I.2 Convergence Properties

In this Chapter, an unified convergence analysis will be developed to cover a large family of Bundle algorithms using direction finding subproblems ($\Delta_{\beta t}$): the basic algorithm, quite similar to the "generic" one presented in §I.1.1, is described in Figure 11.

Bundle Algorithm

⟨ let $t^* > 0$ and $\varepsilon > 0$ be fixed; choose the initial $\bar{y}$, t and β ⟩;
do
⟨ solve ($\Pi_{\beta t}$) and ($\Delta_{\beta t}$) for (v , $\boldsymbol{d}$) and (σ , $\| z \|$) ⟩; /* find a direction */
⟨ move along $\boldsymbol{d}$, generating some new $\boldsymbol{g}_i$ and a trial point $\boldsymbol{y}$ ⟩;
if(a large enough improvement has been obtained)
then $\bar{y} = y$; /* a Serious Step */
⟨ update β, (possibly) including some new $\boldsymbol{g}_i$, deleting old ones and/or
adding convex combination of the present ones (typically z) ⟩; /* the β-strategy */
⟨ update t, depending on the previous iterations ⟩; /* the t-strategy */
while($\sigma + t^* \| z \|^2 > \varepsilon$);

Figure 11: the Bundle Algorithm

Despite its generality, it does not cover all the Bundle-like methods proposed in the literature, most notably

- Bundle methods for nonconvex nondifferentiable optimization [Ki91b], or for problems with "complicated" constraints [Ki91a];
- methods that try to extract some sort of second order information about the curvature of the function, for instance by using a variable matrix-induced norm $\|\cdot\|_M$ for the "stabilizing" term of ($\Pi_{\beta t}$) [Mi82] [Mi91] [GM92] [BGL95] [FQ96] [LS96a] [LS96b] [LS96c] [Mi96] or in some different ways [GM91]: while it seems to be hardly possible to define a proper concept of curvature for polyhedral functions, complicating the structure of the quadratic subproblems may hide some relevant detail of our interpretation;
- methods that use different forms of direction finding subproblems, like Proximal Level algorithms [BKL95] [Ki95] [LNN95].

The proof technique adopted in this paragraph is quite standard: convergence is shown in two steps, making it apparent the "two-level" nature of these algorithms, where sequences of NS and SS "aims"

to different targets; hence, after some preliminaries presented in §I.2.1, the two §I.2.2 and §I.2.3 will be devoted to the separate analysis of the two "levels" of the proof. The main original contribution of our analysis is that, rather than proving the convergence of a single particular algorithm, it focuses on the general properties that any Bundle approach must have: moreover, the proof techniques have been chosen in order to be easily extendible to the new family of (LP)-based Bundle methods described in Chapter I.4. Such an extension might be possible for Proximal Level or Proximal Point algorithms also, and it is currently under research.

I.2.1 Preliminaries and Convergence Conditions

It is possible to prove convergence of any algorithm fitting in the above framework at the only condition that the following five general properties are respected:

(i) $\bar{t} \geq t_i \geq \underline{t} > 0$ for all iterations i and fixed $\bar{t}$, $\underline{t}$.

(ii) If a SS is performed moving to $\boldsymbol{y}$, then $\varphi(\boldsymbol{y}) - \varphi(\bar{\boldsymbol{y}}) \geq m_1 v$ for a fixed $m_1 \in (0, 1)$: the converse is *not* required, in the sense that even when a "considerable" is obtained, there is the choice of not accepting the step, but for what is required by *(iv)* below.

(iii) In case that the algorithm performs an *infinite* sequence of *consecutive* NS, t has to be *finitely convergent*, i.e. for some h it must be $t_i = t_h > 0$ for all $i \geq h$.

(iv) In the case of *(iii)*, at any iteration $i \geq h$ where $\varphi(\boldsymbol{y}_i) - \varphi(\bar{\boldsymbol{y}}_i) \geq m_1 v$ a SS must be performed; in other words, after *finitely many* NS *any* step that improves at least of $m_1 v_i$ must be accepted.

(v) In the case of *(iii)*, at any iteration $i \geq h$ the point $\bar{\boldsymbol{y}}_i + \boldsymbol{d}_i$ must always be among those that are probed by the line search; furthermore, the corresponding subgradient must be added to β.

A brief initial discussion of the above conditions is in order: *(i)* is necessary only for the proof of convergence in the limit, but it also simplifies somewhat the proofs, and its possible relaxations will be discussed each time it will be used. Although from a theoretical viewpoint *(i)* is a rather strong condition, in practice some "external" knowledge about the proper range of values for t should be given to the algorithm anyway, so that it is not difficult to choose a small $\underline{t}$ and a large $\bar{t}$ that does not interfere with the t-strategy; besides, $\bar{t}$ and $\underline{t}$ *are* ultimately fixed by the finite machine precision of any reasonable implementation. The other conditions just describe the basic ideas of the algorithm: at each step, either a "consistent" increase is obtained by performing a SS, or new information enriches the Bundle; to gain flexibility, however, the algorithm is allowed to "to resist the temptation" of moving to a better point if increasing t seems to be preferable *(ii)*, but it is obviously required that this does not happen forever *(iv)*. Analogously, conditions *(iii)* and *(v)* allow for trying to adjust t in response to a "bad" step, but prevent this adjustment phase to impair the overall convergence of the algorithm. Note that the first Bundle methods proposed in the literature used ad-hoc Line Search procedures to ensure that, in case of a NS, the new $(\boldsymbol{g}_i, \alpha_i)$ gives "reliable" information, i.e. that α_i is "small": all the recent methods are designed to visit $\bar{\boldsymbol{y}} + \boldsymbol{d}$ only, since this is generally preferable in terms of performances, but a LS may still be useful under some circumstances, and the convergence analysis is just not affected at all if *(v)* holds.

An important detail of any implementable Bundle method is the management of β (β-*strategy*): in general, at any step the following two possibilities are given for changing β

- some $\boldsymbol{g} \in \beta$ is removed and / or
- given some $\theta \in \Theta$ [usually the latest optimal solution of $(\Delta_{\beta t})$],
  $(\boldsymbol{g}', \alpha') = (\boldsymbol{G}_\beta\theta, \alpha_\beta\theta)$ [$= (\boldsymbol{z}, \sigma)$] is added to β.

Each of these "moves" may, both in theory and in practice, impair finite termination, or even convergence, of the algorithm: hence, other than the rules *(i)* - *(v)*, a (theoretically) convergent method must also respect some restriction on the β-strategy, that will be extensively discussed in the following. The development of suitable assumptions on the β-strategy is a crucial part of the convergence proofs: in fact, this is perhaps the point where the results of this Chapter differ more from the extensions of Chapter I.4.

In the convergence proof, it will be important to distinguish whether, for a given β, the Cutting Plane model $\varphi_\beta()$ is an upper bounded function (briefly, is bounded) or not: the following observation will be repeatedly useful.

Observation I.2.1.1: $\varphi_\beta()$ *is bounded* $\Leftrightarrow \exists\ \theta \in \Theta$ *such that* $\boldsymbol{G}_\beta\theta = \boldsymbol{0}$.

Proof: $\boldsymbol{G}_\beta\theta = \boldsymbol{0}$ ensures that (M_β) has a feasible solution, therefore its dual (L_β) (equivalent to the maximization of $\varphi_\beta()$) cannot be unbounded; conversely, if $\varphi_\beta()$ is bounded then for some sufficiently large ε ($\geq \max_{j \in \beta} \{ \alpha_j \}$) one has $\boldsymbol{0} \in \partial_\varepsilon\varphi(\bar{\boldsymbol{y}})$. ♦

The first thing to prove is that if the algorithm stops, then the provided solution is "satisfactory": in this general setting, this statement takes the form

Theorem I.2.1.2: *if the stopping condition* $\sigma + t^*\| z \|^2 \leq \varepsilon$ *holds, and* $\| \boldsymbol{G}\theta^* \| = 0$ *for any optimal solution* θ^* *of* (Δ_{t^*})*, then* $\varphi(\boldsymbol{y}^*) - \varphi(\bar{\boldsymbol{y}}) = z(\Delta_{t^*}) \leq \varepsilon$.

Proof: First of all, note that θ^* is required to be an optimal solution of (Δ_{t^*}), i.e. perfect knowledge of the problem is assumed in the "neighbourhood" of $\bar{\boldsymbol{y}}$ determined by t^*: for polyhedral functions, this condition can actually be finitely checked. By the formulae (1), (2) and (3) in §I.1.2, $\| \boldsymbol{G}\theta^* \| = 0$ implies $v^* = \varphi(\bar{\boldsymbol{y}} + \boldsymbol{d}^*) = z(\Pi_{t^*}) = z(\Delta_{t^*}) = \alpha\theta^*$; furthermore, in this case θ^* is clearly an optimal solution to (Δ_t) for any $t \geq t^*$, since $z(\Delta_t) \geq z(\Delta_{t^*})$ (being $t\| \boldsymbol{G}\theta \|^2 + \alpha\theta$ a nondecreasing function in t). But $z(\Delta_t) \leq t\cdot 0 + \alpha\theta^* = z(\Delta_{t^*})$, since θ^* is a feasible solution for (Δ_t): this just means that the CP model is bounded, so that $\varphi(\bar{\boldsymbol{y}}) + \varphi_\beta(\boldsymbol{d}^*) = \varphi(\bar{\boldsymbol{y}} + \boldsymbol{d}^*)$ is at the same time an upper bound and a lower bound on $\varphi(\boldsymbol{y}^*)$, hence $\boldsymbol{y}^* = \bar{\boldsymbol{y}} + \boldsymbol{d}^*$ is an optimal solution for (D). Now, since clearly $\sigma + t^*\| z \|^2 \geq z(\Delta_{t^*})$, it is possible to conclude

$$\varphi(\boldsymbol{y}^*) - \varphi(\bar{\boldsymbol{y}}) = \varphi_\beta(\boldsymbol{d}) = z(\Delta_{t^*}) \leq \sigma + t^*\| z \|^2 \leq \varepsilon. \qquad ♦$$

As usual in Bundle methods, the stopping condition is completely satisfactory only if $\| z \| = 0$ (which has to be true for the particular case of $\varepsilon = 0$), since then the above theorem applies and $\boldsymbol{y}^*$ is guaranteed to be an ε-optimal solution; in all the other cases, no real estimate in the "quality" of the accepted solution $\bar{\boldsymbol{y}}$ is available, even though in practice guessing a value of t^* that produces a "true" ε-optimal solution is fairly easy.

Following the typical proof pattern for this class of algorithms, the rest of convergence proof will be divided into two parts: in the first it is assumed that no Serious Steps occur, i.e. we examine sequences of consecutive Null Steps, so that in the second it is possible to disregard what happens between two successive SSs, essentially reducing a run of the algorithm to its Serious Steps alone. The rationale for this proof structure is that *a SS destroys the local view* of the function on which the NS idea is based, making it hard to "collate the pieces" *before* and *after* a SS: this can be made more precise by referring to the Moreau-Yosida Regularization of $\varphi()$ or to its "primal counterpart", as shown in the next paragraph.

I.2.2 Null Steps convergence

This part of the convergence proof will mainly use the dual view of the algorithm, where $(\Pi_{\beta t})$ is the "restriction" to the CP-model $\varphi_\beta()$ of

$$(\Pi_{\bar{y},t}) \qquad \phi_t(\bar{y}) = \max_d \{ \varphi(\bar{y} + d) - 1/2t \| d \|^2 \}$$

i.e. the Moreau-Yosida Regularization of $\varphi()$ (Cf. §I.1.1): the function $\phi_t()$ has many nice properties [HL93 - §XV.4], such as of being differentiable even if $\varphi()$ is not, and that

$$\mathbf{0} \text{ is optimal for } (\Pi_{\bar{y},t}) \quad \Rightarrow \quad \phi_t(\bar{y}) = \varphi(\bar{y}) \quad \Rightarrow \quad \bar{y} \text{ is optimal for (D)}$$

since $\partial[\, \varphi(\cdot) - 1/2t \|\cdot\|^2 \,](\mathbf{0}) = \partial\varphi(\mathbf{0})$, so that the only way for $d = \mathbf{0}$ to be optimal for $(\Pi_{\bar{y},t})$ is that $\mathbf{0} \in \partial\varphi(\bar{y})$, i.e. $\bar{y}$ be optimal for (D). However, the "primal counterparts" of the some of the arguments will be analysed, both for their intrinsic interest and for setting up the notations for §II.4.2: the main observation is that $(\Delta_{\beta t})$ can be viewed as the *restriction to* $\overline{X}_\beta$ of

$$(P_{\bar{y},t}) \qquad \min_x \{ f_{\bar{y},t}(x) = (c - \bar{y}A)x + 1/2t \| b - Ax \|^2 : x \in \overline{X} \} \; [\, + \bar{y}b \,]$$

that is an *Augmented Lagrangean* relaxation of (P), where a linear penalty term $\bar{y}(b - Ax)$ (with Lagrangean multipliers $\bar{y}$) and a quadratic one $1/2t \| b - Ax \|^2$ (with Lagrangean multiplier t) are added to the original cost function cx. Another way of looking at $(P_{\bar{y},t})$ is as the pricing subproblem $(P_{c-\bar{y}A})$ with an extra "stabilizing" quadratic term that makes it more difficult to solve, justifying some sort of *inner linearization* (otherwise called *Simplicial Decomposition* - see II.1.4.2) approach: in fact, this paragraph will be essentially devoted to showing that a *sequence of consecutive NS* is nothing but an inner linearization approach applied to $(P_{\bar{y},t})$ problem. Analogously to what already did for α, in the following the unnecessary references to $\bar{y}$ (that however is fixed during a sequence of NS) will be dropped: hence, until the end of this paragraph 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 v_i, t_i, f_i ... will denote the corresponding values of v, t, $f_t(\bar{x})$ ...

If an infinite sequence of NS is performed, the sequence of the optimal objective function values converges to $z(P_t)$ (for some t that may not be known in advance): the proof makes use of a "compactness" result contained in the following observation.

Observation I.2.2.1: *if an infinite sequence of NS is performed, all the "representative"* y_i *generated belong to a proper ball B centered on* $\bar{y}$.

Proof: According to *(v)*, a "representative" point $\boldsymbol{y}_i = \bar{\boldsymbol{y}} + \boldsymbol{d}_i$ is generated at each iteration, and, in view of *(iii)*, it is possible to disregard whatever had happened before iteration h, when t gets fixed (an alternative statement would be that once a ball containing all the $\boldsymbol{y}_i$, $i \geq h$, has been found, it can always be finitely enlarged to contain the finitely many $\boldsymbol{y}_i$ for $i < h$). Now, it is easy to see that

$$1/2t\| z_i \|^2 \leq 1/2t\| z_i \|^2 + \sigma_i \leq 1/2t\| z_h \|^2 + \sigma_h \qquad \forall\, i \geq h$$

but $\boldsymbol{d}_i = t\boldsymbol{z}_i$, and this completes the proof. ♦

Now, the main result can be proved: an infinite sequence of NS makes $z(\Delta_{\beta t})$ to converge to $z(P_t)$ (for the t to which $\{ t_i \}$ finitely converges) even under the "minimal" conceivable form of β-strategy, where at each step the Bundle only contains the *aggregate subgradient* $\boldsymbol{z}$ (that belongs to the σ-subdifferential of φ() in $\bar{\boldsymbol{y}}$); note how a Bundle method using such a β-strategy closely resembles some of the extended subgradient methods that have been proposed in the literature [Br95].

Theorem I.2.2.2: *if an infinite sequence of NS is performed,* $m_1 < 1$ *and at each step* β *is replaced with the set* $\{ (\boldsymbol{z}, \sigma) \}$*, then* $\{ v_i \} \to 0$ *and* $\varphi(\bar{\boldsymbol{y}}) = \phi(\bar{\boldsymbol{y}}) = z(P)$.

Proof: As in Observation I.2.2.1, everything that has happened before iteration h - that of *(iii)* - is disregarded, so that $t = t_i = t_h$ and the iteration index $i \geq h$ can be dropped: hence, $(\Delta_{\beta t})$ is reduced to the following one-dimensional (QP)

$$\min_\gamma \left\{ h(\gamma) = 1/2t\| \gamma\boldsymbol{z} + (1-\gamma)\boldsymbol{g} \|^2 + \gamma\sigma + (1-\gamma)\alpha : \gamma \in [0, 1] \right\}$$

whose optimality conditions can be algebraically determined. Some boring calculations (or the help of an algebraic package) show that

$$h'(\gamma) = t(\boldsymbol{z} - \boldsymbol{g})(\gamma(\boldsymbol{z} - \boldsymbol{g}) + \boldsymbol{g}) + \sigma - \alpha$$

so that $h'(1) = (t\| \boldsymbol{z} \|^2 + \sigma) - (t\boldsymbol{z}\boldsymbol{g} + \alpha)$ and $\gamma^* = \dfrac{t\boldsymbol{g}(\boldsymbol{z} - \boldsymbol{g}) + \alpha - \sigma}{t\| \boldsymbol{z} - \boldsymbol{g} \|^2}$

where γ^* is the optimal solution disregarding the constraints $\gamma \in [0, 1]$, hence

$$h(1) - h(\gamma^*) = \frac{(t(\| \boldsymbol{z} \|^2 - \boldsymbol{z}\boldsymbol{g}) + \sigma - \alpha)^2}{t\| \boldsymbol{z} - \boldsymbol{g} \|^2} = \frac{h'(1)^2}{t\| \boldsymbol{z} - \boldsymbol{g} \|^2}.$$

Since *(iv)* guarantees $t\boldsymbol{g}\boldsymbol{z} + \alpha = \varphi(\boldsymbol{y}) - \varphi(\bar{\boldsymbol{y}}) < m_1 v = m_1(t\| \boldsymbol{z} \|^2 + \sigma)$, the derivative of $h()$ in 1 is strictly positive, hence $z(\Delta_{\beta t})$ will strictly decrease: in fact, $h'(1) \geq (1 - m_1)(t\| \boldsymbol{z} \|^2 + \sigma) = (1 - m_1)h(1) = (1 - m_1)z(\Delta_{\beta t})$. If $\gamma^* \geq 0$, this suffices to show that the $z(\Delta_{\beta t})$ will (sublinearly, with squared root rate) converge to zero: hence $\{ v_i \}$ will also do, because Observation I.2.2.1 guarantees that $\boldsymbol{y}$ belongs to a compact set B, hence $t\| \boldsymbol{z} - \boldsymbol{g} \|^2$ is upper bonded (φ() is a Lipschitz function). Conversely, if $\gamma^* < 0$ then $h'(0) > 0$, i.e. $\gamma = 0$ is the optimal solution, and since

$$t\boldsymbol{g}\boldsymbol{z} + \alpha < m_1(t\| \boldsymbol{z} \|^2 + \sigma) \quad \text{and} \quad t\boldsymbol{g}\boldsymbol{z} + \sigma > t\| \boldsymbol{g} \|^2 + \alpha$$

just summing up and dividing by 2 gives

$$h(0) = 1/2t\| \boldsymbol{g} \|^2 + \alpha < 1/2m_1 t\| \boldsymbol{z} \|^2 + \frac{1 + m_1}{2}\sigma$$

that (since $m_1 \leq (1 + m_1) / 2 \leq 1$ for each $m_1 \in [0, 1]$) finally allows one to conclude

$$h(0) < \frac{1 + m_1}{2}(t\| z \|^2 + \sigma) = \frac{1 + m_1}{2} h(1)$$

i.e. linear convergence rate of $z(\Delta_{\beta t})$ (hence of v) to 0; clearly, this implies that $\boldsymbol{d} = \boldsymbol{0}$ is optimal for (Π_t), i.e. that $\phi(\bar{\boldsymbol{y}}) = \varphi(\bar{\boldsymbol{y}})$ and therefore that $\bar{\boldsymbol{y}}$ is optimal for (D). ♦

A form of this result was first proved in [Ki85a], with the extra assumption (deriving from the particular form of Bundle algorithm used) that $\alpha_i = 0$; another related result can be found in [SZ92], but the actual form of Theorem I.2.2.2 is new.

A primal interpretation easily follows, because the "aggregated gradient" $\boldsymbol{z}$ corresponds to a "current primal point" $\bar{\boldsymbol{x}}$, and it is not hard to see that $(\boldsymbol{c} - (\bar{\boldsymbol{y}} + \boldsymbol{d})\boldsymbol{A})$ is the gradient of $f_t()$ at $\bar{\boldsymbol{x}}$: hence, the pricing subproblem $(P_{\boldsymbol{c}-(\bar{\boldsymbol{y}}+\boldsymbol{d})\boldsymbol{A}})$ is nothing but the *linearization* of (P_t) using as objective function $\nabla f_t(\bar{\boldsymbol{x}})\boldsymbol{x}$, i.e. a first-order model of $f_t()$ at $\bar{\boldsymbol{x}}$. The new "representative" vertex $\boldsymbol{x}$ corresponding to $\boldsymbol{g}$ is the solution of the linearized subproblem, that is used to find a descent direction; in fact, by using

$$\alpha = (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\boldsymbol{x} - \varphi(\bar{\boldsymbol{y}}) + \bar{\boldsymbol{y}}\boldsymbol{b} \Rightarrow \sigma = (\boldsymbol{c} - \bar{\boldsymbol{y}}\boldsymbol{A})\bar{\boldsymbol{x}} - \varphi(\bar{\boldsymbol{y}}) + \bar{\boldsymbol{y}}\boldsymbol{b} \qquad \text{and}$$

$$\boldsymbol{g} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \Rightarrow \boldsymbol{z} = \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}$$

it is easy to verify that the univariate (QP) in the proof of Theorem I.2.2.2 is

$$\min_\gamma \{ f_t(\gamma\bar{\boldsymbol{x}} + (1 - \gamma)\boldsymbol{x}) : \gamma \in [0, 1] \}$$

i.e. the exact minimization of $f_t()$ along the direction $\boldsymbol{x} - \bar{\boldsymbol{x}}$.

The convergence proof clearly motivates the definition

Definition I.2.2.3: *a* β*-strategy is* ***monotone*** *if, in the case of (iii), for each* $i \geq h$ *either all the* g_j *such that* $\theta_j > 0$ *are kept in* β*, or* (z_i , σ_i) *is added to* β*.*

It is clear that the thesis of Theorem I.2.2.2 follows *a fortiori* by the fact that a monotone β-strategy is used, since in this case the univariate (QP) in the proof of the Theorem is a restriction of $(\Delta_{\beta t})$.

I.2.3 Serious Steps convergence

In this paragraph, all that happens between two consecutive SS is disregarded: therefore, for ease of notation, v_i, t_i ... will rather denote the values of v, t ... at the i-th Serious Step. A weak finite termination proof, requiring $\varepsilon > 0$, is clearly at hand.

Theorem I.2.3.1: *if* $\varphi()$ *is bounded, then the algorithm finitely terminates for any fixed* $\varepsilon > 0$.

Proof: The above discussion has already ruled out the possibility on an infinite sequence of consecutive NS, ensuring that either a SS is performed, or the stopping condition $\sigma + t^*\| z \|^2 \leq \varepsilon$ is met: then, at the i-th SS a lower bound on v_i is available, since $\sigma_i + t^*\| z_i \|^2 > \varepsilon \Rightarrow v_i \geq \varepsilon$ if $t_i \geq t^*$, while if $t_i < t^*$

$$\sigma_i + t^*\| z_i \|_p > \varepsilon \quad \Rightarrow \quad \| z_i \|_p > (\varepsilon - \sigma_i) / t^* \quad \Rightarrow \quad v_i > t_i [(\varepsilon - \sigma_i) / t^*] + \sigma_i$$

$$\Rightarrow \quad v_i > \varepsilon(t_i / t^*) \quad (t_i / t^* < 1).$$

But *(ii)* implies that $\Delta\varphi_i \geq m_1 v_i \Rightarrow \varphi(\bar{\boldsymbol{y}}_i) \geq \varphi(\boldsymbol{y}_0) + m_1(\Sigma_{j<i} v_j) \geq \varphi(\boldsymbol{y}_0) + (m_1\varepsilon / t^*)\Sigma_{j<i} t_j$, and since *(i)* implies that $\Sigma_{j=1..\infty} t_j = \infty$, if infinitely many SS are performed then $\{ \varphi(\bar{\boldsymbol{y}}_i) \} \to \infty$. If a finite upper bound on the optimal value of $\varphi()$ were known, it would be possible to terminate in a finite number of steps, detecting unboundedness of (D), i.e. emptiness of the feasible set of (P). ♦

Here, $t_i \geq \underline{t}$ of *(i)* can be relaxed to the "classical" $\Sigma_{i=1..\infty} t_i = \infty$ (involving only the value of t during SS), eventually allowing $\{ t_i \} \to 0$: also, $\bar{t}$ has never been mentioned yet. Note that the "rate of convergence" that could be desumed from the theorem is adversely affected by a small t: this makes only a little sense, since a too large t may cause a large number of NS to be performed before any SS, but the computational experience seem to show that one of the main practical issues of Bundle algorithms (at least when applied to polyhedral functions) is actually to avoid t getting too small, how we will discuss in details in the next paragraph.

It is now time to turn to the subtler case of convergence in the limit sense, corresponding to a run of the algorithm with $\varepsilon = 0$: the (apparently new) proof presented here is based on an idea of [HL93 Theorem 2.1.5], and has been chosen for "compatibility" with the results of Chapter I.4.

Theorem I.2.3.2: *the sequence* $\{ \varphi(\bar{\boldsymbol{y}}_i) \}$ *converges to a maximizer* φ^* *of* $\varphi()$.

Proof: being a strictly increasing sequence, $\{ \varphi(\bar{\boldsymbol{y}}_i) \}$ converges to some φ^*: if $\varphi^* = \infty$ then it is a maximizer of $\varphi()$ and there is nothing to prove, while if the algorithm finitely terminates then $\varphi^* < \infty$ and the obtained solution is optimal (Theorem I.2.1.2). So, assume that $\varphi^* < \infty$ and the algorithm never stops: it must either perform an infinite sequence of SS or an infinite sequence of consecutive NS, but in the latter case the point $\boldsymbol{y}^*$ reached by the last SS must be optimal (Theorem I.2.2.2).

Hence, assume that infinitely many SS are performed, but that there exist a point $\hat{\boldsymbol{y}}$ and a $\delta > 0$ such that $\varphi(\hat{\boldsymbol{y}}) \geq \varphi(\bar{\boldsymbol{y}}_i) + \delta$ for all i: we claim that, on the limit, $\hat{\boldsymbol{y}}$ must be "farther" from $\boldsymbol{y}_i$ than $\boldsymbol{y}_{i+1}$, i.e. outside the level set $B_i = \{ \boldsymbol{y} : \| \boldsymbol{y} - \bar{\boldsymbol{y}}_i \| = \| \boldsymbol{d}_i \| \}$. Formally, being $\rho_i = \| \boldsymbol{d}_i \|$ the radius of B_i and $\omega_i = \| \boldsymbol{w}_i \| = \| \hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}_i \|$ the distance of $\hat{\boldsymbol{y}}$ from $\bar{\boldsymbol{y}}_i$, the aim is to establish that, for any sufficiently large i, $\omega_i > \rho_i$. As in the proof of Theorem I.2.3.1, it is easy to see that it must be $\Sigma_{j=i..\infty} v_j < \infty$, for otherwise $\varphi^* = \infty$: but ω_i must be bounded away from zero, for otherwise a subsequence of the $\{ \bar{\boldsymbol{y}}_i \}$ would converge to $\hat{\boldsymbol{y}}$ ($\Rightarrow \varphi(\hat{\boldsymbol{y}}) \leq \varphi^*$), while on the contrary ρ_i must converge to zero fast enough, for otherwise $\Sigma_{j=i..\infty} v_j = \infty$ as $v_i \geq t_i \| \boldsymbol{z}_i \|^2 = \| \boldsymbol{d}_i \|^2 / t_i = (\rho_i)^2 / t_i$ and t_i is upper bounded. Now, let $\boldsymbol{r}_i$ be the intersection between the frontier of B_i and $\mathrm{conv}\{ \hat{\boldsymbol{y}}, \bar{\boldsymbol{y}}_i \}$: one has

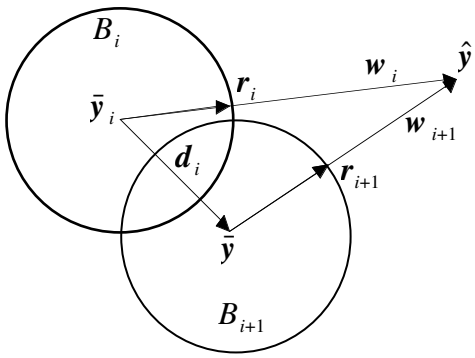

Figure 12: the induction step

- $\| \boldsymbol{r}_i - \bar{\boldsymbol{y}}_i \| = \rho_i < \omega_i$;
- $\varphi(\bar{\boldsymbol{y}}_i) + v_i = \bar{\varphi}_\beta(\bar{\boldsymbol{y}}_{i+1}) \geq \bar{\varphi}_\beta(\boldsymbol{r}_i) \geq \varphi(\boldsymbol{r}_i)$;
- since $\boldsymbol{r}_i = \bar{\boldsymbol{y}}_i + (\rho_i / \omega_i)\boldsymbol{w}_i = (1 - \rho_i / \omega_i)\bar{\boldsymbol{y}}_i + (\rho_i / \omega_i)\hat{\boldsymbol{y}}$ and $\varphi()$ is concave

$$\varphi(\boldsymbol{r}_i) \geq \left(1 - \frac{\rho_i}{\omega_i}\right)\varphi(\bar{\boldsymbol{y}}_i) + \frac{\rho_i}{\omega_i}(\varphi(\bar{\boldsymbol{y}}_i) + \delta) \Rightarrow v_i \geq \delta\frac{\rho_i}{\omega_i}.$$

Reasoning about σ_i as we did for ρ_i, it is easy to show that, for a sufficiently large i, z_i is a σ_i-subgradient of $\varphi()$ in $\bar{y}_i$, hence by the subgradient inequality

$$d_i w_i = t_i z_i (\hat{y} - \bar{y}_i) \geq t_i (\delta - \sigma_i) > t_i (\delta / 2)$$

and this is used twice: the first is to obtain $\rho_i \omega_i > t_i (\delta / 2)$, hence

$$\infty > \Sigma_{j = i..\infty} v_j > \delta \Sigma_{j = i..\infty} \frac{\rho_j}{\omega_j} > \frac{\delta^2}{2} \Sigma_{j = i..\infty} \frac{t_j}{\omega_j^2}$$

and the second is, together with $d_i + w_{i+1} = w_i$ (Figure 12) and the triangle inequality, to obtain

$$\| w_{i+1} \|^2 = \| w_i \|^2 + \| d_i \|^2 - 2 w_i d_i \quad \Rightarrow \quad \omega_{i+1}^2 = \omega_i^2 + \rho_i^2 - t_i \delta < \omega_i^2 + \rho_i^2 .$$

This can be used recursively in the previous relation, together with $t_i \geq \underline{t}$, to obtain

$$\infty > \frac{\delta^2}{2} \Sigma_{j = i..\infty} \frac{t_j}{\omega_j^2} > \underline{t} \frac{\delta^2}{2} \left(\frac{1}{\omega_i^2} + \frac{1}{\omega_i^2 + \rho_i^2} + \frac{1}{\omega_i^2 + \rho_i^2 + \rho_{i+1},^2} + \ldots \right) >$$

$$> \underline{t} \delta \Sigma_{j = i..\infty} \frac{1}{\omega_i^2 + (j - i) \bar{\rho}}$$

that yields the desired contradiction, since the rightmost series is divergent[Δ]; in the formula, $\bar{\rho}$ is any fixed number larger than any ρ_j for $j \geq i$, that exists since $\{ \rho_i \} \rightarrow 0$. ♦

The above proof is not very strong - namely, convergence is only shown for the value of the objective function - especially if compared e.g. to that of [CL93]; however, it has the advantage of being easily extendible to the larger class Bundle methods presented in Chapter I.4, which is *not* true for all the "standard" proof in the literature. Furthermore, an even stronger property can be proved for polyhedral functions: in this case, finding a "true" optimal solution (i.e. not just optimal within some $\varepsilon > 0$) in a finite number of steps is possible (while it is in general hopeless if $\varphi()$ is not polyhedral) provided that the "dangerous" subgradient aggregations are properly limited.

Definition I.2.3.3: *a* β*-strategy is* ***safe*** *if*

- *the total number of aggregations* (*adding* (z_i , σ_i) *to* β) *is finite and*
- *if some* g_j *has been removed from* β *after a SS, no other removal is permitted unless* $\bar{\varphi}_\beta (\bar{y}_k) > \bar{\varphi}_\beta (\bar{y}_i)$*, where* $k < i$ *is the latest iteration where a removal took place.*

This definition is made by two parts, whose role will be made clear by the proof of the following theorem: the desired result also requires the following very simple observation

---

[Δ] The n-th term of the series is (but for a constant term, proportional to) the *digamma* function $\psi()$ evaluated in n, i.e. the logarithmic derivative $\psi(n) = \Gamma'(n) / \Gamma(n)$ of Euler's $\Gamma()$ function: although unbounded, $\psi()$ grows very slowly, making it difficult to prove convergence with anything but a lower bounded t. For instance, if a t proportional to $1 / n$ is used instead, the corresponding series is not divergent any more, although this is not necessarly true for the series *prior* the minorizations.

Observation I.2.3.4: *let* Γ *be a finite set of vectors* $\{ \boldsymbol{g}_1 \dots \boldsymbol{g}_q \}$*: there exists a* $\mu > 0$ *depending only on* Γ *such that, for any* $\beta \subset \Gamma$,

$$\exists\, \theta \in \Theta \text{ such that } \| \boldsymbol{G}_\beta \theta \| < \mu \Rightarrow \exists\, \theta' \in \Theta \text{ such that } \boldsymbol{G}_\beta \theta' = \boldsymbol{0}.$$

Proof: Just denote $\mu(\beta) = \min_\theta \{ \| \boldsymbol{G}_\beta \theta \| : \theta \in \Theta \}$ and let $\underline{\Gamma} = \{ \beta : \mu(\beta) > 0 \} \subseteq \Gamma$: since $\underline{\Gamma}$ is a *finite* set, $\mu = \min\{ \mu(\beta) : \beta \in \underline{\Gamma} \} > 0$, and any β such that $\| \boldsymbol{G}_\beta \theta \| < \mu$ for some $\theta \in \Theta$ (if any) must belong to $\Gamma / \underline{\Gamma}$ - the fact that $\| \boldsymbol{v} \| = 0 \Leftrightarrow \boldsymbol{v} = \boldsymbol{0}$ does the rest. ♦

In view of Observation I.2.1.1, this Observation says that if there exists a small-norm vector in the convex hull of the $\boldsymbol{g}_j \in \beta$, then $\varphi_\beta()$ is bounded.

Theorem I.2.3.5: *let* $\varphi()$ *be a polyhedral function: if a safe* β*-strategy is used, the algorithm run with* $\varepsilon = 0$ *will finitely converge to a pair of respectively primal and dual optimal solutions* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *if and only if* $\varphi()$ *is bounded.*

Proof: We already know that if $\varphi^* = \infty$ then the algorithm will never stop (unless some extra stopping criteria are used), and if the algorithm stops at some iteration h then it has reached the optimum: hence, suppose that $\varphi^* < \infty$ but the algorithm runs forever. As in the proof of Theorem I.2.3.2, it is easy to obtain that $\forall\, \xi > 0\ \exists\, h$ s.t. $\| \boldsymbol{G}_\beta \theta_i \| < \xi\ \forall\ i > h$ (for otherwise $\Sigma_{j = 1..\infty}\, v_j = \infty$, since $v_i \geq \underline{t} \| z_i \|^2$): hence, starting from some step h, the CP-model will always be bounded. In fact, the set Γ_i of all the subgradients generated so far is finite, and Observation I.2.3.4 gives, at each iteration, a lower bound μ_i on the values of $\| \boldsymbol{G}_\beta \theta_i \|$ for which the CP-model can *not* be bounded: although the sequence $\{ \mu_i \}$ can be decreasing, the first part of the definition of a *safe* β-strategy exactly serves to ensure that it will be finitely convergent to some $\bar{\mu} > 0$ (although $\liminf \mu_i = \bar{\mu}$ would suffice), that can be used as the ξ above.
It is now possible to show that, at length, *all the directions* $\boldsymbol{d}$ *are maximizers of the corresponding* $\varphi_\beta()$: in fact, (part of) the Complementary Slackness Conditions for $(\Pi_{\beta t})$ and $(\Delta_{\beta t})$ are

$$(\, \boldsymbol{d}\boldsymbol{G}_\beta + \alpha_\beta - v\boldsymbol{e}\,)\theta = 0$$

so that, calling $B = \{ h : \theta_h > 0 \}$ and $\Gamma_B = \{ \boldsymbol{g}_h : h \in B \}$, one has

$$\varphi_\beta(\boldsymbol{d}) = v = \boldsymbol{d}\boldsymbol{g}_h + \alpha_h \qquad \forall\, h \in B$$

i.e. *every* $\boldsymbol{g}_h$ for $h \in B$ is a 0-*subgradient of* $\varphi_\beta()$ (although not necessarily of $\varphi()$) in $\boldsymbol{d}$. Furthermore, since $\theta_h = 0$ for all $h \notin B$, $\| \Sigma_{i \in B}\, \boldsymbol{g}_i \theta_i \| = \| z \| < \bar{\mu}$ and hence Observation I.2.3.4 can be applied again to Γ_B, yielding a $\theta' \in \Theta$ such that $\boldsymbol{G}_\beta \theta' = \boldsymbol{0}$, i.e. showing that $\boldsymbol{0} \in \text{conv}\{ \Gamma_B \} \subseteq \partial\varphi_\beta(\boldsymbol{d}) \Rightarrow$ the optimal solution of $(\Pi_{\beta t})$ is a maximizer of $\varphi_\beta()$.
Now, the (finite) maximizer φ^* of $\varphi()$ is by definition the minimal among the set of all the maximizers $\bar{\varphi}^*_\beta$ of $\bar{\varphi}_\beta()$, but the *safe* β-strategy ensures that this is a *finite* set: hence, as in the above Observation, $\min\{ \bar{\varphi}^*_\beta - \varphi^* : \bar{\varphi}^*_\beta > \varphi^* \} = \bar{\eta} > 0$. But it is also true that $\{ v_i \} \to 0$, i.e. that

$$\bar{\varphi}^*_\beta - \bar{\varphi}(\bar{\boldsymbol{y}}) \geq \bar{\varphi}^*_\beta - \varphi^*$$

will become smaller than any fixed threshold: hence, we can conclude that *starting from some iteration h, the value* $\bar{\varphi}^*_\beta$ *predicted by the CP model will be exactly the optimal value* φ^*, or in other words that the sequence of the $\bar{\varphi}^*_\beta$ finitely converges to φ^*. Unfortunately, without other assumptions

on $\varphi()$, such as *uniqueness* of the optimal solution, this does not tell us anything about the actual value of $\varphi(\bar{\boldsymbol{y}})$: however, here is where the second part of the definition of a safe β-strategy enters to play, ensuring that after a finite number of steps a complete description of the optimal level set of $\varphi()$ is obtained, hence terminating with an optimal solution. ♦

Note that the second part of the Definition I.2.3.3 kills any possibility of working with a β of bounded dimension: eventually, in the last steps nothing better than accumulating all the subgradients can be done. Alternatively, one could require properties that (typically) guarantee uniqueness of the optimal solution, such as the *Haar constraint qualification* that is postulated in [Ki85a], but these properties are unlikely to hold (and, even in case, difficult to be proved) for practical problems.

It may be instructive to terminate this paragraph by pointing out two other possible schemes that guarantee convergence to a "true optimum" even in presence of a stopping parameter $\varepsilon > 0$: the first, directly based onto Theorem I.2.2.2, is quite obvious and it is described in Figure 13.

```
⟨ run a standard Bundle algorithm until σ + t*‖ z ‖² ≤ εφ(ȳ) ⟩; t = t*;
do
    ⟨ solve (Δ_t) and (Π_t), and let θ, (v, d) be the optimal solutions ⟩;
    if( ‖ Gθ ‖ = 0 )
      then ⟨ y* = ȳ + d is optimal ⟩; Stop;
      else t = 2·t;
while( ~ Stop );
```

Figure 13: a "dual-primal" exact convergence scheme

The "post-processing" of the "ε-optimal" solution found by the standard algorithm, i.e. the exact solution of (Δ_t), can be done by just inhibiting the SS mechanism and performing a sequence of Null Steps with t set to t^*: the t increasing mechanism accounts for the possibility of a wrong choice of t^*. Such a scheme will finitely converge to a true optimum if $\varphi()$ is a polyhedral function and a monotone β-strategy that is *also safe* (in the sense of limiting the total number of aggregations) is used: it will converge in the limit to a maximizer only if t^* has been properly chosen (the t increasing mechanism can enter into the play only after termination), but it will do it for any nondifferentiable function. Interestingly, nothing is assumed about ε, i.e. this process can be started however far from $\boldsymbol{y}^*$: this divides the Bundle algorithm into two parts, the first (*dual*) responsible for obtaining a first approximately optimal point and the second (*primal*) devoted to "refine" this approximation to the true optimum. This is pretty much the converse of a *Dantzig-Wolfe approach* using the *Phase* 0 */ Phase* 1 *mechanism*, where a bounded Cutting Plane model is constructed first by disregarding the value of $\varphi()$, and only subsequently used to optimize the actual function: this idea will be used in the next paragraph to justify some of the new t-strategies that we propose.

Finally, an entirely different approach to "exact" termination with $\varepsilon > 0$ (assuming a polyhedral $\varphi()$ and a *safe* monotone β-strategy) is just to set $m_1 = 1$, i.e. to avoid Serious Steps unless $\varphi_\beta(\boldsymbol{d}) = \varphi(\bar{\boldsymbol{y}} + \boldsymbol{d})$ $[= \phi_t(\bar{\boldsymbol{y}}) + {}^1\!/_{2t}\| \boldsymbol{d} \|^2]$: this is probably *not* a good idea in practice, but it raises the interesting and yet unanswered question of whether *dynamic adjustments* of the m_1 parameter could be useful in practice, that may deserve some research.