, , 1–18 ()

Bundle methods in stochastic optimal power management: a disaggregated approach using preconditioners

LÉONARD BACAUD bacaudl@ensta.fr
EdF, Dépt. Méthodes d'Optimisation et de Simulation, 92141 Clamart, France

CLAUDE LEMARÉCHAL claude.lemarechal@inrialpes.fr
INRIA-Rhône-Alpes, 655 avenue de l'Europe 38330 Montbonnot, France

ARNAUD RENAUD arnaud.renaud@edf.fr
EdF, Dépt. Méthodes d'Optimisation et de Simulation, 92141 Clamart, France

CLAUDIA SAGASTIZÁBAL sagastiz@impa.br
INRIA-Rocquencourt, BP 105, 78153 Le Chesnay, France.
Also IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil.

Editor: Computational Optimization and Applications

Abstract. A specialized variant of bundle methods suitable for large-scale problems with separable objective is presented. The method is applied to the resolution of a stochastic unit-commitment problem solved by Lagrangian relaxation. The model includes hydro- as well as thermal-powered plants. Uncertainties lie in the demand, which evolves in time according to a tree of scenarios. Dual variables are preconditioned by using probabilities associated to nodes in the tree. The approach is illustrated by numerical results, obtained on a model of the French production mix over a time horizon of 10 days and 1 month.
AMS Subject Classification (1991): Primary 65K05, 90B30; Secondary 90C25

Keywords: optimization, bundle methods, stochastic optimization, Lagrangian relaxation, unit-commitment problems, preconditioning

1. Introduction

We are interested in the use of bundle methods to solve large-scale mixed-integer problems, possibly nonconvex. Specifically, consider optimization problems arising in electrical power management. Given an electric generation mix, the aim is to minimize production costs subject to operating constraints of generation units and other external constraints, like network flow capacities. There are many different problems fitting such a large framework. In particular, the time horizon chosen for the scheduling highly determines the specificity of problems. Short, middle and long term decisions have their own peculiarities that need to be reflected in the modeling. Short term problems are generally modeled in a deterministic framework, see [1], [7]. For longer terms, inherent uncertainties may result in poor solutions if a deterministic model is still used. Consider for instance the French case, where winter demand has uncertainties reaching up to several thousands of MW. When comparing this value to typical peak loads (70000 MW), we see that for the modeling to yield any significant values, it must explicitly incorporate the stochastic nature of the problem.

A generic formulation for the optimal power management problem is the following:

$$\begin{cases} \min_p \mathcal{C}(p) \\ p \in \mathcal{D} \cap \mathcal{S}, \end{cases} \tag{1}$$

where p is the production, $\mathcal{C}$ is the operating cost function and $\mathcal{D}$, $\mathcal{S}$ represent the constraints. More precisely, $\mathcal{D}$ includes all the dynamic constraints (expressing operating rules for each type of power plant) while $\mathcal{S}$ contains static constraints. Static constraints can be split into two categories, $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$. The first category gathers *coupling* constraints which relate production of different units (satisfaction of demand, network security, spinning reserve). The second category represents static *non-coupling* constraints, like bounds and integrality constraints. Typically, coupling static constraints are represented by a large-scale linear system.

As a result, (1) is a large-scale mixed integer optimization problem. It can be solved by using Lagrangian relaxation, although other methods had been proposed in the literature, see [22], [7]. An important feature of the Lagrangian relaxation approach is that it provides marginal prices that are useful for the operator of the system. It also gives good starting points for recovering primal solutions. We mention in passing that similar optimization problems in optimal power management (different from the particular monopolistic French case), such as risk minimization in a free market or transactions in a Spot market, can be effectively solved with the same technique.

In this paper, we focus on efficient methods to solve a non-differentiable problem, dual to problem (1), in the French context. For the sake of simplicity, we ignore failures and we drop network constraints but we consider a stochastic model for the demand. The dual problem is obtained by using a Lagrangian relaxation technique, called space decomposition in [16]. We explain this technique in § 3.1, as well as associated master program and subproblems. To solve the (non-differentiable) master program we use a *disaggregated* form of bundle methods. In § 4 we introduce this method and a preconditioner that has proved to be very efficient for numerical purposes. Finally, to assess our approach, we report in § 5 numerical tests on a battery of *Electricité de France* (EDF) power generation problems.

2. Modeling

This section describes the main elements taken into account to define our particular power production problem. First, we specify the model chosen for the demand, knowing that other sources of uncertainty, such as failures on power units or natural inflows, can be treated likewise. In § 2.2 we briefly describe two important models of power generation units.

2.1. *A stochastic framework for the demand*

We choose a dynamic stochastic recourse model for the demand. Contrary to a deterministic approach, the stochastic modeling yields an optimal **set** of schedules, called a *strategy*. There are several possible formulations of stochastic programming adapted to our problem (see for instance [20], [22], [9]).

Essentially, the demand depends on weather, i.e., on temperature, whose expected evolutions or scenarios are provided by forecast centers. Scenarios can also be generated using historical series. We do not go into more detail here, we refer to [23] and [3], where several strategies are discussed.

Assume that for a given demand d at time t, demand at time $t+1$ can only take a finite number of values $d_1, \ldots, d_k$ with probabilities $\pi_1, \ldots, \pi_k$, obtained using the temperature scenarios. Then, by induction, at any given time the demand can only take a finite number of values that can be organized in a tree, as shown in Figure 1. The root corresponds to the first instant for which the demand is known.

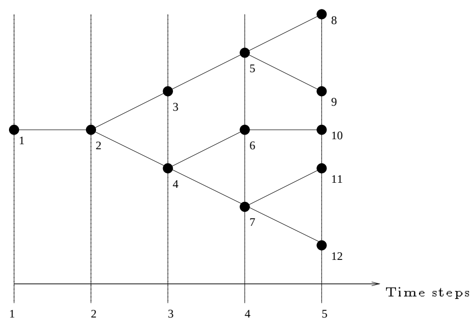

Figure 1. Describing stochastic demands by means of a scenario tree

Ordering these nodes by time steps, we observe that they represent all the possible values for the demand, at any given time. Edges in the tree correspond to feasible transitions between nodes, while one demand scenario corresponds to a path from the root to one particular final node, or *leaf*. Let T be the time horizon and let N be the number of nodes in the tree. To each node n three parameters can be associated:

- the time step corresponding to node n, denoted by $\tau(n)$,
- the demand at node n, denoted by d_n, and
- the probability π_n of reaching node n from the root.

Note that for each $t = 1, \ldots, T$ we have $\sum_{\{n:\tau(n)=t\}} \pi_n = 1$. The stochastic problem is a particular instance of (1), with cost function $\mathcal{C}$ defined as the total expectation of operating costs over all scenarios. More precisely, let our power generation mix be composed of I units and let p_n^i be the production of unit i at node n. Then

$$p := (p_n^i)_{n \le N}^{i \le I} \in \mathbb{R}^N \times \mathbb{R}^I$$

stands for the production of the whole mix. The expected cost has the form

$$\mathcal{C}(p) := \sum_{i \le I} \sum_{n \le N} \pi_n \mathcal{C}_n^i(p_n^i) \,,$$

where $\mathcal{C}_n^i$ is the operating cost function of unit i at node n.

Dynamic constraints describe feasible operating domains for each unit:

$$\mathcal{D} = \prod_{i \le I} \mathcal{D}^i \,,$$

i.e., $\mathcal{D}^i$ is the dynamic operating domain of unit i for the entire scenario tree. Hence, each $\mathcal{D}^i$ is a subset of $\mathbb{R}^N$ whose elements are feasible production schedulings, the same for each scenario in the tree. In § 2.2 we give a short description of the different units i and sets $\mathcal{D}^i$ found in the French mix.

Water inflows or market prices can also be considered as stochastic variables and indexed by the nodes of the scenario tree. However, the size of the problem is proportional to the number of nodes so the number of scenarios needs to be limited. Because in France electrical heating makes the demand the main source of uncertainty, we have chosen to focus on this variable for our modeling. In this respect we set:

$$\mathcal{S}_1 = \left\{ p : \textstyle\sum_{i \le I} p_n^i = d_n \,, \text{for all } n \le N \right\} . \tag{2}$$

It is possible to include in our model spinning reserve constraints. Spinning reserve is used to compensate unexpected events in a specified short-time period. Formally, reserve constraints are the same than demand contraints, so we do not make the former explicit. Notwithstanding, their inclusion leads to higher dimensions in the dual.

Throughout this paper we use superscripts in vector p to denote the stochastic process representing the production of unit i. Its components are nodes of the scenario tree. With this convention, p^i is a vector in $\mathbb{R}^N$.

2.2. *Power generation units*

The model we have considered until now is based on (1), where it was assumed that each power plant can be described straightforwardly in terms of production variables p^i. Actually, sophisticated models use state and control variables, denoted respectively by x and u, so that $p^i = p^i(x, u)$. At first sight, this modification comes just to a change of variables in (1). However, it is important to note that, although (1) is formally the same, the introduction of new variables changes some constraints. We will see next that coupling constraints in (2) may no longer be linear, or even convex, when the dependence of p^i on x and u is made explicit.

Now we analyze two important types of power generation units, namely hydraulic and thermic plants.

Hydro-systems. A hydro-valley is a set of interconnected plants and reservoirs with natural inflows in each reservoir. Figure 2 below illustrates a hydro-valley composed of 5 plants and 4 reservoirs. The state variable $x = (x_j)_{1 \le j \le 4}$ is a vector whose components are the contents of each reservoir. The control variable $u = (u_k)_{1 \le k \le 5}$ describes volumes to be discharged by each plant.

Dynamic constraints are the stream-flow balance equations. In our model, we consider the entire hydro-valley as a single production unit, so its generation is obtained by adding the production of all the plants in the valley. Suppose that each plant production depends bilinearly on the discharge and water-head of the reservoir connected to it. Assume also that water-head is modeled as a quadratic function of the contents. Then, for our hydro-valley in Fig. 2, the power production of hydro-plant # 3 is

$$u_3(\alpha x_2^2 + \beta x_2 + \gamma)$$

where α, β and γ are parameters depending on the characteristics of reservoir # 2.

Finally, operating costs for each reservoir are null for $\tau(n) < T$. Only the water-value for $\tau(n) = T$ is taken into account. It is modeled with a quadratic function of the water contents.

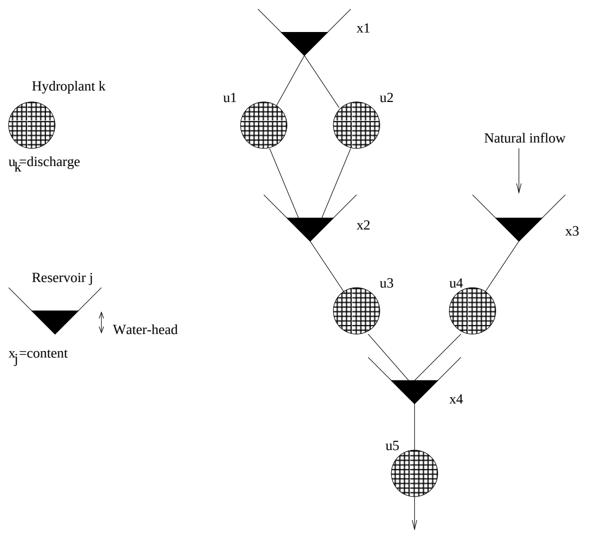

Figure 2. A hydro-valley is a unit formed by interconnected reservoirs.

Thermal units. In the example above, the state variable stands for a volume and the control variable represents an amount of water to be discharged, so variables are continuous. Here, by contrast, the state variable is an integer indexing operating modes (on/off, for instance). The control variable is associated to feasible transitions between states.

Dynamic constraints are modeled by a graph: nodes are associated with an operating mode, a production level and a cost. Clearly, both constraints sets $\mathcal{S}_2$ and $\mathcal{D}$ are non-convex. $\square$

These two examples illustrate two general types of plant modeling. Other generation units, such as nuclear power plants, can be modeled in a similar way, see [5].

3. Lagrangian relaxation

Altogether, our stochastic power management problem has the form

$$\begin{cases} \min_p \sum_{i\leq I}\sum_{n\leq N}\pi_n\mathcal{C}_n^i(p_n^i) & \\ p=(p_n^i)_{n\leq N}^{i\leq I}=(p^i)^{i\leq I} & \\ p^i\in\mathcal{D}^i\cap\mathcal{S}_2^i \text{ for all } i\leq I & \text{(a)} \\ \sum_{i\leq I}p_n^i=d_n \text{ for all } n\leq N & \text{(b).} \end{cases} \tag{3}$$

Both local costs $\sum_{n\leq N}\pi_n\mathcal{C}_n^i(p_n^i)$ and constraints (3)(a) are specified for each generation unit, so they can be considered separately for each $i\in I$. It is not the case of the N static constraints in (3)(b), which describe the state of the whole mix at each time step. Next we explain how to take care of these coupling constraints by using Lagrangian relaxation techniques.

In the sequel we use the terminology "space" decomposition to refer to separability along units $i\in I$. This wording is used in opposition to "space-time" decompositions, that have been proposed to deal with network constraints, see [16].

3.1. A space decomposition scheme

We review here some notions of duality theory and classical Lagrangian relaxation, see for instance [4, Ch. VIII]. The so-called *stochastic Lagrangian* of [8], [19], differs from our approach essentially in the use of a probability-induced scalar product in (4) below (see Remark 2).

Associated to (3) there is the Lagrangian

$$L(p,\lambda):=\sum_{i\leq I}\Big(\sum_{n\leq N}\pi_n\mathcal{C}_n^i(p_n^i)-\langle\lambda,p^i\rangle\Big)+\langle\lambda,d\rangle\,, \tag{4}$$

where $\lambda,d\in\mathbb{R}^N$ are, respectively, the vectors of Lagrange multipliers and of demands. Without any assumptions the weak duality relationship

$$\min_{p\in\mathcal{D}\cap\mathcal{S}_2}\max_{\lambda\in\mathbb{R}^N}L(p,\lambda)\geq\max_{\lambda\in\mathbb{R}^N}\min_{p\in\mathcal{D}\cap\mathcal{S}_2}L(p,\lambda) \tag{5}$$

holds. The left-hand side of (5), or *primal* problem, has the same optimal value as (3). The right-hand side in (5) is the *dual* or *master* problem

$$\max_{\lambda \in \mathbb{R}^N} \theta(\lambda)\,, \tag{6}$$

where the objective is the *dual* function $\theta(\lambda) := \min_{p \in \mathcal{D} \cap \mathcal{S}_2} L(p, \lambda)$.
This concave non-differentiable function inherits the separable structure in (4). Letting

$$\theta^i(\lambda) := \min_{p^i \in \mathcal{D}^i \cap \mathcal{S}_2^i} \sum_{n \leq N} \pi_n \mathcal{C}_n^i(p_n^i) - \langle \lambda, p^i \rangle \text{ for all } i \leq I\,, \tag{7}$$

we see that

$$\theta(\lambda) = \sum_{i \leq I} \theta^i(\lambda) + \langle \lambda, d \rangle\,. \tag{8}$$

A key observation is that, for given $\bar{\lambda}$, any $\bar{p}^i \in \mathcal{D}^i \cap \mathcal{S}_2^i$ realizing the minimum in (7) satisfies the relation

$$\theta^i(\lambda) \leq \theta^i(\bar{\lambda}) + \langle \bar{\lambda} - \lambda, \bar{p}^i \rangle \quad \text{for all } \lambda \in \mathbb{R}^N\,.$$

As a result, $\bar{p}^i$ is a subgradient at $\bar{\lambda}$ of the convex function $-\theta^i$. In this case, we will write $\bar{p}^i \in \partial\theta^i(\bar{\lambda})$ to alleviate notation.

Lagrangian relaxation techniques are appealing when computing the dual function is an easy task, at least when compared to the problem of solving the primal. This is precisely the case of (3). Indeed (8) shows that, to evaluate the dual function at some $\bar{\lambda}$, I subproblems of the form (7) have to be solved. Moreover, these subproblems give for free the following subgradients

$$\begin{array}{rcll} s^i(\bar{\lambda}) & := & \bar{p}^i \in \partial\theta^i(\bar{\lambda}) \text{ for } i \leq I & \text{(a)} \\ s(\bar{\lambda}) & := & \sum_{i \leq I} \bar{p}^i + d \in \partial\theta(\bar{\lambda}) & \text{(b)}. \end{array} \tag{9}$$

Although nonconvexities and integer values in (3) result in a strictly positive duality gap (corresponding to a strict inequality in (5)), a dual solution of (6) can still be used for approximating primal solutions. Besides, dual solutions are useful to estimate marginal prices. Furthermore, solving the master problem by a proximal bundle method allows to find approximate solutions to a relaxed convexified version of the primal problem, see [18].

The next subsection is devoted to the subproblem resolution. In particular, we explain how the separability property of (8) allows us to exploit individual structures and obtain efficient methods of resolution for each type of subproblem.

3.2. Subproblems

Depending on the nature of the unit considered, subproblems in (7) are solved with either interior point methods or dynamic programming.

When constraints are linear and the cost function is convex and C^2, a primal-dual interior point method can be used, [12]. This is a convenient approach to solve hydraulic subproblems, as described in § 2.2. Note also that for hydro-storage problems, an efficient descent algorithm was developed in [19].

Nevertheless, when the modeling yields non-convex constraints and/or cost functions, a (stochastic) dynamic programming approach is preferred, see [5]. Such is the case of thermal subproblems (integer variables) or long term nuclear/hydraulic subproblems (nonconvex constraints).

Both methods (interior points and dynamic programming) have a complexity that depends linearly on the dimension N of the dual space. A known drawback of dynamic programming is that its complexity grows exponentially with the state variable dimension. Hence, it is not a technique to be used for hydraulic subproblems. We refer to [12] and [6] for more details.

4. Solving the master program

Because the dual function is nondifferentiable, to solve (6) we use a variant of the reversal quasi-Newton bundle method in [17]. Our variant exploits the separable structure of the dual function (8) by using a *disaggregated* cutting-planes model. It is related to the multiple cuts technique used by [11]. Specifically in the context of Stochastic Programming, disaggregated cuts were earlier used in [2] and [21].

In [10] the same variant is applied to the resolution of deterministic unit-commitment problems. An outline of the convergence proof of the method can be found in [10, § 6].

4.1. Disaggregated bundle methods

We briefly describe the basics of a bundling methodology. In order to stabilize cutting-planes oscillations, bundle methods generate sampling points λ^j by solving a problem as in (10) below, depending on a given stability center $\hat{\lambda}^k$. Stability centers are selected sampling points, providing a "good enough" improvement in the optimization process. They form a subsequence $\{\hat{\lambda}^k\} \subset \{\lambda^j\}$ that is proved to converge to a solution, see for instance [13, Vol. II].

Specifically, suppose $j-1$ sampling points have been generated and let $\hat{\lambda}^k$ be the current stability center. A certain bundle $\mathcal{B}_j$, collecting the information generated so far, is used to define a cutting-planes model $\hat{\theta}$. Then λ^j is the solution of

$$\max_{\lambda \in \mathbb{R}^N} \hat{\theta}(\lambda) - \frac{1}{2}\mu_k \|\lambda - \hat{\lambda}^k\|^2 , \tag{10}$$

where μ_k is a positive parameter. Let $m_1 \in]0,1[$ be an Armijo-like parameter. If $\theta(\lambda^j) \geq \theta(\hat{\lambda}^k) + m_1 \left(\hat{\theta}(\lambda^j) - \frac{1}{2}\mu_k\|\lambda^j - \hat{\lambda}^k\|^2 - \theta(\hat{\lambda}^k)\right)$, then λ^j provides a sufficient improvement. The sampling point becomes a stability center, μ_k is updated and both j and k are increased.

Otherwise, the model $\hat{\theta}$ is considered not accurate enough and a *null step* is declared. There is no new k-iterate and only j is increased.

Observe that in both cases index j is increased. Accordingly, an enriched bundle generates a new model $\hat{\theta}$. Typically, a bundle is a set of the form

$$\mathcal{B}_{j+1} = \{ \ (\theta(\lambda^1), s(\lambda^1)) \ , \ \ldots, \ (\theta(\lambda^j), s(\lambda^j)) \ \} \ , \tag{11}$$

where we use the notation $s(\lambda)$ for subgradients in $\partial\theta(\lambda)$.

A standard cutting-planes model associated to $\mathcal{B}_{j+1}$ has the general expression

$$\hat{\theta}(\lambda) = \min_{l \leq j} [\theta(\lambda^l) + \langle s(\lambda^l), \lambda - \lambda^l \rangle] \, .$$

For the sake of simplicity, we do not enter here into technicalities such as bundle compression and selection. These techniques are aimed at keeping bounded the size of the bundle by discarding some superfluous elements. Although convergence is not impaired, it has been observed that frequent compressions spoil the speed of convergence. We refer to [14], [15] for further details.

When specialized to the dual function (8), values in the bundle are obtained by solving subproblems (7). From (9) we see that two possibilities are available. Using (9)(b) we obtain the standard (bundle and) cutting-planes model

$$\hat{\theta}(\lambda) = \min_{l \leq j} \left[\textstyle\sum_{i \leq I} \theta^i(\lambda^l) + \left\langle \textstyle\sum_{i \leq I} s^i(\lambda^l), \lambda - \lambda^l \right\rangle \right] + \langle \lambda, d \rangle \, . \tag{12}$$

In addition, (9)(a) gives the disaggregated bundle

$$\mathcal{B}_{j+1} := \left\{ \begin{array}{ccc} (\theta^1(\lambda^1), s^1(\lambda^1)) \ , & \ldots, & (\theta^1(\lambda^j), s^1(\lambda^j)) \ , \\ & \vdots & \\ (\theta^i(\lambda^1), s^i(\lambda^1)) \ , & \ldots, & (\theta^i(\lambda^j), s^i(\lambda^j)) \ , \\ & \vdots & \\ (\theta^I(\lambda^1), s^I(\lambda^1)) \ , & \ldots, & (\theta^I(\lambda^j), s^I(\lambda^j)) \end{array} \right\} , \tag{13}$$

giving birth to the disaggregated cutting-planes model

$$\hat{\theta}(\lambda) = \sum_{i \leq I} \min_{l \leq j} [\theta^i(\lambda^l) + \langle s^i(\lambda^l), \lambda - \lambda^l \rangle] + \langle \lambda, d \rangle \, . \tag{14}$$

When compared to (12), we see that (14) is a tighter approximation of the dual function. As such, (14) can be expected to be a better model, providing better sampling points and converging faster to a dual solution.

However, there is price to pay for this improvement. First, a disaggregated bundle as in (13) uses I times more storage than the standard one obtained from (11)-(9)(b). The management of a bigger bundle may oblige to frequent compressions resulting in a slower speed of convergence. In this respect, the use of sparsity structures in the subgradients is crucial to reduce storage requirements.

Furthermore, the size of stabilized problems (10) also increases. Actually, writing (10) as a nonlinear program (with extra variables for each minimum value in the cutting-planes model) we obtain the following sizes:

variables	constraints	model used in (10)
$1+N$	j	(12)
$I+N$	Ij	(14) .

Accordingly, to compute a sampling point in the second case a bigger program has to be solved. A second drawback of the disaggregated approach is therefore to increase computational times required per iteration.

We see that, for the disaggregated approach to make any sense, the extra computational burden must be compensated by a high reduction in the number of j-iterations. This is the purpose of the preconditioners we describe next.

4.2. The role of preconditioners

In this section, we address the question of finding a good preconditioner for the dual problem. More precisely, we look for a matrix D such that the change of variables

$$\lambda = D\ell \tag{15}$$

in (6) improves the efficiency of our bundle methodology. We restrict ourselves to diagonal scalings, i.e., the only non zero elements of D are its diagonal entries D_{nn}, $n = 1, \ldots, N$.

When minimizing a C^2-function, it is well know that the speed of convergence of a gradient method depends on the conditioning of the Hessian at a solution, say H^*. In this case, an efficient diagonal preconditioning is

$$D_{nn} = \sqrt{(\mathrm{inv} H^*)_{nn}}\,.$$

However, the dual function is often not differentiable at a solution. So we have to find some second order approximation of θ in a neighborhood of a solution. We first make some simplifying assumptions in order to obtain an analytic formulation for θ. We drop dynamic constraints and we take linear generation costs so that we have for each unit i and each node n :

$$\begin{cases} \mathcal{C}_n^i(p_n^i) = c^i p_n^i \\ \mathcal{D}^i = \mathbb{R}^n \\ \mathcal{S}_2^i = [0, \bar{p}^i]^N \end{cases} \tag{16}$$

where $(c^i)_{i=1,\cdots,I}$ are proportional costs which are supposed to be ordered

$$c^1 < c^2 < \cdots < c^I\,.$$

With these assumptions, the dual function from (6) turns out to be separable by nodes,

$$\tilde{\theta}(\ell) = \sum_{n=1,\cdots,N} \tilde{\theta}_n(\ell_n) \tag{17}$$

where

$$\tilde{\theta}_n(\ell_n) := \left(\sum_{i=1,\cdots,I} \min_{p_n^i \in \mathcal{S}_2^i} (\pi_n c^i - D_{nn}\ell_n) p_n^i \right) + D_{nn}\ell_n d_n \,.$$

Moreover, the inner minimization problems have trivial solutions depending on the sign of $\pi_n c^i - D_{nn}\ell_n$. the solution sets are given by

$$\begin{cases} \{0\} & \text{for all } i \in I_- := \{i \leq I : \ \pi_n c^i \ > \ D_{nn}\ell_n\} \\ [0, \bar{p}^{i_=}] & \text{for} \quad i_= \text{ such that} \qquad \pi_n c^{i_=} \ = \ D_{nn}\ell_n \\ \{\bar{p}^i\} & \text{for all } i \in I_+ := \{i \leq I : \ \pi_n c^i \ < \ D_{nn}\ell_n\}\,, \end{cases}$$

where some of the sets I_-,$\{i_=\}$ or I_+ can be empty and form a partition of $\{1, \cdots, I\}$. Thus, it is straightforward to compute for each node n

$$\tilde{\theta}_n(\ell_n) = \sum_{i \in I_+} (\pi_n c^i - D_{nn}\ell_n)\bar{p}^i + D_{nn}\ell_n d_n$$

and

$$\partial\tilde{\theta}_n(\ell_n) = \left[0, D_{nn}\bar{p}^{i_=}\right] + \left\{ D_{nn}(\sum_{i \in I_+} \bar{p}^i - d_n) \right\} \,. \tag{18}$$

Remember that $\partial\tilde{\theta}$ stands for the subdifferential of $-\tilde{\theta}$ which is convex.

A graphic representation of $-\tilde{\theta}_n$ can be found in Figure 3.

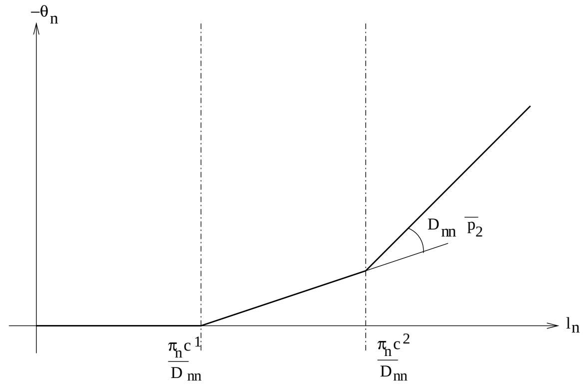

Figure 3. A graphic representation of $-\tilde{\theta}_n$.

Remark 1 Observe that without any preconditioning ($D = \mathcal{I}$, the identity matrix), if (6) has a solution, it lies in the set

$$\pi_1[0, c^I] \times \cdots \times \pi_N[0, c^I] \,.$$

So the set of solutions is contained in a cube dilated by probability factors along each direction of $\mathbb{R}^N$. This remark could be used to set $D_n = \pi_n$. In our setting, however, this choice would dilate the sets of subgradients, see figure (3). $\square$

Now, we find a second-order approximation of the Hessian of $\tilde{\theta}$ with the help of the Moreau-Yosida regularization. Proximal bundle methods can be viewed as an *implementable* proximal point algorithm, [13, Ch. XV]. Indeed, the proximal point algorithm finds the minimum of $-\tilde{\theta}$ which is convex but non-differentiable, by minimizing its Moreau-Yosida regularization,

$$\Theta(\hat{\lambda}) := \min_{l \in \mathbb{R}^n} -\theta(l) + \frac{1}{2}\mu \|l - \hat{\lambda}\|^2 , \tag{19}$$

where μ is a positive parameter. Minimizing $-\tilde{\theta}$ is equivalent to minimize Θ which is differentiable and has a Lipschitz continuous gradient. Denoting by $p(\hat{\lambda})$ the unique minimizer in (19), which is called the proximal point of $-\tilde{\theta}$ at $\hat{\lambda}$, we can write

$$\nabla\Theta(\hat{\lambda}) = \mu(\hat{\lambda} - p(\hat{\lambda})) \in \partial\theta(p(\hat{\lambda})) . \tag{20}$$

Replacing $\hat{\lambda}, p(\hat{\lambda}), \mu$ by $\hat{\lambda}^k, \hat{\lambda}^{k+1}, \mu_k$ gives an iterative procedure for minimizing $-\tilde{\theta}$. This procedure is precisely a gradient method, preconditioned by $1/\mu_k$. However it is not implementable since it requires the computation of Θ. When compared to (10), it follows that proximal bundle methods minimize the Moreau-Yosida regularization of $-\hat{\theta}$, a cutting-planes approximation of $-\theta$. This interpretation allows us to relate the rate of convergence of the subsequence $\{\hat{\lambda}^k\}$ with the conditioning of $\nabla^2\Theta$, whenever this Hessian exists.

The Moreau-Yosida regularization of $-\tilde{\theta}$ is also separable by nodes:

$$\Theta(\hat{\lambda}) = \sum_n \Theta_n(\hat{\lambda}_n) ,$$

where

$$\Theta_n(\hat{\lambda}_n) = \min_{l \in \mathbb{R}} -\tilde{\theta}_n(l) + \frac{1}{2}\mu|\hat{\lambda}_n - l|^2 .$$

In this trivial case, it results that when the Hessian of Θ exists, it is a diagonal matrix whose components are given by the second derivatives of Θ_n, Θ_n''.

Given a particular function Θ_n write (20) with $\Theta, \theta, \hat{\lambda}$ replaced by $\Theta_n, \theta_n, \lambda_i$, for some arbitrary $\lambda_i \in \mathbb{R}$. It follows that

$$\text{for any } \gamma_i \in \partial\tilde{\theta}(\lambda_i) \text{ the point } z_i := \lambda_i + \gamma_i/\mu \text{ satisfies } \begin{array}{l} \lambda_i = p(z_i) , \\ \Theta_n'(z_i) = \gamma_i . \end{array}$$

For $i = 1, 2$, choose

$$\lambda_i := \frac{\pi_n c^i}{D_{nn}} \text{ so that } \partial\tilde{\theta}(\lambda_i) = \begin{cases} [0, D_{nn}\bar{p}^1] - D_{nn}d_n & i = 1 , \\ D_{nn}(\bar{p}^1 - d_n) + [0, D_{nn}\bar{p}^2] & i = 2 . \end{cases}$$

Since the factor D_{nn} appears in every subgradient, let $G>0$ be such that $D_{nn}G$ defines any difference of subgradients $\gamma_1-\gamma_2$. Note that G does not depend on n. By the Mean Value Theorem, there exists $\hat{z}\in(z_1,z_2)$ such that

$$\begin{aligned}\Theta_n''(\hat{z}) &= \frac{\Theta_n'(z_2)-\Theta_n'(z_1)}{z_2-z_1}\\ &= \frac{D_{nn}G}{\frac{\pi_n}{D_{nn}}(c^2-c^1)+D_{nn}G/\mu}\\ &= \mu\frac{1}{1+\frac{\pi_n}{D_{nn}^2}\left(\frac{\mu(c^2-c^1)}{G}\right)}\,.\end{aligned}$$

The last expression discloses the *nature of the beast*: a suitable choice for D_{nn} to stabilize the Hessian needs to be proportional to the **square root** of the corresponding probability. Accordingly, we just take

$$D_{nn} := \sqrt{\pi_n}\,, \tag{21}$$

for our numerical experience.

If generation costs are assumed to be quadratic in (16), it is still possible to derive an analytical expression for the Hessian of Θ that yields the same preconditioning (21). Finally, note that when N from (1) is big the ratio $\frac{\pi_1}{\pi_N}$ may be of the order of 10^4. Our choice of preconditioner eliminates the corresponding disparities by making the updates independent of each particular component n.

Remark 2 As observed by one of the referees, the way of dualization is an important aspect for the conditioning of the dual. Specifically, the choice of preconditioners (as well as their relevance) depends strongly on the Lagrangian used to produce the dual problem (6). It is known that different scalar products pairing constraints with multipliers lead to different forms of dual problem and Lagrangian. Consider for example the stochastic Lagrangian ([8], [19], [20]) obtained when in (4) the Euclidean scalar product $\langle\lambda,g\rangle=\sum_n\lambda_n g_n$ is replaced by $\langle\lambda,g\rangle_\pi:=\sum_n\pi_n\lambda_n g_n$. In this context, (sub)gradients are taken relative to the new scalar product:

$$s=(s_1,\cdots,s_N)\in\partial\theta(\lambda)\iff\pi^{-1}s=(\pi_1^{-1}s_1,\cdots,\pi_N^{-1}s_N)\in\partial_\pi\theta(\lambda)\,;$$

and, similarly, the norm in (10) and (19) is rescaled by π. Accordingly, in (17) we have

$$\begin{aligned}\tilde{\theta}_n(\ell_n) &:= \pi_n\left(\sum_{i=1,\cdots,I}\min_{p_n^i\in\mathcal{S}_2^i}(c^i-D_{nn}\ell_n)p_n^i\right)+\pi_n D_{nn}\ell_n d_n\\ &= \pi_n\textstyle\sum_{i\in I_+}(c^i-D_{nn}\ell_n)\bar{p}^i+\pi_n D_{nn}\ell_n d_n\,.\end{aligned}$$

Here, I_+ is redefined as the set $\{i\le I: c^i<D_{nn}\ell_n\}$ so that $\partial_{\pi_n}\tilde{\theta}_n(\ell_n)$ coincides with the expression in (18). Altogether, using a development analogous to the one leading to (21), it can be seen that the second order derivative $\Theta_n^{''\pi}(\hat{z})$ associated to the new scalar product is **independent** of π_n. In this case, $D=\mathcal{I}$ (i.e. no preconditioning at all) appears as the best choice. $\square$

5. Computational results

EDF manages a generation system with 57 nuclear power plants, 15 hydro-valleys and about 60 thermal units with total capacity of about 102500 MW (1997). Such a big amount, much larger than winter peak loads of 70000 MW, is required to allow nuclear plants maintenance and to ensure power reserves at any time. In 1997, the distribution of power generation among units was

nuclear plants	hydro-valleys	other sources (thermal units)
80 %	15 %	5 %.

We first report on some tests that show the importance of preconditioning. Then we discuss advantages and drawbacks of disagreggation.

For our test problems, we use the mix

nuclear plants	hydro-valleys	thermal units	I
57	3	60	120 .

For all our runs we use warm starts to save factorizations in the resolution of subproblems (10). The size of the bundle was limited to 100 elements, when a compression was necessary, inactive elements at the current iterate were eliminated from the bundle. *Inactive* elements $(\theta(\lambda^i), s(\lambda^i))$ in $\mathcal{B}_j$ are those such that $\hat{\theta}(\lambda^j) > \theta(\lambda^i) + \langle s(\lambda^i), \lambda^j - \lambda^i\rangle$. Optimality is declared by using the stopping criterion in [16].

5.1. *Effect of preconditioning*

In section 4.2 we addressed the question of finding an efficient preconditioner for the dual problem. By developing an analytical approach for a simplified model we obtained the diagonal scaling from (21). The beneficial effect of our preconditioning was confirmed in a battery of 4 test-problems, with trees of size

$$N = 129, 305, 881, 2161\,,$$

respectively.

To assess our approach, we tested the following preconditioners with an aggregated bundle method.

- D^*, given by (21),
- $D_1 = {D^*}^2$, a diagonal scaling with probability factors along each direction (cf. Remark 1) and
- $D_2 = \mathcal{I}$, or no preconditioning at all.

Results are reported on Table 1. For each test we give the total number of iterations needed to reach an optimal value within a tolerance of 10^{-5}. The sign !! was used when more than 1000 iterations were required.

Table 1. Importance of preconditioners.

N	$D^* = (\sqrt{\pi_n})_{nn}$	$D_1 = (\pi_n)_{nn}$	$D_2 = \mathcal{I}$
129	143	160	164
305	139	512	410
881	146	251	!!
2161	176	!!	!!

Inspection of the figures in Table 1 indicates D^* as the best preconditioner, independently of the tree considered. More importantly, when using D^*, the number of iterations needed for convergence is not affected by the size of the tree (i.e., by the size of the dual problem (6)). By contrast, with both D_1 and D_2 the method becomes intolerably slow when N grows.

5.2. *Disagreggation vs. Agreggation*

We already mentioned that a disaggregated cutting-planes model gives a tighter model and thus provides a faster convergence rate. However, we also said that disaggregation results in heavier iterations for the master program. Unlike standard bundle methods, a particular feature of the disaggregated approach is that the total time spent in the master resolution is no more negligible. In fact, disaggregation is an efficient variant only when (in terms of CPU times) the resolution of subproblems (7) is far more demanding that one resolution of (10)-(14). Along these lines, the results reported in [10, §7.2] confirm our remark. Under these circumstances, it is essential for the bundle method to converge in relatively few iterations. For our tests, we consider two different time horizons

days	time steps	stepsize	T	N
10	80	3 h	240	504
30	240	3 h	720	1514 .

Because preliminary tests using a totally disaggregated model resulted in bad performances, we rather chose a strategy of *partial* disaggregation. More precisely, given an arbitrary partition of I, say $\{I_1, \cdots, I_K\}$, we replace θ^i and s^i in(13) respectively by θ^k and s^k where

$$\begin{array}{rcl} \theta^k & := & \sum_{i\in I_k} \theta^i \\ s^k(\bar\lambda) & := & \sum_{i\in I_k} \bar{p}^i \quad \in \partial\theta^k(\bar\lambda) \text{ for } i \leq I\,. \end{array}$$

To choose a partition of I among the $I!$ possibilities, we adopted the following strategies:

- First, we define partitions of I with subsets of the same size in order to find the best number of subsets, K. Increasing K yields a better approximation of the dual function but slows down the master program because the maximal size of each sub-bundle was kept equal to 100 as in the agreggated case. A good compromise was to take a value of K between 5 and 10.

- Having set a suitable value for K, we look for a good partition of K elements. A first possibility, that we call Disaggregate 1, is to have K subsets of the same size. Our second strategy, Disaggregate 2, takes into account the nature of the subproblems. We considered each one of the 3 hydro-valleys individually and assigned one subset for each one. Thermal units were ordered by "convexified costs", stemming from convexified thermal subproblems. The resulting set of nuclear and thermal units was then divided into $K - 3$ subsets of roughly the same size.

Table 2 shows the results obtained with an aggregate version of the bundle method and the two types of partial disaggregation. For the first one, we take $K = 6$ subsets of the same size. For the second one we adopt Disaggregate 2 with $K = 8$. We used for all these tests the preconditioner introduced in 4.2. A failure in convergence after 1000 iterations is again denoted by !!.

Table 2. Aggregate vs Disaggregate.

Test	Tol.	Aggregate It.	Aggregate Time(sec)	Disaggregate 1 It.	Disaggregate 1 Time(sec)	Disaggregate 2 It.	Disaggregate 2 Time(sec)
10 days	10^{-3}	68	920	39	403	36	407
	10^{-4}	112	1450	62	655	59	710
	10^{-5}	!!	!!	170	2370	104	1425
30 days	10^{-3}	75	1780	48	1260	31	840
	10^{-4}	144	3480	122	4192	91	3880
	10^{-5}	!!	!!	!!	!!	174	8430

For the 10 days case, both kinds of disaggregation are more efficient than the aggregated method. Actually, they provide an additional digit of accuracy on the optimal value. On the case with 30 days, however, disaggregation seems to be less interesting. Concerning the different disaggregations we observe that, contrary to our expectations, Disaggregate 2 just provides a slight improvement of the convergence speed. Along these lines, it would be interesting to investigate the effect of other disaggregation techniques, possibly depending on the number of hydro-valleys.

6. Conclusion

In this article we presented a methodology to solve stochastic optimal power management problems. We exploited the separable structure of the dual problem arising from Lagrangian relaxation by using a disaggregated bundle method. We also introduced a preconditioner adapted to space-decomposition of stochastic problems.

Numerical tests confirmed the efficiency of the preconditoner and showed that in some cases, disaggregation techniques can improve the convergence speed. Finally, we used in our test an heuristic of partial disaggregation to improve the numerical efficiency of the method. A potentially improved technique could be devised by assigning specific sizes to each sub-bundle, depending on the nature of the subproblem considered.

References

1. J. Batut and A. Renaud. Daily generation scheduling with transmission constraints: A new class of algorithms. *IEEE Transactions on Power Systems*, 7:982–989, 1992.
2. J.R Birge and F.V Louveaux. A multicut algorithm for two stage stochastic linear programs. *European Journal of Operational Research*, 34:384–392, 1988.
3. J.R. Birge, S. Takriti, and E.Long. Intelligent unified control of unit commitment and generation allocation. EPRI RP8030-13, Department of Industrial and Operations Engineering.
4. J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Optimisation Numérique: aspects théoriques et pratiques*. Springer Verlag, 1997.
5. S. Brignol and A. Renaud. A new model for stochastic optimization of weekly generation schedules. In *APSCOM-97 Proc, Hong-Kong*, pages 656–661, 1997.
6. S. Brignol and G. Ripault. Risk management applied to weekly generation scheduling. In *IEEE Winter Meeting Proceedings*, pages 465–470, 1999.
7. A. J. Conejo and N. Jiménez Redondo. Short-term hydro-thermal coordination by Lagrangian relaxation: solution of the dual problem. *IEEE Transactions on Power Systems*, 1998.
8. D. Dentcheva and W. Römisch. Solving the unit commitment problem in power generation by primal and dual methods, 1996. Humboldt-Universität zu Berlin, Institut für Mathematik.
9. D. Dentcheva and W. Römisch. Optimal power generation under uncertainty via stochastic programming. *Economics and Mathematical Systems*, 458:22–56, 1998.
10. S. Feltenmark and K. C. Kiwiel. Dual applications of proximal bundle methods, including Lagrangian relaxation of nonconvex problems. *siamo*, 1999.
11. J. L. Goffin and J. P. Vial. Multiple cuts in the analytic center cutting plane method, 1998. Logilab, HEC Working Paper 98.10, Departement of Management Studies, University of Geneva.
12. J. P. Goux, A. Renaud, S. Brignol, and J. C. Culioli. Stochastic optimization of weekly generation schedules: solution of the hydraulic subproblems with interior point methods. In N.Flatabo E. Broch, D.K. Lysne and E. Helland-Hansen, editors, *Hydropower'97*, pages 227–2342. Balkema, 1997.
13. J. B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms*. Springer Verlag, 1991.
14. K.C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.
15. K.C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122, 1990.
16. C. Lemaréchal, F. Pellegrino, A. Renaud, and C. Sagastizábal. Bundle methods applied to the unit-commitment problem. In J. Doležal and J. Fidler, editors, *System Modelling and Optimization*, pages 395–402. Chapman and Hall, 1996.
17. C. Lemaréchal and C. Sagastizábal. Variable metric bundle methods: from conceptual to implementable forms. *Mathematical Programming*, 76:393–410, 1997.
18. D. Medhi. *Decomposition of Structured Large-scale Optimization Problems and Parallel Optimization*. PhD thesis, Computer Sciences Department, University of Wisconsin–Madison, 1987.
19. M. P. Nowak and W. Römisch. Stochastic Lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty, 1998. DFG-Schwerpunktprogramm Echtzeit-Optimierung grosser Systeme, Preprint 98-36, submitted to Annals of Operations Research.

20. R. T. Rockafellar and R. J-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16, 1991.
21. A. Ruszczynski. A regularised decomposition method for minimizing a sum of polyhedral functions. *Mathematical Programming*, 35:309–333, 1986.
22. A. Ruszczyński. Decomposition methods in stochastic programming. In *Mathematical Programming*, volume 79, 1997.
23. S. Takriti, J.R. Birge, and E.Long. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, 11:1497–1508, 1996.