in detail in the next section.

4. The algorithm. We now present the algorithm and discuss its main properties. To simplify the notation we will as much as possible avoid the iteration index "i," using the subscript "$+$" for "$i+1$." Differently from standard proximal bundle algorithms, the sequence $\{\bar{x}_i\}$ of stability centers may be almost entirely unrelated from that $\{x_i\}$ of iterates; as a consequence, the algorithm works with the upper approximation $\bar{f}(\bar{x}_i)$ in place of the true value $f(\bar{x}_i)$, and therefore the approximate linearization errors $\bar{\alpha}$, the approximate merit function $\bar{\zeta}_{\mathcal{B},t}$ of (3.16), the corresponding approximate gap function $\bar{\delta}_{\mathcal{B},t}$ of (3.6), and so on. Hence, all references, e.g., to the master problems (2.5)–(2.6), their solutions (2.7) and (2.8), and so on, have to be intended with the approximate quantities in place of the exact ones. These coincide only when $\lambda^* \in \{0, 1\}$, which may (cf. section 4.1) or may not be true.

0. Choose any $\bar{x}$. Initialize $0 < t \leq \bar{t} < \infty$. Set $d^* = 0$, $\mathcal{B} = \emptyset$. Goto 3.
1. Solve (2.6) for θ^*. Find z^*, $\bar{\alpha}^*$, and d^* from (2.7) and (2.8). $\mathcal{B} = \mathcal{B} \cup \{(z^*, \bar{\alpha}^*)\}$.
2. If (2.10) holds, then stop ($\bar{x}$ is ϵ-optimal).
3. Compute $x = \bar{x} + d^*$, $f(x)$, $z \in \partial f(x)$, and α via (2.1). $\mathcal{B} = \mathcal{B} \cup \{(z, \alpha)\}$.
4. For $\mathcal{B}' \subseteq \mathcal{B}$, compute an approximately optimal solution $\lambda^* \in [0, 1]$ to
$$\min\left\{\bar{\zeta}_{\mathcal{B}',t}(x(\lambda)) : \lambda \in [0,1]\right\}. \tag{4.1}$$
5. Set $\bar{x}_+ = x(\lambda^*)$, $\mathcal{B}_+ \subseteq \mathcal{B} \cup \{(z(\lambda^*), \alpha(\lambda^*))\}$, the linearization errors of $\mathcal{B}_+$ according to (2.14), and $0 < t_+ \leq \bar{t}$. Goto 1.

A few comments on the algorithm are useful.

- In the first iteration, where $d^* = 0$, $\bar{x} = x$, $\bar{f}(\bar{x}) = f(x)$, and $\mathcal{B} = \{(z, 0)\}$, (4.1) is "degenerate": any $\lambda \in [0, 1]$ is optimal.
- The fundamental property that we want from the algorithm is
$$\bar{\zeta}_+ = \bar{f}(\bar{x}_+) + t_+\|z_+^*\|^2/2 + \bar{\alpha}_+^* \leq \bar{f}(\bar{x}) + t\|z^*\|^2/2 + \bar{\alpha}^* = \bar{\zeta}. \tag{4.2}$$
This is easy to obtain by requiring monotonicity of t
$$t_+ \leq t (\leq \bar{t}), \tag{4.3}$$
an ever-increasing bundle $\mathcal{B} \subseteq \mathcal{B}_+$, and complete information in (4.1), i.e., $\mathcal{B}' = \mathcal{B}$, since then $(z^*, \bar{\alpha}^*)$ is feasible for (2.16) at $\lambda = 0$, and therefore (4.1) can

only improve on it. However, some form of *bundle management* should be allowed at step 4 (while choosing $\mathcal{B}'$) and at step 5 (while choosing $\mathcal{B}_+$), which still can guarantee (4.2) provided that

- $(z^*, \bar{\alpha}^*)$ is feasible for (2.16) at $\lambda = 0$;
- $(z(\lambda^*), \alpha(\lambda^*))$ is feasible for (2.6) at the subsequent iteration.

This can be obtained by the well-known *aggregation trick*: just add the pair one needs to preserve to $\mathcal{B}$. In the standard case this needs to be done only once per iteration, here—since there are two different bundles $\mathcal{B}$ and $\mathcal{B}'$—it must be done twice, at step 1 and at step 4. Doing that allows us to remove any other pair from $\mathcal{B}/\mathcal{B}'$, provided the critical ones are retained; the most aggressive application of this approach results in $\mathcal{B}' = \{(z^*, \bar{\alpha}^*), (z, \alpha)\}$ in (4.1) and the "poorman's bundle" $\mathcal{B}_+ = \{(z(\lambda^*), \alpha(\lambda^*))\}$ [5], in which case $1 = \theta^*_+ \in \mathbb{R}$ is the only feasible (hence optimal) solution to (2.6) at the next iteration, and therefore

$$(z^*_+, \bar{\alpha}^*_+) = (z(\lambda^*), \alpha(\lambda^*)). \tag{4.4}$$

A milder way to obtain the same result is to ensure that $z(\lambda^*)$ and $\alpha(\lambda^*)$ are still feasible for (2.6) at the beginning of the next iteration by inhibiting removal of any subgradient $h \in \mathcal{B}$ such that $\theta^*_h(\lambda^*) > 0$.

- While the approach seems to solve two problems (2.6) and (4.1), this depends on what exactly "approximately solve" means in step 4 and on the relationships between $\mathcal{B}$ and $\mathcal{B}_+$. For instance, if $\mathcal{B}_+$ is the poorman's bundle, then (4.4) implies that step 1 can be skipped. Furthermore, note that (4.1) is a *difficult* problem as $\bar{\zeta}_{\mathcal{B},t}$ is nonconvex; thus one in principle wants $\mathcal{B}$ as small as possible in the former but not in the latter because working with a very restricted bundle has in general dire consequences on the convergence speed [5, 13]. This requires us to re-solve (2.6) once λ^* has been found, which is why we present it as the default behavior of our algorithm.
- At step 5, one may compute $f(x(\lambda^*))$ (and some subgradient) to avoid using the approximation $\bar{f}(\lambda^*)$ (and possibly improve $\mathcal{B}_+$) and ensure $\bar{\alpha} = \alpha$, $\bar{\zeta}_{\mathcal{B},t} = \zeta_{\mathcal{B},t}$ as in the standard bundle approach. Although this may be worth it if the function computation is quick (e.g., [13]), in general one can work with the approximate quantities, hence we prefer to develop our theory in the more general setting; things can only improve if $\bar{f} = f$. Of course, this is only relevant if $\lambda^* \notin \{0, 1\}$, which may not happen often, if at all (cf. section 4.1).
- Obviously, (4.3) is somewhat harsh: it is well known that online tuning of t is important in practice. Besides, as discussed in section 3, some care is needed to avoid t_i going to zero too fast and thereby rendering $\bar{\zeta}$ useless (cf. (3.14)–(3.15)). These issues require a better grasp of the convergence properties of the approach, and this is why we postpone their discussion to section 5.1.

4.1. Finding λ^*. The solution of (4.1) clearly depends on the specific upper model $\bar{f}^{\mathcal{B}}$ employed; the linear one (2.13) is very easy to compute and available for any f, but it is also the worst-case upper model, and therefore the one relying on the most conservative (hence least accurate) estimates. Furthermore, the corresponding (4.1) is *concave* in λ: indeed, $\delta(\lambda)$ (cf. (2.16)–(2.15)) is concave, and $\bar{f}^{\mathcal{B}}$ is linear. As a consequence, the use of (2.13) leads to $\lambda^* \in \{0, 1\}$, i.e., to only making either NS or SS, although with different rules than in the traditional bundle method. This also implies that (4.1) is then easy to solve: just optimize on θ for $\lambda = 0$ and $\lambda = 1$ separately, then pick the solution giving the best $\bar{\zeta}$-value. While this would require

solving the master problem thrice (or twice if step 1 is skipped as already discussed) at each iteration, one can employ the aggregation trick, in particular with

$$\mathcal{B}' = \{(z^*, \bar{\alpha}^*), (z, \alpha)\}, \tag{4.5}$$

("extreme aggregation") so that the solution to (2.6) can be found by a closed algebraic expression. It is useful to develop this approach in detail, which requires introducing some notation; first and foremost, the two linear functions in λ

$$f_*(\lambda) = \lambda z^* d^* - \bar{\alpha}^* + \bar{f}(\bar{x}), \qquad f(\lambda) = \lambda z d^* - \alpha + \bar{f}(\bar{x})$$

such that $f_{\mathcal{B}'}(x(\lambda)) = \max\{f_*(\lambda), f(\lambda)\}$. In particular, (2.14) then gives

$$\begin{aligned}\bar{\alpha}^{*,\lambda} &= \bar{f}(\lambda) - f_*(\lambda) = \bar{f}(\bar{x}) + \lambda\Delta\bar{f} - \lambda z^* d^* + \bar{\alpha}^* - \bar{f}(\bar{x}) \\ &= \lambda(\Delta\bar{f} - z^* d^*) + \bar{\alpha}^* = \lambda[(z - z^*)d^* - \alpha] + \bar{\alpha}^*, \\ \bar{\alpha}^{\lambda} &= \bar{f}(\lambda) - f(\lambda) = \bar{f}(\bar{x}) + \lambda\Delta\bar{f} - \lambda z d^* + \alpha - \bar{f}(\bar{x}) \\ &= \lambda(\Delta\bar{f} - z d^*) + \alpha = \alpha(1 - \lambda).\end{aligned}$$

One therefore is faced with the special version of (2.16),

$$\delta(\lambda) = \min\left\{\frac{1}{2}t\,\|\theta z^* + (1-\theta)z\|^2 + \theta\bar{\alpha}^{*,\lambda} + (1-\theta)\bar{\alpha}^{\lambda} : \theta \in [0,1]\right\} \tag{4.6}$$

whose optimal solution has the closed-form expression

$$\theta^*(\lambda) = \min\left\{1, \max\left\{0, \tilde{\theta}(\lambda) = \frac{\bar{\alpha}^{\lambda} - \bar{\alpha}^{*,\lambda} - tz(z^* - z)}{t\|z^* - z\|^2}\right\}\right\} \tag{4.7}$$

since $\tilde{\theta}(\lambda)$ is the optimal solution of the *relaxation* of (4.6) where the constraint $\theta \in [0,1]$ is removed. Therefore, one can easily evaluate (an upper estimate of) $\bar{\zeta}(0)$ and $\bar{\zeta}(1)$ in $O(n)$ ($O(1)$ once a few scalar products are computed once and for all), solving (4.1) (under (4.5)) with the same complexity.

It is instructive to examine the result from a different viewpoint. A little algebra (using in particular $\Delta\bar{f} = -tz^*z - \alpha$) shows that (4.1) under (4.5) can be written as

$$\begin{aligned}&\min\left\{h(\theta,\lambda) = h(\theta) + \lambda\big(\Delta\bar{f} + tz^*(z^* - z)\theta - \alpha\big) : \theta \in [0,1], \lambda \in [0,1]\right\}, \\ &\text{where} \qquad h(\theta) = t\|\theta z^* + (1-\theta)z\|^2/2 + \theta\bar{\alpha}^* + (1-\theta)\alpha\end{aligned} \tag{4.8}$$

is the objective function of (4.6) for $\lambda = 0$. For the optimal solution (θ^*, λ^*), then,

$$\begin{aligned}\Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha > 0 &\Longrightarrow \lambda^* = 0, \\ \Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha < 0 &\Longrightarrow \lambda^* = 1,\end{aligned} \tag{4.9}$$

since $h(\theta,\lambda)$ is linear in λ. This first confirms that (2.13) leads to $\lambda^* \in \{0,1\}$, except for the vanishingly small chance that $\Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha = 0$. More tellingly, (4.9) can be interpreted as an ex-post NS/SS rule, to be contrasted to the standard one,

$$\Delta\bar{f} \le m[f_{\mathcal{B}}(x) - \bar{f}(\bar{x})] \Longrightarrow \lambda^* = 1, \tag{4.10}$$

for some arbitrary $m \in (0,1]$. Of course, (4.9) can only be evaluated after (4.1) has been solved; yet, one easily derives the *sufficient* conditions

$$\begin{aligned}\Delta\bar{f} > \max\{tz^*(z - z^*), 0\} + \alpha &\Longrightarrow \lambda^* = 0, \\ \Delta\bar{f} < \min\{tz^*(z - z^*), 0\} + \alpha &\Longrightarrow \lambda^* = 1\end{aligned}$$

that can be evaluated early on to avoid solving (4.1). These conditions show that the process is indeed nonmonotone: while (4.10) requires $\Delta\bar{f}$ to be negative, and "sizably so," to declare an SS, (4.9) only forces an NS when $\Delta\bar{f}$ is "sizably positive," allowing for SS to be taken even when $\Delta\bar{f} > 0$, but "not too large." This is also seen by exploiting the fact that $f(1) - f_*(1) \geq 0$ (f is convex) $\Longrightarrow tz^*(z^* - z) \geq \alpha - \bar{\alpha}^*$: using this (multiplied by θ^*) in (4.9) gives the ex-post

$$\Delta\bar{f} > \bar{\alpha}^*\theta^* + \alpha(1 - \theta^*) \Longrightarrow \lambda^* = 0$$

(which can be made ex-ante using $\bar{\alpha}^*\theta^* + \alpha(1 - \theta^*) \leq \max\{\bar{\alpha}^*, \alpha\}$), showing once again that $\Delta\bar{f}$ need be large positive for an NS to be declared.

Using the simplified problem corresponding to extreme aggregation (4.5) would make (4.1) solvable even when using nonlinear upper models. In fact, it is clear that

$$\delta(\lambda) = t\|\theta^*(\lambda)z^* + (1 - \theta^*(\lambda))z\|^2/2 + \theta^*(\lambda)\bar{\alpha}^{*,\lambda} + (1 - \theta^*(\lambda))\bar{\alpha}^{\lambda}$$

is a concave piecewise function with at most three pieces, since $\theta^*(\lambda)$ is linear function of λ inside one (not necessarily proper) subinterval of $[0, 1]$ and constant outside it. Thus, $\delta(\lambda)$ is quadratic inside that interval and linear outside it, as is easy to verify with somewhat tedious algebra. For any different $f^{\mathcal{B}}$ that is piecewise in $[0, 1]$ with a small number of pieces, each one of them being a simple (say, smooth algebraic) function (cf. section 5.3), (4.1) can still be solved by inspecting all intervals and evaluating $\delta(\lambda) + f^{\mathcal{B}}(\lambda)$ at all extremes and at all points where the derivative vanishes.

This approach could be generalized to $\mathcal{B}$ larger than (4.5), since the optimal solution of (2.16) can be shown to be piecewise-linear in λ as done in [9, section 6] for t. Therefore, the analogue to $\theta^*(\lambda)$ could be constructed, and an explicit concave piecewise form for $\delta(\lambda)$ could be devised; however, the number of pieces would grow exponentially in $|\mathcal{B}|$. While this might be useful for some specific applications, we will concentrate on using the linear upper model (2.13) and the minimal bundle (4.5); all improvements on these would a fortiori converge (hopefully, faster).

5. The convergence proof. The idea of the convergence proof is to show that at all steps $\bar{\zeta}$ decreases enough. (In this section, too, we remove the iteration index i whenever possible.) That is, we need to monitor the crucial quantity

$$\Delta\bar{\zeta} = \bar{\zeta} - \bar{\zeta}_+, \tag{5.1}$$

a nonnegative number due to (4.2), and prove that $\bar{\zeta}_i \to \bar{f}_i (= \bar{f}(\bar{x}_i))$: we will do that by showing that $\Delta\bar{\zeta}$ is at least as large as a nonvanishing fraction of

$$\bar{\zeta} - \bar{f} = \bar{\delta} = t\|z^*\|^2/2 + \bar{\alpha}^* \geq t\|z^*\|^2/2 + \alpha^* = \delta.$$

While we aim at ultimately replacing the standard convergence arguments for proximal bundle methods, we will exploit several of the ideas proposed there. The first, as already discussed, is the aggregation technique leading to the simplified (4.6) of section 4.1. Indeed, it is well known that convergence *of a sequence of NS* is retained even with the poorman's bundle, which allows us to prove convergence even if the maximum size of the bundle is fixed to any number ≥ 2. While the practical usefulness is debatable, here we are interested in the fact that studying $\Delta\bar{\zeta}$ under the extreme aggregation assumption (4.5) allows us to considerably simplify the convergence proof.

The second idea one can exploit, somewhat counterintuitively (or maybe not, given section 4.1), is that of NS/SS. While the algorithm is not—in principle—restricted to these two dichotomic decisions, the fact that standard bundle methods converge and our previous developments clearly suggest that one should be able to prove convergence even if λ^* is restricted to belonging to $\{0,1\}$, i.e., only NS and SS are done. That is, one could interpret (4.10) as a *heuristic for the solution of* (4.1) (taking λ^* when the condition does not hold). We note in passing that (4.10) can usually be replaced with the *weaker*

$$\Delta\bar{f} \le -m\big[t||z^*||^2/2 + \bar{\alpha}^*\big] = -m\big[\bar{\zeta}_{\mathcal{B},t}(\bar{x}) - \bar{f}(\bar{x})\big] = -m\bar{\delta}$$

(cf. (3.11)) because $f_{\mathcal{B}}(x) - \bar{f}(\bar{x}) = v^* = -t||z^*||^2 - \bar{\alpha}^* \le -t||z^*||^2/2 - \bar{\alpha}^*$ (cf. (2.8)). Since at each step either (4.10) holds or it doesn't, we can bound $\Delta\bar{\zeta}$ from below considering separately both cases. To simplify the treatment we assume $t_+ = t$; needless to say, the analysis will a fortiori hold under (4.3), as $\delta_{\beta,t}$ is increasing in t.

We start the case $\lambda^* = 0$ ((4.10) does not hold); this corresponds to bounding the decrease in the master problem value during one regular NS, i.e.,

$$\begin{aligned}\Delta\bar{\zeta} = \bar{\zeta} - \bar{\zeta}_+ &= \bar{f}(\bar{x}) + t\|z^*\|^2/2 + \bar{\alpha}^* - \big(\bar{f}(x(0)) + t\|z(0)\|^2/2 + \alpha(0)\big)\\ &= t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z_+^*\|^2/2 - \bar{\alpha}_+^*,\end{aligned}$$

where $(z_+^*, \bar{\alpha}_+^*)$ are the optimal solution to the standard master problem corresponding to adding (z,α) to $\mathcal{B}$ while keeping all the rest untouched. Estimating $\Delta\bar{\zeta}$ only requires simple algebra using the results of section 4.1; indeed, for $h(\theta)$ of (4.8) one has

$$h'(1) = tz^*(z^* - z) + \bar{\alpha}^* - \alpha.$$

From (2.1) and $tz^* = -d^* = \bar{x} - x$ one has $\Delta\bar{f} = f(x) - \bar{f}(\bar{x}) = -tzz^* - \alpha$; using this in (the contrary of) (4.10) gives

$$h'(1) > (1-m)\big[t||z^*||^2 + \bar{\alpha}^*\big] \ge (1-m)\bar{\delta}.$$

Hence, $h'(1)$ is strictly positive (and large if $\bar{\delta}$ is), which implies that $\theta = 1$ cannot be the optimal point. Thus, as in (4.7), $\theta^* = \theta^*(0)$ is either the unconstrained minimizer $\tilde{\theta}(0)$ of h, or 0. (Note that the latter happens in particular if $z = z^*$, which means that $h(\theta)$ is linear and $\tilde{\theta}$ is not well defined.) The former case gives

$$\Delta\bar{\zeta} \ge h(1) - h(\tilde{\theta}) = \frac{(tz^*(z-z^*) + \alpha - \bar{\alpha}^*)^2}{2t\|z^* - z\|^2} = \frac{h'(1)^2}{2t\|z^*-z\|^2} > \frac{((1-m)\bar{\delta})^2}{2t\|z^*-z\|^2},$$

while the latter gives $\Delta\bar{\zeta} \ge h(1) - h(0) = t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z\|^2/2 - \alpha$. To further develop this, combine $h'(0) = t(z^* - z)z + \bar{\alpha}^* - \alpha \ge 0$ (since h is convex and 0 is the constrained optimum) and (the contrary of) (4.10) to get

$$\begin{aligned}&t\|z\|^2/2 + \alpha < \frac{1+m}{2}(t\|z^*\|^2/2 + \bar{\alpha}^*) \qquad \text{and therefore}\\ &\Delta\bar{\zeta} \ge t\|z^*\|^2/2 + \bar{\alpha}^* - (t\|z\|^2/2 + \alpha) > \frac{1-m}{2}(t\|z^*\|^2/2 + \bar{\alpha}^*) = \frac{1-m}{2}\bar{\delta}.\end{aligned}$$

Combining the two cases we finally obtain

$$\Delta\bar{\zeta} \ge \frac{(1-m)\bar{\delta}}{2}\min\left\{1, \frac{(1-m)\bar{\delta}}{t||z-z^*||^2}\right\}. \tag{5.2}$$

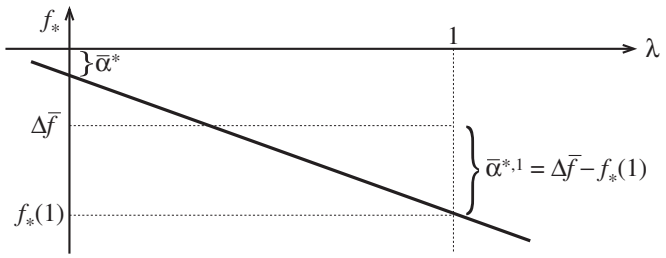

FIG. 5.1. *Estimating* $\bar{\alpha}^{*,1}$

Under (4.10), instead, one has

$$\begin{aligned}\Delta\bar{\zeta} = \bar{\zeta} - \bar{\zeta}_+ &= \bar{f}(\bar{x}) + t\|z^*\|^2/2 + \bar{\alpha}^* - \big(f(x(1)) + t\|z(1)\|^2/2 + \alpha(1)\big) \\ &= -\Delta\bar{f} + t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z(1)\|^2/2 - \alpha(1). \qquad (5.3)\end{aligned}$$

Hence, in this case we do have the positive term $-\Delta\bar{f}$ corresponding to the function value improvement, but we have to estimate the possible *increase* in the value of the master problem corresponding to the change in $\bar{x}$. To simplify this task we assume that $\mathcal{B}' = \{(z^*, \bar{\alpha}^*)\}$, i.e., the newly obtained (z, α) is discarded (this is indeed possible in standard bundle methods, where $\mathcal{B}$ can be reset arbitrarily at each SS); clearly, this corresponds to underestimating $\Delta\bar{\zeta}$. Yet, this also gives

$$\Delta\bar{\zeta} \geq -\Delta\bar{f} + t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z^*\|^2/2 - \bar{\alpha}^{*,1} \geq -\Delta\bar{f} + \bar{\alpha}^* - \bar{\alpha}^{*,1}.$$

The useful relationship is that $\bar{\alpha}^{*,1} = \bar{f}(\bar{x}) + \Delta\bar{f} - f_*(1)$ (see Figure 5.1), where $f_*(1) = v^* + \bar{f}(\bar{x}) = f_{\mathcal{B}'}(x)$, which gives

$$\Delta\bar{\zeta} \geq -2\Delta\bar{f} + \bar{\alpha}^* + f_*(1) - \bar{f}(\bar{x}).$$

Using $\bar{\alpha}^* \geq 0$, $-\Delta\bar{f} \geq -m\big[f_*(1) - \bar{f}(\bar{x})\big]$ (cf. (4.10)) and $f_*(1) - \bar{f}(\bar{x}) = v^* = -t||z^*||^2 - \bar{\alpha}^* \leq -t||z^*||^2/2 - \bar{\alpha}^*$ one finally obtains

$$\Delta\bar{\zeta} \geq (1-2m)\big[f_*(1) - \bar{f}(\bar{x})\big] \geq (2m-1)(t\|z^*\|^2/2 + \bar{\alpha}^*) = (2m-1)\bar{\delta}$$

provided that $2m - 1 > 0$.

Since (4.10) either holds or not, we can conclude that

$$\Delta\bar{\zeta} \geq \bar{\delta} \min\left\{2m-1, \frac{1-m}{2}, \frac{(1-m)^2\bar{\delta}}{2t||z-z^*||^2}\right\} \qquad (5.4)$$

however chosen $m \in (1/2, 1]$; note that standard proximal bundle methods only require $m \in (0, 1]$, but here m is only a technicality in the proof, not an actual algorithmic parameter (possibly requiring tuning). With (5.4) it is easy to prove convergence of the algorithm; to simplify the analysis let us do away first with the obvious case where $\bar{f}_\infty = \liminf_{i\to\infty} \bar{f}_i = -\infty$, since then clearly $f^* = \nu(1.1) = -\infty$: $\{\bar{x}_i\}$ is a minimizing sequence, and there is nothing else to prove.

THEOREM 5.1. *If* (4.1) *is solved by* (4.10), (4.3) *holds,* $\bar{f}_\infty > -\infty$, *and*

$$||z_i|| \leq L < \infty \qquad \forall i, \qquad (5.5)$$

then $\liminf_{i\to\infty} \bar{\delta}_i = 0$ *and*

(i) *if* $t \geq \underline{t} > 0$ *and some subsequence of* $\{\bar{x}_i\}$ *converges to some* x^**, then* x^* *is an optimal solution to* (1.1);

(ii) *if* f *is* **-compact and* (3.14) *holds, then* $\bar{f}_\infty = f^*$.

Proof. Since $\bar{\zeta}_i \geq \bar{f}_i$, the sequence $\{\bar{\zeta}_i\}$ is bounded below and nonincreasing, hence it converges to some $\bar{\zeta}_\infty > -\infty$; thus, $\bar{\delta}_i \to 0$. Indeed, $||z_i^*||$ is bounded above and (5.5) implies that $\{z_i\}$ is also bounded: hence, so is $t_i||z_i - z_i^*||^2$ in the denominator of the rightmost term of (5.4) (given that $t_i \leq \bar{t}$). Therefore, if $\bar{\delta}_i \geq \varepsilon > 0$ for infinitely many indices, then (5.4) would give that $\sum_{i=1}^{\infty} \Delta\bar{\zeta}_i = \infty$, contradicting boundedness of $\bar{\zeta}_i$. This immediately gives part (i) via Lemma 3.1, which requires t bounded away from zero, and part (ii) via Lemma 3.4, which requires f to be *-compact and (3.14). □

Of course, (3.14) requires appropriate management of t; however, t bounded away from zero surely suffice, and more sophisticated strategies are discussed in section 5.1.

Clearly, Theorem 5.1 a fortiori holds if (4.1) is solved by any approach giving a solution at least as good as (4.10); this includes (4.9) and/or any other approach solving accurately enough (4.1) with larger $\mathcal{B}'$ than (4.5). The same idea can also be readily exploited for the announced extension to the constrained case $X \subset \mathbb{R}^n$, which means that one is minimizing the *essential objective* $f_X = f + \imath_X$, with $\imath_X$ the indicator function of X ($\imath_X(x) = +\infty$ for $x \notin X$ and 0 otherwise). To simplify the discussion consider the polyhedral case $X = \{x \in \mathbb{R}^n : \gamma_i \geq g_i x \; i \in \mathcal{V}\}$, i.e., a finite and fixed *bundle of constraints* $\mathcal{V}$, although the concept generalizes. The standard approach in this case is simple: just insert full knowledge of X in the master problems

$$\min\left\{v + \frac{1}{2t}||d||^2 : v \geq z_i d - \bar{\alpha}_i, \quad i \in \mathcal{B}, 0 \geq g_i d - \bar{\gamma}_i, \quad i \in \mathcal{V}\right\},$$
$$\min\left\{\frac{1}{2}t\left\|\sum_{i\in\mathcal{B}} z_i\theta_i + \sum_{i\in\mathcal{V}} g_i\mu_i\right\|^2 + \sum_{i\in\mathcal{B}} \bar{\alpha}_i\theta_i + \sum_{i\in\mathcal{V}} \bar{\gamma}_i\mu_i : \theta \in \Theta, \mu \geq 0\right\},$$

where $\bar{\gamma}_i = \gamma_i^{\bar{x}} = \gamma_i - g_i\bar{x}$. That is, f_X is the sum of two components f and $\imath_X$, where the latter, in the parlance of [13], is an "easy" one: it is completely known from the start (although in practice dynamic handling of $\mathcal{V}$ is also possible).

COROLLARY 5.2. *With the above modifications, Theorem* 5.1 *holds in the constrained case.*

Proof. The primal-dual relationships of the modified master problems are

$$d^* = -t\tilde{z}^*, \quad v^* = -t||\tilde{z}^*||^2 - \tilde{\alpha}^*$$

(cf. (2.8)), where now $\tilde{z}^* = z^* + g^*$, $\tilde{\alpha}^* = \bar{\alpha}^* + \gamma^*$, $g^* = \sum_{i\in\mathcal{V}} g_i\mu_i^*$, and $\gamma^* = \sum_{i\in\mathcal{V}} \bar{\gamma}_i\mu_i^*$. It is easy to realize that $\tilde{f}_*(\lambda) = \lambda\tilde{z}^*d^* - \tilde{\alpha}^* + \bar{f}(\bar{x})$ is a valid lower model of f_X on $\mathcal{L}$: $g^* \in \partial_{\gamma^*}\imath_X(\bar{x})$, since each g_i is a $\bar{\gamma}_i$-subgradient at $\bar{x}$ (use, e.g., the very definition) and it is possible to scale each constraint by any positive value so that μ^* can be convex multipliers (if nonzero). Thus, together with $f(\lambda)$ of section 4.1 they define a valid lower model $\tilde{f}_{\mathcal{B}'} = \max\{\tilde{f}_*(\lambda), f(\lambda)\}$, which is *weaker* than the one actually used by the algorithm, i.e., $\tilde{f}_{\mathcal{B}'} \leq f_{\mathcal{B}',X} = \max\{f_*(\lambda), f(\lambda)\} + \imath_X$. Indeed, in the constrained case aggregation should only happen in $\mathcal{B}$, with the poorman's bundle still being $\{(z^*, \bar{\alpha}^*)\}$ but *all the bundle of constraints* $\mathcal{V}$ *still in the master problems*; instead, $f_{\mathcal{B}'}$ linearizes $\imath_X$ together with f. Note that the upper model does not change (it only uses the value $f_X(x) = f(x)$ at feasible x), and indeed there is no need to approximate $\imath_X$ from above because its value (0) is known exactly. Therefore, all the above development can be repeated with $\tilde{z}^*$ and $\tilde{\alpha}^*$ in place of z^* and $\bar{\alpha}^*$, showing

that even with full aggregation ($f + \imath_X$ being treated as just one function instead of exploiting the knowledge about X) the algorithm is convergent, a fortiori so if the better model $f_{\mathcal{B}',X}$ is used. ∎

To conclude this section, it must be remarked that Theorem 5.1 is somewhat weaker than those available for standard bundle methods, in two main aspects. First, the global boundedness condition (5.5) is usually not required. Something related is needed to show that the denominator in the rightmost term of (5.4) does not grow infinitely large, but this is only required for sequences of consecutive NS. Since $\bar{x}$ is not changing, it is easy to obtain (5.5) as a natural consequence of the algorithm's workings: $\{z_i^*\}$ and $\{\alpha_i^*\}$ are bounded since the objective function of (2.6) is nonincreasing, and hence $\{d_i^*\}$ is bounded (t_i is bounded above), and hence the $\{z_i\}$ all belong to the image of a compact set through the ε-subdifferential mapping of the finite function f for some bounded ε, which is compact. This line of reasoning fails in our case since $\bar{x}$ is changing in a less controlled way, and it does not seem to be easy to directly extend the argument. Also, requiring f to be *-compact is usually not necessary: for instance, [6, Proposition 1.2] guarantees convergence without any assumption on f provided that (in our notation) $||z_i^*|| \to 0$, $\alpha_i^* \to 0$, and

$$\sum_{i=1}^{\infty} \lambda_i^* t_i = \infty. \tag{5.6}$$

While we do have the first two conditions, proving (5.6) for our algorithm, even in the easy case where t_i is bounded away from zero, is not easy. It is so in traditional bundle methods in the case where infinitely many SS are done (the other, where a sequence of infinitely many consecutive NS eventually starts, is dealt with separately): indeed, in that case $\lambda_i^* = 1$ infinitely many times and the requirement is just that t_i goes to zero slowly enough, à la (3.15). One would guess that either (5.6) holds or "$\bar{x}_i$" is moving little and we are converging somewhere (so that Lemma 3.1 applies); however, this would call for proving that $\sum_{i=1}^{\infty} t_i ||z_i^*|| < \infty$, while we only have the (much) weaker $\sum_{i=1}^{\infty} t_i ||z_i^*||^4 < \infty$. Again, a large part of this difference comes from the nonlinear (rightmost) term of (5.4), which is not there in the standard convergence proof since it appears only in the study of consecutive sequences of NS.

Thus, the convergence results for our approach are somehow less satisfactory than these available for standard bundle methods, although they still cover many practical applications; for instance, (5.5) holds when is f globally Lipschitz or inf-compact ($\Longrightarrow$ *-compact) since all $\bar{x}_i$ belong to some level ($\bar{\zeta}_i \geq f_i$ and $\bar{\zeta}_i$ is nonincreasing), and more in general if f is globally Liptchiz when restricted to any sublevel set, even if not compact (think e^x). Allowing intermediate steps (between NS and SS), while simplifying some of the arguments, leaves significantly more freedom to the algorithm, rendering it somewhat more difficult to analyze. This could, however, just be the consequence of an unrefined analysis: it is possible that (5.4) can be improved to make some of these weakness disappear. Furthermore, (3.14) lends itself well to algorithmic treatment, as discussed in the next section.

5.1. Management of t. The above development requires (4.3), which is unrealistic since online tuning of t is well known to be crucial. However, increasing t can easily kill any monotonicity argument in $\bar{\zeta}$, so this is only going to be possible in a controlled way. Similarly, decreases of t must be controlled, in that t must either remain bounded away from zero, or at least (3.14) has to hold. Going toward practical implementations, one should therefore decide when t should be decreased,

and how should exactly t_+ be chosen in order for (3.14) to be satisfied. Standard *proximity control* techniques developed for the proximal bundle method [17] are not immediately applicable here since they are mostly (although not entirely) based on estimating and testing the decrease of f; that is, t is (possibly) increased at SS and decreased at NS, a strategy that cannot be easily replicated here.

As far as increasing t is concerned, the fundamental observation is that all our analysis basically hinges upon (5.4) (which clearly implies (4.2)), so that as long as that condition holds the algorithm converges. This implies that *whenever* $\Delta\bar{\zeta}$ *is large, we can sacrifice some of this decrease to allow increases of* t. Thus, with two appropriate constants $\kappa_1 > \kappa_2 > 0$ we can define

- a *very good step* for which

$$\Delta\bar{\zeta} > \kappa_1\bar{\delta}; \tag{5.7}$$

- a *good step* for which $\kappa_1\bar{\delta} \geq \Delta\bar{\zeta} \geq \kappa_2\bar{\delta}$;
- a *not-so-good step* for which $\Delta\bar{\zeta} < \kappa_2\bar{\delta}$.

Condition (5.7) gives global convergence as (5.4) does (and in particular linear convergence). Hence, for $t_+ \geq t$ one has

$$\begin{aligned}\Delta\bar{\zeta} &= \bar{f} + t\|z^*\|^2/2 + \bar{\alpha}^* - \big(f(\lambda^*) + t\|z(\lambda^*)\|^2/2 + \alpha(\lambda^*)\big)\\ &\geq -\lambda^*\Delta\bar{f} + \bar{\delta} - t_+\|z(\lambda^*)\|^2/2 - \alpha(\lambda^*) > \kappa_1\bar{\delta},\end{aligned}$$

and therefore one can keep (5.7) satisfied by choosing any t_+ such that

$$2\big((1-\kappa_1)\bar{\delta} - \lambda^*\Delta\bar{f} - \alpha(\lambda^*)\big)/\|z(\lambda^*)\|^2 \geq t_+ > t, \tag{5.8}$$

picking the largest possible value being reasonable. Note that (except in the poorman's case) $\bar{\delta}_+ \leq t_+\|z(\lambda^*)\|^2 + \alpha(\lambda^*)$, but this is not an issue for (5.8), as a decreasing $\bar{\delta}_+$ (and hence $\bar{\zeta}_+$) can only help to attain (5.7). One can expect (5.7) to happen mostly when λ^* is large, so this process somewhat mirrors the standard approach to increase t after a successful SS; however, nothing actually prevents t increases from happening after a successful NS that has consistently decreased the optimal value of the master problem. It may be wise to introduce a further damping constraint $t_+ \leq \bar{p}t$ for some fixed $\bar{p} \in [1,\infty)$ to avoid excessive changes of t, together of course with the overarching $t_+ \leq \bar{t}$. For good (but not very good) steps we are content with the current convergence rate and we keep t unchanged, while for not-so-good steps we suspect that the algorithm is stuck in the "bad part with sublinear convergence" of the estimate (5.4); we can then try to improve on this by decreasing t, whence obtaining a smaller $\bar{\delta}_+$. Of course, this decrease of $\bar{\zeta}$ is somewhat fictitious, as discussed in section 3; thus, we may want to only ensure the *minimal target* $\Delta\bar{\zeta} \geq \kappa_2\bar{\delta}$ by

$$t_+ \leq 2\big((1-\kappa_2)\bar{\delta} - \lambda^*\Delta\bar{f} - \alpha(\lambda^*)\big)/\|z(\lambda^*)\|^2 < t, \tag{5.9}$$

where, similarly to (5.8), we may want to select the largest possible value in (5.9) (the minimal possible change of t). The combination of this and (5.8) should ideally produce a linearly convergent process with rate between κ_1 and κ_2, which could in many cases be considered effective. However, we also need to ensure that the relevant hypotheses of Theorem 5.1 hold, i.e., either $t \geq \underline{t} > 0$ or (3.14), and likely to also impose the reasonable damping $t_+ \geq \underline{p}t$ for some fixed $\underline{p} \in (0,1]$. To ensure the weakest hypothesis (3.14) one can select any increasing function $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{s\to 0^+} s/\tau(s) = 0$ (e.g., $\tau(s) = \nu\sqrt{s}$ for a fixed $\nu > 0$) and then require

$$t_+ \geq \tau\big(t\|z(\lambda^*)\|^2 + \alpha(\lambda^*)\big) \geq \tau\big(\bar{\delta}_+\big). \tag{5.10}$$

In fact, this yields $\bar{\delta}_+/t_+ \leq \bar{\delta}_+/\tau(\bar{\delta}_+)$, so that $\bar{\delta}/t \to 0$ as $\bar{\delta} \to 0$ (which it does). Clearly, all this (and a fortiori $t \geq \underline{t} > 0$) can in practice interfere with obtaining the desired rate of convergence; this, if nothing else, justifies increasing t "when the sun shines," in order to buy up some wiggle room to decrease it "when the rain comes."

While (5.7) provides a sensible *aspiration criterion* for managing t during the process, and in particular increasing it against the natural requirement (4.3), it is appropriate (if only for the links with the following section 5.2) to remark that it is not the only one. In particular, our initial motivation for the development of the approach (cf. section 2) was to attain the stopping condition (2.10) as quickly as possible. That condition uses one parameter t^* to weight the term $\|z^*\|^2$ in the x-space (or, better, its dual) with the term $\bar{\alpha}^*$ in the f-value space; while both these need be small, t^* dictates the "relative importance" of the two, which is clearly related to the scaling of the function. Indeed, assume for simplicity that *dom* f (or an appropriate level set) is compact with $D < \infty$ its diameter: the term $t^*\|z^*\|^2 = (t^*z^*)z^*$ in (2.10) is a measure of how much we can decrease at most traveling along the approximate subgradient $-z^*$ by a step t^*, since using $\|x^* - \bar{x}\| \leq D$ and (2.9) gives

$$f(x^*) \geq f(\bar{x}) + z^*(x^* - \bar{x}) - \bar{\alpha}^* \Longrightarrow D\|z^*\| + \bar{\alpha}^* \geq f(\bar{x}) - f^*.$$

When (2.10) holds, one has $\bar{\alpha}^* \leq \varepsilon$ and $\|z^*\| \leq \sqrt{(\varepsilon - \bar{\alpha}^*)/t^*}$ and therefore if t^* is chosen such that

$$\varepsilon \geq D\sqrt{(\varepsilon - \bar{\alpha}^*)/t^*} + \bar{\alpha}^* \geq D\|z^*\| + \bar{\alpha}^* \geq f(\bar{x}) - f^*,$$

then $\bar{x}$ is actually ε-optimal at termination; a sufficient condition is

$$t^* \geq D^2/\varepsilon. \tag{5.11}$$

One could therefore fix $t = t^*$, so that the merit function $\bar{\zeta}$ exactly weights $||z^*||$ as in (2.10), thereby (hopefully) having the latter satisfied as quickly as possible. This is, however, moot in more ways than one. First, (5.11)—assuming D can be estimated—is likely to be a rather large value: fixing t there would result in an almost unstabilized approach, which is likely to be rather inefficient in practice. Furthermore, the initial aim of reaching (2.10) as quickly as possible has had to be changed anyway, since one needs to take into account the decrease of f as well. Indeed, $\bar{\zeta}$ is the sum two terms, $\bar{f}(\bar{x}) + \bar{\alpha}^*$ in the f-value space and $\|z^*\|$ in the dual space, with t^* dictating of the relative importance of having a small z^* w.r.t. having a small f-value. During the algorithm, this should arguably change: intuitively, reducing the f-value is more important at the early stages, where $\bar{x}$ is very far from being optimal, while reducing $\|z^*\|$ is more important at the final stages where $\bar{x}$ is close to being optimal and one only needs to prove it. Thus, t should arguably be allowed to be (much) smaller than t^* during the course of the algorithm but it might be beneficial to have it grow near t^* toward the end; all this could suggest other sensible aspiration criteria for management of t, but the details are left for future research.

Finally, let us remark that, since the theory refers to the asymptotic behavior, everything that only happens finitely many times is allowed. That is, like, e.g., in [10, (4.ii)], t_+ can be set to whatever value one desires provided that some mechanism ensures this stops happening at some point and (5.8)–(5.10) are satisfied from then on, so that the convergence theory applies. In standard bundle methods this mostly applies to sequences of NS: as soon as an SS is performed, everything, among which the counters for allowing irregular behavior, can be reset. This is a convenient consequence

of the otherwise irritating characterization of bundle methods as composed by two almost independent processes (cf. section 1) that is not shared by our approach, which is analyzed in terms of a unique converging process. Yet, as shown in the next section, this allows us to more easily derive efficiency estimates for the method than is possible for the standard approach [18].

5.2. Efficiency estimates. We now turn to deriving efficiency estimates for the method, that is, an upper bound on the number of iterations i required to ensure that $\bar{f}_i - f^* \leq \varepsilon$. Apart from the given (absolute) maximum error $\varepsilon > 0$, the bound typically depends on a pair of other (often unknown in practice) constants: the maximum norm of any subgradient encountered during the process, and the maximum distance between any two iterates. The former is clearly smaller than $L < \infty$ of (5.5) (e.g., the global Lipschitz constant of f), while for the latter one can take $D < \infty$ of (5.11) (e.g., the diameter of *dom* f or of an appropriate level set if f is inf-compact).

We need first to detail the relationships between the stopping condition (2.10) (with $\bar{\alpha}^*$ replacing α^*, of course), $\bar{\zeta}$, and the desired property $\bar{f}_i - f^* \leq \varepsilon$. As previously discussed, the latter is implied by (2.10) whenever (5.11) holds, which we will then assume in the following. Further, it is easy to verify that if $\tau_i = t^*/t_i \geq 1$, then $2\bar{\delta}_i\tau_i \geq s_i^*$; therefore, (2.10) surely holds if

$$2\bar{\delta}_i\tau_i \leq \varepsilon \equiv 2\bar{\delta}_i/t_i \leq \varepsilon/t^* \equiv 2\bar{\delta}_i \leq \varepsilon(t_i/t^*). \tag{5.12}$$

(Note that, due to (3.14), this has to happen eventually.)

The estimate starts from the obvious $\bar{f}_1 - f^* \leq LD$, which, since $||z_1^*||^2 \leq L^2$ and $\bar{\alpha}_1^* = \alpha_1 = 0$, leads to

$$\varepsilon(t_1/t^*) < 2\bar{\delta}_1 \leq t_1 L^2 \tag{5.13}$$

(for otherwise (5.12) would hold for $i = 1$) and therefore

$$\bar{\zeta}_1 - \bar{f}_1 \leq \bar{\delta}_1 \Longrightarrow \bar{\zeta}_1 - f^* \leq LD + \bar{\delta}_1 \Longrightarrow \bar{\zeta}_1 - f^* \leq LD + t_1 L^2/2$$

(note that $\bar{f}_1 = f_1$ and $\bar{\zeta}_1 = \zeta_1$). Thus, we need to estimate the iteration k such that

$$\sum_{i=1}^{k-1} \Delta\bar{\zeta}^i = \bar{\zeta}_1 - \bar{\zeta}_k \geq LD + \bar{\delta}_1 - \varepsilon(t_i/t^*)/2; \tag{5.14}$$

indeed, when this happens we get

$$\bar{\delta}_k = \bar{\zeta}_k - \bar{f}^k \leq \bar{\zeta}_k - f^* \leq \bar{\zeta}_1 - f^* - LD - \bar{\delta}_1 + \varepsilon(t_i/t^*)/2 \leq \varepsilon(t_i/t^*)/2$$

and therefore—via (5.12)—(2.10) holds, and hence $\bar{f}_i - f^* \leq \varepsilon$ due to (5.11). Note that this not only accounts for the iterations required to *reach* an ε-optimal point but also for these required to *certify* its ε-optimality by constructing the appropriate z^* and $\bar{\alpha}^*$.

To simplify the notation we now fix $m = 3/5$, so that $2m - 1 = (1 - m)/2$ and the first two terms in (5.4) are equal; furthermore, we use $\|z - z^*\|^2 \leq 2L^2$ to get

$$\Delta\bar{\zeta} \geq \frac{\bar{\delta}}{5} \min\left\{1, \frac{\bar{\delta}}{5tL^2}\right\}. \tag{5.15}$$

The main issue here is that (5.15) does *not* say how the decrease in $\bar{\zeta}$ is subdivided between its two components. That is, (5.15) may imply either that $\bar{f}$ decreases

(at least) by the given amount, or $\bar{\delta}$ does, or that both decrease by a fraction. Fortunately, it is clear what the worst case is. Indeed, decreasing in $\bar{f}$ while not decreasing $\bar{\delta}$ (or even increasing it) will result in the same (or larger) right-hand side in (5.15) at the next iteration, and therefore at least as large a decrease in $\bar{\zeta}$. Conversely, a decrease in $\bar{\delta}$ results in a smaller right-hand side in (5.15), and therefore to a smaller (estimate, worst-case) decrease in $\bar{\zeta}$ at the next iteration. Thus, from an *adversarial viewpoint* the strategy leading to the slowest possible decrease of $\bar{\zeta}$ is clear:

- first, $\bar{\delta}$ has to be decreased (as slowly as possible) without changing $\bar{f}$, until no further decrease is possible least (5.12) be satisfied;
- then, $\bar{\delta}$ is kept fixed and $\bar{f}$ is improved (as slowly as possible).

Clearly, any other strategy where $\bar{\delta}$ is not decreased as fast as possible will lead to a larger overall worst-case decrease of $\bar{\zeta}$, and therefore faster convergence.

We now estimate the performance corresponding of the above worst-case adversarial strategy, separately for each of the two phases. Actually, the first phase is subdivided into two subphases of large $\bar{\delta}$ and small $\bar{\delta}$; that is, until

$$\bar{\delta}_i/(5t_iL^2) \geq 1 \tag{5.16}$$

the decrease is linear, $\bar{\zeta}_+ \leq \bar{\zeta}_i - \bar{\delta}_i/5$, and since we assume $\bar{f}_+ = \bar{f}_i (= \bar{f}_1)$ this means

$$\bar{\delta}_+ \leq \bar{\delta}_i - \bar{\delta}_i/5 \Longrightarrow \bar{\delta}_i \leq \bar{\delta}_1(4/5)^{i-1}.$$

Unfortunately, (5.16) does not last long: combining it with (5.13) gives $t_1/(10t_i) \geq 1$, i.e., this subphase terminates immediately unless t_i is reduced very quickly. Although this provides an interesting rationale for some of the discussion in section 5.1, it is compatible with the algorithm that this subphase terminates in $O(1)$ iterations (e.g., if t is constant during these) and to simplify the discussion we will assume this happens.

After that, the second subphase begins when instead

$$0 < \varepsilon(t_i/t^*)/2 \leq \bar{\delta}_i \leq 5t_iL^2$$

and therefore the rate of decrease is sublinear:

$$\bar{\zeta}_+ \leq \bar{\zeta}_i - \bar{\delta}_i^2/(25t_iL^2). \tag{5.17}$$

However, since $\varepsilon(t_i/t^*)/2 \leq \bar{\delta}_i$ we can estimate the rate of decrease with the linear

$$\bar{\zeta}_+ \leq \bar{\zeta}_i - \frac{\varepsilon}{50t^*L^2}\bar{\delta}_i,$$

which, using again $\bar{f}_+ = \bar{f}_i (= \bar{f}_1)$ and (5.11), gives

$$\bar{\delta}_+ = \bar{\zeta}_+ - \bar{f}_+ \leq \bar{\zeta}_i - \bar{f}_i - \frac{\varepsilon}{50t^*L^2}\bar{\delta}_i = \bar{\delta}_i\left[1 - \frac{1}{50}\left(\frac{\varepsilon}{DL}\right)^2\right]$$

and therefore

$$\bar{\delta}_i \leq \bar{\delta}_1\left(1 - \frac{\gamma^2}{50}\right)^{i-1} \leq \frac{t_1L^2}{2}\left(1 - \frac{\gamma^2}{50}\right)^{i-1},$$

where $\gamma = \varepsilon/DL$. Hence, the second subphase terminates for the smallest i such that

$$\frac{t_1L^2}{2}\left(1 - \frac{\gamma^2}{50}\right)^{i-1} \leq \varepsilon\frac{t_i}{2t^*},$$

which, using $t_i \le t_1$ and (5.11), gives

$$(1-\gamma^2/50)^{i-1} \le \gamma^2 \Longrightarrow i \ge \xi(\gamma) = \frac{\log(\gamma^2)}{\log(1-\gamma^2/50)} + 1.$$

To estimate how quickly $\xi(\gamma) \to \infty$ when its argument $\varepsilon/DL = \gamma \to 0$, as is easy to verify, one can use the fact that (as easy but tedious calculus shows)

$$\lim_{\gamma \to 0^+} \xi(\gamma)/\gamma^{-k} = 0$$

for all $k > 2$, i.e., that any superquadratic function goes to infinity faster than $\xi(\gamma)$ does. This means that, at least for small values of ε, the second subphase terminates in at most $O((DL/\varepsilon)^k)$ iterations for any $k > 2$.

We are now left to estimate the length of the second (and last) phase, where $\bar{\delta}_i \approx \varepsilon(t_i/t^*)$ is no longer reduced and $\bar{f}_i$ is improved instead. Therefore, denoting by h the last iteration of the first phase, we know that the second phase (and hence the whole algorithm) must terminate when $\bar{\zeta}_h - \bar{\zeta}_i \le LD$. We start again from (5.17), but now we use $\bar{\delta}_i^2 \ge (\varepsilon(t_i/t^*)/2)^2$ and (5.11) to get

$$\bar{\zeta}_+ \le \bar{\zeta}_i - \frac{t_i}{t^*}\frac{\varepsilon^2}{100t^*L^2} = \bar{\zeta}_i - \tau_i^*\frac{\varepsilon^3}{100(DL)^2},$$

where $\tau_i^* = t_i/t^*$. Therefore

$$\bar{\zeta}_i \le \bar{\zeta}_h - \frac{\varepsilon^3}{100(DL)^2}\sum_{j=h}^{i}\tau_i^* \Longrightarrow \sum_{j=h}^{i}\tau_i^* \le 100\left(\frac{DL}{\varepsilon}\right)^3.$$

To obtain the final estimate on i, we need to assume something on how quickly the series of τ_i^* diverges (which of course requires that it indeed diverges in the first place). The simple assumption $\tau_i^* \ge \bar{\tau} > 0$ (which implies t_i bounded away from zero) gives

$$i \le \frac{100}{\bar{\tau}}\left(\frac{DL}{\varepsilon}\right)^3 + h$$

and hence all in all the algorithm has $O((LD/\varepsilon)^3)$ efficiency. This is much worse than the optimal $O((LD/\varepsilon)^2)$ efficiency, that is attained by level methods [21], but comparable to (although somewhat different from) the $O(D^4L^2/\varepsilon^3)$ efficiency of proximal bundle methods [18]. It is hard to say whether the latter difference really reflects a different asymptotic behavior of the two algorithms rather than being a figment of the different complexity analysis techniques; to gain some insight into this issue, the next section is devoted to a preliminary computational comparison between the two.

5.3. Alternative upper models. All the development so far has used the linear upper model (2.13); clearly, the results a fortiori hold for any tighter $f^{\mathcal{B}}$. Intuitively, an upper model which better reflects the behavior of f along λ would lead to a better estimate of the potential decrease of f and therefore to better choices. This is easily seen with an example: consider the function $f(x) = |x|$ with $x_1 = \bar{x} = -1$, and $\mathcal{B} = \{(-1, 0)\}$. For $t = 2$ one would have $x = 2$ with $\Delta\bar{f} = 0$ and the new pair $(1, 2)$ added to $\mathcal{B}'$. Simple symmetry arguments (or a little tedious algebra) show that both $\lambda = 0$ and $\lambda = 1$ are optimal for (4.1), but none of the points $\lambda \in (0, 1)$: $\bar{\zeta}(\lambda)$ is concave (quadratic), with the *maximum* in $x = 0$, precisely where the *minimum* of f lies. Thus, the choice of λ^* is taken on a very bad estimate of the real behavior

of f, which is clearly due to $\bar{f}^{\mathcal{B}}$ assuming f to be constant while it actually has a V-shape. This shows that better upper models may be useful to improve the algorithm's decisions. However, there are no standard ways to construct upper models for convex functions (apart of course from (2.13) itself).

A first possibility could therefore be to rely on specific properties of f. For instance, in many applications, such as in classification problems [2, 3], f is a polyhedral max-function corresponding to an easy optimization problem, that is, either a convex (linear) program [7, 11, 14] or a nonconvex one that can still be solved efficiently [12]. In this case, one can use sensitivity analysis techniques on the problem to determine the minimum value $\bar{\lambda} \leq 1$ so that the optimal solution for the f-problem at x ($\lambda = 1$) is still optimal for $\lambda = \bar{\lambda}$ (and therefore for all values in between). This means that (cf. section 4.1)

$$f(x(\lambda)) = f(\lambda) = \lambda z d^* - \alpha + \bar{f}(\bar{x}) \qquad \forall \lambda \in [\bar{\lambda}, 1]$$

and therefore that one can use the piecewise-linear upper model

$$f^{\mathcal{B}}(\lambda) = \begin{cases} (1-\lambda)\bar{f}^{\mathcal{B}}(\bar{x}) + \lambda f(\bar{\lambda}) & \text{if } 0 \leq \lambda \leq \bar{\lambda}, \\ f(\lambda) & \text{if } \bar{\lambda} \leq \lambda \leq 1. \end{cases}$$

Of course, in the worst case $\bar{\lambda} = 1$ and this gives back (2.13). This approach is easily extended to composite functions where an appropriate mapping is superimposed over the max-function [22].

A different (not necessarily alternative) idea is to use an upper model that estimates the shape of f without requiring $f^{\mathcal{B}} \geq f$. A simple approach could be to mimic a technique developed for heuristically adjusting t [11, 17]: develop the quadratic function $q(\lambda)$ so that $q(0) = 0$, $q(1) = f(1)$, and $q'(1) = f'(1)$. (Note that $q(\lambda) - \delta(\lambda)$ can be convex, concave, or neither.) This possibly estimates the shape of f better than (2.13), but it cannot be used alone because it may overestimate the decrease of f, leading to a step that is actually nonmonotone in ζ. A possible workaround is to combine this model with (2.13) by defining the subinterval of $[0, 1]$ in which (2.13) ensures a sufficient decrease, e.g., in the sense of (5.7). Since one can arbitrarily choose any value of λ in the interval without compromising convergence, the heuristic upper model can be used to drive the selection of λ^* in there.

6. Computational results. We emphasize that the results reported in this section are only meant to be preliminary; the aim is to provide a first look at the performances of the proposed approach as compared to existing ones, and in particular the proximal bundle method. For this purpose we have implemented the proposed approach, which we refer to as NMBundle, in `C++`; in particular we solve (4.1) under (4.5) by means of the closed formulae of section 4.1, and we manage t with the mechanism described in section 5.1. This is compared with the proximal bundle code (which we refer to as PBundle) developed by the second author and already used with success in several other applications [7, 11, 13, 14]. The structure of the code allows for several useful features; for instance, once an *oracle* for a function f is implemented it can be used by both approaches without modifications, and NMBundle can exploit the efficient master problem solver of [9]. All the algorithms have been compiled with GNU `g++ 4.0.1` (with `-O3` optimization option) and run on an Opteron 246 (2 GHz) computer with 2 GB of RAM, under Linux Fedora Core 3.

We have compared the two approaches on two different test sets. The first is a set of 12 academic test functions with small n (up to 48) that have been used

TABLE 6.1
Results for standard test functions.

		PBundle		NMBundle	
f	n	iter	gap	iter	gap
CB2	2	19	3.8e-7	23	1.5e-7
CB3	2	13	2.8e-7	16	8.9e-11
DEM	2	10	2.5e-12	11	6.9e-8
QL	2	17	6.2e-8	18	4.6e-8
LQ	2	11	3.1e-8	6	6.3e-8
Mifflin1	2	31	6.9e-7	26	8.2e-7
Rosen	4	35	3.7e-7	38	1.8e-7
Maxq	20	143	4.6e-7	118	1.4e-6
Maxl	20	32	1.8e-15	41	4.4e-16
Maxquad	10	129	3.3e-7	107	1.8e-7
TR48	48	141	0.0e+0	198	4.7e-15
Shor	5	36	3.1e-7	41	5.0e-7

many times to evaluate the performances of algorithms for (convex) nondifferentiable optimization; one recent instance is [1], to which the interested reader is referred for a detailed description. The results are reported in Table 6.1, where column "iter" reports the number of iterations (function evaluations) and column "gap" returns the relative gap between the best function value found by the algorithm at termination and the true optimum of the function. For both approaches, extensive tuning of the algorithmic parameters has been performed in order to find the single setting that produces the best results across all functions. In particular, for both approaches the initial value of t is set to 0.1, and for NMBundle the two crucial parameters κ_1 and κ_2 in (5.8) and (5.9) are set to 0.1 and 0.06, respectively. The stopping criterion was set to a target *relative* accuracy of `1e-6` ($\varepsilon = 10^{-6} \cdot f(\bar{x})$) in (2.10), with t^* chosen as small as possible to obtain (almost) ex-post satisfaction of the prescribed accuracy (in particular, $t^* = 1$ for all functions except TR48, where $t^* = 100$ is required).

The table shows that the two approaches are roughly comparable; NMBundle requires fewer iterations in 4 cases over 12, in two of them (Maxq and Maxquad) somewhat significantly, while it is significantly slower in the largest TR48. (Note that the cost per iteration of the two approaches is virtually identical, which is why we report only iteration counts.) However, it is difficult to draw significant conclusions out of a test set comprising a few very different test functions of low dimensionality. Therefore we also compared the two approaches on a significant application: the solution of Lagrangian duals for optimization problems with multicommodity flow structure. These, in their many variants, are a staple in Lagrangian optimization; see, e.g., [4, 7, 11, 13, 19] and the references therein. In particular here we employ the simple *weak flow relaxation* of the fixed-charge multicommodity Min-Cost flow problem. We avoid delving into the details of the original problem and of the corresponding Lagrangian relaxation, referring the interested reader to [7] and especially to the recent [13] for the details (comprising a description of the freely available test instances); here we only mention that, contrary to the academic test cases of Table 6.1, these are larger-scale problems where $n = m$ is the number of arcs in the underlying graph, and the problems are constrained ones with (as often happens in Lagrangian relaxation) $X = \mathbb{R}^n_+$.

Also in this case we have performed accurate tuning for both approaches (for PBundle we actually piggybacked on the tuning performed in [13]), which in particular led to selecting $\kappa_1 = 0.6$ and $\kappa_2 = 0.001$ in (5.8) and (5.9), respectively (now $t = 1$ at start). The stopping criterion has been set as in the previous case, and

TABLE 6.2
Results for multicommodity flows.

		PBundle		NMBundle	
m	k	iter	gap	iter	gap
300	100	2404	1.3e-13	773	7.6e-7
300	200	2109	2.9e-14	961	5.6e-7
300	400	1118	6.8e-7	940	2.0e-7
300	800	1644	6.6e-7	1217	1.6e-7
600	100	824	1.9e-7	753	4.6e-7
600	200	671	1.4e-6	640	1.1e-6
600	400	3812	5.7e-7	1880	1.3e-7
600	800	2892	7.0e-7	2066	1.1e-7
1200	100	1598	1.1e-6	1543	2.0e-6
1200	200	1302	7.2e-7	1024	2.1e-6
1200	400	1752	7.9e-7	1932	7.6e-7
1200	800	2980	7.8e-7	2691	3.0e-7

the meaning of the columns in Table 6.2 is the same as in Table 6.1; the number k is that of the commodities (different kinds of flows).

Table 6.2 shows that the two approaches are comparable. The results could be showing a trend whereby NMBundle is more competitive with PBundle for smaller-size problems, while the latter tends to be better on larger-size ones. While this would need to be confirmed by further experiments, it could be explained by the fact that the techniques developed for the management of t in section 5.1 are all *short-term*; that is, they consider only what is happening in the current iteration. For PBundle, *long-term* strategies have been developed (see, e.g., [7] for the problem at hand) which have been shown to be useful for large-scale, difficult problems. Hence, these preliminary results may be an indication that appropriate long-term strategies for t management are required also in the nonmonotone approach.

7. Conclusions and future research. We have developed a new variant of the proximal bundle approach for convex nondifferentiable optimization whose main characteristic is that of being truly nonmonotone while staying very close to the original proximal bundle idea, up to using the very same master problem. This is obtained by employing, instead of the function value, an (apparently) novel merit function, derived from the Moreau–Yoshida regularization, to drive the search toward optimality. This approach would in general lead to a nondichotomic choice for the stepsize, which is intuitively appealing; however, the simple choice of the upper model as a linear function implies that one of the two extreme choices for the step is always optimal, in fact bringing back the algorithm to performing either null steps or serious steps. Yet, the convergence of the approach can now be studied as that of a single process, which provides a much simpler way to deriving efficiency estimates.

These results may provide new insight on the theory of bundle methods for nondifferentiable optimization which may ultimately lead to the development of variants of these algorithms that are actually more efficient in practice. For instance, the efficiency estimates seem to indicate that the main culprit of the low theoretical convergence rate of these algorithms lies in the "bad sublinear part" of the convergence of sequences of consecutive null steps (cf. (5.2)), thus possibly highlighting a crucial target for future improvements: if that part of the convergence could be made faster, a theoretically (and, hopefully, practically) better algorithm would ensue. We remark that that part of the convergence estimate is precisely the one where there is no difference between using a large bundle or a poorman's one, which is perhaps the main

reason behind the wide gap between the appalling worst-case theoretical performance of the algorithm and the much better (at times) observed practical one; this may suggest some direction for future research. Yet, there are several other possible research directions, e.g., extending the results to generalized bundle methods [10] or developing long-term strategies for t management that result in theoretically and/or practically more effective algorithms.

REFERENCES

[1] A. Astorino, A. Frangioni, M. Gaudioso, and E. Gorgone, *Piecewise quadratic approximations in convex numerical optimization*, SIAM J. Optim., 21 (2011), pp. 1418–1438.

[2] A. Astorino, A. Fuduli, and M. Gaudioso, *DC models for spherical separation*, J. Global Optim., 48 (2010), pp. 657–669.

[3] A. Astorino, A. Fuduli, and M. Gaudioso, *Margin maximization in spherical separation*, Comput. Optim. Appl., 53 (2012), pp. 301–322.

[4] F. Babonneau and J.-P. Vial, *ACCPM with a nonlinear constraint and an active set strategy to solve nonlinear multicommodity flow problems*, Math. Program., 120 (2009), pp. 179–210.

[5] O. Briant, C. Lemaréchal, P. Meurdesoif, S. Michel, N. Perrot, and F. Vanderbeck, *Comparison of bundle and classical column generation*, Math. Program., 113 (2008), pp. 299–344.

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

[7] T. 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.

[8] F. Facchinei and S. Lucidi, *Nonmonotone bundle-type scheme for convex nonsmooth minimization*, J. Optim. Theory Appl., 76 (1993), pp. 241–257.

[9] A. Frangioni, *Solving semidefinite quadratic problems within nonsmooth optimization algorithms*, Comput. Oper. Res., 21 (1996), pp. 1099–1118.

[10] A. Frangioni, *Generalized bundle methods*, SIAM J. Optim., 13 (2002), pp. 117–156.

[11] A. Frangioni and G. Gallo, *A bundle type dual-ascent approach to linear multicommodity Min Cost flow problems*, INFORMS J. Comput., 11 (1999), pp. 370–393.

[12] A. Frangioni and B. Gendron, 0-1 *reformulations of the multicommodity capacitated network design problem*, Discrete Appl. Math., 157 (2009), pp. 1229–1241.

[13] A. Frangioni and E. Gorgone, *Generalized bundle methods for sum-functions with "easy" components: Applications to multicommodity network design*, Math. Program., to appear.

[14] A. Frangioni, A. Lodi, and G. Rinaldi, *New approaches for optimizing over the semimetric polytope*, Math. Program., 104 (2005), pp. 375–388.

[15] L. Hou and W. Sun, *On the global convergence of a nonmonotone proximal bundle method for convex nonsmooth minimization*, Optim. Methods Softw., 23 (2008), pp. 227–235.

[16] E. Karas, A. Ribeiro, C. Sagastizábal, and M. Solodov, *A bundle-filter method for nonsmooth convex constrained optimization*, Math. Program., 116 (2009), pp. 297–320.

[17] K. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program., 46 (1990), pp. 105–122.

[18] K. Kiwiel, *Efficiency of proximal bundle methods*, J. Optim. Theory Appl., 104 (2000), pp. 589–603.

[19] K. Kiwiel, *An alternating linearization bundle method for convex optimization and nonlinear multicommodity flow problems*, Math. Program., 130 (2011), pp. 59–84.

[20] K. Kiwiel and C. Lemaréchal, *An inexact bundle variant suited to column generation*, Math. Program., 118 (2009), pp. 177–206.

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

[22] C. Sagastizábal, *Composite proximal bundle method*, Math. Program., 140 (2013), pp. 189–233.