Journal of Convex Analysis
Volume 30 (2023), No. 2, 453–479

A Bundle-Like Progressive Hedging Algorithm

Felipe Atenas, Claudia Sagastizábal
IMECC/UNICAMP, Campinas, São Paulo, Brazil
f262900@dac.unicamp.br, sagastiz@unicamp.br

Dedicated to Roger J-B Wets on the occasion of his 85th birthday.

Received: July 23, 2022
Accepted: November 17, 2022

For convex multistage programming problems, we propose a variant for the Progressive Hedging algorithm inspired from bundle methods. Like in the original algorithm, iterates are generated by first solving separate problems for each scenario, and then performing a projective step to ensure non-anticipativity. An additional test checks the quality of the approximation, splitting iterates into two subsequences, akin to the dichotomy between bundle serious and null steps. The method is shown to converge in both cases, and the convergence rate is linear for the serious subsequence. Our bundle-like approach endows the Progressive Hedging algorithm with an implementable stopping test. Moreover, it is possible to vary the augmentation parameter along iterations without impairing convergence. Such enhancements with respect to the original Progressive Hedging algorithm are obtained at the expense of the solution of additional subproblems at each iteration, one per scenario.

Keywords: Bundle methods, progressive Hedging algorithm, projective descent schemes, Douglas-Rachford splitting.

2010 Mathematics Subject Classification: 90C30, 90C33, 90C55, 65K05.

1. Introduction and motivation

Multistage stochastic programs represent an important source of large-scale optimization problems. This is because the number of variables and constraints needed to formulate the problem grows with the number of scenarios. As illustrated by energy applications in [33], decomposition methods are essential for solving effectively this type of problems; see also [37].

A very popular approach that deals with scenario decomposition is the Progressive Hedging (PH) algorithm, introduced in [31] and later extended to handle risk measures in [30]. The setting is such that the problem uncertainty, represented by a set of S scenarios, reveals progressively, in T stages. For each scenario s, the decision variable is $x_s \in \mathbb{R}^n$, the convex objective function is $f_s : \mathbb{R}^n \to \mathbb{R}$ and a nonempty convex compact feasible set $C_s \subseteq \mathbb{R}^n$ is given. Consider the following stochastic multistage optimization problem,

$$\begin{cases} \min\limits_{x} & \mathbb{E}[f(x)] := \sum_{s=1}^{S} p_s f_s(x_s) \\ \text{s.t.} & x_s \in C_s, \quad \text{for all } s = 1, \ldots, S, \quad x \in \mathcal{N}, \end{cases} \tag{1}$$

ISSN 0944-6532 / \$ 2.50

where, in the last inclusion, the linear subspace $\mathcal{N}$ gathers the so-called non-anticipativity constraints. Such constraints ensure that, at each stage t, the decision making process depends only on information that is available at time t.

Non-anticipativity constraints couple decisions along scenarios in a structured manner. The PH algorithm decouples those constraints so that, at each iteration, individual subproblems can be solved separately for each scenario. This makes the approach very suitable for parallel implementations. The parallel phase of separate scenario subproblems is followed by a synchronization step that yields a non-anticipative vector by projecting onto the linear subspace $\mathcal{N}$.

The PH algorithm is in fact a Douglas-Rachford splitting, in a space endowed with a weighted scalar product; see for example [13]. Early work on splitting methods, dealing with the classical problem of finding a zero of a sum of maximal monotone operators, can be traced back to [9, 23, 24]. The more recent family of projective splitting methods [7, 10, 11, 12] expanded significantly the reach of Douglas-Rachford approaches. The projective hedging algorithm [13], in particular, can operate in block-iterative and partially asynchronous manner. For a randomized asynchronous variant of the PH algorithm, we refer to [14].

Another important advance of projective splitting methods is the ability to dynamically update certain proximal parameter involved in the calculations. Along iterations, the PH algorithm also generates a dual sequence, that lies in $\mathcal{N}^\perp$, the orthogonal complement of the non-anticipative space. The update of the dual variable amounts to maximizing certain dual function, by applying a gradient method with fixed stepsize (related to the aforementioned proximal parameter). Empirically, splitting methods are generally observed to converge linearly, but at a rate that becomes asymptotically slow and is heavily dependent on the parameter choice. In order to speed up the process, practitioners resort to various heuristic techniques that can be computationally expensive and may not always prove successful [36, 15, 21]. This difficulty is not a surprise as, for the specific PH context, the parameter in question must strike a good balance between optimality and feasibility. Clearly, such a goal is not easy to attain with a value that is kept fixed along iterations.

Our proposal is to employ, instead, a proximal bundle method [6, Part II], tailored to maximize the dual function over $\mathcal{N}^\perp$. To this aim, we consider a more general setting, the constrained minimization of weakly convex functions. For tackling such problems, assuming that projecting onto the feasible set is an easy operation, we propose to apply a bundle algorithm of projective type.

This work is organized as follows. After introducing some notation and definitions, in Section 2 we recall the PH algorithm and formulate a problem dual to (1). This dual problem is a particular instance of a constrained weakly convex problem considered in Section 3. Therein, the subspace $\mathcal{N}^\perp$ is replaced by a general constraint set $\mathcal{M}$, onto which projecting points is assumed to be easy. In such a setting, we introduce a new bundle-like approach with projections, that exploits separable structures as those present in multistage programs like (1). The convergence analysis of the serious step subsequence of this new bundle method is an extension to the projective setting of the theory in [4]. The corresponding results are proven in Appendix A.

The general methodology developed in Section 3 is particularized to (1) in Section 4. The resulting Bundle Progressive Hedging algorithm is formulated in both primal and dual forms. Convergence to solutions of (1) and (6) is shown in Section 5, using the dual form. An equivalent primal formulation is useful to compare our new approach with the original PH algorithm. The final Section 6 discusses similarities and differences between the original Progressive Hedging algorithm and our proposal.

2. The PH algorithm

Before stating the Progressive Hedging algorithm when applied to the multistage program (1), we fix our notation and discuss the crucial non-anticipativity constraint in this stochastic setting.

2.1. Notation and some definitions

Our notation is standard, following mainly [32] and [18]. Given S scenarios, for a certain scenario realization $s \in \{1, \dots, S\}$, the probability of occurrence is denoted by $p_s > 0$. For a vector $v \in \mathbb{R}^{nS}$ with components $v_s \in \mathbb{R}^n$ for all $s = 1, \dots, S$, the expected value and the conditional expectation at stage t are respectively denoted by

$$\mathbb{E}[v] = \sum_{s=1}^{S} p_s v_s\,, \quad \text{and} \quad \mathbb{E}_{[t-1]}[v]\,,$$

where the uncertainty realization at the stages $1, \dots, t-1$ is known.

Considering that all vectors are column vectors, the inner product employed in the space of decision variables is

$$\forall u, v \in \mathbb{R}^{nS} : \langle u, v \rangle_S = \sum_{s=1}^{S} p_s u_s^\top v_s\,.$$

Note that this inner product uses the (nonnegative) probability of each scenario as a weight, a crucial feature in the analysis. We denote by $\|\cdot\|_S$ its corresponding induced norm in $\mathbb{R}^{nS}$, while $\|\cdot\|$ stands for the usual Euclidean norm in $\mathbb{R}^n$. Throughout the text, the symbol $\langle \cdot, \cdot \rangle$ refers to any inner either product (with or without weights), and similarly for the corresponding norms.

On non-anticipativity

In the stochastic setting, decisions have to be taken progressively, as uncertainty is revealed: at stage t, decisions should only depend on the information that became available in stages $1, \dots, t-1$. If $x_{t,s}$ denotes the decision made in stage t for scenario s, the following relations, called of *non-anticipativity*, need to hold for all t and s:

$$x_{t,s} = \mathbb{E}_{[t-1]}[x_t]\,.$$

For the first stage, this constraint states that $x_{1,s}$ are equal for all scenario s. Non-anticipativity constraints define a linear subspace $\mathcal{N}$ of decision variables characterized by conditional expected values. In a manner similar, the projection operator $P_{\mathcal{N}}$ onto $\mathcal{N}$ is characterized by the following simple algebraic relations:

$$\text{for each stage } t,\ P_{\mathcal{N}}[x]_t = \mathbb{E}_{[t-1]}[x_t].$$

Being a self-adjoint operator, the following relation holds for the projection:

$$P_{\mathcal{N}^\perp} = I - P_{\mathcal{N}}.$$

2.2. Primal and dual formulations of the multistage program

To induce separability, the PH algorithm relaxes the non-anticipativity constraint in (1), namely

$$x \in \mathcal{N} \iff x = P_{\mathcal{N}}[x]\,,$$

by means of the following Lagrangian:

$$\mathcal{L}(x, w) = \mathbb{E}[f(x)] + \langle w, x - P_{\mathcal{N}}[x]\rangle_S\,.$$

Because of the identity $P_{\mathcal{N}^\perp} = I - P_{\mathcal{N}}$, the Lagrangian multiplier $w = (w_s)_{s=1}^S \in \mathbb{R}^{nS}$ can be assumed to satisfy

$$w \in \mathcal{N}^\perp \iff P_{\mathcal{N}}[w] = 0\,. \tag{2}$$

Furthermore, by perpendicularity, the linear term in the relaxation can be simplified:

$$\langle w, x - P_{\mathcal{N}}[x]\rangle_S = \langle w, x\rangle_S - \langle w, P_{\mathcal{N}}[x]\rangle_S = \langle w, x\rangle_S.$$

Therefore, the Lagrangian $\mathcal{L}(x, w) = \mathbb{E}[f(x)] + \langle w, x\rangle_S$ is decomposable along scenarios. More specifically,

$$\mathcal{L}(x, w) = \sum_{s=1}^{S} p_s \mathcal{L}_s(x_s, w_s)\,,$$

where for each scenario $s = 1, \ldots, S$, we defined the s-Lagrangian

$$\mathcal{L}_s(x_s, w_s) = f_s(x_s) + w_s^\top x_s\,. \tag{3}$$

Throughout our development, subindices s refer to scenario components (of functions, sets, or vectors). To ease the understanding, Table 2.1 below summarizes the main elements in our notation and clarifies their dimensionality.

In Algorithm 1, with the PH algorithm, the s-Lagrangians (3) are used to construct separate subproblems in the primal space, one per scenario.

Algorithm 1: Progressive Hedging (PH) Algorithm [31]

1: **Initialization**: Choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^\perp$.
2: **for** $k = 0, 1, \ldots$ **do**
3: **Primal subproblems:** solve

$$x_s^{k+\frac{1}{2}} = \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2 \text{ for } s = 1, \ldots, S\,. \tag{4}$$

4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}(x^{k+\frac{1}{2}})$.
5: **Dual update:** $w_s^{k+1} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^{k+1})$ for $s = 1, \ldots, S$.

Notice that, by (3) and (4), the dual updating rule can be interpreted as a gradient step for maximizing an augmented s-dual function.

A notable feature of the PH algorithm is that feasibility is achieved both in the primal and dual iterates by performing simple calculations. On the primal space, the vector formed by collecting the subproblem solutions is projected onto $\mathcal{N}$. On the dual space, the difference $x^{k+\frac{1}{2}} - x^{k+1}$, which measures primal feasibility, lies in $\mathcal{N}^{\perp}$. As a result, dual feasibility is guaranteed throughout the iterative process, as long as the starting dual point belongs to $\mathcal{N}^{\perp}$. The Bundle Progressive Hedging algorithm proposed in this work preserves this characteristic, introducing some modifications in the PH scheme by resorting to duality. More precisely, consider the convex dual function derived from the Lagrangian, that is

$$h(w) = \sum_{s=1}^{S} p_s h_s(w_s) \,, \quad \text{where} \quad h_s(w_s) = \max_{x_s \in C_s} (-\mathcal{L}_s)(x_s, w_s) \,. \tag{5}$$

With this notation, the problem dual to (1) has the expression

$$\min_{w} \; h(w) \,, \quad \text{s.t. } w \in \mathcal{N}^{\perp} \,. \tag{6}$$

Since the multistage stochastic program (1) is convex with a linear constraint, there is no duality gap. As a result, a convergent dual method applied to (6) yields solutions to the original problem. Note that in view of weak duality and the fact that the feasible set of the primal problem is compact and the primal objective function is continuous, then $\inf_{\mathcal{N}^{\perp}} h > -\infty$.

	vector in $\mathbb{R}^{nS}$	scenario s subvector in $\mathbb{R}^n$
primal variable	x	x_s
dual variable	w	w_s
	full function	scenario s subfunction
primal objective	$f : \mathbb{R}^{nS} \to \mathbb{R}$	$f_s : \mathbb{R}^n \to \mathbb{R}$
Lagrangian	$\mathcal{L} : \mathbb{R}^{nS} \times \mathbb{R}^{nS} \to \mathbb{R}$	$\mathcal{L}_s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$
dual objective	$h : \mathbb{R}^{nS} \to \mathbb{R}$	$h_s : \mathbb{R}^n \to \mathbb{R}$

Table 2.1: Notation and dimensionality of variables and functions.

3. A projective bundle-like approach

In (6), the objective function is nonsmooth and convex while the feasible set is a subspace. To handle constrained problems, many bundle methods include easy constraints directly in the subproblems that define iterates, usually a quadratic programming problem, [28, Section 3]. An alternative is to make use of an improvement function or a filter method, as in [2, 3, 20, 35].

The drawback of those approaches is that they destroy decomposable structures, like the one present in (5). We now explain how to put in place a new algorithmic scheme that preserves separability in the iterate subproblems while keeping the well-known convergence properties of bundle methods.

On subdifferentials and weak convexity

Consider a function $H : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, and a point $\overline{w}$ where $H(\overline{w})$ is finite.

If H is a locally Lipschitz function, then the Clarke subdifferential of H at $\overline{w}$ is the set $\partial H(\overline{w})$ given by the convex hull of all the points of the form $\lim_{k\to+\infty} \nabla H(w_k)$, where $\{w_k\}$ is a sequence of points where H is differentiable, and $w_k \to \overline{w}$.

If H is a proper convex lower semicontinuous (lsc) function, then $\partial H(\overline{w})$ coincides with the subdifferential of H at $\overline{w}$ of convex analysis:

$$\partial H(\overline{w}) = \{x \in \mathbb{R}^N : H(w) \geqslant H(\overline{w}) + \langle x, w - \overline{w}\rangle \text{ for all } w \in \mathbb{R}^N\}.$$

We shall work with a special class of nonconvex objective functions, defined below.

– Given $\rho > 0$, the function $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is said to be *ρ-weakly convex*, if the augmented function $h(\cdot) + \frac{\rho}{2}|\cdot|^2$ is convex. Recall that $|\cdot|$ could either be $\|\cdot\|$ or $\|\cdot\|_S$ (in the latter case, the dimension is $N = nS$).

Weakly convex functions are related to other nonconvex notions in the literature, such as prox-regular or lower-$\mathcal{C}^2$ functions [32, Definition 13.27], generalized differentiable functions in the sense of Norkin [27], and semismooth functions [26]. Furthermore, weakly convex functions can be characterized using the Clarke subdifferential. In particular, from [4, Proposition 2.2], h is ρ-weakly convex if and only if, for all $w, \tilde{w} \in \mathbb{R}^N$ with nonempty $\partial h(\tilde{w})$, any $g \in \partial h(\tilde{w})$ satisfies

$$h(w) + \frac{\rho}{2}|w - \tilde{w}|^2 \geqslant h(\tilde{w}) + \langle g, w - \tilde{w}\rangle.$$

Considered setting for the bundle method

Given $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ a ρ-weakly convex function and $\mathcal{M}$ a nonempty closed convex set, onto which computing a projection is not costly, we are interested in solving the following constrained minimization problem

$$\min_{w \in \mathcal{M}} h(w) \tag{7}$$

by means of a bundle-like method. For the multistage program (1), the constraints are given by $\mathcal{M} = \mathcal{N}^\perp$ in problem (6), but any easy-to-project feasible set could be considered (for instance a nonnegative orthant, $\mathcal{M} = \{w \geqslant 0\}$).

Accordingly, assumptions (A1)–(A5) in Theorem 3.2 shall make use of the following definitions, stated for a function H given by the dual function h plus the indicator function of the feasible set.

– A lsc function $H : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ satisfies the *proper separation of isocost surfaces* property, if there exists $\varepsilon > 0$ such that

$$\bar{x}_1, \bar{x}_2 \in (\partial H)^{-1}(0),\ |\bar{x}_1 - \bar{x}_2| < \varepsilon \implies H(\bar{x}_1) = H(\bar{x}_2)\,. \tag{8}$$

– The *subdifferential error bound* holds for problem $\min_w H(w)$, where the function $H : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is bounded below, if for every $v \geqslant \inf_{w \in \mathbb{R}^N} H(w)$, there exist $\epsilon, \ell > 0$ such that whenever $w \in \mathbb{R}^N$, $H(w) \leqslant v$, and $x \in \partial H(w) \cap B(0, \epsilon)$, the following is true:

$$d(w, \mathcal{S}) \leqslant \ell |x|\,, \tag{9}$$

where $\mathcal{S} = (\partial H)^{-1}(0)$ is the set of critical points of H.

The subdifferential error bound, sometimes also called metric subregularity, is related to other concepts in the literature. For convex functions, condition (9) is equivalent to the Kurdyka–Łojasiewicz inequality with exponent $\frac{1}{2}$, and also equivalent to a quadratic growth condition with respect to the distance to the set of minimizers ([38, Proposition 2]).

In the weakly convex case, the subdifferential error bound implies the Kurdyka–Łojasiewicz inequality with exponent $\frac{1}{2}$, and the quadratic growth condition for the set of critical points [8]. For nonconvex structured problems, the subdifferential error bound combined with the assumption of properly separated isocost surfaces implies the Kurdyka–Łojasiewicz inequality with exponent $\frac{1}{2}$ [22, Theorem 4.1].

All these regularity conditions relate primal and dual information, in particular points in the primal space or their function values, with subgradients at a critical point or near a critical point. These conditions are usually used in the literature to prove linear rate of convergence of different methods, see [4, 5, 25].

3.1. A general algorithmic pattern for weakly convex problems

In the setting of problem (7), a family of *model functions* $\{\varphi^k_{w^k}\}$, built along iterations, is available. Denoting by $i_{\mathcal{M}}$ the indicator function of the feasible set, given $w^k \in \mathcal{M}$, the model is a

$$\text{convex function } \ \varphi^k_{w^k}(\cdot) \ \text{ approximating } \ h(\cdot) + i_{\mathcal{M}}(\cdot) - h(w^k)\,. \tag{10}$$

If the objective function h is convex, a typical construct is to create a cutting-plane model for the sum $h + i_{\mathcal{M}}$. For nonconvex h, cutting-plane models need to be tilted and/or downshifted to be adequate. If the parameter of weak convexity ρ is known, this is an easy task. Otherwise, estimates of such parameter must be "guessed" and suitably updated along the iterative scheme, as in the redistributed proximal bundle method from [16, 17].

In our development, to generate iterates with a special subsequence converging with linear rate, the following property should be satisfied by the models.

Definition 3.1. (1QA models) The family of finite-valued convex functions $\varphi^k_{w^k} : \mathbb{R}^N \to \mathbb{R}$ satisfying (10) is one-sided with quadratic accuracy for $h + i_{\mathcal{M}}$ if there exists $q > 0$ such that

$$\forall w^k \in \mathcal{M}\,, \forall w \in \mathbb{R}^N\,, \quad \varphi^k_{w^k}(w) \leqslant h(w) + i_{\mathcal{M}}(w) - h(w^k) + \frac{q}{2}|w - w^k|^2\,. \quad \square \tag{11}$$

It is shown in [4, Proposition 5.2] that the redistributed approach from [16, 17], which defines piecewise linear models for $h(w) + \frac{q}{2}\|w - w^k\|^2$, satisfies the 1QA property.

Model functions are useful to define iterates with low computational burden. Finding a new iterate in our proposal amounts to solving

$$\min_w \left\{\varphi^k_{w^k}(w) + \frac{1}{2t_k}|w - w^k|^2\right\}, \tag{12}$$

a problem that is a simple quadratic program if the model is piecewise linear. The first-order optimality condition for problem (12),

$$0 \in \partial\varphi^k_{w^k}(w) + \frac{1}{t_k}(w - w^k),$$

characterizes the implicit updating rule (13a) given below.

Given parameters $m \in (0,1)$, $t_{\min} > 0$, and some initial $w^0 \in \mathcal{M}$, for all $k \geqslant 0$, our algorithmic scheme defines

$$w^{k+\frac{1}{2}} = w^k - t_k G^k, \qquad \text{for some } G^k \in \partial\varphi^k_{w^k}(w^{k+\frac{1}{2}})\,, \tag{13a}$$

$$u^{k+1} = P_{\mathcal{M}}(w^{k+\frac{1}{2}}), \tag{13b}$$

$$\text{if } \; h(u^{k+1}) - h(w^k) \leqslant m\varphi^k_{w^k}(w^{k+\frac{1}{2}}),\; \text{ declare a serious step.}$$

$$\text{Set } \; w^{k+1} = u^{k+1} \; \text{ and choose } \; t_{k+1}\,. \tag{13c}$$

$$\text{Otherwise, declare a null step.}$$

$$\text{Set } \; w^{k+1} = w^k \text{ and choose } \; t_{k+1}\,. \tag{13d}$$

$$\text{In both cases select a new model } \; \varphi^{k+1}_{w^{k+1}} \text{ and } \; t_{k+1} \geqslant t_{\min}\,.$$

In this algorithmic pattern, iterates satisfying (13a)–(13c) form the so-called *serious-step subsequence* in bundle methods. As indicated by (13c), at serious steps the objective functional values decrease. Iterates satisfying (13a) and (13b) but not (13c) form the subsequence of *nullsteps*.

Since iterates in bundle methods can be of two types, the convergence theory splits the asymptotic analysis into two parts, depending on whether the serious step sequence is finite or infinite. We state below the corresponding result for the latter, referring for the proof in the appendix. Note that the result extends the theory for descent methods in [4], to take into consideration the extra projective step in (13b).

For the next result, by construction (cf. (13c)), recall that the stepsizes t_k are bounded from below by $t_{\min} > 0$.

Theorem 3.2. (Global convergence and local linear rate of serious subsequence) *Consider problem* (7) *with the following assumptions:*

$$h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\} \textit{ is a } \rho - \textit{weakly convex function;} \tag{A1}$$

$$\mathcal{M} \textit{ is a nonempty closed convex set;} \tag{A2}$$

$$h \textit{ is bounded below on } \mathcal{M},\; \textit{that is } \inf_{\mathcal{M}} h > -\infty\,; \tag{A3}$$

$$h + i_{\mathcal{M}} \textit{ satisfies the proper separation of isocost surfaces property (8);} \tag{A4}$$

$$h + i_{\mathcal{M}} \textit{ satisfies the subdifferential error bound (9).} \tag{A5}$$

Assume that in the pattern (13) *the functions* $\varphi^k_{w^k}$ *are 1QA models having parameter* $q \leqslant \rho$ *as in Definition* 3.1*, and that the stepsizes* t_k *are bounded above by* $t_{\max} > 0$ *in step* (13c)*.*

If there is an infinite subsequence of serious steps, that is $\{w^k\}$ *satisfying* (13a)–(13c)*, the following holds.*

(i) *The subsequence of functional values* $\{h(w^k)\}$ *monotonically converges to some critical value* h^* *of* $h + i_{\mathcal{M}}$*, such that the sequence of functional errors* $\{v^k = h(w^k) - h^*\}$ *converges to 0 with Q-linear rate: there exists* $r \in (0,1)$ *such that for all sufficiently large* k*,* $v^{k+1} \leqslant r v^k$ *.*

(ii) *The subsequences of iterates* $\{w^k\}$ *and intermediate points* $\{w^{k+\frac{1}{2}}\}$ *converge to a critical point* w^* *of* $h + i_{\mathcal{M}}$ *with R-linear rate: there exists* $r \in (0,1)$*, and* $c > 0$ *such that for all sufficiently large* k

$$|w^k - w^*| \leqslant c\sqrt{r}^k, \quad |w^{k+\frac{1}{2}} - w^*| \leqslant c(2-\sqrt{r})\sqrt{r}^k$$

It remains to analyze the asymptotic behavior of the "tail of null steps". To do so, the specific definition of the model functions plays a crucial role. Since this feature is problem dependent, we consider the particular instance of interest, that is problem (1)–(6), and give a complete convergence analysis for the bundle PH algorithm presented in the next section.

4. Bundle progressive hedging

Our motivation to consider algorithmic schemes of the form (13a)–(13c) is to exploit decomposable structures in (7). The challenge is to define model functions that inherit h's structure and, at the same time, incorporate information on the feasible set $\mathcal{M}$ without destroying separability. We now explain how to build suitable model functions for (6) and present the bundle variant of the PH algorithm.

4.1. Building separable models

Usually bundle methods define linearizations for the sum $h + i_{\mathcal{M}}$, but such model functions are not separable. To illustrate this difficulty, consider the following simple instance of (7), where $w_1, w_2 \in \mathbb{R}$ are decision variables and $A \in \mathbb{R}$ is given:

$$\begin{cases} \displaystyle\min_{w_1,w_2} & h_1(w_1) + h_2(w_2) \\ \text{s.t.} & w_1 + w_2 = A\,. \end{cases}$$

A classical model for this problem takes cutting-plane approximations for each term, say $\check{h}_s(w_s)$ for $s = 1, 2$, and adds the indicator function:

$$\varphi^k_{w^k}(w_1, w_2) := \check{h}_1(w_1) + \check{h}_2(w_2) - h_1(w_1^k) - h_2(w_2^k) + i_{\mathcal{M}}(w_1, w_2)\,.$$

With such a model, problem (12), whose solution yields (13a), is not separable:

$$\min_{w_1+w_2=A} \left\{ \varphi^k_{w^k}(w_1, w_2) + \frac{1}{2t_k} \left| \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} w_1^k \\ w_2^k \end{pmatrix} \right|^2 \right\}.$$

By contrast, if a separable model function was available for the indicator function $i_{\mathcal{M}}(w_1, w_2)$, the calculations required in (13a) could be performed separately, by solving, for $s = 1, 2$,

$$\min_{w_s} \left\{ \varphi^k_{w_s^k}(w_s) + \frac{1}{2t_k} \|w_s - w_s^k\|^2 \right\}. \tag{15}$$

When the objective function h involves the sum of more than two terms, as in the case of multistage programs with many scenarios, using separable models improves significantly the computational performance (the solution to (15) can be computed in parallel for the different scenarios).

Following the spirit of Table 2.1, and because the ability of defining separable models depends on the constraints $\mathcal{M}$ under consideration, we recall in Table 4.1 the feasible sets involved in the primal and dual problems.

Regarding the PH setting, that is when (7) has the format (5)–(6), the feasible set therein is given by the non-anticipativity constraints written in dual form.

Hence, as in (2), $\mathcal{M} = \mathcal{N}^\perp = \{w : P_{\mathcal{N}}[w] = 0\}$.

space	nature	notation
general framework of subsection 3.1	-	$\mathcal{M}$
nonanticipative	primal	$\mathcal{N}$
nonanticipative orthogonal	dual	$\mathcal{N}^\perp$

Table 4.1: Spaces involved in the general framework of subsection 3.1, and in the BPH analysis.

This subspace has a favorable structure and the projection step (13b) is straightforward:

$$P_{\mathcal{M}}[w] = w - P_{\mathcal{N}}[w]\,.$$

Recall that, by (5), $h(w) = \sum_{s=1}^{S} p_s h_s(w_s)$. Our proposal is to work with the weighted sum of special model functions, defined for each scenario and derived from (4).

Proposition 4.1. (Separable 1QA models) *Given $(x^k, w^k) \in \mathcal{N} \times \mathcal{N}^\perp$, for $s = 1, \ldots, S$ consider the approximate s-Lagrangian*

$$\mathcal{L}_s^k(x_s, w_s) = f(x_s) + w_s^\top (x_s - x_s^k) = \mathcal{L}_s(x_s, w_s) - w_s^\top x_s^k\,, \tag{16}$$

and the individual functions

$$\varphi^k_{w_s^k}(w_s) := \max_{x_s \in C_s} (-\mathcal{L}_s^k)(x_s, w_s) - h_s(w_s^k)\,. \tag{17}$$

The corresponding weighted sum

$$\varphi^k_{w^k}(w) = \sum_{s=1}^{S} p_s \varphi^k_{w_s^k}(w_s) = h(w) - h(w^k) + \left\langle w, x^k \right\rangle_S$$

defines a model for $(h + i_{\mathcal{N}^\perp})(w)$ at w^k that is separable, convex and of type 1QA with $q = 0$.

Proof. First, the model $\varphi^k_{w^k}$ is separable by construction, and convex, since each $\varphi^k_{w_s^k}$ is the maximum of affine functions $-\mathcal{L}_s(x_s, w_s) - h_s(w_s^k)$ in w_s. From (17) it follows that

$$\varphi^k_{w_s^k}(w_s) \leqslant h_s(w_s) + w_s^\top x_s^k - h_s(w_s^k).$$

After multiplying by p_s and adding all scenarios, this means that

$$\varphi^k_{w^k}(w) \leqslant h(w) + \langle w, x^k \rangle_S - h(w^k).$$

Note that since $x^k \in \mathcal{N}$, then $\langle w, x^k \rangle_S \leqslant i_{\mathcal{N}^\perp}(w)$ for all w. Indeed, if $w \in \mathcal{N}^\perp$, $\langle w, x^k \rangle_S = 0 \leqslant 0 = i_{\mathcal{N}^\perp}(w)$. On the other hand, if $w \in \mathcal{N} \backslash \{0\}$, $\langle w, x^k \rangle$ is a finite value, thus $\langle w, x^k \rangle_S < +\infty = i_{\mathcal{N}^\perp}(w)$. Therefore,

$$\varphi^k_{w^k}(w) \leqslant h(w) + i_{\mathcal{N}^\perp}(w) - h(w^k).$$

Hence, $\varphi^k_{w^k}$ is a 1QA model for $h + i_{\mathcal{N}^\perp}$ at w^k with $q = 0$. □

The approximate s-Lagrangians put together build a 1QA model for $h + i_{\mathcal{N}^\perp}$. In fact, as

$$\mathcal{L}_s^k(x_s, w_w) = \mathcal{L}_s(x_s, w_s) - (x_s^k)^\top w_s\,,$$

the term $\langle x^k, w \rangle_S$ corresponds to a lower linearization of the indicator function $i_{\mathcal{N}^\perp}$. Using the wording from [2], the models $\varphi^k_{w^k}$ are of lower type; see also [1].

4.2. Comparison with subproblems in the PH algorithm

To understand the given definition for the individual models, first recall we are interested in dealing with separate subproblems (15). Plugging (17) therein yields

$$\begin{aligned}
&\min_{w_s}\left\{\max_{x_s\in C_s}(-\mathcal{L}_s^k)(x_s,w_s)-h_s(w_s^k)+\frac{1}{2t_k}\|w_s-w_s^k\|^2\right\}\\
&=\min_{w_s}\left\{\max_{x_s\in C_s}(-\mathcal{L}_s)(x_s,w_s)+(x_s^k)^\top w_s-h_s(w_s^k)+\frac{1}{2t_k}\|w_s-w_s^k\|^2\right\}.
\end{aligned}$$

For this (convex-concave) saddle-point problem it is equivalent to solve

$$\begin{aligned}
&\max_{x_s\in C_s}\min_{w_s}\left\{(-\mathcal{L}_s)(x_s,w_s)+(x_s^k)^\top w_s-h_s(w_s^k)+\frac{1}{2t_k}\|w_s-w_s^k\|^2\right\}\\
&=-h_s(w_s^k)+\max_{x_s\in C_s}\left\{\min_{w_s}\left[-f_s(x_s)-w_s^\top(x_s-x_s^k)+\frac{1}{2t_k}\|w_s-w_s^k\|^2\right]\right\}.
\end{aligned}$$

The expression between brackets is minimized at $w_s=w_s^k+t_k(x_s-x_s^k)$. Therefore, (15) is equivalent to

$$\begin{aligned}
&\max_{x_s\in C_s}\left\{-f_s(x_s)-(w_s^k+t_k(x_s-x_s^k))^\top(x_s-x_s^k)+\frac{t_k}{2}\|x_s-x_s^k\|^2\right\}\\
&=\max_{x_s\in C_s}\left\{-f_s(x_s)-(w_s^k)^\top(x_s-x_s^k)-t_k\|x_s-x_s^k\|^2+\frac{t_k}{2}\|x_s-x_s^k\|^2\right\}\\
&=-\min_{x_s\in C_s}\left\{f_s(x_s)+(w_s^k)^\top(x_s-x_s^k)+\frac{t_k}{2}\|x_s-x_s^k\|^2\right\}\\
&=(w_s^k)^\top x_s^k-\min_{x_s\in C_s}\mathcal{L}_s^k(x_s,w_s^k)+\frac{t_k}{2}\|x_s-x_s^k\|^2. \qquad (18)
\end{aligned}$$

This last minimization problem is practically identical to the PH original subproblem (4). The difference is that the stepsize t_k, in this case, can be dynamically updated, as in a bundle method.

Additionally, note that the Lagrangian $\mathcal{L}_s^k$ is an approximation of the original Lagrangian obtained when relaxing the non-anticipativity constraint. More specifically, the Lagrangian

$$\mathcal{L}(x,w)=\sum_{s=1}^{S}p_s\left(f_s(x_s)+w_s^\top(x_s-P_{\mathcal{N}}[x]_s)\right)$$

is approximated with

$$\mathcal{L}^k(x,w)=\sum_{s=1}^{S}p_s\left(f_s(x_s)+w_s^\top(x_s-x_s^k)\right),\quad x_s^k\in\mathcal{N}.$$

Consequently, the non-separable augmented Lagrangian subproblems

$$\sum_{s=1}^{S}p_s\min_{x_s\in C_s}\left\{f_s(x_s)+w_s^\top(x_s-P_{\mathcal{N}}[x]_s)+\frac{t_k}{2}\|x_s-P_{\mathcal{N}}[x]_s\|^2\right\}$$

are approximated (in both PH and BPH) by the separable subproblems

$$\sum_{s=1}^{S}p_s\min_{x_s\in C_s}\left\{f_s(x_s)+w_s^\top(x_s-x_s^k)+\frac{t_k}{2}\|x_s-x_s^k\|^2\right\}$$

With our approach, the quality of this approximation is measured by means of the serious/null step test (13c). Performing the descent test is not free, as it requires an extra dual function evaluation, that is, solving another set of subproblems in parallel. This is made clear in the descent test in Algorithm 2 given below.

4.3. Statement of the algorithm in dual form

Algorithm 2: Bundle Progressive Hedging (BPH) Algorithm in dual form

---

1: **Initialization**: Given a stopping tolerance `TOL` $\geqslant 0$ and parameters $m \in (0,1)$ and $t_{\min} > 0$, choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^\perp$ and an initial stepsize $t_0 > 0$. Compute the dual value $h_s(w_s^0) = -\mathcal{L}_s(\hat{y}_s^0, w_s^0)$, by finding

$$\hat{y}_s^0 \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, w_s^0) \text{ for } s = 1, \ldots, S\,.$$

2: **for** $k = 0, 1, \ldots$ **do**

3: **Dual subproblems:** given $\varphi^k_{w_s^k}(w_s) = \max_{x_s \in C_s}(-\mathcal{L}_s^k)(x_s, w_s) - h_s(w_s^k)$, solve

$$w_s^{k+\frac{1}{2}} = \arg\min_{w_s} \left\{ \varphi^k_{w_s^k}(w_s) + \frac{1}{2t_k}\|w_s - w_s^k\|^2 \right\} \text{ for } s = 1, \ldots, S\,. \tag{19}$$

The nominal decrease $\varphi^k_{w_s^k}(w_s^{k+\frac{1}{2}})$ is available.

4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$, where

$$x_s^{k+\frac{1}{2}} = x_s^k + \frac{1}{t_k}(w_s^{k+\frac{1}{2}} - w_s^k) \ \text{ for } \ s = 1, \ldots, S\,. \tag{20}$$

5: **Stopping test**: if $-\varphi^k_{w^k}(w^{k+\frac{1}{2}}) \leqslant$ `TOL`, stop and return (x^k, w^k).

6: **Dual projection:** $u^{k+1} = P_{\mathcal{N}^\perp}(w^{k+\frac{1}{2}}) = w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}]$.

7: **Descent test:** Compute the dual value $h_s(u_s^{k+1}) = -\mathcal{L}_s(y_s^{k+1}, u_s^{k+1})$, by finding

$$y_s^{k+1} \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, u_s^{k+1}) \text{ for } s = 1, \ldots, S\,.$$

If $-\mathcal{L}(y^{k+1}, u^{k+1}) \leqslant -\mathcal{L}(\hat{y}^k, w^k) + m\varphi^k_{w^k}(w^{k+\frac{1}{2}})$ declare a serious step:
$w^{k+1} = u^{k+1},\ \hat{y}^{k+1} = y^{k+1},\ t_{k+1} \geqslant t_{\min}$.
Otherwise, declare a null step: set $w^{k+1} = w^k$, $\hat{y}^{k+1} = \hat{y}^k$ and $t_{k+1} \geqslant t_{\min}$.

Recall that the PH algorithm keeps separate subproblems for each scenario, performing afterwards a coordination step, projecting the primal candidates onto the linear subspace of non-anticipative constraints. The Bundle Progressive Hedging Algorithm 2 maintains those features, adding a descent test to measure the quality of the model approximation. The main steps of the algorithm are as follows.

- Iteration k starts by solving, for each scenario $s = 1, \ldots, S$, a dual subproblem, yielding dual intermediate iterates.
- Intermediate primal points are derived from the dual intermediate solutions. As shown in Lemma 4.2(i), such primal points actually correspond to minimizing an augmented s-Lagrangian in the primal space.

- The projection of the dual intermediate points onto the orthogonal subspace of the non-anticipativity ensures dual feasibility.
- The projected dual iterates are evaluated, to determine if there was some decrease in the dual function h:
  - if there is sufficient descent, a serious step is made. This point then becomes the best candidate point generated so far, or
  - if there is no sufficient descent, a null step is made.

The Bundle Progressive Hedging Algorithm 2 results from applying the scheme (13) with the 1QA model in Proposition 4.1, to the dual problem (6). Being a particular instance of the general algorithmic pattern given in Section 3.1, convergence of the serious dual subsequence generated by BPH is therefore ensured by Theorem 3.2.

Table 4.2 compares the notation employed in Algorithms 2 and 3 to generate, respectively, the primal and dual sequences of the bundle approach.

	vector in $\mathbb{R}^{nS}$	scenario s subvector in $\mathbb{R}^n$
intermediate primal	$x^{k+\frac{1}{2}}$	$x_s^{k+\frac{1}{2}}$
projected primal	x^k	x_s^k
intermediate dual	$w^{k+\frac{1}{2}}$	$w_s^{k+\frac{1}{2}}$
projected dual	u^k	u_s^k
dual serious step	w^k	w_s^k
minimizer of $\mathcal{L}_s(\cdot, u_s^k)$ over C_s	y^k	y_s^k

Table 4.2: Notation for primal and dual sequences generated by Algorithms 2 and 3.

Before passing to BPH's convergence analysis, some remarks are in order.

- Notice that the descent test is performed with the full Lagrangian $\mathcal{L}$, and not the individual s-Lagrangians.
- In (13), consider h to be the dual function defined in (5), the linear subspace $\mathcal{M} = \mathcal{N}^{\perp}$, and $N = nS$. First, the optimality conditions of the problem in (19) correspond to (13a), for
$$G^k = -(x^k - x^{k+\frac{1}{2}}), \tag{21}$$
according to (20). Furthermore, step 6 of Algorithm 2 is exactly (13c), while the descent test in step 7 of Algorithm 2 is equivalent to (13c), by using the definition of the dual function h in (5), the construction of y^{k+1}, and the definition of $\hat{y}^k$. Barring the stopping test, Algorithm 2 is a particular instance of the pattern in (13).
- The stopping test of step 5 in Algorithm 3 yields indeed approximate solutions. To see this, recall that the BPH method stops when the aggregate error and the aggregate subgradient defined in Lemma A.1(i) are sufficiently small. Since, see Lemma A.1(ii), $G^k \in \partial_{E^k}\varphi^k_{w_s^k}(w_s^k)$ for $E^k \geqslant 0$, and $\varphi^k_{w_s^k}(w_s^k) = 0$, this means that for all w
$$h(w) - h(w^k) + \left\langle w, x^k \right\rangle_S \geqslant \left\langle G^k, w - w^k \right\rangle - E^k\,, \text{ where } t_k\|G^k\|^2\,, E^k \leqslant \texttt{TOL}.$$

In particular, for all $w \in \mathcal{N}^\perp$ the linear term in the left-hand side vanishes, and w^k is an approximate minimizer of the dual problem (6), for

$$\eta = \max\left\{\texttt{TOL}, \sqrt{\frac{\texttt{TOL}}{t_{\min}}}\right\}: \quad h(w) \geqslant h(w^k) - \eta\|w - w^k\| - \eta\,.$$

4.4. Relation with progressive hedging and primal formulation

The BPH Algorithm 2 differs from the PH Algorithm 1 in the implementation of the descent test. Notwithstanding, both methods also share some features. As shown in (18), for each scenario, the subproblem (19) solved by BPH is dual to the subproblem of the PH Algorithm 1. In both PH and BPH, primal points are projected onto the set of non-anticipativity constraints; and the nature of the dual update is the same, projecting onto $\mathcal{N}^\perp$ to guarantee dual feasibility.

The resemblance between the two methods becomes more apparent after formulating Algorithm 2 in primal terms, exploiting in addition to (18), primal-dual relations stated in Lemma 4.2. The primal formulation of the Bundle Progressive Hedging given in Algorithm 3 allows a straightforward comparison with the PH Algorithm 1. However, the dual form Algorithm 2 is more handy for the convergence analysis.

Lemma 4.2. (From dual to primal BPH formulations) *Consider the approximate s-Lagrangian $\mathcal{L}_s^k$ defined in* (16)*, and the model $\varphi_{w^k}^k$ of* (17) *that defines subproblems* (19) *in Algorithm* 2*. The following holds.*

(i) *For all scenario $s = 1, \ldots, S$, the intermediate step $x_s^{k+\frac{1}{2}}$ defined in* (20) *can be equivalently computed as follows*
$$x_s^{k+\frac{1}{2}} \in \operatorname*{arg\,min}_{x_s \in C_s}\left\{\mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2\right\}.$$
In particular, $x_s^{k+\frac{1}{2}} \in C_s$.

(ii) *The dual projection rule $u^{k+1} = w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}]$ can be equivalently performed by doing $u^{k+1} = w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1})$.*

(iii) *The dual model $\varphi_{w^k}^k$ evaluated at the dual intermediate point $w^{k+\frac{1}{2}}$ can be written in primal-dual terms as follows*
$$\varphi_{w^k}^k(w^{k+\frac{1}{2}}) = \mathcal{L}(\hat{y}^k, w^k) - \mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}).$$

Proof. To show item (i), recall the relations (18). Since $w_s^{k+\frac{1}{2}}$ minimizes the expression $\varphi_{w_s^k}^k(w_s) + \frac{1}{2t_k}\|w_s - w_s^k\|^2$, and it also has the form $w_s^{k+\frac{1}{2}} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^k)$, then $x_s^{k+\frac{1}{2}}$ minimizes over C_s

$$f_s(x_s) + (w_s^k)^\top(x_s - x_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2 = \mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2\,.$$

Regarding item (ii), since $w^{k+\frac{1}{2}} = w^k + t_k(x^{k+\frac{1}{2}} - x^k)$, and $w^k \in \mathcal{N}^\perp$, then

$$\begin{aligned} w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}] &= w^k + t_k(x^{k+\frac{1}{2}} - x^k) - P_{\mathcal{N}}[w^k + t_k(x^{k+\frac{1}{2}} - x^k)] \\ &= w^k + t_k(x^{k+\frac{1}{2}} - x^k) - t_k(x^{k+1} - x^k) = w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1}), \end{aligned}$$

where in the second equality we use $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$.

Finally, note that (20) and (19) imply that $x_s^{k+\frac{1}{2}}$ solves $\max_{x_s\in C_s}(-\mathcal{L}_s^k)(x_s, w_s^{k+\frac{1}{2}})$. Therefore, from the definition of $\varphi^k_{w_s^k}$, it holds that

$$\varphi^k_{w_s^k}(w_s^{k+\frac{1}{2}}) = (-\mathcal{L}_s^k)(x_s^{k+\frac{1}{2}}, w_s^{k+\frac{1}{2}}) - h_s(w_s^k).$$

Moreover, by construction, $\widehat{y}_s^k$ solves $\max_{x_s\in C_s}(-\mathcal{L}_s)(x_s, w_s^k)$, then

$$\varphi^k_{w_s^k}(w_s^{k+\frac{1}{2}}) = (-\mathcal{L}_s^k)(x_s^{k+\frac{1}{2}}, w_s^{k+\frac{1}{2}}) + \mathcal{L}_s(\widehat{y}_s, w_s^k).$$

Taking the expected value in this last formula gives item (iii). □

Algorithm 3: Bundle Progressive Hedging (BPH) Algorithm in primal form

---

1: **Initialization**: Given a stopping tolerance `TOL` $\geqslant 0$ and parameters $m \in (0,1)$ and $t_{\min} > 0$, choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^\perp$. Compute

$$\widehat{y}_s^0 \in \arg\min_{x_s\in C_s} \mathcal{L}_s(x_s, w_s^0) \text{ for } s = 1, \ldots, S\,.$$

2: **for** $k = 0, 1, \ldots$ **do**

3: **Primal subproblems:** solve

$$x_s^{k+\frac{1}{2}} = \arg\min_{x_s\in C_s}\left\{f_s(x_s) + (w_s^k)^\top x_s + \frac{t_k}{2}\|x_s - x_s^k\|^2\right\}, \text{ for } s = 1, \ldots, S.$$

4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$.

5: **Dual update:** Compute $w_s^{k+\frac{1}{2}} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^k)$ for $s = 1, \ldots, S$.

6: **Stopping test:** if $\mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}) - \mathcal{L}(\widehat{y}^k, w^k) \leqslant$ `TOL`, stop and return (x^k, w^k).

7: **Dual projection:** $u^{k+1} = w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1})$.

8: **Descent test:** Compute

$$y_s^{k+1} \in \arg\min_{x_s\in C_s} \mathcal{L}_s(x_s, u_s^{k+1}) \text{ for } s = 1, \ldots, S\,.$$

If $\mathcal{L}(\widehat{y}^k, w^k) - \mathcal{L}(y^{k+1}, u^{k+1}) \leqslant m(\mathcal{L}(\widehat{y}^k, w^k) - \mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}))$, declare a serious step: set $w^{k+1} = u^{k+1}$, $\widehat{y}^{k+1} = y^{k+1}$, and take $t_{k+1} \geqslant t_{\min}$.
Otherwise, declare a null step: set $w^{k+1} = w^k$, $\widehat{y}^{k+1} = \widehat{y}^k$ and $t_{k+1} \geqslant t_{\min}$.

By Lemma 4.2, the intermediate primal point $x_s^{k+\frac{1}{2}}$ can be computed as either a minimizer over C_s of the augmented approximate s-Lagrangian

$$\mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2,$$

evaluated at the serious dual point $w_s = w_s^k$, or as a minimizer over C_s of the Lagrangian

$$\mathcal{L}_s(x_s, w_s^{k+\frac{1}{2}}),$$

evaluated at the intermediate dual point $w_s = w_s^{k+\frac{1}{2}}$. Together with the primal updating rule (20), this means that the dual subproblems of Algorithm 2 can be written in primal terms, followed by a dual updating rule for $w_s^{k+\frac{1}{2}}$ deduced from (20). Algorithm 3 is the primal version of Algorithm 2.

We now show that Bundle Progressive Hedging algorithm finds primal and dual solutions, in the case of finite termination, infinite number of serious steps or when there is a tail of null steps.

5. Convergence analysis of the Bundle Progressive Hedging

The primal and dual versions of the Bundle Progressive Hedging algorithm are equivalent because:

- Steps 3 and 4 of Algorithm 2 are equivalent to steps 3 and 4 of Algorithm 3, due to Lemma 4.2(i).
- Step 5 of Algorithm 2, the stopping test in dual terms, is equivalent to step 5 of Algorithm 3, due to Lemma 4.2(iii).
- Step 6 of Algorithm 2 is equivalent to step 6 of Algorithm 3, due to Lemma 4.2(ii).
- Step 7 of Algorithm 2 is equivalent to step 7 of Algorithm 3. Indeed, the descent test is exactly the same by using Lemma 4.2(iii). As for the construction of the model, in the dual version of the algorithm (Algorithm 2) it is explicitly done by using the serious step w^{k+1}, while in the primal case, Algorithm 3, it is implicitly performed by using the projection x^{k+1} in the quadratic term of the primal subproblem, and using w^{k+1} as the cost of the linear term of the objective of the subproblem.

Throughout this section the stopping tolerance is set to $\texttt{TOL} = 0$. By the stopping test in Algorithm 2, this means that when the test is triggered, it must hold that

$$0 \leqslant -\varphi^k_{w^k}(w^{k+\frac{1}{2}}) \leqslant \texttt{TOL} = 0\,,$$

by Lemma A.1(i). With this setting, either Algorithm 2 stops after a finite number of iterations with $\varphi^k_{w^k}(w^{k+\frac{1}{2}}) = 0$, or the algorithm runs indefinitely. In this case, either the serious subsequence is infinite (after each serious iterate, only a finite number of null iterations occur), or a last serious iterate is generated at iteration $\hat{k}-1$, say $\hat{w} = w^{\hat{k}}$, and afterwards all iterates are declared null steps.

5.1. Cases of finite termination and infinite number of serious steps

We first consider the case of a finite termination. To this aim, it is useful to characterize the subdifferential of the dual function h in terms of primal information. Specifically, given a scenario $s \in \{1, \dots, S\}$, each function $h_s(w_s)$ defined in (5) is the maximum of a family of affine functions of w_s, therefore it is convex [18, Chapter I, Proposition 2.1.2]. Furthermore, letting

$$C_s(w_s) = \{x_s \in C_s : -\mathcal{L}_s(x_s, w_s) = h_s(w_s)\}\,,$$

then according to [18, Chapter VI, Theorem 4.4.2],

$$\partial h_s(\bar{w}_s) = \mathrm{co}\Big(\bigcup_{\bar{x}_s \in C_s(\bar{w}_s)} \partial(-\mathcal{L}_s)(\bar{x}_s, \bar{w}_s)\Big)\,.$$

Moreover, since $\mathcal{L}_s$ is differentiable with respect to w_s, then $\partial_w(-\mathcal{L}_s)(\bar{x}_s, \bar{w}_s) = \{-\bar{x}_s\}$, for any $\bar{x}_s \in C_s(\bar{w}_s)$, that is, any $\bar{x}_s \in C_s$ that solves the problem in (5) for

$w_s = \bar{w}_s$. In other words, $C_s(\bar{w}_s)$ is exactly the set of maximizers of the problem in (5) for $w_s = \bar{w}_s$. Therefore,

$$\partial h_s(\bar{w}_s) = \mathrm{co}\{-\bar{x}_s : \bar{x}_s \text{ solves the problem in (5) for } w_s = \bar{w}_s\}\,. \tag{22}$$

Theorem 5.1. (Finite termination of BPH) *Let $h : \mathbb{R}^{nS} \to \mathbb{R}$ in* (5) *be the dual function associated with the primal problem* (1)*. If Algorithm* 2 *stops after finitely many iterations, then x^k and $x^{k+\frac{1}{2}}$ are equal and both are a solution of primal problem* (1)*, and w^k and $w^{k+\frac{1}{2}}$ are equal, and both are a solution of dual problem* (6)*.*

Proof. If the algorithm stops, then $\varphi^k_{w^k}(w^{k+\frac{1}{2}}) = 0$. Thus, the aggregate error and gradient defined in Lemma A.1 (stated in Appendix A) are null: $E^k = 0$ and $G^k = 0$. In particular, we also have that $w^k = w^{k+\frac{1}{2}}$ and $x^k = x^{k+\frac{1}{2}}$. Then, from Lemma A.1(ii), we have that $0 \in \partial(h + i_{\mathcal{N}^\perp})(w^k)$, that is, $w^k = w^{k+\frac{1}{2}} \in \mathcal{N}^\perp$ is a dual solution. Furthermore, since $G^k \in \partial\varphi^k_{w^k}(w^{k+\frac{1}{2}})$, we have that $G^k - x^k \in \partial h(w^{k+\frac{1}{2}})$. Thus $-x^k \in \partial h(w^{k+\frac{1}{2}})$, because $G^k = 0$. It follows from (22) that for all $s = 1, \dots, S$, $x^k_s \in C_s$ and it also solves $\min_{x_s \in C_s} \mathcal{L}_s(x_s, w^{k+\frac{1}{2}})$. Note that this means that x^k is primal feasible. Hence, since $w^{k+\frac{1}{2}}$ is a dual solution, then $x^k = x^{k+\frac{1}{2}}$ is a primal solution, as stated. □

We continue with the analysis of an infinite serious subsequence, and show the generated primal and dual points asymptotically solve the primal and dual problems (1) and (6). Recall that the dual form of the BPH algorithm fits the framework (13). In this case, the proper separation of isocost surfaces (8) is trivially satisfied because the function $H := h + i_{\mathcal{N}^\perp}$ is convex. Regarding the error bound (9), the condition is equivalent to requiring that, for every $v \geqslant \inf_{w \in \mathcal{N}^\perp} h(w)$, there exists $\varepsilon, \ell > 0$, such that whenever $w \in \mathcal{N}^\perp$, $h(w) \leqslant v$, and $x \in \mathbb{R}^{nS}$, with $\|x - P_{\mathcal{N}}[x]\|_S < \varepsilon$, and the s-component x_s solves the problem in (5), there holds that

$$d(w, \mathcal{S}) \leqslant \ell \|x\|_S, \tag{23}$$

where $\mathcal{S}$ is the set of minimizers of the dual problem.

Theorem 5.2. (Convergence of serious steps) *Let $h : \mathbb{R}^{nS} \to \mathbb{R}$ in* (5) *be the dual function associated with the primal problem* (1)*. Suppose, in addition, that the subdifferential error bound* (23) *is valid. In Algorithm* 2*, let $K_s := \{k : w^{k+1}$ was declared a serious step$\}$ and recall that, by construction $\{t_k\}_{K_s}$ is bounded below by $t_{\min} > 0$. Assume, additionally, that $\{t_k\}_{K_s}$ is bounded above by $t_{\max} > 0$.*

If the set K_s is infinite, the following holds for $k \in K_s$.

(i) *$\{h(w^k)\}_{K_s}$ monotonically converges to the dual optimal value h^*, such that the sequence of functional errors $\{v^k = h(w^k) - h^*\}_{K_s}$ converges to 0 with Q-linear rate: there exists $r \in (0, 1)$, such that for all sufficiently large $k \in K_s$,*

$$v^{k+1} \leqslant r v^k.$$

(ii) *The sequence of serious-step iterates $\{w^k\}_{K_s}$, as well as the intermediate points $\{w^{k+\frac{1}{2}}\}_{K_s}$, converge to a minimizer w^* of the dual problem* (6) *with R-linear rate: there exists $r \in (0, 1)$, and $c > 0$ such that for all sufficiently large $k \in K_s$,*

$$\|w^k - w^*\| \leqslant c\sqrt{r}^k, \quad \|w^{k+\frac{1}{2}} - w^*\| \leqslant c(2 - \sqrt{r})\sqrt{r}^k.$$

(iii) *The primal sequences* $\{x^k\}_{K_s}$ *and* $\{x^{k+\frac{1}{2}}\}_{K_s}$ *sub-sequentially converge to a solution of the primal problem* (1). *The functional values* $\{f(x^k)\}_{K_s}$ *and* $\{f(x^{k+\frac{1}{2}})\}_{K_s}$ *sub-sequentially converge to the optimal value of problem* (1).

Proof. Throughout, iterations parse $k \in K_s$. Items (i) and (ii) follow from Theorem 3.2, applied to the convex function h, which is $0-$weakly convex, and $\mathcal{M} = \mathcal{N}^\perp$. For item (iii), note that from Lemma A.1(iii), $G^k \to 0$. Therefore, the primal update in Algorithm 2 and (13a) imply $x^k - x^{k+\frac{1}{2}} \to 0$. Since both sequences are bounded, because $\{x^{k+\frac{1}{2}}\}$ belongs to a compact set and $P_{\mathcal{N}}$ is continuous, then both sequences $\{x^k\}$ and $\{x^{k+\frac{1}{2}}\}$ have the same accumulation points. In particular, any accumulation point x^* of these sequences belongs to $\mathcal{N}$, and for all scenario $s = 1, \ldots, S$, $x^*_s \in C_s$, from Lemma 4.2(i).

Furthermore, (13a) implies $G^{k_i} - x^{k_i} \in \partial h(w^{k_i+\frac{1}{2}})$, where $\{x^{k_i}\}$ is a subsequence that converges to x^*. Therefore, since ∂h is outer semicontinuous, $-x^* \in \partial h(w^*)$. In particular,

$$-f(x^{k_i+\frac{1}{2}}) - \langle w^{k_i+\frac{1}{2}}, x^{k_i+\frac{1}{2}} \rangle_S = h(w^{k_i+\frac{1}{2}}) \geqslant h(w^*) + \langle -x^*, w^{k_i+\frac{1}{2}} - w^* \rangle_S.$$

Taking the limit when $i \to +\infty$, it follows that $-f(x^*) \geqslant h(w^*)$, because we have $(x^*, w^*) \in \mathcal{N} \times \mathcal{N}^\perp$. Therefore, weak duality implies $-f(x^*) = h(w^*)$, with x^* being primal feasible. Hence, x^* is primal optimal.

Finally, sub-sequential convergence of $\{f(x^k)\}$ and $\{f(x^{k+\frac{1}{2}})\}$ to the optimal primal value $f(x^*)$ follows from continuity. □

To conclude, we show that when there are finitely many serious steps, the generated sequences also provide solutions for problems (1) and (6).

5.2. Tail of null steps

It remains to analyze the case when there is a last serious step $\widehat{w}$, performed at iteration $k = \widehat{k} - 1$. Accordingly, we let $K_n := \{k > \widehat{k} : w^k \text{ was declared a null step}\}$ denote the corresponding iteration index set. Note that for all $k \in K_n$, $w^k = \widehat{w}$. We will also assume that the stepsizes eventually stabilize along the tail of null steps. The next result shows that the last serious step is a solution of the dual problem, and that the accumulation points of the primal sequences are solutions of the primal problem.

Theorem 5.3. (Convergence of null steps) *Let* $h : \mathbb{R}^{nS} \to \mathbb{R}$ *in* (5) *be the dual function associated with the primal problem* (1). *Assume there is a last serious step* $\widehat{w}$*, followed by a tail of nulls steps. If the corresponding stepsizes eventually stabilize* ($t_k = \bar{t} \geqslant t_{\min}$ *for* $k \in K_n$ *sufficiently large), the following holds for Algorithm 2.*

(i) *Both* $\{w^{k+\frac{1}{2}}\}_{K_n}$ *and* $\{u^k\}_{K_n}$ *converge to* $\widehat{w}$*, which is also a solution for the dual problem* (6). *Furthermore, the sequences* $\{h(w^{k+\frac{1}{2}})\}_{K_n}$ *and* $\{h(u^k)\}_{K_n}$ *converge to the optimal value of the dual problem* (6).

(ii) *The primal sequences* $\{x^k\}_{K_n}$ *and* $\{x^{k+\frac{1}{2}}\}_{K_n}$ *subsequentially converge to a solution of the primal problem. The corresponding functional values* $\{f(x^k)\}_{K_n}$ *and* $\{f(x^{k+\frac{1}{2}})\}_{K_n}$ *subsequentially converge to the optimal value of problem* (1).

Proof. Throughout, consider $k \in K_n$ for which $t_k = \bar{t}$. We claim that $\{x^k\}$ is the result of applying a projected gradient method with constant stepsize $\frac{1}{\bar{t}}$ to the following problem

$$\begin{cases} \min & F(x) \\ \text{s.t.} & x \in \mathcal{N} \end{cases} \quad \text{for} \quad F(x) = \min_{y \in X} \left\{ f(y) + \langle \widehat{w}, y - x \rangle + \frac{\bar{t}}{2} \|y - x\|^2 \right\}. \tag{24}$$

By convexity of f, the objective function in (24) is strongly convex with modulus $\frac{1}{\bar{t}}$, and thus F has a Lipschitz continuous gradient with modulus $\bar{t}$. In particular, when $x = x^k$, by Lemma 4.2(i),

$$x^{k+\frac{1}{2}} = \arg\min_{y \in X} \left\{ f(y) + \langle \widehat{w}, y - x^k \rangle + \frac{\bar{t}}{2} \|y - x^k\|^2 \right\}$$

and

$$\nabla F(x^k) = -\widehat{w} + \bar{t} \left(x^k - x^{k+\frac{1}{2}} \right) .$$

Combined with the identity $\nabla F(x^k) = -w^{k+\frac{1}{2}}$ from (20), we see that

$$x^k - \frac{1}{\bar{t}} \nabla F(x^k) = x^k + \frac{1}{\bar{t}} w^{k+\frac{1}{2}} = x^{k+\frac{1}{2}} + \frac{1}{\bar{t}} \widehat{w}.$$

Projecting over $\mathcal{N}$ and recalling that $\hat{w} \in \mathcal{N}^\perp$ yields the claim, because

$$P_{\mathcal{N}} \left[x^k - \frac{1}{\bar{t}} \nabla F(x^k) \right] = P_{\mathcal{N}}[x^{k+\frac{1}{2}}] + \frac{1}{\bar{t}} P_{\mathcal{N}}[\widehat{w}] = P_{\mathcal{N}}[x^{k+\frac{1}{2}}] = x^{k+1} .$$

By compacteness of X, the sequence $\{x^k\}$ is bounded and the assumptions in [19, Theorem 4.1.3], stating convergence properties of gradient projected methods, are satisfied. Thus, $x^{k+1} - x^k \to 0$, and all accumulation points of $\{x^k\}$ are solutions to (24). Passing to the limit as $k \to +\infty$ in the identity $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] = \bar{t}(x^{k+1} - x^k)$ ensures that $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0$. In turn,

$$u^{k+1} - w^{k+\frac{1}{2}} = P_{\mathcal{N}^\perp}[w^{k+\frac{1}{2}}] - w^{k+\frac{1}{2}} = -P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0.$$

The sequence $\{w^{k+\frac{1}{2}} = \widehat{w} + \bar{t}(x^{k+\frac{1}{2}} - x^k)\}$ is bounded and so is $\{u^{k+1}\}$. Being convex and finite, h is Lipschitz continuous in any compact set that contains u^{k+1} and $w^{k+\frac{1}{2}}$, therefore

$$|h(u^{k+1}) - h(w^{k+\frac{1}{2}})| \leqslant L\|u^{k+1} - w^{k+\frac{1}{2}}\| \to 0, \text{ as } k \to +\infty.$$

Non-satisfaction of (13c) amounts to

$$h(u^{k+1}) - h(\widehat{w}) > m\big(h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}} \rangle - h(\widehat{w})\big),$$

which is equivalent to $h(u^{k+1}) - h(w^{k+\frac{1}{2}}) - m\langle x^k, w^{k+\frac{1}{2}} \rangle > (m-1)\big(h(w^{k+\frac{1}{2}}) - h(\widehat{w})\big)$. On the left-hand side, the inner product satisfies the relations

$$\langle x^k, w^{k+\frac{1}{2}} \rangle = \langle x^k, P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \rangle \to 0, \tag{25}$$

because $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0$ and $\{x^k\}$ is bounded.

Passing to the limit,

$$0 \geqslant (m-1)\liminf_k \left\{h(w^{k+\frac{1}{2}}) - h(\widehat{w})\right\} \iff h(\widehat{w}) \leqslant \liminf_k h(w^{k+\frac{1}{2}}), \tag{26}$$

because $m \in (0,1)$. On the other hand, by definition of $w^{k+\frac{1}{2}}$,

$$h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle + \frac{1}{2\bar{t}}\|w^{k+\frac{1}{2}} - \widehat{w}\|^2 \leqslant h(\widehat{w}).$$

Passing once again to the limit, using (26) and (25),

$$h(\widehat{w}) \leqslant \liminf_k h(w^{k+\frac{1}{2}}) \leqslant \liminf_k \left\{h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle + \frac{1}{2\bar{t}}\|w^{k+\frac{1}{2}} - \widehat{w}\|^2\right\} \leqslant h(\widehat{w}).$$

Take any accumulation point $w^\star$ of the bounded $\{w^{k+\frac{1}{2}}\}$ and consider a subsequence $w^{k_j+\frac{1}{2}} \to w^\star$, whenever $j \to +\infty$. Then $h(w^\star) + \frac{1}{2\bar{t}}\|w^\star - \widehat{w}\|^2 = h(\widehat{w})$. Moreover, from (26) it also holds that $h(\widehat{w}) \leqslant h(w^\star)$, and thus $h(w^\star) + \frac{1}{2\bar{t}}\|w^\star - \widehat{w}\|^2 \leqslant h(w^\star)$, which necessarily implies $w^\star = \widehat{w}$. This means that any accumulation point of $\{w^{k+\frac{1}{2}}\}$ is equal to the last serious step $\widehat{w}$, and thus $\{w^{k+\frac{1}{2}}\}$ converges to $\widehat{w}$, and since $u^{k+1} - w^{k+\frac{1}{2}} \to 0$, then $\{u^k\}$ also converges to $\widehat{w}$.

To prove that $\widehat{w}$ solves the dual problem, note that from (21) and (20),

$$G^k = x^{k+\frac{1}{2}} - x^k = \frac{1}{\bar{t}}(w^{k+\frac{1}{2}} - \widehat{w}) \to 0, \text{ as } k \to +\infty.$$

By continuity of h and (25), $\varphi^k_{\widehat{w}}(w^{k+\frac{1}{2}}) \to 0$. Lemma A.1(i) ensures that $E^k \to 0$ and Lemma A.1(ii) yields in the limit that $0 \in \partial(h + i_{\mathcal{N}^\perp})(\widehat{w})$, which means that $\widehat{w}$ minimizes $h + i_{\mathcal{N}^\perp}$.

As for the primal problem, since $x^{k+\frac{1}{2}} - x^k \to 0$, both sequences $\{x^k\}$ and $\{x^{k+\frac{1}{2}}\}$ have the same accumulation points. The remaining assertions follow as in the proof of Theorem 5.2(iii), now taking $k \in K_n$. ☐

6. Final remarks

We have introduced a new projective bundle method that, when applied to multistage programs, exploits parallelism and generates a serious subsequence converging with linear rate. The resulting Bundle Progressive Hedging, both in primal (Algorithm 3) and dual (Algorithm 2) forms, preserves the main features of the Progressive Hedging algorithm by T. Rockafellar and R. Wets, [31]. In particular, both methods solve separate scenario subproblems per iteration. These subproblems are strongly convex, thanks to the addition of a quadratic term related to an approximate augmented Lagrangian of the problem. Furthermore, the PH algorithm and its bundle BPH variant both use projections onto $\mathcal{N}$ and $\mathcal{N}^\perp$ to respectively ensure feasibility in the primal and dual spaces.

Besides these similarities, our proposal adds features typical from the bundle methodology to measure the quality of the approximation of the augmented Lagrangian of the full problem. By contrast, the original PH method uses the dual information obtained with an approximate Lagrangian without further ado.

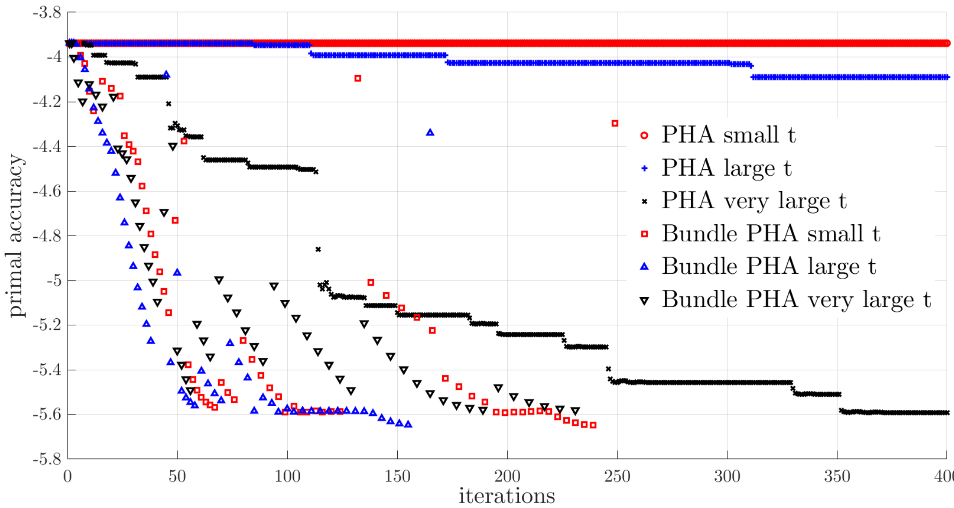

Figure 6.1: Accuracy in the primal solution for the PH and the BPH algorithms (the ordinate reports the negative of the number of digits obtained at a given iteration). Both methods were run for the same random instance with 50 scenarios, taking initial proximal stepsizes $t_0 \in \{1, 10, 100\}$, being small t, large t, and very large t, respectively. The values of t_k are maintained fixed to t_0 with the PH variant, yielding good accuracy only for the largest value $t_0 = 100$. With BPH, t_k varies according to the serious/null step rules and the performance is more stable.

Thanks to the bundle techniques, it is possible to dynamically adjust the augmentation parameter in the PH algorithm without impairing its convergence. Figure 6.1 gives a simple, yet illustrative, instance of a randomly generated linear problem with 50 scenarios. The impact of keeping t_k fixed along the iterative procedure is clear. With $t_k = t_0$, the PH approach only reaches a good accuracy if the parameter is sufficiently large ($t_0 = 100$, named "very large t" in the figure). The BPH method, with its adaptive adjustment of stepsizes, seems less sensitive to the initial value of t_0.

With the bundle approach, optimality primal and dual certificates are available, based on the aggregate information constructed along iterations (the PH algorithm, by contrast, measures distance to the primal-dual set of minimizers). However, those enhancements require additional evaluations of the dual functions at each iteration. This involves a new set of computations, that can be done in parallel and are similar to the PH subproblems, barring the quadratic term in (15).

The projective bundle method in Section 3 is general and can be used to extend existing approaches to the weakly convex setting. As often in bundle methods, the crucial point is the ability to show that, when the method enters an infinite loop of null steps, the family of model functions drives iterates to the last generated serious point. The stumbling block lies in the suitable definition of the models, as it was done in [34] for Taylor models in composite optimization, or in Section 4 for the particular dual function (5).

Acknowledgements. The authors are grateful to the reviewer and editor for beneficial comments. Research of the first author is supported by FAPESP Grant 2019/20023-1. The second author is partly funded by CEPID CeMEAI, CNPq Grant 306089/2019-0, and PRONEX–Optimization.

A. Projective model-based schemes

We show that if the scheme presented in section 3.1 generates an infinite sequence of serious iterates, they globally converge with local linear rate. The following results correspond to an extension of the analysis in [4] that takes into account the additional projection step in (13b), to deal with a feasible set.

A.1. Technical results preliminary to Theorem 3.2

Similarly to [4], the first step is to exploit the model convexity to transport the model subgradient to an approximate subgradient of the "convexified" function associated to weak convexity. Then, a Brøndsted-Rockafellar's like relation makes the connection with the objective function in (7). The corresponding results are gathered in the following statement, listing several technical relations. None of those relations involves the projection step, they can be shown following the arguments in [4, Proposition 5.4 and Lemma 5.6], applied to the function $H = h + i_{\mathcal{M}}$.

Lemma A.1. (Transportation of subgradients, descent, theoretical sequence) *Let* $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ *and* $\mathcal{M}$ *satisfy* (A1)−−(A2). *Consider solving problem* (7) *applying the model-based proximal scheme in* (13). *If the models* $\varphi^k_{w^k}$ *are of type 1QA with parameter* $q \leqslant \rho$*, as in Definition 3.1, the following holds for all* k*.*

(i) *The aggregate error* $E^k := -t_k\|G^k\|^2 - \varphi^k_{w^k}(w^{k+\frac{1}{2}})$ *satisfies* $E^k \geqslant 0$.

(ii) $G^k \in \partial_{E^k} H_{w^k}(w^k)$ *where* $H_{w^k}(\cdot)$ *denotes the "convexified" function* $h(\cdot) + i_{\mathcal{M}}(\cdot) + \frac{\rho}{2}\|\cdot - w^k\|^2$

If (A3) *holds, and the proximal stepsizes* t_k *are bounded,* $t_k \in [t_{\min}, t_{\max}]$*, then*

(iii) *the sequences of model subgradient and errors* $\{G^k\}$ *and* $\{E^k\}$ *converge to* 0 *as* $k \to \infty$*, and*

(iv) *condition* (13c) *is equivalent to*

$$h(w^{k+1}) + \frac{m}{t_{\max}}\left(\|w^{k+\frac{1}{2}} - w^k\|^2 + t_k E^k\right) \leqslant h(w^k)\,. \tag{27}$$

If, in addition, (A5) *holds, then*

(v) *there exists a theoretical auxiliary sequence* $\{z^k\}$ *such that* $\|z^k - w^k\| \leqslant \sqrt{\ell E^k}$ *and, for* $b > 0$,

$$\exists g^k \in \partial h(z^k) \text{ and } \nu^k \in N_{\mathcal{M}}(z^k),\ \|g^k + \nu^k\| \leqslant b\Big(\|w^{k+\frac{1}{2}} - w^k\| + \|w^k - z^k\|\Big)\,. \tag{28}$$

Proof. Items (i)–(iv) follow directly from [4, Proposition 5.4(i)–(iv)] applied to $h + i_{\mathcal{M}}$. Item (v) follows from [4, Lemma 5.6] applied to $h + i_{\mathcal{M}}$. Indeed, apply [29, Theorem 2] to the function H_{w^k} and the pair of points (G^k, w^k) such that $G^k \in \partial_{E^k} H_{w^k}(w^k)$, it follows that there exists d^k such that $\|d^k\| \leqslant \sqrt{E^k}$ and

$$G^k - \frac{1}{\sqrt{\ell}} d^k \in \partial H_{w^k}(w^k + \sqrt{\ell} d^k),$$

where $\ell > 0$ is the constant of condition (A5). Defining $z^k := w^k + \sqrt{\ell} d^k$, we have that $\|z^k - w^k\| \leqslant \sqrt{\ell E^k}$, and $\partial H_{w^k}(z^k) = \partial(h + i_{\mathcal{M}})(z^k) + \rho\sqrt{\ell} d^k$.

Therefore, for $g^k_{\mathcal{M}} := G^k + (1+\rho\ell)\sqrt{\ell^{-1}}d^k$, it holds that

$$g^k_{\mathcal{M}} \in \partial(h + i_{\mathcal{M}})(z^k) = \partial h(z^k) + N_{\mathcal{M}}(z^k).$$

Thus, there exists $g^k \in \partial h(z^k)$ and $\nu^k \in N_{\mathcal{M}}(z^k)$ such that $g^k_{\mathcal{M}} = g^k + \nu^k$. Moreover, from the definition of $g^k_{\mathcal{M}}$, d^k, and (13a), we also have

$$\begin{aligned}\|g^k + \nu^k\| &\leqslant \|G^k\| + (1+\rho\ell)\sqrt{\ell^{-1}}\|d^k\| \\ &= \frac{1}{t_k}\|w^k - w^{k+\frac{1}{2}}\| + (1+\rho\ell)\ell^{-1}\|z^k - w^k\| \\ &\leqslant \frac{1}{t_{\min}}\|w^k - w^{k+\frac{1}{2}}\| + (1+\rho\ell)\ell^{-1}\|z^k - w^k\|\,,\end{aligned}$$

then condition (28) holds for $b := \max\{t_{\min}^{-1}, (1+\rho\ell)\ell^{-1}\}$. □

Lemma A.2. *Let $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{M}$ satisfy* (A1)–(A3), *and that the stepsizes t_k remain in $[t_{\min}, t_{\max}]$. Then the following holds:*

(i) *$\{h(w^k)\}$ monotonically converges to some value $\tilde{h} \in \mathbb{R}$.*

(ii) *$w^{k+1} - w^k \to 0$, $z^{k+1} - z^k \to 0$ and $g^k + \nu^k \to 0$, as $k \to +\infty$.*

Suppose, in addition, (A4) *and* (A5) *hold. Then*

(iii) *$\{h(z^k)\}$ and $\{h(w^k)\}$ both converge to h^*, where $h^* \in \mathbb{R}$ is a critical value (i.e., $h^* = h(w)$ for some $w \in \mathcal{S}$).*

(iv) *Defining $\tilde{p}^k \in P_{\mathcal{S}}(z^k)$, for all k sufficiently large, the distance from z^k to S can be estimated as*

$$\|z^k - \tilde{p}^k\|^2 \leqslant \frac{2\ell^2 b^2}{a}(h(w^k) - h(w^{k+1})) + 2\ell^2 b^2 \|w^k - z^k\|^2\,.$$

(v) *For the functional value errors $v^k := h(w^k) - h^*$, it holds that*

$$v^{k+1} \leqslant \frac{2\ell b^2}{a}(v^k - v^{k+1}) + 2\ell b^2 \|w^k - z^k\|^2 + \Theta^k\,,$$

where $$\Theta^k := h(w^k) - h(z^k) + \frac{\rho}{2}\|\tilde{p}^k - z^k\|^2\,.$$

Proof. The following proof is an adaptation of the proof of [4, Lemma 4.2], taking into consideration the extra projection step in the present setting. To see item (i), notice that, since $\mathcal{M}$ is convex, then $P_{\mathcal{M}}$ is a non-expansive operator, and together with $w^k \in \mathcal{M}$, imply

$$\|w^{k+1} - w^k\| = \|P_{\mathcal{M}}(w^{k+\frac{1}{2}}) - P_{\mathcal{M}}(w^k)\| \leqslant \|w^{k+\frac{1}{2}} - w^k\|$$

Combined with (27), it yields for some constant $a > 0$

$$h(w^{k+1}) + a(\|w^{k+1} - w^k\|^2 + \varepsilon_k) \leqslant h(w^k) \tag{29}$$

which, in particular, implies that the sequence of function values of $\{w^k\}$ is non-increasing. Since h is bounded below on $\mathcal{M}$, monotonicity of $\{h(w^k)\}$ implies that

$h(w^k) \to \tilde{h}$, for some $\tilde{h} \in \mathbb{R}$. Note that inequality (29) corresponds to [4, (2a)] for $\varepsilon_k = t_{\min} E_k$.

For item (ii), observe that (29) also implies, combined with the fact that $h(w^{k+1}) - h(w^k) \to 0$, that $w^{k+1} - w^k \to 0$. Note that $w^{k+\frac{1}{2}} - w^k \to 0$ follows similarly from (27). Together with Lemma A.1(v), imply $g^k + \nu^k \to 0$.

Item (iii) follows directly from [4, Lema 4.2 (iii)], because this original result does not depend directly on (28). In particular, since $g^k + \nu^k \to 0$ and $h(z^k) \to h^*$, we can apply the error bound to obtain, for all sufficiently large k,

$$\|z^k - \tilde{p}^k\| \leqslant \ell \|g^k + \nu^k\| . \tag{30}$$

Regarding item (iv), from (28) and (29), there holds

$$\begin{aligned}
\|g^k + \nu^k\|^2 &\leqslant b^2 (\|w^{k+\frac{1}{2}} - w^k\| + \|w^k - z^k\|)^2 \\
&\leqslant 2b^2 \|w^{k+\frac{1}{2}} - w^k\|^2 + 2b^2 \|w^k - z^k\|^2 \qquad (31)\\
&\leqslant \frac{2b^2}{a} (h(w^k) - h(w^{k+1})) + 2b^2 \|w^k - z^k\|^2 , \qquad (32)
\end{aligned}$$

and the results follows from (30). The final item (v) is just [4, Lema 4.2 (v)]. □

A.2. Proof of Theorem 3.2

The results of global convergence and local rate of convergence are similar to [4, Theorem 5.7], now taking into account the projection step (13b).

First, let $\{z^k\}$ be defined as in Lemma A.1(v), and consider the sequence of functional errors $\{v^k\}$, defined in Lemma A.2(v). Then, there exist constants $C_1, C_2 > 0$, such that for all sufficiently large k, there holds

$$h(w^k) - h(z^k) \leqslant C_1 (v^k - v^{k+1}) \tag{33}$$

and

$$\|w^k - z^k\|^2 \leqslant C_2 (v^k - v^{k+1}) . \tag{34}$$

Indeed, these two estimates directly follow from the proof of [4, Theorem 5.7]. Estimate (34) comes from the definition of z^k, and the descent condition (27), while estimate (33) follows from the E^k-subgradient inequality of H_{w^k} for G^k and (34).

The proof of item (i) in Theorem 3.2 is exactly the same as for [4, Theorem 4.3(i)], due to the fact that it depends on Lemma A.2. Indeed, use (33), (34), and Lemma A.2(iv) in the estimate of Lemma A.2(v), and then rearrange terms to deduce (i) for $r = M/(1+M) \in (0,1)$, and $M = C_1 + \ell b^2 (2 + \rho \ell)(1/a + C_2)$.

To deduce Theorem 3.2(ii), we can apply [4, Lemma 4.1], using (29) (which is exactly the same as [4, (2a)]) and item (i). Therefore, $\{w^k\}$ converges to some w^* with a R-linear rate: for all sufficiently large k,

$$\|w^k - w^*\| \leqslant c \sqrt{r}^k , \text{ where } \quad c = \frac{\sqrt{v^0}}{\sqrt{a}\,(1 - \sqrt{r})}.$$

The fact that w^* is critical can be similarly proven as in the original analysis, using Lemma A.1(v) and the fact that the subdifferential ∂h is an upper semicontinuous multi-function.

Regarding the sequence of intermediate points, for all k,

$$\|w^{k+\frac{1}{2}} - w^*\| \leqslant \|w^{k+\frac{1}{2}} - w^k\| + \|w^k - w^*\|$$

In the right-hand side, the second term is bounded by $c\sqrt{r}^k$, therefore it only remains to bound the first one. From (29) and the fact that $\varepsilon_k, v^k \geqslant 0$, it follows that

$$\|w^{k+\frac{1}{2}} - w^k\|^2 \leqslant \frac{1}{a}(h(w^k) - h(w^{k+1})) = \frac{1}{a}(v^k - v^{k+1}) \leqslant \frac{1}{a}v^k$$

From (ii), it holds that $$\|w^{k+\frac{1}{2}} - w^k\|^2 \leqslant \frac{v^0}{a}r^k$$

Therefore, $$\|w^{k+\frac{1}{2}} - w^*\| \leqslant c(2 - \sqrt{r})\sqrt{r}^k,$$

which concludes the proof of Theorem 3.2.

References

[1] W. van Ackooij, V. Berge, W. de Oliveira, C. Sagastizábal: *Probabilistic optimization via approximate p-efficient points and bundle methods*, Comp. Oper. Res. 77 (2017) 177–193.

[2] W. van Ackooij, C. Sagastizábal: *Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems*, SIAM J. Optimization 24/2 (2014) 733–765.

[3] P. Apkarian, D. Noll, A. Rondepierre: *Mixed H_2/H_∞ control via nonsmooth optimization*, SIAM J. Optimization Control 47/3 (2008) 1516–1546.

[4] F. Atenas, C. Sagastizábal, P. J. S. Silva, M. Solodov: *A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods*, SIAM J. Optimization 33/1 (2023) 89–115.

[5] H. Attouch, J. Bolte, B. F. Svaiter: *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Programming 137/1-2 (2013) 91–129.

[6] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, C. Sagastizábal: *Numerical Optimization: Theoretical and Practical Aspects*, 2nd ed., Springer, Berlin (2006).

[7] P. L. Combettes, J. Eckstein: *Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions*, Math. Programming 126/1-2 (2018) 645–672.

[8] D. Drusvyatskiy, A. D. Ioffe, A. S. Lewis: *Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria*, Math. Programming 185/1 (2021) 357–383.

[9] J. Eckstein: *Splitting Methods for Monotone Operators, with Applications To Parallel Optimization*, PhD Thesis, MIT, Cambridge (1989).

[10] J. Eckstein: *A simplified form of block-iterative operator splitting and an asynchronous algorithm resembling the multi-block alternating direction method of multipliers*, J. Optimization Theory Appl. 173/1 (2017) 155–182.

[11] J. Eckstein, B. F. Svaiter: *A family of projective splitting methods for the sum of two maximal monotone operators*, Math. Programming 111/1-2 (2008) 173–199.

[12] J. Eckstein, B. F. Svaiter: *General projective splitting methods for sums of maximal monotone operators*, SIAM J. Control Optimization 48/2 (2009) 787–811.

[13] J. Eckstein, J.-P. Watson, D. L. Woodruff: *Projective hedging for stochastic programming*, Optimization Online, published 2022-05-06 (2022), 39p.

[14] B. Gilles, Y. Laguel, D. Grishchenko, F. Iutzeler, J. Malick: *Randomized progressive hedging methods for multi-stage stochastic programming*, Ann. Oper. Res. 295/2 (2020) 535–560.

[15] R. E. C. Gonçalves, E. C. Finardi, E. L. da Silva: *Applying different decomposition schemes using the progressive hedging algorithm to the operation planning problem of a hydrothermal system*, Electric Power Syst. Res. 83/1 (2012) 19–27.

[16] W. Hare, C. Sagastizábal: *Computing proximal points of nonconvex functions*, Math. Programming 116/1-2 (2009) 221–258.

[17] W. Hare, C. Sagastizábal: *A redistributed proximal bundle method for nonconvex optimization*, SIAM J. Optimization 20/5 (2010) 2442–2473.

[18] J. B. Hiriart-Urruty, C. Lemaréchal: *Convex Analysis and Minimization Algorithms I, II*, Grundlehren der mathematischen Wissenschaften Band 305 und 306, Springer, Berlin (1996).

[19] A. Izmailov, M.Solodov: *Optimization: Computational Methods (Portuguese)*, 3rd ed., IMPA, Rio de Janeiro (2018).

[20] E. Karas, A. Ribeiro, C. Sagastizábal, M. V. Solodov: *A bundle-filter method for nonsmooth convex constrained optimization*, Math. Programming B 116/1-2 (2009) 297–320.

[21] B. Knueven, D. T. Mildebrath, C. Muir, J. D. Siirola, J.-P. Watson, D. L. Woodruff: *A parallel hub-and-spoke system for large-scale scenario-based optimization under uncertainty*, Optimization Online, published 2022-09-09 (2022), 30 p.

[22] G. Li, T. K. Pong: *Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods*, Found. Comp. Math. 18/5 (2018) 1199–1232.

[23] P. L. Lions, B. Mercier: *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Analysis 16/6 (1979) 964–979.

[24] P. L. Lions, B. Mercier: *Augmented Lagrangian methods for the solution of variational problems*, in: *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, R. Glowinski et al. (eds.), Society for Industrial and Applied Mathematics, Philadelphia (1989) 45–121.

[25] Z.-Q. Luo, P. Tseng: *Error bounds and convergence analysis of feasible descent methods: a general approach*, Ann. Oper. Res. 46/1 (1993) 157–178.

[26] R. Mifflin: *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optimization 15/6 (1977) 959–972.

[27] V. I. Norkin: *Generalized-differentiable functions*, Cybernetics 16/1 (1980) 10–12.

[28] W. de Oliveira, C. Sagastizábal: *Bundle methods in the XXIst century: a bird's-eye view*, Pesquisa Operacional 34/3 (2014) 647–670.

[29] S. M. Robinson: *Linear convergence of epsilon-subgradient descent methods for a class of convex functions*, Math. Programming 86/1 (1999) 41–50.

[30] R. T. Rockafellar: *Solving stochastic programming problems with risk measures by progressive hedging*, Set-Valued Var. Analysis 26/4 (2018) 759–768.

[31] R. T. Rockafellar, R. J.-B. Wets: *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res. 16/1 (1991) 119–147.

[32] R. T. Rockafellar, R. J.-B. Wets: *Variational Analysis*, Grundlehren der mathematischen Wissenschaften Band 317, Springer, Berlin (2009).

[33] C. Sagastizábal: *Divide to conquer: decomposition methods for energy optimization*, Math. Programming 134/1 (2012) 187–222.

[34] C. Sagastizábal: *Composite proximal bundle method*, Math. Programming 140/1 (2013) 189–233.

[35] C. Sagastizábal, M. V. Solodov: *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optimization 16/1 (2005) 146–169.

[36] M. L. L. dos Santos, E. Luiz da Silva, E. C. Finardi, R. E. C. Gonçalves: *Practical aspects in solving the medium-term operation planning problem of hydrothermal power systems by using the progressive hedging method*, Int. J. Electrical Power Energy Syst. 31/9 (2009) 546–552.

[37] P. Tseng: *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optimization 29/1 (1991) 119–138.

[38] J. J. Ye, X. Yuan, S. Zeng, J. Zhang: *Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems*, Set-Valued Var. Analysis 29/4 (2021) 803–837.