47

Bundle methods applied to the unit-commitment problem

C. Lemaréchal, C. Sagastizábal
INRIA, BP 105, 78153 Le Chesnay, France.
F. Pellegrino, A. Renaud
EdF, Dépt. Méthodes d'Optimisation et de Simulation, 92141 Clamart, France.

Abstract

On a daily-time scale, the power generation management poses a large-size and composite problem. For this reason, price decomposition has become one of the most commonly used strategies in this field. Classical spatial decomposition, obtained by dualizing the coupling constraints, is not the only way of decomposition. Some recent tests have shown that a space/time decomposition is particularly well-suited to deal with transmission constraints. The resulting dual problem is large but tractable (more than ten thousand variables), while the spatial decomposition would yield exceedingly many dual variables. This paper is devoted to the application of bundle methods to such large-scale nondifferentiable problems. To assess the approach we present some numerical tests which show that, in practice, bundle methods can cope with such large problems.

1 THE POWER UNIT-COMMITMENT PROBLEM

1.1 General formulation

The optimal scheduling of many power units is a challenging problem, from both economical and mathematical points of view. In particular, to solve a short-term unit-commitment problem means to consider simultaneously:

- thermal power-plants with high start-up costs and discontinuous operation domain, introducing non-convex constraints;
- hydro-valleys of interconnected reservoirs, leading to tight dynamic constraints;
- various types of integral and logical constraints.

More precisely, we study a mix I of power-generation units (hydraulic valleys, nuclear plants and classical thermal units) for a discrete period of time $\{1,2,\ldots,T\}$ (typically, 48 half-hour steps). Call p_i^t the production level of unit $i \in I$ at time $t \leq T$. Througout this paper, we will denote by $p^t := (p_1^t, p_2^t, \ldots, p_{|I|}^t)$ the production vector of all the units at time t; likewise, $p_i := (p_i^1, p_i^2, \ldots, p_i^T)$ will be the production vector of unit i for the time period.

Altogether, our problem is

$$\begin{cases} \min \sum_{i \in I} c_i(p_i) \\ p_i \in \mathcal{D}_i\,, \quad \text{for all } i \in I\ , \\ p^t \in \mathcal{S}^t\,, \quad \text{for } t = 1,2,\ldots,T\ , \end{cases} \tag{1}$$

J. Doležal et al. (eds.), *System Modelling and Optimization*

where $\mathcal{D}_i$ describes the operating dynamic constraints of unit i and $\mathcal{S}^t$ stands for the static constraints such as satisfaction of demand and network transmission. As for the objective function, each c_i represents the whole production cost of each unit i over the time period.

1.2 The model

Dynamic constraints

Problem (1) can be considered as a general modelling of the unit-commitment problem. Yet, the operating constraints are not exactly the same from one utility to the other. Without entering into details, a general description of the $\mathcal{D}_i$'s for the French generation mix includes the following features:

Nuclear power plants: A minimum delay of two hours is necessary between two output variations; some output ranges are forbidden; the number of output variations is limited.

Classical thermal units: If the plant is online, then $p_i^t \in [\underline{p_i}, \overline{p_i}]$, with $\underline{p_i} > 0$; a plant must not have more than one start-up per day; during shutdown or start-up phases, fixed output curves must be followed.

Needless to say, the description of the operating constraints for these thermal plants involves 0-1 variables, to describe starts-up, number of output variations, etc.

Hydro-valleys: A valley consists of a set of interconnected reservoirs. During low-demand periods, water in a downstream reservoir can be pumped upstream, so that it can be discharged later, when generation costs become higher.

More formally, let $\mathcal{H}_v$ denote the set of water reservoirs of valley v, and let $\Gamma(h \leftarrow)$ be the set of plants upstream a particular reserve $h \in \mathcal{H}_v$.Then the following must hold for all $h \in \mathcal{H}_v$ and $t = 1, 2, \ldots, T$:

$$V_h^{t+1} - V_h^t = a_h^t - T_h^t + Q_h^t - D_h^t + \sum_{g \in \Gamma(h \leftarrow)} \left(T_g^{t_{g \to h}} - Q_g^{t_{g \to h}} + D_g^{t_{g \to h}} \right) ,$$

where

- $V_h^t \in [\underline{V_h}, \overline{V_h}]$ and a_h^t are respectively the content and inflow of reservoir h at time t;
- $T_h^t \in [\underline{T_h^t}, \overline{T_h^t}]$ is the discharge, $D_h^t \in [\underline{D_h^t}, \overline{D_h^t}]$ the spillage and $Q_h^t \in [\underline{Q_h^t}, \overline{Q_h^t}]$ the quantity of water pumped by plant h at time t;
- $t_{g \to h} := t - d(g \to h)$, with $d(g \to h)$ denoting the discharge delay from plant g to reserve h.

Static constraints

Concerning the description of $\mathcal{S}^t$, we distinguish two types of constraints.

Demand: For each time-step $t = 1, 2, \ldots, T$, the demand D^t is known and has to be satisfied: $\sum_{i \in I} p_i^t = D^t$.

"Spinning reserve" constraints could also be introduced, guaranteeing that there is enough power started up to quickly face most random events (unit outage, error in demand forecast, ...). These constraints are of the same type as above; for simplicity, we do not consider them here.

Transmission: The capacity of each arc in the network is limited. "Security constraints" can also be included, to ensure the satisfaction of the demand even after the outage of

one line. In the flow or the DC approximation model, these constraints are linear; but, as opposed to demand or reserve, they are numerous: the network has several hundreds of nodes. With security constraints, there will be several hundred thousand linear constraints at each time step.

2 DECOMPOSITION SCHEMES

Abstractly, our problem can be written

$$\min c(p)\,, \quad p \in \mathcal{D} \cap \mathcal{S}\,, \tag{2}$$

where $p = \{p_i^t\} \in R^I \times R^T$ represents the vector of productions, and the cost function c is a sum over $i \in I$ of "local costs" $c_i(p_i)$. As for the feasible sets $\mathcal{D}$ and $\mathcal{S}$, they both have the product form $\mathcal{D} = \prod_{i\in I} \mathcal{D}_i$ and $\mathcal{S} = \prod_{t=1}^{T} \mathcal{S}^t$; but $\mathcal{D}$ is nonconvex and has a rather complicated description, while $\mathcal{S}$ is described by linear constraints, say

$$\text{for each } t \in T, \quad p^t \in \mathcal{S}^t \Leftrightarrow A^t p^t = b^t\,. \tag{3}$$

Here, each matrix A^t is either a single row $(1, \cdots, 1)$ or an awfully large full matrix, depending on whether or not transmission constraints are considered.

Clearly, (2) is fairly difficult: it is a large-scale, nonlinear, mixed integer programming problem. However, it has a highly decomposable structure which can be exploited through *Lagrangian relaxation*. We study here two decomposition schemes, which lend themselves to a handy (approximate) numerical treatment of (2). For more details on dualization techniques see Chap. XII in (Hiriart-Urruty, Lemaréchal, 1993); see also (Grinold, 1970).

2.1 Spatial decomposition

Exploiting the linearity (hence decomposability) of the constraints (3), it is natural to write (2) as:

$$\min \sum_{i\in I} c_i(p_i) \quad \text{subject to } \sum_{i\in I} A_i p_i = b \text{ and } p_i \in \mathcal{D}_i \quad \text{for } i \in I\,, \tag{4}$$

where b denotes the compound vector $(b^1, \cdots, b^T)$ and similarly for the compound matrices A_i.

The writing in (4) exhibits the role of the production units p_i as "local agents", coupled by the constraints (3). For the present unit-commitment problem, the local agents are the power-plants, coupled by the "spatial" static constraints.

To achieve decomposition, associate a multiplier μ^t to the coupling constraints in (4) and form the Lagrangian

$$L_s(p, \mu) := \sum_{i\in I} [c_i(p_i) - \mu^\top A_i p_i] + \mu^\top b\,.$$

The associated dual function is then defined by

$$\Sigma(\mu) := \min_{p\in\mathcal{D}} L_s(p,\mu) = \mu^\top b + \min_{p\in\mathcal{D}} \sum_{i\in I}[c_i(p_i) - \mu^\top A_i p_i] \tag{5}$$

and can be computed in a decomposed way: each local agent $i \in I$ has to solve

$$\min\{c_i(p_i) - \mu^\top A_i p_i : p_i \in \mathcal{D}_i\}\,.$$

In other words, each power-plant has to optimize its own production, for a given "shadow price" μ. The whole issue is then to find appropriate values for μ; by duality theory, what we have to do is to maximize the function $\Sigma(\mu)$. This is the so-called *dual problem* associated with the formulation (4).

The above decomposition scheme has been used for rather long for the unit-commitment problem (Fisher, 1973), (Merlin, Sandrin, 1983).

2.2 Space-time decomposition

Our second scheme starts from a naive stratagem: artificially duplicating the variable p, we can write (2) in the "cross-form"*

$$\min c(p) \quad \text{subject to } p \in \mathcal{D},\ q \in \mathcal{S},\quad p = q\,. \tag{6}$$

An obvious decomposition has thus appeared, which is actually enhanced by the particular form of $c, \mathcal{D}$ and $\mathcal{S}$. In fact, the variables $\{p_i^t\}$ and $\{q_i^t\}$ are "local agents", coupled by the constraints $p_i^t = q_i^t$.

To achieve decomposition, associate a multiplier λ_i^t to each coupling constraint and form the Lagrangian

$$L_\psi(p,q,\lambda) := c(p) - \lambda p + \lambda q\,.$$

The dual function is now defined by

$$\Psi(\lambda) := \min_{p\in\mathcal{D},q\in\mathcal{S}} L_\psi(p,q,\lambda) = \min_{p\in\mathcal{D}} \sum_i [c_i(p_i) - \lambda_i p_i] + \min_{q\in\mathcal{S}} \sum_t \lambda^t q^t\,. \tag{7}$$

Here again, it can be computed in a decomposed way: the local agents have to solve respectively

$$\min\{c_i(p_i) - \lambda_i p_i : p_i \in \mathcal{D}_i\} \quad \text{and} \quad \min\{\lambda^t q^t : q^t \in \mathcal{S}^t\}\,. \tag{8}$$

The above p-problems are essentially the same as in spatial decomposition. As for the q-problems, they are just linear programs. When only demand-constraints are present, the feasible domain is a mere simplex. As for transmission constraints, they result in classical static network problems. For this scheme also, duality theory tells us that appropriate values of λ have to maximize $\Psi(\lambda)$.

*the objective function could also be $c(q)$.

For nonlinear unit-commitment problems, this decomposition scheme has been proposed in (Batut, Renaud, 1992). It somehow generalizes an earlier approach for mixed integer linear programming, given in (Guignard,Kim, 1987).

3 THE DUAL PROBLEMS

In the previous section, we have defined two approaches to solve (1)=(2), namely the two dual problems $\max_\mu \Sigma(\mu)$ of (5), and $\max_\lambda \Psi(\lambda)$ of (7). We now study these two new problems.

3.1 Theoretical assessment

A first question is: are the dual problems any good in terms of their respective primals (4) or (6)? Call $\bar{c}$ the optimal objective value in the primal problem (1)=(2)=(4)=(6). It is known that both dual optimal objective values $\bar{\Sigma}$ and $\bar{\Psi}$ are lower than $\bar{c}$, and a crucial property is whether any of them is equal to $\bar{c}$. However, this has no reason to be true, because (1) is a nonconvex optimization problem. In this sense, the dual approaches proposed in § 2 can only yield approximate solutions. Clearly, for such approaches to be performant, it is crucial that the (positive) *duality gaps* $\bar{c} - \bar{\Sigma}$ and $\bar{c} - \bar{\Psi}$ are sufficiently small.

Here comes an important result that we state here without proof. Because the static constraints $\mathcal{S}$ are linear,

$$\textit{there always holds } \bar{c} \geq \bar{\Psi} = \bar{\Sigma}.$$

When (2) is an integer linear program, a proof can be found in (Guignard, Kim, 1987). For the general case it is a consequence of Proposition XII.5.1.3 of (Hiriart-Urruty, Lemaréchal, 1993). A detailed proof can be found in (Lemaréchal, Renaud, 1996), together with an interpretation of the optimal dual solutions $\bar{\mu}$ or $\bar{\lambda}$, as marginal prices associated with particular convexified forms of (1). As a result, if some nonconvex constraints are duplicated (taken into account in $\mathcal{S}$ *and* in $\mathcal{D}$), then the duality gap is reduced for the space-time decomposition.

Thus, for our problem, the space-time decomposition scheme can only give better dual solutions than the spatial one.

Another issue is the complexity of the resulting duals $\max \Sigma(\mu)$ and $\max \Psi(\lambda)$. We recall that a dual function such as Σ or Ψ, defined via a minimization problem such as (5) or (7), is concave; and that any optimal solution such as $\bar{p}$ in (5) or $(\bar{p}, \bar{q})$ in (7) gives a subgradient of the corresponding dual function:

$$\nabla_\mu L_s(\bar{p}, \mu) = b - \sum_{i \in I} A_i \bar{p}_i \in \partial\Sigma(\mu) \quad \text{and} \quad \nabla_\lambda L_\psi(\bar{p}, \bar{q}, \lambda) = \bar{q} - \bar{p} \in \partial\Psi(\lambda)\,. \tag{9}$$

Because, due to nonconvexity, such solutions $\bar{p}$ and $\bar{q}$ have no reason to be unique, $\nabla\Sigma$ and $\nabla\Psi$ have no reason to exist: the dual functions are nonsmooth. Now, the complexity of a nonsmooth optimization problem depends crucially on the number of variables.

In the case of space-time decomposition, the dual variables are indexed in $I \times \{1, \ldots, T\}$. For the French generation mix, we have $|I| \simeq 150$; with $T = 48$, this makes about 7000 dual variables λ_i^t.

For the spatial decomposition, according to our description of static constraints, there are two cases:

- If $\mathcal{D}$ involves only demand satisfaction, there is one linear constraint for each time-period t. The dual function Σ has then 48 variables μ^t, a reasonable number†
- If transmission constraints are taken into account, each righthand side b^t in (3) may have hundreds of thousands of coordinates; this results in a huge vector μ.

Thus, space-time decomposition becomes a must when transmission constraints are present; indeed, this has been the initial motivation of such an approach (Batut, Renaud, 1992).

3.2 Dual algorithms

Traditionally, subgradient optimization (Shor, 1985) has been used to maximize the non-smooth function Σ or Ψ; this is only conceivable for the simplest situation of spatial decomposition without transmission constraints. To handle the 10^4-10^5 variables appearing in the other cases, sophisticated methods must be applied: we choose bundle methods. They are a certain improvement of the classical cutting-plane algorithm, using ideas from the proximal regularization; for a general review, see Chapter XV in (Hiriart-Urruty, Lemaréchal,1993). We have conducted experiments with a particular implementation, namely Algorithm 19 of (Lemaréchal, Sagastizábal, 1993), using the objective regularization (g-f), diagonal quasi-Newton updates (dqN) and curved search. For the reader's convenience, we briefly sketch the algorithm here.

3.3 General Scheme

Consider the maximization of $\Psi(\lambda)$, with the help of local solvers able to compute $\Psi(\lambda)$ in (7), yielding $g(\lambda) \in \partial\Psi(\lambda)$ from (9). We choose a positive integer **MEMAX**. At the current iteration of the algorithm, we have on hand a "bundle" $\{\Psi(\lambda_k), g(\lambda_k)\}_{k\leq n}$ of $n \leq$**MEMAX** function- and subgradient-values. Therefore **MEMAX** is the maximum size of the bundle. We also have a current prox-center $\hat{\lambda}$. We define the polyhedral model $\hat{\Psi}_n(\lambda) := \min_{k\leq n}[\Psi(\lambda_k) + g(\lambda_k)(\lambda - \lambda_k)]$ and, for given prox-coefficient $t > 0$, we call λ^c the solution of the quadratic problem

$$\max_{\lambda}[\hat{\Psi}_n(\lambda) - \frac{1}{2t}|\lambda - \hat{\lambda}|^2]\,. \tag{10}$$

Step 1 *(t-adjustment)* Solve (10) and decide between the following actions:
- Declare "t too large": decrease t and solve (10) again.
- Declare "t too small": increase t and solve (10) again.
- Declare "descent-step": go to Step 2.
- Declare "null-step": go to Step 3.

Step 2 *(descent-step)* Update $\hat{\lambda} = \lambda^c$ and compute a new $t > 0$.

Step 3 *(bundle management)* If the size of the bundle has reached its maximum **MEMAX**, delete one element $\{\Psi(\lambda_k), g(\lambda_k)\}$. Append the new element $\{\Psi(\lambda^c), g(\lambda^c)\}$ to the bundle and loop to Step 1.

---

†if the "spinning" reserve constraints were taken into account, this number would be doubled.

We refer to (Lemaréchal, Sagastizábal, 1993) for details concerning the decision in Step 1 and the computation of t in Step 2. The quadratic code to solve (10) is due to K.C. Kiwiel, and is described in (Kiwiel, 1986).

3.4 Stopping test

The above description of the algorithm skips the question of the stopping criterion. This question is not straightforward in nonsmooth optimization; it becomes particularly crucial for the present problem, especially because of its large number of variables.

The resolution of (10) gives not only the candidate λ^c but also a certain $\varepsilon \geq 0$ and a certain "regularized subgradient" G, which is an ε-subgradient of Ψ at the prox-center $\hat{\lambda}$. Denoting by $\bar{\lambda}$ a maximizer of Ψ, we can write

$$\Psi(\bar{\lambda}) \leq \Psi(\hat{\lambda}) + \left\langle G, \bar{\lambda} - \hat{\lambda} \right\rangle + \varepsilon \leq \Psi(\hat{\lambda}) + \|G\| \, \|\bar{\lambda} - \hat{\lambda}\| + \varepsilon \,. \tag{11}$$

The rationale is then to request from the bundle algorithm a point $\hat{\lambda}$ which maximizes Ψ up to a tolerance $\underline{\varepsilon}$, chosen beforehand. Then we stop when both ε and $\|G\|$ are small, say $\varepsilon \leq \underline{\varepsilon}$ and $\|G\| \leq \mathtt{B}_G$. Ideally, (11) shows that we should have $\mathtt{B}_G = \underline{\varepsilon}/\|\hat{\lambda} - \bar{\lambda}\|$. Since this quantity is unknown, we need a value accounting for some "equanimity-in-the-adversity". For this, we assume that each coordinate $\|(\bar{\lambda} - \hat{\lambda})_t^t\|$ is smaller than some d; reasonable values of d are given by the nature of the problem. Then we obtain $\mathtt{B}_G = \underline{\varepsilon}/(\sqrt{n}d)$, where $n = |I|T$ is the number of (dual) variables.

4 NUMERICAL ILLUSTRATIONS

For numerical illustration, we have tested 6 datasets corresponding to 6 different days in the French mix. The set $\mathcal{S}$ is described by 48 demand-constraints in (3); the spatial dual problem has therefore 48 variables in all examples. Examples 1-4 have 207 plants in production, Examples 5 and 6 have 226 plants. For the space-time dual problems, this makes respectively $48 \times 207 = 9936$ and $48 \times 226 = 10848$ dual variables.

The table below compares the numbers of iterations with the two dualization schemes. The first column gives the number of resolutions of the decomposed problems (5) and (7) respectively. Between parentheses, we also give the number of descent-steps in the bundle algorithm of § 3.3. Then come the dual objective value after convergence, and the corresponding norm of regularized gradient. Note that $\mathtt{B}_G$ of § 3.4 corresponds to $d \simeq 0.1$FF/kW. In all our tests, the tolerance $\underline{\varepsilon}$ was set to 0.1%; this accuracy is good enough for a duality gap of 0.5%, typical for our unit-commitment problems. The aim of the column "partial" is to illustrate the difficulty to reach a reliable stopping criterion in nonsmooth optimization. For space-time decomposition (which has many variables), it gives the number of iterations and descent-steps necessary to obtain the required 3 digit accuracy. One sees how a bundle algorithm proceeds: first it improves objective values, and then it spends a rather long time reducing $\|G\|$ to appropriate values. Finally the last two columns give computing times: the proportion of CPU spent by the bundle algorithm itself (the rest being taken by Lagrangian minimizations) and the total CPU time on a SUN Sparc 20. Let us explain the drastic differences appearing in the last but one column. Most of the time used by a bundle algorithm is naturally due to the resolution of (10); but the complexity of the latter is far from what is commonly believed. Actually, one solves

Test	# iter	Dual value	$\|G\|$	Partial	Bundle	CPU(s)
1Σ	33(9)	5195575	150		0.02%	351
1Ψ	250(37)	5194556	100	85(22)	22%	1490
2Σ	19(5)	7842647	700		0.01%	327
2Ψ	98(16)	7841180	150	41(9)	11%	718
3Σ	25(70)	6499580	160		0.02%	237
3Ψ	280(33)	6499297	130	60(14)	27%	1480
4Σ	20(4)	8310985	690		0.02%	238
4Ψ	131(14)	8310235	160	37(8)	14%	590
5Σ	58(15)	4153291	110		0.04%	480
5Ψ	561(56)	4152531	70	209(49)	31%	3880
6Σ	34(10)	5679565	160		0.03%	326
6Ψ	216(28)	5678200	110	83(20)	22%	1480

the dual of (10), which is of the type $\min_\alpha \left[\sum_{k,\ell=1}^n \alpha_k\alpha_\ell \langle g(\lambda_k), g(\lambda_\ell)\rangle + \sum_{k=1}^n p_k\alpha_k\right]$. This requires first the computation of the scalar products $\langle g(\lambda_k), g(\lambda_\ell)\rangle$, which become fairly expensive when the dual space has 10^4 variables. Afterwards, the minimization problem is quickly solved: its number of variables does not exceed the number of iterations, which is itself reasonable; besides, it differs very little from the quadratic problem of the previous iteration.

REFERENCES

Batut, J., Renaud, A. (1992) Daily Generation Scheduling with Transmission Constraints: A New Class of Algorithms. *IEEE Transactions on Power Systems*, **7**, 982–9.

Fisher, M.L. (1973) Optimal Solution of Scheduling Problems Using Lagrangian Multipliers: Part I. *Operation Research*, **21**, 1114–27.

Grinold, R.C. (1970) Lagrangian subgradients. *Management Science*, **17**, 185–8.

Guignard, M., Kim, S. (1987) Lagrangean decomposition: a model yielding stronger lagrangean bounds. *Mathematical Programming*, **39**, 215–28.

Hiriart-Urruty, J.-B., Lemaréchal, C. (1993) *Convex Analysis and Minimization Algorithms*. Springer-Verlag.

Kiwiel, K.C. (1986) A Method for Solving Certain Quadratic Programming Problems Arising in Nonsmooth Optimization. *IMA Journal of Numerical Analysis*, **6**, 137–52.

Lemaréchal, C., Renaud, A. forthcoming paper.

Lemaréchal, C., Sagastizábal, C. (1994) An Approach to Variable Metric Bundle Methods in *Lecture Notes in Control and Information Sciences*, 197: System Modelling and Optimization.

Merlin, A., Sandrin, P. (1983) A New Method for Unit Commitment at Electricité de France. *IEEE Transactions on Power Apparatus and Systems*, **5**, 1218–25.

Shor, N. (1985) *Optimization methods for non-differentiable functions*, Springer-Verlag.