Paper

Direct method of hierarchical nonlinear optimization – reassessment after 30 years

Andrzej Karbowski

Abstract—We consider the optimization problems which may be solved by the direct decomposition method. It is possible when the performance index is a monotone function of other performance indices, which depend on two subsets of decision variables: an individual for every inner performance index and a common one for all. Such problems may be treated as a generalization of separable problems with the additive cost and constraints functions. In the paper both the underlying theory and the basic numerical techniques are presented and compared. A special attention is paid to the guarantees of convergence in different classes of problems and to the effectiveness of calculations.

Keywords—hierarchical optimization, decomposition, the direct method, Benders method, cutting plane method, distributed computations.

1. Introduction

We consider the following optimization problem:

$$\min_{x_1,x_2,\ldots,x_{p-1},v} \psi\Big(f_1(x_1,v), f_2(x_2,v),\ldots \\ \ldots, f_{p-1}(x_{p-1},v), f_p(v)\Big), \tag{1}$$

$$v \in V \subseteq \mathbb{R}^{n_v},\ x_i \in X_i \subseteq \mathbb{R}^{n_i},\ i=1,\ldots,p-1, \tag{2}$$

$$(x_i,v) \in XV_i = \big\{(x_i,v) : g_{ij}(x_i,v) \le 0, \\ j=1,\ldots,m_i\big\},\ i=1,\ldots,p-1, \tag{3}$$

where $\psi : \mathbb{R}^p \to \mathbb{R}$ is an order preserving (i.e., monotonically increasing with all its arguments), continuous function and all functions f_i, g_{ij} are convex and differentiable. We want to solve this problem applying hierarchical two-level approach with the decomposition of (1)–(3) in the direct way (so-called direct method). That is, we would like to apply the following computational scheme:

coordination problem (*CP*):

$$\min_{v\in V\cap V_0} \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v),\ldots \\ \ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big), \tag{4}$$

$$V_0 = \{v | \forall i \in \{1,\ldots,p-1\}\ \exists x_i \in X_i : \\ g_{ij}(x_i,v) \le 0 \quad \forall j=1,\ldots,m_i\big\}, \tag{5}$$

***i*th local problem** (LP_i), $i=1,\ldots,p-1$:

$$\hat{x}_i(v) = \arg\min_{x_i\in X_i} f_i(x_i,v), \tag{6}$$

$$g_{ij}(x_i,v) \le 0, \quad j=1,\ldots,m_i. \tag{7}$$

We will call the variables forming vector v coordinating or complicating variables (the last name stems from the observation, that when they are temporarily fixed the remaining optimization problem is considerably more tractable). They have to belong to a given explicitly set V and to an unknown set V_0, which is the set of admissible values of these variables from the point of view of the local problems. The set V_0 is called solvability set.

Such problems have been considered for more than 30 years in more [10, 17] or less [2] general statement. Surprisingly, they are often treated in some isolation from other problems, which are, in the autor's opinion, very close to them [1, 3, 5, 11]. The latter works were devoted to general problems with two (or more) sets of variables and the possibilities to iterate them in Gauss-Seidel manner to obtain the global optimum. There were no assumptions concerning specific structural properties of the performance index and the constraints' functions. Even terminology is different in these two types of problems. In the first case the variables forming the v vector are called the coordinating variables, while in the second–complicating variables.

The methods proposed depend on the presence of mixed constraints defining sets XV_i (3). If there are no such constraints, the theory considerably simplifies. It will be shown later on, that in this case the coordinating variables v stop to be complicating and there is no need to treat them in a different way than the others. It leads to plane (one level) decision structure, that is without the coordination level, even with some possibilities of desynchronization of calculations between different local units. When such constraints are present, the situation is more complicated and the coordination level is necessary, where the unknown set V_0 has to be taken into account when calculating new values of the coordinating vector v. In the article it will be shown, that actually, it is not necessary to look for a general method of determining the set V_0, and an efficient algorithm based on Kelley's cutting plane method [14], Benders decomposition [1] and ellipsoid method [15, 16] will be proposed.

2. The case of independent constraints on local and coordinating variables

If there are no mixed constraints on local and coordinating variables (7), that is in the definitions (3), (5) of sets XV_i and V_0 $m_i = 0\ \forall i$ we may take as these sets full domains, and as the consequence

$$V \cap V_0 = V \cap \mathbb{R}^{n_v} = V\,. \tag{8}$$

In such circumstances the coordination problem takes the form:

coordination problem for independent sets $(CP-I)$:

$$\min_{v\in V} \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v),\ldots \\ \ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big) \tag{9}$$

and the local problem

***i*th local problem for independent sets** (LP_i-I), $i = 1,\ldots,p-1$:

$$\hat{x}_i(v) = \arg\min_{x_i\in X_i} f_i(x_i,v)\,. \tag{10}$$

Such problem for additive cost function ψ was considered, e.g., in [2, p. 270]. However it seems, that there are possibilities to solve this and a more general problem with (1) performance index more effectively. First of all, let us take that the coordinating vector does not differ qualitatively from the other vectors x_i and denote it by x_p, and its set by X_p that is:

$$x_p = v, \qquad X_p = V, \qquad n_p = n_v\,. \tag{11}$$

Now denoting

$$n = \sum_{i=1}^{p} n_i \tag{12}$$

we will define the performance index $f : \mathbb{R}^n \mapsto \mathbb{R}$ hiding the structure of the function ψ, as:

$$f(x_1,x_2,\ldots,x_p) = \psi\Big(f_1(x_1,x_p), f_2(x_2,x_p),\ldots \\ \ldots, f_{p-1}(x_{p-1},x_p), f_p(x_p)\Big)\,. \tag{13}$$

For typographical convenience the partitioned column vectors:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

will be written in the form $(x_1,x_2,\ldots,x_p)$.

In this notation we deal with the following optimization problem:

$$\min_{x\in X} f(x)\,, \tag{14}$$

where

$$X = X_1 \times X_2 \times \ldots \times X_p\,, \tag{15}$$

$$x = (x_1,x_2,\ldots,x_p) \tag{16}$$

and $x_i \in \mathbb{R}^{n_i}$, $i = 1,\ldots,p$.

For problems with such general structure it is possible to propose two types of optimization algorithms:

- Jacobi algorithm:

$$x_i^{k+1} = \arg\min_{x_i\in X_i} f\Big(x_1^k,\ldots,x_{i-1}^k,x_i,x_{i+1}^k, \\ \ldots,x_p^k\Big),\ i = 1,\ldots,p\,, \tag{17}$$

- Gauss-Seidel algorithm:

$$x_i^{k+1} = \arg\min_{x_i\in X_i} f\Big(x_1^{k+1},\ldots,x_{i-1}^{k+1},x_i,x_{i+1}^k, \\ \ldots,x_p^k\Big),\ i = 1,\ldots,p\,, \tag{18}$$

where k denotes subsequent iterations.

So, in Jacobi algorithm the new values of subvector x_i, that is x_i^{k+1} for every i are obtained on the basis of the same information, that is they may be determined independently of each other. In Gauss-Seidel algorithm to determine new x_i the previous values of subvectors $x_{i+1},\ldots,x_p$ are used, but already new values of subvectors $x_1,\ldots,x_{i-1}$. We may say, that although both these algorithms use decomposition, Jacobi algorithm is parallel, while Gauss-Seidel sequential from its nature.

The following theorems concerning the convergence of these two algorithms have been formulated:

Proposition 1 [3, Prop. 3.9, p. 219]: Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously differentiable and convex function on the set X. Furthermore, suppose that for each i f is strictly convex function of x_i, when the values of the other components of x are held constant. Let $\{x^k\}$ be the sequence generated by the nonlinear Gauss-Seidel algorithm (18), assumed to be well defined. Then every limit point of $\{x^k\}$ minimizes f over X.

Proposition 2 [7]: Let $\{x^k\}$ be the sequence generated by the proximal Gauss-Seidel method:

$$x_i^{k+1} = \arg\min_{x_i\in X_i}\Big[f\Big(x_1^{k+1},\ldots,x_{i-1}^{k+1},x_i,x_{i+1}^k,\ldots,x_p^k\Big) \\ +\frac{1}{2}\tau_i||x_i - x_i^k||^2\Big],\quad i = 1,\ldots,p\,, \tag{19}$$

where $\tau_i > 0$, $i = 1,\ldots,p$. Then, if f is pseudoconvex on X, every limit point of $\{x^k\}$ is a global minimizer of problem (14).

Proposition 3 [3, Prop. 3.10, p. 221]: Let $f:\mathbb{R}^n \mapsto \mathbb{R}$ be a continuously differentiable function, let γ be a positive scalar, and suppose that the mapping $h: X \mapsto \mathbb{R}^n$, defined by

$$h(x) = x - \gamma \cdot \nabla f(x) \tag{20}$$

is a contraction with respect to the block-maximum norm $||x|| = ||(x_1, x_2, \ldots, x_p)|| = \max_i \frac{||x_i||_i}{w_i}$, where each $||.||_i$ is the Euclidean norm on $\mathbb{R}^{n_i}$ and each w_i is a positive scalar. Then there exists a unique vector $\hat{x}$ which minimizes f over X. Furthermore, the nonlinear Jacobi and Gauss-Seidel algorithms are well defined, that is, a minimizing x_i in Eqs. (17) and (18) always exists. Finally, the sequence $\{x^k\}$ generated by either of these algorithms converges to $\hat{x}$ geometrically.

The first two propositions concern Gauss-Seidel algorithm. Although, as it was written earlier, it is sequential from its nature, in the case of specific structural properties of the optimized functions, like in our case (13), it may be used to obtain the solution of the optimization problem. Owing to monotonicity of the function ψ, when it is strictly convex, the block coordinate problems (18) for $i = 1, \ldots, p-1$ may be simplified to:

$$x_i^{k+1} = \arg\min_{x_i \in X_i} f_i(x_i, x_p^k) \tag{21}$$

and solved independently. Only the last coordinate x_p has to be modified according to formula (18), for new optimal values of $x_1, x_2, \ldots, x_{p-1}$. Strict convexity is necessary to get from the local problems unique solutions. When this function is not strictly convex, but convex or pseudoconvex, according to Proposition 2, we may force the uniqueness of local solutions by adding quadratic proximal terms. Unfortunately, Grippo and Sciandrone theory [7] allows for independent, parallel solutions of block coordinate problems for $i = 1, \ldots, p-1$ only in the case of additive functions ψ.

The third proposition concerns a specific subclass of convex problems. The contraction condition for the mapping h (20) is satisfied for example (for functions from the $C_2(\mathbb{R}^n)$ class) when the Hessian of the function f is constrained, that is there exists such a constant K, that:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \leq K \quad \forall x \in \mathbb{R}^n, \forall i, j \tag{22}$$

and the domination of the main diagonal condition is fulfilled for a positive weights vector $[w_1, w_2, \ldots, w_n]$ (usually we take $w_i = 1, \forall i$):

$$w_i \cdot \frac{\partial^2 f}{\partial x_i^2} > \sum_{j \neq i} w_j \cdot \left| \frac{\partial^2 f}{\partial x_i \partial x_j} \right|. \tag{23}$$

In such conditions, if we take a sufficiently small coefficient γ, more precisely:

$$0 < \gamma < \frac{1}{K} \tag{24}$$

then the mapping h is a contraction in the maksimum norm. The functions whose Hessian is diagonally dominated are a subclass of the set of all convex functions. It results from the Gershgorin's circle theorem (saying that all eigenvalues of the matrix are contained within the union of n disks $K(a_{ii}, \sum_{j \neq i} |a_{ij}|)$, with each disk centered at a diagonal entry of the matrix and having radius equal to the sum of absolute values of off-diagonal entries in that row) and the equivalence of the positive signs of eigenvalues and positive definiteness in the class of symmetric matrices [6].

Unfortunately, not all convex functions have diagonally dominated Hessian. For example a quadratic form $f(x) = \frac{1}{2}x'Ax$ with the matrix:

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix} \tag{25}$$

is convex, but the diagonal dominance condition will never take place, i.e., there are no positive weights w_1, w_2, w_3 for which the condition (23) will be satisfied (it is easy to prove it by a contradiction).

Let us return now to our hierarchical algorithm and state the conclusions from the Proposition 3. We may say, that in the case of functions of the class $C_2(X)$ whose Hessian satisfies conditions (22), (23), and when there are no mixed constraints on local and coordinating variables, it is not necessary to realize the hybrid version of calculations: Gauss-Seidel iterations between coordination and local level and Jacobi iteration between different units of the local level. It is possible and should be useful to treat the coordination problem in the same way as the local problems. Due to the structural properties of the function f–see Eq. (13)–(that it grows monotonically with all functions f_i) the iterations (17) for $i = 1, \ldots, p-1$ will be equivalent to LP_i (6), that is:

$$\begin{aligned} x_i^{k+1} &= \arg\min_{x_i \in X_i} f\left(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_p^k\right) \\ &= \arg\min_{x_i \in X_i} \psi\left(f_1(x_1^k, x_p^k), \ldots, f_i(x_i, x_p^k), \ldots,\right. \\ &\qquad \left. \ldots, f_{p-1}(x_{p-1}^k, x_p^k), f_p(x_p^k)\right) \\ &= \arg\min_{x_i} f_i(x_i, x_p^k), \qquad i = 1, \ldots, p-1 \end{aligned} \tag{26}$$

while for $i = p$ (earlier it was problem CP-I)

$$\begin{aligned} x_p^{k+1} &= \arg\min_{x_p \in X_p} f\left(x_1^k, \ldots, x_{p-1}^k, x_p\right) \\ &= \arg\min_{x_p \in X_p} \psi\left(f_1(x_1^k, x_p), \ldots, f_{p-1}(x_{p-1}^k, x_p), f_p(x_p)\right). \end{aligned} \tag{27}$$

Until now nothing was said about the numerical optimization algorithm solving local problems. Since all local decision variables x_i have to belong to given sets X_i, they have to be constrained optimization procedures. The simplest way is to apply directly the steepest descent algorithm adding to it, to take into account the constraints, the orthogonal projection (with respect to Euclidean norm) of a vector

onto the convex set X_i. Let us define the projection operator $[y]_Z^+$ by:

$$[y]_Z^+ = \arg\min_{z\in Z} ||z-y||\,, \tag{28}$$

where $||.||$ is the Euclidean norm. The simplest constrained optimization algorithm implementing Jacobi iterations (17) will be then:

$$x_i := [h_i(x)]_{X_i}^+ = [x_i - \gamma\nabla_i f(x)]_{X_i}^+,\ \ i=1,\ldots,p\,. \tag{29}$$

Since the projection does not change nonexpansive property [3], this mapping will be a contraction when the mapping h is a contraction. Moreover, different x_i may be calculated totally asynchronously [3], that is without the need to make a new calculation or communication in any finite window.

But it was all about such convex problems with independent admissible sets, where the mapping h defined in (20) was contractive in the maximum block norm. What about these situations, rather more common, where this feature does not take place? Surprisingly, the last algorithm (29) is still valid. The only differences are in the restriction on γ coefficient and in the time dependencies between subsequent iterations of the ith local subvector x_i and in the exchange of information between different local units. If we denote by B the window (measured in the number of iterations of the whole algorithm) in which at least one iteration of each local units and the communication updating their values in the buffers of other units should take place, and by K_1 the Lipschitz constant for the gradient of a convex, nonnegative function f:

$$||\nabla f(x) - \nabla f(y)|| \le K_1 \cdot ||x-y||\ \ \forall x,y\in\mathbb{R}^n \tag{30}$$

the assessment on γ will be as follows [3]:

$$\gamma < \gamma_0(B) = \frac{1}{K_1(1+B+nB)}\,. \tag{31}$$

In this case we deal with so-called partially asynchronous implementation of the algorithm, where B is the measure of asynchronism. For functions f belonging to class $C_2(X)$ the constant K_1 equals K from the assessment (22).

It means, that in the case when the admissible sets are independent, it may be useful to abandon the hierarchical manner of solving the problem (1). In the "peer-to-peer" (Jacobi) version of the algorithm it might be possible to find the solutions faster and even in an asynchronous implementation.

3. The case of mixed constraints on local and coordinating variables

In this case the biggest problem with the above two-level algorithm (4)–(5), (6)–(7), which seems to be quite natural and promising, is that it is very difficult to calculate two things: the set V_0 and the functions $\hat{x}_i(v)$. Because of that the algorithm (4)–(7) is completely impractical–it cannot be directly applied. First of all, solving CP involves the reactivation of all local problems $LP_i, i=1,\ldots,p-1$ after every change of the v vector, that is after every movement in its optimization. It is so, because only in this way we may guarantee the proper first arguments of functions $f_i(\hat{x}_i(v),v)$. It is fast only in these rare cases when we may solve analytically local problems. Yet more difficult situation is with the solvability set V_0. This set is not given explicitly. The direct formula to calculate it was presented by Geoffrion [5] and is the following:

$$V_0 = \left\{ v\in\mathbb{R}^{n_v} : \max_{\lambda\in\Lambda} \min_{x_i\in X_i\, i=1,\ldots,p-1} \sum_{i=1}^{p-1}\sum_{j=1}^{m_i} \lambda_{ij}\cdot g_{ij}(x_i,v) \le 0 \right\}, \tag{32}$$

where

$$\Lambda = \left\{ \lambda\in\mathbb{R}^{m_1+m_2+\ldots+m_{p-1}} : \lambda \ge 0 \sum_{i=1}^{m_1+m_2+\ldots+m_{p-1}} \lambda_i = 1 \right\}. \tag{33}$$

So, it is rather difficult to estimate it and the computational effort to assess whether a given v belongs to this set is comparable with that of solving the whole optimization problem. It would be better to estimate this set by some additional constraints, possibly simple. In the book [10, p. 87] it is written, that: "In general the problem of defining inequalities and equations describing the set V_0 is unsolved" and as the only remedy the penalty function method is suggested:

coordination problem for penalty function method (CP-PFM):

$$\min_{v\in V} \psi\Big(f_1(\hat{x}_1(v),\hat{v}_1(v)) + \rho_{1k}||v-\hat{v}_1(v)||^2, \ldots$$
$$\ldots, f_{p-1}(\hat{x}_{p-1}(v),\hat{v}_{p-1}(v)) + \rho_{(p-1)k}||v-\hat{v}_{p-1}(v)||^2, f_p(v)\Big) \tag{34}$$

***i*th local problem for penalty function method** (LP_i-PFM) $i=1,\ldots,p-1$:

$$[\hat{x}_i(v),\hat{v}_i(v)] = \arg\min_{x_i\in X_i, v_i} \left[f_i(x_i,v_i) + \rho_{ik}||v-v_i||^2 \right]$$
$$g_{ij}(x_i,v_i) \le 0,\ j=1,\ldots,m_i\,. \tag{35}$$

However there is a possibility to estimate both the set V_0 and the function

$$\varphi(v) = \psi\Big(f_1(\hat{x}_1(v),v), f_2(\hat{x}_2(v),v), \ldots$$
$$\ldots, f_{p-1}(\hat{x}_{p-1}(v),v), f_p(v)\Big) \tag{36}$$

by a set of inequalities, growing as the computation progresses. This is a decomposition method proposed by Benders in early sixties [1]. He considered problems (called by him "mixed-variables programming" problems) where both the performance index and the constraints were sums of two components: one linear depending on one set of variables

and one nonlinear (they were called complicating variables; also because in many practical problems, e.g. [12], they are discrete). He proposed an iterative procedure for solving this problem by optimization with respect to either the first or the second group of variables in some auxiliary problems, related to dual representation of the initial problem and to optimality conditions. In the latter–the outer–the number of constraints on the variables corresponding to nonlinear part of the problem was gradually growing. They were delivered by the other–the inner–problem in the way dependent on the existence or not of the feasible solutions in the space of variables corresponding to linear components. Hence, in the decision space of nonlinear part variables either an "optimality cut" or "feasibility cut" was made.

In seventies the procedure proposed by Benders was generalized by Geoffrion [5] to the case of continuous nonlinear problems with performance indices and constraint functions being convex functions for fixed values of complicating variables. In later works Floudas *et al.* [12, 13] presented methods of transformation of many practical nonconvex and mixed continuous-discrete problems to apply this theory. A good review of these methods and well presentation of the algorithms may be found in [11]. We will present the basic procedure on the general problem:

$$\min_{x\in X, v\in V} f(x,v)\,, \tag{37}$$

$$g_j(x,v)\le 0,\ \ j=1,\ldots,m. \tag{38}$$

The solution algorithm is an iterative procedure where every iteration (let us say kth) consists of two parts:

1. Solving the primal problem for the current value of coordinating/complicating variables:

$$\min_{x\in X} f(x,v^k)\,, \tag{39}$$

$$g_j(x,v^k)\le 0,\ \ j=1,\ldots,m. \tag{40}$$

If the problem is feasible (i.e., there exists at least one point $x\in X$ for which all constraints (40) are satisfied) the optimal values of decision variables x^k and Lagrange multipliers λ_o^k are memorized (to be used in optimality cut later on). If not, the following problem assessing the departure from feasibility is solved:

$$\min_{x\in X,\alpha} \sum_{j=1}^{m} \alpha_j\,, \tag{41}$$

$$g_j(x,v^k)\le \alpha_j,\ \ j=1,\ldots,m, \tag{42}$$

$$\alpha_j\ge 0,\ \ j=1,\ldots,m. \tag{43}$$

The optimal values of decision variables $x\in X$ and the Lagrange multipliers in this problem λ_f^k are also memorized (to be used in feasibility cut in the next phase).

2. Solving the relaxed master problem:

$$\min_{\mu, v\in V} \mu\,, \tag{44}$$

$$L_o(x^k,v,\lambda_o^k)\le \mu,\ \ k\in K_o\,, \tag{45}$$

$$L_f(x^k,v,\lambda_f^k)\le 0,\ \ k\in K_f\,, \tag{46}$$

where

$$L_o(x^k,v,\lambda_o^k) = f(x^k,v)+{\lambda_o^k}^T g(x^k,v)\,, \tag{47}$$

$$L_f(x^k,v,\lambda_f^k) = {\lambda_f^k}^T g(x^k,v)\,. \tag{48}$$

Symbols K_o and K_f denote the sets of indices of iterations in which, respectively, the optimal solution of the primal problem existed or not. Functions L_o and L_f are Lagrange functions for primal (39)–(40) and feasibility (41)–(43) problems (the latter restricted to admissible solutions that is for $\hat{\alpha}_j = 0, \forall j$). The assessments on $\varphi(v)$ then result directly from the duality theory.

In the terms of the direct method of hierarchical optimization the first set of inequalities delivers the assessment of the function (36), while the second–the assessment of the set V_0 (32).

The most important classes of problems where this algorithm is proved to converge to the optimum are [5, 11] variable factor programming problems and problems with f, g_j, $j=1,\ldots,m$ linearly separable and convex in x and y, where X is a polyhedron.

The basic drawback of this method is the growing number of constraints of nonlinear type. In the next section we will show how to cope with it.

4. Combining Benders decomposition and Kelley's cutting plane method

Even in Benders' article at the end [1, p. 250] there is a remark on the solution of the relaxed master problem (44)–(46), that if the complicating variables (i.e., nonlinear) components are "convex and differentiable functions (...) problem becomes a convex programming problem that can be solved by-well known methods, e.g., by Kelley cutting plane technique...". It seemed attractive, because in the case when the set V is a polyhedron, if this method is used we actually deal with a linear programming problem.

Let us define:

$$\varphi(v) = L_o\big(\hat{x}_o(v), v, \hat{\lambda}_o(v)\big), \tag{49}$$

$$\xi(v) = L_f\big(\hat{x}_f(v), v, \hat{\lambda}_f(v)\big)\,, \tag{50}$$

where $\hat{x}_o(v)$ is solution of the primal problem (39)–(40), $\hat{x}_f(v)$ is the solution of the feasibility problem (41)–(43), and $\hat{\lambda}_o(v)$, $\hat{\lambda}_f(v)$ are Lagrange multipliers corresponding to

them for given complicating vector. While linearizing the constraints the following expressions may be used [9]:

$$\frac{\partial \varphi(v)}{\partial v} = \frac{\partial f}{\partial v} + \lambda_o^T \frac{\partial g}{\partial v}, \tag{51}$$

$$\frac{\partial \xi(v)}{\partial v} = \lambda_f^T \frac{\partial g}{\partial v}. \tag{52}$$

When we apply Kelley's cutting plane method together with the Benders decomposition, the relaxed master problem (44)–(46) will be replaced by:

$$\min_{\mu, v \in V} \mu, \tag{53}$$

$$\varphi(v^k) + \frac{\partial \varphi^T}{\partial v}(v^k)(v - v^k) \le \mu, \;\; k \in K_o, \tag{54}$$

$$\xi(v^k) + \frac{\partial \xi^T}{\partial v}(v^k)(v - v^k) \le 0, \;\; k \in K_f. \tag{55}$$

The problem is, that at this point Benders was wrong. This algorithm may fail and end in nonoptimal points even in convex problems. The counterexample was shown in Grothey *et al*. [8]. The convex NLP there was:

$$\min_{x_1, x_2, v} v^2 - x_2, \tag{56}$$

$$(x_1 - 1)^2 + x_2^2 \le \ln v, \tag{57}$$

$$(x_1 + 1)^2 + x_2^2 \le \ln v, \tag{58}$$

$$v \ge 1. \tag{59}$$

The optimal solution of this problem is $[x_1, x_2, v] \simeq [0, 0.0337568, 2.721381]$. Starting with the feasible $v = e^2$ we obtain in the first step $\hat{x}_o(e^2) = [0,1]$ and the optimality cut:

$$(e^4 - 1) + \left(2e^2 - \frac{1}{2e^2}\right)(v - e^2) \le \mu. \tag{60}$$

From the relaxed master problem we obtain the new optimal $v = 1 < e$. For this value, however, the primal problem is infeasible and we will get from the feasibility problem $\hat{x}_f = [0,0]$. In general, if $v^k < e$, the following feasibility cut is generated and added to the master problem:

$$(2 - 2\ln v^k) + \left(-\frac{2}{v^k}\right)(v - v^k) \le 0 \;\Leftrightarrow\; v \ge (2 - \ln v^k) v^k.$$

The next values of v from the master problem will be calculated according to the formula:

$$v^{k+1} = (2 - \ln v^k) v^k.$$

They all will be from the interval $(1, e)$ giving the whole time infeasibility and the same optimal values in feasibility problems (actually the sequence v^k will approach e from the left hand side). The authors explain that "the failure of Benders decomposition to converge is due to the fact that the Benders cuts only approach feasibility in the limit and never collect subgradient information from the objective" function of the problem $\varphi(v)$. As the remedy they propose, as they call, "feasibility restoration algorithm", where in the case of infeasibility, after solving feasibility problem, the modified primal problem is solved again with the modified inequalities (40) in such a way, that on the right hand side of them there are positive numbers being values of constraints in the feasibility problem multiplied by some coefficient bigger than 1. Then both the previously obtained feasibility as well as the optimality cuts from this relaxed problem are added to the master problem constraints.

This procedure overcomes the basic disadvantage of the Benders method combined with Kelley's cutting plane algorithm–converging to nonoptimal points, but still has one drawback: the growing number of constraints in master problems as the calculations proceed. One has to wait longer and longer for new values of complicating variables v. How to overcome this difficulty and even to replace the optimization on the upper level with calculation of values of two simple analytic expressions will be shown in the next section.

5. Integration of Benders decomposition with cutting plane and ellipsoid algorithms

The main idea lies in the application (instead of solving relaxed master problem as the optimization problem) one of the simplest algorithms of nondifferentiable optimization, which was proposed by Shor [16], Nemirovski and Yudin [15], namely the ellipsoid algorithm. It will deliver in subsequent iterations the centers of the smallest volume ellipsoids, containing smaller and smaller sets of admissible points, in which the performance index may have a better value than in points it was calculated so far. It is obtained by cutting off the halfspaces of points in which, owing to convexity, for sure the value of performance index is worse than in the current point (if it is feasible) or the value of functions defining constraints is worse than the present one (if the current point is infeasible).

What concerns optimality cuts, we use the same formulae for derivatives as before, that is, it is defined by:

$$\left\langle \frac{\partial \varphi}{\partial v}(v^k), v - v^k \right\rangle \le 0. \tag{61}$$

The calculation of feasibility cuts may be simplified by making individual cuts for the most violated constraint. It is described in the next subsection.

5.1. Feasibility cuts

We will perform for every query point $v^k \in V$ (that is the current value of coordinating variables) verification of the feasibility from the point of view of the constraints (7), independently for all local problems LP_i, $i = 1, \ldots, p-1$, and adding the corresponding linear constraints to coordination problem in the case of a failure.

The verification of feasibility consists in calculation for every LP_i the "constraint index" $g_i(v^k)$, by solving a preliminary optimization problem:

$$g_i(v^k) = \min_{x_i \in X_i} \max_{j=1,\ldots,m_i} g_{ij}(x_i, v^k) \tag{62}$$

before the principal optimization problem.

In the case when $g_i(v^k) > 0$, we draw a conclusion, that for the query point v^k there are no admissible points $x_i \in X_i$ and the rational thing is cutting off a halfspace containing inadmissible values of coordinating variables. This will be obtained by the condition:

$$g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v - v^k \right\rangle \leq 0\,, \tag{63}$$

where $x_{i_{\min}}, j^*, v^k$ are such that

$$g_{ij^*}(x_{i_{\min}}, v^k) = g_i(v^k) > 0\,. \tag{64}$$

Proposition 4: Condition (63) assures the elimination from the admissible set V points not belonging to the solvability set V_0.

Proof: To prove the proposition we have to show that:

$$\forall v^* \in V \quad g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle > 0$$
$$\Rightarrow \forall x_i \in \mathbb{R}^{n_i} \quad g_{ij^*}(x_i, v^*) > 0\,. \tag{65}$$

From the convexity and smoothness of functions g_{ij} we have for any given pair $(\tilde{x}_i, v^k)$ and all x_i,v:

$$g_{ij}(x_i, v) \geq$$
$$g_{ij}(\tilde{x}_i, v^k) + \left\langle \frac{\partial g_{ij}}{\partial x_i}(\tilde{x}_i, v^k), x_i - \tilde{x}_i \right\rangle + \left\langle \frac{\partial g_{ij}}{\partial v}(\tilde{x}_i, v^k), v - v^k \right\rangle. \tag{66}$$

Setting in (66) $j = j^*$, $\tilde{x}_i = x_{i_{\min}}$ and $v = v^*$ we will get $\forall x_i \in \mathbb{R}^{n_i}$

$$g_{ij^*}(x_i, v^*) \geq g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial x_i}(x_{i_{\min}}, v^k), x_i - x_{i_{\min}} \right\rangle$$
$$+ \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle. \tag{67}$$

That is

$$g_{ij^*}(x_i, v^*) \geq \left[g_{ij^*}(x_{i_{\min}}, v^k) + \left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle \right]$$
$$+ \left\langle \frac{\partial g_{ij^*}}{\partial x_i}(x_{i_{\min}}, v^k), x_i - x_{i_{\min}} \right\rangle. \tag{68}$$

Let us notice, that from the assumption, the term in square brackets is positive. The second component is nonnegative, because we assumed, that $x_{i_{\min}}$ is the solution of the minimax problem. This means that

$$g_{ij^*}(x_i, v^*) > 0 \quad \forall x_i \in \mathbb{R}^{n_i} \tag{69}$$

what completes the proof. □

The interpretation of the Proposition 4 is, that by cutting off from the set V more and more points, we get a better estimate of the set $V \cap V_0$, that is the admissible set in the CP problem (4)–(5).

So, if we restrict our attention to these points of the decision space in which the value of the most violated constraint function g_{ij^*} is better than in the current point, it is sufficient to add a constraint:

$$\left\langle \frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k), v^* - v^k \right\rangle \leq 0\,. \tag{70}$$

5.2. Ellipsoid algorithm

The presented algorithm was proposed by Shor [16], Nemirovski and Yudin [15]. At every step we obtain an ellipsoid

$$E_k = \left\{ v | (v - v^k)^T W_k^{-1} (v - v^k) \leq 1 \right\}. \tag{71}$$

It is characterized by two parameters: a matrix W_k and a center v^k. It is assumed, that we start from an ellipsoid E_0 containing the admissible set V. The subsequent ellipsoids E_k are such that E_{k+1} is the minimum volume ellipsoid containing $E_k \cap \{v | \langle h_k, v - v^k \rangle \leq 0\}$. It is defined by:

$$v^{k+1} = v^k - \frac{1}{n_v + 1} \frac{W_k h_k}{\sqrt{h_k^T W_k h_k}}\,, \tag{72}$$

$$W_{k+1} = \frac{n_v^2}{n_v^2 - 1} \left(W_k - \frac{2}{n_v + 1} \frac{W_k h_k h_k^T W_k}{h_k^T W_K h_k} \right), \tag{73}$$

where n_v is the dimension of v. It can be shown, that the volume of E_{k+1} equals the volume of E_k reduced by the factor $(1 - 1/(n_v + 1)^2)$.

5.3. Integration

If we use as the vector h_k in expressions modifying ellipsoids (72), (73) the gradient $\frac{\partial \varphi}{\partial v}(v^k)$ from optimality cut expressions (61), (51) or the gradient $\frac{\partial g_{ij^*}}{\partial v}(x_{i_{\min}}, v^k)$ from feasibility cut expression (70), we will have what we need–a very simple and fast metod of delivering subsequent values of coordinating (complicating) variables with the convergence guarantee.

This approach seems to be the most promising among all, because the calculations on the coordination level are the simplest one can imagine: only two direct formulas without any optimization, iterative process, etc. There are other techniques from cutting plane family generating queries

at new points inside the admissible area (e.g., center of gravity, largest inscribed sphere, volumetric, analytic center methods–see [4]), which prevents the algorithm against blocking, but none of them has so simple and fast master problem iteration.

6. Conclusions

In the paper the basic approaches to solving optimization problems with generalized separable structure, where the performance index is a monotone function of other performance indices depending on individual and common subvectors of decision variables, were presented and compared. It was shown, that in the case when the admissible set is a Cartesian product of individual domains and a domain of common variables (that is coordinating or complicating variables), the problem may be solved by application of hybrid Gauss-Seidel (between coordination and local level) and Jacobi (between different units of the local level) algorithms or in a completely symmetric (Jacobi) version, even with asynchronous iterations. The degree of asynchronism depends on the features of the overall performance index. If its Hessian is restricted and diagonally dominated the steepest descent type iterations and the exchange of information may be totally asynchronous, otherwise they may be partially asynchronous, that is with iterations and communication between local units in a given finite window, dependent on the length of the step in optimization iterations, the dimension of the problem and the assessment on the Hessian elements.

The situation is much more complicated when the admissible set is not a Cartesian product of local and common variables domains. The most natural seems to be the Benders decomposition, where so-called optimality and feasibility cuts obtained after, respectively, admissible or inadmissible queries of complicating/coordinating variables are used to estimate the value function (i.e., the function whose value is the optimal value of the original problem for fixed values of coordinating variables) and the solvability set (i.e., the set of complicating variables for which all mixed constraints can be satisfied for at least one combination of the primal variables). This approach, based on duality relations, although very general and elegant, has one serious drawback–since the estimates of both the value function and the solvability set have to be more accurate as the computations progress, the number of constraints defining them systematically grows. It means, that the problems solved in subsequent iterations are more are more complicated and the time needed for one iteration of master problem is longer and longer. An attempt to simplify calculations by combining Benders decomposition with Kelley's cutting plane method and transform the master problem to LP problem is not a good idea, because, as it was shown in an example, the optimization process even in the convex case may converge to a nonoptimal point. It is possible to avoid it by either so-called feasibility restoration algorithm, which adds an additional optimality cut in an extended domain, or the application on the master (i.e., coordination) level an algorithm which delivers query points lying inside the admissible area, e.g., center of gravity method, the largest inscribed sphere or ellipsoid method, volumetric center method, analytic center method (ACCPM) or the smallest circumscribing ellipsoid method. The latter approach seems to be the most attractive due to its simplicity, noniterative character (that is, the new values of complicating variables are not determined, as for example in optimization, via an iterative process, but directly from two simple formulas) and converges to optimal solution with the geometric rate.

References

[1] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems", *Numer. Math.*, vol. 4, pp. 238–252, 1962.

[2] D. P. Bertsekas, *Nonlinear Programming.* 2nd ed. Belmont: Athena Scientific, 1999.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs: Prentice Hall, 1989.

[4] S. Elhedhli, J. L. Goffin, and J.-P. Vial, "Nondifferentiable optimization: cutting plane methods", in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Dordrecht: Kluwer, 2001, vol. 4, pp. 40–45.

[5] A. M. Geoffrion, "Generalized Benders decomposition", *J. Opt. Theory Appl.*, vol. 10, pp. 237–260, 1972.

[6] G. Golub and J. M. Ortega, *Scientific Computing: an Introduction with Parallel Computing.* San Diego: Academic Press, 1993.

[7] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints", Report R. 467, Istituto di Analisi dei Sistemi ed Informatica, CNDR, Settembre 1998.

[8] A. Grothey, S. Leyffer, and K. I. M. McKinnon, "A note on feasibility in Benders decomposition", Numerical Analysis Report NA/188, Dundee University, 1999.

[9] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Mathematics in Science and Engineering. New York: Academic Press, 1983, vol. 165.

[10] W. Findeisen, F. N. Bailey , M. Brdyś, K. Malinowski, P. Tatjewski, and A. Woźniak, *Control and Coordination in Hierarchical Systems*. Chichester: Wiley, 1980.

[11] C. A. Floudas, "Generalized Benders decomposition, GBD", in *Encyclopedia of Optimization*, C. A. Floudas and P. M. Pardalos, Eds. Dordrecht: Kluwer, 2001, vol. 2, pp. 207–218.

[12] C. A. Floudas, A. Aggarwal, and A. R. Ciric, "Global optimum search for nonconvex NLP and MINLP problems", *Comput. Chem. Eng.*, vol. 13, no. 10, pp. 1117–1132, 1989.

[13] C. A. Floudas and V. Visweswaran, "A primal-relaxed dual global optimization approach", *J. Opt. Theory Appl.*, vol. 78, no. 2, pp. 187–225, 1993.

[14] J. E. Kelley, "The cutting-plane method for solving convex programs", *J. Soc. Indust. Appl. Math.*, vol. 8, pp. 703–712, 1960.

[15] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Chichester: Wiley, 1983.

[16] N. Z. Shor, *Minimization Methods for Non-differentiable Functions*. Berlin: Springer Verlag, 1985.

[17] *Optimization Methods for Large-Scale Systems with Applications*, D. A. Wismer, Ed. New York: McGraw-Hill, 1971.

Andrzej Karbowski received his M.Sc. degree in electronic engineering (specialization automatic control) from Warsaw University of Technology (Faculty of Electronics) in 1983. He received the Ph.D. in 1990 in automatic control and robotics. He works as adjunct both at Research and Academic Computer Network (NASK) and at the Faculty of Electronics and Information Technology (at the Institute of Control and Computation Engineering) of Warsaw University of Technology. His research interests concentrate on data networks management, optimal control in risk conditions, decomposition and parallel implementation of numerical algorithms.
e-mail: A.Karbowski@ia.pw.edu.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland