International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Augmented Lagrangian Decomposition for Sparse Convex Optimization

Ruszczynski, A.

IIASA Working Paper

WP-92-075

October 1992

Working Paper

Augmented Lagrangian Decomposition for Sparse Convex Optimization

Andrzej Ruszczyński

WP-92-75
October 1992
(revised April 1993)

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

Augmented Lagrangian Decomposition for Sparse Convex Optimization

Andrzej Ruszczyński

WP-92-75
October 1992
(revised April 1993)

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 715210 □ Telex: 079 137 iiasa a □ Telefax: +43 2236 71313

Abstract

A decomposition method for large-scale convex optimization problems with block-angular structure and many linking constraints is analysed. The method is based on a separable approximation of the augmented Lagrangian function. Weak global convergence of the method is proved and speed of convergence analysed. It is shown that convergence properties of the method are heavily dependent on sparsity of the linking constraints. Application to large scale linear programming and stochastic programming is discussed.
Keywords: Large-Scale Optimization, Decomposition, Augmented Lagrangians.

1. Introduction

Rapid development of computing technology, emergence of parallel, massively parallel and distributed computing systems provides us with an increasing computing power but also creates a need for specialized approaches that can use it efficiently. The principal objective of this paper is to analyse properties of a decomposition method for very large optimization problems which can be easily implemented on a variety of parallel and distributed computer architectures.

Let $X_1, X_2, \ldots, X_L$ be nonempty closed convex subsets of $R^{n_1}, R^{n_2}, \ldots, R^{n_L}$, respectively, and let $f_i : R^{n_i} \to R$, $i = 1, 2, \ldots, L$ be convex functions. Next, let A_i be matrices of dimension $m \times n_i$, $i = 1, 2, \ldots, L$ and let $b \in R^m$. We consider the convex programming problem

$$\min \left[f(x) = \sum_{i=1}^{L} f_i(x_i) \right] \tag{1.1a}$$

$$\sum_{i=1}^{L} A_i x_i = b, \tag{1.1b}$$

$$x_i \in X_i, \ i = 1, 2, \ldots, L. \tag{1.1c}$$

There is a vast literature devoted to decomposition methods in linear and nonlinear programming (see, e.g., [11]). They are usually in one way or another related to the famous decomposition principle of Dantzig and Wolfe [4] and to the duality theory based on the ordinary Lagrangian function.

A much smaller number of works is devoted to decomposition-type methods based on the theory of augmented Lagrangians. Although the augmented Lagrangian function does not possess decomposability properties of the ordinary Lagrangian, some special tricks and problem transformations can be used to allow decomposition (see [3, 6, 5, 18, 19, 20, 21]).

A promising decomposition method based on the augmented Lagrangian function - the Diagonal Quadratic Approximation Method (DQA) - has been succesfully applied to large-scale stochastic optimization in [12] and [13]. Its basic idea of separable quadratic approximation of the augmented Lagrangian can be traced back to [19]. The method proved well-suited for parallel and distributed computation because it has modest communication requirements and allows for a distributed implementation of the coordination procedure [13]. It has found successful application in stochastic programming and appears to have potential to solve a much broader class of problems.

The objective of this paper is to present the method in a more general form for large-scale convex optimization problems and to carry out a detailed analysis of its convergence properties.

The paper is organized as follows.

In section 2 we remind some basic definitions and results associated with augmented Lagrangians and the multiplier method. Section 3 is devoted to the description of the DQA method; we also introduce some sparsity measures for the linking constraints. In section 4 we prove global convergence of DQA (in terms of objective function values). In particular, we formulate conditions on the stepsize that directly involve sparsity

properties of the problem. Speed of convergence is analysed in section 5: we show that for some important classes of problems (such as linear or quadratic programs) the speed is heavily dependent on the number of blocks linked by any single constraint. In this way we relate sparsity to the number of iterations necessary to solve the problem rather than to the cost of one iteration. Finally, in section 6 we have some examples that illustrate the potential of the method.

Our approach and results differ significantly from earlier works on similar ideas. The results of the pioneering work [19] are mainly empirical. In [5] and [20] there is a local convergence analysis for problems with twice differentiable functions under second order sufficient conditions. Finally, [13] considers linear stochastic programming problems. Our results generalize and improve those of [13] by broadening the class of problems under consideration (convex problems with general linking constraints), by providing detailed convergence rates and by introducing the issue of sparsity as the key factor in the assessment of efficiency of DQA. In the special case of stochastic linear programs they are sharper than those of [13].

In our work we use elementary notions and results of convex analysis (for an extensive treatment see [15, 14, 9]). For every x in a convex set $X \subset R^n$ we use $K_X(x)$ to denote the cone of feasible directions

$$K_X(x) = \{d \in R^n : \ d = \beta(y - x), \ y \in X, \ \beta \geq 0\}.$$

Its conjugate cone $(K_X(x))^*$ is defined by

$$\begin{aligned} (K_X(x))^* &= \{g \in R^n : \langle g, d\rangle \geq 0 \text{ for all } d \in K_X(x)\} \\ &= \{g \in R^n : \langle g, y - x\rangle \geq 0 \text{ for all } y \in X\}. \end{aligned}$$

For a convex function $f : R^n \to R$ and $\epsilon \geq 0$ we define the ϵ-subdifferential at x by

$$\partial_\epsilon f(x) = \{g \in R^n : \ f(y) \geq f(x) + \langle g, y - x\rangle - \epsilon \ \text{ for all } y \in R^n\};$$

its elements are called ϵ-subgradients. For $\epsilon = 0$ we shall call $\partial_0 f(x)$ the subdifferential, denote it by $\partial f(x)$, and call its elements - subgradients of f at x.

2. Preliminaries

The *ordinary Lagrangian* associated with (1.1) has the form

$$L(x, \pi) = \sum_{i=1}^{L} f_i(x_i) + \langle \pi, b - \sum_{i=1}^{I} A_i x_i\rangle = \langle b, \pi\rangle + \sum_{i=1}^{L} \left(f_i(x_i) - \langle A_i^T \pi, x_i\rangle\right). \qquad (2.1)$$

We can use it to derive the *dual problem*

$$\max_{\pi \in R^m} g(\pi), \qquad (2.2)$$

where g is the *dual functional*,

$$g(\pi) = \inf_{x \in X} L(x, \pi) = \langle b, \pi\rangle + \sum_{i=1}^{L} g_i(\pi) \qquad (2.3)$$

with $X = X_1 \times X_2 \times \cdots \times X_L$ and

$$g_i(\pi) = \inf_{x_i \in X_i} \left[f_i(x_i) - \langle A_i^T \pi, x_i \rangle \right] \quad i = 1, 2, \ldots, L. \tag{2.4}$$

Relations between (1.1) and (2.2) are based on fundamental results of the duality theory for convex programming (see, e.g., [15, 14]).

Proposition 1. *Assume that (1.1) has an optimal solution and at least one of the following conditions is satisfied:*

(i) $\operatorname{ri} K_X(x^0) \cap \{d : Ad = 0\} \neq \emptyset$ *at some* $x^0 \in X$ such that $Ax^0 = b$; *or*

(ii) X *is a polyhedral set.*

Then (2.2) has an optimal solution and

(a) for every optimal solution $\hat{x}$ *of (1.1) and every optimal solution* $\hat{\pi}$ *of (2.2)*

$$f(\hat{x}) = g(\hat{\pi});$$

(b) for every optimal solution $\hat{\pi}$ *of (2.2) a point* $\hat{x}$ *is a solution of (1.1) if and only if*

$$L(\hat{x}, \hat{\pi}) = \min_{x \in X} L(x, \hat{\pi}), \tag{2.5a}$$

$$A\hat{x} = b. \tag{2.5b}$$

Condition (2.5a) can be also expressed in an equivalent subdifferential form: *there exist* $g_i \in \partial f_i(\hat{x})$, $i = 1, 2, \ldots, L$ *such that*

$$g_i - A_i^T \hat{\pi} \in (K_{X_i}(\hat{x}_i))^*, \quad i = 1, 2, \ldots, L. \tag{2.6}$$

It is exactly Proposition 1 that motivates the classical decomposition methods for (1.1). Calculating the dual function (2.3) and its subgradients simplifies by decomposition into independent problems in (2.4), so (2.2) can be easier to solve than (1.1).

However, there are well-known disadvantages of the dual approach based on the ordinary Lagrangian (2.1). They are associated with the non-uniqueness of the solutions of subproblems (2.4) and more precisely with the non-uniqueness of the value of Ax at these solutions. It results in non-differentiability of the dual functional (2.3) and calls for an application of rather involved nonsmooth optimization methods for solving the dual problem (2.2). Even in the linear case, the Dantzig-Wolfe method requires constructing the master problem, which may be interpreted as a cutting plane method for solving (2.2) (see [11]). Recovery of the primal solution by (2.5) is not easy, too. For very large problems with many linking constraints (1.1b) these difficulties make the ordinary dual approach impractical.

There are two closely related ways of overcoming this difficulty, both based on regularization. The primal regularization method, known as the *proximal point method* adds to the objective of (1.1) the quadratic term $\frac{1}{2}\rho\|x - \xi\|^2$ with some penalty parameter $\rho > 0$. It makes (1.1) strictly convex which implies existence and uniqueness

of solutions to the regularized versions of (2.4). This results in an improved behavior of the dual functional, but has the drawback that an additional outer iteration eoop over ξ is necessary.

A corresponding dual approach is the regularization in the space of multipliers π. In place of the ordinary Lagrangian (2.1) we introduce the *augmented Lagrangian*

$$\begin{aligned}\Lambda(x,\pi) &= f(x) + \langle \pi, b - Ax\rangle + \frac{1}{2}\rho\|b - Ax\|^2 \\ &= \langle b,\pi\rangle + \sum_{i=1}^{L}\left(f_i(x_i) - \langle A_i^T\pi, x_i\rangle\right) + \frac{1}{2}\rho\left\|b - \sum_{i=1}^{L} A_i x_i\right\|^2, \qquad (2.7)\end{aligned}$$

with a penalty parameter $\rho > 0$. For the augmented Lagrangian we can formally copy the duality results from Proposition 1 with the regularized dual function

$$\gamma(\pi) = \inf_{x\in X}\Lambda(x,\pi)$$

and the regularized dual problem

$$\max_{\pi\in R^m}\gamma(\pi). \qquad (2.8)$$

There are many theoretical and computational advantages of the augmented Lagrangian approach over the ordinary dual method. The most important one is the possibility of solving the dual problem (2.8) by the following algorithm.

Method of Multipliers

Step 1. For fixed multipliers π^k find a solution x^k of the problem

$$\min_{x\in X}\Lambda(x,\pi^k). \qquad (2.9)$$

Step 2. If $Ax^k = b$ then stop (optimal solution found); otherwise set

$$\pi^{k+1} = \pi^k + \rho(b - Ax^k), \qquad (2.10)$$

increase k by 1 and go to Step 1.

The following two propositions summarize the fundamental properties of the method of multipliers.

Proposition 2. *Let the assumptions of Proposition 1 be satisfied. Then the sequence $\{\pi^k\}$ generated by the method of multipliers is convergent to a solution $\hat{\pi}$ of (2.2).*

Proposition 3. *Assume that f_i, $i = 1,2,\ldots,L$ are convex polyhedral functions and X_i, $i = 1,2,\ldots,L$ are convex polyhedral sets. Then, if (1.1) has a solution, the method of multipliers is convergent in finitely many iterations.*

However, a serious disadvantage of the method of multipliers is that (2.7) is not separable, so problem (2.9) cannot be split into independent subproblems for x_i, $i = 1,2,\ldots,L$. One possibility to overcome this difficulty is the use of *alternating direction methods* (cf. [8, 7, 6]).

In the next section we shall present another method for decomposing the augmented Lagrangian. It is based on successive separable approximations of (2.7) and extends and refines the earlier ideas of [19] and [12] (for a related work see [3, 20]).

3. The separable approximation

Clearly, non-separability of (2.7) is due to the existence of the quadratic penalty term, which contains products $\langle A_i x_i, A_j x_j \rangle$. To overcome this difficulty we introduce for $i = 1, 2, \ldots, L$ the functions $\Lambda_i : R^{n_i} \times R^n \times R^m \to R$,

$$\Lambda_i(x_i, \tilde{x}, \pi) = f_i(x_i) - \langle A_i^T \pi, x_i \rangle + \frac{1}{2}\rho \left\| b - A_i x_i - \sum_{j \neq i} A_j \tilde{x}_j \right\|^2, \tag{3.1}$$

where $\tilde{x} \in R^n$ is an additional parameter, $n = \sum_{i=1}^{L} n_i$. The main idea of our approach is to replace problem (2.9) by L problems

$$\min_{x_i \in X_i} \Lambda_i(x_i, \tilde{x}, \pi^k), \quad i = 1, 2, \ldots, L \tag{3.2}$$

and to iteratively update the parameter $\tilde{x}$ by making steps towards the solutions of (3.2). It is not difficult to see that (3.2) is equivalent to the minimization of (2.7) with respect to x_i with x_j, $j \neq i$ frozen at $\tilde{x}_j$. However, we are not going to use (3.2) in a Gauss-Seidel fashion, but we shall rather solve it for each i in parallel and then update $\tilde{x}$. This approach is called in [2] a *nonlinear Jacobi algorithm.*

It is well known that for $\hat{x}_i$ to be a solution of (3.2) it is necessary and sufficient that

$$\partial \Lambda_i(\hat{x}_i, \tilde{x}, \pi^k) \cap (K_{X_i}(\hat{x}_i))^* \neq \emptyset, \tag{3.3}$$

where the subdifferential is with respect to the first argument (see, e.g., [15, 14]). However, it is in general rather difficult to solve (3.2) with perfect accuracy, especially when f_i is a nonsmooth function. Therefore, we introduce the set of approximate subgradients of Λ_i

$$D_i^\epsilon(x_i, \tilde{x}, \pi) = \partial_\epsilon f_i(x_i) - A_i^T \pi - \rho A_i^T \left(b - A_i x_i - \sum_{j \neq i} A_j \tilde{x}_j \right). \tag{3.4}$$

It is obvious that $D_i^\epsilon(x_i, \tilde{x}, \pi) \subset \partial_\epsilon \Lambda_i(x_i, \tilde{x}, \pi)$ but we assume that the gradient of the quadratic term is exact and only subgradients of f_i are subject to errors. Basing on that, we introduce the set of approximate solutions

$$S_i^\epsilon(\tilde{x}, \pi) = \{x_i \in X_i : \ D_i^\epsilon(x_i, \tilde{x}, \pi) \cap (K_{X_i}(x_i))^* \neq \emptyset\}. \tag{3.5}$$

In other words, for every $x_i \in S_i^\epsilon(\tilde{x}, \pi)$ it must be possible to find

$$g_i \in D_i^\epsilon(x_i, \tilde{x}, \pi) \tag{3.6}$$

such that

$$g_i \in (K_{X_i}(x_i))^*. \tag{3.7}$$

We are now ready to describe the method in detail. It should be noted that DQA is a sub-algorithm for carrying out Step 1 of the method of multipliers in a decomposed fashion.

In what follows $\tau > 0$ and $0 \leq \mu < 1$ are parameters of the method.

The DQA Method

Step 0. Set $\tilde{x}^{k,0} = x^{k-1}$ and $s = 0$.

Step 1. For $i = 1, 2, \ldots, L$ find

$$x_i^{k,s} \in S_i^{\epsilon_s}(\tilde{x}^{k,s}, \pi^k) \tag{3.8}$$

with

$$\epsilon_s \leq \frac{1}{2}\mu\rho\|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2. \tag{3.9}$$

Step 2. If $A_i x_i^{k,s} = A_i \tilde{x}_i^{k,s}$, $i = 1, 2, \ldots, L$, then stop; otherwise set for $i = 1, 2, \ldots, L$

$$\tilde{x}_i^{k,s+1} = \tilde{x}_i^{k,s} + \tau(x_i^{k,s} - \tilde{x}_i^{k,s}), \tag{3.10}$$

increase s by 1 and go to Step 1.

Remark. It is worth noting that the above method is (abstractly speaking) implementable in the sense that the stopping criteria (3.8)-(3.9) depend on the current point $x_i^{k,s}$, not on the solution of (3.2). As an illustration, let us assume that (3.2) is solved by minimizing a function

$$\underline{\Lambda}_i(x_i, \tilde{x}^{k,s}, \pi^k) = \underline{f}_i(x_i) - \langle A_i^T \pi^k, x_i \rangle + \frac{1}{2}\rho \left\| b - A_i x_i - \sum_{j \neq i} A_j \tilde{x}_j^{k,s} \right\|^2,$$

where $\underline{f}_i$ is a piecewise-linear lower approximation of f_i. If $x_i^{k,s}$ is a solution of this approximate problem, then there is a subgradient $h_i \in \partial \underline{f}_i(x_i^{k,s})$ such that

$$h_i - A_i^T \pi^k - \rho A_i^T \left(b - A_i x_i^{k,s} - \sum_{j \neq i} A_j \tilde{x}_j^{k,s} \right) \in (K_{X_i}(x_i^{k,s}))^*.$$

Obviously, $h_i \in \partial_\epsilon f_i(x_i^{k,s})$ with

$$\epsilon = f_i(x_i^{k,s}) - \underline{f}_i(x_i^{k,s}).$$

This is the value of ϵ_s in (3.8) at $x_i^{k,s}$. If (3.9) holds, we can stop the minimization procedure; otherwise the piecewise-linear approximation has to be improved (e.g., by ading a cut derived at $x_i^{k,s}$, see [10]).

In the next section we shall show that if the parameters τ and μ fulfill some simple conditions, then the DQA method generates sequences $\{x^{k,s}\}_{s=0}^{\infty}$ and $\{\tilde{x}^{k,s}\}_{s=0}^{\infty}$ whose accumulation points are solutions of (2.9). But before proceeding to the convergence analysis we shall make a simple observation that will allow us to obtain much stronger convergence results and estimates of the speed of convergence than earlier works.

Let m_i be the number of nonempty rows of A_i and let us define a zero-one matrix E_i of dimension $m \times m_i$ as follows: it has a 1 at position (k, l) iff the k-th row of A_i is the l-th consecutive nonempty row of A_i. Thus, columns of E_i are orthonormal

unit vectors, $E_i^T E_i = I$. Introducing additional variables $z_i \in R^{m_i}$ we can equivalently reformulate (1.1) as follows

$$\min \sum_{i=1}^{L} f_i(x_i) \tag{3.11a}$$

$$\sum_{i=1}^{L} E_i z_i = b, \tag{3.11b}$$

$$(x_i, z_i) \in Z_i, \quad i = 1, 2, \ldots, L, \tag{3.11c}$$

with

$$Z_i = \{(x_i, z_i) : \ x_i \in X_i, \ A_i x_i = E_i z_i\}. \tag{3.12}$$

The subproblems (3.2) for the new formulation have the form

$$\min_{(x_i, z_i) \in Z_i} \left\{ f_i(x_i) - \langle E_i^T \pi^k, z_i \rangle + \frac{1}{2}\rho \left\| b - E_i z_i - \sum_{j \neq i} E_j \tilde{z}_j^{k,s} \right\|^2 \right\}. \tag{3.13}$$

The approximation of the augmented Lagrangian terms in (3.13) is quadratic with a diagonal Hessian.

Suppose that

$$A_i \tilde{x}_i^{k,s} = E_i \tilde{z}_i^{k,s}, \quad i = 1, 2, \ldots, L. \tag{3.14}$$

Then, in view of (3.12), we can substitute in (3.13) $A_i \tilde{x}_i^{k,s}$ and $A_i x_i^{k,s}$ for $E_i \tilde{z}_i^{k,s}$ and $E_i z_i^{k,s}$ and arrive at (3.2) for the original problem (1.1). With (3.14) Steps 1 and 2 of the DQA method are identical for both formulations. Finally, if we define $\tilde{z}_i^{k,0}$ such that $A_i \tilde{x}_i^{k,0} = E_i \tilde{z}_i^{k,0}$ then by (3.8) we shall have (3.14) for all s. Therefore, the DQA algorithm is invariant under the transformation (3.11).

We shall use transformation (3.11) to introduce some measures of sparsity that will be relevant for further analysis. For every $i = 1, 2, \ldots, L$ and $j = 1, 2, \ldots, m_i$ we define the set of *neighbors* of variable z_{ij}

$$V(i, j) = \{(k, l) : \ k \neq i, \ \langle E_{kl}, E_{ij} \rangle \neq 0\}, \tag{3.15}$$

where E_{ij} denotes the jth column of E_i. In terms of the original formulation (1.1), $V(i, j)$ is the set of (k, l) such that the l-th nonempty row of A_k has the same position as the j-th nonempty row of A_i.

With the number of elements in $V(i, j)$ denoted by $|V(i, j)|$, we define the maximum number of neighbors

$$N = \max_{i,j} |V(i, j)|. \tag{3.16}$$

By the definition of E_i, for every (i, j) and every $k \neq i$ there is at most one l such that $(k, l) \in V(i, j)$. Thus N is the maximum number of blocks linked by any single constraint, decremented by one. It is worth noting that this definition is invariant under the transformation (3.11).

In the next two sections we shall show that convergence properties of the DQA method heavily depend on the number of neighbors N.

4. Convergence

Let us define the function

$$\tilde{\Lambda}(x,\tilde{x},\pi) = \langle b,\pi\rangle + \sum_{i=1}^{L}\Lambda_i(x_i,\tilde{x},\pi) - \frac{1}{2}\rho(L-1)\left\|b-\sum_{i=1}^{L}A_i\tilde{x}_i\right\|^2.$$

Clearly, $\tilde{\Lambda}$ is the function that is (approximately) minimized in Step 1 of the DQA method.

We shall prove convergence of DQA by estimating the improvement in $\tilde{\Lambda}$ at each step and by bounding the difference between $\tilde{\Lambda}$ and the true augmented Lagrangian (2.7). In this way we shall be able to show that for appropriately chosen stepsizes each step of DQA improves the values of the augmented Lagrangian as well.

We start from estimating the difference between the augmented Lagrangian and its separable approximation $\tilde{\Lambda}$.

Lemma 1. *For all x, $\tilde{x}$ and π the following inequality holds*

$$|\Lambda(x,\pi) - \tilde{\Lambda}(x,\tilde{x},\pi)| \le \frac{1}{2}\rho N\sum_{i=1}^{L}\|A_i(x_i-\tilde{x}_i)\|^2. \tag{4.1}$$

Proof. By direct calculation we obtain

$$\begin{aligned}
&\Lambda(x,\pi) - \tilde{\Lambda}(x,\tilde{x},\pi)\\
&= \frac{1}{2}\rho\left\|b-\sum_{i=1}^{L}A_ix_i\right\|^2 + \frac{1}{2}\rho(L-1)\left\|b-\sum_{i=1}^{L}A_i\tilde{x}_i\right\|^2 - \frac{1}{2}\rho\sum_{i=1}^{L}\left\|b-A_ix_i-\sum_{k\neq i}A_k\tilde{x}_k\right\|^2\\
&= \frac{1}{2}\rho\left\|\sum_{i=1}^{L}A_ix_i\right\|^2 + \frac{1}{2}\rho(L-1)\left\|\sum_{i=1}^{L}A_i\tilde{x}_i\right\|^2 - \frac{1}{2}\rho\sum_{i=1}^{L}\left\|A_ix_i+\sum_{k\neq i}A_k\tilde{x}_k\right\|^2.
\end{aligned}$$

Expansion of the quadratic terms yields

$$\begin{aligned}
\Lambda(x,\pi) - \tilde{\Lambda}(x,\tilde{x},\pi) &= \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{k\neq i}\langle A_ix_i, A_kx_k\rangle + \frac{1}{2}\rho(L-1)\sum_{i=1}^{L}\sum_{k\neq i}\langle A_i\tilde{x}_i, A_k\tilde{x}_k\rangle\\
&\quad - \rho\sum_{i=1}^{L}\sum_{k\neq i}\langle A_ix_i, A_k\tilde{x}_k\rangle - \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{k\neq i}\sum_{s\neq k,i}\langle A_k\tilde{x}_k, A_s\tilde{x}_s\rangle.
\end{aligned}$$

Noting that in the last sum each product $\langle A_k\tilde{x}_k, A_s\tilde{x}_s\rangle$ appears exactly $L-2$ times we can rewrite the last equation as follows

$$\begin{aligned}
\Lambda(x,\pi) - \tilde{\Lambda}(x,\tilde{x},\pi) &= \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{k\neq i}\left(\langle A_ix_i, A_kx_k\rangle + \langle A_i\tilde{x}_i, A_k\tilde{x}_k\rangle - 2\langle A_ix_i, A_k\tilde{x}_k\rangle\right)\\
&= \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{k\neq i}\langle A_i(x_i-\tilde{x}_i), A_k(x_k-\tilde{x}_k)\rangle.
\end{aligned}$$

Using transformation (3.11) we obtain

$$\begin{aligned}\Lambda(x,\pi)-\tilde{\Lambda}(x,\tilde{x},\pi) &= \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{k\neq i}\langle E_i(z_i-\tilde{z}_i),E_k(z_k-\tilde{z}_k)\rangle \\ &= \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{j=1}^{m_i}\sum_{k\neq i}\sum_{l=1}^{m_k}\langle E_{ij}(z_{ij}-\tilde{z}_{ij}),E_{kl}(z_{kl}-\tilde{z}_{kl})\rangle.\end{aligned}$$

By (3.15),

$$\begin{aligned}|\Lambda(x,\pi)-\tilde{\Lambda}(x,\tilde{x},\pi)| &\leq \frac{1}{2}\rho\sum_{i=1}^{L}\sum_{j=1}^{m_i}\sum_{(k,l)\in V(i,j)}|\langle E_{ij}(z_{ij}-\tilde{z}_{ij}),E_{kl}(z_{kl}-\tilde{z}_{kl})\rangle| \\ &\leq \frac{1}{4}\rho\sum_{i=1}^{L}\sum_{j=1}^{m_i}\sum_{(k,l)\in V(i,j)}\left(\|E_{ij}(z_{ij}-\tilde{z}_{ij})\|^2+\|E_{kl}(z_{kl}-\tilde{z}_{kl})\|^2\right).\end{aligned}$$

Let us observe that each of the terms $\|E_{ij}(z_{ij}-\tilde{z}_{ij})\|^2$ appears in this sum at most $2N$ times, with N given by (3.16). Indeed, for fixed i and j there are at most N neighbors $(k,l)\in V(i,j)$; conversely, the pair (i,j) itself may be a neighbor of at most N other pairs. Therefore

$$|\Lambda(x,\pi)-\tilde{\Lambda}(x,\tilde{x},\pi)|\leq\frac{1}{2}\rho N\sum_{i=1}^{L}\sum_{j=1}^{m_i}\|E_{ij}(z_{ij}-\tilde{z}_{ij})\|^2.$$

Since the columns E_{ij}, $j=1,\ldots,m_i$ are orthogonal, the last inequality yields

$$|\Lambda(x,\pi)-\tilde{\Lambda}(x,\tilde{x},\pi)|\leq\frac{1}{2}\rho N\sum_{i=1}^{L}\|E_i(z_i-\tilde{z}_i)\|^2.$$

Putting $E_i(z_i-\tilde{z}_i)=A_i(x_i-\tilde{x}_i)$ we obtain (4.1). The proof is complete.

The progress in the minimization of Λ_i at Step 1 of the DQA method can be estimated as follows.

Lemma 2. *If* $x_i^{k,s}\in S_i^{\epsilon}(\tilde{x}^{k,s},\pi)$, $i=1,2,\ldots,L$ *then*

$$\Lambda_i(x_i^{k,s},\tilde{x}^{k,s},\pi)-\Lambda_i(\tilde{x}_i^{k,s},\tilde{x}^{k,s},\pi)\leq-\frac{1}{2}\rho(1-\mu)\|A_i(x_i^{k,s}-\tilde{x}_i^{k,s})\|^2.$$

Proof. For brevity we shall skip the superscripts k,s from $x_i^{k,s}$ and $\tilde{x}_i^{k,s}$, because they do not change here.

Directly from the definitions of Λ_i and $\partial_\epsilon f_i$,

$$\begin{aligned}&\Lambda_i(x_i,\tilde{x},\pi)-\Lambda_i(\tilde{x}_i,\tilde{x},\pi) \\ &= f_i(x_i)-f_i(\tilde{x}_i)-\langle A_i^T\pi,x_i-\tilde{x}_i\rangle-\rho\left\langle b-\sum_{j=1}^{L}A_j\tilde{x}_j,A_i(x_i-\tilde{x}_i)\right\rangle+\frac{1}{2}\rho\|A_i(x_i-\tilde{x}_i)\|^2 \\ &\leq \langle h_i,x_i-\tilde{x}_i\rangle+\epsilon_s-\langle A_i^T\pi,x_i-\tilde{x}_i\rangle-\rho\left\langle b-\sum_{j=1}^{L}A_j\tilde{x}_j,A_i(x_i-\tilde{x}_i)\right\rangle+\frac{1}{2}\rho\|A_i(x_i-\tilde{x}_i)\|^2,\end{aligned}$$

with $h_i \in \partial_\epsilon f_i(x_i)$. By (3.4) we can rewrite this inequality as

$$\Lambda_i(x_i, \tilde{x}, \pi) - \Lambda_i(\tilde{x}_i, \tilde{x}, \pi) \leq \langle g_i, x_i - \tilde{x}_i \rangle + \epsilon - \frac{1}{2}\rho \|A_i(x_i - \tilde{x}_i)\|^2, \tag{4.2}$$

with $g_i \in D_i^\epsilon(x_i, \tilde{x}, \pi)$. Since $\tilde{x}_i - x_i \in K_{X_i}(x_i)$, from (3.7) we get

$$\langle g_i, \tilde{x}_i - x_i \rangle \geq 0.$$

Using this inequality in (4.2), in view of (3.9), we obtain the required result.

We are now ready to prove convergence of the DQA method.

Theorem 1. *Assume that the sets $X_i, i = 1, 2, \ldots, L$ are bounded. If in the DQA method*

$$0 < \tau < \frac{1 - \mu}{N}, \tag{4.3}$$

where N is given by (3.16), then:

(a) for all $i = 1, 2, \ldots, L$ $\lim_{s\to\infty} A_i(x_i^{k,s} - \tilde{x}_i^{k,s}) = 0$;

(b) each accumulation point of the sequence $\{x^{k,s}\}_{s=0}^\infty$ is a solution of (2.9).

Proof. By Lemma 1,

$$\begin{aligned} \Lambda(\tilde{x}^{k,s} + \tau(x^{k,s} - \tilde{x}^{k,s}), \pi^k) - \tilde{\Lambda}(\tilde{x}^{k,s} + \tau(x^{k,s} - \tilde{x}^{k,s}), \tilde{x}^{k,s}, \pi^k) \\ \leq \frac{1}{2}\rho N \tau^2 \sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2. \end{aligned} \tag{4.4}$$

Next, by Lemma 2 and the convexity of $\tilde{\Lambda}(\cdot, \tilde{x}, \pi)$,

$$\begin{aligned} \tilde{\Lambda}(\tilde{x}^{k,s} + \tau(x^{k,s} - \tilde{x}^{k,s}), \tilde{x}^{k,s}, \pi^k) - \Lambda(\tilde{x}^{k,s}, \pi^k) \\ &= \tilde{\Lambda}(\tilde{x}^{k,s} + \tau(x^{k,s} - \tilde{x}^{k,s}), \tilde{x}^{k,s}, \pi^k) - \tilde{\Lambda}(\tilde{x}^{k,s}, \tilde{x}^{k,s}, \pi^k) \\ &\leq \tau\left(\tilde{\Lambda}(x^{k,s}, \tilde{x}^{k,s}, \pi^k) - \tilde{\Lambda}(\tilde{x}^{k,s}, \tilde{x}^{k,s}, \pi^k)\right) \\ &\leq -\frac{1}{2}\rho\tau(1 - \mu)\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2. \end{aligned}$$

Combining the last inequality with (4.4) we see that for $s = 0, 1, 2, \ldots$

$$\Lambda(\tilde{x}^{k,s+1}, \pi^k) - \Lambda(\tilde{x}^{k,s}, \pi^k) \leq -\frac{1}{2}\rho\tau(1 - \mu - \tau N)\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2. \tag{4.5}$$

Thus for τ satisfying (4.3) the sequence $\{\Lambda(\tilde{x}^{k,s}, \pi^k)\}_{s=0}^\infty$ is decreasing. By the boundedness of the sets X_i, $\Lambda(x, \pi^k)$ is bounded below for all $x \in X$. Therefore the sequence $\{\Lambda(\tilde{x}^{k,s}, \pi^k)\}_{s=0}^\infty$ is convergent. Since the left hand side of (4.5) converges to 0, so is the right one, which proves assertion (a).

Let x^* be a limit of a convergent subsequence $\{x^{k,s}\}_{s\in\mathcal{S}}$ of the sequence $\{x^{k,s}\}_{s=0}^{\infty}$. Let us consider the sequence $\{g_i^{k,s}\}_{s\in\mathcal{S}}$ such that (3.6) and (3.7) hold:

$$g_i^{k,s} \in \left(K_{X_i}(x_i^{k,s})\right)^* .$$

By the compactness of the sets X_i and by the boundedness and upper semicontinuity of the ϵ-subdifferential (cf. [9]) the sets $D_i^{\epsilon_s}(x_i^{k,s}, \tilde{x}^{k,s}, \pi^k)$ are uniformly bounded for all s. Therefore the sequence $\{g_i^{k,s}\}$ is bounded. Then for every accumulation point g_i^* of $\{g_i^{k,s}\}_{s\in\mathcal{S}}$ after passing to the limit in the last relation (over an appropriately chosen subsequence) we get

$$g_i^* \in (K_{X_i}(x_i^*))^* . \tag{4.6}$$

In the last relation we additionally used the upper semicontinuity of the conjugate cone $(K_{X_i}(x_i))^*$ with respect to x_i.

Next, by (3.9) and (a), $\epsilon_s \to 0$. Then (by upper semicontinuity of the ϵ-subdifferential) each accumulation point of the subsequence $h_i^s \in \partial_{\epsilon_s} f_i(x_i^{k,s})$, $s \in \mathcal{S}$, is an element of $\partial f_i(x_i^*)$. In view of (3.4),

$$g_i^* \in \partial\Lambda_i(x_i^*, \tilde{x}^*, \pi^k), \tag{4.7}$$

with some accumulation point $\tilde{x}^*$ of $\{\tilde{x}^{k,s}\}_{s\in\mathcal{S}}$. Since (a) implies $A_i x_i^* = A_i \tilde{x}_i^*$, then the subdifferential of the augmented Lagrangian (2.7) with respect to x_i equals

$$\partial_{x_i}\Lambda(x^*, \pi^k) = \partial\Lambda_i(x_i^*, \tilde{x}^*, \pi^k). \tag{4.8}$$

Combining (4.6), (4.7) and (4.8) we conclude that

$$g^* \in \partial\Lambda(x^*, \pi^k) \cap (K_X(x^*))^* ,$$

which implies optimality of x^* for problem (2.9). The proof is complete.

5. Speed of convergence

Our analysis of the speed of convergence will be based on inequality (4.5). Technically speaking, our aim is to relate the right side of (4.5) to some measure of the distance to the solution.

We start from the following relation between ϵ-subgradients of the augmented Lagrangian (2.7) and ϵ-subgradients of our approximation $\tilde{\Lambda}$.

Lemma 3. *For every $g = (g_1, g_2, \ldots, g_L)$ such that $g_i \in D_i^{\epsilon}(x_i, \tilde{x}, \pi)$ there is $w \in \partial_{\epsilon}\Lambda(x, \pi)$ such that for every $d = (d_1, d_2, \ldots, d_L)$ with $d_i \in R^{n_i}$*

$$|\langle w - g, d\rangle| \le \rho N \left(\sum_{i=1}^{L} \|A_i(x_i - \tilde{x}_i)\|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{L} \|A_i d_i\|^2\right)^{\frac{1}{2}} . \tag{5.1}$$

Proof. By (3.4) there exists $h_i \in \partial_\epsilon f_i(x_i)$ such that

$$\begin{aligned} g_i &= h_i - A_i^T\pi - \rho A_i^T\left(b - A_i x_i - \sum_{k\neq i} A_k\tilde{x}_k\right) \\ &= h_i - A_i^T\pi - \rho A_i^T\left(b - \sum_{k=1}^{L} A_k x_k\right) - \rho A_i^T\sum_{k\neq i} A_k(x_k - \tilde{x}_k). \end{aligned}$$

Defining

$$w_i = h_i - A_i^T\pi - \rho A_i^T\left(b - \sum_{k=1}^{L} A_k x_k\right) \in \partial_\epsilon\Lambda(x,\pi)$$

we get

$$w_i - g_i = \rho A_i^T\sum_{k\neq i} A_k(x_k - \tilde{x}_k).$$

Thus

$$\langle w - g, d\rangle = \rho\sum_{i=1}^{L}\left\langle A_i d_i, \sum_{k\neq i} A_k(x_k - \tilde{x}_k)\right\rangle.$$

By the definition of E_i we can find η_i and ξ_i such that

$$A_i(x_i - \tilde{x}_i) = E_i\eta_i, \tag{5.2}$$

$$A_i d_i = E_i\xi_i. \tag{5.3}$$

Then

$$\langle w - g, d\rangle = \rho\sum_{i=1}^{L}\left\langle E_i\xi_i, \sum_{k\neq i} E_k\eta_k\right\rangle = \rho\sum_{i=1}^{L}\sum_{j=1}^{m_i}\left\langle E_{ij}\xi_{ij}, \sum_{k\neq i}\sum_{l=1}^{m_k} E_{kl}\eta_{kl}\right\rangle.$$

Let us observe that

$$\langle E_{ij}, E_{kl}\rangle = \begin{cases} 1 & \text{if } (k,l) \in V(i,j) \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$\langle w - g, d\rangle = \rho\sum_{i=1}^{L}\sum_{j=1}^{m_i}\xi_{ij}\sum_{(k,l)\in V(i,j)}\eta_{kl} = \rho\langle\xi, r\rangle, \tag{5.4}$$

with

$$r_{ij} = \sum_{(k,l)\in V(i,j)}\eta_{kl}.$$

By the orthogonality of E_{ij}, $j = 1, 2\ldots, m_i$,

$$\|\xi\|^2 = \sum_{i=1}^{L}\sum_{j=1}^{m_i}\xi_{ij}^2 = \sum_{i=1}^{L}\left\|\sum_{j=1}^{m_i} E_{ij}\xi_{ij}\right\|^2 = \sum_{i=1}^{L}\|E_i\xi_i\|^2. \tag{5.5}$$

Next, we have

$$r_{ij}^2 = \left(\sum_{(k,l)\in V(i,j)}\eta_{kl}\right)^2 \leq N\sum_{(k,l)\in V(i,j)}\eta_{kl}^2.$$

Thus

$$\|r\|^2 = \sum_{i=1}^{L}\sum_{j=1}^{m_i} r_{ij}^2 \le N \sum_{i=1}^{L}\sum_{j=1}^{m_i} \sum_{(k,l)\in V(i,j)} \eta_{kl}^2.$$

Since $(k,l) \in V(i,j)$ iff $(i,j) \in V(k,l)$, each term η_{kl}^2 appears in this sum at most N times, so

$$\|r\|^2 \le N^2 \sum_{k=1}^{L}\sum_{l=1}^{m_k} \eta_{kl}^2.$$

By the orthogonality of E_{kl}, $l = 1, 2 \ldots, m_k$, similarly to (5.5), we can rewrite the last inequality as follows

$$\|r\|^2 \le N^2 \sum_{k=1}^{L} \|E_k \eta_k\|^2. \tag{5.6}$$

Putting together (5.4), (5.5) and (5.6) we obtain

$$|\langle w - g, d\rangle| \le \rho N \left(\sum_{i=1}^{L} \|E_i \xi_i\|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{L} \|E_i \eta_i\|^2\right)^{\frac{1}{2}}.$$

After subtitution of (5.2) and (5.3) we arrive to the required result. The proof is complete.

In our further analysis we shall denote by $\hat{X}(\pi)$ the set of solutions of (2.9). The next lemma provides us with the desired relation of the expression at the right side of (4.5) and the distance to the solution.

Lemma 4. *For every $x = (x_1, \ldots, x_L)$ such that $x_i \in S_i^{\epsilon}(\tilde{x}, \pi)$, with ϵ satisfying (3.9), we have*

$$\begin{aligned} \Lambda(x,\pi) - \hat{\Lambda}(\pi) \;\le\; & \rho N \min_{\hat{x}\in\hat{X}(\pi)} \left(\sum_{i=1}^{L} \|A_i(x_i - \hat{x}_i)\|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{L} \|A_i(x_i - \tilde{x}_i)\|^2\right)^{\frac{1}{2}} \\ & + \frac{1}{2}\rho\mu \sum_{i=1}^{L} \|A_i(x_i - \tilde{x}_i)\|^2, \end{aligned} \tag{5.7}$$

where

$$\hat{\Lambda}(\pi) = \min_{x\in X} \Lambda(x,\pi).$$

Proof. By the definition of the ϵ-subdifferential, for every $x \in X$, every $w \in \partial_{\epsilon}\Lambda(x,\pi)$ and every $\hat{x} \in \hat{X}(\pi)$

$$\Lambda(\hat{x},\pi) \ge \Lambda(x,\pi) + \langle w, \hat{x} - x\rangle - \epsilon. \tag{5.8}$$

We shall estimate $\langle w, \hat{x} - x\rangle$ for a selected w.

If $x_i \in S_i^{\epsilon}(\tilde{x},\pi)$ then we can find $g_i \in D_i^{\epsilon}(x_i, \tilde{x}, \pi)$ such that (3.7) holds. Therefore

$$\langle g_i, \hat{x}_i - x_i\rangle \ge 0.$$

By Lemma 3, there is $w \in \partial_{\epsilon}\Lambda(x,\pi)$ such that (5.1) is satisfied. Then the last inequality yields

$$\langle w_i, \hat{x}_i - x_i\rangle \ge \langle w_i - g_i, \hat{x}_i - x_i\rangle,$$

so

$$\langle w, \hat{x} - x \rangle \geq \langle w - g, \hat{x} - x \rangle \geq -\rho N \min_{\hat{x} \in \hat{X}(\pi)} \left(\sum_{i=1}^{L} \|A_i(x_i - \hat{x}_i)\|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{L} \|A_i(x_i - \tilde{x}_i)\|^2 \right)^{\frac{1}{2}}.$$

We can now go back to (5.8) to get (with the help of (3.9)) the required inequality (5.7). The proof is complete.

To estimate the speed of convergence we shall need the following assumption on the growth rate of the augmented Lagrangian function.

Quadratic Growth Condition. There is $\gamma > 0$ such that for every $x \in X$

$$\Lambda(x, \pi) - \hat{\Lambda}(\pi) \geq \gamma [\mathrm{dist}(x, \hat{X}(\pi))]^2. \tag{5.9}$$

It is clear that (5.9) is satisfied by linear and quadratic problems (1.1).

We are now ready to prove our main result on the speed of convergence.

Theorem 2. *Let the assumptions of Theorem 1 and the Quadratic Growth Condition (5.9) be satisfied. Then there exists $q \in (0,1)$ such that for all $s = 0, 1, 2, \ldots$ the following inequality holds*

$$\Lambda(\tilde{x}^{k,s+1}, \pi^k) - \hat{\Lambda}(\pi^k) \leq q(\Lambda(\tilde{x}^{k,s}, \pi^k) - \hat{\Lambda}(\pi^k)). \tag{5.10}$$

For $\mu = 0$

$$q = 1 - \frac{\tau(1 - \tau N)}{2\rho\alpha^2 N^2 \gamma^{-1} + 1}, \tag{5.11}$$

where $\alpha = \max_{1 \leq i \leq L} \|A_i\|$.

Proof. Let us denote

$$\tilde{\Delta}_s = \Lambda(\tilde{x}^{k,s}, \pi^k) - \hat{\Lambda}(\pi^k),$$
$$\Delta_s = \Lambda(x^{k,s}, \pi^k) - \hat{\Lambda}(\pi^k).$$

From (4.5) we obtain

$$0 \leq \tilde{\Delta}_{s+1} \leq \tilde{\Delta}_s - \frac{1}{2}\rho\tau(1 - \tau N) \sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2. \tag{5.12}$$

We shall estimate the last term in the above inequality. From Lemma 4 we get

$$\begin{aligned}
\Delta_s &\leq \rho N \min_{\hat{x} \in \hat{X}(\pi)} \left(\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \hat{x}_i^{k,s})\|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \right)^{\frac{1}{2}} \\
&\quad + \frac{1}{2}\rho\mu \sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \\
&\leq \rho N \alpha \mathrm{dist}(x^{k,s}, \hat{X}(\pi)) \left(\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \right)^{\frac{1}{2}} \\
&\quad + \frac{1}{2}\rho\mu \sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2.
\end{aligned}$$

By (5.9), $\operatorname{dist}(x^{k,s}, \hat{X}(\pi)) \leq \Delta_s^{\frac{1}{2}} \gamma^{-\frac{1}{2}}$, so

$$\frac{\Delta_s}{\rho} \leq C \left(\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \right)^{\frac{1}{2}} \Delta_s^{\frac{1}{2}} + \frac{1}{2}\mu \sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2, \tag{5.13}$$

where $C = \alpha N \gamma^{-\frac{1}{2}}$.

We shall consider two cases.

Case 1: $\mu = 0$.

From (5.13) we obtain

$$\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \geq \frac{\Delta_s}{\rho^2 C^2}. \tag{5.14}$$

Substituting the last estimate into (5.12) we get

$$\tilde{\Delta}_{s+1} \leq \tilde{\Delta}_s - \frac{1}{2\rho C^2} \tau(1 - \tau N)\Delta_s. \tag{5.15}$$

By convexity of Λ,

$$\tau \Delta_s \geq \tilde{\Delta}_{s+1} - (1 - \tau)\tilde{\Delta}_s.$$

Combining the last two inequalities and rearranging terms we obtain

$$\tilde{\Delta}_{s+1} \left(1 + \frac{1 - \tau N}{2\rho C^2} \right) \leq \tilde{\Delta}_s \left(1 + \frac{1 - \tau N}{2\rho C^2} \right) - \frac{\tau(1 - \tau N)}{2\rho C^2} \tilde{\Delta}_s.$$

This yields (for $\tau < 1/N$)

$$\tilde{\Delta}_{s+1} \leq \tilde{\Delta}_s - \frac{\tau(1 - \tau N)}{2\rho C^2 + 1} \tilde{\Delta}_s,$$

which completes the proof in this case.

Case 2: $\mu > 0$.

Multiplying both sides of (5.13) by μ and adding to them $C^2 \Delta_s / 2$ we obtain

$$\left(\frac{\mu}{\rho} + \frac{C^2}{2} \right) \Delta_s \leq \frac{1}{2} \left(C \Delta_s^{\frac{1}{2}} + \mu \left(\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \right)^{\frac{1}{2}} \right)^2.$$

This immediately yields an inequality similar to (5.14):

$$\sum_{i=1}^{L} \|A_i(x_i^{k,s} - \tilde{x}_i^{k,s})\|^2 \geq C_1 \Delta_s, \tag{5.16}$$

with

$$C_1 = \frac{C^2}{\mu^2}\left(\sqrt{1+\frac{2\mu}{\rho C^2}}-1\right)^2. \tag{5.17}$$

The rest of the analysis is similar to Case 1, only the constants differ slightly here.

While one could expect that the convergence rate of DQA (as an algorithm involving simple iterations (3.10)) can be at most linear, it is interesting to analyse the constants that appear in (4.5), (5.15) and (5.11).

Let us at first note that a small $\mu > 0$ does not significantly change convergence properties of the method, because using in (5.17) the approximation $\sqrt{1+2\mu/\rho C^2} \approx 1+\mu/\rho C^2$ we get $C_1 \approx 1/\rho^2 C^2$ and (5.16) becomes very close to (5.14). It is therefore sufficient (and much easier) to look at the case of $\mu = 0$.

We see that in order to make the estimates of the one-step decrease (4.5) and (5.15) as large as possible, the stepsize τ should maximize the expression

$$\phi(\tau) = \tau(1-\tau N),$$

which yields

$$\hat{\tau} = \frac{1}{2N}. \tag{5.18}$$

So, the recommended stepsize does not depend on the penalty parameter ρ, but it depends on the number of neighbors N.

The estimated covergence ratio q in (5.10) exhibits even stronger dependence on N. After substituting (5.18) into (5.11) we obtain (for $\rho\alpha^2 N^2/\gamma \gg 1$)

$$q \approx 1 - \frac{\gamma}{8\rho\alpha^2 N^3}. \tag{5.19}$$

While there is no surprise in the negative influence of the penalty parameter ρ on the speed of convergence, the dependence of q on the number of neighbors N is astonishing. In direct approaches to large scale linear programming (like the simplex method or interior point methods) sparsity of the problem influences the cost of one iteration. Here it improves the rate of convergence of the decomposition method. It is clear that we can profit a lot from having very loosely linked blocks.

If (1.1a) is a polyhedral function and the sets X_i are polyhedral, we can further sharpen the above estimates, by noting that in the neighborhood of $\hat{X}(\pi)$ we have $\gamma = \rho\beta$ with some $\beta > 0$ independent of ρ (β may be taken equal to the value of γ for $\rho = 1$). Then (5.19) reads

$$q \approx 1 - \frac{\beta}{8\alpha^2 N^3},$$

and the asymptotic speed of convergence is fully determined by the properties of the original problem, not by the penalty parameter.

Obviously, we have derived here only upper bounds on q, so we should not conclude that large N must result in slow convergence. It is, however, possible, as the following example shows.

Example 1.

Consider the problem

$$\min \frac{1}{2}\sum_{i=1}^{L}(x_i)^2$$

$$\sum_{i=1}^{L} x_i = 1,$$

$$0 \le x_i \le 1,\ i = 1, 2, \ldots, L,$$

and the associated augmented Lagrangian

$$\Lambda(x, \pi) = \frac{1}{2}\sum_{i=1}^{L}(x_i)^2 + \pi(1 - \sum_{i=1}^{L} x_i) + \frac{1}{2}\rho(1 - \sum_{i=1}^{L} x_i)^2.$$

Clearly, the number of neighbors N equals $L - 1$ here.

Suppose that $\pi = 0$. The separable approximations take on the form

$$\Lambda_i(x_i, \tilde{x}, \pi) = \frac{1}{2}(x_i)^2 + \frac{1}{2}\rho(1 - x_i - \sum_{j \ne i} \tilde{x}_j)^2.$$

By straightforward calculation we obtain the solution of (3.2):

$$x_i^{k,s} = \frac{\rho}{\rho + 1}(1 - \sum_{j \ne i} \tilde{x}_j^{k,s}).$$

Thus (3.10) reads

$$\tilde{x}_i^{k,s+1} = (1 - \tau)\tilde{x}_i^{k,s} - \frac{\tau\rho}{\rho + 1}\sum_{j \ne i} \tilde{x}_j^{k,s} + \frac{\tau\rho}{\rho + 1}.$$

It follows that convergence properties of DQA are determined by the spectrum of the matrix

$$\begin{bmatrix} 1 - \tau & -\frac{\tau\rho}{\rho+1} & \cdots & -\frac{\tau\rho}{\rho+1} \\ -\frac{\tau\rho}{\rho+1} & 1 - \tau & \cdots & -\frac{\tau\rho}{\rho+1} \\ & & \ddots & \\ -\frac{\tau\rho}{\rho+1} & -\frac{\tau\rho}{\rho+1} & \cdots & 1 - \tau \end{bmatrix}.$$

For even L the spectrum is contained between

$$\lambda_{\max} = 1 - \tau + \frac{\tau\rho}{\rho + 1}$$

(corresponding to the vector $[1\ \ -1\ 1\ \ -1\ \cdots\ ]^T$) and

$$\lambda_{\min} = 1 - \tau - \frac{\tau\rho(L - 1)}{\rho + 1}$$

(corresponding to the vector $[1\ 1\ \cdots\ 1]^T$). It is obvious that we must have

$$0 < \tau < \frac{2(\rho+1)}{1+\rho L}$$

which makes

$$1 - \frac{2}{1+\rho L} < \lambda_{\max} < 1.$$

For large L the DQA method becomes very slow.

Summing up, although our estimates are not sharp, they show that DQA can substantially profit from sparsity of the linking constraints.

6. Applications

We start from two straightforward applications to general linear programming problems.

Example 2: *Decomposition of decisions in linear programming*

Consider a linear program in the standard form

$$\begin{aligned} &\min\ c^T x \\ &Ax = b, \\ &x \geq 0. \end{aligned} \tag{6.1}$$

It is already in form (1.1), with x_i denoting the i-th component of x and with the augmented Lagrangian

$$\Lambda(x,\pi) = c^T x + \langle \pi, b - Ax\rangle + \frac{1}{2}\rho\|b - Ax\|^2.$$

Subproblems (3.2) are then simple one-dimensional minimization problems:

$$\min_{x_j \geq 0}\left\{\bar{c}_j x_j + \frac{1}{2}\rho\|\bar{b} - A_j(x_j - \tilde{x}_j)\|^2\right\},\quad j = 1,2,\ldots,n,$$

where A_j is the j-th column of A and

$$\bar{b} = b - A\tilde{x},$$

$$\bar{c} = c - A^T\pi.$$

They have a closed-form solution

$$x_j = \left[\tilde{x}_j + \frac{1}{\|A_j\|^2}\left(\langle A_j, \bar{b}\rangle - \frac{\bar{c}_j}{\rho}\right)\right]_+,\quad j = 1,2\ldots,n, \tag{6.2}$$

which can be substituted for Step 1 of the DQA method. As a result, we obtain a very simple iterative procedure with alternating steps made by (6.2) and (3.10), well suited to massively parallel computing systems.

It follows from Theorem 1 that the stepsize τ in (3.10) can be chosen from the interval

$$0 < \tau < \frac{1}{l_r(A) - 1}, \tag{6.3}$$

where $l_r(A)$ is the maximum number of nonzeros in a row of A. Best estimates of the speed of convergence can be obtained for τ in the middle of the interval.

The first example has been included mainly for illustrative purposes; for the approach to be efficient we need to have the number of nonzeros in rows to be bounded by small number. But applying the same trick to the dual problem yields a more interesting method.

Example 3: *Decomposition of constraints in linear programming*

Let us consider the dual to (6.1):

$$\begin{aligned} &\max\ b^T\pi \\ &A^T\pi + \mu = c, \\ &\mu \geq 0. \end{aligned} \tag{6.4}$$

Clearly, the optimal Lagrange multipliers x associated with (6.4) solve the primal problem (6.1). The augmented Lagrangian has the form:

$$\Lambda(\pi, \mu, x) = b^T\pi + x^T(c - A^T\pi - \mu) - \frac{1}{2}\rho\|c - A^T\pi - \mu\|^2.$$

Denoting the approximation point for (π, μ) by $(\tilde{\pi}, \tilde{\mu})$ we obtain for $i = 1, 2, \ldots, m$ the subproblems

$$\max_{\pi_i} \left\{ \bar{b}_i\pi_i - \frac{1}{2}\rho\|\bar{c} - a_i(\pi_i - \tilde{\pi}_i)\|^2 \right\}, \tag{6.5a}$$

where a_i is the i-th row of A and

$$\bar{b} = b - Ax,$$

$$\bar{c} = c - A^T\tilde{\pi} - \tilde{\mu}.$$

For the dual slacks μ_j, $j = 1, 2 \ldots, n$, the subproblems are simpler

$$\max_{\mu_j \geq 0} \left\{ -x_j\mu_j - \frac{1}{2}\|\bar{c} - e_j(\mu_j - \tilde{\mu}_j)\|^2 \right\}, \tag{6.5b}$$

with e_j denoting the j-th unit vector in R^n. Again, subproblems (6.5) have closed-form solutions

$$\pi_i = \tilde{\pi}_i + \frac{1}{\|a_i\|^2}\left(\langle a_i, \bar{c} \rangle - \frac{\bar{b}_i}{\rho} \right), \quad i = 1, 2, \ldots, m,$$

$$\mu_j = \left[\tilde{\mu}_j + \bar{c}_j + \frac{x_j}{\rho}\right]_+, \quad j = 1, 2, \ldots, n.$$

This can be substituted for Step 1 of DQA, with Step 2 of the form

$$\tilde{\pi}^{k,s+1} = \tilde{\pi}^{k,s} + \tau(\pi^{k,s} - \tilde{\pi}^{k,s}), \tag{6.6a}$$

$$\tilde{\mu}^{k,s+1} = \tilde{\mu}^{k,s} + \tau(\mu^{k,s} - \tilde{\mu}^{k,s}). \tag{6.6b}$$

Again, only $\bar{c}$ changes from iteration to iteration, so implementation of such an iterative procedure on massively parallel computing systems can be quite efficient.

By Theorem 1, the stepsize τ in (6.6) can be chosen from the interval

$$0 < \tau < \frac{1}{l_c(A)},$$

where $l_c(A)$ is the maximum number of nonzeros in a column of A. To obtain best estimates of the speed of convergence we should choose τ in the middle of this interval.

In Example 3 we need to have the number of nonzeros in columns to be bounded by a small number which is far more practical than a bound for row lengths. There are many problems with short columns, such as generalized networks, dynamic inventory-type problems, etc.

Our third example is more serious; it describes an application of the method to stochastic programming, which proved successful for very large problems with hundreds of thousands of variables [13].

Example 4: *Scenario decomposition in stochastic programming*

In a multistage stochastic programming problem (see, e. g., [16, 17]) each x_i represents a sequence of decisions at time stages $t = 1, 2, \ldots, T$:

$$x_i = (x_i(1), x_i(2), \ldots, x_i(T)),$$

that has to be made in scenario $i = 1, 2, \ldots, L$. The sets X_i, $i = 1, 2 \ldots, L$, are given by scenario-dependent constraints that describe the evolution of the system. In the simplest case it may be

$$D_i(t)x_i(t-1) + H_i(t)x_i(t) = b_i(t), \quad t = 1, 2, \ldots, T,$$

$$x_i(t) \geq 0, \quad t = 1, 2, \ldots, T.$$

However, the scenario subproblems cannot be solved independently, because at time t only the scenario data

$$s_i(\theta) = (D_i(\theta), H_i(\theta), b_i(\theta)), \quad \theta = 1, 2, \ldots, t,$$

are known. Therefore we have to impose on the sequences $x_i(t)$, $t = 1, 2, \ldots, T$ an additional *nonanticipativity constraint*: if for some t scenarios i and j have common past and present, i.e.

$$s_i(\theta) = s_j(\theta), \quad \theta = 1, 2, \ldots, t,$$

then we must have

$$x_i(t) = x_j(t).$$

The set of all scenarios j that coincide with scenario i up to time t is denoted $\mathcal{A}(i,t)$. The multistage stochastic programming problem can be stated as follows:

$$\min \sum_{i=1}^{L} p_i \langle c_i, x_i \rangle \tag{6.7a}$$

$$x_i(t) = x_j(t) \quad \text{for all } j \in \mathcal{A}(i,t), \quad i = 1,2,\ldots,L. \tag{6.7b}$$

$$x_i \in X_i, \quad i = 1,2,\ldots,L; \tag{6.7c}$$

p_i denotes here the probability of scenario i. The problem has form (1.1) with very many linking constraints. Clearly, many of the constraints (6.7b) are redundant and we can work with a carefully selected subset of them. In [13] the following approach has been applied. If at time stage t scenarios $i_1, i_2, \ldots, i_k$ form a group with common past data, then we can enforce nonanticipativity by the following constraints:

$$\begin{aligned} x_{i_1}(t) &= x_{i_2}(t), \\ x_{i_2}(t) &= x_{i_3}(t), \\ &\vdots \\ x_{i_{k-1}}(t) &= x_{i_k}(t). \end{aligned} \tag{6.8}$$

The number of linking constraints is still large, but the maximum number of neighbors N is only 1 (each constraint links variables from two blocks). Therefore the stepsize in (3.10) can be chosen from

$$0 < \tau < 1 - \mu \tag{6.9}$$

and the best speed of convergence can be quite high.

Thus, our analysis allowed us to improve the results of [13], where $0 < \tau < \frac{1}{2}$ was required (with $\mu = 0$). It also explains good behavior of the method with $\tau = \frac{1}{2}$ observed for large-scale stochastic programming problems.

Our examples make it easy to identify classes of problems to which our techniques apply directly; we can also indicate problems for which DQA may be slow. One of them is the multicommodity network flow problem with many commodities, because its linking constraints relate decisions (flows) from all blocks. Then the number of neighbors N is large, we have to use very small stepsizes and convergence is slow.

References

[1] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, (Academic Press,1982).

[2] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation* (Prentice-Hall, Englewood Cliffs, 1989).

[3] G Cohen and D.L. Zhu, "Decomposition-coordination methods in large scale optimization problems: the nondifferentiable case and the use of augmented Lagrangians," in: *Advances in Large Scale Systems, vol. 1*, J. B. Cruz (ed.), JAI Press 1984, pp. 203-266.

[4] G.B. Dantzig and P. Wolfe, "Decomposition principle for linear programs", *Operations Research* 8(1960) 101–111.

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

[6] M. Fortin and R. Glowinski, "On decomposition-coordination methods using an augmented Lagrangian," in: *Augmented Lagrangian Methods: Applications to the Numerical Solution of Bocudery-Value Problems*, M. Fortin and R. Glowinski (eds.), North-Holland, Amsterdam, 1983, pp. 97-146.

[7] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comput. and Math. Appl.* 2(1976), pp. 17-40.

[8] R. Glowinski and A. Marocco, "Sur l'approximation par éléments finis d'ordre un et la résolution par pénalisation dualité d'une classe de problèmes de Dirichlet non linéaires," *Revue Française d'Automatique Informatique Recherche Opérationelle, Analyse Numérique*, R-2(1975), pp. 41-76.

[9] J.-B. Hiriart-Urruty, *ϵ-Subdifferential Calculus*, in: *Convex Analysis and Optimization, Research Notes in Mathematics* 57, PITMAN Publishers, 1982, pp. 1-44.

[10] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization* (Springer-Verlag, Berlin, 1985).

[11] L.S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York, 1970.

[12] J.M. Mulvey and A. Ruszczyński, "A diagonal quadratic approximation method for large scale linear programs," technical report SOR 90-08, Department of Civil Engineering and Operations Research, Princeton University, Princeton 1990 (to appear in *Operation Research Letters*).

[13] J.M. Mulvey and A. Ruszczyński, "A new scenario decomposition method for large-scale stochastic optimization," technical report SOR 91-19, Department of Civil Engineering and Operations Research, Princeton University, Princeton 1991.

[14] B.N. Pshenichnyi, *Convex Analysis and Extremal Problems*, Nauka, Moskva, 1980 (in Russian).

[15] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton 1973.

[16] R.T. Rockafellar and R.J.-B. Wets, "Scenarios and policy aggregation in optimization under uncertainty," *Mathematics of Operations Research* **16**(1991) 1-23.

[17] A. Ruszczyński, "Parallel decomposition of multistage stochastic programs", *Mathematical Programming* 1992 (forthcoming).

[18] A. Ruszczyński, "An augmented Lagrangian decomposition method for block diagonal linear programming problems", *Operations Research Letters* 8(1989) 287-294.

[19] G. Stephanopoulos and W. Westerberg, "The use of Hestenes' method of multipliers to resolve dual gaps in engineering system optimization", *Journal of Optimization Theory and Applications*, 15(1975), pp. 285-309.

[20] P. Tatjewski, "New dual-type decomposition algorithm for nonconvex separable optimization problems", *Automatica*, 25(1989), pp. 233-242.

[21] N. Watanabe, Y. Nishimura and M. Matsubara, "Decomposition in large system optimization using the method of multipliers," *Journal of Optimization Theory and Applications*, 25(1978), pp. 181-193.