Università di Pisa

Dipartimento di Informatica

Technical Report : TR-98-04

Generalized Bundle Methods

Antonio Frangioni

March 17, 1998

ADDR: Corso Italia 40,56125 Pisa,Italy. TEL: +39-50-887111. FAX: +39-50-887226

Generalized Bundle Methods

Antonio Frangioni[‡]

Abstract: We propose a class of generalized Bundle methods where the stabilizing term can be any closed convex function satisfying certain properties. We prove finite termination, asymptotic convergence and finite convergence to an optimal point of different variants of the algorithm: for some of these proofs, f-regularity of the function (a new generalization of inf-compactness) is required. Our analysis gives a unified convergence proof for several classes of Bundle methods that have been so far regarded as distinct, enhancing on the results known for some of them: furthermore, it covers methods proposed in the literature that had not previously been recognised as Bundle methods. A novelty in our approach is the proposal of a dual for the minimization problem: we show that Bundle methods can be seen as a dual ascent approach to one single nonlinear problem in the dual space, where nonlinear subproblems are approximately solved at each step with an inner linearization approach. This interpretation is particularly interesting in the case of Lagrangean Duals of structured problems. Differently from previous descriptions, we are able to describe and justify the changes in the subproblems during the Serious Steps, since the subproblems are no more (necessarily) tied to the local concept of ε-subdifferential.

Keywords: NonDifferentiable Optimization, Bundle methods, Lagrangean Duals

0. Introduction

During the last 20 years, several variants of Bundle methods for NonDifferentiable Optimization (NDO) have been proposed. The firsts have been ε-descent Bundle methods [HL93b - Chap. XIII], conceived to find an ε-ascent direction by properly approximating the ε-subdifferential in the current point and performing a line search along it. Other methods are rather conceived as "stabilizations" of the Cutting Plane approach: they essentially differ in the stabilization technique, that can be either a trust region [HL93b - §XV.2.1] or a penalty term in the objective function, which is usually based on the L_2 norm [HL93b - §XV.2.2] but may also be based on different functions [KCL95]. Different modifications of the basic Cutting Plane idea are the Proximal Level method [HL93b - §XV.2.3] [LNN95] and Analytic Centre-based algorithms [AV95] [Ne95] [GV97].

[‡] Department of Computer Science, University of Pisa • Corso Italia 40, 56125 Pisa - Italy
e-mail: frangio@di.unipi.it • Ph. +39 50 887289 • Fax. +39 50 887226

In a previous work [Fr97], we have shown how the "classical" Bundle methods with a stabilizing term based on the L_2 norm and certain Bundle methods with a trust region in the L_1 and L_∞ norms are much similar if seen from a "dual" viewpoint: as a result, their convergence could be proved almost independently from the norm employed. Here, we extend this idea to a more general setting: we prove convergence of a family of Bundle methods with a generic stabilizing term which only satisfies certain basic properties. This setting comprises many (even if not all) of the Bundle algorithms proposed so far, as well as other methods that have not previously been recognised as Bundle methods [PZ94]; furthermore, it is related with recent ε-approximation algorithms for block-structured problems [GK95].

For some convergence proofs, we require the function to be *f-regular*, apparently a new type of assumption properly generalizing inf-compactness. This seems to be necessary for proving convergence in a unified way: although f-regularity is not needed for some algorithms in this class, in fact, their convergence proofs do not seem to generalize. F-regularity seems to be a reasonable assumption: many functions are f-regular even if they are not inf-compact, such as all the polyhedral ones. Furthermore, the assumption can be dropped in certain variants of the method, e.g. when Serious Steps are properly inhibited.

A novelty in our approach is the identification of what we think should be considered "the dual" of the minimization problem: we show that Bundle methods can be seen as a dual ascent approach to one single nonlinear problem in the dual space, where nonlinear subproblems are approximately solved at each step with an inner linearization approach. Bundle methods with inhibited Serious Steps are rather penalty function approaches to the same problem. This formally establish the substantial identity between (generalized) Bundle methods applied to the Lagrangean Dual of an optimization problem and (generalized) Augmented Lagrangean approaches [PZ94]. Furthermore, we are able to describe and justify the changes in the subproblems during the Serious Steps, since the subproblems are no more tied to the "local" concept of ε-subdifferential.

The structure of this paper is the following: an overview of Bundle methods is given in §1, where the main concepts and notations are introduced. §2 is devoted to the derivation and justification of the dual of the minimization problem: the main properties of pairs of primal and dual solutions of the restricted subproblems are also described. Convergence proofs are given in §3, with an extensive discussion of the required properties for the stabilizing term, reset rules and aggregation. F-regular functions are studied in §4, while §5 is dedicated to extensions of the Bundle approach and to comparison with previously known results from the literature. Finally, conclusions are drawn in §6.

Throughout the paper, the following notation is used: symbols are written in *italic*, vectors are intended as columns and are written in **bold**, the scalar product between $\boldsymbol{v}$ and $\boldsymbol{w}$ is denoted by $\boldsymbol{v}\boldsymbol{w}$. $\| \boldsymbol{v} \|_1$, $\| \boldsymbol{v} \|_2$ and $\| \boldsymbol{v} \|_\infty$ stand respectively for $\Sigma_i | v_i |$, $(\boldsymbol{v}\boldsymbol{v})^{1/2}$ and $\max_i \{ | v_i | \}$, i.e. the L_1, L_2 and L_∞ norms of the vector $\boldsymbol{v}$. The ball around $\mathbf{0}$ of radius δ in the L_p norm will be denoted with $B_p(\delta)$. Given a set C, we denote by

$$I_C(\boldsymbol{y}) = \begin{cases} 0 & \text{if } \boldsymbol{y} \in C \\ +\infty & \text{otherwise} \end{cases}$$

its (convex) *indicator function*. Given the problem

(P) $\quad \inf\ [\sup]\ \{\ f(\boldsymbol{x}) : \boldsymbol{x} \in X\ \}$

where f is an extended-valued function, $z(\mathrm{P})$ denotes the optimal value of f over X: with the usual assumption that the inf [sup] over an empty set is $+\infty$ $[-\infty]$, $X = \emptyset \Rightarrow z(\mathrm{P}) = +\infty\ [-\infty]$.

1. Overview of Bundle methods

We are concerned about the numerical solution of the problem

(Π) $\quad \inf_y \{\ \varphi(\boldsymbol{y}) : \boldsymbol{y} \in Y\ \}$

where φ is a (nondifferentiable) *closed* (i.e. lower semicontinuos) *convex* function that is *finite everywhere*, and Y is a convex nonempty subset of $\Re^n$: for sake of simplicity, we will initially assume $Y = \Re^n$. Knowledge of φ is only available under the form of a "black box" that, given $\bar{\boldsymbol{y}} \in Y$, returns the value $\varphi(\bar{\boldsymbol{y}})$ and one *subgradient* $\boldsymbol{g}$ of φ in $\bar{\boldsymbol{y}}$, i.e. a vector in $\Re^n$ such that

$$\varphi(\boldsymbol{y}) \geq \varphi(\bar{\boldsymbol{y}}) + \boldsymbol{g}(\boldsymbol{y} - \bar{\boldsymbol{y}}) \qquad \forall\ \boldsymbol{y}.$$

Subgradients can be visualised as gradients of the supporting hyperplanes of the *epigraph* of φ

$$Epi\ \varphi = \{\ (\alpha, \boldsymbol{y}) : \varphi(\boldsymbol{y}) \leq \alpha\ \} \subseteq \Re^{n+1}$$

in $(\varphi(\bar{\boldsymbol{y}}), \bar{\boldsymbol{y}})$; alternatively, they may be regarded to as the directions $\boldsymbol{g}$ such that the *level set*

$$S_\delta(\varphi) = \{\ \boldsymbol{y} : \varphi(\boldsymbol{y}) \leq \delta\ \} \qquad \text{for } \delta = \varphi(\bar{\boldsymbol{y}})$$

(and therefore the set of the optimal solutions of (Π)), lies in the half-space pointed by $-\boldsymbol{g}$ w.r.t. $\bar{\boldsymbol{y}}$. Only to simplify some notation, we also assume w.l.o.g. that $\varphi(\boldsymbol{0}) = 0$: in practice, there is no need for explicitly scaling the function.

Generally speaking, Bundle methods are characterised by sampling the function in a sequence of points $\{\ \boldsymbol{y}_i\ \}$ to gather information in the form $\{\ (\varphi(\boldsymbol{y}_i), \boldsymbol{g}_i)\ \}$, where $\boldsymbol{g}_i$ is the subgradient returned by the black box when called at $\boldsymbol{y}_i$. Also, they usually keep a distinguished vector $\bar{\boldsymbol{y}}$ (usually, the best point found so far) as the *current point*: at the generic iteration k, the new information acquired about φ is understood to be "near" $\bar{\boldsymbol{y}}_k$. In fact, (a subset of) the information obtained so far (the *Bundle* β) is used in order to compute a *tentative descent direction* $\boldsymbol{d}_k$ along which the next points will be generated. In general, $\boldsymbol{d}_k$ can not even be guaranteed to be of descent: hence, Bundle methods can decide whether to update the current point (a so-called *Serious Step*), or rather leave it unchanged and use the new information to enlarge β and obtain a - hopefully - better $\boldsymbol{d}_{k+1}$ (a *Null Step*). A generic Bundle algorithm is described in Figure 1.

Most Bundle methods can be regarded to as "stabilized" versions of the *Cutting Plane algorithm* [Ke60] [HL93b - Algorithm XII.4.2.1], that is based on the following idea: the *Cutting Plane model*

$$\varphi_\beta(\boldsymbol{y}) = \max\{\ \varphi(\boldsymbol{y}_i) + \boldsymbol{g}_i(\boldsymbol{y} - \boldsymbol{y}_i) : (\varphi(\boldsymbol{y}_i), \boldsymbol{g}_i) \in \beta\ \}$$

is a polyhedral lower approximation of φ which is "tight" at least in every point $\boldsymbol{y}_i$. A straightforward idea is then to minimize φ_β instead of φ, i.e. to solve

(Π_β) $\quad \inf_y \{\ \varphi_\beta(\boldsymbol{y})\ \}$

which is a Linear Programming subproblem with $|\beta|$ constraints. Any optimal solution $\hat{\boldsymbol{y}}$ of (Π_β) can be used as the next trial point: if $\varphi_\beta(\hat{\boldsymbol{y}}) = \varphi(\hat{\boldsymbol{y}})$ then $\hat{\boldsymbol{y}}$ is optimal for (Π), otherwise a new cutting plane $(\varphi(\hat{\boldsymbol{y}}), \hat{\boldsymbol{g}})$ has been generated that can be added to β to "improve" the CP model.

Generic Bundle Algorithm

⟨ choose the initial $\bar{\boldsymbol{y}}$ and β ⟩;
do
⟨ find a tentative direction $\boldsymbol{d}$ ⟩;
⟨ move along $\boldsymbol{d}$, generating some new subgradients and a trial point $\boldsymbol{y} = \bar{\boldsymbol{y}} + \tau\boldsymbol{d}$ ⟩;
if(a large enough improvement has been obtained)
then $\bar{\boldsymbol{y}} = \boldsymbol{y}$; /* SS */
/* **else** it is a NS */
⟨ update β, adding new subgradients and deleting old ones ⟩;
⟨ update some algorithmic parameter, depending on the previous iterations ⟩;
while(⟨ some stopping condition ⟩);

Figure 1: a generic Bundle algorithm

Unfortunately, this simple approach is not satisfactory for several reasons:

- at least in the early stage of the algorithm, (Π_β) is likely *not* to have a finite solution;
- in practice, this must be faced by adding "artificial" (e.g. box) constraints in order to obtain a bounded feasible region that still contains some optimal point: since finding "tight" such constraints is difficult, $\hat{\boldsymbol{y}}$ often turns out to be unreasonably far away from $\bar{\boldsymbol{y}}$, in a region where the CP model is a poor representation of the real function $\varphi()$;
- in theory, the trajectory followed by the trial point is "unstable", in the sense that the distance between $\bar{\boldsymbol{y}}$ and some optimal point for (Π) has no monotonicity properties: in fact, the current point play no role in the CP algorithm.

The idea behind most Bundle methods is that of adding some "stabilizing device" to the CP algorithm: in many (even if not all) cases, this takes the form of a penalty term D_t that discourages points "far away" from $\bar{\boldsymbol{y}}$, yielding the subproblem

$$(\Pi_{\beta(\bar{\boldsymbol{y}}),t}) \qquad \inf_{\boldsymbol{d}} \{ \varphi_{\beta(\bar{\boldsymbol{y}})}(\boldsymbol{d}) + D_t(\boldsymbol{d}) \}$$

where $\boldsymbol{d}$ has to be intended as the displacement from $\bar{\boldsymbol{y}}$ and t is a nonnegative parameter dictating the "strength" of D_t. Here, $\varphi_{\beta(\bar{\boldsymbol{y}})}$ is the *translated* CP model w.r.t. $\bar{\boldsymbol{y}}$: we can equivalently consider β as formed by pairs $(\alpha_i, \boldsymbol{g}_i)$, where (due to the assumption $\varphi(\boldsymbol{0}) = 0$) $\alpha_i = \boldsymbol{g}_i\boldsymbol{y}_i - \varphi(\boldsymbol{y}_i) \geq 0$ is the value of the i-th linear function in $\boldsymbol{0}$ (the "default current point"). In general

$$\varphi_{\beta(\bar{\boldsymbol{y}})}(\boldsymbol{d}) = \max_{\sigma,z}\{ \boldsymbol{d}z - \sigma : (\sigma, z) \in \beta(\bar{\boldsymbol{y}}) \} = \varphi_\beta(\bar{\boldsymbol{y}} + \boldsymbol{d}) - \varphi(\bar{\boldsymbol{y}})$$

where $(\sigma, z) \in \beta = \beta(\boldsymbol{0}) \Leftrightarrow (\sigma - z\bar{\boldsymbol{y}} + \varphi(\bar{\boldsymbol{y}}),\ z) \in \beta(\bar{\boldsymbol{y}})$: $\sigma - z\bar{\boldsymbol{y}} + \varphi(\bar{\boldsymbol{y}}) \geq 0$ is called the *linearization error* of z w.r.t. $\bar{\boldsymbol{y}}$, and it is easy to see that it is exactly the difference between $\varphi(\bar{\boldsymbol{y}})$ and the value of the corresponding linear function in $\bar{\boldsymbol{y}}$.

The first stabilizing term ever proposed is $D_t = {}^1/_{2t}\|\cdot\|_2^2$ [HL93b - Algorithm XV.3.3.4]: this establishes a connection with the *Proximal Point* algorithm [Ro76], which is nowadays interpreted in terms of the *Moreau-Yosida Regularization* of φ

$$(\Pi^2_{y,t}) \qquad \phi_t(\boldsymbol{y}) = \inf_{\boldsymbol{d}} \{ \varphi(\boldsymbol{y}+\boldsymbol{d}) + {}^1/_{2t}\|\boldsymbol{d}\|^2_{\boldsymbol{M}} \}$$

($\|\boldsymbol{d}\|^2_{\boldsymbol{M}} = \boldsymbol{d}^T\boldsymbol{M}\boldsymbol{d}$ for some positive definite matrix $\boldsymbol{M}$). ϕ_t is the "ideal" version of φ, since it has the same set of minima but it is differentiable: hence, a conceptually attractive way of maximizing φ would be that of applying any (efficient) smooth optimization algorithm to ϕ_t. Unfortunately, computing $\phi_t(\bar{\boldsymbol{y}})$ requires the solution of $(\Pi^2_{\bar{y},t})$ that is, in terms of the original "black box", no less difficult than (Π): however, $(\Pi^2_{\bar{y},t})$ may be solved only *approximately*, and the resulting information be used to move the current point towards the optimum. In fact, the Bundle algorithm can be regarded to as a "two-level" approach to the minimization of ϕ_t, where the approximated version of $(\Pi^2_{y,t})$

$$(\Pi^2_{\beta(\bar{y}),t}) \qquad \min_{d,v} \{ v + {}^1/_{2t}\|\boldsymbol{d}\|^2 : v \geq \boldsymbol{d}\boldsymbol{z} - \sigma \,,\, (\sigma, z) \in \beta(\bar{\boldsymbol{y}}) \}$$

is solved at each iteration: sequences of consecutive NS aim at approximately computing $\phi_t(\bar{\boldsymbol{y}})$, while Serious Steps use the resulting information for the actual minimization.

This approach has the following nice *dual* interpretation: the *quadratic dual* of $(\Pi^2_{\beta(\bar{y}),t})$ is

$$(\Delta^2_{\beta(\bar{y}),t}) \qquad \min_{\theta} \{ {}^1/_2\, t\|\boldsymbol{G}_\beta\theta\|^2 + \alpha_{\beta(\bar{y})}\theta : \theta \in \Theta \}$$

where $\Theta = \{ \theta : \sum_{i \in \beta} \theta_i = 1 \,,\, \theta \geq \boldsymbol{0} \}$ is the unitary simplex of proper dimension and $\boldsymbol{G}_\beta$, $\alpha_{\beta(\bar{y})}$ are respectively the matrix of subgradients and the vector of the corresponding linearization errors w.r.t. $\bar{\boldsymbol{y}}$. $(\Delta^2_{\beta(\bar{y}),t})$ can be viewed as the Lagrangean relaxation of

$$(\Delta^{2,\varepsilon}_{\beta(\bar{y})}) \qquad \min_{\theta} \{ {}^1/_2\|\boldsymbol{G}_\beta\theta\|^2 : \alpha_{\beta(\bar{y})}\theta \leq \varepsilon \,,\, \theta \in \Theta \}$$

where the (fixed) Lagrangean multiplier $1/t$ is used to dualize the constraint $\alpha_{\beta(\bar{y})}\theta \leq \varepsilon$: the latter is the *steepest* ε*-descent direction finding problem* for the CP model $\varphi_{\beta,\bar{y}}$. In fact [HL93b - Chap. VI]

- an ε*-subgradient* of φ in $\bar{\boldsymbol{y}}$ is any vector such that

  $$\varphi(\boldsymbol{y}) \geq \varphi(\bar{\boldsymbol{y}}) + \boldsymbol{g}(\boldsymbol{y} - \bar{\boldsymbol{y}}) - \varepsilon \qquad \forall\ \boldsymbol{y}$$

  and the ε*-subdifferential* $\partial_\varepsilon\varphi(\bar{\boldsymbol{y}})$ the set of all the ε*-subgradients*;
- the *least-norm* ε*-subgradient* is the (opposite of the) *steepest* ε*-descent direction*, i.e.

  $$-\operatorname{argmin}_{\boldsymbol{g}} \{ \|\boldsymbol{g}\| : \boldsymbol{g} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}}) \} \equiv \operatorname{argmin}_{\boldsymbol{d}} \{ \varphi'_\varepsilon(\bar{\boldsymbol{y}}, \boldsymbol{d}) : \|\boldsymbol{d}\| = 1 \}$$

  where $\varphi'_\varepsilon(\bar{\boldsymbol{y}}, \boldsymbol{d})$ is the ε*-directional derivative* (of φ in $\bar{\boldsymbol{y}}$ along direction $\boldsymbol{d}$)

  $$\varphi'_\varepsilon(\bar{\boldsymbol{y}}, \boldsymbol{d}) = \inf_{t\,>\,0} \frac{\varphi(\bar{\boldsymbol{y}} + t\boldsymbol{d}) - \varphi(\bar{\boldsymbol{y}}) + \varepsilon}{t} = \inf_{v}\{ \boldsymbol{v}\boldsymbol{d} : \boldsymbol{v} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}}) \}$$

  and $\varphi'_\varepsilon(\bar{\boldsymbol{y}}, \boldsymbol{d}) < 0 \Leftrightarrow \boldsymbol{d}$ *is of* ε*-descent*, i.e. it allows a descent strictly larger than ε

  $$\exists\ t > 0 \text{ s.t. } \varphi(\bar{\boldsymbol{y}} + t\boldsymbol{d}) < \varphi(\bar{\boldsymbol{y}}) - \varepsilon$$

  (as a consequence, $\bar{\boldsymbol{y}}$ is ε-optimal for $\varphi \Leftrightarrow \boldsymbol{0} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}})$);
- the ε-subdifferential of the Cutting Plane model $\varphi_{\beta,\bar{y}}$ in $\boldsymbol{0}$ is [HL93b - Example XI.3.5.3]

$$\partial_\varepsilon \varphi_{\beta(\bar{y})}(\,\mathbf{0}\,) = \{\; \boldsymbol{g} = \boldsymbol{G}_\beta \theta : \alpha_{\beta(\bar{y})}\theta \le \varepsilon \;\;, \theta \in \Theta \;\}.$$

Hence, $(\Delta^{2,\varepsilon}_{\beta(\bar{y})})$ either finds the steepest ε-descent direction for $\varphi_{\beta(\bar{y})}$ in the current point, or returns $\mathbf{0}$ if $\bar{\boldsymbol{y}}$ is ε-optimal: alternatively, since $\partial_\varepsilon \varphi(\,\bar{\boldsymbol{y}}\,) \supseteq \partial_\varepsilon \varphi_{\beta(\bar{y})}(\,\mathbf{0}\,)$ one can say that the steepest ε-descent direction finding problem for φ is solved *approximately* by replacing the unknown $\partial_\varepsilon \varphi(\,\bar{\boldsymbol{y}}\,)$ with its *inner linearization* $\partial_\varepsilon \varphi_{\beta(\bar{y})}(\,\mathbf{0}\,)$. For each ε, there exist a t such that $(\Delta^{2}_{\beta(\bar{y}),t})$ gives the same solution: hence, Bundle methods can be seen from the dual side as searching for approximate steepest ε-descent directions (for varying ε). In this case, Null Steps aim at refining the approximation of $\partial_\varepsilon \varphi(\,\bar{\boldsymbol{y}}\,)$ when it is "so bad" that it does not identifies a descent (enough) direction: a SS is instead the result of a correctly approximated ε-subdifferential.
There are reasons to prefer $(\Delta^{2}_{\beta(\bar{y}),t})$ to $(\Delta^{2,\varepsilon}_{\beta(\bar{y})})$: the former is always nonempty, while the latter may not have feasible solutions - although it surely has if one $(0,\, z) \in \beta(\bar{\boldsymbol{y}})$, i.e. z is "tight" in $\bar{\boldsymbol{y}}$. Furthermore, rules for choosing t are easier to develop than rules for choosing ε, both in theory (compare [HL93b - Algorithm XIV.3.4.2] with [HL93b - Algorithm XV.3.1.4]) and in practice [Ki90] [SZ92] [HL93b - §XIV.3.3.(c)] [CFN96] [Fr97]. However, $(\Delta^{2,\varepsilon}_{\beta(\bar{y})})$ was considered easier to solve than $(\Delta^{2}_{\beta(\bar{y}),t})$ [Ki85 - p. 299]: in fact, ten years separate the first specialised algorithm for the former [Mi77] from the first specialised algorithm for the latter [Ki86]. This is probably not true anymore [Fr96].

The dual interpretation of Bundle methods motivated the first developments and it is very important for their analysis: for instance, being θ^* the optimal solution of $(\Delta^{2}_{\beta(\bar{y}),t})$, the optimal solution $(v^*, \boldsymbol{d}^*)$ of $(\Pi^{2}_{\beta(\bar{y}),t})$ is

$$(v^*, \boldsymbol{d}^*) = (t\| z^* \|^2 + \sigma^*, -tz^*) \qquad \text{where} \qquad z^* = \boldsymbol{G}_\beta \theta^* \text{ and } \sigma^* = \alpha_{\beta(\bar{y})}\theta^*$$

i.e. the new tentative point is found along the *direction* z^* - a convex combination of the currently available subgradients - using $(-)t$ as *predefinite stepsize*. Furthermore, $z^* \in \partial_{\sigma^*}\varphi(\,\bar{\boldsymbol{y}}\,)$: hence, β could be replaced by just $\{\; (\sigma^*, z^*) \;\}$ while still obtaining the same $\boldsymbol{d}^*$, and this is the basis of the aggregation (or compression) techniques that allow to keep the size of β low without impairing the convergence. Finally, the typical stopping criterion of Bundle algorithms

$$\mu\| z^* \|^2 + \sigma^* \le \varepsilon$$

can be easily understood in terms of the above ε-subgradients machinery, since $\bar{\boldsymbol{y}}$ would actually be σ^*-optimal if $\| z^* \|$ would be 0.

However, this dual viewpoint of Bundle algorithms suffers from some limitations: the first is that the dual problem is described in terms of a "local" object, the ε-subdifferential of φ (or $\varphi_{\beta(\bar{y})}$) in $\bar{\boldsymbol{y}}$, so that it is difficult to relate two subproblems corresponding to different current points, e.g. before and after a SS. The second is that the interpretation does not easily extend to other forms of Bundle methods where the quadratic proximal term is replaced by different convex functions, like

- the *characteristic function* of the ball of radius t under some norm $\|\cdot\|$: this amounts at restricting the next trial point inside a *trust region* [HL93b - Algorithm XV.2.1.1];
- *norm-like* nonnegative convex functions h such that $h(\,\boldsymbol{y}\,) = 0 \Leftrightarrow \boldsymbol{y} = \mathbf{0}$: in practice, the L_1 and L_∞ norms [KCL95].

However, a first step towards the developments of more general primal-dual relationships was made in [Fr97], where it was shown that

- when $B_p(t)$ is used as Trust Region for $p = 1$ and $p = \infty$, the subproblems are

  $(\Pi^p_{\beta(\bar{y}),t})$ $\quad \min_{\boldsymbol{d}} \{ \varphi_{\beta(\bar{y})}(\boldsymbol{d}) : 1/t \| \boldsymbol{d} \|_p \le 1 \}$

  $(\Delta^p_{\beta(\bar{y}),t})$ $\quad \min_{\theta} \{ t \| \boldsymbol{G}_\beta \theta \|_q + \alpha_{\beta(\bar{y})}\theta : \theta \in \Theta \}$

  for $q = \infty$ and $q = 1$ respectively: that is, a primal trust region in the L_1 norm corresponds to a dual penalty term in the L_∞ norm (and vice-versa).

- when $\|\cdot\|_p$ is used as the stabilizing term for $p = 1$ and $p = \infty$, the subproblems are

  $(\underline{\Pi}^p_{\beta(\bar{y}),t})$ $\quad \min_{\boldsymbol{d}} \{ \varphi_{\beta(\bar{y})}(\boldsymbol{d}) + 1/t \| \boldsymbol{d} \|_p \}$

  $(\underline{\Delta}^p_{\beta(\bar{y}),t})$ $\quad \min_{\theta} \{ \alpha_{\beta(\bar{y})}\theta : t \| \boldsymbol{G}_\beta \theta \|_q \le 1 , \theta \in \Theta \}$

  for $q = \infty$ and $q = 1$ respectively: that is, a primal penalty term in the L_1 norm corresponds to a dual trust region in the L_∞ norm (and vice-versa).

Hence, we were led to think that more general primal-dual relationships could be devised for a generic stabilizing term D_t: this will be the subject of the next paragraph.

To finish up this introduction, let us mention some important classes of Bundle methods that do not fall under our analysis: the first are *Proximal Level* methods [HL93b - Algorithm XV.2.3.1], that use

$$\min \{ \| \boldsymbol{d} \|^2 : \varphi_{\beta(\bar{y})}(\boldsymbol{d}) \le t \}$$

as the direction finding problem: the quadratic dual of this problem is

$$\min_{\theta} \{ 1/2 \| \boldsymbol{G}_\beta \theta \|^2 + (t\boldsymbol{e} + \alpha_{\beta(\bar{y})})\theta , \theta \in \Theta \}$$

where $\boldsymbol{e}$ is the all-1 vector. The second are *Analytic Centre Cutting Plane* methods [AV95] [Ne95] [GV97], that look at the polyhedron

$$\{ (v, \boldsymbol{y}) : \varphi_{\beta(\bar{y})}(\boldsymbol{y}) \le v \le 0 \} \subseteq \mathcal{R}^{m+1}$$

and compute its *analytic centre*, i.e. the solution (π, ω) of

$$\max_{\boldsymbol{d},v} \{ \Sigma_{(\sigma, z) \in \beta(\bar{y})} \ln(v - \boldsymbol{d}z + \sigma) + \ln(-v) : (v, \boldsymbol{d}) \in \mathcal{R}^{m+1} \}$$

to use ω as the next trial point. The third are ε-*descent algorithms* [HL93b - Algorithm XIV.3.4.2] that solve $(\Delta^{2,\varepsilon}_{\beta(\bar{y})})$ to find a direction along which to perform a line search: the quadratic dual of $(\Delta^{2,\varepsilon}_{\beta(\bar{y})})$ is

$(\Pi^{2,\varepsilon}_{\beta(\bar{y})})$ $\quad \min_{\boldsymbol{d},v,w} \{ v + \varepsilon w + 1/2 \| \boldsymbol{d} \|_2^2 : v \ge \boldsymbol{d}z - w\sigma , (\sigma, z) \in \beta(\bar{y}) , w \ge 0 \}$

and it is hard to directly recover a term related to $\varphi_{\beta(\bar{y})}(\boldsymbol{d})$ in this objective function. Finally, other algorithms have been proposed where $\varphi_{\beta(\bar{y})}$ is replaced by a *quadratic* approximation of φ [GM91]. However, the extension of the theory developed in this article to some (or all) of the above methods might be possible, and it is currently under research.

2. Primal-Dual relationships in Bundle methods

To develop primal-dual relationships for our class of generalized Bundle methods we need some tools from convex analysis: the first is the *conjugate* of the function φ

$$\varphi^*(z) = \sup_y \{ zy - \varphi(y) \}$$

that characterises the set of vectors that are slope of linear functions minorizing φ. If φ is a closed proper convex function, then φ^* is also a closed convex function with the following properties (among others) [HL93b - Chap. X]:

2.i) $(\varphi^*)^* = \varphi$ (the conjugate operation is self-dual)

2.ii) $\varphi_1 \le \varphi_2 \Rightarrow \varphi_1^* \ge \varphi_2^*$ ("monotonicity" of the conjugate operation)

2.iii) $(\varphi(\cdot + \bar{y}))^*(z) = \varphi^*(z) - z\bar{y}$ (effect of simple variable change)

2.iv) $z \in \partial_\varepsilon \varphi(y) \Leftrightarrow y \in \partial_\varepsilon \varphi^*(z)$ (duality of the subdifferential mappings)

2.v) $z \in \partial_\varepsilon \varphi(y) \Leftrightarrow \varphi(y) + \varphi^*(z) \le zy + \varepsilon$ (Fenchel's inequality)

2.vi) $zy = \varphi(y) + \varphi^*(z) \Leftrightarrow z \in \partial\varphi(y)$ (basic relation between the function values)

An important object is the *epigraph* of φ^*, in view of the following property [HL93a - Proposition IV.1.2.8]: any convex function φ can be expressed as the sup of the linear functions minorizing it

$$\varphi(y) = \sup_{\sigma,z} \{ zy - \sigma : (\sigma, z) \in \Sigma_\varphi \} \quad \text{where} \quad \Sigma_\varphi = \{ (\sigma, z) : zy - \sigma \le \varphi(y) \ \forall y \}.$$

But it is easy to see that $\Sigma_\varphi = Epi\ \varphi^*$, since by the definition of φ^*

$$zy - \sigma \le \varphi(y) \ \forall y \Leftrightarrow zy - \varphi(y) \le \sigma \ \forall y \Leftrightarrow \sup_d \{ zy - \varphi(y) \} \le \sigma \Leftrightarrow (\sigma, z) \in Epi\ \varphi^*$$

hence φ can always be considered as

$$(P_y) \qquad \varphi(y) = \sup_{\sigma,z} \{ zy - \sigma : (\sigma, z) \in Epi\ \varphi^* \} = -\inf_{\sigma,z} \{ \sigma - zy : (\sigma, z) \in Epi\ \varphi^* \}.$$

Furthermore, the conjugate is immediately related with the minimization of φ by the obvious

$$z(\Pi) = \inf_y \{ \varphi(y) \} = -\sup_y \{ \mathbf{0}y - \varphi(y) \} = -\varphi^*(\mathbf{0})$$

and $\varphi^*(\mathbf{0})$ can be rewritten in the following (apparently fancy) way:

$$\varphi^*(\mathbf{0}) = \inf_\sigma \{ \sigma : \sigma \ge \varphi^*(\mathbf{0}) \} = \inf_{\sigma,z} \{ \sigma : (\sigma, z) \in Epi\ \varphi^*, z = \mathbf{0} \}.$$

Hence, we can propose a dual for (Π):

$$(\Delta) \qquad \inf_{\sigma,g} \{ \sigma : (\sigma, z) \in Epi\ \varphi^*, z = \mathbf{0} \}.$$

This is a reasonable dual, since $z(\Pi) = -z(\Delta)$ and (Δ) deals with "dual" objects: $(\sigma, z) \in Epi\ \varphi^* \Leftrightarrow z$ is a subgradient of φ in some point and σ is its linearization error in $\mathbf{0}$. Furthermore, the problem (P_y) of computing $\varphi(y)$ is the Lagrangean Relaxation of (Δ) w.r.t. the constraints $z = \mathbf{0}$ using y as Lagrangean multipliers: hence, (Π) is the Lagrangean Dual of (Δ).

The same technique can also be used to develop a dual for the "stabilized" problem

$$(\Pi_t) \qquad \phi_t(\mathbf{0}) = \inf_y \{ \varphi(y) + D_t(y) \}$$

(where $\mathbf{0}$ is assumed as the current point): in fact, it can be shown that the dual of (Π_t) is

(Δ_t) $\quad \inf_{\sigma,z} \{ \sigma + D_t^*(-z) : (\sigma, z) \in Epi\ \varphi^* \}.$

due to the following general result.

Lemma 2.1 *If f is a closed convex function, the following problems are duals of each other:*

(PP) $\quad \inf_{x,w} \{ \boldsymbol{cx} + \boldsymbol{dw} + f(\boldsymbol{x}) : \boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{D}_i\boldsymbol{w} \geq b_i\ ,\ i \in V \}$

(DP) $\quad \sup_s \{ \Sigma_{i \in V}\, b_i s_i - f^*(\Sigma_{i \in V}\, \boldsymbol{A}_i s_i - \boldsymbol{c}) : \Sigma_{i \in V}\, \boldsymbol{D}_i s_i = \boldsymbol{d}\ ,\ \boldsymbol{s} \geq \mathbf{0} \}$

Proof: rewrite (DP) as

$$\sup_{s,z} \{ \Sigma_{i \in V}\, b_i s_i - f^*(\boldsymbol{z}) : \Sigma_{i \in V}\, \boldsymbol{D}_i s_i = \boldsymbol{d}\ ,\ \boldsymbol{z} = \Sigma_{i \in V}\, \boldsymbol{A}_i s_i - \boldsymbol{c}\ ,\ \boldsymbol{s} \geq \mathbf{0} \}$$

and consider its Lagrangean Dual w.r.t. both the constraints $\Sigma_{i \in V}\, \boldsymbol{D}_i s_i = \boldsymbol{d}$ (with multipliers $\boldsymbol{w}$) and the constraints $\boldsymbol{z} = \Sigma_{i \in V}\, \boldsymbol{A}_i s_i - \boldsymbol{c}$ (with multipliers $\boldsymbol{x}$):

(DP*) $\quad \inf_{x,w} \{ \boldsymbol{cx} + \boldsymbol{dw} + \sup_{s \geq \mathbf{0}} \{ \Sigma_{i \in V}\, (b_i - \boldsymbol{x}\boldsymbol{A}_i - \boldsymbol{w}\boldsymbol{D}_i) s_i \} + \sup_z \{ \boldsymbol{xz} - f^*(\boldsymbol{z}) \} \}.$

The first inner sup is $+\infty$ if $b_i - \boldsymbol{x}\boldsymbol{A}_i - \boldsymbol{w}\boldsymbol{D}_i > 0$ for some $i \in V$ and 0 otherwise, hence it is equivalent to the constraints in (PP): the second inner sup is just the definition of $(f^*)^* = f$, hence (DP*) is the same problem as (PP). But (DP*) is equivalent to (DP), since the conditions for the absence of duality gap [HL93b - Corollary XII.2.3.6.(ii)] are fulfilled: hence, (PP) is the dual of (PD).
Note that, at least if V is finite, one can obtain the same result starting with (PP) and writing its Lagrangean Dual w.r.t. the constraints $\boldsymbol{Ax} + \boldsymbol{Dw} \geq \boldsymbol{b}$ (with multipliers $\boldsymbol{s}$)

(PP*) $\quad \sup_{s \geq \mathbf{0}} \{ \Sigma_{i \in V}\, b_i s_i + \inf_x \{ (\boldsymbol{c} - \Sigma_{i \in V}\, \boldsymbol{A}_i s_i)\boldsymbol{x} + f(\boldsymbol{x}) \} + \inf_w \{ (\boldsymbol{d} - \Sigma_{i \in V}\, \boldsymbol{D}_i s_i)\boldsymbol{w} \} \}$

showing that the dual optimal solution $\boldsymbol{s}^*$ is the vector of optimal Lagrangean multipliers for the primal constrains and vice-versa. ♦

This result is an application of *Fenchel's Duality* [HL93b - §XII.5.4]: however, some arguments used in our proof will be useful later on. Note that the set of constraints V need not to be finite: hence, Lemma 2.1 can be directly applied to (Π_t) rewritten as

(Π_t) $\quad \inf_{v,y} \{ v + D_t(\boldsymbol{y}) : v \geq \boldsymbol{z}\boldsymbol{y} - \sigma : (\sigma, z) \in Epi\ \varphi^* \}$

yielding $\quad \inf_\theta \{ \Sigma_{(\sigma,z) \in Epi\ \varphi^*}\, \sigma\theta_{(\sigma,z)} + D_t^*(-\Sigma_{(\sigma,z) \in Epi\ \varphi^*}\, z\theta_{(\sigma,z)}) : \Sigma_{(\sigma,z) \in Epi\ \varphi^*}\, \theta_{(\sigma,z)} = 1\ ,\ \theta \geq \mathbf{0} \}$

i.e. (Δ_t) as desired, since $Conv(\ Epi\ \varphi^*\ ) = Epi\ \varphi^*$. Of course, the same arguments apply when the cutting plane model is substituted to φ: hence, the dual of

$(\Pi_{\beta,t})$ $\quad \inf_y \{ \varphi_\beta(\boldsymbol{y}) + D_t(\boldsymbol{y}) \}$

can be easily seen to be

$(\Delta_{\beta,t})$ $\quad \inf_{\sigma,z} \{ \sigma + D_t^*(-z) : (\sigma, z) \in Epi\ \varphi_\beta^* \}$

or, equivalently,

$(\Delta_{\beta,t})$ $\quad \inf_\theta \{ \alpha_\beta\theta + D_t^*(-\boldsymbol{G}_\beta\theta) : \theta \in \Theta \}$

(in fact, $Epi\ \varphi^*_\beta = \{\ (\sigma, z) : \sigma \geq \alpha_\beta\theta\ ,\ z = \boldsymbol{G}_\beta\theta\ ,\ \theta \in \Theta\ \}$ [HL93b - Proposition X.3.4.1]).

It is now possible to introduce the concept of "current point" directly in the dual problem, with a quite interesting construction: assuming that $D_t^*(\ \mathbf{0}\ ) = 0$ (i.e. that $\mathbf{0}$ is a minimum of D_t, cfr. 2.iv), (Δ) is also obviously equivalent to

$$\inf_{\sigma,g}\ \{\ \sigma + D_t^*(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*\ ,\ z = \mathbf{0}\ \}$$

and therefore to the Lagrangean Dual of the above problem w.r.t. the constraints $z = \mathbf{0}$ (cfr. the proof of Lemma 2.1), i.e.

$$(\Delta_t^*) \qquad \sup_y\ \{\ \inf_{\sigma,g}\ \{\ \sigma - \boldsymbol{y}z + D_t^*(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*\ \}\ \}.$$

The inner problem

$$(\Delta_{\bar{y},t}) \qquad \inf_{\sigma,g}\ \{\ \sigma - \bar{\boldsymbol{y}}z + D_t^*(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*\ \}$$

for a fixed $\bar{\boldsymbol{y}}$ has a nice interpretation in terms of the scaled function

$$\varphi_{\bar{y}}(\ \boldsymbol{d}\ ) = \varphi(\ \bar{\boldsymbol{y}} + \boldsymbol{d}\ ) - \varphi(\ \bar{\boldsymbol{y}}\ )$$

due to the following result:

Lemma 2.2: $(\sigma, z) \in\ Epi\ \varphi^* \Leftrightarrow (\sigma - \bar{\boldsymbol{y}}z + \varphi(\ \bar{\boldsymbol{y}}\ ), z) \in\ Epi\ \varphi^*_{\bar{y}}$.

Proof: it is easy to see that $\varphi^*_{\bar{y}}(\ z\ ) = \varphi^*(\ z\ ) - z\bar{\boldsymbol{y}} + \varphi(\ \bar{\boldsymbol{y}}\ )$ (use 2.iii and the definition of φ^*): then

$$(\sigma, z) \in\ Epi\ \varphi^* \Leftrightarrow \sigma \geq \varphi^*(\ z\ ) \Leftrightarrow \sigma - z\bar{\boldsymbol{y}} + \varphi(\ \bar{\boldsymbol{y}}\ ) \geq \varphi^*(\ z\ ) - z\bar{\boldsymbol{y}} + \varphi(\ \bar{\boldsymbol{y}}\ )$$

$$\Leftrightarrow \quad \sigma - z\bar{\boldsymbol{y}} + \varphi(\ \bar{\boldsymbol{y}}\ ) \geq \varphi^*_{\bar{y}}(\ z\ ).$$ ♦

Therefore, $(\Delta_{\bar{y},t})$ is equivalent (but for a constant term $-\varphi(\ \bar{\boldsymbol{y}}\ )$) to

$$(\Delta_t^{\bar{y}}) \qquad \inf_{\sigma,g}\ \{\ \sigma + D_t^*(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*_{\bar{y}}\ \}$$

whose dual is

$$(\Pi_t^{\bar{y}}) \qquad \phi_t(\ \bar{\boldsymbol{y}}\ ) = \inf_d\ \{\ \varphi_{\bar{y}}(\ \boldsymbol{d}\ ) + D_t(\ \boldsymbol{d}\ )\ \}$$

i.e. the calculation of the *generalized Moreau-Yosida Regularization* of φ in $\bar{\boldsymbol{y}}$: analogously, the "pricing" subproblem of computing $\varphi_{\bar{y}}(\ \boldsymbol{d}\ )$ can be rewritten as

$$(P_d^{\bar{y}}) \qquad \varphi_{\bar{y}}(\ \boldsymbol{d}\ ) = -\ \inf_{\sigma,g}\ \{\ \sigma - z(\ \bar{\boldsymbol{y}} + \boldsymbol{d}\ ) : (\sigma, z) \in\ Epi\ \varphi^*\ \} - \varphi(\ \bar{\boldsymbol{y}}\ )$$

$$= \ -\ \inf_{\sigma,g}\ \{\ (\ \sigma - z\bar{\boldsymbol{y}} + \varphi(\ \bar{\boldsymbol{y}}\ )\ ) - z\boldsymbol{d} : (\sigma, z) \in\ Epi\ \varphi^*\ \}$$

$$= \ -\ \inf_{\sigma,g}\ \{\ \sigma - z\boldsymbol{d} : (\sigma, z) \in\ Epi\ \varphi^*_{\bar{y}}\ \}.$$

Obviously, the same game can be played with $\varphi_{\bar{y}}$ replaced by its "approximated" version $\varphi_{\beta(\bar{y})}$, ending in the following pair of dual problems

$$(\Pi_{\beta,t}^{\bar{y}}) \qquad \inf_d\ \{\ \varphi_{\beta(\bar{y})}(\ \boldsymbol{d}\ ) + D_t(\ \boldsymbol{d}\ )\ \}$$

$$(\Delta_{\beta,t}^{\bar{y}}) \qquad \inf_{\sigma,z}\ \{\ f_t(\sigma, z) = \sigma + D_t^*(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*_{\beta(\bar{y})}\ \}$$

that can be equivalently rewritten as

$(\Pi^{\bar{y}}_{\beta,t})$ $\quad \inf_{v,\boldsymbol{d}} \{\ v + D_t(\ \boldsymbol{d}\ ) : v\boldsymbol{e} \geq \boldsymbol{d}\boldsymbol{G}_\beta - \alpha_{\beta(\bar{y})}\ \}$

$(\Delta^{\bar{y}}_{\beta,t})$ $\quad \inf_\theta \{\ \alpha_{\beta(\bar{y})}\theta + D^*_t(\ -\boldsymbol{G}_\beta\theta\ ) : \theta \in \Theta\ \}.$

However, these problems have a "local flavour", since they are written in terms of $\varphi_{\beta(\bar{y})}$ and *Epi* $\varphi^*_{\beta(\bar{y})}$ that change with the current point $\bar{\boldsymbol{y}}$: their "global" versions are (start from (Δ^*_t) restricted to β or use Lemma 2.2 in the previous definition)

$(\Pi_{\beta,\bar{y},t})$ $\quad \inf_{\boldsymbol{d}} \{\ \varphi_\beta(\ \bar{\boldsymbol{y}} + \boldsymbol{d}\ ) + D_t(\ \boldsymbol{d}\ )\ \}$

$(\Delta_{\beta,\bar{y},t})$ $\quad \inf_{\sigma,z} \{\ f_{\bar{y},t}(\sigma, z) = \sigma - \bar{\boldsymbol{y}}z + D^*_t(\ -z\ ) : (\sigma, z) \in\ Epi\ \varphi^*_\beta\ \}.$

The next paragraph will heavily rely on these primal-dual relationships: hence, let us introduce some useful relations between the optimal solution $(\sigma^*,\ z^*) = (\alpha_{\beta(\bar{y})}\theta^*,\ \boldsymbol{G}_\beta\theta^*)$ of $(\Delta^{\bar{y}}_{\beta,t})$ and the optimal solution $(v^* = \varphi_{\beta(\bar{y})}(\ \boldsymbol{d}^*\ ),\ \boldsymbol{d}^*)$ of $(\Pi^{\bar{y}}_{\beta,t})$ (assumed to exist).

Lemma 2.3: $v^* = \varphi_{\beta(\bar{y})}(\ \boldsymbol{d}^*\ ) = \boldsymbol{d}^*\boldsymbol{z}^* - \sigma.$

Proof: from Lemma 2.1, θ^* are the optimal Lagrangean multipliers of the constraints $v\boldsymbol{e} \geq \boldsymbol{d}\boldsymbol{G}_\beta - \alpha_{\beta(\bar{y})}$ in $(\Pi_{\beta(\bar{y}),t})$, hence the *Complementary Slackness condition* requires

$$(\ v^*\boldsymbol{e} + \alpha_{\beta(\bar{y})} - \boldsymbol{d}^*\boldsymbol{G}_\beta\ )\theta^* = 0 \quad \Rightarrow \quad v^* = \boldsymbol{d}^*\boldsymbol{z}^* - \sigma^*. \qquad \blacklozenge$$

Lemma 2.4: $-z^* \in \partial D_t(\ \boldsymbol{d}^*\ )$ *and* $\boldsymbol{d}^* \in \partial D^*_t(\ -z^*\ ).$

Proof: from Lemma 2.3, $\varphi_{\beta(\bar{y})}(\ \boldsymbol{d}^*\ ) = \boldsymbol{d}^*\boldsymbol{z}^* - \sigma^*$, but from duality

$$z(\Pi_{\beta(\bar{y}),t}) = \varphi_{\beta(\bar{y})}(\ \boldsymbol{d}^*\ ) + D_t(\ \boldsymbol{d}^*\ ) = -\ z(\Pi_{\beta(\bar{y}),t}) = -\ \sigma^* - D^*_t(\ -z^*\ )$$

so that $\quad D^*_t(\ -z^*\ ) + D_t(\ \boldsymbol{d}^*\ ) = -\ \boldsymbol{d}^*\boldsymbol{z}^*$

which, in view of 2.vi and 2.iv, implies the thesis. $\blacklozenge$

Lemma 2.5: $(1, -\boldsymbol{d}^*) \in \partial f_t(\ \sigma^*, z^*\ )$; *furthermore, it "cuts away" all Epi* $\varphi^*_{\beta(\bar{y})}$*, i.e.*

$$\begin{pmatrix} 1 \\ -\boldsymbol{d}^* \end{pmatrix}\left[\ \begin{pmatrix} \alpha \\ \boldsymbol{g} \end{pmatrix} - \begin{pmatrix} \sigma^* \\ z^* \end{pmatrix}\ \right] \geq 0 \qquad \forall(\alpha, \boldsymbol{g}) \in\ Epi\ \varphi^*_{\beta(\bar{y})}.$$

Proof: the first relation comes from Lemma 2.4 plus direct calculation:

$$D^*_t(\ -\boldsymbol{g}\ ) \geq D^*_t(\ -z^*\ ) + \boldsymbol{d}^*(\ (-\boldsymbol{g}) - (-z^*)\ )\ \forall \boldsymbol{g} \Rightarrow$$

$$\alpha + D^*_t(\ -\boldsymbol{g}\ ) \geq (\ \sigma^* + D^*_t(\ -z^*\ )\ ) - \boldsymbol{d}^*(\ \boldsymbol{g} - z^*\ ) + (\ \alpha - \sigma^*\ )\ \forall(\alpha, \boldsymbol{g}).$$

For the second, assume by contradiction that $\exists(\bar{\alpha}, \bar{\boldsymbol{g}}) \in\ Epi\ \varphi^*_{\beta(\bar{y})}$ s.t. $(\ \bar{\alpha} - \sigma^*\ ) - \boldsymbol{d}^*(\ \bar{\boldsymbol{g}} - z^*\ ) < 0$ and define $\bar{\boldsymbol{w}} = z^* - \bar{\boldsymbol{g}}$: from Lemma 2.1, $\boldsymbol{d}^*$ are the optimal Lagrangean multipliers for the constraints $\boldsymbol{s} = \boldsymbol{G}_\beta\theta$ in

$$\inf_{\theta,\boldsymbol{s}} \{\ \alpha_{\beta(\bar{y})}\theta + D^*_t(\ -\boldsymbol{s}\ ) : \boldsymbol{s} = \boldsymbol{G}_\beta\theta\ ,\ \theta \in \Theta\ \}.$$

The *value function* for this problem w.r.t. the constraints $\boldsymbol{s} = \boldsymbol{G}_\beta\theta$ is

$$v(\ \boldsymbol{w}\ ) = \inf_{\theta,\boldsymbol{s}} \{\ \alpha_{\beta(\bar{y})}\theta - \bar{\boldsymbol{y}}\boldsymbol{s} + D^*_t(\ -\boldsymbol{s}\ ) : \boldsymbol{s} - \boldsymbol{G}_\beta\theta = \boldsymbol{w}\ ,\ \theta \in \Theta\ \}$$

$$= \inf_{\sigma,z} \{ f_t(\sigma, z + \boldsymbol{w}) : (\sigma, z) \in Epi\ \varphi^*_{\beta(\bar{y})} \};$$

note that $v(\bar{\boldsymbol{w}}) \le f_t(\bar{\alpha}, \bar{\boldsymbol{g}} + \bar{\boldsymbol{w}}) = \bar{\alpha} + D_t^*(-(\bar{\boldsymbol{g}} + (z^* - \bar{\boldsymbol{g}})))$ just because $(\bar{\alpha}, \bar{\boldsymbol{g}}) \in Epi\ \varphi^*_{\beta(\bar{y})}$, hence

$$v(\bar{\boldsymbol{w}}) \le \bar{\alpha} + D_t^*(-z^*) = \sigma^* + D_t^*(-z^*) + (\bar{\alpha} - \sigma^*)$$

$$< \sigma^* + D_t^*(-z^*) + \boldsymbol{d}^*(\bar{\boldsymbol{g}} - z^*) = f_t(\sigma^*, z^*) - \boldsymbol{d}^*\bar{\boldsymbol{w}} = v(\boldsymbol{0}) + (-\boldsymbol{d}^*)(\bar{\boldsymbol{w}} - \boldsymbol{0}).$$

But it is well-known [Ge69 - Lemma 4] that $-\boldsymbol{d}^* \in \partial v(\boldsymbol{0})$, which is contradicted by the above inequality: hence, the Lemma is proved. ♦

Lemma 2.6: $z(P^{\bar{y}}_{\boldsymbol{d}^*}) + f_t(\sigma^*, z^*) - (\sigma^* - \boldsymbol{d}^* z^*) \le z(\Delta^{\bar{y}}_t) \le z(\Delta^{\bar{y}}_{\beta,t})$

$$\Rightarrow\ 0 \le (\sigma^* - \boldsymbol{d}^* z^*) - z(P^{\bar{y}}_{\boldsymbol{d}^*}) = \varphi_{\bar{y}}(\boldsymbol{d}^*) - v^* \le z(\Delta^{\bar{y}}_{\beta,t}) - z(\Delta^{\bar{y}}_t)$$

Proof: from the first part Lemma 2.5

$$h(\alpha, \boldsymbol{g}) = f_t(\sigma^*, z^*) + \begin{pmatrix} 1 \\ -\boldsymbol{d}^* \end{pmatrix} \left[\begin{pmatrix} \alpha \\ \boldsymbol{g} \end{pmatrix} - \begin{pmatrix} \sigma^* \\ z^* \end{pmatrix} \right] \le f_t(\alpha, \boldsymbol{g})\ \forall (\alpha, \boldsymbol{g});$$

take the inf over $Epi\ \varphi^*_{\bar{y}}$ on both sides, recognise the definition of $(P^{\bar{y}}_{\boldsymbol{d}^*})$ and use $z(\Delta^{\bar{y}}_t) \le z(\Delta^{\bar{y}}_{\beta,t}) = f_t(\sigma^*, z^*)$ and Lemma 2.4 to complete the proof. ♦

The last two Lemmas could be equivalently restated in "global" notation by substituting $(\bar{\boldsymbol{y}} + \boldsymbol{d}^*)$ for $\boldsymbol{d}^*$, $f_{\bar{y},t}$ for f_t, $Epi\ \varphi^*_{\beta}$ for $Epi\ \varphi^*_{\beta(\bar{y})}$ and the "global" problems for their "local" versions.
Let us now summarise what has been obtained so far:

- $(\Delta_{\bar{y},t})$ is a *generalised Augmented Lagrangean relaxation* of (Δ), where the constraints $\boldsymbol{z} = \boldsymbol{0}$ are replaced with the *linear term* $-\bar{\boldsymbol{y}}\boldsymbol{z}$ (with Lagrangean multipliers $\bar{\boldsymbol{y}}$) and the *nonlinear term* $D_t^*(-\boldsymbol{z})$ in the objective function;
- minimizing $f_{\bar{y},t}$ over $Epi\ \varphi^*$ (equivalently: minimizing f_t over $Epi\ \varphi^*_{\bar{y}}$) is "difficult": however, the black box allows to minimize *any linear function* over $Epi\ \varphi^*$;
- the linear function h of Lemma 2.6 minorizes f_t: hence, the value of $z(\Delta^{\bar{y}}_t)$ can be "bracketed" using $(P_{\bar{y}})$ as a lower bound and $(\Delta^{\bar{y}}_{\beta,t})$ as an upper bound;
- $(\Delta^{\bar{y}}_{\beta,t})$ is a restriction of $z(\Delta^{\bar{y}}_t)$ where the unknown set $Epi\ \varphi^*_{\bar{y}}$ is approximated with its subset $Epi\ \varphi^*_{\beta(\bar{y})}$; $(P^{\bar{y}}_{\boldsymbol{d}^*})$ can be used to generate new $(\alpha, \boldsymbol{g}) \in Epi\ \varphi^*$, and the above Lemmas guarantee that they must be "new" (not already belonging to $Epi\ \varphi^*_{\beta(\bar{y})}$) unless $z(\Delta^{\bar{y}}_{\beta,t}) = z(\Delta^{\bar{y}}_t)$.

All the above suggest the following dual interpretation of Bundle methods: we are applying an ascent approach to the solution of (Δ_t^*) (that is equivalent to (Δ)), where we are not able to solve the Lagrangean subproblem $(\Delta_{\bar{y},t})$ exactly, but only to approximate it from above *and* below. Clearly, sequences of NS are aimed to refining the approximation in the solution of $(\Delta_{\bar{y},t})$, i.e. closing the gap between $z(\Delta^{\bar{y}}_{\beta,t})$ and $z(P_{\bar{y}})$, while SS do the real ascent. In the next paragraph, we are going to prove convergence of the method: the "dual language" just developed will be instrumental.

3. Convergence

In this paragraph, we will provide convergence analysis for the generic Bundle algorithm of Figure 2.

Bundle Algorithm

⟨ let $\mu \geq 1$ and $\varepsilon > 0$ be fixed; choose the initial $\bar{y}$, t and β ⟩; /* initialization */
do
⟨ solve $(\Pi^{\bar{y}}_{\beta,t})$ and $(\Delta^{\bar{y}}_{\beta,t})$ for $(v^*, \boldsymbol{d}^*)$ and (σ^*, z^*) ⟩; /* find a direction */
⟨ move along $\boldsymbol{d}^*$, generating some new (σ, z) and a trial point y ⟩; /* evenctual LS */
if(a large enough improvement has been obtained)
then $\bar{y} = y$; /* a Serious Step */
⟨ update β, (possibly) including some new (σ, z), deleting old ones and/or adding convex combination of the present ones (typically (σ^*, z^*)) ⟩; /* the β-strategy */
⟨ update t, depending on the previous iterations ⟩; /* the t-strategy */
while($\sigma^* + \mu D_t^*(-z^*) > \varepsilon$);

Figure 2: the Bundle Algorithm

Obviously, to obtain a convergent algorithm some assumptions are needed about

- properties of the stabilizing term D_t,
- properties of the function φ,
- handling of the Bundle (the β-strategy),
- handling of the t parameter (the t-strategy),
- the NS/SS decision (definition of the "large enough improvement").

The last two will be discussed together later on, while the β-strategy will be discussed during the convergence proof: hence, let's start with the assumptions on D_t and φ. For D_t, we require five basic properties to hold: all of them can be equivalently expressed in terms of the conjugate D_t^*. Two other properties eventually allow to refine some convergence proofs.

P1) for each t, $D_t(\mathbf{0}) = 0$ and $\mathbf{0} \in \partial D_t(\mathbf{0})$.

P*1) for each t, $D_t^*(\mathbf{0}) = 0$ and $\mathbf{0} \in \partial D_t^*(\mathbf{0})$.

In view of 2.iv and 2.vi, P1 and P*1 are clearly equivalent: note that, having a minimizer in $\mathbf{0}$ where they evaluate to 0, both D_t and D_t^* are nonnegative functions for each t.

Observation 3.1: $S_\varepsilon(D_t) = \partial_\varepsilon D_t^*(\mathbf{0})$ *and* $S_\varepsilon(D_t^*) = \partial_\varepsilon D_t(\mathbf{0})$.

Proof: by 2.v, $z \in \partial_\varepsilon D_t(\mathbf{0}) \Leftrightarrow D_t^*(z) + D_t(\mathbf{0}) - z\mathbf{0} \leq \varepsilon \Leftrightarrow z \in S_\varepsilon(D_t^*)$; the rest follows by symmetry. ♦

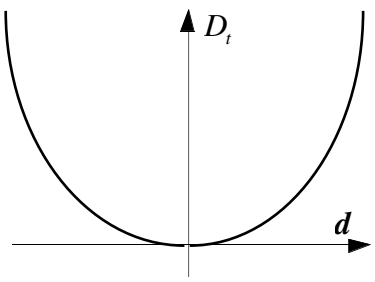
P2) for each t and $\varepsilon > 0$, $S_\varepsilon(D_t)$ is *compact* and $\mathbf{0} \in$ *int* $S_\varepsilon(D_t)$ (it is *full-dimensional*).

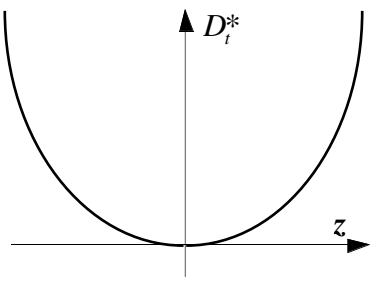
P*2) for each t and $\varepsilon > 0$, $S_\varepsilon(D_t^*)$ is *compact* and $\mathbf{0} \in$ *int* $S_\varepsilon(D_t^*)$ (it is *full-dimensional*).

Again, P2 and P*2 are equivalent: full-dimensionality of $S_\varepsilon(D_t)$ implies compactness of $S_\varepsilon(D_t^*) = \partial_\varepsilon D_t(\mathbf{0})$, since [HL93b - Theorem XI.1.1.4] $\partial_\varepsilon D_t(\mathbf{0})$ is compact $\Leftrightarrow \mathbf{0} \in$ *int Dom* D_t. On the other hand, if $\partial_\varepsilon D_t(\mathbf{0})$ is *not* full-dimensional then there exist a direction along which D_t is constant(ly 0), contradicting the compactness of $S_\varepsilon(D_t)$. A rephrasing of P2 is therefore that *both* the level sets of D_t and its ε-subdifferentials must be compact. The first half of P2 (compactness) rules out the possibility of using the constant zero function $0(\boldsymbol{d}) = 0$ as D_t (i.e. $I_{\{\mathbf{0}\}}$ as D_t^*): we will see that it is necessary only if the next property - which guarantees that $(\Delta_{\beta,\bar{\boldsymbol{y}},t})$ and $(\Pi_{\beta,\bar{\boldsymbol{y}},t})$ attain finite solution - is *not* satisfied.

Note that the above properties are stated for $\varepsilon > 0$: nothing is said about full-dimensionality of $S_0(D_t)$ or $S_0(D_t^*)$. In practice, they may or may not be full-dimensional, as shown in the following examples:

Ex. 3.1

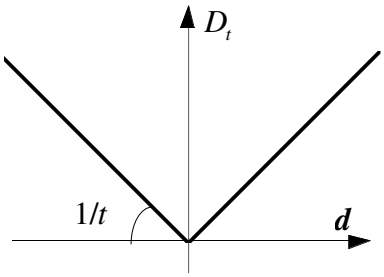


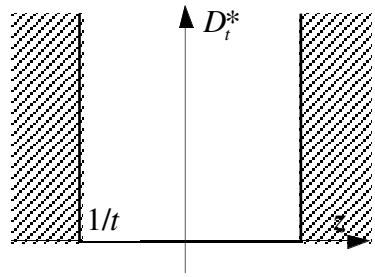


$D_t(\boldsymbol{d}) = {}^1\!/_{2t}\|\boldsymbol{d}\|_2^2$
$D_t^*(\boldsymbol{z}) = {}^1\!/_{2t}\|\boldsymbol{z}\|_2^2$

none is full-dimensional

Ex. 3.2

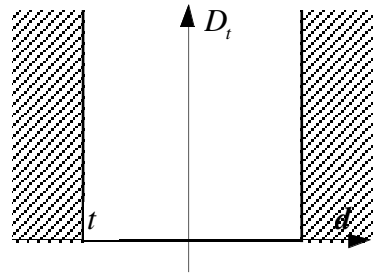


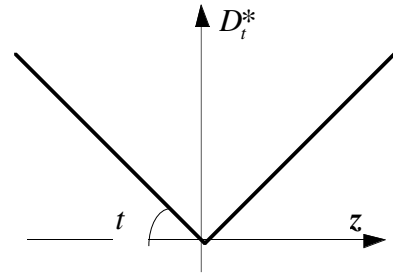


$D_t(\boldsymbol{d}) = {}^1\!/_t\|\boldsymbol{d}\|_1$
$D_t^*(\boldsymbol{z}) = I_{B_\infty(1/t)}(\boldsymbol{z})$

$S_0(D_t^*)$ *but not* $S_0(D_t)$ is full-dimensional

Ex. 3.3





$D_t(\boldsymbol{d}) = I_{B_\infty(t)}(\boldsymbol{d})$
$D_t^*(\boldsymbol{z}) = t\|\boldsymbol{d}\|_1$

$S_0(D_t)$ *but not* $S_0(D_t^*)$ is full-dimensional

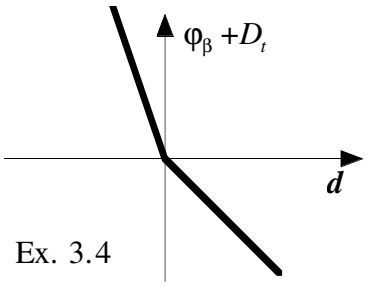
In Ex. 3.1, both D_t and its conjugate are differentiable (in $\mathbf{0}$) $\Rightarrow$ both have unique minimizer $\mathbf{0}$, while in Ex. 3.2 and Ex. 3.3 only one of the properties hold. Although apparently symmetric, the latter two cases are *not* equivalent (cfr. P6 below).

P3) for each t, $\lim_{\|\boldsymbol{d}\|\to\infty} D_t(\boldsymbol{d}) / \|\boldsymbol{d}\| = +\infty$ (D_t is *strongly coercive*).

P*3) for each t, D_t^* is *finite everywhere*.

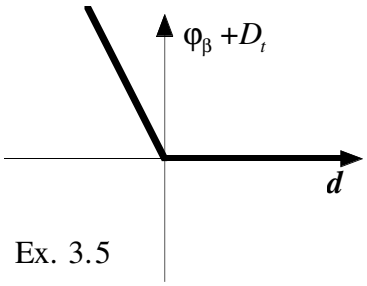
The equivalence between P3 and P*3 is [HL93b - Remark X.1.3.10]: strongly coercive functions (sometimes called 1-coercive) increase faster than any linear function at infinity. P3 guarantees that $(\Pi_{\beta(\bar{\boldsymbol{y}}),t})$ has a bounded nonempty set of optimal solutions [HL93a - Remark IV.3.2.8], since $\varphi_{\beta(\bar{\boldsymbol{y}})}(\boldsymbol{d}) + D_t(\boldsymbol{d})$ is strongly coercive too ($\varphi_{\beta(\bar{\boldsymbol{y}})}$ is polyhedral): from the dual viewpoint, the feasible

set of ($\Delta_{\beta(\bar{y}),t}$) is compact and its objective function is finite everywhere. Note that the stabilising term in Ex. 3.2 is *not* strongly coercive: in fact, it is easy to construct examples (see Ex 3.4-5-6) where the optimal solutions set of ($\Pi_{\beta(\bar{y}),t}$) is either empty or unbounded.



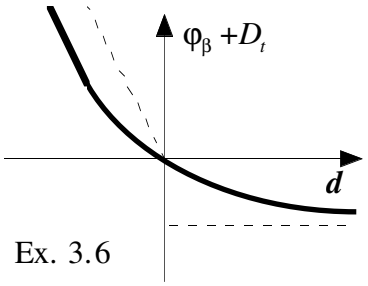
Ex. 3.4

$\varphi_\beta(d) = -2d,\ D_t(d) = |d|$



Ex. 3.5

$\varphi_\beta(d) = -d,\ D_t(d) = |d|$



Ex. 3.6

$\varphi_\beta(d) = -d,\ D_t(d) = \dfrac{d^2}{1 + |d|}$

It is possible, both in theory and in practice, to use stabilizing functions that are not strongly coercive, provided that other assumptions guarantee that the subproblems attain finite solutions: one possibility is that the CP model be always bounded below (from the dual viewpoint, that one element in the form (α, **0**) belongs to β).

Assumption 3.2: φ *is* bounded below, *a* $\delta \le z(\Pi)$ *is* known *and* $(\varphi(\bar{\boldsymbol{y}}) - \delta, \mathbf{0}) \in \beta(\bar{\boldsymbol{y}})$.

Lemma 3.3: *under Assumption 3.2,* $\boldsymbol{d}^* \in S_\alpha(D_t)$ *where* $\alpha = \varphi(\bar{\boldsymbol{y}}) - \delta$.

Proof: just note that $\varphi_{\beta(\bar{y})}(\boldsymbol{d}) \ge -\alpha\ \forall \boldsymbol{d}$ so that $\varphi_{\beta(\bar{y})}(\boldsymbol{d}) + D_t(\boldsymbol{d}) > 0$ outside $S_\alpha(D_t)$: the latter is a compact set by P2, hence the infimum must be finitely attained. ♦

Note that there are three separate conditions in Assumption 3.2: a suitable δ must *exist*, must *be known* and the corresponding "flat" subgradient must *be explicitly kept* in the Bundle. Obviously, a more general condition would be requiring that $\varphi_{\beta(\bar{y})}$ itself be always bounded below - in this case, the CP algorithm could be directly applied without stabilization. The first part of P2 (compactness) is only used in the proof of Lemma 3.3, hence it is not required if P3 is enforced: however, strong coercivity implies boundedness of the level sets [HL93a - Proposition IV.3.2.5.(ii)], hence there is no loss of generality - and a gain in symmetry - in requiring it to hold. In the following, we will always assume that at least one among P3 and Assumption 3.2 is in force, to be guaranteed of the existence of finite optimal solutions for the subproblems.
Now, let us move to consider the role of *t*.

P4) $D_t \ge D_\tau$ for each $\tau \ge t$ (D_t is *nonincreasing* in *t*).

P*4) $D_t^* \le D_\tau^*$ for each $\tau \ge t$ (D_t is *nondecreasing* in *t*).

The equivalence between these two properties is just 2.ii: in the practical cases seen so far (see Ex. 3.1-2-3 above) a factor of $1/t$ enters in D_t while a factor of t enters in D_t^*.

P5) $\lim_{t\to\infty} D_t(\boldsymbol{d}) = 0\ \forall \boldsymbol{d}$ ($\{ D_t \}$ *converges pointwise* to the constant zero function $0(\cdot)$).

P*5) $\lim_{t\to\infty} D_t^*(\boldsymbol{z}) = +\infty\ \forall \boldsymbol{z} \neq \mathbf{0}$ ($\{ D_t^* \}$ *converges pointwise* to $I_{\{\mathbf{0}\}}$).

As usual, P5 and P*5 are equivalent: would a $\bar{\boldsymbol{z}} \neq \mathbf{0}$ exist such that $D_t^*(\bar{\boldsymbol{z}}) = \sup_{\boldsymbol{d}} \{ \boldsymbol{d}\bar{\boldsymbol{z}} - D_t(\boldsymbol{d}) \} \le$

$\delta < +\infty\ \forall t$, it would be $\boldsymbol{d}\bar{z} - D_t(\boldsymbol{d}) \le \delta\ \forall \boldsymbol{d} \forall t \Rightarrow D_t(\boldsymbol{d}) \ge \boldsymbol{d}\bar{z} - \delta\ \forall \boldsymbol{d} \forall t$, hence $D_t(\alpha / \| \bar{z} \|_2^2)\bar{z}) \ge \alpha - \delta\ \forall t$ which contradicts P5 whenever $\alpha > \delta$.
This condition is intuitively clear: as the "strength" of the stabilizing term goes to zero, the stabilized problem ($\Pi_{\bar{y},t}$) must "go" to the original problem (Π). The following two results give a better picture of the kind of convergence that can be expected, while the subsequent (easy) one establish that P5 is well-suited for our purposes, at least if no SS were done and we would be able to solve ($\Delta_{\bar{y},t}$) exactly.

Observation 3.4: *under P5,* $\{ D_t \}$ *converges* uniformly *to* $0(\cdot)$ *on every compact set C, i.e.* $\forall \varepsilon > 0$ $\exists \bar{t}$ *such that* $D_t(\boldsymbol{d}) \le \varepsilon\ \forall \boldsymbol{d} \in C\ \forall t \ge \bar{t}$.

Proof: use [HL93a - Theorem IV.3.1.5], the fact that *ri Dom* $0(\cdot) = \Re^n$ and P3. ♦

Observation 3.5: *under P5,* $\forall \varepsilon > 0\ \forall \delta > 0\ \exists \bar{t}$ *such that* $S_\varepsilon(D_t) \supseteq B_2(\delta)\ \forall t \ge \bar{t}$.

Proof: use Observation 3.4 with $C = B_2(\delta)$: $S_\varepsilon(D_t)$ are nondecreasing in t. ♦

Observation 3.6: *under P5,* $\lim_{t\to\infty} z(\Pi_{\bar{y},t}) = z(\Pi)$.

Proof: if $\lim_{t\to\infty} z(\Pi_{\bar{y},t}) = \bar{\varphi} = -\infty$ we are clearly done, hence assume $z(\Pi) < \bar{\varphi}$: there must be one $\bar{\boldsymbol{d}}$ such that $\varphi(\bar{\boldsymbol{y}} + \bar{\boldsymbol{d}}) < \bar{\varphi}$, for otherwise their inf would be $\ge \bar{\varphi}$. Then

$$\bar{\varphi} = \lim_{t\to\infty} z(\Pi_{\bar{y},t}) \le \lim_{t\to\infty} \varphi(\bar{\boldsymbol{y}} + \bar{\boldsymbol{d}}) + D_t(\bar{\boldsymbol{d}}) = \lim_{t\to\infty} \varphi(\bar{\boldsymbol{y}} + \bar{\boldsymbol{d}}) < \bar{\varphi}. \qquad ♦$$

With D_t satisfying P1 .. P5, we will be able to prove convergence of the algorithm: however, in some cases we will also need the following assumption about φ.

Assumption 3.7: φ *is* f-regular*, i.e.* $\forall \delta, \Delta$ *s.t.* $\delta \le \Delta$ *and* $S_\delta(\varphi) \ne \emptyset\ \exists K(\delta, \Delta) < +\infty$ *such that*

$$\sup_{\bar{y}} \{ \inf_{y} \{ \| \boldsymbol{y} - \bar{\boldsymbol{y}} \|_2 : \boldsymbol{y} \in S_\delta(\varphi) \} : \bar{\boldsymbol{y}} \in S_\Delta(\varphi) \} < K(\delta, \Delta).$$

In plain words, φ is *f-regular* if any two of its level sets are always at a finite distance; alternatively, one could say that φ never becomes "infinitely flat" (the "f" stands for "flatness"). We will see in §4 that f-regularity properly generalises *inf-compactness*, in the sense that all inf-compact functions are f-regular while the converse is not true: for instance, all polyhedral functions are f-regular.

As for the *t*-strategy and the NS/SS decision, they are required to obey to the following five rules:

3.i $t \ge \underline{t} > 0$ for a fixed $\underline{t}$;

3.ii if a SS is performed, then $\varphi_{\bar{y}}(\boldsymbol{d}^*) \le m_1 v^* = m_1 \varphi_{\beta(\bar{y})}(\boldsymbol{d}^*)$ for a fixed $m_1 \in (0,1)$: the converse is *not* required, i.e. even if a "considerable" improvement has been obtained there is the choice of not accepting the step, but for what is required by 3.iv below;

3.iii if an *infinite* sequence of *consecutive* NS is performed, t must be *finitely convergent*, i.e. there must be an h such that $t_i = t_h > 0$ for all $i \ge h$;

3.iv in the case of 3.iii, $\varphi_{\bar{y}}(\boldsymbol{d}_i^*) > m_1 v_i^* = m_1 \varphi_{\beta(\bar{y})}(\boldsymbol{d}_i^*)$ for all $i \ge h$, i.e. after finitely many NS we are obliged to accept *any* step that improves at least of $m_1 v_i^*$;

3.v in the case of 3.iii, at any iteration $i \ge h$ φ must always be evaluated at least in $\bar{\boldsymbol{y}}_i + \boldsymbol{d}_i^*$, and the

corresponding subgradient must be added to β.

Let us briefly discuss the above conditions: 3.i is needed to avoid that the stabilizing term becomes "infinitely heavy" as the algorithm proceeds, "imprisoning" the sequence of current points in a bounded region of the space. In practice, it is not difficult to choose a small $\underline{t}$ that does not interfere with the t-strategy; besides, $\underline{t}$ *is* ultimately fixed by the finite machine precision. The other conditions just describe the basic ideas of the algorithm: at each step, either performing a SS a "consistent improvement" is obtained, or new information enrich the Bundle. Moving to a better point is not mandatory if some alternative strategy - typically increasing t - appears to be preferable (3.ii), but this must not happen forever (3.iv). Analogously, conditions 3.iii and 3.v allow for adjusting t in response to a "bad" step, but prevent this to impair the overall convergence of the algorithm.

An important detail of any implementable Bundle method is the β-*strategy*: at each step one can

- remove some $(\alpha, \boldsymbol{g})$ from β and / or
- given some $\theta \in \Theta$ [usually θ^*], add $(\alpha', \boldsymbol{g}') = (\boldsymbol{G}_\beta\theta , \alpha_{\beta(\bar{y})}\theta)$ [$= (\sigma^*, z^*)$] to β.

These moves may, both in theory and in practice, impair finite termination or even convergence of the algorithm: hence, we will state the first results temporarily assuming the most conservative β-strategy, i.e. that the above operations are never performed.

The convergence proof will be divided into two parts: in the first, we will pretend that no Serious Step occur - i.e. examine an infinite sequence of consecutive Null Steps - so that in the second we will be allowed to disregard what happens between two successive Serious Steps, essentially reducing a run of the algorithm to its Serious Steps alone. We will show that a(n infinite) *sequence of consecutive NS* can be seen from the dual viewpoint as an *inner linearization approach* applied to the solution of $(\Delta_t^{\bar{y}})$, while from the primal viewpoint it is a *Cutting Plane approach* applied to the computation of $\phi_t(\bar{y})$ (the solution of $(\Pi_t^{\bar{y}})$). In the following theorems, the iteration index i will denote the i-th Null Step of the (only) infinite sequence of consecutive NS that the algorithm is supposed to perform, so that (Δ_i) and (Π_i) are the subproblems solved in that iteration, (σ_i, z_i) and $(v_i, \boldsymbol{d}_i)$ are their solutions, φ_i is the corresponding CP model and so on.

Lemma 3.8: *after some iteration h, all the* (σ_i, z_i) $i \geq h$ *belong to a compact set.*

Proof: the h is the one in 3.iii, such that $t_i = t_h = t$ for $i \geq h$: since clearly $z(\Delta_i) \leq z(\Delta_h)$, one has

$$D_t^*(-z_i) \leq \sigma_i + D_t^*(-z_i) = z(\Delta_i) \leq z(\Delta_h) \text{ for } i \geq h$$

but the level sets of D_t^* are compact from P2. ♦

Lemma 3.9: *all the* $\boldsymbol{d}_i$ *belong to a compact set.*

Proof: the result has already been established by Lemma 3.3 if Assumption 3.2 holds; otherwise, P3 holds and therefore D_t^* is finite everywhere. By Lemma 2.4 $\boldsymbol{d}_i \in \partial D_t^*(-z_i)$, and by Lemma 3.8 all the z_i $i \geq h$ belong to a compact set (that can always be finitely enlarged to comprise the first $h - 1$), but the image of a compact set under the subdifferential mapping of a function that is finite everywhere is compact [HL93a - Remark VI.6.2.3]. ♦

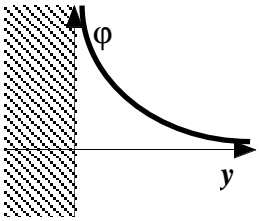
Theorem 3.10: *all the* $(\alpha_i, \boldsymbol{g}_i) \in Epi\ \varphi_{\bar{y}}^*$ *returned by the evaluation of* $\varphi_{\bar{y}}(\boldsymbol{d}_i)$ *belong to a compact set.*

Proof: we know from Lemma 3.9 that all the $\{ \boldsymbol{d}_i \}$ belong to a compact set: since $\varphi_{\bar{y}}$ is finite everywhere and $\boldsymbol{g}_i \in \partial\varphi_{\bar{y}}(\boldsymbol{d}_i)$, we can invoke again [HL93a - Remark VI.6.2.3] to conclude that all the $\boldsymbol{g}_i$ belong to a compact set. Also, $\varphi_{\bar{y}}$ is upper bounded over any compact set (an easy consequence of [HL93a - Theorem VI.3.1.2]), hence

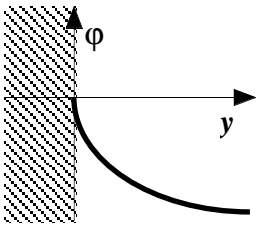
$$\varphi_{\bar{y}}(\boldsymbol{d}_i) = \boldsymbol{g}_i\boldsymbol{d}_i - \alpha_i \le \Delta < +\infty$$

for some Δ, which implies the thesis since both $\{ \boldsymbol{d}_i \}$ and $\{ \boldsymbol{g}_i \}$ "live" in a compact. ♦

Theorem 3.10 is the main point where the assumption "φ is finite everywhere" enters into play: it is actually a Lipschitz requirement that is needed, ruling out situations like



$$\varphi(y) = \begin{cases} 1/y & \text{if } y > 0 \\ +\infty & \text{otherw.} \end{cases}$$



$$\varphi(y) = \begin{cases} -\sqrt{y} & \text{if } y > 0 \\ +\infty & \text{otherw.} \end{cases}$$

where the set of all possible subgradients relative to the compact set $[0, \varepsilon]$ ($\not\subset$ *ri Dom* φ) is *not* compact. Let us mention here that this does not mean that Bundle methods cannot be used for *constrained* minimization: more on this subject will be said in §5.
Now, the main result is at hand: an infinite sequence of NS actually *solves* $(\Delta_t^{\bar{y}})$.

Theorem 3.11: $\{ (\sigma_i, z_i) \} \to (\sigma^*, z^*)$ *optimal for* $(\Delta_t^{\bar{y}})$, *and*

$\{ \boldsymbol{d}_i \} \to \boldsymbol{d}^*$ *optimal for* $(\Pi_t^{\bar{y}})$.

Proof: obviously, the t in the statement is the one "fixed" after iteration h. From Lemma 2.5

$$(1, -\boldsymbol{d}_i)[\ (\alpha, \boldsymbol{g}) - (\sigma_i, z_i)\ ] \ge 0 \qquad \forall (\alpha, \boldsymbol{g}) \in Epi\ \varphi_i^*.$$

Since β_i is monotonically increasing, $(\alpha_k, \boldsymbol{g}_k) \in Epi\ \varphi_i^*$ for $i > k \ge h$ and therefore

$$(1, -\boldsymbol{d}_i)(\alpha_k, \boldsymbol{g}_k) \ge (1, -\boldsymbol{d}_i)(\sigma_i, z_i) \qquad \forall\ i > k.$$

Now, use Lemma 2.6 and $\varphi_{\bar{y}}(\boldsymbol{d}_i) = \boldsymbol{g}_i\boldsymbol{d}_i - \alpha_i$ to write

$$(1, -\boldsymbol{d}_i)[\ (\alpha_i, \boldsymbol{g}_i) - (\sigma_i, z_i)\ ] + z(\Delta_i) \le z(\Delta_t^{\bar{y}}) \le z(\Delta_i)$$

and substitute $(\alpha_k, \boldsymbol{g}_k)$ for (σ_i, z_i): taking the sup over all $i > k$ we obtain

$$\sup_{i>k} \{ (1, -\boldsymbol{d}_i)[\ (\alpha_i, \boldsymbol{g}_i) - (\alpha_k, \boldsymbol{g}_k)\ ] \} + \sup_{i>k} \{ z(\Delta_i) \} \le z(\Delta_t^{\bar{y}}) \le \sup_{i>k} \{ z(\Delta_i) \}.$$

Sending $k \to \infty$ the leftmost sup goes to 0, since $\{ \boldsymbol{d}_i \}$ are bounded (Lemma 3.9) and $(\alpha_i, \boldsymbol{g}_i)$ belong to a compact set (Theorem 3.10): hence, $\{ z(\Delta_i) \} \to z(\Delta_t^{\bar{y}})$ (and $\{ z(\Pi_i) \} \to z(\Pi_t^{\bar{y}})$ by duality) and the convergence of the optimal solutions follows by compactness (Lemma 3.8-3.9). ♦

Note that the "critical" term $(1, -\boldsymbol{d}_i)[\ (\alpha_i, \boldsymbol{g}_i) - (\sigma_i, z_i)\ ]$ in the above proof is the opposite of $\Delta\varphi_i = \varphi_{\bar{y}}(\boldsymbol{d}_i) - v_i$, i.e. the gap between the estimated and the real value of φ in $\bar{\boldsymbol{y}} + \boldsymbol{d}_i$: hence, $\{ \Delta\varphi_i \} \to 0$.

Lemma 3.12: $\{ \boldsymbol{d}_i \} \to \boldsymbol{d}^*$ *such that* $\varphi_{\bar{y}}(\boldsymbol{d}^*) = 0$*: hence,* $\mathbf{0}$ *is also optimal for* $(\Pi_t^{\bar{y}})$.

Proof: from 3.iv, $\Delta\varphi_i > -(1 - m_1)v_i$ for all $i \geq h$. Now, assume by contradiction that $0 > \delta \geq \varphi_{\bar{y}}(\boldsymbol{d}_i) \geq v_i$ for all i: one would get $\Delta\varphi_i > -(1 - m_1)\delta > 0$, while $\{ \Delta\varphi_i \} \to 0$. Since $\varphi_{\bar{y}}(\mathbf{0}) = \varphi_{\bar{y}}(\boldsymbol{d}^*)$ and $D_t(\boldsymbol{d}^*) \geq D_t(\mathbf{0})$, then $\mathbf{0}$ is also optimal. ♦

Corollary 3.13: $\{ \sigma_i \} \to 0$ *and* $\{ D_t^*(-z_i) \} \to 0$.

Proof: by Lemma 3.12, $z(\Pi_t^{\bar{y}}) = z(\Delta_t^{\bar{y}}) = 0$: the results follows by nonnegativity of σ_i and D_t^*. ♦

Corollary 3.14: *if* $\varepsilon > 0$, *then only* finitely many consecutive *NS can be performed*.

The proof of Theorem 3.11 is essentially that of [HL93b - Theorem XII.4.2.3] for the Cutting Plane method and requires that all the information computed during the course of the algorithm is kept until termination. For the "standard" Bundle method with $D_t^* = {}^1\!/_2 t\|\cdot\|_2^2$, this assumption can be released: in fact, convergence can be proved under the "minimal" conceivable form of β-strategy, where at each step the Bundle is replaced with $\{ (\sigma_i, z_i) \}$. Interestingly enough, most of the arguments for this proof (but the crucial one) can be carried along in a general form, as follows.
Assume that D_t^* is actually *differentiable* and consider the function

$$h(\gamma) = f_t(\sigma_i + \gamma\omega, z_i + \gamma \boldsymbol{w}) \qquad \text{where } (\omega, \boldsymbol{w}) = (\alpha_i, \boldsymbol{g}_i) - (\sigma_i, z_i):$$

$\inf_{\gamma \in [0, 1]} h(\gamma) = z(\Delta_{\beta,t}^{\bar{y}})$ when $\beta = \{ (\alpha_i, \boldsymbol{g}_i), (\sigma_i, z_i) \}$. From Lemma 2.5, $(1, -\boldsymbol{d}_i) = \nabla f_t(\sigma_i, z_i)$, hence

$$h'(0) = (1, -\boldsymbol{d}_i)[(\alpha_i, \boldsymbol{g}_i) - (\sigma_i, z_i)] = -\Delta\varphi_i = v_i - \varphi_{\bar{y}}(\boldsymbol{d}_i) \leq 0.$$

From 3.iv, $-\Delta\varphi_i < (1 - m_1)v_i = -(1 - m_1)[(1, -\boldsymbol{d}_i)(\sigma_i, z_i)]$, furthermore

$$f_t(0, \mathbf{0}) - f_t(\sigma_i, z_i) \geq \nabla f_t(\sigma_i, z_i)[(0, \mathbf{0}) - (\sigma_i, z_i)] \Rightarrow f_t(\sigma_i, z_i) \leq (1, -\boldsymbol{d}_i)(\sigma_i, z_i)$$

hence $\quad h'(0) < -(1 - m_1) f_t(\sigma_i, z_i) = -(1 - m_1)h(0)$.

In plain words, the derivative of h at zero is at least as negative as $-(1 - m_1)h(0)$: being able to bound the *obtained* improvement in terms of the *estimated* improvement $h'(0)$, it would be possible to estimate the speed of convergence of h to 0. This can actually be done when $D_t^* = {}^1\!/_2 t\|\cdot\|_2^2$, as

$$\inf_{\gamma \in [0, 1]} h(\gamma) - h(0) \leq -\lambda h'(0)^2$$

[Fr97 - Theorem I.2.2.2] for some constant $\lambda > 0$: hence, the "standard" Bundle method suffers of a sublinear rate of convergence during the sequences of NS. Within our general framework, it may be possible to find other stabilizing terms that give better convergence properties: for instance, would the obtained improvement be a constant fraction of $h'(0)$, linear convergence would be obtained. This clearly motivates the following definition.

Definition 3.15: *a* β*-strategy is* monotone *if for each* $i \geq h$ *either all the* $\boldsymbol{g}_j$ *such that* $\theta_j > 0$ *are kept in* β*, or* (σ_i, z_i) *is added to* β.

For $D_t^* = {}^1\!/_2 t\|\cdot\|_2^2$, a *monotone* β-strategy is enough to guarantee that $\{ z(\Delta_i) \} \to 0$, i.e. $z(\Delta_t^{\bar{y}}) = 0$,

which implies back Theorem 3.11: this is likely to happen for many other differentiable D_t^* (strictly convex D_t [HL93b - Theorem X.4.1.3]). Unfortunately, a *monotone* β-strategy is not enough to guarantee convergence in a nondifferentiable case [Fr97 - §I.4.2]: however, this motivates a further definition that will not give any finite bound on the dimension of β, but can help to design practical β-strategies.

Definition 3.16: *a* β*-strategy is* strictly monotone *if it is monotone and, if some* (σ, z) *has been removed from* β *at iteration k, no other removal is permitted unless* $v_k - \mu \le v_i$ $(\mu > 0,\ k < i)$.

A *strictly monotone* β-strategy ensures $\{ v_i \} \to 0 \Rightarrow \{ z(\Delta_i) \} \to 0$ ($-v_i \ge z(\Delta_i)$ by duality), preventing removals when $v_i \ge -\mu$ (for a large enough i) so that Theorem 3.11 can be used: hence, we assume from now on that a (strictly, if necessary) monotone β-strategy is used. Now we disregard everything that happens between two successive SS, so that v_i, t_i ... rather denote the values of v, t ... at the i-th Serious Step; a "weak" finite termination proof is at hand.

Theorem 3.17: *if* infinitely many *(Serious) steps are done and* $\varepsilon > 0$*, then* $\{ \varphi(\bar{\boldsymbol{y}}_i) \} \to -\infty$.

Proof: the above theorems rule out the possibility on an infinite sequence of consecutive NS: if infinitely many steps are done, the Serious Steps must be an infinite number. At the i-th SS an upper bound on v_i is available, since

$$\sigma_i + \mu D_{t_i}^*(-z_i) > \varepsilon \quad \Rightarrow \quad \sigma_i/\mu + D_{t_i}^*(-z_i) > \varepsilon/\mu = \varepsilon' \quad \Rightarrow \quad \sigma_i + D_{t_i}^*(-z_i) > \varepsilon'$$

but $\quad \sigma_i + D_{t_i}^*(-z_i) = z(\Delta_{\beta_i,t_i}^{\bar{y}_i}) = -z(\Pi_{\beta_i,t_i}^{\bar{y}_i}) = -v_i - D_{t_i}(\boldsymbol{d}_i) \le -v_i \quad \Rightarrow \quad v_i < -\varepsilon'.$

From 3.ii we know that $\Delta\varphi_i \le m_1 v_i \Rightarrow \varphi(\bar{\boldsymbol{y}}_i) \le \varphi(\boldsymbol{y}_0) + m_1(\sum_{j<i} v_j) \le \varphi(\boldsymbol{y}_0) - i\varepsilon'$: when i goes to infinity, $\{ \varphi(\bar{\boldsymbol{y}}_i) \} \to -\infty$. ♦

Immediate consequences of Theorem 3.17 are:

- if $\varepsilon > 0$ and φ is bounded, then the algorithm finitely terminates;
- if $\varepsilon > 0$ and the algorithm never terminates, then φ is unbounded below.

The natural question is then: is it possible that the algorithm terminates but still φ is unbounded below? In other words: what does the stopping condition exactly say about the last $\bar{\boldsymbol{y}}$? One should not expect too much: 3.i has not been used yet, hence nothing prevents t do converge to 0 quite fast during sequences of SS (even if t is lower bounded during infinite sequences of NS by 3.iii). Thinking to Ex. 3.3 where t is the radius of a trust region around $\bar{\boldsymbol{y}}$, it is clear that $\{ \bar{\boldsymbol{y}}_i \}$ may be forced to converge to a nonoptimal point if t converges to 0 fast enough. In fact, to prove convergence one needs t to be "large": we supplement Observation 3.6 showing that, if φ is f-regular, convergence can be proved even if $\bar{\boldsymbol{y}}$ is allowed to change in any "reasonable" way for a descent method. Note that f-regularity is not needed for "weak" convergence (Theorem 3.17).

Lemma 3.18: *assume that* φ *is f-regular, choose two sequences* $\{ t_i \} \to +\infty$ *and* $\{ \bar{\boldsymbol{y}}_i \}$ *such that* $\varphi(\bar{\boldsymbol{y}}_i) \ge \varphi(\bar{\boldsymbol{y}}_{i+1})$ *for all i: then,* $\lim_{i\to\infty} z(\Pi_{\bar{\boldsymbol{y}}_i,t_i}) = z(\Pi)$.

Proof: first of all, note that we are dealing with the "exact" problem $(\Pi_{\bar{\boldsymbol{y}}_i,t_i})$. As in Observation 3.6,

assume by contradiction that $z(\Pi) < \bar{\varphi} = \lim_{t\to\infty} z(\Pi_{\bar{y}_i,t_i})$: there must be one y^* such that $\varphi(y^*) = \delta < \bar{\varphi}$, i.e. $S_\delta(\varphi) \neq \emptyset$. Call $\Delta = \varphi(\bar{y}_0) \geq \bar{\varphi} > \delta$: from f-regularity, $\exists K(\delta, \Delta) < +\infty$ such that

$$\sup_{\bar{y}} \left\{ \inf_y \left\{ \left\| y - \bar{y} \right\|_2 : y \in S_\delta(\varphi) \right\} : \bar{y} \in S_\Delta(\varphi) \right\} < K(\delta, \Delta).$$

In other words, for each $\bar{y}_i$ there exist one $\hat{y}_i \in S_\delta(\varphi)$ such that $\left\| \hat{y}_i - \bar{y}_i \right\|_2 < K(\delta, \Delta)$: it must be

$$\varphi(\hat{y}_i) + D_{t_i}(\hat{y}_i - \bar{y}_i) \geq z(\Pi_{\bar{y}_i,t_i}) \geq \bar{\varphi}.$$

Now, call $\varepsilon = (\bar{\varphi} - \delta) / 2$: by Observation 3.5, $\exists \bar{t}$ such that $S_\varepsilon(D_t) \supseteq B_2(K(\delta, \Delta)) \ \forall t \geq \bar{t}$. But since $\{ t_i \} \to +\infty$, $t_i \geq \bar{t}$ from some index i on: hence

$$\bar{\varphi} \leq \varphi(\hat{y}_i) + D_{t_i}(\hat{y}_i - \bar{y}_i) \leq \delta + (\bar{\varphi} - \delta) / 2 = (\bar{\varphi} + \delta) / 2 < \bar{\varphi}. \qquad \blacklozenge$$

In the above proof, it is not actually needed that $\{ t_i \} \to +\infty$: it is enough that t is "large" ($\geq \bar{t}$). Obviously, the problem is that the threshold value $\bar{t}$ is not known in advance: however, with the above Lemma we are ready to prove convergence in the following sense.

Theorem 3.19: *assume that* φ *is f-regular, choose two sequences* $\{ \underline{t}_i \} \to +\infty$ *and* $\{ \varepsilon_i \} \to 0$ *and call* $\bar{y}_i$ *the current point when the Bundle algorithm ran with* $\underline{t} = \underline{t}_i$ *and* $\varepsilon = \varepsilon_i$ *stops: then,* $\lim_{i\to\infty} \varphi(\bar{y}_i) = z(\Pi)$.

Proof: call t_i ($\geq \underline{t}_i$) the value of t when the code stops: clearly, $\{ t_i \} \to +\infty$. Hence

$$-z(\Delta^{\bar{y}_i}_{t_i}) + \varphi(\bar{y}_i) = z(\Pi_{\bar{y}_i,t_i}) \Rightarrow \varphi(\bar{y}_i) \leq z(\Pi_{\bar{y}_i,t_i}) + \varepsilon_i$$

since $z(\Delta^{\bar{y}_i}_{t_i}) \leq z(\Delta^{\bar{y}_i}_{\beta_i,t_i}) \leq \varepsilon_i$. Take the limit for $i \to \infty$ and apply Corollary 3.18. $\blacklozenge$

The above theorem suggests the following implementation: start with any (ε, $\underline{t}$) and run the Bundle algorithm until (finite) convergence, then decrease ε, increase $\underline{t}$ and restart. As long as ε and $\underline{t}$ are driven respectively to 0 and $+\infty$ simultaneously, $\varphi(\bar{y}_i)$ will converge to $z(\Pi)$. It is not possible to directly set $\varepsilon = 0$ since the algorithm may not finitely stop: this could be done if we would be allowed to increase $\underline{t}$ to $+\infty$ during the run, but 3.iii requires that t is not changed (too many times) during sequences of NS. In other words, we need to be sure that $(\Delta_{\bar{y},t})$ is solved with "enough precision" before to increase t, and allowing t to "settle down" during sequences of NS is crucial. This version can be seen as a "triply iterative" approach to (Δ) where a sequence of (Δ_t^*) with (globally) increasing t are approximately solved: the approximation is driven to 0 while t is driven to $+\infty$. This can also suggest several practical t-strategies [CFN96] [Fr97 - §I.5.3].

Of course, the remarks to Lemma 3.18 also apply to Theorem 3.19: it is not actually needed that $\underline{t}$ goes all the way up to $+\infty$, but one should ensure that it is larger than some (unknown) value $\bar{t}$. In practice, it is not difficult to find a value that guarantees the required precision: hence, the "third level" of the algorithm is usually not needed. When D_t is a real penalty term, e.g. in Ex. 3.1 and Ex. 3.2 above (but not in Ex. 3.3), part of the job can be done by the extra stopping parameter $\mu \geq 1$: setting a large μ can make the stopping condition be that of a larger t than the current one. For instance, for $D_t^* = t\left\|\cdot\right\|_p$ the stopping conditions are equivalent to those with $t^* = t\mu \geq t$: in our experience [CFN96] [Fr97], guessing a value of μ that produces an ε-optimal solution is fairly easy.

Stronger results can be proved for particular choices of D_t: for the "standard" case of Ex. 3.1, for

instance, convergence can be obtained without the f-regularity assumption and even if t converges to zero "not too fast" [HL93b - Theorem XV.3.2.2]. This is unlikely to hold in general: essentially, assumptions would be needed about the "rate of growth" of the level sets of D_t in t.

Yet, stronger results can be obtained with a mild extra assumption about D_t.

P6) $\partial D_t(\mathbf{0}) = \{\mathbf{0}\}$ (D_t is *differentiable* in $\mathbf{0}$).

P*6) $S_0(D_t^*) = \{\mathbf{0}\}$ (D_t^* has *unique minimizer* in $\mathbf{0}$).

In view of Observation 3.1, P6 and P*6 are equivalent: with this property, we can show asymptotic convergence (still under Assumption 3.7) with just any fixed $\underline{t} > 0$ and directly setting $\varepsilon = 0$.
To do that, let us describe the "dual" interpretation of a SS: remind from §2 that, at least if t is fixed, the whole process can be seen as solving (Δ_t^*) with a dual ascent approach. At each step, we know that $z(\Delta_{\beta,\bar{y},t}) \geq z(\Delta_{\bar{y},t}) \geq z(P_{\bar{y}})$ and we compute $z(P_{\bar{y}+d^*})$ that gives a lower bound on $z(\Delta_{\bar{y}+d^*,t})$. Now assume that $z(P_{\bar{y}+d^*}) > z(P_{\bar{y}})$, i.e. $\varphi(\bar{y}+d^*) < \varphi(\bar{y})$:

- either we believe that $z(\Delta_{\bar{y}+d^*,t}) > z(\Delta_{\bar{y},t})$ as well, hence we shall move the current point (SS): we can be sure about it only if $z(P_{\bar{y}+d^*}) \geq z(\Delta_{\beta,\bar{y},t})$;
- or we don't believe it, because $z(\Delta_{\beta,\bar{y},t}) >> z(P_{\bar{y}+d^*})$: the current $z(P_{\bar{y}})$ may be a "bad" estimate of the real $z(\Delta_{\bar{y},t})$, and we are better to improve our knowledge before moving (NS).

The SS condition in 3.ii, $\varphi(\bar{y}+d^*) - \varphi(\bar{y}) \leq m_1 v^*$, can be seen (using Lemma 2.3 and $-v^* \geq z(\Delta_{\beta,\bar{y},t})$) to be equivalent to

$$z(P_{\bar{y}+d^*}) \geq m_1 z(\Delta_{\beta,\bar{y},t}) + (1-m_1)\, z(P_{\bar{y}}) \qquad \text{or}$$

$$z(P_{\bar{y}+d^*}) \geq m_1 \delta + z(P_{\bar{y}}) \qquad \text{where } \delta = z(\Delta_{\beta,\bar{y},t}) - z(P_{\bar{y}}).$$

That is, the current point is moved if the increase in the lower bound is at least a constant fraction of the maximum possible increase $\delta = z(\Delta_{\beta,\bar{y},t}) - z(P_{\bar{y}})$: $m_1 = 1$ is the conservative choice of moving only when $z(\Delta_{\bar{y}+d^*,t}) \geq z(\Delta_{\bar{y},t})$. We will exploit this interpretation in the proof of the next theorem.

Theorem 3.20: *under P6,* $\{\varphi(\bar{y}_i)\} \to z(\Pi)$ *in the Bundle algorithm ran with* $\varepsilon = 0$.

Proof: if an infinite sequence of consecutive NS is performed, we know from Lemma 3.12 that $\mathbf{0}$ is optimal for $(\Pi_t^{\bar{y}})$, hence $\mathbf{0} \in \partial[\varphi(\bar{y}+\cdot) + D_t(\cdot)](\mathbf{0}) \Rightarrow \mathbf{0} \in \partial\varphi(\bar{y})$ by P6: since $\varepsilon = 0$, this is also true if the algorithm finitely terminates. If an infinite number of SS is performed, $\{\varphi(\bar{y}_i)\}$ (a nondecreasing sequence) converges to some $\bar{\varphi}$: if $\bar{\varphi} = -\infty$ we are done, hence the only case to take care of is $\bar{\varphi} > -\infty$. As usual, let us restrict our attention (hence the indices i) to SS only: from the dual restatement of the SS condition

$$z(P_{\bar{y}_{i+1}}) \geq m_1\delta_i + z(P_{\bar{y}_i}) \Rightarrow \Sigma_{i=0..\infty}\, \delta_i \leq +\infty \Rightarrow \{\delta_i\} \to 0.$$

But $$\delta_i = z(\Delta_{\beta_i,\bar{y}_i,t_i}) - z(P_{\bar{y}_i}) = (\sigma_i - \bar{y}_i z_i + D_{t_i}^*(-z_i)) - (\alpha_i - \bar{y}_i g_i) \geq D_{t_i}^*(-z_i)$$

since (α_i, g_i) is the inf over *Epi* φ^*: since $t_i \geq \underline{t}$, $\{D_{t_i}^*(-z_i)\} \to 0 \Rightarrow \{z_i\} \to \mathbf{0}$ due to P*6.
Now, assume by contradiction that one y^* exists such that $\varphi(y^*) = \delta < \bar{\varphi}$, call $\Delta = \varphi(\bar{y}_0)$ and use f-regularity of φ to show that for each $\bar{y}_i$ there exist one $\hat{y}_i$ such that $\|\hat{y}_i - \bar{y}_i\|_2 < K(\delta, \Delta)$ and $\varphi(\hat{y}_i) \leq \delta$. Let us call $\sigma_i^{\bar{y}_i} = \sigma_i - \bar{y}_i z_i + \varphi(\bar{y}_i) \geq 0$, and note that $\delta_i = \sigma_i^{\bar{y}_i} + D_{t_i}^*(-z_i)$ (so that $\{\sigma_i^{\bar{y}_i}\} \to 0$ as

well): z_i is a $\sigma_i^{\bar{y}}$-subgradient of φ in $\bar{y}_i$, hence

$$\bar{\varphi} > \delta \geq \varphi(\hat{y}_i) \geq \varphi(\bar{y}_i) + z_i(\hat{y}_i - \bar{y}_i) - \sigma_i^{\bar{y}}$$

$$\Rightarrow \quad 0 > \delta - \bar{\varphi} \geq -\| z_i \|_2 K(\delta, \Delta) - \sigma_i^{\bar{y}}$$

that yields the desired contradiction. ♦

Still under the "smoothness" assumption P6, we can also prove finite convergence to a "true" optimal solution - not just optimal within some $\varepsilon > 0$ - provided that φ is polyhedral ($\Rightarrow$ f-regular, see §4) and that the subgradient aggregations are properly limited.

Definition 3.21: *a* β*-strategy is* safe *if*

- *the total number of aggregations (adding* (z_i, σ_i) *to* β*) is finite and*
- *if some* (σ, z) *has been removed from* β *after a SS at iteration k, no other removal is permitted unless* $\varphi_{\beta_k}(\bar{y}_k) < \varphi_{\beta_i}(\bar{y}_i)$ *(*$k < i$*).*

This definition is made by two parts, whose role will be made clear by the proof of the following theorems: for now, note how the second part is equivalent to the definition of a strictly monotone β-strategy applied to the SS rather than to the NS (but for the fact that no guarantee on the decrease has to be given) since $\varphi_{\beta_k}(\bar{y}_k) - \varphi_{\beta_i}(\bar{y}_i) = v_k - v_i$ if $\bar{y}$ has not changed in the meantime.

Observation 3.22: *let* Γ *be a finite set of vectors: there exist a* $\mu > 0$ *depending only on* Γ *such that, for any* $\Sigma \subset \Gamma$, $\exists z \in Conv(\Sigma)$ *such that* $D_t^*(-z) < \mu \Rightarrow \mathbf{0} \in Conv(\Sigma)$.

Proof: just denote $\mu(\Sigma) = \min_z \{ D_t^*(-z) : z \in Conv(\Sigma) \}$ and let

$$\mu = \min_\Sigma \{ \mu(\Sigma) : \mu(\Sigma) > 0 , \Sigma \subseteq \Gamma \}:$$

since 2^Γ is a finite set, $\mu > 0$ and $\mu(\Sigma) = 0$ for any Σ such that $D_t^*(-z) < \mu$ for some $z \in Conv(\Sigma)$. The fact that $D_t^*(-z) = 0 \Leftrightarrow z = \mathbf{0}$ does the rest. ♦

Applying Observation 3.22 to *Epi* φ^*, one shows that there exist a μ *independent from* β (dependent only on φ) such that whenever a $(\sigma, z) \in Epi\, \varphi_\beta^*$ is found s.t. $D_t^*(-z) < \mu$, φ_β is *bounded*. Actually, one should rather apply Observation 3.22 to φ "plus" *all* the aggregated subgradients, but due to the first part of the definition of a safe β-strategy this is still a polyhedral function.

Theorem 3.23: *let* φ *be a polyhedral function: if a safe* β*-strategy is used, the algorithm ran with* $\varepsilon = 0$ *will finitely converge to an optimal solution* $\boldsymbol{y}^*$ *if* φ *is bounded.*

Proof: if φ is not bounded then the algorithm never stops, and if the algorithm stops we are done: hence, suppose that $\bar{\varphi} > -\infty$ but still the algorithm runs forever. From the proof of Theorem 3.20, $\{ D_{t_i}^*(-z_i) \} \to 0$: hence, φ_{β_i} is *always* bounded from some step h on. At length, *every* $\boldsymbol{d}_i$ *is a minimizer of the corresponding* φ_{β_i}: in fact, calling $\beta_i^+ = \{ (\sigma, z) \in \beta_i : \theta_{(\sigma, z)} > 0 \}$, we have from Complementary Slackness (Cfr. Lemma 2.3) that *every* z such that $(\sigma, z) \in \beta_i^+$ is a 0-*subgradient of* φ_{β_i} (although not necessarily of φ) in $\boldsymbol{d}_i$. Hence, z_i is a convex combination of the vectors in β_i^+ whose norm becomes arbitrarily small, showing that (at length) $\mathbf{0} \in \partial\varphi_{\beta_i}(\boldsymbol{d}_i)$. Now, the (finite) minimizer $\bar{\varphi}$ of φ is the maximum in the set of the minimizers $\bar{\varphi}_\beta$ of φ_β for all possible β: but since we are using a

safe β-strategy this is a *finite* set. From the above theorems we know that $\{ v_i \} \to 0$, i.e. that $\bar{\varphi}_{\beta_i} - \bar{\varphi}(\bar{\boldsymbol{y}}) \geq \bar{\varphi}_{\beta_i} - \bar{\varphi}$ will become smaller than any fixed threshold: we can conclude that $\{ \bar{\varphi}_{\beta_i} \}$ *finitely converges* to $\bar{\varphi}$. Unfortunately, unless the optimal solution of φ is unique, this does not imply anything about the actual value of $\varphi(\bar{\boldsymbol{y}})$: however, the second part of Definition 3.21 ensures that, after a finite number of steps, a *complete* description of the optimal level set of φ is available. ♦

The second part of Definition 3.21 rules out any possibility of keeping the size of β bounded: eventually, in the last steps we can do nothing better than accumulating all the subgradients. Alternatively, one could make assumptions that guarantee uniqueness of the optimal solution, such as the *Haar constraint qualification* [Ki85]. Of course, this is potentially relevant only if D_t is smooth: for nonsmooth stabilizing terms, infinite memory is already needed for convergence of the NS sequences.

Finally, another possible approach to convergence is suggested by contrasting Observation 3.6 with Lemma 3.18: in the former - where no SS are done - convergence is proved without the f-regularity assumption, while in the latter - where SS are allowed - it is required. Hence, inhibiting SS may allow to eliminate the f-regularity assumption: in this case, t may in principle need to go all the way up to $+\infty$. Such a variant of the Bundle method would require the following additional rule:

3.vi only *finitely many* SS are done: after the last one, t is fixed, rule 3.iv is inhibited and the stopping condition becomes $\varphi_{\bar{\boldsymbol{y}}}(\boldsymbol{d}^*) - v^* \geq -\varepsilon$.

This rule is rather abstract, but several practical implementations can be imagined: the obvious one is that SS are *never* performed, but alternatively one can choose some $\underline{\varepsilon} > 0$ and inhibit SS if $v^* \geq -\underline{\varepsilon}$ (the optimal solution is "near"), as long as the t-strategy is managed accordingly.

Theorem 3.24: *with rule 3.vi in force, choose two sequences* $\{ \underline{t}_i \} \to +\infty$ *and* $\{ \varepsilon_i \} \to 0$ *and call* $\boldsymbol{d}_i$ *the optimal solution of* $(\Pi^{\bar{\boldsymbol{y}}_i}_{\beta_i,t_i})$ *when the Bundle algorithm ran with* $\underline{t} = \underline{t}_i$ *and* $\varepsilon = \varepsilon_i$ *stops: then,* $\lim_{i\to\infty} \varphi(\bar{\boldsymbol{y}}_i + \boldsymbol{d}_i) = z(\Pi)$ *even if* φ *is* not *f-regular.*

Proof: if $t_i < \underline{t}_{i+1}$, when the algorithm is restarted t is increased to satisfy 3.i: hence, $\{ t_i \} \to +\infty$. Wait for the last SS to be performed: since Theorem 3.11 is not harmed by inhibiting SS (rule 3.iv is *not* used) it is still true that $\{ \Delta\varphi_k = \varphi_{\bar{\boldsymbol{y}}}(\boldsymbol{d}_k) - v_k = (1, -\boldsymbol{d}_k)[\, (\alpha_k, \boldsymbol{g}_k) - (\sigma_k, \boldsymbol{z}_k) \,] \} \to 0$, hence for any fixed ε_i the new stopping condition will be finitely attained. From Lemma 2.6

$$-\varepsilon_i + z(\Delta_{\beta_i,t_i}) \leq \varphi(\boldsymbol{d}_i) - v_i + z(\Delta_{\beta_i,t_i}) \leq z(\Delta_{t_i}) \leq z(\Delta_{\beta_i,t_i})$$

(we have dropped reference to $\bar{\boldsymbol{y}}$ that is fixed), hence by taking the limit for $i \to \infty$

$$\lim_{i\to\infty} -\varepsilon_i + z(\Delta_{\beta_i,t_i}) = \lim_{i\to\infty} z(\Delta_{t_i}) = z(\Delta)$$

(we have used Observation 3.6) which clearly implies the result. ♦

Provided that $z(\Pi) > -\infty$, with one extra assumption on D_t it is possible to guarantee that t need not go all the way up to $+\infty$.

P7) $\forall \delta > 0\ \exists t(\delta)$ such that $S_0(D_{t(\delta)}) = \partial D^*_{t(\delta)}(\mathbf{0}) \supseteq B_2(\delta)$.

P7 is a strengthened form of Observation 3.5 (which in turn derived from P5), and it is easy to see that P7 ⇒ P6: such a D_t must be "like" the characteristic function of a set growing with t, see e.g. Ex.

3.3. Under P7, there exist a large enough t such that $z(\Delta_{\bar{y},t}) = z(\Delta_t) = z(\Delta)$: all these three problems has the same feasible set $Epi\ \varphi^*$, and $(\Delta_{\bar{y},t})$

$$\inf_{\sigma,z} \{ \sigma - \bar{y}z + D_t^*(-z) : (\sigma, z) \in Epi\ \varphi^* \}$$

can be regarded as

$$\inf_{\sigma,z} \{ \sigma - \bar{y}z : (\sigma, z) \in Epi\ \varphi^* , z = \mathbf{0} \}$$

(that is clearly equivalent to (Δ)) where the constraints $z = \mathbf{0}$ has been replaced by the penalty function $D_t^*(-z)$. Since $\partial D_t^*(\mathbf{0})$ grows with t, there exist a large enough t such that $\bar{y} \in \partial D_t^*(\mathbf{0})$: by the boundedness assumption, there must exist one $(\sigma, \mathbf{0}) \in Epi\ \varphi^*$ with minimal σ, which is therefore optimal. This is a standard result: nondifferentiability is the crucial property for D_t^* to be an *exact* penalty function, a smooth D_t^* would eventually need $t \to \infty$ [HL93a - Remark VII.1.2.4].

4. f-regular functions

In this paragraph we will address the following question: how restrictive is the f-regularity assumption required for some of the proofs in §3? The answer will be "not much". To begin our study, let us state a general sufficient condition for the property to hold.
For a convex set C, the *asymptotic cone* of C is the set

$$C_\infty = \{ \boldsymbol{d} : \boldsymbol{d} + \alpha \boldsymbol{x} \in C\ \forall \alpha \geq 0 \};$$

a non-empty closed C is compact $\Leftrightarrow C_\infty = \{ \mathbf{0} \}$ [HL93a - Proposition III.2.2.3]. For a closed convex function φ, all its nonempty level sets have the same asymptotic cone [HL93a - Proposition IV.3.2.5]: we shall denote that cone with φ_∞. A sufficient condition for a function to be f-regular is that each level set can be surrounded by φ_∞ plus a compact set.

Property 4.1: $\forall \delta$ *there exist a* compact *set* C_δ *such that* $S_\delta(\varphi) \subseteq C_\delta + \varphi_\infty$.

Theorem 4.2: *Property 4.1* $\Rightarrow$ φ *is f-regular.*

Proof: note that the case $S_\delta(\varphi) = \emptyset$ is not a problem, since it is covered by $C_\delta = \emptyset$; also, note that C_δ is *not* required to be convex. Select δ and Δ such that $\delta \leq \Delta$ and $S_\delta(\varphi) \neq \emptyset$, and any $\bar{y} \in S_\Delta(\varphi)$: from the hypothesis, there exist a $\boldsymbol{y}_\Delta \in C_\Delta$ and a $\boldsymbol{d} \in \varphi_\infty$ such that $\bar{y} = \boldsymbol{y}_\Delta + \boldsymbol{d}$.
Now, take any $\boldsymbol{y}_\delta \in S_\delta(\varphi)$: clearly

$$\inf_y \{ \| y - \bar{y} \|_2 : y \in S_\delta(\varphi) \} \leq \| (\boldsymbol{y}_\delta + \boldsymbol{d}) - (\boldsymbol{y}_\Delta + \boldsymbol{d}) \|_2 = \| \boldsymbol{y}_\delta - \boldsymbol{y}_\Delta \|_2.$$

Hence $\sup_{\bar{y}} \{ \inf_y \{ \| y - \bar{y} \|_2 : y \in S_\delta(\varphi) \} : \bar{y} \in S_\Delta(\varphi) \}$

$\leq$ $\sup_{y_\Delta} \{ \| \boldsymbol{y}_\delta - \boldsymbol{y}_\Delta \|_2 : \boldsymbol{y}_\Delta \in C_\Delta \} < +\infty$

since C_Δ is compact. ♦

Note that Property 4.1 is obviously true if $\varphi_\infty = \{ \mathbf{0} \}$, i.e. all *inf-compact* functions are f-regular: however, many other functions are f-regular.

Corollary 4.3: *all polyhedral functions are f-regular.*

Proof: the level sets of polyhedral functions are polyhedra, which can always be expressed minimally as the sum of a polytope ($\Rightarrow$ compact) and a cone: obviously, the cone is φ_∞. ♦

Corollary 4.3: *all quadratic functions* $\varphi(\boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{a})^T \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{a})$ *with* $\boldsymbol{A}$ *symmetric and positive semidefinite are f-regular.*

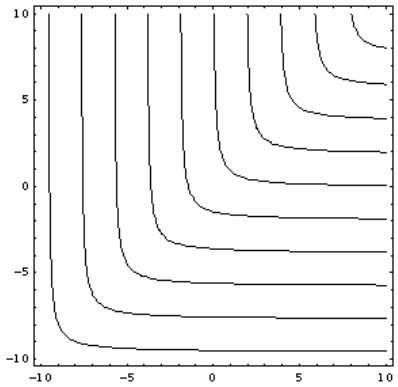
Proof: it is well-known that the level sets of these functions are ellipsoids ($\Rightarrow$ compact) if $\boldsymbol{A}$ is positive definite, and degenerate ellipsoids that can be expressed as the sum of an ellipsoid in *Im* $\boldsymbol{A}$ and the cone (subspace) *Ker* $\boldsymbol{A}$ otherwise. ♦

In the above Corollaries, the set C_δ is a subset of $S_\delta(\varphi)$: hence, one may wonder if Property 4.1 might have been equivalently stated in terms of a *generator* of $S_\delta(\varphi)$, that is a minimal (w.r.t. $\subseteq$) *subset* G of $S_\delta(\varphi)$ such that $S_\delta(\varphi) \subseteq G + \varphi_\infty$. In other words, one may wonder whether f-regular functions need to have compact generators of all the level sets: interestingly, the answer is no.

Observation 4.4: *there are f-regular functions whose generators of the level sets are not compact.*

Proof: an example is the function

$$\varphi(y_1, y_2) = \frac{\sqrt{(y_1 + y_2)^2 + 4} - (y_1 - y_2)}{2}$$



that is differentiable, convex (the eigenvalues of $\nabla^2\varphi(y_1, y_2)$ are 0 and $4(4 + (y_1 - y_2)^2)^{-3/2}$) and finite everywhere. Its level sets

$$S_t(\varphi) = \{ (y_1, y_2) : y_2 \geq 1/(y_1 + t) - t \}$$

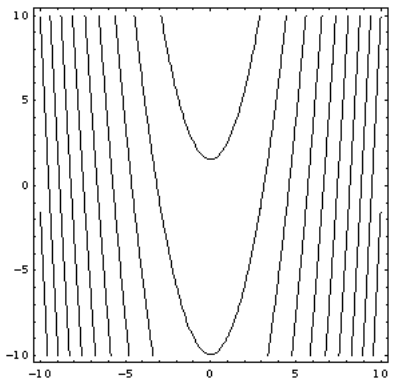
are unbounded and $\varphi_\infty = \Re_+ \times \Re_+$: clearly, their generators (boundaries) are *not* compact. Yet, Property 4.1 holds: just take

$$C_t = \{ (-t, -t) \}.$$ ♦

Another interesting observation is that Property 4.1 is *not* necessary.

Observation 4.5: *there are f-regular functions that do not enjoy Property 4.1.*

Proof: an example is the function $\varphi(y_1, y_2) = y_1^2 - y_2$ that is differentiable, convex and finite everywhere. Its level sets



$$S_t(\varphi) = \{ (y_1, y_2) : y_2 \geq y_1^2 - t \}$$

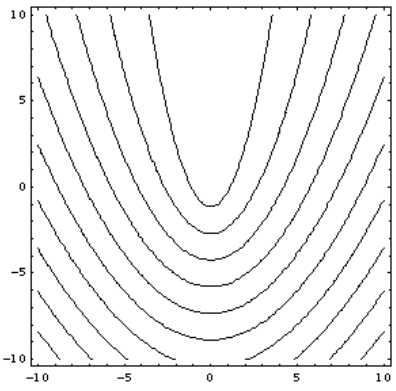
are a family of parabolas translated along the y_2 axe: obviously, $\varphi_\infty = \{ \mathbf{0} \} \times \Re_+$, due to the superlinearity of the parabolas. Since each parabola is "infinitely wide", there cannot be any compact set C such that $C + \varphi_\infty$ surrounds it: however, φ is clearly f-regular. In fact, for each point $((y - \Delta)^{1/2}, y)$ on the boundary of $S_\Delta(\varphi)$ at "height" y there is a point of $S_\delta(\varphi)$ (at the same height) no farther than $(y - \Delta)^{1/2} - (y - \delta)^{1/2}$, and this distance is bounded even if $y \to +\infty$. ♦

There are also convex functions that are *not* f-regular, although they are not very easy to construct.

Observation 4.6: *there are closed convex (differentiable) functions that are finite everywhere but not f-regular.*

Proof: an example is the convex function

$$\varphi(y_1, y_2) = \frac{\sqrt{4y_1^2 + y_2^2} - y_2}{2}$$



(the eigenvalues of $\nabla^2\varphi(y_1, y_2)$ are 0 and $2(y_1^2 + y_2^2)(4y_1^2 + y_2^2)^{-3/2}$): its level sets are

$$S_t(\varphi) = \{ (y_1, y_2) : y_2 \geq (1/t)y_1^2 - t \}$$

for $t > 0$ and $S_0(\varphi) = \{ \mathbf{0} \} \times \Re_+ = \varphi_\infty$, i.e. family of parabolas that become "infinitely steep" until they collapse on nonnegative half-line of the y_2 axe. This function is *not* f-regular, since by taking $0 = \delta < \Delta$ one can always find points in $S_\Delta(\varphi)$ farther from the y_2 axe more than any given threshold: however, note that this is possible *only* with $\delta = 0$. ♦

The function in Observation 4.6 allows to construct an ideal counter-example to the convergence of the Bundle algorithm in the case of non f-regular function. One should assume that the $\bar{\boldsymbol{y}}_i$ converge (from the outside) to the boundary of $S_\Delta(\varphi)$ while their y_2 component goes to $+\infty$. With a proper choice of the stabilizing term, it is possible that the sequence of optimal solutions of $(\Delta_{\bar{y}_i,t_i})$ stay away from the optimal region even if $t_i \to +\infty$. The easier way of constructing a "bad" stabilizing term is to take it nonsmooth in $\mathbf{0}$ (*not* enjoying P6): however, one may also take a D_t that is constant on a tiny ball around $\mathbf{0}$ - hence enjoying P6 - but nonsmooth on the boundary of the ball. This also confirms Theorem 3.24: to impair convergence, it is necessary that the SS are done (in the worst possible way). When SS are inhibited the counter-example fails, indicating that f-regularity is no more needed.

A standard way for describing a class of functions is to give a set of "basic" objects and some operations that preserve membership: for instance, it is easy to see that φ is f-regular $\Rightarrow \alpha\varphi$ is f-regular for each $\alpha > 0$ (since $S_\delta(\alpha\varphi) = S_{\delta/\alpha}(\varphi)$). Unfortunately, few results of this kind can be proved for f-regularity: for instance, the sum of two f-regular functions need not to be f-regular. The counterexample is again the function in Observation 4.6: both $\sqrt{4y_1^2 + y_2^2}$ and $-y_2$ are convex and f-regular (the former is inf-compact and the latter linear), while their sum is not.
However, f-regular functions are likely to be to norm in applications.

5. Extensions and special cases

Bundle algorithms of our class can cope with constraints $\boldsymbol{y} \in Y$, if Y is a convex set. For the sake of simplicity, assume that Y is a polyhedron $\boldsymbol{h} \geq \boldsymbol{y}\boldsymbol{H}$: the two (local) subproblems become

$(\Pi^{\bar{y}}_{\beta,t})$ $\quad \inf_{\boldsymbol{d}} \{ \varphi_{\beta(\bar{y})}(\boldsymbol{d}) + D_t(\boldsymbol{d}) : (\bar{\boldsymbol{y}} + \boldsymbol{d}) \in Y \}$

$(\Delta^{\bar{y}}_{\beta,t})$ $\quad \inf_\theta \{ \alpha_{\beta(\bar{y})}\theta + (\boldsymbol{h} - \bar{\boldsymbol{y}}\boldsymbol{H})\mu + D_t^*(-\boldsymbol{G}_\beta\theta - \boldsymbol{H}\mu) : \theta \in \Theta , \mu \geq \mathbf{0} \}$

and the latter can be equivalently rewritten in terms of *Epi* $(\varphi_{\beta(\bar{y})} + I_Y)^*$. The special case of nonnegativity constraints $\boldsymbol{y} \geq \boldsymbol{0}$ gives

$$(\Delta^{\bar{y}+}_{\beta,t}) \qquad \inf_\theta \{ \alpha_{\beta(\bar{y})}\theta + D_t^*(-[\, \boldsymbol{G}_\beta\theta \,]_+) : \theta \in \Theta , \mu \geq \boldsymbol{0} \}$$

where $[\, \boldsymbol{x} \,]_+ = \max\{ \boldsymbol{x}, \boldsymbol{0} \}$. The convergence proofs of §3 can be extended to this case: first of all, $\bar{\boldsymbol{y}} + \boldsymbol{d}^*$ is *feasible*, so that condition 3.v is satisfied. The only problem could come from Theorem 3.10 where we assume that φ is finite everywhere: hence, the Bundle method cannot be directly applied to $\varphi + I_Y$. However, the black box returns subgradients of φ, that is finite everywhere, and *not* of $\varphi + I_Y$: hence, the convergence analysis is not harmed as long as *full knowledge of Y* is inserted in $(\Pi^{\bar{y}}_{\beta,t})$ *from the beginning*. When Y is a polyhedron, this requirement can be further released: it is sufficient that, when an unfeasible point is probed, the black box returns $+\infty$ and one of the defining hyperplanes of Y. Clearly, only finitely many steps are required to eventually acquire full knowledge about Y: note that aggregation techniques *cannot* be used for constraints. If Y is *not* a polyhedron, it could take infinitely many iterations only to generate the second feasible point: the only "easy" case is when φ itself is linear, since then the algorithm essentially becomes the original Cutting Plane algorithm.

Another important extension is a different treatment of *decomposable* functions

$$\varphi(\boldsymbol{y}) = \Sigma_{i \in K} \varphi^i(\boldsymbol{y})$$

where $|K| = k < \infty$: the obvious example with $k = 2$ is the case of *Lagrangean Decomposition*, where a problem is decomposed by introducing two copies of the same set of variables and dualizing the equality constraints. Separable functions with larger k also appear in cost-decomposition approaches to block-structured problems [GK95], such as the Multicommodity Min Cost Flow problem [FG96]. In this case, computation of each $\varphi^i(\bar{\boldsymbol{y}})$ corresponds to an independent maximization problem, that gives a $\boldsymbol{g}^i \in \partial\varphi^i(\bar{\boldsymbol{y}})$: rather than aggregating this information into an unique subgradient $\boldsymbol{g} = \Sigma_{i \in K} \boldsymbol{g}^i$, one may keep it in a disaggregated form. The Bundle is then partitioned into k disjoint subsets $\beta(i)$, $i \in K$: the "disaggregated subproblems"

$$(\Pi^{\bar{y}}_{\beta,t}) \qquad \min_{d,v} \{ \Sigma_{i \in K} v^i + D_t(\boldsymbol{d}) : v^i \geq z\boldsymbol{d} - \sigma , (\sigma, z) \in \beta(i) , i \in K \}$$

$$(\Delta^{\bar{y}}_{\beta,t}) \qquad \min_\theta \{ \Sigma_{(\sigma,z) \in \beta} \sigma\theta_{(\sigma,z)} + D_t^*(- \Sigma_{(\sigma,z) \in \beta} z\theta_{(\sigma,z)}) : \Sigma_{(\sigma,z) \in \beta(i)} \theta_{(\sigma,z)} = 1 , \theta \geq \boldsymbol{0} , i \in K \}$$

can be solved instead of the aggregated versions (note that the k separate Bundles $\beta(i)$ only interact in the objective function). Use of disaggregate information might clearly be beneficial: the sum of the *k disaggregated CP models* is at least a better description of φ than the aggregate one. In fact, convincing proofs of the benefits of disaggregation are given in a comparison of aggregate and disaggregate Dantzig-Wolfe (i.e. Cutting Plane) approaches for the solution of Multicommodity Min Cost Flow problems [JLF93]. The idea has already been proposed for specific classes of Bundle methods [Ki95] [GV97], but it is notable that it fits perfectly with our general framework, independently on the stabilizing term D_t.

Generalized Bundle methods can, like the "standard" ones [Ki95] [GV97], also work if φ is not calculated exactly: if $\varphi(\bar{\boldsymbol{y}})$ is known with an absolute error of δ, the analysis essentially applies to the approximated function $\tilde{\varphi}$ returned by the black box - whose minimizer is a δ-approximation of the real one - rather than to the original φ. If δ is user-specified and can be driven to 0, the algorithm can be

made to converge to the optimal solution. Of course, some care is needed to ensure consistency in the data of the subproblem, i.e. avoiding negative linearization errors, and/or to ensure that at each step the newly obtained subgradient changes the previous optimal solution of $(\Delta^{\bar{y}}_{\beta,t})$.

Finally, let us mention the interesting interpretation of generalized Bundle methods when applied to the solution of Lagrangean Duals. Assume that we are interested to solve

(P) $\quad \sup_x \{ \boldsymbol{cx} : \boldsymbol{x} \in X , \boldsymbol{Ax} = \boldsymbol{b} \}$

where X is a "simple" convex set and $\boldsymbol{Ax} = \boldsymbol{b}$ are "complicating" constraints, in the sense that (P) could be "efficiently" solved if they could be removed. Under proper assumptions [HL93b - Chap. XII], (P) is equivalent to its Lagrangean Dual

(D) $\quad \inf_y \{ \varphi(\boldsymbol{y}) = \sup_x \{ \boldsymbol{cx} + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{Ax}) : \boldsymbol{x} \in X \} \}.$

In this case, there is a natural interpretation for the set *Epi* φ*: from the definition

$$\varphi^*(z) = \sup_y \{ \inf_x \{ -\boldsymbol{cx} + \boldsymbol{y}(z - (\boldsymbol{b} - \boldsymbol{Ax})) : \boldsymbol{x} \in X \} \}$$
$$= \inf_x \{ -\boldsymbol{cx} : \boldsymbol{b} - \boldsymbol{Ax} = z , \boldsymbol{x} \in X \}.$$

(the former is the Lagrangean Dual of the latter), i.e. φ* *is the value function of* (P) w.r.t. the constraints $\boldsymbol{Ax} = \boldsymbol{b}$ (cfr. Lemma 2.5.) Hence, $\varphi^*(z) = +\infty$ for z not belonging to the image if X under the mapping $\boldsymbol{b} - \boldsymbol{Ax}$: the set $\{ (-\boldsymbol{cx}, \boldsymbol{b} - \boldsymbol{Ax}) : \boldsymbol{x} \in X \}$ can clearly be substituted to *Epi* φ* in the definition of the various problems of §2. Doing that, one obtains

$(\Delta_{\bar{y},t})$ $\quad \sup_x \{ f_{\bar{y},t}(\boldsymbol{x}) = \boldsymbol{cx} + \bar{\boldsymbol{y}}(\boldsymbol{b} - \boldsymbol{Ax}) - D_t^*(\boldsymbol{Ax} - \boldsymbol{b}) : \boldsymbol{x} \in X \}$

i.e. $(\Delta_{\bar{y},t})$ is a *generalised Augmented Lagrangean relaxation* of (P), where both the linear penalty term $\bar{\boldsymbol{y}}(\boldsymbol{b} - \boldsymbol{Ax})$ (with Lagrangean multipliers $\bar{\boldsymbol{y}}$) and the nonlinear one $D_t^*(\boldsymbol{Ax} - \boldsymbol{b})$ are added to the original cost function $\boldsymbol{cx}$. In this case, the Bundle β can be equivalently considered a set of (extreme) points of X: hence, the subproblem that is solved at each step is

$(\Delta_{\beta,\bar{y},t})$ $\quad \sup_x \{ f_{\bar{y},t}(\boldsymbol{x}) : \boldsymbol{x} \in Conv(\{ \boldsymbol{x} \in \beta \}) = X_\beta \}.$

while the "pricing subproblem" is

$(P^{\bar{y}}_{d^*})$ $\quad \sup_x \{ \boldsymbol{cx} + (\boldsymbol{b} - \boldsymbol{Ax})(\bar{\boldsymbol{y}} + \boldsymbol{d}) : \boldsymbol{x} \in X \}.$

In this case, the "dual" interpretation of the Bundle algorithm goes as follows: (P) is replaced by

$$\sup_x \{ \boldsymbol{cx} - D_t^*(\boldsymbol{Ax} - \boldsymbol{b}) : \boldsymbol{x} \in X , \boldsymbol{Ax} = \boldsymbol{b} \}$$

and we are solving its Lagrangean Dual w.r.t. the constraints $\boldsymbol{Ax} = \boldsymbol{b}$, ending in the generalised Augmented Lagrangean relaxation $(\Delta_{\bar{y},t})$. The latter problem is "difficult", but its linearization $(P^{\bar{y}}_{d^*})$ can be easily solved: hence we resort to an *inner linearization approach*. Having collected a set β of (extreme) points of X, we solve the problem restricted to X_β obtaining a tentative optimal solution $\boldsymbol{x}^*$: then, we use $\boldsymbol{c} + (\bar{\boldsymbol{y}} + \boldsymbol{d})\boldsymbol{A}$ (that can be easily shown to be a *subgradient* of $f_{\bar{y},t}$ in $\boldsymbol{x}^*$, cfr. Lemma 2.5) in the linear approximation to obtain a new point that enlarges X_β.

This formally establish the *substantial identity* between (generalized) Bundle methods applied to the Lagrangean Dual of an optimization problem and (generalized) Augmented Lagrangean approaches. In

case of inequality constraints $\boldsymbol{Ax} \le \boldsymbol{b}$

$$(\Delta_t^*) \qquad \sup_{\boldsymbol{x}} \{ f_{\bar{y},t}(\boldsymbol{x}) = \boldsymbol{cx} + \bar{\boldsymbol{y}}(\boldsymbol{b} - \boldsymbol{Ax}) - D_t^*(-[\boldsymbol{Ax} - \boldsymbol{b}]_+) : \boldsymbol{x} \in X_\beta \}$$

i.e. the penalty function only counts the violated constraints; alternatively, one could use a $D_t^*(\boldsymbol{Ax} - \boldsymbol{b})$ that is zero everywhere but on the nonnegative orthant and adjust the convergence theory accordingly. This kind of approach has been proposed independently from the Bundle idea: one example is [PZ94], where $D_t^*(z) = t\Sigma_{i=1..n} \Phi_\varepsilon(z_i)$ for

$$\Phi_\varepsilon(z) = \begin{cases} 0 & \text{if } z \le 0 \\ z^2/2\varepsilon & \text{if } 0 \le z \le \varepsilon \\ z - \varepsilon/2 & \text{if } \varepsilon \le z \end{cases}$$

and some fixed $\varepsilon > 0$. Φ_ε is meant to " smoothly approximate" the nonsmooth exact penalty term $t\max\{0, z\}$: the corresponding $(\Delta_{\beta,\bar{y},t})$ is solved with a *simplicial decomposition*, i.e. inner linearization, approach. The proposed algorithm is therefore *exactly* a Bundle method of our class, of the type where no Serious Steps are ever done: the only relevant variant is that the ε in the penalty function is essentially the one in the stopping condition, that is driven to 0 each time that "convergence" is reached - however, this is covered by our theory.
Presumably, the choice of not performing SS was not conscious: the dual viewpoint alone does *not* suggest the SS mechanism at all. In fact, although the primal problem $(\Pi_{\beta,\bar{y},t})$ is not described, the conjugate of the penalty term

$$\Phi_\varepsilon^*(d) = \begin{cases} \varepsilon/2t\, d^2 & \text{if } 0 \le d \le t \\ +\infty & \text{otherwise} \end{cases}$$

is mentioned in [PZ94]. This might have been unfortunate, because the primal problem is just a variant of $(\Pi^2_{\beta(\bar{y}),t})$ (cfr. §1) with box constraints, that can be solved with specialised codes [Fr96] possibly more efficiently than with the truncated Newton method employed in [PZ92]. Hence, the primal *and* dual viewpoints together seem to give a larger set of tools upon which to construct (hopefully) more efficient Augmented Lagrangean approaches.

To conclude this section, let us mention that generalized Bundle methods are also closely related with recent ε-approximation algorithms for (block-)structured convex problems like (P) [GK95]. In these approaches, (P) is substituted with $(\Delta_{\beta,t})$ where

$$D_t^*(z) = t/\varepsilon \ln(1 + \Sigma_{i=1..n} e^{\varepsilon z_i})$$

for some fixed $\varepsilon > 0$: in this case, D_t^* is conceived as an approximation of $t\max_i \{ z_i \}$ (i.e. of the L_∞ norm rather than of the L_1 norm). The subproblem is solved with the usual inner linearization approach, where the gradient of D_t^* (called *potential function*) in the solution $\boldsymbol{z}^*$ of the previous iterate is used to generate a new point $\boldsymbol{g}$: at each step, only the "minimal" Bundle containing $\boldsymbol{z}^*$ and $\boldsymbol{g}$ is kept - actually, the k disaggregate representative of the former and of the latter, since the method is proposed for block-structured problems, i.e. decomposable functions. As in [PZ94], a proper choice of ε ensures that the obtained solution be approximately optimal for (P): in this case, however, t can be kept fixed (to a properly chosen value) and a sophisticated analysis gives a bound on the number of iterations needed to achieve the result. This approach is *not* precisely a generalized Bundle method for

several reasons: for example, D_t^* is not zero in the feasible region and the algorithm starts with a feasibility phase where D_t^* is optimized alone, without the original objective function (something like setting $t = +\infty$). Furthermore, the analysis of the approach is based on reducing the problem to a "resource sharing"-type problem and applying an approximation algorithm to it. However, the two methods are clearly strictly related, and it could be possible to exchange ideas between the two: for instance, the concepts of current point and Serious Step could be introduced in the approximation algorithm, while part of the complexity analysis of the latter could be applied to versions of the former using (slight variants of) the exponential penalty in the dual.
Here, note one further example of the importance of the primal-dual relationships: the conjugate of the exponential penalty function is not easy to characterise, hence $(\Pi_{\beta,t})$ is even difficult to write - still, a Bundle method using it can be implemented.

6. Conclusions

The use of non-quadratic proximal terms in convex programming is definitely not a new idea: see e.g. [IST94] or the very recent [PT97] and the references therein. However, it is usually employed within Proximal Point algorithms, that assume *exact* solution of the proximal subproblem ($(\Delta_{\bar{y},t})$ in our notation) and therefore lack implementability in many cases: on the contrary, the Bundle approach proposed in this article is implementable in virtually every circumstance, provided that the stabilizing term is chosen in such a way that $(\Delta_{\beta,\bar{y},t})$ is efficiently solvable.

We have given a convergence proof for a class of Bundle methods using a generic stabilizing term: the statements of the properties needed for convergence allow a great flexibility in the implementation of the method, and several variants have been given with different convergence properties according to the characteristics of the function to be minimized and of the stabilizing term employed. Our proofs are somehow weaker than some of those known for specific choices of D_t [HL93b - §XV.2/3]: for $D_t = {}^1/_{2t}\|\cdot\|_2^2$, for instance, t can (slowly) converge to zero during infinite sequences of NS and/or SS, f-regularity is not needed and also the $\{ \bar{y}_i \}$ sequence converges whenever the φ has a nonempty set of minimizers. For $D_t = I_{B_p(t)}$, asymptotic convergence can be proved without the f-regularity assumption. However, our results are stronger than those in [KCL95] and more general than those in [PZ94]; furthermore, they are satisfactory for practical applications. Besides, this seems to be a price to pay for generality: only the very basic conditions has been imposed on D_t, hence our convergence analysis works also for "strange" functions with little regularity, that do not seem to allow stronger results (think to a D_t constant in a tiny ball around 0 and nondifferentiable on its boundary, or smooth in 0 and growing very little in a ball of radius t).

The results presented in this paper suggest practical implementations of Bundle algorithms with several different non-quadratic stabilizing terms: examples are primal and/or dual trust regions based on "linear" (L_1 or L_∞) norms, with the possibility important advantage of requiring the solution of a Linear Programming subproblem rather than a Quadratic Programming one at each step. Other stabilizing terms, maybe entropy-like or exponential, may have better theoretical or practical convergence properties than the L_2 norm that compensate the more difficult subproblem to be solved: computational comparisons will be needed to assess the relative efficiency of these variants.

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

References

[AV95] D.S. ATKINSON, P.M. VAIDYA "**A Cutting Plane Algorithm for Convex Programming that Uses Analytic Centers**" *Math. Prog.* **69**, p. 1-43, 1995

[CFN96] P. CARRARESI, A. FRANGIONI AND M. NONATO "**Applying Bundle Methods to Optimization of Polyhedral Functions: An Applications-Oriented Development**" *Ricerca Operativa* Vol. **25** n. 74, p. 5-49, 1996

[FG96] A. FRANGIONI, G. GALLO "**A Bundle type Dual-ascent Approach to Linear MultiCommodity Min Cost Flow Problems**" *Technical Report* **TR 01/96**, Dip. di Informatica, Università di Pisa, 1996

[Fr96] A. FRANGIONI "**Solving Semidefinite Quadratic Problems Within Nonsmooth Optimization Algorithms**" *Computers & O.R.* Vol. **23**(11), p. 1099-1118, 1996

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

[Ge69] A.M. GEOFFRION "**Duality in Nonlinear Programming: a Simplified Applications-Oriented Development**" in *Prespective on Optimization: a collection of expository articles* (A.M. Geoffrion ed.), Addison-Welsey, 1972

[GK95] M.D. GRIGORIADIS, L.G. KAHCHIYAN "**An Exponential-Function Reduction Method for Block-Angular Convex Programs**" *Networks* **26**(2), p. 59-68, 1995

[GM91] M. GAUDIOSO, M.F. MONACO "**Quadratic Approximations in Convex Nondifferentiable Optimization**" *SIAM J. on Control and Optimization* **29**(1), p. 58-70, 1991

[GV97] J. GONDZIO, J.-P. VIAL "**Warm Start and ε-subgradients in Cutting Plane Scheme for Block-angular Linear Programs**" *Logilab Technical Report* **1997.1**, 1997

[HL93a] J-B. HIRIART-URRUTY, C. LEMARÉCHAL "**Convex Analysis and Minimization Algorithms I - Fundamentals**" *A Series of Comprehensive Studies in Mathematics*, vol. **306**, Springer-Verlag, 1993

[HL93b] J-B. HIRIART-URRUTY, C. LEMARÉCHAL "**Convex Analysis and Minimization Algorithms II - Advanced Theory and Bundle Methods**" *A Series of Comprehensive Studies in Mathematics*, vol. **306**, Springer-Verlag, 1993

[IST94] A.N. IUSEM, B.F. SVAITER AND M. TEBOULLE "**Entropy-like Proximal Methods in Convex Programming**" *Math. of O.R.* **19**(4), p. 790-814, 1994

[JLF93] K.L. JONES, I.J. LUSTIG, J.M. FARVOLDEN AND W.B. POWELL "**Multicommodity Network Flows: The Impact of Formulation on Decomposition**" *Math. Prog.* **62**, p. 95-117, 1993

[Ke60] J.E. KELLEY "**The Cutting-Plane Method for Solving Convex Programs**" *Journal of the SIAM* **8**, p. 703-712, 1960

[Ki85] K.C. KIWIEL "**Methods of Descent for Nondifferentiable Optimization**" *Lecture Notes in Mathematics*, vol. **1133**, Springer-Verlag, 1985

[Ki86] K.C. KIWIEL "**A Method for Solving Certain Quadratic Programming Problems Arising in Nonsmooth Optimization**" *IMA J. Numer. Anal.* **6**, p. 137-152, 1986

[Ki90] K.C. KIWIEL "**Proximity Control in Bundle Methods for Convex Nondifferentiable Optimization**" *Math. Prog.* **46**, p. 105-122, 1990

[Ki95] K.C. KIWIEL "**Approximations in Proximal Bundle Methods and Decomposition of Convex Programs**" *JOTA* **84**(3), p. 529-548, 1997

[KCL95] S. KIM, K.N. CHANG AND J.Y. LEE "**A Descent Method with Linear Programming Subproblems for Nondifferentiable Convex Optimization**" *Math. Prog.* **71**, p. 17-28, 1995

[LNN95] C. LEMARÉCHAL, A. NEMIROVSKII AND Y. NESTEROV "**New Variants of Bundle Methods***" Math. Prog.* **69**, p. 111-147, 1995

[Mi77] R. MIFFLIN "**A Feasible Descent Algorithm for Linearly Constrained Least Squares Problems**" *in Nonsmooth Optimization*, vol. **3** of IIASA Proceedings Series (C. Lemaréchal, R. Mifflin eds.), Pergamon Press, 1977

[Ne95] Y. NESTEROV "**Complexity Estimates of Some Cutting Plane Methods Based on the Analytic Barrier**" *Math. Prog.* **69**, p. 149-176, 1995

[PT97] R. POLYAK, M. TEBOUULE "**Nonlinear Rescaling and Proximal-like Methods in Convex Optimization**" *Math. Prog.* **76**(2), p. 265-284, 1997

[PZ92] M.C. PINAR, S.A. ZENIOS "**Parallel Decomposition of Multicommodity Network Flows Using a Linear-Quadratic Penalty Algorithm**" *ORSA J. on Computing* **4**(3), p. 235-249, 1992

[PZ94] M.C. PINAR, S.A. ZENIOS "**On Smoothing Exact Penalty Functions for Convex Constrained Optimization**" *SIAM J. on Optimization* **4**(3), p. 486-511, 1994

[Ro76] R.T. ROCKAFELLAR "**Monotone Operators and the Proximal Point Algorithm**" *SIAM J. on Control and Optimization* **14**(5), 1976

[SZ92] H. SCHRAMM, J. ZOWE "**A Version of the Bundle Idea for Minimizing a Nonsmooth Function: Conceptual Idea, Convergence Analysis, Numerical Results**" *SIAM J. Optimization* **2**(1), p. 121-152, 1992