Università di Pisa
Dipartimento di Informatica

Technical Report: TR-12-09

A Nonmonotone Proximal Bundle Method With (Potentially) Continuous Step Decisions

A. Astorino A. Frangioni A. Fuduli E. Gorgone

ADDRESS: Largo B. Pontecorvo 3, 56127 Pisa, Italy. TEL: +39 050 2212700 FAX: +39 050 2212726

---

[*] Istituto di Calcolo e Reti ad Alte Prestazioni C.N.R., c/o Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: astorino@icar.cnr.it

[†] Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italia. E-mail: frangio@di.unipi.it

[‡] Dipartimento di Matematica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: antonio.fuduli@unical.it

[§] Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: egorgone@deis.unical.it

A Nonmonotone Proximal Bundle Method With (Potentially) Continuous Step Decisions

A. Astorino* A. Frangioni† A. Fuduli‡ E. Gorgone§

Abstract

We discuss a numerical algorithm for minimization of a convex nondifferentiable function belonging to the family of proximal bundle methods. Unlike all of its brethren, the approach does not rely on measuring descent of the objective function at the so-called "serious steps", while "null steps" only serve at improving the descent direction in case of unsuccessful 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. Simple choices for the step selection actually reproduce the dichotomic 0/1 behavior of standard proximal bundle methods, but shedding new light on the rationale behind the process, and ultimately with different rules. Yet, using nonlinear upper models of the function in the step selection process can lead to actual fractional steps.

Keywords*: Nonsmooth optimization, bundle methods, nonmonotone algorithm*

1 Introduction

We are concerned with the numerical solution of the problem

$$f^* = \inf \{ f(x) : x \in X \} \quad , \tag{1}$$

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* (for the sake of conciseness, just "bundle methods" in the following unless otherwise stated) are known to be among the most efficient implementable algorithms for solving this class of problems. They only require that f is known through an oracle ("black box") that, given any $x \in X$, returns the values $f(x)$ and $z \in \partial f(x)$. They work by generating a sequence of *tentative points* $\{x_i\}$ where the oracle provides the f-value $f_i = f(x_i)$ and any subgradient $z_i \in \partial f(x_i)$. Then, at each iteration, a *bundle* of information $\mathcal{B} = \{(x_i, f_i, z_i)\}$ is maintained to construct a *model* $f_\mathcal{B}$ of the function f, that is exploited to construct the next tentative point. This is done by taking a distinguished vector $\bar{x}$ as the *current point*, and solving a (primal) *master problem*

$$\phi_{\mathcal{B},t}(\bar{x}) = \inf \{ f_\mathcal{B}(\bar{x} + d) + \tfrac{1}{2t}||d||^2 \} \quad , \tag{2}$$

which provides a *tentative descent direction* d^* along which the next tentative point is generated. The master problem seeks for a minimum of the current model $f_\mathcal{B}$ plus the *stabilizing term* $\frac{1}{2t}||d||^2$ that discourages points "far away" (in the standard Euclidean norm) from $\bar{x}$, where $f_\mathcal{B}$ is presumably a "bad approximation" of f; $t > 0$ is the *proximal parameter* dictating the "strength" of the stabilization.

In all proximal bundle methods 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, $(x, f(x), z \in \partial f(x))$ 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 of t, these approaches can be shown to minimize f. However, the convergence proofs are somewhat complicated because two basically distinct processes are going on:

- During sequences of consecutive NS, one is actually aiming at solving the *stabilized primal problem*

  $$\phi_t(\bar{x}) = \inf \left\{ \; f(\bar{x} + d) + \tfrac{1}{2t}||d||^2 \; \right\} \quad . \tag{3}$$

  This amounts at computing the *Moreau–Yosida regularization* ϕ_t of f in $\bar{x}$, which has the same set of minima as f (cf. §3) but is smooth, making it easier to construct descent algorithms. However, solving (3) with the sole help of the black box for f is, in principle, as difficult as solving (1); therefore, bundle methods solve (3) *approximately and iteratively* during sequences of NS by a *standard cutting-plane-type approach* [20]. Indeed, a NS happens when $f_{\mathcal{B}}(x) \ll f(\bar{x})$ but $f(x) \not\ll f(\bar{x})$; in this case, the new information $(x, f(x), z)$ will "significantly enrich" $\mathcal{B}$, leading to a new master problem (2) that will eventually provide a better approximation to the solution of (3).

- After SS is declared, the master problem (2) bears basically no relation with that before the SS. Indeed, in the convergence proofs one is typically allowed to completely change $\mathcal{B}$ after each SS.

In other words, convergence of (proximal) bundle methods is usually analyzed as if being made by two almost unrelated processes: sequences of consecutive NS, where an approximate solution to (3) is computed for the given $\bar{x}$ and t, (possibly) interrupted by the sought-after SS. However, at each SS the algorithm can basically be restarted from scratch (apart from keeping $\bar{x}$), so each sequence of NSs is basically seen as being almost completely unrelated with all the others.

It can be argued that, at least in practice, this leads to a substantial underestimation of the rate of convergence of these methods. Indeed, the (few) available theoretical estimates [22] depict an almost hopelessly slow method, even worse than what methods of this kind need necessarily be [25]. While indeed practical convergence can be slow, and tailing-off often rears its ugly head [6], 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 [15, 16]. 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" decision to be made at each step between NS and SS 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, the convergence of the algorithm is proven by monitoring *one single* merit function, and showing that it is improved at any step, so that eventually optimality is reached. Doing so requires doing away with monotonicity in the form that is usually found in bundle methods, i.e., that of the objective function value between two consecutive SS (which is, anyway, a somewhat awkward notion). Thus, the algorithm we propose is nonmonotone in that particular sense, although of course it *is* monotone w.r.t. its newly defined merit function. Note that this is substantially different from the previously proposed nonmonotone bundle methods [9, 18], that are based on the well-known (and somewhat "dirty") trick of setting a fixed k and requiring that the SS improves over the worst among the last k values of the current point. Even the filter-bundle approach of [19] for constrainted optimization, which is arguably nonmonotone in some sense, uses the objective function value to define the "filter"; the nonmonotone behavior is dictated by the need to balance function value improvements with constraint satisfaction improvements, while our approach is nonmonotone even in the unconstrained case. There are "truly nonmonotone" bundle methods, in particular the *level* ones [25] which always move the current point to the last iterate, making up for convergence with a clever update of the "stabilizing device", that is quite different from that of the proximal bundle methods we study here. Perhaps not by chance, level bundle methods have the best efficiency estimates among all bundle methods; however they require some compactness assumptions that may not be easily 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 bring 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 from the standard proximal bundle method in order to motivate our approach; we also show why a "naïve" initial version of our main idea does not work. To overcome this latter problem, in Section 3 we introduce and analyze the merit function we employ, deriving some useful results. 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 on-line management of the proximal parameter (§5.1), provide speed-of-convergence estimates (§5.2), and briefly discuss the impact of employing nonlinear *upper* models of f (§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 on some significant test function in relation to that of the standard proximal bundle method, and in Section 7 we draw some conclusions.

2 Motivation

In this section we motivate the key ideas in our development. To do that, we need to introduce some notation. In the following, the *lower model* $f_{\mathcal{B}}$ has to be intended as the ordinary *cutting plane model*

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

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

$$\alpha_i(\bar{x}) = f(\bar{x}) - [\ f_i + z_i(\bar{x} - x_i)\ ] \tag{4}$$

is the *linearization error* of the subgradient z_i obtained at x_i w.r.t. $\bar{x}$; in other words,

$$z_i \in \partial_{\alpha_i(\bar{x})} f(\bar{x})\ \ . \tag{5}$$

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

$$\alpha_i(x) = z_i\big(\ \bar{x} - x\ \big) + \alpha_i(\bar{x}) + \big(\ f(x) - f(\bar{x})\ \big) \tag{6}$$

(just write (4) for x and $\bar{x}$ and simplify out common terms). In standard approaches, the current point is mostly regarded as being fixed, and thus there is no need to stress the fact that the linearization errors depend on $\bar{x}$; 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 (6) gives

$$\alpha_i(x(\lambda)) = \alpha_i(\bar{x}) + f(x(\lambda)) - f(\bar{x}) - \lambda z_i d^*\ \ , \tag{7}$$

to simplify the notation we will denote $\alpha_i(x(\lambda))$ simply as α_i^{λ}. Note, however, that (7) requires knowledge of $f(x(\lambda))$, that usually is known only for $\lambda \in \{0, 1\}$. To further ease notation, in the following we will often use the shorthand "$i \in \mathcal{B}$" for "$(\alpha_i, z_i) \in \mathcal{B}$"; let us also remark immediately that while the index "i" can upon a first reading be considered that of the iteration, in general the bundle evolves in rather different ways. All this allows us to rewrite $\hat{f}_{\mathcal{B}}$ as

$$\hat{f}_{\mathcal{B}}(x) = \max\{\ z_i(x - \bar{x}) - \alpha_i\ :\ i \in \mathcal{B}\ \}\ +\ f(\bar{x})\ \ .$$

With this choice, the master problem (2) becomes the simple QP

$$\min\ \{\ v + \tfrac{1}{2t}||d||^2\ :\ v \geq z_i d - \alpha_i \quad i \in \mathcal{B}\ \} \quad [+f(\bar{x})]\ \ . \tag{8}$$

Note that the constant term "$+f(\bar{x})$" is most often disregarded, but it is crucial in our development, as we shall see. Solving (8) is equivalent to solving its dual

$$\min\ \Big\{\ \tfrac{1}{2}t\ \big\|\textstyle\sum_{i\in\mathcal{B}} z_i\theta_i\big\|^2 + \sum_{i\in\mathcal{B}} \alpha_i\theta_i\ :\ \theta \in \Theta\ \Big\} \quad [-f(\bar{x})]\ \ , \tag{9}$$

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. Standard duality results show that $v(8) = -v(9)$ (where $v(\cdot)$ denotes the optimal value of an optimization problem), and that the dual optimal solution θ^* of (9), appropriately "translated in the (z, α)-space" by

$$z^* = \textstyle\sum_{i\in\mathcal{B}} z_i\theta_i^* \qquad\qquad \alpha^* = \sum_{i\in\mathcal{B}} \alpha_i\theta_i^*\ \ , \tag{10}$$

gives the primal optimal solution (v^*, d^*) of (8) as

$$d^* = -tz^* \qquad\qquad v^* = -t||z^*||^2 - \alpha^* \ . \tag{11}$$

The dual form of the master problem is not only useful for algorithmic purposes [10], but it is also closely related to the stopping criterion of the method. In fact, from

$$z^* \in \partial_{\alpha^*} f(\bar{x}) \tag{12}$$

(an immediate consequence of (5)) one immediately realizes that, whenever $z^* = 0$ and $\alpha^* = 0$, $\bar{x}$ is optimal. In practice, one would therefore stop when the two non-negative numbers $\|z^*\|$ and α^* are "small". One possible way of implementing this is to choose a scaling factor t^* and stop whenever

$$s^* = t^* \|z^*\|^2 + \alpha^* \leq \varepsilon, \tag{13}$$

where ε is the (absolute) accuracy required to the objective function value. In general one wants to on-line tune t, allowing it to shrink and grow to adapt to the scaling of f in the neighborhood of the current $\bar{x}$, while leaving t^* 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 $t^* = t$; further discussion is provided later on.

In the standard approach, after that (8)/(9) are solved, one probes $f(x)$ (with $x = \bar{x} + d^*$) and decides whether or not moving $\bar{x}$ to x. Of course, setting $\bar{x} = x$ is quite natural whenever $f(x) \ll f(\bar{x})$ (after all, we are minimizing f); on the other hand, it is not entirely obvious that this is necessarily the best choice. Indeed, what one would really want is that s^* *decreases as fast as possible*, so that (13) is attained as early as possible. But choosing $\bar{x} = x$ is not necessarily the best way to decrease it; indeed, any decrease in s^* is brought about at best indirectly by the fact that the α_i's change when the SS is performed. In particular, the new pair (z, α) obtained evaluating $f(x)$ has $\alpha(x) = \alpha^1 = 0$, i.e., z has a "small linearization error" if the SS is done. However, there is no guarantee at all that (13) has improved after the SS.

Therefore, one may wonder whether, even if $f(x) \ll f(\bar{x})$, a different move than a standard SS could be better in terms of improvements of (13). A possibility is to consider the line segment $\mathcal{L} = conv(\,\{\bar{x}, x\}\,) = \{\, x(\lambda) \,:\, \lambda \in [0,1] \,\}$, and try to *determine the optimal value* λ^* *of* λ *that minimizes* s^* *at the next iteration*; then the new current point $\bar{x}$ is set to $x(\lambda^*)$. We observe in passing, that, instead of $\mathcal{L}$, one could consider $conv(\,\{x_i \,:\, i \in \mathcal{B}\} \cup \{x\}\,)$, which requires keeping track of the iterates x_i. Because this would surely make the development much more complex, we refrain from exploring this option; clearly, any approach that works with $\mathcal{L}$ will *a fortiori* work if more information is provided.

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. (7)). 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:

$$\begin{array}{rlr} \bar{\alpha}_i^\lambda = & f^{\mathcal{B}}(\lambda) - [\; f_i + z_i(\, x(\lambda) - x_i \,) \;] & \geq \\ & f^{\mathcal{B}}(x(\lambda)) - [\; f_i + z_i(\, x(\lambda) - x_i \,) \;] & = \alpha_i^\lambda \end{array} \tag{14}$$

(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 [23]). In particular, the "obvious" upper model is

$$f^{\mathcal{B}}(\lambda) = (1-\lambda)f(\bar{x}) + \lambda f(x) = f(\bar{x}) + \lambda \Delta f \tag{15}$$

where $\Delta f = f(x) - f(\bar{x})$ is the all-important difference between the f-values of x and $\bar{x}$. This is clearly the "worst case" model: no upper model need (and therefore should) be above (15), and being below only improves the accuracy of (14), so we will mainly work with (15) since then results will *a fortiori* hold with any more accurate upper model (besides is has some computational advantage, cf. §4.1). Note that (15) can only be computed after that x has been selected and $f(x)$ evaluated: hence, strictly speaking one can only do that after that $(f(x), z)$ has been added to $\mathcal{B}$, but we will allow ourselves this small abuse of notation. A more serious consequence is that one cannot use the upper model to drive the selection of the next iterate x; yet, for this task one can still use (as customary in bundle methods) the lower model $f_{\mathcal{B}}$, which is well-defined everywhere, restricting the use of $f^{\mathcal{B}}$ to the selection of λ, i.e., of the next current point $x(\lambda) \in \mathcal{L}$.

An issue with (15) is that $\bar{x}$ itself may be the $x(\lambda)$ of the previous iteration: hence $f(\bar{x})$ is actually unknown, and the above formula cannot be directly applied. However, one can obviously use $f^{\mathcal{B}}(\bar{x})$ in place of $f(\bar{x})$ in (15) and still obtain a valid upper model; that is the "recursive" definition

$$\bar{f}^{\mathcal{B}}(\lambda) = (1-\lambda)\bar{f}^{\mathcal{B}}(\bar{x}) + \lambda f(x) = \bar{f}^{\mathcal{B}}(\bar{x}) + \lambda \Delta \bar{f} \enspace , \tag{16}$$

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

$$\bar{\alpha}_i^{\lambda} \;=\; \bar{f}(\lambda) - [\, f_i + z_i(x(\lambda) - x_i)\,] \;=\; \bar{\alpha}_i(\bar{x}) + \lambda \Delta \bar{f} - \lambda z_i d^* \geq \alpha_i^{\lambda} \tag{17}$$

(cf. (7)). Thus, we can always assume knowledge of an upper model like (16), which for notational convenience we will most often denote simply as $\bar{f}$: any way to construct an upper model based on exact information should easily extend if some of the data actually comes from an upper estimate of f instead. Doing so allows us to consider the *family of QP dual pairs, parametric over* λ

$$-\delta(\lambda) = \quad \min \{\, v + \tfrac{1}{2t}\|d\|^2 \;:\; v \geq z_i d - \bar{\alpha}_i^{\lambda} \quad i \in \mathcal{B} \,\} \qquad [\, +f(x(\lambda))\,] \tag{18}$$

$$\delta(\lambda) = \quad \min \left\{ \tfrac{1}{2}t \left\| \textstyle\sum_{i \in \mathcal{B}} z_i \theta_i \right\|^2 + \textstyle\sum_{i \in \mathcal{B}} \bar{\alpha}_i^{\lambda} \theta_i \;:\; \theta \in \Theta \right\} \quad [\, -f(x(\lambda))\,] \enspace . \tag{19}$$

The optimal solution $\theta^*(\lambda)$ of (19), or better its representative in the (z, α)-space

$$z(\lambda) = \textstyle\sum_{i \in \mathcal{B}} z_i \theta_i^*(\lambda) \qquad\qquad \alpha(\lambda) = \textstyle\sum_{i \in \mathcal{B}} \bar{\alpha}_i^{\lambda} \theta_i^*(\lambda) \enspace , \tag{20}$$

immediately gives (cf. (12) and (14))

$$z_i \in \partial_{\bar{\alpha}_i^{\lambda}} f(x(\lambda)) \quad \Longrightarrow \quad z(\lambda) \in \partial_{\alpha(\lambda)} f(x(\lambda)) \enspace . \tag{21}$$

Note that (21) refers to the "true" $f(x(\lambda))$ rather than its approximation $\bar{f}(\lambda)$. That is, while linearization errors are only estimated, the estimate is "kept in check" by the fact that at each application of (17) we do use the "true" value of $f(x)$ to compute $\Delta \bar{f}$. Said otherwise, applying (17) twice cancels out the term "$\bar{f}^{\mathcal{B}}(\bar{x})$" corresponding to the middle point, and therefore any error in that term: 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, each time (if ever) a SS is done ($\lambda^* = 1 \Rightarrow x(\lambda) = x \Rightarrow f(\lambda) = f(x)$) the linearization errors all become "exact" again. Also, we remark for future reference that $\delta(\lambda) = v(19) = -v(18)$ is a *concave* function. Indeed, λ appears linearly in the right-hand-side of the constraints in (18); therefore, $-\delta(\lambda)$ is the *value function* of a convex problem, and hence convex.

Due to (21), the stopping condition (13)—with $z(\lambda)$ and $\alpha(\lambda)$ replacing z^* and α^*, respectively—can still be applied. The idea of the approach is then to seek the value of λ such that $z(\lambda)$ and $\alpha(\lambda)$ are "the best possible" for (13). In other words, one would like to find the optimal solution λ^* of

$$\min \left\{\, s^*(\lambda) = t^* \|z(\lambda)\|^2 + \alpha(\lambda) \;:\; \lambda \in [0,1] \,\right\} \tag{22}$$

and set $\bar{x} = x(\lambda^*)$, in order to (hopefully) obtain (13) faster. However, it is easy to prove that the approach, as it is stated, might not work in general. Indeed, consider the linear case $f(x) = rx$ for one fixed $r \in \mathbb{R}^n$. Clearly this problem is unbounded below, and a standard bundle algorithm would "prove it" by making infinitely many consecutive SSs along the direction $-r$. It is also easy to see that $\mathcal{B}$ would contain 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 (22), and nothing can prevent the approach, as previously stated, to always select $\lambda^* = 0$, thereby dramatically failing to solve the problem.

The example shows that $s^*(\lambda)$ of (22) is not an appropriate merit function for our problem. In the next paragraph 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* (18)/(19) (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. (3)) of f. We start recalling some of its well-know properties:

$$\phi_t \leq f \tag{23}$$

$$\phi_t(x) = f(x) \quad \Longleftrightarrow \quad x \text{ optimal for (1)} \tag{24}$$

Indeed, $d = 0$ is a feasible solution to (3), which gives (23). Furthermore, the optimality conditions of (3)

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

clearly imply that $d^* = 0 \iff x$ is optimal for (1), but $d^* = 0 \iff \phi_t(x) = f(x)$, whence (24). As already remarked, ϕ_t is only a "conceptual" object because it is difficult to compute (the oracle being for f); however, owing to $f_{\mathcal{B}} \leq f$, for the "readily available" $\phi_{\mathcal{B},t}$ (cf. (2)) one clearly has

$$\phi_{\mathcal{B},t} \leq \phi_t \; [\, \leq f \,] \tag{25}$$

$$\phi_{\mathcal{B},t} = f(x) \qquad \Longrightarrow \quad x \text{ optimal for (1).} \tag{26}$$

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

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

and the "implementable" form

$$\zeta_{\mathcal{B},t}(x) = 2f(x) - \phi_{\mathcal{B},t}(x) \;\; [\, \geq \zeta_t \,] \;\; . \tag{27}$$

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)

$$\delta_t(x) = \zeta_t(x) - f(x) = f(x) - \phi_t(x) \; [\, \geq 0 \,] \tag{28}$$

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

$$\zeta_t \geq f \tag{29}$$

$$\zeta_t(x) = f(x) \quad \Longleftrightarrow \quad x \text{ optimal for (1)} \tag{30}$$

via (23)–(24). This and (27) then immediately give

$$\zeta_{\mathcal{B},t} \geq \zeta_t \; [\, \geq f \,] \tag{31}$$

$$\zeta_{\mathcal{B},t} = f(x) \qquad \Longrightarrow \quad x \text{ optimal for (1)} \tag{32}$$

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 (18)/(19) is

$$\begin{aligned} \zeta_{\mathcal{B},t}(x(\lambda)) = 2f(x(\lambda)) - \phi_{\mathcal{B},t}(x(\lambda)) & \qquad\qquad = 2f(x(\lambda)) - v(18) \\ = f(x(\lambda)) + t\|z(\lambda)\|^2/2 + \alpha(\lambda) & \qquad\qquad = f(x(\lambda)) + \delta_{\mathcal{B},t}(x(\lambda)) \end{aligned} \tag{33}$$

due to $v(18) = -v(19)$ and the fact that the constant term "$-f(x(\lambda))$" in (19) cancels out with the factor of 2.

Hence, ζ_t is a "complementary" function to ϕ_t: both coincide with f only at optimality, but the former is an upper approximation, whereas the latter is a lower approximation. Of course, a difference of much greater magnitude is that while ϕ_t is convex, ζ_t is, in general, *not* (although clearly it is a DC function, as is δ_t); this is shown in Figure 1 for the simple piecewise-linear function $f(x) = \max\{\, -3x + 8, -x + 4, 1, x - 3, 2x - 9 \,\}$. Thus, while minimizing $\phi_{\mathcal{B},t}$ (and refining $\mathcal{B}$ as needed) is an attractive option, the same cannot be said for $\zeta_{\mathcal{B},t}$. Yet, as a merit function ζ_t is "no (much) less powerful" than ϕ_t, at least on converging (sub) sequences.

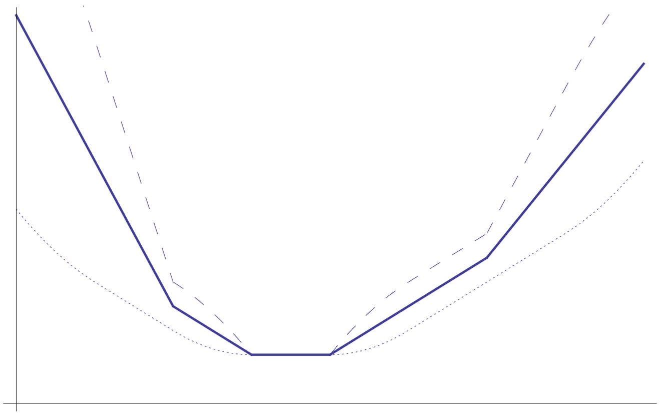

Figure 1: ϕ_t (dotted line) and ζ_t (dashed line) for a simple f (thick solid line).

Lemma 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).*

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 semi-continuous,

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

and the result follows from (32). ■

The interesting aspect of Lemma 1 is that nothing specific is required 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, it is important to remark as to why $t_i \geq \underline{t} > 0$ is required. The point is that there is an easy but "fictitious" way to ensure $\zeta_t(\bar{x}) = f(\bar{x})$: just take $t = 0$ (or, in a sequence setting, have $t_i \to 0$ fast). This does not harm much the Moreau–Yosida regularization, 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, this setting is disastrous for ζ_t: $\zeta_0 - f = \delta_0 \equiv 0$, hence the merit function is of no use whatsoever for detecting minima of f. However, requiring t to be bounded away from zero is less than ideal: bundle algorithms (and not only them) converge even if $t_i \to 0$, provided this happens "slowly enough" [7]. Furthermore, Lemma 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, 11], is the following:

Definition 2 *Let* $S_\delta(f) = \{\, x \,:\, f(x) \leq \delta \,\}$ *be the level set corresponding to the* f*-value* δ*: a function* f *is* **-compact if* $\forall\, \bar{l} \geq \underline{l} > f^* \geq -\infty$

$$e(\underline{l}, \bar{l}) = \sup_x \{\, dist(x, S_{\underline{l}}(f)) \,:\, x \in S_{\bar{l}}(f) \,\} < \infty \quad . \tag{34}$$

A function is *-compact if 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 [11] 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 1 to non-converging sequences. To do that, we must first bound below the gap function in terms of the remaining "distance" to the optimal value.

Lemma 3 *For each* x *and* $\tilde{x}$ *such that* $f(\tilde{x}) < f(x)$ *and* $g \in \partial_\varepsilon f(x)$*, it holds*

$$\delta_t(x) \geq \min \left\{ \frac{(f(x) - f(\tilde{x}))^2 t}{2||\tilde{x} - x||^2} \,,\, \frac{(f(x) - f(\tilde{x}) - \varepsilon)^2}{2t||g||^2} \right\} \quad . \tag{35}$$

Proof. By the very definition (3) we have

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

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

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

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

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

where $\Delta f = f(x) - f(\tilde{x}) > 0$ and $\Delta x = ||\tilde{x} - x||^2 > 0$. The unconstrained maximum of ψ is

$$\tilde{\beta} = t\Delta f / \Delta x > 0 \quad \text{with} \quad \psi(\tilde{\beta}) = \frac{t(\Delta f)^2}{2\Delta x} \ .$$

Therefore, the optimal solution to the maximization problem is

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

which immediately gives that its optimal value is

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

For the case $\Delta f \geq \Delta x / t$, we have that

$$\psi(\beta^*) = \Delta f - \Delta x/(2t) \geq \Delta x/(2t)$$

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:

$$f(\tilde{x}) \geq f(x) + g(\tilde{x} - x) - \varepsilon \quad \Rightarrow \quad ||g|| \cdot ||x - \tilde{x}|| \geq \Delta f - \varepsilon \ .$$

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

Lemma 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, $0 < t_i \leq \bar{t} < \infty$, and*

$$\liminf_{i \to \infty} \delta_{\mathcal{B}_i, t_i}(\bar{x}_i)/t_i = 0 \tag{36}$$

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

Proof. Because $\zeta_i \geq \bar{f}_i$ and ζ_i is nonincreasing, the sequence $\{\ \bar{f}_i\ \}$ is bounded above: $\bar{f}_i \leq \bar{l} < \infty$. If $f_\infty = -\infty$, then the theorem is proven: clearly, $f^* = -\infty$ and $\{\ \bar{x}_i\ \}$ is an optimizing sequence. Otherwise, assume by contradiction that $\bar{f}_i - f^* > \gamma > 0$ and set $-\infty < \underline{l} = \inf\{\ \bar{f}_i\ \} - \gamma \leq \bar{l} - \gamma$. Since $(\delta_i = \delta_{\mathcal{B}_i, t_i}(\bar{x}_i))/t_i = ||z_i^*||^2/2 + \alpha_i^*/t_i$, (36) 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 with $\tilde{x}$ the projection of $\bar{x}_i$ over the level set $S_{\underline{l}}(f)$ (so that $\bar{f}_i - f(\tilde{x}_i) \geq \gamma$), $g = z_i^*$ and $\varepsilon = \alpha_i^*$: using (34) in (35) gives

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

Since $\alpha_i^* \to 0$, eventually $\gamma - \alpha_i^*$ 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

$$\delta_i = \delta_{t_i}(\bar{x}_i)/t_i \geq \gamma^2 / (2e(\underline{l}, \bar{l})^2)$$

which contradicts (36) and hence proves the result. ■

Hence, under mild conditions on f any sequence of $\{\bar{x}_i\}$ that nullifies the gap function is a minimizing one; t_i can even be allowed to 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 more familiar way to obtain (36) is

$$\textstyle\sum_{i=1}^\infty t_i = \infty \quad \text{and} \quad \sum_{i=1}^\infty \delta_i < \infty \quad : \tag{37}$$

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 (37) implies (36) since $\liminf_{i\to\infty} \delta_i/t_i > 0$ implies that for some $\delta > 0$, a large enough h and all $i \geq h$ one has $\delta_i \geq \delta t_i$, hence $\sum_{i=h}^{\infty} \delta_i \geq \delta \sum_{i=h}^{\infty} t_i$ which contradicts (37). Alternatively, if t_i is bounded away from zero then (36) 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 1.

All this gives us confidence that, under appropriate conditions, $\zeta_{\mathcal{B},t}$ can be used to construct a non-monotone bundle method along the lines of §2. At least, using $\zeta_{\mathcal{B},t}$ solves the linear function counterexample. Conceptually, this first requires working with the further approximate version of (33) employing the upper model

$$\bar{\zeta}_{\mathcal{B},t}(x(\lambda)) = \bar{f}(\lambda) + t\|z(\lambda)\|^2/2 + \alpha(\lambda) \geq \zeta_{\mathcal{B},t}(x(\lambda)) \tag{38}$$

(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 (38) w.r.t. (22), 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 details in the next paragraph.

4 The algorithm

We are now in the position to formally presenting the algorithm and discussing its main properties. To simplify on 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 (38), the corresponding approximate gap function $\bar{\delta}_{\mathcal{B},t}$ of (28), and so on. Hence, all references e.g. to the master problems (8)/(9), their solutions (10) and (11), 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. §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 (9) for $\theta^*$. Find $z^*$, $\bar{\alpha}^*$ and $d^*$ from (10) and (11). $\mathcal{B} = \mathcal{B} \cup \{ (z^*, \bar{\alpha}^*) \}$.
2. If (13) holds then stop ($\bar{x}$ is $\epsilon$-optimal).
3. Compute $x = \bar{x} + d^*$, $f(x)$, $z \in \partial f(x)$ and $\alpha$ via (4). $\mathcal{B} = \mathcal{B} \cup \{ (z, \alpha) \}$.
4. For $\mathcal{B}' \subseteq \mathcal{B}$, compute an approximately optimal solution $\lambda^* \in [0, 1]$ to
       $\min \{ \bar{\zeta}_{\mathcal{B}',t}(x(\lambda)) \; : \; \lambda \in [0, 1] \}$        (39)
5. Set $\bar{x}_+ = x(\lambda^*)$, $\mathcal{B}_+ \subseteq \mathcal{B} \cup \{ (z(\lambda^*), \alpha(\lambda^*)) \}$, the linearization errors of $\mathcal{B}_+$
   according to (17) and $0 < t_+ \leq \bar{t}$. Goto 1.
```

A few comments on the algorithm are useful.

- In the first iteration, where $d^* = 0$, $\bar{x} = x$ and $\mathcal{B} = \{ (z, \alpha) \}$, (39) is "degenerate": any $\lambda \in [0, 1]$ is optimal. Also, $\alpha = 0$ (which is what (4) gives if $f(\bar{x})$ is "silently" initialized to $f(x)$).
- The fundamental property that we want from the algorithm is

$$\bar{\zeta}_+ = \bar{f}(\bar{x}_+) + t_+\|z_+^*\|^2/2 + \bar{\alpha}_+^* \leq \bar{f}(\bar{x}) + t\|z^*\|^2/2 + \bar{\alpha}^* = \bar{\zeta} \ . \tag{40}$$

  This is easy to obtain by requiring monotonicity of t

$$t_+ \leq t \tag{41}$$

  (which clearly implies $t_+ \leq \bar{t}$), an ever-increasing bundle $\mathcal{B} \subseteq \mathcal{B}_+$ and "complete" information in (39), i.e. $\mathcal{B}' = \mathcal{B}$, since then $(z^*, \bar{\alpha}^*)$ is feasible for (19) at $\lambda = 0$, and therefore (39) can only improve on it. Things are somewhat more complicated if, as it should, some form of *bundle management* is allowed at Step 4. (while choosing $\mathcal{B}'$) and at Step 5. (while choosing $\mathcal{B}_+$). However, what is needed is clearly:

  - that $(z^*, \bar{\alpha}^*)$ is feasible for (19) at $\lambda = 0$;
  - that $(z(\lambda^*), \alpha(\lambda^*))$ is feasible for (9) 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 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 the so-called "poorman's Bundle method" [6]; here, this translates to $\mathcal{B}' = \{(z^*, \bar{\alpha}^*), (z, \alpha)\}$ in (39) and the "poorman's bundle" $\mathcal{B}_+ = \{\ (z(\lambda^*), \alpha(\lambda^*))\ \}$, in which case $1 = \theta^*_+ \in \mathbb{R}$ is the only feasible (hence optimal) solution to (9) at the next iteration, and therefore

$$(z^*_+, \bar{\alpha}^*_+) = (z(\lambda^*), \alpha(\lambda^*)) \ . \tag{42}$$

  A milder way to obtain the same result is to ensure that $z(\lambda^*)$ and $\alpha(\lambda^*)$ are still feasible for (9) at the beginning of the next iteration by inhibiting removal of any subgradient $h \in \mathcal{B}$ such that $\theta^*_h(\lambda^*) > 0$. However, this requires that these subgradients are actually "considered" during Step 4., which may not necessarily be the case (cf. the next point and §4.1).

- According to what exactly "approximately solve" means in Step 4., one could actually entirely avoid Step 1. and pass directly to Step 2. This is the case if $(z(\lambda^*), \alpha(\lambda^*))$ is already the optimal solution to (9) at the next iteration, which depends on exactly how (39) is solved and on the relationships between $\mathcal{B}$ and $\mathcal{B}_+$. For instance, it surely happens when $\mathcal{B}_+$ is the "poorman's bundle" (cf. (42)). However, this need not always happen. Indeed, since (39) is a "difficult" problem as $\bar{\zeta}_{\mathcal{B},t}$ is nonconvex, one reasonable strategy is to simplify it by making $\mathcal{B}$ "as small as possible" (cf. §4.1). However, working with a very restricted bundle has in general dire consequences on the convergence speed [6, 15], which makes it only advisable when the cost of the master problem would be unbearable otherwise [8]. So, in general one may want a "thin" $\mathcal{B}$ in (39) and a "fat" one in (9), which in turn requires to re-solve (9) once λ^* has been found. This is why we present this as the "default behavior" of our algorithm.

- At Step 5., one may actually 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. Clearly this can only improve things, but it comes at the cost of an extra function computation. Although there are cases where this may be worth, since the function computation is only a negligible part of the overall time (e.g. [15]), in general one can work with the approximate quantities, hence we prefer to develop our theory in the more general setting. Of course, this is only relevant if $\lambda^* \notin \{0, 1\}$, which may not happen often, if at all (cf. §4.1).

- Obviously, (41) is somewhat harsh. It is well-known that on-line tuning of t is important in practice, and ensuring that t_1 is chosen "large enough" may not be entirely trivial (even though $t_1 = t^*$ from (13) should in principle do). Besides, as already discussed in §3, some care is needed to avoid t_i going to zero too fast and thereby rendering $\bar{\zeta}$ useless (cf. (36)/(37)). These issues require a better grasp of the convergence properties of the approach, this is why we postpone their discussion to §5.1.

4.1 Finding λ^*

The solution of (39) clearly depends on the specific upper model $\bar{f}^{\mathcal{B}}$ employed. The linear upper model (16) is clearly one interesting possibility: it is very easy to compute and available for any f. However, as already discussed it is also the "worst case" upper model, and therefore the one relying on the most conservative (hence least accurate) estimates. Furthermore, it immediately leads to the fact that (39) is *concave* in λ: indeed, $\delta(\lambda)$ (cf. (19)/(18)) is concave, and $\bar{f}^{\mathcal{B}}$ is linear. As a consequence, the use of (16) leads to $\lambda^* \in \{0, 1\}$, i.e., to only making either NSs or SSs, although with different "rules" than in the traditional bundle method.

Indeed, a computationally useful consequence is that (39) is then easy: just solve it for $\lambda = 0$ and $\lambda = 1$ and pick the solution giving the best $\bar{\zeta}$-value. This is clearly possible, but requires solving the master problem twice at each iteration (although then step 1. could be skipped entirely, as already discussed). To avoid this it is possible to employ the already recalled aggregation technique, in particular with "extreme aggregation"

$$\mathcal{B}' = \{\ (z^*, \bar{\alpha}^*)\ ,\ (z, \alpha)\ \}\ , \tag{43}$$

so that the solution to (9) can be found by a closed algebraic expression. It is useful to develop the approach in detail, which requires introducing some notation; first and foremost, the two linear functions in λ

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

such that

$$f_{\mathcal{B}'}(x(\lambda)) = \max\ \{\ f_*(\lambda)\ ,\ f(\lambda)\ \}\ .$$

In particular, (17) 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 (19)

$$\delta(\lambda) = \min\ \left\{\ \tfrac{1}{2}t\,\|\theta z^* + (1-\theta)z\|^2 + \theta\bar{\alpha}^{*,\lambda} + (1-\theta)\bar{\alpha}^{\lambda}\ :\ \theta \in [0,1]\ \right\} \tag{44}$$

whose optimal solution has the closed-form expression

$$\theta^*(\lambda)\ =\ \min\left\{\ 1\ ,\ \max\left\{\ 0\ ,\ \tilde{\theta}(\lambda) = \frac{\bar{\alpha}^{\lambda} - \bar{\alpha}^{*,\lambda} - tz(z^* - z)}{t\|z^* - z\|^2}\ \right\}\right\}\quad , \tag{45}$$

since $\tilde{\theta}(\lambda)$ is the optimal solution of the *relaxation* of (44) 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 (39) (under (43)) 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 (39) under (43) can be written as

$$\begin{aligned}
&\min\ \left\{\ h(\theta,\lambda) = h(\theta) + \lambda\big(\,\Delta\bar{f} + tz^*(z^* - z)\theta - \alpha\,\big)\ :\ \theta \in [0,1]\ ,\ \lambda \in [0,1]\ \right\} \\
&\text{where} \qquad\qquad h(\theta) = t\|\theta z^* + (1-\theta)z\|^2/2 + \theta\bar{\alpha}^* + (1-\theta)\alpha
\end{aligned} \tag{46}$$

is the objective function of (44) for $\lambda = 0$. For the optimal solution (θ^*, λ^*) one then has

$$\begin{aligned}
\Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha > 0 \quad &\Longrightarrow \quad \lambda^* = 0 \\
\Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha < 0 \quad &\Longrightarrow \quad \lambda^* = 1
\end{aligned} \tag{47}$$

due since $h(\theta, \lambda)$ is linear in λ. This first confirms that (16) leads to $\lambda^* \in \{0, 1\}$, except for the vanishingly small chance that $\Delta\bar{f} + tz^*(z^* - z)\theta^* - \alpha = 0$. More tellingly, (47) can be interpreted as an "ex-post NS/SS rule", to be contrasted to the standard SS condition

$$\Delta\bar{f} \leq m[\ f_{\mathcal{B}}(x) - \bar{f}(\bar{x})\ ] \tag{48}$$

for some arbitrary $m \in (0, 1]$. Of course, (47) can only be evaluated after that (39) has been solved (albeit this is not a large computational burden); yet, one easily derives the *sufficient* conditions

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

that can be evaluated early on to avoid solving (39) (and computing some of the required scalar products in the process). These conditions show that the process is indeed nonmonotone: while (48) requires $\Delta\bar{f}$ to be negative, and "sizably so", to declare a SS, (47) only forces a 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 (47) gives the "ex-post"

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

(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 a NS to be declared.

Using the simplified problem corresponding to "extreme" aggregation (43) would make (39) solvable even when using non-linear 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) sub-interval of $[0, 1]$, and constant outside it. Thus, $\delta(\lambda)$ is quadratic inside that interval and linear outside it, as it is easy to verify with somewhat tedious algebra. Thus, if $f^{\mathcal{B}}$ is piecewise in $[0, 1]$ with a small number of pieces, each one of them being a simple (say, smooth algebraic) function (cf. §5.3), (39) can 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 "larger" $\mathcal{B}$ than (43). In fact, it is possible to show that the optimal solution of (19) is piecewise-linear in λ; this only requires some sensitivity analysis arguments along the lines of [10, §6] (there the parameter was t, but the result easily extends). Therefore, the analogous to $\theta^*(\lambda)$ can be constructed, and an explicit concave piecewise form for $\delta(\lambda)$ can be devised. However, the number of pieces grows with the set of all possible "bases" (optimal active sets) of (9), and therefore exponentially. Thus, while applying the idea would be possible with a "small" $\mathcal{B}$, the approach would rapidly become impractical as $|\mathcal{B}|$ grows. There could also be intermediate approaches between assuming (43) and complete enumeration of all possible bases. For instance, one may find the optimal λ^* under (43), then solve (19) with $\lambda = \lambda^*$, substitute $(z^*, \bar{\alpha}^*)$ with $(z(\lambda^*), \alpha(\lambda^*))$ in (43) and keep iterating as long as improvement is obtained. This could make sense if function evaluation is much more expensive than the master problem, i.e., the opposite situation as the one where computing $f(x(\lambda^*))$ makes sense. While all these modifications could be useful for some specific applications, we will concentrate on the simplest approach using the linear upper model (16) and the minimal bundle (43); all improvements on these would *a fortiori* converge (hopefully, faster).

5 The convergence proof

The idea of the convergence proof is, clearly, 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

$$\Delta\bar{\zeta} = \bar{\zeta} - \bar{\zeta}_+ \ , \tag{49}$$

a non-negative number due to (40), and prove that $\bar{\zeta}_i \to \bar{f}_i$ $(= \bar{f}(\bar{x}_i))$: we will do that by showing, basically, that $\Delta\bar{\zeta}$ at least as large as an appropriate 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 (44) of §4.1. Indeed, it is well-known that convergence *of a sequence of NSs* is retained even with the "poorman's bundle", which allows to prove convergence of the approach even if the maximum size of the bundle is fixed to any number ≥ 2. While this is very useful in practice, here we are rather interested in the fact that it allows to considerably simplify the convergence proof. In particular, we will study $\Delta\bar{\zeta}$ under the "extreme aggregation" assumption (43).

The second idea one can exploit, somewhat counter-intuitively (or maybe not, given §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 belong to $\{0, 1\}$, i.e., only NS and SS are done. That is, one could interpret (48) as a *heuristic for the solution of* (39):

$$\lambda^* = \begin{cases} 1 & \text{if (48) holds} \\ 0 & \text{otherwise} \end{cases} \ . \tag{50}$$

We note in passing that (48) can usually be replaced with the *weaker*

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

(cf. (33)) because $f_{\mathcal{B}}(x) - \bar{f}(\bar{x}) = v^* = -t\|z^*\|^2 - \bar{\alpha}^* \leq -t\|z^*\|^2/2 - \bar{\alpha}^*$ (cf. (11)). Since at each step either (48) 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 (41), as $\delta_{\beta,t}$ is increasing in t.

We start the case where (48) does not hold, and hence $\lambda^* = 0$; of course, 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 this decrease is possible by simple algebraic means, that have already been developed in §4.1. Indeed, for $h(\theta)$ of (46) one has

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

From (4) 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) (48) gives

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

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 (45), $\theta^* = \theta^*(0)$ 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

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

while the latter gives

$$\Delta\bar{\zeta} \geq h(1) - h(0) = t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z\|^2/2 - \alpha \quad .$$

To further develop this, combine $h'(0) = t(z^* - z)z + \bar{\alpha}^* - \alpha \geq 0$ (since h is convex and 0 is the constrained optimum) and (the contrary of) (48) to get

$$t\|z\|^2/2 + \alpha < \frac{1 + m}{2}(\, t\|z^*\|^2/2 + \bar{\alpha}^* \,)$$

and therefore

$$\Delta\bar{\zeta} \geq t\|z^*\|^2/2 + \bar{\alpha}^* - (\, t\|z\|^2/2 + \alpha \,) > \frac{1 - m}{2}(\, t\|z^*\|^2/2 + \bar{\alpha}^* \,) = \frac{1 - m}{2}\bar{\delta} \quad .$$

Combining the two cases we finally obtain

$$\Delta\bar{\zeta} \geq \frac{(1 - m)\bar{\delta}}{2} \min\left\{\, 1 \,,\, \frac{(1 - m)\bar{\delta}}{t\|z - z^*\|^2} \,\right\} \quad . \tag{51}$$

If (48) holds instead, and therefore we heuristically choose $\lambda^* = 1$, one has

$$\begin{aligned}\Delta\bar{\zeta} = \bar{\zeta} - \bar{\zeta}_+ &= \bar{f}(\bar{x}) + t\|z^*\|^2/2 + \bar{\alpha}^* - \big(\, f(x(1)) + t\|z(1)\|^2/2 + \alpha(1) \,\big) \\ &= -\Delta\bar{f} + t\|z^*\|^2/2 + \bar{\alpha}^* - t\|z(1)\|^2/2 - \alpha(1) \quad .\end{aligned} \tag{52}$$

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 2), 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. (48)) 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 $2m - 1 > 0$.

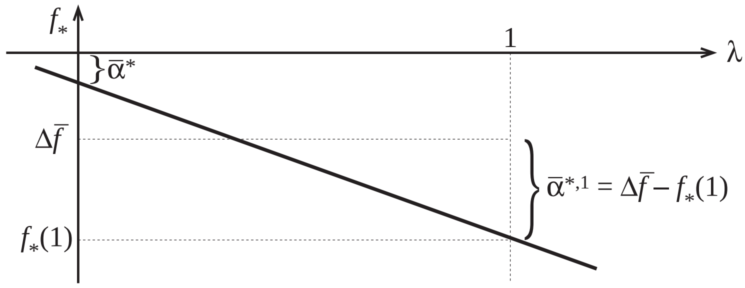

Figure 2: Estimating $\bar{\alpha}^{*,1}$

Since (48) either holds or not, we can conclude that

$$\Delta\bar{\zeta} \geq \bar{\delta} \min\left\{ 2m-1 \ , \ \frac{1-m}{2} \ , \ \frac{(1-m)^2\bar{\delta}}{2t||z-z^*||^2} \right\} \tag{53}$$

however chosen $m \in (1/2, 1]$. Note that this latter condition may seem restrictive when compared with that of standard proximal bundle methods, which only require $m \in (0, 1]$, but here m is not an actual algorithmic parameter (possibly requiring tuning) as in that case; rather, it is only a technicality in the proof.

We can now prove convergence of the algorithm. To simplify the proof, let us do away first with the obvious case where $\bar{f}_\infty = \liminf_{i\to\infty} \bar{f}_i = -\infty$, since then clearly $v(1) = -\infty$: $\{\bar{x}_i\}$ is a minimizing sequence, and there is nothing else to prove.

Theorem 5 *If (39) is solved by (50), (41) holds,* $\bar{f}_\infty > -\infty$ *and*

$$||z_i|| \leq L < \infty \qquad \forall i \ , \tag{54}$$

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);*
- *ii) if* f *is* **-compact and (36) 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 finite value. This ensures that $\bar{\delta}_i \to 0$. Indeed, $||z_i^*||$ is bounded above and (54) 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 (53) (given that $t_i \leq \bar{t}$). Therefore, if $\bar{\delta}_i \geq \varepsilon > 0$ for infinitely many indices, then (53) 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 1, which requires t bounded away from zero, and part ii) via Lemma 4 which requires f to be *-compact and (36). ■

A few comments on Theorem 5 are appropriate. Of course, (36) does not come for free: some form of appropriate management of t is required. Yet, this cannot be too difficult: t bounded away from zero surely suffice, and more sophisticated strategies are discussed in §5.1. In general, the result is somewhat weaker than those available for standard bundle methods, in two main aspects:

- The global boundedness condition (54) is usually not required. Something related is actually needed to show that the denominator in the rightmost term of (53) does not grow infinitely large, but this is only required for sequences of consecutive NS. Since in that case $\bar{x}$ is not changing, it is easy to obtain (54) as a natural consequence of the algorithm's workings: $\{z_i^*\}$ and $\{\alpha_i^*\}$ are bounded since the objective function of (9) is nonincreasing, hence $\{d_i^*\}$ is bounded (t_i is bounded above), 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. Yet, there are plenty of cases where (54) holds, such as f globally Lipschitz, f inf-compact (all $\bar{x}_i$ belong to some level set since $\bar{\zeta}_i \geq f_i$ and $\bar{\zeta}_i$ is nonincreasing), and even more general cases where the f is globally Liptchiz when restricted to any sublevel set, even if not compact (think e^x).
- Requiring f to be *-compact is also usually not necessary: [7, Proposition 1.2], for instance, guarantees convergence without any assumption on f provided that (in our notation) $||z_i^*|| \to 0$, $\alpha_i^* \to 0$, and
$$\textstyle\sum_{i=1}^{\infty} \lambda_i^* t_i = \infty \quad . \tag{55}$$
While we do have the first two conditions, proving (55) 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, a-la (37). One would guess that either (55) holds, or "$\bar{x}_i$ is moving little and we are converging somewhere" (so that Lemma 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 (53), which is not there in the standard convergence proof since it only appears 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. Apparently, 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 (53) can be improved as to make some of these weakness disappear. Furthermore, (36) lends itself well to algorithmic treatment, as discussed in the next section.

5.1 Management of t

The above development requires (41), which is unrealistic since on-line 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 (36) has to hold. Going towards practical implementations, one should therefore decide when is that t should be decreased, and how should exactly t_+ be chosen in order for (36) to be satisfied. Standard "proximity control" techniques developed for the proximal bundle method [21] 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 (53) (which clearly implies (40)), 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* one for which
$$\Delta\bar{\zeta} > \kappa_1 \bar{\delta} \quad ; \tag{56}$$

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

Condition (56) gives global convergence as (53) does (and is gives the “nice part” where convergence is linear). Hence, for $t_+ \geq t$ one has

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

and therefore any value

$$\frac{2\big(\,(1-\kappa_1)\bar{\delta} - \lambda^* \Delta \bar{f} - \alpha(\lambda^*)\,\big)}{\|z(\lambda^*)\|^2} \geq t_+ > t \tag{57}$$

keeps (56) satisfied, and thus we may want to pick the largest possible value in (57). Note, that (except in the “poorman” case), $\bar{\delta}_+ \leq t_+\|z(\lambda^*)\|^2 + \alpha(\lambda^*)$, but this is not an issue for (57), as a decreasing $\bar{\delta}_+$ (and hence $\bar{\zeta}_+$) can only help to attain (56). One can expect (56) 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 to happen 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 (53). 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 amply discussed in §3; thus, we may want to only ensure the “minimal target” $\Delta\bar{\zeta} \geq \kappa_2 \bar{\delta}$. Reasoning as for (57) we have that any

$$t_+ \leq \frac{2\big(\,(1-\kappa_2)\bar{\delta} - \lambda^* \Delta \bar{f} - \alpha(\lambda^*)\,\big)}{\|z(\lambda^*)\|^2} < t \tag{58}$$

ensures that the “minimal target” is reached, and therefore we may want to select the largest possible value in (58) (the minimal possible change of t). The combination of this and (57) 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 hold, i.e., either $t \geq \underline{t} > 0$ or (36), and likely to impose the reasonable damping $t_+ \geq \underline{p}t$ for some fixed $\underline{p} \in (0, 1]$. To ensure the weakest hypothesis (36) 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

$$t_+ \geq \tau\big(\, t\|z(\lambda^*)\|^2 + \alpha(\lambda^*) \,\big) \geq \tau\big(\, \bar{\delta}_+ \,\big) \ . \tag{59}$$

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 (56) provides a sensible “aspiration criterion” for managing t during the process, and in particular increasing it against the natural requirement (41), it is appropriate (if only for the links with the following §5.2) to remark that it is not the only one. In particular, our initial motivation for the development of the approach (cf. §2) was to attain the stopping condition (13) 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 α^* (actually, $\bar{\alpha}^*$) in the f-value space. Indeed, while both these need be “small”, t^* dictates the “relative importance” of the two, which is clearly somewhat related to the scaling of the function. Indeed, assume for the sake of 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 (13) can be seen as a measure of how much we can decrease at most traveling along the approximate subgradient $-z^*$ by a step t^*. Indeed, using $\|x^* - \bar{x}\| \leq D$ the (approximate) subgradient inequality then gives

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

When (13) holds, one has $\bar{\alpha}^* \leq \varepsilon$ and

$$\|z^*\| \leq \sqrt{(\varepsilon - \bar{\alpha}^*)/t^*}$$

and therefore one can seek for the value t^* such that

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

and therefore guarantees that $\bar{x}$ is actually ε-optimal at termination. Simple algebra shows that a sufficient condition is

$$t^* \geq D^2/\varepsilon \ . \tag{60}$$

One could therefore run the algorithm with fixed $t = t^*$; this, conceptually, would have the advantage that the merit function $\bar{\zeta}$ exactly weights the norm of z^* as the stopping condition (13), thereby hopefully reaching the goal of having the latter satisfied as quickly as possible. This is, however, moot in more ways than one. First, (60)—even assuming it can be precisely estimated—is likely to be a rather large value: fixing t there would result in an almost un-stabilized approach, which is likely to be rather inefficient in practice. Furthermore, the initial aim of reaching (13) as quickly as possible has had to be changed anyway, since one needs to take into account the decrease of f as well. Indeed, $\bar{\zeta}$ can be seen as being made of two terms, $\bar{f}(\bar{x}) + \bar{\alpha}^*$ in the f-value space and $\|z^*\|$ in the dual space, with t^* weighting the two. That is, t^* can be seen as a measure 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, one one needs "proving" it. This means that t should arguably be allowed to be (much) smaller than t^* during the course of the algorithm, although it could be beneficial to have it grow and reach near t^* as the computation nears its end. In other words, $t \approx t^*$ is another sensible aspiration criterion for management of t, but not necessarily uniformly along all the iterations. Different heuristics to this effect (that do not counter (40)/(53)) can be developed, but this is not the appropriate venue to discuss them.

Finally, let us remark that, since the theory refers to the asymptotic behavior, "everything than only happens finitely many times is allowed". That is, like e.g. in [11, (4.ii)], t_+ can be set to whatever value one desires provided that some mechanism ensures this stops happening "at some point" and (57)/(59) are satisfied "from that point onwards", so that the convergence theory applies to the iterates "after that tipping point". Yet, in standard bundle methods this mostly applies to sequences of NS: as soon as a SS is performed, everything can be basically reset, comprised the "counters for allowing irregular behavior". This is a somehow convenient consequence of the otherwise irritating characterization of bundle methods as composed by two almost independent processes (cf. §1): for the "inner" process, i.e., sequences of consecutive NS, everything restarts anew (in terms of asymptotic convergence) with any new SS (if any). Such convenience is not shared by our approach, that is analyzed in terms of a unique converging process. Yet, as shown in the next section, this allows to more easily derive efficiency estimates for the method than what is possible for the standard approach [22].

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 (even non consecutive) iterates. The former is clearly smaller than the $L < \infty$ of (54) (e.g., the global Lipschitz constant of f); for the latter one can take the $D < \infty$ of (60) (e.g., the diameter of *dom* f, or of an appropriate level set if f is inf-compact).

We need first detail the relationships between the stopping condition (13) (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 (13) whenever (60) 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, (13) surely holds if

$$2\bar{\delta}_i\tau_i \leq \varepsilon \quad \equiv \quad 2\bar{\delta}_i/t_i \leq \varepsilon/t^* \quad \equiv \quad 2\bar{\delta}_i \leq \varepsilon(t_i/t^*) \ ; \tag{61}$$

note that, due to (36), this has to happen eventually.

In order to estimate how many iterations at most this will take, we use the fact that

$$\bar{f}_1 - f^* \leq LD \ .$$

Furthermore, since $||z_1^*||^2 \leq L^2$ and $\bar{\alpha}_1^* = \alpha_1 = 0$ we can assume

$$\varepsilon(t_1/t^*) < 2\bar{\delta}_1 \leq t_1 L^2 \tag{62}$$

(for otherwise (61) would hold for $i = 1$) and therefore

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

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

$$\textstyle\sum_{i=1}^{k-1} \Delta\bar{\zeta}^i = \bar{\zeta}_1 - \bar{\zeta}_k \geq LD + \bar{\delta}_1 - \varepsilon(t_i/t^*)/2 \ ; \tag{63}$$

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 (61)—that (13) holds, and hence that $\bar{f}_i - f^* \leq \varepsilon$ due to (60). Note that in this way we not only account for the time it takes to *reach* a ε-optimal point, but also for the time it takes to *certify* its ε-optimality having constructed appropriate z^* and $\bar{\alpha}^*$. Thus, this is a very fair evaluation of the complexity of the algorithm.

To simplify the notation somewhat we now fix $m = 3/5$, so that $2m - 1 = (1 - m)/2$ and the first two terms in the min of (53) are equal; furthermore, we use $\|z - z^*\|^2 \leq 2L^2$ to get

$$\Delta\bar{\zeta} \geq \frac{\bar{\delta}}{5} \min\left\{ \ 1 \ , \ \frac{\bar{\delta}}{5tL^2} \ \right\} \ . \tag{64}$$

The main issue at this point is that (64) does *not* say is how the decrease in $\bar{\zeta}$ is "subdivided between its two components". That is, (64) may imply either that $\bar{f}$ decreases (at least) by the given amount, or $\bar{\delta}$ does, or that both decrease, each by a fraction of the total amount. 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 (64) 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 (64), 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 $\bar{f}$, up until no further decrease is possible least (61) 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 therefore 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 sub-phases: that of "large $\bar{\delta}$" and that of "small $\bar{\delta}$". That is, until

$$\frac{\bar{\delta}_i}{5t_i L^2} \geq 1 \tag{65}$$

the rate of 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

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

Unfortunately, (65) does not last long: combining it with (62) gives

$$\frac{t_1 L^2}{10 t_i L^2} \geq 1 \;\; .$$

i.e., the "large $\bar{\delta}$" sub-phase can well terminate immediately unless t_i is reduced very quickly. Although this provides an interesting rationale for some of the discussion in §5.1, it is compatible with the algorithm that this sub-phase 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 sub-phase begins where instead

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

and therefore the rate of decrease is sublinear:

$$\bar{\zeta}_+ \leq \bar{\zeta}_i - \frac{\bar{\delta}_i^2}{25 t_i L^2} \;\; . \tag{66}$$

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

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

which, using again $\bar{f}_+ = \bar{f}_i$ ($= \bar{f}_1$) and (60) gives

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

and therefore

$$\bar{\delta}_i \leq \bar{\delta}_1 \left(1 - \frac{\gamma^2}{50}\right)^{i-1} \leq \frac{t_1 L^2}{2}\left(1 - \frac{\gamma^2}{50}\right)^{i-1} \;\; .$$

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

$$\frac{t_1 L^2}{2}\left(1 - \frac{\gamma^2}{50}\right)^{i-1} \leq \varepsilon \frac{t_i}{2t^*}$$

which, using $t_i \leq t_1$ and (60), gives

$$(1 - \gamma^2/50)^{i-1} \leq \gamma^2 \quad \Longrightarrow \quad i \geq \xi(\gamma) = \frac{\log(\gamma^2)}{\log(1 - \gamma^2/50)} + 1 \;\; .$$

We have therefore to estimate how quickly $\xi(\gamma)$ goes to infinity (as it is easy to verify it does) when its argument $\gamma = \varepsilon/DL$ goes to zero. Some easy but tedious calculus shows that

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

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

We are now left to estimating 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

$$\bar{\zeta}_h - \bar{\zeta}_i \leq LD \;\; .$$

We start again from (66), but now we use $\bar{\delta}_i^2 \geq (\varepsilon(t_i/t^*)/2)^2$ and (60) to get

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

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

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

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 it indeed diverges in the first place). The simple assumption $\tau_i^* \geq \bar{\tau} > 0$ (which implies t_i bounded away from zero) gives

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

Therefore, 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 [25], but comparable to (although somewhat different than) the $O(D^4L^2/\varepsilon^3)$ efficiency of proximal bundle methods [22]. It is difficult to say whether this difference in the efficiency estimate really reflects a different asymptotic behavior of the proposed algorithm w.r.t. the proximal bundle method rather than being a figment of the different complexity analysis techniques. To gain some insight upon 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 (16); 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 (39), 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 (16) itself).

A first possibility could therefore be to rely on specific properties of f. For instance, in many applications, such as in classification problems [4, 2, 3], f is a polyhedral max-function corresponding to an "easy" optimization problem, that is, either a convex (linear) program [8, 12, 14, 16] or a nonconvex one that can still be solved efficiently [13]. 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. §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

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

Of course, in the worst case $\bar{\lambda} = 1$ and this gives back (16). This approach is easily extended to composite functions where an appropriate mapping is superimposed over the max-function [26].

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 [12, 21]: 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 (16), 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 (16) by defining the sub-interval of $[0, 1]$ in which (16) ensures a "sufficient decrease", e.g. in the sense of (56). 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 start by stressing 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 the existing ones (and in particular the proximal bundle method from which the approach derives). For this purpose we have implemented the proposed approach (which we refer to as "NMBundle") in `C++` and compared it 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 [12, 8, 16, 15]. The structure of the existing code allows the new approach to inherit several useful features. For instance, once an "oracle" for a function f is implemented it can be used by both approaches without modifications, and the new approach can exploit the efficient solver for the master problem described in [10]. All the algorithms have been compiled with GNU `g++ 4.0.1` (with -O3 optimization option) and ran 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 "classical academic test functions" with small n (up to 48) that have been used 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 of the test set. The results are reported in Table 1. For each approach, 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 (57) and (58) 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 (13), 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).

Table 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

The Table shows that the two approaches are roughly comparable; NMBundle requires less iterations in 4 cases over 12, in two of them (Maxq and Maxquad) somewhat significantly, while it is significantly slower in the largest TR48. 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. [5, 8, 12, 15, 17, 24] for a partial list of reasonably recent publications. In particular here we employ the simple *weak flow relaxation* of the Fixed-Charge Multicommodity Min-Cost Flow (FC-MMCF) problem. We avoid delving into the details of the original problem and of the corresponding Lagrangian relaxation, referring the interested reader to [8] and especially to the recent [15] for the details (comprised a description of the freely available test instances); here we only mention that, contrary to the "academic" test cases of Table 1, there are larger-scale problems whose dimensionality n is given by the product of the number of arcs (m) and the number of commodities (k) in the underlying graph. Hence, the largest instances reported in Table 2 have $n = 960000$, although this is partly mitigated by the fact that the

Lagrangian multipliers are constrained to be non-negative ($x \geq 0$) which somewhat decreases the "true" dimensionality of the problem, since a fraction of the variables can be expected to be zero at optimality and along all the optimization process. Again in this case we have performed accurate tuning for both approaches (for PBundle we actually piggy-backed on the tuning performed in [15]), which in particular led to selecting $\kappa_1 = 0.6$ and $\kappa_2 = 0.001$ in (57) and (58), respectively (now $t = 1$ at start). The stopping criterion has been set as in the previous case, and the meaning of the columns in Table 2 is the same as in Table 1.

Table 2: Results 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

Also Table 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 §5.1 are all "short-term"; that is, they only "look" at what is happening in the current iteration. For PBundle, "long-term" strategies have been developed (see e.g. [8] 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 using, instead of the function value, an (apparently) novel merit function, derived from the Moreau-Yoshida regularization, to drive the search towards optimality. This approach would in general lead to a non-dichotomic 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. (51)), 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 "full" 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 [11] or developing "long-term" strategies for t management that result into 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 Journal on Optimization*, 21(4):1418–1438, 2011.

[2] A. Astorino, A. Fuduli, and M. Gaudioso. DC Models for Spherical Separation. *Journal of Global Optimization*, 48(4):657–669, 2010.

[3] A. Astorino, A. Fuduli, and M. Gaudioso. Margin Maximization in Spherical Separation. *Computational Optimization and Applications*, to appear, 2012.

[4] A. Astorino, A. Fuduli, and E. Gorgone. Non-smoothness in Classification Problems. *Optimization Methods and Software*, 23(5):675–688, 2008.

[5] F. Babonneau and J.-P. Vial. ACCPM with a Nonlinear Constraint and an Active Set Strategy to Solve Nonlinear Multicommodity Flow Problems. *Mathematical Programming*, 120(1):179–210, 2009.

[6] O. Briant, C. Lemaréchal, P. Meurdesoif, S. Michel, N. Perrot, and F. Vanderbeck. Comparison of Bundle and Classical Column Generation. *Mathematical Programming*, 113(2):299–344, 2008.

[7] R. Correa and C. Lemaréchal. Convergence of Some Algorithms for Convex Minimization. *Mathematical Programming*, 62(2):261–275, 1993.

[8] T.G. Crainic, A. Frangioni, and B. Gendron. Bundle-based Relaxation Methods for Multicommodity Capacitated Fixed Charge Network Design Problems. *Discrete Applied Mathematics*, 112:73–99, 2001.

[9] F. Facchinei and S. Lucidi. Nonmonotone Bundle-Type Scheme for Convex Nonsmooth Minimization. *Journal of Optimization Theory and Applications*, 76(2):241–257, 1993.

[10] A. Frangioni. Solving Semidefinite Quadratic Problems Within Nonsmooth Optimization Algorithms. *Computers & Operations Research*, 21:1099–1118, 1996.

[11] A. Frangioni. Generalized Bundle Methods. *SIAM Journal on Optimization*, 13(1):117–156, 2002.

[12] A. Frangioni and G. Gallo. A Bundle Type Dual-Ascent Approach to Linear Multicommodity Min Cost Flow Problems. *INFORMS Journal on Computing*, 11(4):370–393, 1999.

[13] A. Frangioni and B. Gendron. 0-1 Reformulations of the Multicommodity Capacitated Network Design Problem. *Discrete Applied Mathematics*, 157(6):1229–1241, 2009.

[14] A. Frangioni and C. Gentile. Solving Nonlinear Single-Unit Commitment Problems with Ramping Constraints. *Operations Research*, 54(4):767–775, 2006.

[15] A. Frangioni and E. Gorgone. Generalized Bundle Methods for Sum-Functions with "Easy" Components: Applications to Multicommodity Network Design. Technical Report 3049, Optimization Online, http://www.optimization-online.org, 2011.

[16] A. Frangioni, A. Lodi, and G. Rinaldi. New Approaches for Optimizing Over the Semimetric Polytope. *Mathematical Programming*, 104(2-3):375–388, 2005.

[17] K. Holmberg and D. Yuan. A Lagrangean Heuristic Based Branch-and-Bound Approach for the Capacitated Network Design Problem. *Operations Research*, 48:461–481, 2000.

[18] L. Hou and W. Sun. On the Global Convergence of a Nonmonotone Proximal Bundle Method for Convex Nonsmooth Minimization. *Optimization Methods & Software*, 23:227–235, 2008.

[19] E. Karas, A. Ribeiro, C. Sagastizábal, and M. Solodov. A Bundle-filter Method for Nonsmooth Convex Constrained Optimization. *Mathematical Programming*, 116:297–320, 2009.

[20] J.E. Kelley. The Cutting Plane Method for Solving Convex Programs. *Journal of the SIAM*, 8:703–712, 1960.

[21] K. Kiwiel. Proximity Control in Bundle Methods for Convex Nondifferentiable Minimization. *Mathematical Programming*, 46:105–122, 1990.

[22] K. Kiwiel. Efficiency of Proximal Bundle Methods. *Journal of Optimization Theory and Applications*, 104(3):589–603, 2000.

[23] K. Kiwiel and C. Lemaréchal. An Inexact Bundle Variant Suited to Column Generation. *Mathematical Programming*, 118:177–206, 2009.

[24] K.C. Kiwiel. An Alternating Linearization Bundle Method for Convex Optimization and Nonlinear Multicommodity Flow Problems. *Mathematical Programming*, 130(1):59–84, 2011.

[25] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov. New Variants of Bundle Methods. *Mathematical Programming*, 69:111–147, 1995.

[26] C. Sagastizábal. Composite Proximal Bundle Method. *Mathematical Programming*, to appear, 2013.