I.1 A View of Bundle Algorithms

During the last 20 years, Bundle methods for Nondifferentiable Optimization (NDO) have become an important research area: stemming from the pioneering work of Lemaréchal, a large number of papers and books devoted to several variants of Bundle methods have appeared, both for unconstrained and constrained optimization. Nowadays, Bundle methods have a solid theoretical background as well as important applications (one for all, the solution of Lagrangean Duals): the main aim of all Part I is to present an application-oriented development of Bundle methods, with a convergence proof for a general form of the algorithm that explicates the very basic properties needed for the convergence analysis, thus allowing greater flexibility in the design of practical algorithms. When Bundle methods are applied to the optimization of Lagrangean functions, a "primal" interpretation¶ can be given to their basic ideas and algorithmic tools: such an interpretation leads to the development of a new class of algorithms with Linear Programming direction finding subproblems, which can be viewed as "stabilizations" of the Dantzig-Wolfe method, and it also suggests some new ideas for the most delicate task in any Bundle algorithm, i.e. the handling of the "trust region parameter". This Chapter is devoted to the presentation of the relevant terminology and ideas, together with an applications-oriented review of the Bundle approaches proposed in the literature: in particular, §I.1.1 deals with the "dual" viewpoint that is usually assumed in NDO theory, i.e. the maximization of a NonDifferentiable function provided as a "black box", while §I.1.2 illustrates the "primal" viewpoint.

I.1.1 The Dual viewpoint

Except for the treatment of the stopping criteria at the end, the material presented in this paragraph is not new: however, an attempt has been done to resume the large body of knowledge about Bundle methods in an original way. Bundle methods are perhaps the most important class of implementable algorithms for the numerical solution of the problem (that, for reasons to become clear in §I.1.2, is considered of a "dual" nature)

(D) $\max\{\ \varphi(\boldsymbol{y}) : \boldsymbol{y} \in Y\ \}$

where $\varphi()$ is a *concave* and *finite everywhere* (nondifferentiable) function and Y is a *convex nonempty* subset of $\mathcal{R}^n$. In the following development, we will adopt the usual assumption that $\varphi()$ be given as a "black box", which inputs a vector $\bar{\boldsymbol{y}} \in Y$ and outputs the value of $\varphi(\bar{\boldsymbol{y}})$ and one *subgradient*

¶ In a different sense from the one usually found in textbooks, e.g. [HL93 Chapter XV].

$\boldsymbol{g} = \boldsymbol{g}(\bar{\boldsymbol{y}})$ of $\varphi()$ in $\bar{\boldsymbol{y}}$, i.e. a vector of $\mathcal{R}^m$ that satisfies the *subgradient inequality*

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

It should be noted that the definition of subgradient is usually given for *convex* functions, and is therefore slightly different from the above (which is that of a *supergradient* in the standard terminology); analogously, concepts like the *epigraph* of $\varphi()$

$$Epi(\varphi) = \{ (\boldsymbol{y}, v) \in \mathcal{R}^{m+1} : v \le \varphi(\boldsymbol{y}) \},$$

that are defined here for convenience in the setting of concave functions, may be found defined with a reversed inequality sign in a Convex Analysis text.

Geometrically, a subgradient is the gradient of a *supporting hyperplane* of $Epi(\varphi)$ in $\bar{\boldsymbol{y}}$ (Figure 1): unlike the gradient of a differentiable function, however, subgradients are not guaranteed to be directions of ascent (Figure 2). Although apparently poor, this first-order information is enough to drive convergent algorithms for the solution of (D): from the subgradient inequality, it easily follows that the level set $\{ \boldsymbol{y} : \varphi(\boldsymbol{y}) \ge \varphi(\bar{\boldsymbol{y}}) \}$, and therefore the set of the optimal solutions of (D), lies in the half-space pointed by $\boldsymbol{g}$ w.r.t. $\bar{\boldsymbol{y}}$ (Figure 2).

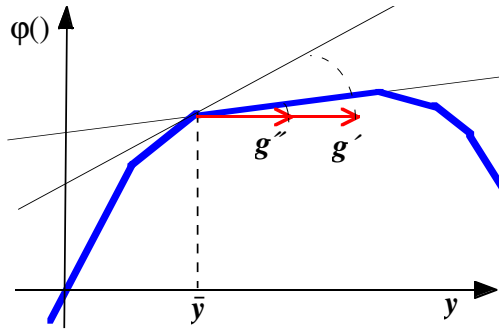

Figure 1: two subgradients $\boldsymbol{g}'$ and $\boldsymbol{g}''$ in a kink point

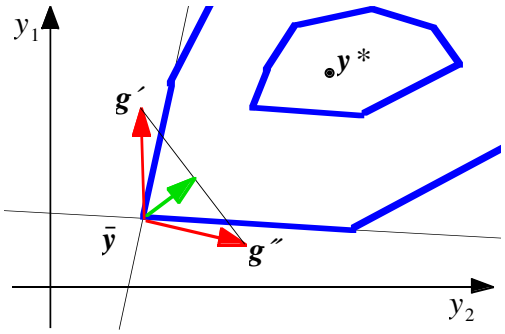

Figure 2: contour plots of $\varphi()$ and an optimal solution $\boldsymbol{y}^*$

It is well-known that any local optima of (D) is also a global optima, and that the problem enjoys many useful properties, such as continuity of $\varphi()$ on the relative interior of Y and existence of a subgradient in any point of Y: the proofs of these and the other basic properties stated throughout this paragraph, as well as a complete treatment of the underlying ε-subgradient theory, are beyond the scope of this work and will not be reported, the interested reader being referred to some of the available excellent books on Convex Analysis [Ro70] [HL93].

A particular class of concave functions is in fact the main target of the present work: *polyhedral functions* mostly arise from *Lagrangean Relaxations* of (Integer) Linear Programs, and have the form

$$\varphi(\boldsymbol{y}) = \min_{i=1..l} \{ b_i + \boldsymbol{y}\boldsymbol{h}_i \}$$

i.e. the pointwise minimum of a finite (but very large) set of linear functions. In this case, the black box typically returns the gradient $\boldsymbol{h}_i$ of one of the defining hyperplanes that are "tight" in $\bar{\boldsymbol{y}}$: this requires the solution of the possibly "difficult" optimization problem of choosing among the (exponentially many) facets that define the function.

For the sake of simplicity, the first two Chapters will focus on unconstrained NonDifferentiable Optimization problems, i.e. to the case $Y = \mathcal{R}^m$: by allowing $\varphi()$ to assume the value $-\infty$ somewhere,

this could be done without loss of generality, since any constrained problem can be turned into an unconstrained one by adding to $\varphi()$ the (concave) *characteristic function* of the set Y

$$\rho_Y(\boldsymbol{y}) = \begin{cases} 0 & \text{if } \boldsymbol{y} \in Y \\ -\infty & \text{otherwise} \end{cases}$$

The unconstrained case is interesting in itself, since for instance the Lagrangean Relaxation of equality constraints only gives rise to problems of this form: however, extensions to the constrained case will be discussed in §I.3.3, §I.3.5 and §I.4.3 for the case where Y is a polyhedron, that covers most practical applications.

The basic idea of Bundle methods is to collect all the information $\langle \varphi(\boldsymbol{y}_i), \boldsymbol{g}(\boldsymbol{y}_i) \rangle$ generated in the previous iterates in a set β, the *Bundle*, that is used at each step to compute a *tentative ascent direction* $\boldsymbol{d}$ along which the next points are generated. Since $\boldsymbol{d}$ is not guaranteed to be a direction of ascent, Bundle methods can usually decide whether to accept the step and move the *current point* $\bar{\boldsymbol{y}}$ (*SS*, a *Serious Step*), or rather use the new information to enlarge β and obtain a (hopefully) better $\boldsymbol{d}$ at the next iteration (*NS*, a *Null Step*): a very general Bundle scheme is stated in Figure 3.

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 3: a "generic" Bundle algorithm.

Such a general description allows the *Cutting Plane* algorithm to be included among the Bundle approaches: it is based on the simple observation that the *Cutting Plane model*

$$\bar{\varphi}_\beta(\boldsymbol{y}) = \min_{i \in \beta} \{ \varphi(\boldsymbol{y}_i) + \boldsymbol{g}(\boldsymbol{y}_i)(\boldsymbol{y} - \boldsymbol{y}_i) \}$$

(Figure 4) is a polyhedral upper approximation of $\varphi()$ - formed, when $\varphi()$ itself is polyhedral, of a subset of its facets - which is tight at least in every point $\boldsymbol{y}_i$: a straightforward idea is then to maximize the known $\bar{\varphi}_\beta()$ in place of the unknown $\varphi()$, and use (one of) the maximand(s) as the next trial point where to evaluate the function. An attractive feature of the CP algorithm is that a maximizer of $\bar{\varphi}_\beta()$ can be found by means of a Linear Programming problem: in fact, $\bar{\varphi}_\beta()$ can be rewritten with respect to $\boldsymbol{d} = \boldsymbol{y} - \bar{\boldsymbol{y}}$ (the displacement from the current point $\bar{\boldsymbol{y}}$), obtaining

$$\varphi_{\beta,\bar{\boldsymbol{y}}}(\boldsymbol{d}) = \min_{i \in \beta} \{ \alpha_i + \boldsymbol{d}\boldsymbol{g}_i \} + \varphi(\bar{\boldsymbol{y}})$$

where $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{y}_i)$, $\alpha_i = \varphi(\boldsymbol{y}_i) + \boldsymbol{g}(\boldsymbol{y}_i)(\bar{\boldsymbol{y}} - \boldsymbol{y}_i) - \varphi(\bar{\boldsymbol{y}}) \geq 0$ and $\varphi_{\beta,\bar{\boldsymbol{y}}}(\boldsymbol{d}) = \bar{\varphi}_\beta(\bar{\boldsymbol{y}} + \boldsymbol{d})$ (Figure 5);

then, the desired point is $\bar{\boldsymbol{y}} + \boldsymbol{d}^*$ where $\boldsymbol{d}^*$ is an optimal solution of

$(D_{\beta,\bar{y}})$ $\quad \max_{\boldsymbol{d}} \{ \varphi_{\beta,\bar{y}}(\boldsymbol{d}) \} = \max_{\boldsymbol{d},v} \{ v : v \le \boldsymbol{d}\boldsymbol{g}_i + \alpha_i \; i \in \beta \} + \varphi(\bar{\boldsymbol{y}})$.

Since $(D_{\beta,\bar{y}})$ has a limited number of constraints, it can be efficiently solved: note that, for the sake of simplicity, β will denote the set of pairs $\langle \alpha_i , \boldsymbol{g}(\boldsymbol{y}_i) \rangle$ as well as a set of their indices i, and $\varphi_\beta(\boldsymbol{d})$ will be used instead of $\varphi_{\beta,\bar{y}}(\boldsymbol{d})$ whenever the current point $\bar{\boldsymbol{y}}$ is clear from the context.

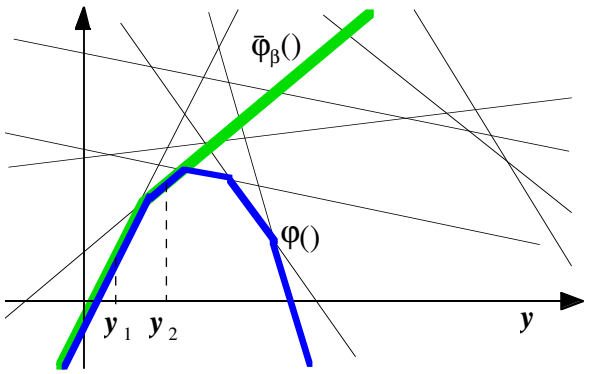

Figure 4: $\bar{\varphi}_\beta()$ approximates $\varphi()$, and is tight in $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$

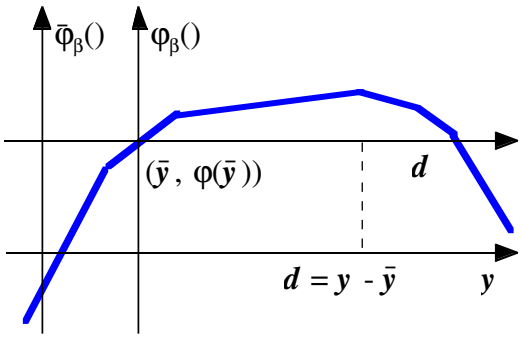

Figure 5: $\varphi_{\beta,\bar{y}}(\boldsymbol{d})$ is a translation of $\bar{\varphi}_\beta(\boldsymbol{y})$

If $\varphi_\beta(\boldsymbol{d}^*) = \varphi(\bar{\boldsymbol{y}} + \boldsymbol{d}^*)$ then $\boldsymbol{y} = \bar{\boldsymbol{y}} + \boldsymbol{d}^*$ is optimal for the entire problem (D), otherwise a new cutting plane $\langle \varphi(\boldsymbol{y}) , \boldsymbol{g}(\boldsymbol{y}) \rangle$ is generated and added to β to improve the CP model: this gives (in principle) a very simple globally convergent approach, requiring the solution of a (LP) at each step. Unfortunately, the CP approach is known to have a bad practical behaviour, mainly because of the "artificial" (usually box) constraints that have to be added to (D_β), at least in the initial iterations, to ensure that it has a finite solution: it is difficult to find "tight" such constraints that does not "cut away" $\boldsymbol{y}^*$, so that the new $\boldsymbol{y}$ is often unreasonably "far" from $\bar{\boldsymbol{y}}$, where $\varphi_\beta()$ does not correctly approximate $\varphi()$, resulting in a slow convergence.

A "dual" approach w.r.t. CP is based on the properties of the *subdifferential* $\partial\varphi(\bar{\boldsymbol{y}})$, i.e. the set of all the subgradients of $\varphi()$ in $\bar{\boldsymbol{y}}$: $\partial\varphi(\bar{\boldsymbol{y}})$ is a *closed* and *convex* set, and if $\varphi()$ is finite everywhere in a neighbourhood of $\bar{\boldsymbol{y}}$, then $\partial\varphi(\bar{\boldsymbol{y}})$ is also *compact*. The point-to-set mapping $\partial\varphi(\bar{\boldsymbol{y}})$ generalises the point-to-point mapping $\nabla\varphi(\bar{\boldsymbol{y}})$ since

- if $\varphi()$ is differentiable in $\bar{\boldsymbol{y}}$, then $\partial\varphi(\bar{\boldsymbol{y}}) = \{ \nabla\varphi(\bar{\boldsymbol{y}}) \}$;
- let $\varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d})$ be the directional derivative at $\bar{\boldsymbol{y}}$ along the direction $\boldsymbol{d}$: the relation

  $$\varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d}) = \nabla\varphi(\bar{\boldsymbol{y}})\boldsymbol{d}$$

  that holds if $\varphi()$ is differentiable in $\bar{\boldsymbol{y}}$, extends to

  $$\varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d}) = \min_v \{ \boldsymbol{v}\boldsymbol{d} : \boldsymbol{v} \in \partial\varphi(\bar{\boldsymbol{y}}) \}$$

  so that $\varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d})$ exists and it is *finite* if $\varphi()$ is finite everywhere in a neighbourhood of $\bar{\boldsymbol{y}}$ (in general, the concavity of $\varphi()$ implies that $\varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d})$ exists for *any* $\bar{\boldsymbol{y}}$ and $\boldsymbol{d}$, but it can be $-\infty$);
- $\nabla\varphi(\bar{\boldsymbol{y}})$ (when existent) is the *steepest ascent direction* for $\varphi()$ in $\bar{\boldsymbol{y}}$, i.e.

  $$\nabla\varphi(\bar{\boldsymbol{y}}) = \operatorname{argmax}_{\boldsymbol{d}} \{ \varphi'(\bar{\boldsymbol{y}} , \boldsymbol{d}) : \| \boldsymbol{d} \| = 1 \}$$

but more in general

($Q_{\bar{y}}$) $\quad \operatorname{argmax}_d \{ \varphi'(\bar{y}, d) : \| d \| = 1 \} = \operatorname{argmin}\{ \| g \| : g \in \partial\varphi(\bar{y}) \};$

– finally, from the above relations it follows that

$$y^* \text{ is } \textit{optimal} \text{ for } \varphi() \Leftrightarrow 0 \in \partial\varphi(y^*) \Leftrightarrow 0 = \min\{ \| g \| : g \in \partial\varphi(y^*) \}.$$

At a first sight, nondifferentiable concave functions seem not to be much harder to maximize than differentiable ones: first-order information about $\varphi()$ of (apparently) the same "strength" as $\nabla\varphi()$ can be obtained at the "only" cost of solving the Least-Square *steepest ascent direction finding* subproblem ($Q_{\bar{y}}$) on the convex set $\partial\varphi()$. However, the rationale under the "black box" form of $\varphi()$ is exactly that it is impossible, or too costly, to work with the whole $\partial\varphi()$: the simplest idea is then to use *any* subgradient, i.e. the one returned by the black box, as an approximation of an optimal solution of ($Q_{\bar{y}}$). The result is the classic *Subgradient Algorithm* [Po77] [BNS87], that directly generalises the Steepest Descent method for the smooth case: its main advantages are its low cost and straightforward implementations, that have made it very popular among researchers in practical optimization. In the Bundle terminology, β here is made of the last subgradient only, and therefore movements to the trial point y must always be performed in order to avoid cycling when d is not a direction of ascent; the stepsize along the current direction is usually taken according to a predefined rule, several of which have been proposed in the literature [HVC74] [CFM75] [Br95], rather than dynamically adjusted via a Line Search. Unfortunately, the rules that work better in practice (e.g. geometric decrease of the stepsize every p iterations) does not usually guarantee convergence to the optimal solution; furthermore, although the method can be extended by taking into account the previous direction [Cr76] [Br91] or the last two subgradients [KU93], its fundamental Markovian nature heavily limits its efficiency: at the k-th step, almost no trace is left of what happened at steps $0 \,..\, k-1$.

The main theoretical problem of Subgradient algorithms is that (see Figure 2) subgradients are not guaranteed to be directions of ascent: in principle, however, this can (in theory) be easily detected by an exact Line Search, as if the current direction is not of ascent, the LS will just return the current point $\bar{y}$. During such a LS, a new subgradient at $\bar{y}$ (different from the current one) could be generated: this suggest to resort to some form of *inner approximation* of the subdifferential, where the Bundle β is meant to contain some subgradients of the current point. A *tentative* steepest ascent direction could be found by solving ($Q_{\bar{y}}$) restricted to the convex hull of the known subgradients in $\bar{y}$, i.e.

(Δ_β) $\quad \min_\theta \{ \| \Sigma_{j \in \beta} g_j \theta_j \| : \theta \in \Theta \} = \min_g \{ \| g \| : g \in \operatorname{conv}(\{ g_i : i \in \beta \}) \}$

where $\Theta = \{ \theta : \Sigma_{j \in \beta} \theta_j = 1 , \theta \geq 0 \}$ is the unitary simplex in the $|\beta|$-dimensional space: if $d = \Sigma_{j \in \beta} g_j \theta_j$ is not of ascent, the new subgradient is used to refine the current approximation of $\partial\varphi(\bar{y})$. Such an approach is only *conceptual*, however, since an exact LS may not finitely terminate on generic functions - although it would on polyhedral ones - and, even if it did, there is no guarantee that a new subgradient at $\bar{y}$ would be generated: note that calling again the black box on $\bar{y}$ does not solve the problem. However, one may hope that, if the LS terminates "close" to $\bar{y}$, then the last subgradient will be "close" to a subgradient at $\bar{y}$: this idea leads to the *Conjugated Subgradient* method [Wo75], that uses (Δ_β) a specialized inexact Line Search along d to find either a significantly better

point $\boldsymbol{y}$, or a new subgradient $\boldsymbol{g}$ which is *not* guaranteed to be a subgradient in $\bar{\boldsymbol{y}}$, but to yield a different search direction $\boldsymbol{d}$ at the next iteration; since it is obtained "close" to $\bar{\boldsymbol{y}}$, it still gives relevant information on the behaviour of $\varphi()$ around $\bar{\boldsymbol{y}}$. A drawback of such an approach is that every time a good improvement is obtained and the current point is moved, β should be emptied, loosing all the information about $\varphi()$: since the elements of β do not really need to be subgradients of the new current point, a more attractive alternative is to keep them, provided that some sort of *reset rule* avoids outdated information to interfere with the determination of a correct $\boldsymbol{d}$. Several such rules may be imagined, such as the *exponential smoothing* technique where at each step an old $\boldsymbol{g}_i$ is multiplied by a factor $\mu^{A(i)}$, where $\mu < 1$ and $A(i)$ is its *age* (how many iterations ago it was found): however, a more robust approach calls for theoretically justifiable rules, that can be obtained by appropriately defining *approximate subgradients*.

An ε-*subgradient* (of $\varphi()$ in $\bar{\boldsymbol{y}}$) has to satisfy the obvious inequality

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

i.e. to be the gradient of a supporting hyperplane $\boldsymbol{g}$ of the epigraph of $\varphi()$ in $(\bar{\boldsymbol{y}},\ \varphi(\bar{\boldsymbol{y}}) + \varepsilon)$ (Figure 6): the previous results for the special case $\varepsilon = 0$ can be generalized by using the notions of ε-*subdifferential* set $\partial_\varepsilon\varphi(\bar{\boldsymbol{y}})$ and ε-*ascent direction* for $\varphi()$ in $\bar{\boldsymbol{y}}$, that is a direction $\boldsymbol{v}$ such that

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

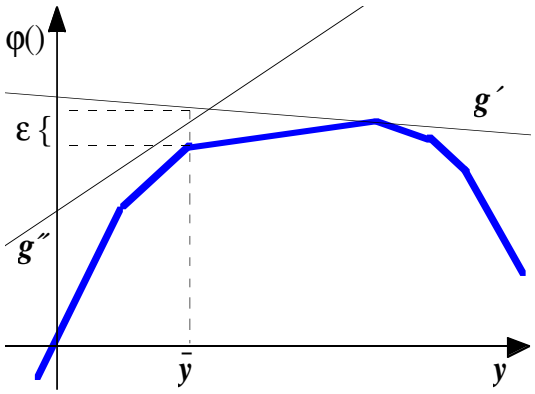

Figure 6: two ε-subgradients for $\varphi()$ in $\bar{\boldsymbol{y}}$: $\boldsymbol{g}'$ gives the (right) ε-directional derivative in $\bar{\boldsymbol{y}}$

i.e. an improvement strictly larger than ε is *possible* along $\boldsymbol{v}$. A strictly related concept is that of ε-*directional derivative* of $\varphi()$ in $\bar{\boldsymbol{y}}$ along $\boldsymbol{d}$, that is, roughly speaking, the slope of the steepest line contained in the restriction of $\varphi()$ along $\boldsymbol{d}$, passing from $(\bar{\boldsymbol{y}}\ ,\ \varphi(\bar{\boldsymbol{y}}) + \varepsilon)$ and touching the epigraph of $\varphi()$ somewhere (Figure 6): more formally

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

(the formula justifies for the name), and it can be easily seen that $\boldsymbol{d}$ *is of* ε-*ascent* $\Leftrightarrow \varphi'_\varepsilon(\bar{\boldsymbol{y}}\ ,\boldsymbol{d}) > 0$.

The previous results gracefully extend to the new case, i.e.

- $\varphi'_\varepsilon(\bar{\boldsymbol{y}}\ ,\boldsymbol{d}) = \min_{\boldsymbol{v}}\{\ \boldsymbol{v}\boldsymbol{d} : \boldsymbol{v} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}})\ \}$;
- $(Q^\varepsilon_{\bar{\boldsymbol{y}}})$ $\operatorname{argmin}_{\boldsymbol{g}}\{\ \|\boldsymbol{g}\| : \boldsymbol{g} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}})\ \} = \operatorname{argmax}_{\boldsymbol{d}}\{\ \varphi'_\varepsilon(\bar{\boldsymbol{y}}\ ,\boldsymbol{d}) : \|\boldsymbol{d}\| = 1\ \}$;
- $\bar{\boldsymbol{y}}$ is ε-optimal for $\varphi() \Leftrightarrow \boldsymbol{0} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}}) \Leftrightarrow 0 = \min_{\boldsymbol{g}}\{\ \|\boldsymbol{g}\| : \boldsymbol{g} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}})\ \}$;

but ε-subgradients also enjoy the so-called *information transport property*

$$\boldsymbol{g} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}}) \ \Rightarrow\ \boldsymbol{g} \in \partial_\mu\varphi(\bar{\boldsymbol{y}}) \quad \forall\ \boldsymbol{y}\ \forall\ \mu \ge \alpha = \varphi(\bar{\boldsymbol{y}}) + \boldsymbol{g}(\boldsymbol{y} - \bar{\boldsymbol{y}}) - \varphi(\boldsymbol{y}) + \varepsilon$$

saying, in plain words, that a vector $\boldsymbol{g}$ belonging to the ε-subdifferential of $\varphi()$ in some point $\bar{\boldsymbol{y}}$ also belongs to the α-subdifferential of $\varphi()$ in any other point $\boldsymbol{y}$, where $\alpha \ge 0$ (known as *linearization error*) can be calculated from $\boldsymbol{g}$ and $\varphi(\boldsymbol{y})$. Hence, the first-order information at hand about $\varphi()$ in $\bar{\boldsymbol{y}}$

can be "translated" into an information about a completely unrelated point $\boldsymbol{y}$, at the cost of "deteriorating" it of a known measure α. Therefore, one possibility is just to fix some $\varepsilon > 0$ and solve (Δ_β) restricted to the (known) ε-subgradients of the current point

$$\min \{ \| \boldsymbol{g} \| : \boldsymbol{g} \in \text{conv}(\{ \boldsymbol{g}_i : \alpha_i \leq \varepsilon , i \in \beta \}) \}$$

i.e. the steepest ε-ascent direction finding problem restricted to an inner linearization of the ε-subdifferential $\partial_\varepsilon \varphi(\bar{\boldsymbol{y}})$: the very first algorithms that were referred to as *Bundle methods* [Le75] [Le77] [BNS87] [HL93 Algorithm XIV.3.4.2] can be seen as a "smoothing" of the above reset rule, since they are based on the same ideas but on the slightly different (QP)

$$(\Delta_\beta^\varepsilon) \quad \min_\theta \{ \| \Sigma_{j \in \beta} \boldsymbol{g}_j \theta_j \| : \Sigma_{j \in \beta} \alpha_j \theta_j \leq \varepsilon , \theta \in \Theta \}$$

where, rather than eliminating all the $\boldsymbol{g}_i$ with $\alpha_i > \varepsilon$, the α_i are used as "weights" of the corresponding vector, and the constraint ensures that any $\boldsymbol{g}_i$ with a "large" α_i will have a "small influence" (multiplier θ_i) on the direction. Other than being heuristically clear, this approach has a neat theoretical justification [HL93 Example 3.5.3] since the ε-subdifferential of the polyhedral function

$$\varphi(\boldsymbol{y}) = \min_{i = 1 \,..\, l} \{ b_i + \boldsymbol{y} \boldsymbol{h}_i \}$$

is the polytope

$$\partial_\varepsilon \varphi(\bar{\boldsymbol{y}}) = \{ \boldsymbol{g} = \Sigma_{i = 1..l} \boldsymbol{h}_i \theta_i : \Sigma_{i = 1..l} \xi_i \theta_i \leq \varepsilon \}$$

where $\xi_i = \bar{\boldsymbol{y}} \boldsymbol{h}_i + b_i - \varphi(\bar{\boldsymbol{y}})$; therefore, if each $\boldsymbol{h}_i$ is read as the subgradient of $\varphi()$ in a point $\boldsymbol{y}_i$ such that $\varphi(\boldsymbol{y}_i) = b_i + \boldsymbol{y}_i \boldsymbol{h}_i$, ξ_i is exactly the linearization error α_i w.r.t. the current point $\bar{\boldsymbol{y}}$, i.e. $(\Delta_\beta^\varepsilon)$ corresponds to the *steepest ε-ascent direction finding problem* (at $\bar{\boldsymbol{y}}$) if the CP-model $\bar{\varphi}_\beta()$ is used in place of the real function $\varphi()$. The critical part of this approach is the introduction of the algorithmic parameter ε, that must be dynamically updated in order to achieve good practical performances: unfortunately, it is not clear what sort of rules should be devised for the task. Therefore, the more recent approaches [Ki90] [SZ92] [HL93 Algorithm XV.3.3.4] use the further "smoothing"

$$(\Delta_{\beta t}) \quad \min_\theta \{ 1/2 \, t \| \Sigma_{j \in \beta} \boldsymbol{g}_j \theta_j \|^2 + \Sigma_{j \in \beta} \alpha_j \theta_j : \theta \in \Theta \}$$

that may viewed as a Lagrangean relaxation of $(\Delta_\beta^\varepsilon)$ where a (fixed) Lagrangean multiplier $1/t$ is used to dualize the constraint $\Sigma_{j \in \beta} \alpha_j \theta_j \leq \varepsilon$; actually, such a relaxation would be

$$\min_\theta \{ 1/2 \| \Sigma_{j \in \beta} \boldsymbol{g}_j \theta_j \|^2 + (1/t) \Sigma_{j \in \beta} \alpha_j \theta_j : \theta \in \Theta \} \; [- \varepsilon t]$$

that differs from $(\Delta_{\beta t})$ for a multiplicative factor of t in the objective function, but the Khun-Tucker conditions easily show that the two problems are essentially the same, i.e. they have the same optimal primal solution θ and the respective dual optimal solutions $(v, \boldsymbol{d})$ are scaled of a factor of t. Historically, $(\Delta_\beta^\varepsilon)$ was considered the "basic" (QP) subproblem to be solved, and in fact it is possible to develop the whole theory of Bundle algorithms in terms of $(\Delta_\beta^\varepsilon)$ rather than of $(\Delta_{\beta t})$ [HL93]: this choice had a practical rationale, since the constrained least square problem $(\Delta_\beta^\varepsilon)$ was considered easier to solve than the general (QP) with one constraint $(\Delta_{\beta t})$ [Ki85a Introduction to Chapter 7] - in fact, ten years separate [Mi77] from [Ki86] where the first specialized algorithms for the two different problems have been proposed. As shown in Chapter I.3, this is not true any longer, and besides (Cf. §I.1.2) $(\Delta_\beta^\varepsilon)$ can be solved for any fixed ε by means of $(\Delta_{\beta t})$: hence, the development in terms of $(\Delta_{\beta t})$

seem to be preferable for several reasons, the main being that rules for choosing t are apparently easier to develop than rules for choosing ε. Such rules (the *t-strategies*), which are the critical part of any Bundle algorithm, will be extensively discussed in Chapter I.5: however, all of them are related to the *trust radius nature* of the parameter t, that becomes apparent by writing the Quadratic Dual of $(\Delta_{\beta t})$

$$(\Pi_{\beta t}) \qquad \max_{d,v} \{ v - 1/2t \| \boldsymbol{d} \|^2 : v \le \boldsymbol{d}\boldsymbol{g}_j + \alpha_j \; j \in \beta \}.$$

$(\Pi_{\beta t})$ is equivalent to $(\Delta_{\beta t})$ in the sense that, being θ the optimal solution of $(\Delta_{\beta t})$, the optimal solutions $\boldsymbol{d}$ and v of $(\Pi_{\beta t})$ can be computed as

$$\boldsymbol{d} = t\boldsymbol{z} \qquad \text{where} \qquad \boldsymbol{z} = \Sigma_{j \in \beta} \boldsymbol{g}_j \theta_j$$

$$v = t \| \boldsymbol{z} \|^2 + \sigma \qquad \text{where} \qquad \sigma = \Sigma_{j \in \beta} \alpha_j \theta_j$$

i.e. the new tentative point $\boldsymbol{y}$ is found along the *direction* $\boldsymbol{z}$ (a convex combination of the currently available subgradients), using t as *predefinite stepsize*: this naturally leads to the idea of *Bundle/Trust Region* approaches, where t is increased or decreased according to the outcome of the last (few) step(s). In case of a SS, increasing t may be preferred to moving to $\boldsymbol{y}$ whenever there is the chance that $\varphi_\beta()$ be a "good" approximation of $\varphi()$ also outside the current "trust region"; in case of a NS, decreasing t may be preferred to inserting the new $\boldsymbol{g}$ if the latter is not believed to carry "reliable information" about $\varphi()$ in a neighbourhood of $\bar{\boldsymbol{y}}$ (e.g. due to a large α). However, Bundle methods of this form are *not* pure Trust Region algorithms for several reasons, the simplest being that t is not a trust radius since $\boldsymbol{z}$ is not in principle of unitary norm, and $\| \boldsymbol{z} \|$ depends on t in a "complicated" way so that the *actual displacement* $t \| \boldsymbol{z} \|$ is only known *a posteriori*: furthermore, using a Line Search within such an approach is still possible, and may also be advantageous in some cases.

Interestingly enough the Conjugated Subgradient and the Cutting Plane algorithms can be obtained as the "extreme cases" of a Bundle method using $(\Delta_{\beta t})$: in fact, CP and CS are respectively the "limit for $t \to 0$" and the "limit for $t \to \infty$" of the above scheme, since for example the - possibly unbounded - optimal solution of (D_β) is just the limit for $t \to 0$ of the optimal solution of $(\Pi_{\beta t})$. A good way of seing this equivalence is rewriting $(\Pi_{\beta t})$ as

$$(\Pi_{\beta t}) \qquad \max_d \{ \varphi_\beta(\boldsymbol{d}) - 1/2t \| \boldsymbol{d} \|^2 \}$$

i.e. the same as (D_β) but for a simple quadratic *stabilizing term*: in fact, Bundle algorithms are often regarded to as "stabilized" versions of the CP algorithm that avoid its inherent "instability" problems, i.e. the fact that $\bar{\boldsymbol{y}} + \boldsymbol{d}^*$ can be very "far" from $\bar{\boldsymbol{y}}$, even if the latter is very near to the the optimum. The stabilizing term introduces a measure of distance from the *prox. center* $\bar{\boldsymbol{y}}$, discouraging directions leading to "far away" points: assuming that $\varphi_\beta()$ has been constructed as to be a "local" model of $\varphi()$, this can be interpreted as avoiding "far away" points where $\varphi_\beta()$ is probably a "bad" approximation of $\varphi()$.

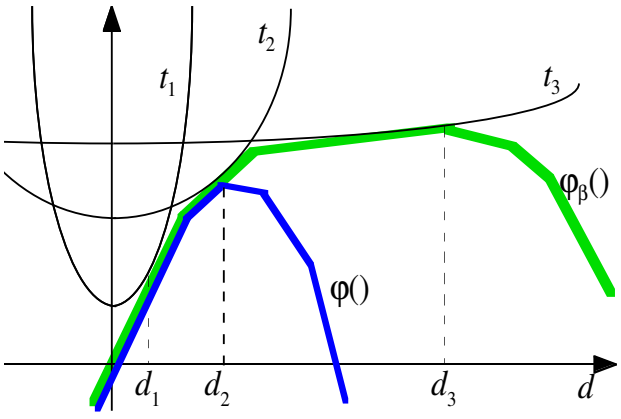

Figure 7: effect of the stabilizing term; for $t_1 < t_2 < t_3$, d_i are the optimal solutions

Again, t is a parameter that measures the "trust" in $\varphi_\beta()$ as a model of $\varphi()$ far from the current point $\bar{\boldsymbol{y}}$: the direction $\boldsymbol{d}$ is a "function" of t and, for sufficiently large values of t, $(\Pi_{\beta t})$ is equivalent to the "naive" (D_β) (Figure 7), provided that the latter problem has finite maximum.

Hence, Bundle methods belong to the class of *Proximal Point* approaches, originally introduced in [Ro76], that in the "modern" view are connected to the *Moreau-Yosida Regularization* of $\varphi()$

$$\phi_{M,t}(\boldsymbol{y}) = \max_{\boldsymbol{d}} \{ \varphi(\boldsymbol{y}+\boldsymbol{d}) - 1/2t \| \boldsymbol{d} \|_M^2 \}$$

where $\boldsymbol{M}$ is a positive semi-definite matrix intended to take the place of $\nabla^2\varphi()$ (whatever the Hessian of a nonsmooth function could mean); $\phi_{M,t}()$ is in some sense the "ideal" version of $\varphi()$, since it has the same set of minimizers but it is differentiable, so that a conceptually attractive way of maximizing $\varphi()$ would be that of applying any (efficient) smooth optimization algorithm to $\phi_{M,t}()$. Unfortunately, the computation of just a single value of $\phi_{M,t}(\boldsymbol{y})$ requires the solution of a NonDifferentiable Optimization problem, in terms of the original "black box", in principle as difficult as the overall maximization of $\varphi()$: however, it is quite easy to see (Cf. §I.2.2) that a sequence of consecutive Null Steps is just a Cutting Plane approach applied to the computation of $\phi_{M,t}(\bar{\boldsymbol{y}})$, so that the "bundling" mechanism provides a natural framework for algorithms based on these concepts.
In other words, any Bundle algorithm can be seen as a "two-level" approach to the maximization of $\phi_{M,t}()$, where sequences of consecutive NS aim at approximately compute $\phi_{M,t}(\bar{\boldsymbol{y}})$ while Serious Steps use the resulting information for the actual maximization: hence, the challenge is to adapt the efficient Quasi-Newton methods, that could be directly used if $\phi_{M,t}()$ were known exactly, to this new setting, where the smallest possible number of $\varphi()$ computations (NS) shall be used between two "upper level" moves (SS). Quite a lot of work is currently being done on this subject, as attested by several recent papers (e.g. [Mi82] [Mi91] [GM92] [BGL95] [FQ96] [LS96a] [LS96b] [LS96c] [Mi96] and many others, few of which already in the open literature): part of this renewed interest is due to some results that show how superlinearly convergent Quasi-Newton-type methods can be obtained in spite of the fact that $\phi_{M,t}()$ is *not twice continuously differentiable*. In fact, under some assumptions on the "original" function $\varphi()$, the directional derivatives of $\phi_{M,t}()$ can be shown to be *semismooth*, and this property can be enough for superlinear convergence to hold if the matrix $\boldsymbol{M}$ is properly handled; although some encouraging computational results have already been presented, the state of the art on this subject is very rapidly evolving, and a discussion of the related technical details will not be issued here.

Different forms of Proximal Point method has been recently proposed, where the quadratic proximal term is replaced by some different function, entropy-like [IST94] or even nonsmooth such as the L_1 or L_∞ norms [KCL95]: the latter is related to the new family of Bundle methods proposed in Chapter I.4, and will be discussed in some more details in §I.4.1.

Another recent entry in the family of Bundle methods is the *Proximal Level* method [HL93 Algorithm 2.3.1] [BKL95] [Ki95] [LNN95], that uses

$$(\Lambda_{\beta l}) \qquad \min \{ \| \boldsymbol{d} \|^2 : \varphi_\beta(\boldsymbol{d}) \geq l \}$$

as the direction finding problem, where l is the algorithmic parameter that plays a role similar to that of t: Proximal Level methods have been reported to be less sensitive to the choice of the l parameter than other forms of Bundle method be with theirs, but no decisive evidence of clear superiority seem to

exist yet.

Finally, a quite different form of Bundle approach uses the concept of *analytic centre* to locate the next trial point [GHV92] [RV92] [GHV93] [AV95] [Ne95] [AK96] [MGV96]: applied to the optimization of a concave function, the methods look at the polyhedron

$$P_{\bar{y},\beta} = \{ (v, \boldsymbol{y}) : \varphi_\beta(\boldsymbol{d}) \geq v \geq \varphi(\bar{y}) \} \subseteq \mathcal{R}^{m+1}$$

and compute its analytic centre, i.e. the point (π, ω) maximizing the product of the slacks from the boundaries, or equivalently the solution of

$$(A_{\bar{y},\beta}) \qquad \max_{d,v} \{ \Sigma_{j \in \beta} \ln(\boldsymbol{d}\boldsymbol{g}_j + \alpha_j - v) + \ln(v) : (v, \boldsymbol{d}) \in \mathcal{R}^{m+1} \}.$$

The analytic centre (π, ω) lies in the interior of $P_{\bar{y},\beta}$, and it is related in a well-known way to Interior Point methods for linearly constrained convex optimization [CLM93]: the ω component is then given to the "black box" to compute $\langle \varphi(\omega), \boldsymbol{g}(\omega) \rangle$, that can be interpreted as a separating hyperplane (a *cut*) in $\mathcal{R}^{m+1}$. Hence, these methods essentially keep "shrinking" $P_{\bar{y},\beta}$ by adding new cuts, and are in fact somehow better understood as algorithms for proving nonemptyness of a convex body given by a separation oracle: the properties of the analytic centre, together with a judicious choice of the cuts to be conserved, allow to derive polynomial bounds on the number of iterations required to reduce the volume (or the "width") of $P_{\bar{y},\beta}$ down to any required threshold, i.e. to achieve any required precision. Clearly, the solution of the $(A_{\bar{y},\beta})$ subproblem can be in principle a heavy task, but it is possible to exploit the analytic centre of the previous iterate to "warm-start" a Newton method, that can be shown to converge in a constant (and very small) number of iterations to a "close enough" approximation of the new analytic centre: the resulting overall complexity estimates are quite attractive, and the methods have an interesting behaviour in practice [GGS94], especially if "hybridized" with traditional CP techniques [MGV96] (Cf. Sec. II.1.4.2).

The remainder of this Chapter will essentially focus on "standard" Bundle algorithms, i.e. on methods that use subproblems (similar to) $(\Pi_{\beta t})$ and $(\Delta_{\beta t})$: these (QP)s have emerged from the very beginning [Le77] [Le89] [Le91] as interesting candidates for the direction finding subproblem within Bundle methods, probably due to the fact that they can be viewed from a surprisingly high number of different viewpoints, each one possibly suggesting new ways of approaching some of the open questions in the development of a good overall algorithm. In fact, a part of this work is exactly aimed at describing the interpretations of $(\Pi_{\beta t})$ and $(\Delta_{\beta t})$ when $\varphi()$ is a polyhedral function: taking into account all these viewpoints naturally leads to the definition of a new, broader class of Bundle methods and also to the development of new t management strategies.

However, the "dual" viewpoint of Bundle methods is very important for the understanding of the algorithms and for proving results: for instance, it is easy to show that, for any $\theta \in \Theta$, the vector z belongs to the σ-subdifferential of the current point $\bar{y}$, and therefore the relation

$$\| z \| = 0 \qquad \Rightarrow \qquad \bar{y} \text{ is } \sigma\text{-optimal}$$

can be easily understood and proved in terms of the above ε-subgradients theory. It also gives a rationale for the "standard" stopping conditions used within Bundle algorithms, i.e.

$$\| z \|^2 \leq \varepsilon_Q \qquad \text{and} \qquad \sigma \leq \varepsilon_L$$

where ε_Q and ε_L are two fixed tolerances: however, choosing a suitable value for ε_Q can be a delicate task, and a full understanding of the above stopping conditions requires some analysis. Let $\boldsymbol{y}^*$ be an optimal solution of (D): the "ideal" stopping criterion is

$$\varphi(\boldsymbol{y}^*) - \varphi(\bar{\boldsymbol{y}}) \leq \varepsilon$$

i.e. ε-optimality for a fixed absolute tolerance ε, but this is not an implementable rule since in general the value of $\varphi(\boldsymbol{y}^*)$ is unknown. Since $\boldsymbol{z} \in \partial_\sigma \varphi(\bar{\boldsymbol{y}})$, i.e. $\varphi(\boldsymbol{y}^*) \leq \varphi(\bar{\boldsymbol{y}}) + (\boldsymbol{y}^* - \bar{\boldsymbol{y}})\boldsymbol{z} + \sigma$,

$$\sigma + (\boldsymbol{y}^* - \bar{\boldsymbol{y}})\boldsymbol{z} \leq \varepsilon$$

would an even stronger stopping rule, if only $\boldsymbol{y}^*$ was known; analogously, by defining

$$t^*(\boldsymbol{y}, \boldsymbol{z}) = \| \boldsymbol{y}^* - \boldsymbol{y} \| / \| \boldsymbol{z} \|$$

(the distance from $\boldsymbol{y}$ and the optimum, taking $\| \boldsymbol{z} \|$ as the unitary distance), the even stronger

$$\sigma + t^*(\bar{\boldsymbol{y}}, \boldsymbol{z}) \| \boldsymbol{z} \|^2 \leq \varepsilon$$

could be turned into the "standard" form by letting $\varepsilon_L = \varepsilon / 2$ and choosing at each step some $\varepsilon_Q \leq \varepsilon_L / t^*(\bar{\boldsymbol{y}}, \boldsymbol{z})$, if the function $t^*()$ would be known. Since they all contains an unknown quantity, these formulae only give an idea on how ε_L and ε_Q should be set, saying that ε_Q should be related to both ε_L and to some measure of the distance to the optimum measured "along" $\boldsymbol{z}$: however, they suggest an alternative approach, i.e. to stop when

$$\sigma + t^* \| \boldsymbol{z} \|^2 \leq \varepsilon$$

where t^* is a user-provided fixed parameter, intended to serve as an upper approximation of the function $t^*()$. Usually, a viable such t^* can be easily choosen by observing a few runs of the algorithm, and seems to be fairly stable for problems within the same class: furthermore, this stopping criteria will show to be useful in Chapter I.2 when dealing with convergence. In practice, the *absolute* tolerance ε is better replaced by a *relative* tolerance, e.g. by using $\varepsilon | \varphi(\bar{\boldsymbol{y}}) |$ as the right hand side of the inequality, where $\varphi(\bar{\boldsymbol{y}})$ serves as an (usually pretty good) estimate of the magnitude of $\varphi(\boldsymbol{y}^*)$; the only event to take care of is $\varphi(\boldsymbol{y}^*) \approx \varphi(\bar{\boldsymbol{y}}) \approx 0$, that can be faced for instance by using $\max\{ | \varphi(\bar{\boldsymbol{y}}) |, 1 \}$ instead of $| \varphi(\bar{\boldsymbol{y}}) |$.

I.1.2 The Primal viewpoint

Part of the interest of NDO algorithms comes from the fact that they can be used as a tool for efficiently solving problems of an apparently very different nature: the typical example is

$$\text{(H)} \qquad \min\{ \boldsymbol{cx} : \boldsymbol{x} \in X, \boldsymbol{Ax} = \boldsymbol{b} \}$$

where $X \subseteq \mathcal{R}^n$ is a "simple" compact set and $\boldsymbol{Ax} = \boldsymbol{b}$ are "complicating" constraints, in the sense that

$$\text{(P}_\mu\text{)} \qquad \min\{ \mu\boldsymbol{x} : \boldsymbol{x} \in X \}$$

can be "efficiently" solved for any choice of μ while the whole (H) is significantly harder. In many approaches, (H) is "replaced" by the convex program

$$\text{(P)} \qquad \min\{ \boldsymbol{cx} : \boldsymbol{x} \in \overline{X}, \boldsymbol{Ax} = \boldsymbol{b} \}$$

where $\bar{X} = \text{conv}(\boldsymbol{x}_1, .., \boldsymbol{x}_p)$, each $\boldsymbol{x}_i$ being an extreme point of X; this general framework covers several different cases, such as

- X is a structured polyhedron itself, such as a flow and/or a separable polytope, so that (H) = (P) is a structured (LP) with "complicating" constraints;
- X is a discrete set, (P_μ) is a relaxation of (H) that can be efficiently solved (either to optimality or with an heuristic) and (P) is a relaxation of (H) that is used to derive lower bounds;
- X is a generic convex set that can be finitely inner approximated for the purpose of (approximately) solving (H);
- combinations of the above.

Following the classic Dantzig-Wolfe decomposition approach [DW60], the Master Problem is

$$(M) \qquad \min_\theta \{ \boldsymbol{c}(\Sigma_{j=1..p} \boldsymbol{x}_j\theta_j) : \boldsymbol{A}(\Sigma_{j=1..p} \boldsymbol{x}_j\theta_j) = \boldsymbol{b}, \theta \in \Theta \}$$

where Θ is the unitary simplex of proper dimension: (M) is obviously equivalent to (P), where each $\boldsymbol{x} \in \bar{X}$ is explicitly described as a convex combination of the *vertices* $\boldsymbol{x}_j$ using (non unique) *combinators* θ. The fundamental strategy of the DW algorithm is to solve (M) by generating only a(n hopefully small) subset of the vertices, so that from now on the subscript β will be meant to restrict something to the vertices in the subset β, as in

$$(M_\beta) \qquad \min_\theta \{ \boldsymbol{c}(\Sigma_{j\in\beta} \boldsymbol{x}_j\theta_j) : \boldsymbol{A}(\Sigma_{j\in\beta} \boldsymbol{x}_j\theta_j) = \boldsymbol{b}, \theta \in \Theta \} = \min_\theta \{ \boldsymbol{cx} : \boldsymbol{x} \in \bar{X}_\beta, \boldsymbol{Ax} = \boldsymbol{b} \}.$$

The linear dual problem of (M_β) is

$$(L_\beta) \qquad \max_{\boldsymbol{y},v} \{ \boldsymbol{yb} + v : v \le (\boldsymbol{c} - \boldsymbol{yA})\boldsymbol{x}_i \;\; i \in \beta \}$$

and its optimal solutions (v, $\boldsymbol{y}$) are used in the DW approach for the so called "mechanised pricing", i.e. for obtaining an optimal solution $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})$ to the *pricing subproblem*

$$(P_{\boldsymbol{c}-\boldsymbol{yA}}) \qquad \min_{\boldsymbol{x}} \{ (\boldsymbol{c} - \boldsymbol{yA})\boldsymbol{x} : \boldsymbol{x} \in X \} \; [+ \boldsymbol{yb}]$$

that gives a new "tentative optimal" vertex of $\bar{X}$: if the corresponding constraint $v \le (\boldsymbol{c} - \boldsymbol{yA})\boldsymbol{x}$ is satisfied, then (v, $\boldsymbol{y}$) are clearly dual optimal solutions for the whole (L) and $\bar{\boldsymbol{x}} = \Sigma_{j\in\beta} \boldsymbol{x}_j\theta_j$ (θ being the optimal solution of (M_β)) is optimal for (P), otherwise $\boldsymbol{x}$ is added to β and the process is repeated. It is well-known that the DW approach is exactly equivalent to the CP approach applied to the Lagrangean Dual of (P) with respect to the complicating constraints $\boldsymbol{Ax} = \boldsymbol{b}$: in fact, by letting $\varphi(\boldsymbol{y}) = z(P_{\boldsymbol{c}-\boldsymbol{yA}})$, (L) can be written as the unconstrained maximization of the concave polyhedral function $\varphi()$ and $\boldsymbol{g}(\bar{\boldsymbol{y}}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}(\bar{\boldsymbol{y}})$ is a subgradient for $\varphi()$ in $\bar{\boldsymbol{y}}$. Hence, the optimal dual prices $\boldsymbol{y}$ are used from a *primal viewpoint* to obtain a new $\boldsymbol{x}_i = \boldsymbol{x}(\boldsymbol{y}_i)$ to enrich the inner approximation $\bar{X}_\beta$, and from the *dual viewpoint* to obtain a new $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{y}_i)$ to enrich the CP-model $\varphi_\beta()$. For any dual vector $\boldsymbol{y}$, $\varphi(\boldsymbol{y})$ can be written as $\varphi(\boldsymbol{y}) = \boldsymbol{cx}(\boldsymbol{y}) + \boldsymbol{yg}(\boldsymbol{y})$, and the linearization errors α_j of $\varphi()$ w.r.t. any current point $\bar{\boldsymbol{y}}$ are

$$\alpha_j = \varphi(\boldsymbol{y}_j) + (\bar{\boldsymbol{y}} - \boldsymbol{y}_j)\boldsymbol{g}_j - \varphi(\bar{\boldsymbol{y}}) = (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})(\boldsymbol{x}_j - \boldsymbol{x}(\bar{\boldsymbol{y}}))$$
$$= (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\boldsymbol{x}_j - \varphi(\bar{\boldsymbol{y}}) + \bar{\boldsymbol{y}}\boldsymbol{b} = \boldsymbol{cx}_j + \bar{\boldsymbol{y}}\boldsymbol{g}_j - \varphi(\bar{\boldsymbol{y}})$$

so that, rewriting (L_β) as

(L_β) $\quad \max_{y,v} \{\ v : v \le (\ \boldsymbol{c} - \boldsymbol{yA}\ )\boldsymbol{x}_j + \boldsymbol{yb} - \varphi(\ \bar{\boldsymbol{y}}\ )\ \ j \in \beta\ \} + \varphi(\ \bar{\boldsymbol{y}}\ )$

the above relations on the α_i, the definition $\boldsymbol{y} = \bar{\boldsymbol{y}} + \boldsymbol{d}$ and some simple algebra yield

(L_β) $\quad \max_{y,v} \{\ v : v \le \boldsymbol{dg}_j + \alpha_j\ \ j \in \beta\ \} + \varphi(\ \bar{\boldsymbol{y}}\ )$

i.e. (L_β) and (D_β) are exactly the same problem modulo a proper translation w.r.t. the current point $\bar{\boldsymbol{y}}$, and (M_β) can be rewritten as

(M_β) $\quad \min_\theta \{\ \alpha\theta : \Sigma_{j \in \beta}\, \boldsymbol{g}_j\theta_j = \boldsymbol{0}\ ,\ \theta \in \Theta\ \}\ [\ + \varphi(\ \bar{\boldsymbol{y}}\ )\ ].$

If (M_β) has a nonempty feasible region, any feasible solution θ gives an upper bound $\sigma = \alpha\theta$ $[+ \varphi(\ \bar{\boldsymbol{y}}\ )]$ to the optimal value $\varphi(\ \boldsymbol{y}^*\ )$: therefore, (M_β) can be seen from the dual viewpoint as the problem of finding the best of such bounds. However, the relations between primal and dual optimality can be made more precise by means of the "sensitivity" (concave and nondifferentiable) function

$$\psi(\ \gamma\ ) = \min_x \{\ (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x} : \boldsymbol{x} \in X\ \} = z(\ (P_{c-\gamma})\ )$$

such that $\varphi(\ \boldsymbol{y}\ ) = \psi(\ \boldsymbol{yA}\ ) + \boldsymbol{yb}$: the following theorem gives a "dual flavour" even to the vectors $\boldsymbol{x}$ of the primal space, showing how [ε-]optimality for $(P_{c-\gamma})$ is equivalent to the condition of being an [ε-]subgradient for $\psi()$.

Theorem I.1.2.1: *$\boldsymbol{x}$ is an ε-optimal solution for $(P_{c-\gamma}) \Leftrightarrow -\boldsymbol{x}$ is an ε-subgradient for $\psi()$ in γ.*

Proof: [⇒] By ε-optimality of $\boldsymbol{x}$, $\psi(\ \gamma\ ) + \varepsilon \ge (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x}$: clearly, for any γ', $\boldsymbol{x}$ is in general suboptimal, i.e. $(\ \boldsymbol{c} - \gamma'\ )\boldsymbol{x} \ge \psi(\ \gamma'\ )$, hence

$$\psi(\ \gamma'\ ) \le (\ \boldsymbol{c} - \gamma'\ )\boldsymbol{x} + \gamma\boldsymbol{x} - \gamma\boldsymbol{x} = (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x} + (\ \gamma - \gamma'\ )\boldsymbol{x} \le \psi(\ \gamma\ ) + (\ \gamma' - \gamma\ )(\ -\boldsymbol{x}\ ) + \varepsilon.$$

[⇐] Conversely, suppose $\forall\ \gamma'\ \psi(\ \gamma'\ ) \le \psi(\ \gamma\ ) + (\ \gamma' - \gamma\ )(\ -\boldsymbol{x}\ ) + \varepsilon$ holds, but $\psi(\ \gamma\ ) < (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x} - \varepsilon$ (i.e. $\boldsymbol{x}$ is not ε-optimal): then, by taking $\gamma' = \boldsymbol{c}$ in the previous relation, we get

$$0 = \psi(\ \boldsymbol{c}\ ) \le \psi(\ \gamma\ ) + (\ \boldsymbol{c} - \gamma\ )(\ -\boldsymbol{x}\ ) + \varepsilon < [\ (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x} - \varepsilon\ ] - (\ \boldsymbol{c} - \gamma\ )\boldsymbol{x} + \varepsilon = 0. \qquad \blacklozenge$$

The linearization error $\alpha_i = \varphi(\ \boldsymbol{y}_i\ ) + \boldsymbol{g}_i(\ \bar{\boldsymbol{y}} - \boldsymbol{y}_i\ ) - \varphi(\ \bar{\boldsymbol{y}}\ )$ of the i-th subgradient $\boldsymbol{g}_i = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_i$ w.r.t. the current point $\bar{\boldsymbol{y}}$ is

$$\begin{aligned}
\alpha_i &= \varphi(\ \boldsymbol{y}_i\ ) + \boldsymbol{g}_i(\ \bar{\boldsymbol{y}} - \boldsymbol{y}_i\ ) - \varphi(\ \bar{\boldsymbol{y}}\ ) = \\
&= [\ \psi(\ \boldsymbol{y}_t\boldsymbol{A}\ ) + \boldsymbol{y}_t\boldsymbol{b}\ ] + (\ \bar{\boldsymbol{y}} - \boldsymbol{y}_i\ )(\ \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_i\ ) - [\ \psi(\ \bar{\boldsymbol{y}}\boldsymbol{A}\ ) + \bar{\boldsymbol{y}}\boldsymbol{b}\ ] = \\
&= \psi(\ \boldsymbol{y}_t\boldsymbol{A}\ ) + (\ -\boldsymbol{x}_i\ )(\ \bar{\boldsymbol{y}}\boldsymbol{A} - \boldsymbol{y}_t\boldsymbol{A}\ ) - \psi(\ \bar{\boldsymbol{y}}\boldsymbol{A}\ ) = \psi(\ \gamma_i\ ) + (\ \bar{\gamma} - \gamma_i\ )(\ -\boldsymbol{x}_i\ ) - \psi(\ \bar{\gamma}\ )
\end{aligned}$$

(where $\bar{\gamma} = \bar{\boldsymbol{y}}\boldsymbol{A}$ and $\gamma_i = \boldsymbol{y}_t\boldsymbol{A}$), i.e. it is as well the linearization error of $-\boldsymbol{x}_i$ (as a subgradient of $\psi()$ in γ_i) w.r.t. $\bar{\gamma}$: in other words, $\boldsymbol{x}_i$ is an α_i-*optimal solution* for the pricing subproblem $(P_{c-\bar{\gamma}})$ corresponding to the current point $\bar{\boldsymbol{y}}$. This immediately implies that, for any $\theta \in \Theta$, $\bar{\boldsymbol{x}} = \Sigma_{j \in \beta}\, \boldsymbol{x}_j\theta_j$ is a σ-subgradient of $\psi()$ in $\bar{\gamma}$, i.e. that it is σ-*optimal for* $(P_{c-\bar{\gamma}})$: hence, it is possible to prove in a much simpler and more general form two theorems that have been proposed [Me94] to measure the "quality" of the current point $\bar{\boldsymbol{y}}$.

Lemma I.1.2.2: *let* $\boldsymbol{x}^*$ *be the optimal solution for (M) and* $\boldsymbol{z} = \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}} = \Sigma_{i \in \beta} \boldsymbol{g}_i \theta_i$ *(for any* $\theta \in \Theta$*): hence,* $\boldsymbol{c}\boldsymbol{x}^* \geq \boldsymbol{c}\bar{\boldsymbol{x}} + \bar{\boldsymbol{y}}\boldsymbol{z} - \sigma$.

Proof: σ-optimality of $\bar{\boldsymbol{x}}$ implies $(\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\boldsymbol{x}^* + \sigma \geq (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\bar{\boldsymbol{x}}$, i.e.

$$\boldsymbol{c}\boldsymbol{x}^* \geq \boldsymbol{c}\bar{\boldsymbol{x}} + \bar{\boldsymbol{y}}(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{A}\bar{\boldsymbol{x}}) - \sigma$$

but $\boldsymbol{x}^*$ is feasible, i.e. $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{b}$, and this completes the proof. ♦

Lemma I.1.2.3: *for any* $\theta \in \Theta$, $\boldsymbol{c}\bar{\boldsymbol{x}} - \varphi(\bar{\boldsymbol{y}}) \leq \sigma - \bar{\boldsymbol{y}}\boldsymbol{z}$.

Proof: By σ-optimality of $\bar{\boldsymbol{x}}$, $\varphi(\bar{\boldsymbol{y}}) = \psi(\bar{\boldsymbol{y}}\boldsymbol{A}) + \bar{\boldsymbol{y}}\boldsymbol{b} \geq (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\bar{\boldsymbol{x}} - \sigma + \bar{\boldsymbol{y}}\boldsymbol{b}$. ♦

However, the above theorems are of little (if any) use in practice, and the stopping rules discussed in §I.1.1 are typically used instead; since $\boldsymbol{z} = \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}$ is σ-optimal for subproblem $(P_{\boldsymbol{c}-\bar{\boldsymbol{y}}\boldsymbol{A}})$, these rules are a generalization of the stopping condition of DW algorithm, since they hold when a point $\bar{\boldsymbol{x}}$ is found such that constraints $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ are *slightly violated* ($\| \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}} \|^2 \leq \varepsilon_Q$) and $\bar{\boldsymbol{x}}$ is *almost* optimal for the current pricing subproblem. This finally gives a clear interpretation to ε_Q as a measure of the maximum violation of the $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ constraint that can be accepted in the final *primal* solution reported by the algorithm: however, when the stress is more on the *dual* side, i.e. on the determination of the optimal value $\varphi(\boldsymbol{y}^*)$, this interpretation is not of much use in guessing a correct ε_Q, and the alternative stopping rule presented at the end of §I.1.1 can be more effective.

A *primal interpretation* of $(\Delta_{\beta t})$ can be obtained from

(M_β^ε) $\quad \min_\theta \{ \alpha\theta : \| \boldsymbol{G}_\beta\theta \|^2 \leq \varepsilon , \theta \in \Theta \}$,

where $\boldsymbol{G}_\beta$ is the matrix having $\{ \boldsymbol{g}_i : i \in \beta \}$ as columns, which is the nonlinear relaxation of (M_β) where the constraints $\boldsymbol{A}(\Sigma_{j \in \beta} \boldsymbol{x}_j\theta_j) = \boldsymbol{b}$ are relaxed in a *Least-Square fashion* to $\| \boldsymbol{A}(\Sigma_{i \in \beta} \boldsymbol{x}_i\theta_i) - \boldsymbol{b} \|^2 \leq \varepsilon$: $(\Delta_{\beta t})$ is a Lagrangean relaxation (with Lagrangean multiplier t) of (M_β^ε). Obviously, (M_β^0) is equivalent to (M_β), since their objective functions and feasible regions are exactly the same: in general, $\{ (M_\beta^\varepsilon) \}$ is a family of relaxations of (M_β) that gets weaker as ε increases. Even from a practical point of view, using (M_β^ε) instead of (M_β) may be reasonable, since it means extending the linear (affine) constraint region $\{ \boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \}$ to a cylinder of ray $(2\varepsilon)^{1/2}$ around it, and it is always possible to find a large enough ε which makes (M_β^ε) feasible for every nonempty β (Figure 8), while (M_β) can stay unbounded until a large number of vertices is added: in this sense, the former can be viewed as a "stabilized" version of the latter, that might help in enhancing the poor practical performances of the classical DW method, caused by (the primal interpretation of) the instability problems of the CP algorithm.

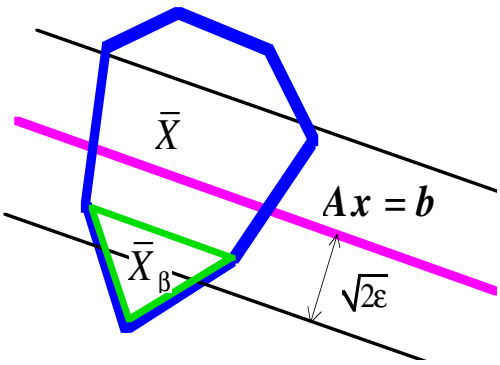

Figure 8: (M_β) can be empty while (M_β^ε) it is not

It is important to remark that the *same* problem $(\Delta_{\beta t})$ can be seen as a Lagrangean relaxation of the *two different* problems (M_β^ε) and $(\Delta_\beta^\varepsilon)$, with Lagrange multiplier t in the first case and $(1 / t)$ in the second: the constraint relaxed from $(\Delta_\beta^\varepsilon)$ is aimed to ensure that all the information used to compute the

direction is reliable, i.e. (almost) in the ε-subdifferential, while the one relaxed from (M_β^ε) is aimed to ensure (almost) feasibility, so that it is possible to see t as a "knob" that can be tuned to privilege either the dual aspect (searching for an ascent direction) or the primal one (searching for a feasible primal solution). This latter interpretation of t, that will be used later to justify new t-strategies in Chapter I.5, can be "verified" by referring to both the viewpoints: from the dual side, t switches from a "local" to a "global" view of the function, that is if $\boldsymbol{d}$ has to be an ascent direction in a small neighbourhood of $\bar{\boldsymbol{y}}$, or rather if $\boldsymbol{d}$ has to point towards the maximum of $\varphi_\beta()$. However, finding a feasible $\bar{\boldsymbol{x}}$ means in fact finding a null subgradient in the σ-subdifferential of the current point, i.e. *demonstrating global* σ*-optimality*: there is a clear difference (typical of many NDO methods) between *reaching* an ε-optimal solution and *demonstrating* its ε-optimality, but both conditions are required to stop, so that t can be seen as a "relative weight" given to the decrease of $\| \boldsymbol{G}_\beta \theta \|^2$ w.r.t. the increase of $\varphi(\bar{\boldsymbol{y}})$. Furthermore, the role of t can be interpreted in exactly the same way by only using primal concepts, which can be clearly seen by considering that in practical DW codes the computation is usually divided in a *Phase* 0 and a *Phase* 1, where the first apply the DW technique to the feasibility (linear) problem

$$(M^p) \qquad \min_\theta \{ \| \boldsymbol{A}(\Sigma_{j \in \beta} \boldsymbol{x}_j \theta_j) - \boldsymbol{b} \|_p : \theta \in \Theta \} \qquad p \in \{1, \infty\}$$

in order to generate a nonempty (M_β), and the second actually optimizes the objective function $\boldsymbol{c}\boldsymbol{x}$ (= $\alpha\theta$); if using $p = 2$ instead, t in ($\Delta_{\beta t}$) can be interpreted as a weight given to the "Phase 0 portion" of a composite objective function trying to balance between the two different requirements.

The above observation suggests the possibility of a "purely primal" Bundle algorithm, closely resembling the DW approach, that be constructed using the (QP) subproblem ($\Delta_{\beta t}$) rather than the (LP) subproblem (M_β): it would not make any use of the typical concepts (such as the Null Step - Serious Step) of "dual" Bundle methods, and even completely disregards the value of the (dual) function $\varphi()$. Although such an approach is probably inefficient in practice, it is a nice illustration of how the conceptual machinery of Bundle methods can be used from a completely different point of view: furthermore, some ideas introduced in the discussion will be useful later on. The basic observation is that any (efficient) algorithm that solves ($\Delta_{\beta t}$) can be used to approach the quadratically constrained problem (M_β^ε): in fact, being a pointwise minimum of an infinite set of linear functions, the univariate function $\delta_\beta(t) = z(\Delta_{\beta t})$ is concave, so that the Lagrangean Dual

$$(T_\beta^\varepsilon) \qquad \max_{t \geq 0} \{ \delta_\beta(t) - \varepsilon t \}$$

can be solved by means of a suitable line search procedure on t; (T_β^ε) is also well known to be equivalent to (M_β^ε), so that the former has finite optimum if and only if the latter is nonempty. More in detail, the nonnegative halfline $\mathcal{R}_+$ can be subdivided into a finite number of intervals $[\, t_j \,,\, t_{j+1} \,]$ (each one being the stability interval of an optimal *base* of ($\Delta_{\beta t}$), Cf. §I.3.6) such that for each $t \in [\, t_j \,,\, t_{j+1} \,]$ the optimal solutions of ($\Delta_{\beta t}$) and ($\Pi_{\beta t}$) are completely described by three numbers $s_j \geq 0$, q_j and $r_j \geq 0$ and two vectors θ_j' and θ_j'', in the sense that $\theta(t) = \theta_j' + (1/t)\theta_j''$ is the optimal solution of ($\Delta_{\beta t}$) with

$$\| \boldsymbol{G}_\beta \theta(t) \|^2 = \| z(t) \|^2 = s_j + (1/t)^2 r_j \qquad \text{and} \qquad \alpha\theta(t) = q_j - (1/t) r_j \qquad (1)$$

while for ($\Pi_{\beta t}$)

$$\| \boldsymbol{d}(t) \|^2 = \| tz(t) \|^2 = t^2 s_j + r_j \qquad \text{and} \qquad v(t) = t s_j + q_j \qquad (2)$$

so that

$$\delta_j(t) = v(t) - {}^1/_2 t \left\| \boldsymbol{d}(t) \right\|^2 = {}^1/_2 [\, t s_j + 2q_j - (1/t) r_j \,] = ({}^1/_2) t \left\| \boldsymbol{G}_\beta \theta(t) \right\|^2 + \alpha\theta(t) \qquad (3)$$

is equal to $\delta_\beta(t)$ for each $t \in [\, t_j , t_{j+1} \,]$. Being the sum of the linear function $({}^1/_2) t s_j + 2q_j$ and of the concave function $-({}^1/_{2t}) r_j$, $\delta_j(t)$ is a (strictly for $r_j > 0$) smooth concave upper approximation of $\delta_\beta(t)$, while the latter is a piecewise concave possibly nonsmooth function (see Figure 9): therefore

$$\delta_j'(t) = {}^1/_2 (s_j + (1/t)^2 r_j) = {}^1/_2 \left\| z(t) \right\|^2 \geq 0$$

is the derivative of $\delta_\beta()$ for any $t \in (\, t_j , t_{j+1} \,)$ and its left and right derivative in the (resp. right and left) extremes of the interval, so that $\delta_\beta()$ is a non decreasing function. If $\delta_\beta()$ is differentiable in the optimal point $\bar{t}$, then $\left\| z(\bar{t}) \right\|^2 = 2\varepsilon$; otherwise, $\bar{t} = t_i$ for some i, and two equivalent optimal solutions $\theta' = \theta_j(\bar{t})$ and $\theta'' = \theta_{j+1}(\bar{t})$ of $(\Delta_{\beta\bar{t}})$ exist such that

$$\varepsilon \geq {}^1/_2 \left\| \boldsymbol{G}_\beta \theta' \right\|^2 \quad \text{and} \quad {}^1/_2 \left\| \boldsymbol{G}_\beta \theta'' \right\|^2 \leq \varepsilon$$

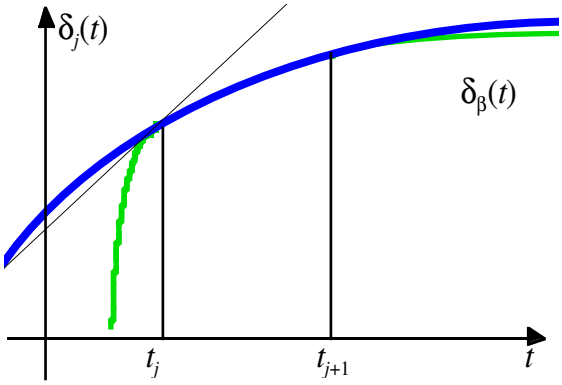

Figure 9: $\delta^+(t)$ is a smooth concave upper approximation of the nonsmooth $\delta_\beta(t)$

where at least one of the two inequalities strictly holds: hence, $\mu\theta' + (1-\mu)\theta''$ is still an optimal solution of $(\Delta_{\beta\bar{t}})$ for each $\mu \in [0, 1]$, and it is easy to find a $\bar{\mu}$ such that

$$\left\| \boldsymbol{G}_\beta [\, \bar{\mu}\theta' + (1-\bar{\mu})\theta'' \,] \right\|^2 = \left\| \boldsymbol{G}_\beta \theta(\bar{t}) \right\|^2 = \left\| z(\bar{t}) \right\|^2 = 2\varepsilon$$

(as $\mu \left\| \theta' \right\|^2 + (1-\mu) \left\| \theta'' \right\|^2 \geq \left\| \mu\theta' + (1-\mu)\theta'' \right\|^2$ for each μ). In other words, $\bar{t}$ is the smallest value of t, if any, for which an optimal solution of $(\Delta_{\beta t})$ exists (and can be constructively found) that is feasible for (M_β^ε): conversely, the existence of an $\bar{\boldsymbol{x}} \in \bar{X}_\beta$ such that $\left\| \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b} \right\|^2 \leq \varepsilon$ is equivalent to the existence of a $\bar{t}$ such that the derivative of $\delta_\beta(t) - \varepsilon t$ in $\bar{t}$ is ≤ 0. All the above facts lead to the definition of a Line Search procedure that can be sketched as follows: for given ε, β and t, $(\Delta_{\beta t})$ is solved and its sensitivity analysis parameters are used to find the new t or to detect unboundedness as

- if $s_j < \varepsilon$, then $t' = (r_j / (\varepsilon - s_j))^{1/2}$ is a good value to test: if $t' \in [\, t_j , t_{j+1} \,]$ then it is optimal and explicit formulae for obtaining the optimal θ, v ... exist, otherwise a new interval is visited (the piecewise nature of the function clearly ensures finite termination);
- if $s_j \geq \varepsilon$, then $\left\| z(t) \right\|^2 > 2\varepsilon$ for each $t \in [\, t_j , t_{j+1} \,]$ and some $t' > t_{j+1}$ must be tried in order to see whether a suitable interval exists: if $t_{j+1} = +\infty$ than (T_β^ε) is unbounded, and the current value of t and θ can be used in order to obtain new information.

Hence, in a finite number of steps either the optimal $\bar{t}$ (and the associated $\theta(\bar{t})$, $v(\bar{t})$ ...) is found, or unboundedness is detected; interestingly enough, this is *not* a LS in the space of the dual variables $\boldsymbol{y}$, since $\varphi()$ is *not* evaluated on the corresponding dual points $\{ \boldsymbol{y}_h \}$ (the interest is rather in what the current model says), although one could imagine a procedure that does: moreover, the points $\boldsymbol{y}_h$ do not lie along the same line, so that it would eventually result in a *Curve-Search* approach analogous to the Bundle Trust-Region algorithm of §I.5.2.

A possible algorithm is stated in Figure 10, and its proof of convergence is simple and can be carried

along without ever mentioning the "dual" side. When the new $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{y})$ gives no new information, i.e. $v \le \alpha + \boldsymbol{g}\boldsymbol{y}$, the current value of $\delta_\beta(t)$ is equal to $\delta(t)$ [$= z(\Delta_t)$], since $(v, \boldsymbol{y})$ solves (Π_t): in fact, being $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y})$ the solution of the pricing subproblem, one has

$$(\boldsymbol{c} - \boldsymbol{y}\boldsymbol{A})\boldsymbol{x} \le (\boldsymbol{c} - \boldsymbol{y}\boldsymbol{A})\boldsymbol{x}' \qquad \forall\, \boldsymbol{x}' \in X \supseteq \bar{X}$$

i.e. $$v \le \alpha + \boldsymbol{y}\boldsymbol{g} = \boldsymbol{c}\boldsymbol{x} + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \le \boldsymbol{c}\boldsymbol{x}' + \boldsymbol{y}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}') = \alpha' + \boldsymbol{y}\boldsymbol{g}' \quad \forall\, \boldsymbol{g}'$$

so that $(v, \boldsymbol{y})$ is feasible for (Π_t), $\theta(t)$ is feasible for (Δ_t) and complementary slackness hold.

Stabilized Dantzig-Wolfe

```
t = t0; ε = ε0; β = { g( 0 ) };                    /* initialization */
do
   ( Bounded , t , v,  y ) = SolveT( ε , β , t );  /* the above line search on t */
   g = g( y ); α = cx( y );                        /* solve the pricing subproblem */
   if( v > α + gy )                                /* "new information" obtained */
         β = β ∪ { g };                            /* exploit it (enlarge X̄β) */
   else  if( Bounded )                             /* (Mβε) equivalent to (Mε) */
               ε = sj + ( 1 / tj+1 )²rj;           /* increase the required primal accuracy */
         else  t = 2t;
while( ε > 0 );
```

Figure 10: a "stabilized" Dantzig-Wolfe-type algorithm

Therefore, provided that the original problem (P) has a nonempty feasible region, for any fixed ε a "Bounded" condition must be obtained in a finite number of iterations: in fact, at each step either β is enlarged or t is augmented, but there cannot be infinitely many consecutive t augments as this would imply that $\delta()$ goes to infinity (contradicting the boundedness of (T^0)), and for any $\varepsilon \ge 0$ a finite β exists such that (M^ε_β) is nonempty and (T^ε_β) is bounded. The latter argument also shows that there cannot be infinitely many ε decreases, since when β contains enough vertices it exists a t such that $\| z(t) \|^2 = s_j + (1/t)^2 r_j = 0$ ($\Rightarrow s_j = r_j = 0$) that must be at some point reached (as the number of intervals is finite and does not depend on ε) forcing ε to become 0; some subtle details, as how t_{j+1} has to be defined when t is a kink point of $\delta_\beta()$, can be omitted. Now, if the current t solves (T^ε_β), and $v \le \alpha + \boldsymbol{y}\boldsymbol{g}$, then $\theta(t)$ is optimal for the whole (M^ε), since it solves $(\Delta_{\bar t})$ and satisfies $\| \boldsymbol{G}_\beta \theta(\bar t) \|^2 = 2\varepsilon$, so that it also solves (T^ε), that is known to be equivalent to (M^ε): therefore, when the algorithm stops (with $\varepsilon = 0$) the obtained solution is optimal to (M^0), hence to (P).

Note that several questions of relevant importance in practice have not been addressed, such as the "β-strategy" (how and when to eliminate items from β in order to keep it to a reasonable size) and the possibility of changing the current point from $\mathbf{0}$ to any general $\bar{\boldsymbol{y}}$, that can be obviously embedded in this framework: the interest was in making apparent the deep relations between Bundle and Dantzig-Wolfe methods, and to clarify the primal role of the subproblem $(\Delta_{\beta t})$. In the next Chapter, however, some of ideas introduced above will be used in the convergence proof of a very general dual Bundle algorithm.