SIAM J. OPTIM.
Vol. 8, No. 1, pp. 187–196, February 1998

009

ON THE CONVERGENCE OF CONSTRAINED PARALLEL VARIABLE DISTRIBUTION ALGORITHMS*

MICHAEL V. SOLODOV†

Abstract. We consider the parallel variable distribution (PVD) approach proposed by Ferris and Mangasarian [*SIAM J. Optim.*, 4 (1994), pp. 815–832] for solving optimization problems. The problem variables are distributed among p processors with each processor having the primary responsibility for updating its block of variables while allowing the remaining "secondary" variables to change in a restricted fashion along some easily computable directions. For constrained nonlinear programs, convergence in [M. C. Ferris and O. L. Mangasarian, *SIAM J. Optim.*, 4 (1994), pp. 815–832] was established in the special case of convex block-separable constraints. For general (inseparable) constraints, it was suggested that a dual differentiable exact penalty function reformulation of the problem be used. We propose to apply the PVD approach to problems with general convex constraints directly and show that the algorithm converges, provided certain conditions are imposed on the change of secondary variables. These conditions are both natural and practically implementable. We also show that the original requirement of exact global solution of the parallel subproblems can be replaced by a less stringent sufficient descent condition. The first rate of convergence result for the class of constrained PVD algorithms is also given.

Key words. parallel optimization, nonlinear programming, linear convergence

AMS subject classifications. 90C30, 49D27

PII. S1052623495293949

1. Introduction.

We consider the general nonlinear programming problem

$$\min_{x \in C} f(x), \tag{1.1}$$

where C is a nonempty closed convex set in $\Re^n$ and $f : \Re^n \to \Re$ is a continuously differentiable function. We first describe the PVD algorithm proposed in [4] and, in the unconstrained case, further studied and extended in [25, 7]. Let the problem variables $x \in \Re^n$ be partitioned into p blocks $x_1, \ldots, x_p$ such that $x_l \in \Re^{n_l}$, $\sum_{l=1}^{p} n_l = n$. These blocks of variables are then distributed among p parallel processors. Each processor has the primary responsibility for updating its block of variables by solving the parallelization problem (see Algorithm 1.1 below). The remaining "secondary" variables are allowed to change in a restricted fashion along some easily computable directions. The distinctive novel feature of this algorithm is the presence of the "forget-me-not" term $x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}$ in the parallel subproblems (1.2) which allows for a change in "secondary" variables. This makes PVD fundamentally different from the block Jacobi [2], coordinate descent [26], and parallel gradient distribution algorithms [16]. The forget-me-not approach improves robustness and accelerates convergence of the algorithm and is the key to its success. The parallelization phase in Algorithm 1.1 is followed by a simple synchronization step which selects a point with the objective function value at least as good as the smallest among all the new points computed by the parallel processors.

We now formally state the algorithm.

*Received by the editors October 27, 1995; accepted for publication (in revised form) November 13, 1996. This material is based on research supported by CNPq grant 300734/95-6.
http://www.siam.org/journals/siopt/8-1/29394.html
†Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br).

ALGORITHM 1.1 (PVD). *Start with any $x^0 \in C$. Having x^i, check a stopping criterion (to be specified later). If it is not satisfied, compute x^{i+1} as follows:*

• **Parallelization:** *For each processor $l \in \{1, \ldots, p\}$ compute*

$$(y_l^i, \mu_{\bar{l}}^i) \in \arg\min_{x_l, \mu_{\bar{l}}} \psi_l^i(x_l, \mu_{\bar{l}}) := f(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}) \quad \text{subject to } (x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}) \in C. \tag{1.2}$$

• **Synchronization:** *Compute x^{i+1} such that*

$$f(x^{i+1}) \le \min_{l \in \{1,\ldots,p\}} \psi_l^i(y_l^i, \mu_{\bar{l}}^i). \tag{1.3}$$

In the above algorithm $\bar{l}$ denotes the complement of l in the set $\{1, \ldots, p\}$ and $\mu_{\bar{l}} \in \Re^{p-1}$. The matrix $D_{\bar{l}}^i$ is an $n_{\bar{l}} \times (p-1)$ block-diagonal matrix formed by placing the blocks $d_1^i, \ldots, d_{p-1}^i$ ($d_t^i \in \Re^{n_t}$, $t = 1, \ldots, p-1$) of an arbitrary direction $d^i \in \Re^n$ (e.g., quasi-Newton or steepest descent directions in the case $C = \Re^n$) along its block diagonal as follows:

$$D_{\bar{l}}^i := \begin{pmatrix} d_1^i & & & & & & \\ & d_2^i & & & & & \\ & & \ddots & & & & \\ & & & d_{l-1}^i & & & \\ & & & & d_{l+1}^i & & \\ & & & & & \ddots & \\ & & & & & & d_p^i \end{pmatrix}.$$

The synchronization step in Algorithm 1.1 may consist of minimizing the objective function in the affine hull, subject to feasibility, of all the points computed in parallel by the p processors. In principle, any point with the objective function value at least as good as the smallest computed by all the processors is acceptable. We refer the reader to [4] for a study of computational efficiency of the PVD approach and other practical issues.

In [4] it was shown that in the special case when the feasible set C has block-separable structure (i.e., C is a Cartesian product of closed convex sets), every accumulation point of the PVD iterates satisfies the minimum principle necessary optimality condition [15]. It was further stated that in the case of inseparable constraints, the PVD approach may fail. The following example was employed to illustrate this conclusion.

$$\min_{x_1, x_2} f(x) := x_1^2 + x_2^2 \quad \text{subject to} \quad 0 \le g(x) := x_1 + x_2 - 2.$$

This strongly convex program has a unique global solution at $x_1 = x_2 = 1$. However, the point $x_1 = 0.5$, $x_2 = 1.5$ is a global minimum with respect to both x_1 and x_2, separately, subject to the problem constraint

$$0.5 = \arg\min_{x_1} \{x_1^2 + 2.25 \mid x_1 \ge 0.5\},$$
$$1.5 = \arg\min_{x_2} \{0.25 + x_2^2 \mid x_2 \ge 1.5\}.$$

Therefore, if we use the above point as x^0 in Algorithm 1.1 and we set $D_1^0 = D_2^0 = 0$, then the algorithm will fail to make any progress towards solving the original problem. Moreover, every point in the set $\{x \mid g(x) = 0, x \geq 0\}$ has the same property, while only one of them is the true solution of the original problem. It was therefore argued that having a stationary point that results from minimizing the objective function with respect to individual blocks of variables and subject to the problem constraints does not result in a useful point unless the constraints are separable [4]. While this is true for an arbitrary choice of directions in D_l^i (note that in the example above, zero directions were used), it seems clear that one of the keys to success of any PVD algorithm should be precisely a reasonable choice of directions for change in the secondary variables. In fact, the presence of forget-me-not terms is what separates PVD from other related parallel techniques [2]. It is therefore natural to choose the directions in a special way and study their effect on the properties of the algorithm. In [25], by imposing natural restrictions on these directions, we were able to strengthen convergence results for unconstrained PVD. A reasonable choice of directions also allowed us to propose useful generalizations, such as algorithms with inexact subproblem solution and a certain degree of asynchronization [25].

We now define the following *projected gradient* residual function

$$r(x) := x - P_C[x - \nabla f(x)], \tag{1.4}$$

where $P_C[\cdot]$ stands for the orthogonal projection map onto the closed convex set C; i.e., $P_C[y] := \arg\min_{x \in C} \|x - y\|$. It is well known that some $x \in \Re^n$ satisfies the minimum principle necessary optimality condition [15] for problem (1.1)

$$x \in C, \quad \langle \nabla f(x), y - x \rangle \geq 0 \quad \forall\, y \in C$$

if and only if $r(x) = 0$. We shall call such x a stationary point of (1.1). Our stopping criterion for Algorithm 1.1 will consist of checking whether at the current iterate x^i the minimum principle optimality condition is satisfied; i.e., $r(x^i) = 0$.

In this paper, we propose to employ the residual (projected gradient) function (1.4) for the change of secondary variables in problems with general convex constraints. In particular, we set $d^i := r(x^i)$. Then the matrices D_l^i are comprised of the portions of the projected gradient direction, that is

$$d_l^i = r_l(x^i) := [r(x^i)]_l, \quad l = 1, \ldots, p, \tag{1.5}$$

where $[r(x^i)]_l \in \Re^{n_l}$ is the portion of $r(x^i)$ corresponding to $x_l \in \Re^{n_l}$. Let us go back to the example above. As before, consider the point $x_1^0 = 0.5,\ x_2^0 = 1.5$. It is easy to check that $r_1(x^0) = -1$ and $r_2(x^0) = 1$. Thus the two parallel subproblems in the PVD algorithm are

$$\begin{aligned} &\min_{x_1, \mu_2} f(x_1, x_2^0 + \mu_2 r_2(x^0)) = (x_1)^2 + (\mu_2)^2 + 2x_2^0\mu_2 + (x_2^0)^2 \\ &\text{subject to} \quad 0 \leq g(x_1, x_2^0 + \mu_2 r_2(x^0)) = x_1 + \mu_2 - 0.5 \end{aligned}$$

and

$$\begin{aligned} &\min_{x_2, \mu_1} f(x_1^0 + \mu_1 r_1(x^0), x_2) = -2x_1^0\mu_1 + (\mu_1)^2 + (x_2)^2 + (x_1^0)^2 \\ &\text{subject to} \quad 0 \leq g(x_1^0 + \mu_1 r_1(x^0), x_2) = x_2 - \mu_1 - 1.5. \end{aligned}$$

The solution of the first subproblem is $y_1^0 = 1,\ \mu_2^0 = -0.5$, and the solution of the second is $y_2^0 = 1,\ \mu_1^0 = -0.5$. It is easy to see that $x_1^0 + \mu_1^0 r_1(x^0) = 1$ and $x_2^0 + \mu_2^0 r_2(x^0) = 1$. So both subproblems produce the solution of the original problem.

In section 2 we show that, for the choice of directions given by (1.5), the PVD algorithm converges under no special assumptions on the structure of the feasible set (see Theorem 2.1). Moreover, we prove convergence for the case of Hölder continuous gradient of the objective function (Lipschitz continuity was assumed in [4] and [25]). We are further able to give the first rate of convergence result for the constrained PVD methods (Theorem 2.2). Finally, in section 2 we also show that the exact global solution requirement for parallel subproblems (1.2) can be replaced by a less stringent condition of sufficient descent (see (2.3)). This yields a more practical approach and makes it easier to achieve good load balancing among the parallel processors. Some concluding remarks are given in section 3.

We briefly describe our notation now. The usual inner product of two vectors $x \in \Re^n$, $y \in \Re^n$ is denoted by $\langle x, y\rangle$. The Euclidean 2-norm of $x \in \Re^n$ is given by $\|x\|^2 = \langle x, x\rangle$. For a nonempty (closed) set $S \subset \Re^n$, $d(\cdot, S)$ denotes the Euclidean distance to the set S. For a differentiable function $f : \Re^n \to \Re$, ∇f will denote the n-dimensional vector of partial derivatives with respect to x. If a function $f(\cdot)$ has Hölder continuous partial derivatives on $\Re^n$ with some constant $L > 0$ and modulus $\alpha \in (0, 1]$, that is

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|^\alpha \ \ \forall\, x, y \in \Re^n,$$

we write $f(\cdot) \in C_L^{1,\alpha}(\Re^n)$. By R-linear convergence and Q-linear convergence, we mean linear convergence in the root sense and in the quotient sense, respectively, as defined in [19].

2. Convergence of constrained PVD algorithms. In this section we show that, in the general convex-constrained case, for the PVD approach to be effective it is crucial to impose certain conditions on the change of secondary variables. Fortunately, those conditions are natural and easily implementable in practice. The importance of change in the secondary variables was mentioned and experimentally established in [4]. However, convergence analysis in [4] fails to take advantage of forget-me-not terms. This inevitably weakens the properties of PVD algorithms, especially in the constrained case. In fact, it was suggested that the only sensible way to distribute variables for problems with inseparable constraints is to convert them to unconstrained problems. This can be done in several ways, but, unfortunately, each has some disadvantages. The exterior penalty approach [5] requires the unboundedness of the penalty parameter, while the augmented Lagrangian methods [24, 1] change the minimization problem into a saddle-point problem which is, in general, more difficult. In [4] it was suggested to use a certain dual differentiable exact penalty function formulation [8] for which the penalty parameter remains finite. This is a reasonable approach, but it also has some drawbacks. In particular, by using this dual penalty function formulation we increase the dimensionality of the problem and may lose any special structure if it was present in the problem. We propose an alternative approach in which we deal directly with the original problem (1.1).

Theorem 2.1 below shows that PVD methods can be successfully applied to *any* optimization problem with general convex constraints. Furthermore, considering specific directions and making explicit use of their properties in our analysis, we are able to obtain stronger convergence results, including rate of convergence results. This

was also the case for the analysis of unconstrained PVD algorithms [25]. Another improvement consists of extending the analysis to functions in the class $C_L^{1,\alpha}(\Re^n)$ where $\alpha \in (0,1]$ (in [4, 25] the case of $\alpha = 1$ is considered).

We are now ready to prove our main results.

THEOREM 2.1. *Let $f(\cdot) \in C_L^{1,\alpha}(\Re^n)$. Suppose $\{x^i\}$ is any sequence generated by PVD Algorithm* 1.1 *with the choice of directions for secondary variables given by* (1.5). *Then either $f(\cdot)$ is unbounded from below on C or the sequence $\{f(x^i)\}$ converges, the sequence $\{r(x^i)\}$ converges to zero and every accumulation point of the sequence $\{x^i\}$ satisfies the minimum principle necessary optimality condition.*

Proof. Take any $\beta \in (0,1)$. We first show that for any iteration $i = 0, 1, \ldots$ and any processor $l = 1, \ldots, p$, the point

$$(x_l^i - \eta_i r_l(x^i), -\eta_i e_{\bar{l}}) \in \Re^{n_l+p-1}$$

is feasible for the corresponding subproblem (1.2) of minimizing the function $\psi_l^i(x_l, \mu_{\bar{l}})$. In the above, $e_{\bar{l}}$ is a vector of ones of appropriate dimension, and

$$\eta_i := \min\left\{1,\ \beta\left(\frac{1+\alpha}{L}\right)^{1/\alpha}\|r(x^i)\|^{(1-\alpha)/\alpha}\right\}.$$

Indeed,

$$\begin{aligned}(x_l^i - \eta_i r_l(x^i), x_{\bar{l}}^i - \eta_i D_{\bar{l}}^i e_{\bar{l}}) &= (x_l^i - \eta_i r_l(x^i), x_{\bar{l}}^i - \eta_i r_{\bar{l}}(x^i))\\ &= x^i - \eta_i r(x^i)\\ &= (1-\eta_i)x^i + \eta_i P_C[x^i - \nabla f(x^i)] \in C,\end{aligned}$$

where the first equality follows from the block diagonal structure of $D_{\bar{l}}^i$ and (1.5), the last equality follows from (1.4), and the inclusion is by convexity of the set C. We further obtain

$$\begin{aligned}f(x^i) - f(y_l^i, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}^i) &= \psi_l^i(x_l^i, 0) - \psi_l^i(y_l^i, \mu_{\bar{l}}^i)\\ &\geq \psi_l^i(x_l^i, 0) - \psi_l^i(x_l^i - \eta_i r_l(x^i), -\eta_i e_{\bar{l}})\\ &= f(x^i) - f(x_l^i - \eta_i r_l(x^i), x_{\bar{l}}^i - \eta_i D_{\bar{l}}^i e_{\bar{l}})\\ &= f(x^i) - f(x^i - \eta_i r(x^i)),\end{aligned}$$

where the inequality follows from (1.2). Furthermore,

$$\begin{aligned}f(x^i - \eta_i r(x^i)) - f(x^i) &= -\int_0^{\eta_i} \langle \nabla f(x^i - t r(x^i)), r(x^i)\rangle dt\\ &= -\eta_i \langle \nabla f(x^i), r(x^i)\rangle\\ &\quad - \int_0^{\eta_i} \langle \nabla f(x^i - t r(x^i)) - \nabla f(x^i), r(x^i)\rangle dt\\ &\leq -\eta_i \langle \nabla f(x^i), r(x^i)\rangle + L\int_0^{\eta_i} t^\alpha \|r(x^i)\|^{1+\alpha} dt\\ &= -\eta_i \langle \nabla f(x^i), r(x^i)\rangle + \frac{L\eta_i^{1+\alpha}}{1+\alpha}\|r(x^i)\|^{1+\alpha},\end{aligned}$$

where the inequality follows from the Hölder continuity of $\nabla f(\cdot)$ and the Cauchy–Schwarz inequality. Hence

$$f(x^i) - f(y_l^i, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}^i) \geq \eta_i \langle \nabla f(x^i), r(x^i)\rangle - \frac{L\eta_i^{1+\alpha}}{1+\alpha}\|r(x^i)\|^{1+\alpha}. \tag{2.1}$$

By properties of the projection operator [22, p. 121], for any $x \in \Re^n$ and any $y \in C$ it holds that

$$\langle x - P_C[x], y - P_C[x] \rangle \leq 0.$$

Taking $x = x^i - \nabla f(x^i)$ and $y = x^i \in C$, we have

$$\begin{aligned} 0 &\geq \langle x^i - \nabla f(x^i) - P_C[x^i - \nabla f(x^i)], x^i - P_C[x^i - \nabla f(x^i)] \rangle \\ &= \langle r(x^i) - \nabla f(x^i), r(x^i) \rangle. \end{aligned}$$

Hence

$$\langle \nabla f(x^i), r(x^i) \rangle \geq \|r(x^i)\|^2.$$

The latter relation combined with (2.1) yields

$$\begin{aligned} f(x^i) - f(y_l^i, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}^i) &\geq \eta_i \|r(x^i)\|^2 - \frac{L\eta_i^{1+\alpha}}{1+\alpha} \|r(x^i)\|^{1+\alpha} \\ &= \eta_i \|r(x^i)\|^2 \left(1 - \frac{L\eta_i^{\alpha}}{1+\alpha} \|r(x^i)\|^{\alpha-1}\right) \\ &\geq (1 - \beta^\alpha) \eta_i \|r(x^i)\|^2 \\ &\geq (1 - \beta^\alpha) \min \left\{ \|r(x^i)\|^2, \ \beta \left(\frac{1+\alpha}{L}\right)^{1/\alpha} \|r(x^i)\|^{(1+\alpha)/\alpha} \right\}, \end{aligned}$$

where the last two inequalities follow from the choice of η_i. By the synchronization step (1.3), we have

$$(2.2) \quad f(x^i) - f(x^{i+1}) \geq (1 - \beta^\alpha) \min \left\{ \|r(x^i)\|^2, \ \beta \left(\frac{1+\alpha}{L}\right)^{1/\alpha} \|r(x^i)\|^{(1+\alpha)/\alpha} \right\}.$$

Thus the sequence $\{f(x^i)\}$ is nonincreasing. If $f(\cdot)$ is bounded below on C, then $\{f(x^i)\}$ is bounded and hence it converges. In the latter case, $\{f(x^i) - f(x^{i+1})\} \to 0$ and therefore $\{r(x^i)\} \to 0$ by (2.2). By continuity of $r(\cdot)$, it follows that for every accumulation point $\bar{x}$ of the sequence $\{x^i\}$, $r(\bar{x}) = 0$. Thus all accumulation points of $\{x^i\}$ satisfy the minimum principle necessary optimality condition. □

Some remarks are in order on the practical issues related to computing the projected gradient directions. When C is a polyhedral set (i.e., the constraints are linear), computing these directions requires solving a single quadratic programming problem at every synchronization step of Algorithm 1.1. For this, a wealth of fast and reliable algorithms is available [9]. In the case of nonlinear convex constraints, the task of computing the projected gradient directions is considerably more difficult. One of the referees for this paper made an interesting suggestion of computing the secondary directions by projecting onto the linearization of a nonlinear constraint set, much in the spirit of the iterative quadratic programming approach [6]. At this time, this is an open question as we do not have a proof verifying that this approach will work. However, it certainly deserves further investigation.

Careful reexamination of the proof of Theorem 2.1 shows that there is no need to compute exact global solutions for parallel subproblems (1.2). In particular, it is clear that the analysis is still valid if we accept any point (y_l^i, μ_l^i) that yields at least as good a descent for the function $\psi_l^i(\cdot, \cdot)$ as the point $(x_l^i - \eta_i r_l(x^i), -\eta_i e_{\bar{l}})$. We

can therefore replace the minimization problem in (1.2) with the following *sufficient descent* condition:

$$\text{Compute} \quad (y_l^i, \mu_{\bar{l}}^i) \quad \text{such that} \quad \psi_l^i(y_l^i, \mu_{\bar{l}}^i) \le \psi_l^i(x_l^i - \eta_i r_l(x^i), -\eta_i e_{\bar{l}}). \tag{2.3}$$

This observation is significant for several reasons. First of all, the global solution requirement is impractical if the objective function is not convex. And in any case, insisting on exact subproblem solution can be undesirable because it is likely to result in considerable idle times for processors that have already completed their work. It can be especially wasteful on the initial stages of minimization process. The sufficient descent criterion above shows that we are allowed a lot of flexibility in devising PVD algorithms. In particular, we can allow each of the p-parallel processors to take as many steps as desired. Synchronization can be performed at any time provided every processor has achieved the sufficient descent condition. This provides a flexible framework for effective load balancing. Finally, we note that (2.3) can be easily satisfied in practice if we take $(x_l^i, 0)$ as starting points when minimizing $\psi_l^i(\cdot, \cdot)$, $l = 1, \ldots, p$, in subproblems (1.2).

In [4], the following *optimality function* was used to monitor the progress of the algorithm in the case of block-separable constraints:

$$\varphi(x) := -\min\{\langle \nabla f(x), h\rangle \mid x + h \in C, \ \|h\| \le 1\}.$$

Then convergence results similar to our Theorem 2.1 follow from the relation

$$f(x^i) - f(x^{i+1}) \ge c\, \varphi(x^i), \quad c > 0,$$

and the fact that $\varphi(x) = 0$ if x satisfies the minimum principle necessary optimality condition and $\varphi(x) > 0$ otherwise. The use of a different optimality function

$$\varphi(x) := \|r(x)\|^2$$

allows us furthermore to give a rate of convergence result under the conditions similar to those in [13].

THEOREM 2.2. *Let $f(\cdot) \in C_L^{1,1}(\Re^n)$ and $\{x^i\}$ be any sequence generated by Algorithm* 1.1. *Let accumulation points of the sequence $\{x^i\}$ exist and belong to the set $S := \arg\min_{x \in C} f(x) \ne \emptyset$. Suppose there exist positive constants c_j, $j = 2, \ldots, 5$, such that*

$$\|r(x)\| \ge c_2 d(x, S) \quad \textit{for all } x \textit{ with} \quad \|r(x)\| \le c_3 \tag{2.4}$$

and

$$f(x) - \min_{x \in C} f(x) \le c_4 d(x, S)^2 \quad \textit{for all } x \textit{ with} \quad d(x, S) \le c_5. \tag{2.5}$$

Then the sequence $\{f(x^i)\}$ converges Q-linearly and the sequences $\{r(x^i)\}$ and $\{d(x^i, S)\}$ converge to zero R-linearly. If, furthermore, the solution set is a singleton ($S = \{\bar{x}\}$), then the sequence $\{x^i\}$ converges to $\bar{x}$ at least R-linearly.

Proof. By Theorem 2.1, $\{r(x^i)\} \to 0$ and hence also $\{d(x^i, S)\} \to 0$. It follows that relations (2.4) and (2.5) are satisfied for i sufficiently large, say $i \ge i_1$. Then, for $i \ge i_1$, from (2.2) with $\alpha = 1$ and (2.4), (2.5) we obtain

$$\begin{aligned} f(x^i) - f(x^{i+1}) &\ge c_1 c_2^2 d(x^i, S)^2 \\ &\ge \frac{c_1 c_2^2}{c_4} (f(x^i) - f^*), \end{aligned}$$

where $f^* = \min_{x \in C} f(x)$. Rearranging terms gives

$$f(x^{i+1}) - f^* \leq \left(1 - \frac{c_1 c_2^2}{c_4}\right)(f(x^i) - f^*).$$

Hence the sequence $\{f(x^i)\}$ converges Q-linearly. By (2.2), the sequence $\{r(x^i)\}$ converges to zero R-linearly and, by (2.4), $\{d(x^i, S)\}$ also converges to zero R-linearly.

If S is a singleton, that is $S = \{\bar{x}\}$, then $d(x^i, S) = \|x^i - \bar{x}\|$ and hence $\{x^i\}$ converges to $\bar{x}$ R-linearly. □

Remark 2.1. It is an open question whether the sequence $\{x^i\}$ itself converges linearly under the assumptions of Theorem 2.2 when S is not a singleton. Note that if we had a serial algorithm where

$$x^{i+1} - x^i = -\eta_i r(x^i)$$

with the sequence of stepsizes $\{\eta_i\}$ uniformly bounded away from zero, then the linear convergence rate of $\{x^{i+1} - x^i\}$ (and hence also of $\{x^i\}$) would immediately follow from the linear convergence to zero of $\{r(x^i)\}$. The difficulty with the parallel algorithm is that we cannot explicitly relate $\{r(x^i)\}$ to $\{x^{i+1} - x^i\}$.

Condition (2.4) is sometimes referred to as a *projection-type error bound* [13, 14]. This error bound is known to hold when $f(\cdot)$ is quadratic and C is polyhedral [12, 23] or when $f(\cdot)$ is strongly convex [20] or when $f(\cdot)$ is a certain convex (not necessarily strictly convex) function and C is polyhedral [13] (see [14, Theorem 2.1] for a summary). Moreover, in certain situations, this condition holds globally (with $c_3 = \infty$) [10, 11, 17, 20]. Condition (2.5) states that $f(\cdot)$ does not grow faster than quadratically near its set of minimizers.

In conclusion, we note that the implicit Lagrangian reformulation [18] of the nonlinear complementarity problem (NCP) satisfies (2.4)–(2.5) under certain assumptions. Consider the following NCP [3, 21] of finding an $x \in \Re^n$ such that

$$F(x) \geq 0, \quad x \geq 0, \quad \langle x, F(x) \rangle = 0,$$

where $F : \Re^n \to \Re^n$. In [18] it was established that the NCP can be solved via minimization of the following implicit Lagrangian function:

$$M(x, \alpha) := 2\alpha \langle x, F(x) \rangle + \|[x - \alpha F(x)]^+\|^2 - \|x\|^2 + \|[F(x) - \alpha x]^+\|^2 - \|F(x)\|^2,$$

where $\alpha > 1$ is a parameter and $[\cdot]^+$ denotes the orthogonal projection onto the nonnegative orthant $\Re^n_+$. In particular, the implicit Lagrangian is nonnegative everywhere in $\Re^n$ and assumes the value of zero precisely at the solutions of the NCP. One of the advantages of this reformulation is that if $F(\cdot)$ is continuously differentiable then so is $M(\cdot, \alpha)$.

First note that $\nabla M(\cdot, \alpha)$ is Lipschitz continuous on any bounded set if $F(\cdot)$ and $\nabla F(\cdot)$ are Lipschitz continuous and it is Lipschitz continuous everywhere if $F(\cdot)$ is affine. In [11] it was established that for all $x \in \Re^n$ (and any $F(\cdot)$)

$$2(\alpha - 1)\|x - [x - F(x)]^+\|^2 \leq M(x, \alpha) \leq 2\alpha(\alpha - 1)\|x - [x - F(x)]^+\|^2. \tag{2.6}$$

Let $S := \{x \in \Re^n \mid x - [x - F(x)]^+ = 0\}$ be the (nonempty) solution set of the NCP. It can be easily seen that (2.5) is satisfied if the mapping $F(\cdot)$ is Lipschitz continuous

(with some modulus $\mu > 0$):

$$\begin{aligned} M(x,\alpha)^{\frac{1}{2}} - \min_{x\in\Re^n} M(x,\alpha)^{\frac{1}{2}} &= M(x,\alpha)^{\frac{1}{2}} \\ &\le \sqrt{2\alpha(\alpha-1)}\|x-[x-F(x)]^+\| \\ &= \sqrt{2\alpha(\alpha-1)}\|x-[x-F(x)]^+ - P_S[x] \\ &\qquad +[P_S[x]-F(P_S[x])]^+\| \\ &\le \sqrt{2\alpha(\alpha-1)}\|x-P_S[x]\| \\ &\quad +\sqrt{2\alpha(\alpha-1)}\|[x-F(x)]^+ - [P_S[x]-F(P_S[x])]^+\| \\ &\le \sqrt{2\alpha(\alpha-1)}\,(2\|x-P_S[x]\| + \|F(x)-F(P_S[x])\|) \\ &\le \sqrt{2\alpha(\alpha-1)}(2+\mu)d(x,S), \end{aligned}$$

where the first inequality follows from (2.6), the equality follows from $P_S[x] \in S$, the second inequality follows from the Cauchy–Schwarz inequality, the third from the nonexpansiveness of $P_S[\cdot]$, and the last from the Lipschitz continuity of $F(\cdot)$.

It is further known that, when $F(\cdot)$ is affine [12, 23] or $F(\cdot)$ has certain strong monotonicity structure [27, Theorem 2], then the following error bound holds:

$$d(x,S) \le \rho\|x-[x-F(x)]^+\| \quad \forall x \text{ with } \|x-[x-F(x)]^+\| \le \epsilon,$$

where ρ and ϵ are positive constants (independent of x). In that case, from (2.6),

$$M(x,\alpha) - \min_{x\in\Re^n} M(x,\alpha) = M(x,\alpha) \ge \nu d(x,S)^2, \quad \nu := 2(\alpha-1)/\rho^2 > 0. \tag{2.7}$$

By a well-known lemma [22, p. 6], we have that

$$M(x,\alpha) - M(P_S[x],\alpha) \le \langle \nabla M(x,\alpha), x-P_S[x]\rangle + \frac{L}{2}\|x-P_S[x]\|^2.$$

Hence, from (2.7),

$$\begin{aligned} \langle \nabla M(x,\alpha), x-P_S[x]\rangle &\ge M(x,\alpha) - \frac{L}{2}d(x,S)^2 \\ &\ge (\nu - L/2)M(x,\alpha). \end{aligned}$$

Therefore, if $\nu > L/2$,

$$\begin{aligned} \|\nabla M(x,\alpha)\|d(x,S) &\ge \langle \nabla M(x,\alpha), x-P_S[x]\rangle \\ &\ge (\nu - L/2)M(x,\alpha) \\ &\ge \nu(\nu-L/2)d(x,S)^2. \end{aligned}$$

Hence

$$\|\nabla M(x,\alpha)\| \ge \nu(\nu-L/2)d(x,S).$$

Since in this case $C = \Re^n$, it follows that $r(x) = \nabla M(x,\alpha)$ and (2.4) holds.

3. Concluding remarks. It was established that the parallel variable distribution approach can be successfully applied to solve optimization problems with general convex constraints. Furthermore, conditions imposed on solving parallel subproblems were considerably relaxed, thus yielding a more practical framework. A rate of convergence result was also presented.

Acknowledgment. I am grateful to the two anonymous referees for careful reading and constructive suggestions which led to numerous improvements in the paper.

REFERENCES

[1] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1989.
[3] R. W. Cottle, F. Giannessi, and J.-L. Lions, *Variational Inequalities and Complementarity Problems: Theory and Applications*, Wiley, New York, 1980.
[4] M. C. Ferris and O. L. Mangasarian, *Parallel variable distribution*, SIAM J. Optim., 4 (1994), pp. 815–832.
[5] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.
[6] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, Chichester, UK, 1981.
[7] M. Fukushima, *Parallel variable transformation in unconstrained optimization*, SIAM J. Optim., 8 (1998), to appear.
[8] S.-P. Han and O. L. Mangasarian, *A dual differentiable exact penalty function*, Math. Programming, 25 (1983), pp. 293–301.
[9] Y. Y. Lin and J.-S. Pang, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383–411.
[10] X.-D. Luo and P. Tseng, *On global projection-type error bound for the linear complementarity problem*, Linear Algebra Appl., to appear.
[11] Z.-Q. Luo, O. L. Mangasarian, J. Ren, and M. V. Solodov, *New error bounds for the linear complementarity problem*, Math. Oper. Res., 19 (1994), pp. 880–892.
[12] Z.-Q. Luo and P. Tseng, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43–54.
[13] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.
[14] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178.
[15] O. L. Mangasarian, *Nonlinear Programming*, McGraw–Hill, New York, 1969.
[16] O. L. Mangasarian, *Parallel gradient distribution in unconstrained optimization*, SIAM J. Control Optim., 33 (1995), pp. 1916–1925.
[17] O. L. Mangasarian and J. Ren, *New improved error bounds for the linear complementarity problem*, Math. Programming, 66 (1994), pp. 241–255.
[18] O. L. Mangasarian and M. V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 277–297.
[19] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[20] J.-S. Pang, *A posteriori error bounds for the linearly–constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.
[21] J.-S. Pang, *Complementarity problems*, in Handbook of Global Optimization, R. Horst and P. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 271–338.
[22] B. T. Polyak, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.
[23] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Study, 14 (1981), pp. 206–214.
[24] R. T. Rockafellar, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control, 12 (1974), pp. 268–285.
[25] M. V. Solodov, *New inexact parallel variable distribution algorithms*, Comput. Optim. Appl., 7 (1997), pp. 165–182.
[26] P. Tseng, *Dual coordinate ascent methods for non-strictly convex minimization*, Math. Programming, 59 (1993), pp. 231–248.
[27] P. Tseng, *On linear convergence of iterative methods for the variational inequality problem*, J. Comput. Appl. Math., 60 (1995), pp. 237–252.