Computational Optimization and Applications 7, 165–182 (1997)


# New Inexact Parallel Variable Distribution Algorithms*

MICHAEL V. SOLODOV solodov@impa.br
*Instituto de Matematica Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botanico, Rio de Janeiro, RJ, CEP 22460-320, Brazil*

*Received June 26, 1995; Revised October 24, 1995*

**Abstract.** We consider the recently proposed parallel variable distribution (PVD) algorithm of Ferris and Mangasarian [4] for solving optimization problems in which the variables are distributed among $p$ processors. Each processor has 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. We propose useful generalizations that consist, for the general unconstrained case, of replacing exact global solution of the subproblems by a certain natural sufficient descent condition, and, for the convex case, of inexact subproblem solution in the PVD algorithm. These modifications are the key features of the algorithm that has not been analyzed before. The proposed modified algorithms are more practical and make it easier to achieve good load balancing among the parallel processors. We present a general framework for the analysis of this class of algorithms and derive some new and improved linear convergence results for problems with weak sharp minima of order 2 and strongly convex problems. We also show that nonmonotone synchronization schemes are admissible, which further improves flexibility of PVD approach.

**Keywords:** parallel optimization, asynchronous algorithms, load balancing, unconstrained minimization, linear convergence

## 1. Introduction

We consider the general unconstrained optimization problem

$$\min_{x\in\Re^n} f(x), \tag{1}$$

where $f:\Re^n\to\Re$. We first state the original PVD algorithm [4]. Let $x\in\Re^n$ 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 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 $x_l^i + D_l^i\mu_{\bar{l}}$ in the parallel subproblems (2).

*This work was started when the author was with the Computer Sciences Department, University of Wisconsin-Madison, U.S.A., and was supported by Air Force Office of Scientific Research Grant F49620-94-1-0036 and National Science Foundation Grant CCR-9322479.

The presence of this term allows for a change in "secondary" variables. This makes PVD fundamentally different from the block Jacobi [1], coordinate descent [20] and parallel gradient distribution algorithms [10]. The directions $D_{\bar{l}}^i$ are typically easily computable steepest descent or quasi-Newton directions in the space of the corresponding variables. The "forget-me-not" approach improves robustness and accelerates convergence of the algorithm and is the key to its success. The parallelization phase is followed by a simple synchronization step which picks up a point with the objective function value at least as good as the smallest among all the new points computed by the parallel processors.

**Algorithm 1 (PVD).** *Start with any* $x^0 \in \Re^n$. *Having* $x^i$, *stop if* $\nabla f(x^i) = 0$. *Otherwise, compute* $x^{i+1}$ *as follows*:

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

$$\left(y_l^i, \mu_{\bar{l}}^i\right) \in \arg\min_{x_l, \mu_{\bar{l}}} \psi_l^i(x_l, \mu_{\bar{l}}) := f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}\right). \tag{2}$$

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

$$f(x^{i+1}) \leq \min_{l \in \{1,\ldots,p\}} \psi_l^i\left(y_l^i, \mu_{\bar{l}}^i\right). \tag{3}$$

We will sometimes refer to $x^i$ as the base point at the $(i+1)$-st iteration. 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 $d^i \in \Re^n$ along its block diagonal as follows:

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

In the original PVD algorithm the proposed synchronization step consists of minimizing the objective function in the affine hull of all the points computed in parallel by the $p$ processors.

In [4] it was shown that every accumulation point of the PVD iterates is a stationary point of $f(\cdot)$ if an *exact global solution* to subproblems (2) is computed at every iteration. It was also established that, in the strongly convex case, the iterates converge to the problem solution at a linear rate.

We point out that the *global solution* requirement in the general (nonconvex) case is impractical. In Section 3 we show that it is possible to get rid of this requirement by imposing a certain sufficient descent condition instead. Section 3 also contains some new convergence results for problems with weak sharp minima of order 2. We note that the original requirement of *exact subproblem solution* is also undesirable. In Section 2 we describe an

algorithm with inexact subproblem solution in the convex case and derive a sharper linear convergence result than the one given in [4]. We emphasize that the sufficient descent and inexact subproblem solution approaches provide a flexible framework that allows for effective load balancing among the parallel processors. In Section 3 we also exhibit that synchronization step can be combined with nonmonotone stabilization schemes, if needed.

One of the keys to our analysis is imposing certain reasonable conditions on the choice of directions for the change in secondary variables. The choice of those directions is very important for the success of the PVD approach. This fact was empirically observed in [4]. It can also be vividly illustarted by theoretical considerations for the constrained optimization problems [19].

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 $\|x\|^2 = \langle x, x \rangle$. The closed unit ball in $\Re^n$ is denoted by $B := \{x \in \Re^n \mid \|x\| \leq 1\}$. For a nonempty (closed) set $X \subset \Re^n$, $d(\cdot, X)$ denotes the Euclidean distance to the set $X$. For a real-valued matrix $A$ of any dimension, $A^\top$ denotes its transpose. For a differentiable function $f : \Re^n \to \Re$, $\nabla f$ will denote the $n$-dimensional vector of partial derivatives with respect to $x$, and $\nabla_l f$ will denote the $n_l$-dimensional vector of partial derivatives with respect to $x_l \in \Re^{n_l}$, $l = 1, \ldots, p$. If a function $f(\cdot)$ has Lipschitz continuous partial derivatives on $\Re^n$ with some constant $L > 0$, that is

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

we write $f(\cdot) \in C^1_L(\Re^n)$. 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 [13].

We now state a classical lemma ([16], p. 6), as well as another lemma (a slight modification of [16], p. 44) that will be used later.

**Lemma 1.** *Let $\varphi(\cdot) \in C^1_L(\Re^n)$, then*

$$|\varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2 \quad \forall x, y \in \Re^n.$$

**Lemma 2.** *Let $\{a^i\}$ and $\{\epsilon^i\}$ be two sequences of real numbers such that $\epsilon^i \geq 0$, $\sum_{i=0}^{\infty} \epsilon^i < \infty$, and $a^{i+1} \leq a^i + \epsilon^i$ for $i = 0, 1, \ldots$. It follows that either the sequence $\{a^i\}$ is unbounded below, or it converges.*

## 2. PVD with inexact subproblem solution

In this section we propose a computationally important modification of the PVD algorithm in which the subproblems (2) in the Algorithm 1 are solved approximately. It is clear that in practice insisting on exact solution of those subproblems is undesirable, and often unrealistic. Even when it is possible to compute these solutions accurately, it can be wasteful doing so, especially in the initial stages of the minimization process.

Our results show that there is no need to wait until exact solutions to all the subproblems are found (which can result in considerable idle times for processors that have already

completed their work). Instead, we can accept the current approximations to solutions of the subproblems and proceed to the synchronization step, provided those approximations are reasonably good. This approach is more robust and allows for flexible synchronization schemes thus making it easier to achieve good load balancing among the parallel processors. In particular, we show that we can solve the subproblems to within *$\varepsilon$-stationarity* (see (5)), and yet guarantee the linear convergence rate if $f(\cdot)$ is strongly convex. The tolerance for an $l$-th parallel subproblem depends linearly on the the norm of the corresponding portion of the gradient at the current base point (see (5) and (10)).

By making an explicit use of the "forget-me-not" terms in the subproblems, we also improve on the linear convergence result given in [4]. In [4] it is established that, for the strongly convex case, the following estimate is valid

$$\|x^i - \bar{x}\| \le c_1 \left(1 - \frac{c_2}{p}\right)^{\frac{i}{2}},$$

where $\bar{x}$ is the (unique) solution of the problem, $p$ is the number of parallel processors, and $c_1, c_2$ are positive constants. This result is not quite satisfactory because the presense of $p$ in the denominator suggests that the convergence speed goes down as the number of processors used increases. We point out that the proof given in [4] fails to make use of the "forget-me-not" terms which are the key to the algorithm. By refining the proof, we obtain a better convergence speed estimate

$$\|x^i - \bar{x}\| \le c_1 (1 - c_3)^{\frac{i}{2}},$$

where $c_3 > 0$ does not depend on $p$. Therefore convergence speed of the algorithm does not deteriorate as the number of processors used increases, provided certain natural conditions are imposed on the "forget-me-not" terms.

We consider the following algorithm.

**Algorithm 2.** *Start with any $x^0 \in \Re^n$. Having $x^i$, stop if $\nabla f(x^i) = 0$. Otherwise, compute $x^{i+1}$ as follows*:

**(•)** ***Parallelization:*** *For each processor $l \in \{1, \ldots, p\}$ compute $(y_l^i, \mu_{\bar{l}}^i)$ as an $\varepsilon_{i,l}$-approximate solution (see (5)) of*

$$\min_{x_l, \mu_{\bar{l}}} \psi_l^i(x_l, \mu_{\bar{l}}) := f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}\right).$$

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

$$f(x^{i+1}) \le \min_{l \in \{1, \ldots, p\}} \psi_l^i\left(y_l^i, \mu_{\bar{l}}^i\right). \tag{4}$$

To make the parallelization step precise, we say that the current approximation to the solution of a subproblem is admissible if it belongs to an $\varepsilon$-stationary set [18] of this subproblem. The parallelization subproblems are therefore equivalent to computing a point

$$\left(y_l^i, \mu_{\bar{l}}^i\right) \in X_s^{l,i}(\varepsilon_{i,l}) := \left\{(x_l, \mu_{\bar{l}}) \in \Re^{n_l + p - 1} \;\middle|\; \left\|\nabla \psi_l^i(x_l, \mu_{\bar{l}})\right\| \le \varepsilon_{i,l}\right\}. \tag{5}$$

We first establish some preliminary results. Let $A_l^i$ be an $n \times (n_l + p - 1)$ matrix defined by

$$A_l^i = \begin{pmatrix} I_l & 0 \\ 0 & D_{\bar{l}}^i \end{pmatrix},$$

where $I_l$ is an $n_l \times n_l$ identity matrix. We assume that every block $d_t^i$ of $D_{\bar{l}}^i$ is normalized, that is $\|d_t^i\| = 1,\ t = 1, \ldots, p$. Then for any $y \in \Re^{n_l+p-1}$ we have

$$\begin{aligned} \left\|A_l^i y\right\|^2 &= \sum_{j=1}^{n_l} y_j^2 + \sum_{j=n_l+1}^{n_l+p-1} y_j^2 \left\|d_j^i\right\|^2 \\ &= \sum_{j=1}^{n_l+p-1} y_j^2 \\ &= \|y\|^2, \end{aligned} \tag{6}$$

where the first equality follows from the block diagonal structure of $D_{\bar{l}}^i$. Hence $\|A_l^i\| = \|(A_l^i)^\top\| = 1$.

**Lemma 3.** *If $f(\cdot) \in C_L^1(\Re^n)$ then $\psi_l^i(\cdot,\cdot) \in C_L^1(\Re^{n_l+p-1})$ for any $i = 0, 1, \ldots$ and $l = 1, \ldots, p$.*

**Proof:** Note that

$$\begin{aligned} \nabla \psi_l^i(x_l, \mu_{\bar{l}}) &= \begin{pmatrix} \nabla_l f\left(x_l, x_{\bar{l}}^i + \mu_{\bar{l}} D_{\bar{l}}^i\right) \\ \left(D_{\bar{l}}^i\right)^\top \nabla_{\bar{l}} f\left(x_l, x_{\bar{l}}^i + \mu_{\bar{l}} D_{\bar{l}}^i\right) \end{pmatrix} \\ &= \left(A_l^i\right)^\top \nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}\right) \end{aligned} \tag{7}$$

For any $(x_l, \mu_{\bar{l}}), (z_l, \nu_{\bar{l}}) \in \Re^{n_l+p-1}$ we have

$$\begin{aligned} \left\|\nabla \psi_l^i(x_l, \mu_{\bar{l}}) - \nabla \psi_l^i(z_l, \nu_{\bar{l}})\right\| &= \left\|\left(A_l^i\right)^\top \left(\nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}\right) - \nabla f\left(z_l, x_{\bar{l}}^i + D_{\bar{l}}^i \nu_{\bar{l}}\right)\right)\right\| \\ &\le \left\|\left(A_l^i\right)^\top\right\| \left\|\nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i \mu_{\bar{l}}\right) - \nabla f\left(z_l, x_{\bar{l}}^i + D_{\bar{l}}^i \nu_{\bar{l}}\right)\right\| \\ &\le L \left\| A_l^i \begin{pmatrix} x_l - z_l \\ \mu_{\bar{l}} - \nu_{\bar{l}} \end{pmatrix} \right\| \\ &= L \|(x_l, \mu_{\bar{l}}) - (z_l, \nu_{\bar{l}})\|, \end{aligned}$$

where the second inequality follows from the fact that $\|(A_l^i)^\top\| = 1$, and $f(\cdot) \in C_L^1(\Re^n)$; the last equality follows from (6). We thus established that $\psi_l^i(\cdot,\cdot) \in C_L^1(\Re^{n_l+p-1})$, for all $l = 1, \ldots, p,\ i = 0, 1, \ldots$. □

**Lemma 4.** *If $f(\cdot)$ is strongly convex with modulus $\theta > 0$ then $\psi_l^i(\cdot,\cdot)$ is strongly convex with modulus $\theta > 0$ for any $i = 0, 1, \ldots$ and $l = 1, \ldots, p$.*

**Proof:** Making use of (7), we have

$$\begin{aligned}
&\left\langle \nabla\psi_l^i(x_l,\mu_{\bar{l}}) - \nabla\psi_l^i(z_l,\nu_{\bar{l}}), (x_l,\mu_{\bar{l}}) - (z_l,\nu_{\bar{l}})\right\rangle \\
&\quad = \left((A_l^i)^\top\left(\nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i\mu_{\bar{l}}\right) - \nabla f\left(z_l, x_{\bar{l}}^i + D_{\bar{l}}^i\nu_{\bar{l}}\right)\right)\right)^\top \begin{pmatrix} x_l - z_l \\ \mu_{\bar{l}} - \nu_{\bar{l}}\end{pmatrix} \\
&\quad = \left(\nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i\mu_{\bar{l}}\right) - \nabla f\left(z_l, x_{\bar{l}}^i + D_{\bar{l}}^i\nu_{\bar{l}}\right)\right)^\top A_l^i\begin{pmatrix} x_l - z_l \\ \mu_{\bar{l}} - \nu_{\bar{l}}\end{pmatrix} \\
&\quad = \left(\nabla f\left(x_l, x_{\bar{l}}^i + D_{\bar{l}}^i\mu_{\bar{l}}\right) - \nabla f\left(z_l, x_{\bar{l}}^i + D_{\bar{l}}^i\nu_{\bar{l}}\right)\right)^\top \begin{pmatrix} x_l - z_l \\ D_{\bar{l}}^i(\mu_{\bar{l}} - \nu_{\bar{l}})\end{pmatrix} \\
&\quad \geq \theta\left\|\begin{pmatrix} x_l - z_l \\ D_{\bar{l}}^i(\mu_{\bar{l}} - \nu_{\bar{l}})\end{pmatrix}\right\|^2 \\
&\quad = \theta\left\|A_l^i\begin{pmatrix} x_l - z_l \\ \mu_{\bar{l}} - \nu_{\bar{l}}\end{pmatrix}\right\|^2 \\
&\quad = \theta\|(x_l,\mu_{\bar{l}}) - (z_l,\nu_{\bar{l}})\|^2,
\end{aligned}$$

where the inequality follows from strong convexity of $f(\cdot)$, and the last equality follows from (6). Hence $\psi_l^i(\cdot,\cdot)$ is strongly convex with modulus $\theta$. □

For simplicity of presentation, from now on we assume that

$$d_t^i = \frac{\nabla_t f(x^i)}{\|\nabla_t f(x^i)\|}, \quad t = 1,\ldots,p.$$

For this choice of directions, we have

$$\left(A_l^i\right)^\top\nabla f(x^i) = \begin{pmatrix} \nabla_l f(x^i) \\ \left\langle d_1^i, \nabla_1 f(x^i)\right\rangle \\ \vdots \\ \left\langle d_{l-1}^i, \nabla_{l-1} f(x^i)\right\rangle \\ \left\langle d_{l+1}^i, \nabla_{l+1} f(x^i)\right\rangle \\ \vdots \\ \left\langle d_p^i, \nabla_p f(x^i)\right\rangle \end{pmatrix} = \begin{pmatrix} \nabla_l f(x^i) \\ \|\nabla_1 f(x^i)\| \\ \vdots \\ \|\nabla_{l-1} f(x^i)\| \\ \|\nabla_{l+1} f(x^i)\| \\ \vdots \\ \|\nabla_p f(x^i)\| \end{pmatrix}.$$

Hence, by (7),

$$\begin{aligned}
\left\|\nabla\psi_l^i\left(x_l^i, 0\right)\right\| &= \left\|\left(A_l^i\right)^\top\nabla f(x^i)\right\| \\
&= \|\nabla f(x^i)\|. \qquad (8)
\end{aligned}$$

The latter property enables us to explicitly relate solutions of the parallel subproblems (2) to the progress being made towards solving the original problem (1). This is the key to our generalizations as well as improved convergence results.

We note that instead of the scaled gradient directions we could take any other directions satisfying the natural conditions

$$\left|\left\langle d_t^i, \nabla_t f(x^i)\right\rangle\right| \geq \sigma_t(\|\nabla_t f(x^i)\|), \quad t = 1, \ldots, p,$$

where $\sigma_t(\cdot)$ are forcing functions (see [13], p. 479). Depending on the particular forcing functions, some arguments in the subsequent analysis may need to be changed.

We finally state a useful lemma which is the basis for devising algorithms with inexact subproblem solution in the convex case. This result is a simplification of [18] Lemma 2.4 for smooth unconstrained case. We include the simplified proof for completeness.

**Lemma 5.** *Let $\varphi(\cdot)$ be convex and differentiable. Let $x^* \in X_s := \arg\min_{x\in\Re^n} \varphi(x)$ and $x \in X_s(\varepsilon) := \{x \in \Re^n \mid \|\nabla\varphi(x)\| \leq \varepsilon\}$, $\varepsilon \geq 0$. Then*

$$\varphi(x) - \varphi(x^*) \leq \varepsilon d(x, X_s).$$

*If $\varphi(\cdot)$ is strongly convex with modulus $\theta > 0$, then*

$$\varphi(x) - \varphi(x^*) \leq \frac{\varepsilon^2}{2\theta}.$$

**Proof:** Let $x \in X_s(\varepsilon)$ and $x^*$ be the orthogonal projection of $x$ onto $X_s$. By convexity of $\varphi(\cdot)$, we have

$$\begin{aligned}\varphi(x) - \varphi(x^*) &\leq \langle -\nabla\varphi(x), x^* - x\rangle \\ &\leq \|\nabla\varphi(x)\|\|x - x^*\| \\ &\leq \varepsilon d(x, X_s).\end{aligned}$$

For the second assertion, just note that ([16], p. 24) for any $x \in \Re^n$

$$2\theta(\varphi(x) - \varphi(x^*)) \leq \|\nabla\varphi(x)\|^2.$$

The proof is complete. □

We are now ready to prove our main results.

**Theorem 1.** *Suppose $f(\cdot)$ is strongly convex with modulus $\theta > 0$ and $f(\cdot) \in C_L^1(\Re^n)$. If*

$$\sum_{i=0}^{\infty} \max_{l\in\{1,\ldots,p\}} \varepsilon_{i,l}^2 < \infty, \tag{9}$$

*then every sequence $\{x^i\}$ generated by Algorithm 2 converges to the solution $\bar{x}$ of (1). Moreover, if*

$$\varepsilon_{i,l} \leq \beta\|\nabla_l f(x^i)\|, \quad 0 \leq \beta < \sqrt{\frac{\theta}{L}} \tag{10}$$

*then* $\{x^i\}$ *converges to* $\bar{x}$ *R-linearly*:

$$\|x^i - \bar{x}\| \le \left(\frac{2}{\theta}(f(x^0) - f(\bar{x}))\right)^{\frac{1}{2}} \left(1 - \frac{\theta(\theta - L\beta^2)}{L^2}\right)^{\frac{i}{2}}.$$

**Proof:** For any iteration $i = 0, 1, \ldots$ and any processor $l = 1, \ldots, p$, by (5) and Lemma 5, we have that

$$\psi_l^i\left(y_l^i, \mu_l^i\right) \le \bar{\psi}_l^i + \frac{\varepsilon_{i,l}^2}{2\theta}, \tag{11}$$

where $\bar{\psi}_l^i$ is the exact optimal value of the corresponding subproblem. Define an auxiliary point

$$\Re^{n_l+p-1} \ni \left(z_l^i, \nu_l^i\right) := \left(x_l^i, 0\right) - \frac{1}{L}\nabla\psi_l^i\left(x_l^i, 0\right).$$

We further obtain

$$\begin{aligned} f(x^i) - f\left(y_l^i, x_{\bar{l}}^i + D_{\bar{l}}^i\mu_{\bar{l}}^i\right) &= \psi_l^i\left(x_l^i, 0\right) - \psi_l^i\left(y_l^i, \mu_l^i\right) \\ &\ge \psi_l^i\left(x_l^i, 0\right) - \bar{\psi}_l^i - \frac{\varepsilon_{i,l}^2}{2\theta} \\ &\ge \psi_l^i\left(x_l^i, 0\right) - \psi_l^i\left(z_l^i, \nu_l^i\right) - \frac{\varepsilon_{i,l}^2}{2\theta} \\ &\ge \frac{1}{2L}\left\|\nabla\psi_l^i\left(x_l^i, 0\right)\right\|^2 - \frac{\varepsilon_{i,l}^2}{2\theta} \\ &= \frac{1}{2L}\|\nabla f(x^i)\|^2 - \frac{\varepsilon_{i,l}^2}{2\theta}, \end{aligned} \tag{12}$$

where the first inequality follows from (11), the third inequality from Lemma 1, and the last equality from (8). By (4), we have

$$\begin{aligned} f(x^i) - f(x^{i+1}) &\ge f(x^i) - f\left(y_l^i, x_{\bar{l}}^i + D_{\bar{l}}^i\mu_{\bar{l}}^i\right) \\ &\ge \frac{1}{2L}\|\nabla f(x^i)\|^2 - \frac{1}{2\theta}\max_{l\in\{1,\ldots,p\}}\varepsilon_{i,l}^2. \end{aligned} \tag{13}$$

From (13) we have

$$f(x^{i+1}) \le f(x^i) + \frac{1}{2\theta}\max_{l\in\{1,\ldots,p\}}\varepsilon_{i,l}^2.$$

Note that, by strong convexity of $f(\cdot)$, the sequence $\{f(x^i)\}$ is bounded below. Hence, by Lemma 2 and (9), it follows that the sequence $\{f(x^i)\}$ converges. Therefore $\{f(x^i) -$

$f(x^{i+1})\} \to 0$. Since, by (9),

$$\lim_{i\to\infty} \max_{l\in\{1,\ldots,p\}} \varepsilon_{i,l}^2 = 0,$$

we conclude from (13) that $\{\|\nabla f(x^i)\|\} \to 0$. Since $\bar{x}$, the solution of (1), is the unique stationary point, it follows that $\{x^i\}$ converges to $\bar{x}$.

If (10) holds, then from (12) we obtain

$$\begin{aligned} f(x^i) - f(x^{i+1}) &\geq \frac{1}{2L}\|\nabla f(x^i)\|^2 - \frac{\beta^2}{2\theta}\|\nabla_l f(x^i)\|^2 \\ &\geq \frac{1}{2L}\|\nabla f(x^i)\|^2 - \frac{\beta^2}{2\theta}\|\nabla f(x^i)\|^2 \\ &= \frac{\theta - L\beta^2}{2L\theta}\|\nabla f(x^i)\|^2, \end{aligned} \tag{14}$$

where the second inequality follows from monotonicity of the 2-norm. Note that by (10), $\frac{\theta - L\beta^2}{2L\theta} > 0$. The rest of the proof is standard. By Lemma 1, it follows that

$$\begin{aligned} \frac{L}{2}\|x^i - \bar{x}\|^2 &\geq f(x^i) - f(\bar{x}) - \langle \nabla f(\bar{x}), x^i - \bar{x}\rangle \\ &= f(x^i) - f(\bar{x}) \end{aligned} \tag{15}$$

By the Cauchy-Schwartz inequality and strong convexity of $f(\cdot)$, it follows that

$$\begin{aligned} \|\nabla f(x^i)\|\|x^i - \bar{x}\| &= \|\nabla f(x^i) - \nabla f(\bar{x})\|\|x^i - \bar{x}\| \\ &\geq \langle \nabla f(x^i) - \nabla f(\bar{x}), x^i - \bar{x}\rangle \\ &\geq \theta\|x^i - \bar{x}\|^2. \end{aligned}$$

Hence

$$\|\nabla f(x^i)\| \geq \theta\|x^i - \bar{x}\|.$$

Combining the last inequality with (14), we obtain

$$f(x^i) - f(x^{i+1}) \geq \frac{\theta(\theta - L\beta^2)}{2L}\|x^i - \bar{x}\|^2.$$

This together with (15) yields

$$f(x^i) - f(x^{i+1}) \geq \frac{\theta(\theta - L\beta^2)}{L^2}(f(x^i) - f(\bar{x})).$$

Rearranging terms gives

$$f(x^{i+1}) - f(\bar{x}) \leq \left(1 - \frac{\theta(\theta - L\beta^2)}{L^2}\right)(f(x^i) - f(\bar{x})).$$

Hence the sequence $\{f(x^i)\}$ converges $Q$-linearly. Successive application of the last inequality yields

$$f(x^i) - f(\bar{x}) \le \left(1 - \frac{\theta(\theta - L\beta^2)}{L^2}\right)^i (f(x^0) - f(\bar{x})).$$

By strong convexity of $f(\cdot)$, we have

$$\begin{aligned}\frac{\theta}{2}\|x^i - \bar{x}\|^2 &\le f(x^i) - f(\bar{x}) - \langle \nabla f(\bar{x}), x^i - \bar{x}\rangle \\ &= f(x^i) - f(\bar{x}).\end{aligned}$$

Hence the sequence $\{x^i\}$ converges $R$-linearly. In particular, we have

$$\|x^i - \bar{x}\| \le \left(\frac{2}{\theta}(f(x^i) - f(\bar{x}))\right)^{\frac{1}{2}},$$

and

$$\|x^i - \bar{x}\| \le \left(\frac{2}{\theta}(f(x^0) - f(\bar{x}))\right)^{\frac{1}{2}} \left(1 - \frac{\theta(\theta - L\beta^2)}{L^2}\right)^{\frac{i}{2}}.$$

This completes the proof. □

For the convex case, we have the following result.

**Theorem 2.** *Suppose $f(\cdot)$ is convex and $f(\cdot) \in C^1_L(\Re^n)$. Let $\mathcal{L}(f, x^0) := \{x \mid f(x) \le f(x^0)\}$. Suppose $\mathcal{L}(f, x^0) \subset x^0 + rB$, $r > 0$. If*

$$\varepsilon_{i,l} \le \beta \|\nabla_l f(x^i)\|^2, \quad 0 \le \beta < \frac{1}{2Lr},$$

*or*

$$\sum_{i=0}^{\infty} \max_{l \in \{1,\ldots,p\}} \varepsilon_{i,l} < \infty,$$

*then every accumulation point of any sequence $\{x^i\}$ generated by Algorithm 2 is a solution of (1).*

**Proof:** First note that under our assumptions $\mathcal{L}(f, x^0)$ is bounded and hence $X_s$ is nonempty. Furthermore, for all $i$

$$d(x^i, X_s) \le r.$$

Applying Lemma 5, similarly to the proof of Theorem 1, we obtain

$$f(x^i) - f(x^{i+1}) \geq \frac{1}{2L}\|\nabla f(x^i)\|^2 - r\varepsilon_{i,l}.$$

The rest of the proof can be patterned after that of Theorem 1. □

## 3. PVD with a sufficient descent condition

In this section, we present a practical version of the PVD algorithm for the general (nonconvex) case. In particular, we show that there is no need to find an exact global solution for the subproblems. Any point that satisfies a natural sufficient descent condition can be accepted for the next iteration. We note, in the passing, that the proof given in [4] makes use of exact global solutions in an essential way and breaks down if, for example, only stationary points in the subproblems are available. We further point out that a certain degree of asynchronization among the $p$ parallel processors is possible by allowing each of the $p$ processors to take as many steps as desired by individually updating its base point. Synchronization can be performed at any time provided every processor has achieved the sufficient descent condition. Furthermore, we show that synchronization step need not be monotone and can be combined with nonmonotone stabilization schemes similar to [6].

We also derive some new convergence results for weakly sharp problems of order 2 (see Definition below). This class of problems can be viewed as a generalization of strongly convex problems and a certain unconstrained smooth analogue of weak sharp minima [2].

We begin by imposing a natural sufficient descent condition on an algorithm (Algorithm A below) used to solve the subproblems (2) generated by the PVD Algorithm 1.

**Algorithm A.** *Given any function $\varphi(\cdot) \in C^1_L(\Re^m)$ and any starting point $t^0 \in \Re^m$ generate a point $t^* \in \Re^m$ such that*

$$\varphi(t^0) \geq \varphi(t^*) + \gamma\|\nabla\varphi(t^0)\|^2, \tag{16}$$

*where $\gamma > 0$ depends on $L$ and does not depend on $t^0$.*

Note that the above condition is satisfied by a single iteration of any reasonable descent algorithm [16], [10] applied to the problem of minimizing $\varphi(\cdot)$ with $t^0$ as a starting point. Hence it is also satisfied for a minimum or a stationary point computed by some descent algorithm provided it uses $t^0$ as a starting point.

We now state our new PVD algorithm.

**Algorithm 3.** *Start with any $x^0 \in \Re^n$. Having $x^i$, stop if $\nabla f(x^i) = 0$. Otherwise, compute $x^{i+1}$ as follows*:

**(•)** ***Parallelization:*** *for each processor $l \in \{1, \ldots, p\}$ generate $(y^i_l, \mu^i_l)$ by applying Algorithm A one or more times to the problem*

$$\min_{x_l, \mu_{\bar{l}}} \psi^i_l(x_l, \mu_{\bar{l}}) := f\left(x_l, x^i_{\bar{l}} + D^i_{\bar{l}}\mu_{\bar{l}}\right) \tag{17}$$

*using $(x^i_l, 0)$ as the first starting point.*

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

$$f(x^{i+1}) \leq \max_{l \in \{1,\ldots,p\}} \psi_l^i\left(y_l^i, \mu_l^i\right) + \lambda\gamma\|\nabla f(x^i)\|^2, \tag{18}$$

*where* $\lambda \in (0, 1)$.

Note that once the sufficient descent condition (16) with respect to $f(x^i) = \psi_l^i(x_l^i, 0)$ is satisfied, each processor can independently update its base point, generate new directions $D_l^i$ and proceed to find a point with better objective function value. After these parallel steps are performed by each processor then an eventual synchronization step is taken. Note that our synchronization step may increase rather than decrease the objective function when compared to the values obtained by the parallel processors. This provides the algorithm with more flexibility and is known to be sometimes useful in nonlinear nonconvex optimization [5, 6]. Of course, only computational experiments can give an insight into the usefulness of nonmonotone synchronization schemes for PVD algorithms.

We next introduce a notion of weak sharp minima of order 2 which allows us to strengthen some of the traditional convergence results.

*Definition.* We say that a set of (local) minima $X_s$ is weakly sharp of order 2 if there exist positive constants $\rho$ and $\epsilon$ such that

$$f(x) - f([x]^+) \geq \rho d(x, X_s)^2 \quad \forall x \in X_s + \epsilon B, \tag{19}$$

where $[\cdot]^+$ denotes the orthogonal projection map onto $X_s$.

The class of problems with weak sharp minima of order 2 can be thought of as a certain unconstrained smooth analogue of weak sharp minima (of order 1) [16, 2]. Note that it subsumes strongly convex programs. Let $f(\cdot)$ be strongly convex with modulus $2\rho$. Then its unique optimal point $\bar{x}$ is globally (with $\epsilon = \infty$) weakly sharp of order 2. This can be easily verified as follows. By strong convexity, for any $x \in \Re^n$

$$\begin{aligned} f(x) - f(\bar{x}) &\geq \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{2\rho}{2}\|x - \bar{x}\|^2 \\ &= \rho\|x - \bar{x}\|^2 \\ &= \rho d(x, X_s)^2. \end{aligned}$$

Hence the growth property of $f(\cdot)$ (near the solution set) in the above Definition is a generalization of strong convexity. It is clear that there exist functions with weak sharp minima of order 2 which are not strongly convex (or even convex) in any neighborhood of their solution sets. One example is

$$f(x) := \left(x_1^2 + x_2^2 - 1\right)^2, \quad x \in \Re^2.$$

The stationary set of this function is

$$X_s = \left\{x \mid x_1^2 + x_2^2 = 1\right\} \cup \{(0,0)\} := X_s^1 \cup X_s^2$$

with $X_s^1$ being the set of minima. It is easy to see that $X_s^1$ is a set of weak sharp minima of order 2 (with $\rho = 1$ and $\epsilon = 1/2$). Indeed, for any $x \in X_s^1 + \frac{1}{2}B$

$$\begin{aligned} d\left(x, X_s^1\right)^2 &= \left|1 - \left(x_1^2 + x_2^2\right)\right|^2 \\ &= \left(x_1^2 + x_2^2 - 1\right)^2 \\ &= f(x) - f(\bar{x}) \quad \forall \bar{x} \in X_s^1. \end{aligned}$$

Obviously, even locally (in any neighborhood of $X_s^1$) $f(\cdot)$ in this example is neither strongly convex nor convex. However, we are able to strengthen standard convergence results for problems of this class (see Theorem 3 below). As an aside, we note that $X_s^2 = \{(0,0)\}$ is a set of weak sharp maxima in the sense of the same definition (with the sign of the left-hand-side of (19) reversed).

A remark in the end of this section contains further examples of problems with weak sharp minima of order 2.

**Theorem 3.** *Let $f(\cdot) \in C_L^1(\Re^n)$. Suppose $\{x^i\}$ is any sequence generated by Algorithm* 3. *Then either $f(\cdot)$ is unbounded from below on $\Re^n$ or the sequence $\{f(x^i)\}$ converges, the sequence $\{\nabla f(x^i)\}$ converges to zero and for every accumulation point $\bar{x}$ of the sequence $\{x^i\}$ it follows that $\nabla f(\bar{x}) = 0$.*

*Suppose the sequence $\{x^i\}$ is bounded (this holds, for example, if the level set $\mathcal{L}(f, x^0) := \{x \mid f(x) \leq f(x^0)\}$ is bounded). Let the subset $X_s$ of stationary points of $f(\cdot)$ that contains accumulation points of $\{x^i\}$ be a set of weak sharp minima of order* 2, *and let* (19) *hold with $\rho > L/2$. Then the sequence $\{f(x^i)\}$ converges Q-linearly, and the sequences $\{\nabla f(x^i)\}$ and $\{d(x^i, X_s)\}$ converge to zero R-linearly.*

**Proof:** By Lemma 3, for any iteration $i = 0, 1, \ldots$ and any processor $l = 1, \ldots, p$, $\psi_l^i(\cdot,\cdot) \in C_L^1(\Re^{n_l+p-1})$ (with the same $L$). By (16) and (8), it follows that

$$\begin{aligned} \psi_l^i\left(x_l^i, 0\right) - \psi_l^i\left(y_l^i, \mu_l^i\right) &\geq \gamma \left\|\nabla \psi_l^i\left(x_l^i, 0\right)\right\|^2 \\ &= \gamma \|\nabla f(x^i)\|^2. \end{aligned}$$

Since the last inequality holds for all $l = 1, \ldots, p$, we have

$$f(x^i) - \max_{l \in \{1,\ldots,p\}} \psi_l^i\left(y_l^i, \mu_l^i\right) \geq \gamma \|\nabla f(x^i)\|^2.$$

Hence, by the synchronization step (18),

$$f(x^i) - \left(f(x^{i+1}) - \lambda\gamma \|\nabla f(x^i)\|^2\right) \geq \gamma \|\nabla f(x^i)\|^2$$

and

$$f(x^i) - f(x^{i+1}) \geq (1-\lambda)\gamma\|\nabla f(x^i)\|^2. \tag{20}$$

We immediately conclude that $\{f(x^i)\}$ is a monotonically nonincreasing sequence. If this sequence is bounded from below then it converges. In the latter case, $\{f(x^i) - f(x^{i+1})\} \to 0$ and consequently $\{\nabla f(x^i)\} \to 0$. Hence, by continuity of $\nabla f(\cdot)$, if there exist accumulation points of $\{x^i\}$, all of them are stationary points of $f(\cdot)$.

Suppose now the sequence $\{x^i\}$ is bounded. The preceding discussion immediately implies that the set of stationary points of $f(\cdot)$ is nonempty. Denote by $X_s$ its subset that contains accumulation points of $\{x^i\}$. Clearly, $\{d(x^i, X_s)\} \to 0$. Hence $x^i \in X_s + \epsilon B$ for $i$ sufficiently large, say $i \geq i_0$. Suppose $X_s$ is weakly sharp of order 2. Then (19) is satisfied for all $i \geq i_0$.

By Lemma 1,

$$\begin{aligned} f(x^i) - f([x^i]^+) &\leq \langle \nabla f(x^i), x^i - [x^i]^+\rangle + \frac{L}{2}\|x^i - [x^i]^+\|^2 \\ &= \langle \nabla f(x^i), x^i - [x^i]^+\rangle + \frac{L}{2}d(x^i, X_s)^2, \end{aligned}$$

where $[\cdot]^+$ denotes the orthogonal projection onto $X_s$. Hence for all $i \geq i_0$, by (19), we obtain

$$\begin{aligned} \langle \nabla f(x^i), x^i - [x^i]^+\rangle &\geq f(x^i) - f([x^i]^+) - \frac{L}{2}d(x^i, X_s)^2 \\ &\geq f(x^i) - f([x^i]^+) - \frac{L}{2\rho}(f(x^i) - f([x^i]^+)) \\ &= \left(1 - \frac{L}{2\rho}\right)(f(x^i) - f([x^i]^+)), \end{aligned} \tag{21}$$

By the Cauchy-Schwartz inequality and (19), we further obtain

$$\begin{aligned} \|\nabla f(x^i)\| d(x^i, X_s) &\geq \langle \nabla f(x^i), x^i - [x^i]^+\rangle \\ &\geq \left(1 - \frac{L}{2\rho}\right)(f(x^i) - f([x^i]^+)) \\ &\geq \rho\left(1 - \frac{L}{2\rho}\right) d(x^i, X_s)^2. \end{aligned}$$

Hence

$$\|\nabla f(x^i)\| \geq \rho\left(1 - \frac{L}{2\rho}\right) d(x^i, X_s). \tag{22}$$

By (22), the Cauchy-Schwartz inequality and (21) we have

$$\begin{aligned}\|\nabla f(x^i)\|^2 &\geq \|\nabla f(x^i)\|\rho\left(1-\frac{L}{2\rho}\right)d(x^i, X_s)\\ &\geq \rho\left(1-\frac{L}{2\rho}\right)\langle\nabla f(x^i), x^i-[x^i]^+\rangle\\ &\geq \rho\left(1-\frac{L}{2\rho}\right)^2(f(x^i)-f([x^i]^+))\end{aligned} \tag{23}$$

Combining (20) and (23) gives

$$\begin{aligned}f(x^i)-f(x^{i+1}) &\geq \gamma(1-\lambda)\|\nabla f(x^i)\|^2\\ &\geq \gamma\rho(1-\lambda)\left(1-\frac{L}{2\rho}\right)^2(f(x^i)-f([x^i]^+)).\end{aligned}$$

Rearranging terms, we obtain

$$f(x^{i+1})-f([x^i]^+) \leq \left(1-\gamma\rho(1-\lambda)\left(1-\frac{L}{2\rho}\right)^2\right)(f(x^i)-f([x^i]^+)).$$

We already established that the sequence $\{f(x^i)\}$ converges. Let $\bar{f} := \lim_{i\to\infty} f(x^i)$. Since all accumulation points of the sequence $\{x^i\}$ belong to the set $X_s$ and $\{x^i\}$ is bounded, it follows that accumulation points of the sequences $\{x^i\}$ and $\{[x^i]^+\}$ are the same. Therefore, by continuity of $f(\cdot)$, we obtain

$$\lim_{i\to\infty} f([x^i]^+) = \lim_{i\to\infty} f(x^i) = \bar{f}.$$

Because $X_s$ is a set of (local) minima and $[x^i]^+ \in X_s$, it must be the case that $f([x^i]^+) = \bar{f}$ for all $i$ sufficiently large, say $i \geq i_1$. Therefore, for $i \geq \max\{i_0, i_1\}$, we obtain

$$f(x^{i+1})-\bar{f} \leq \left(1-\gamma\rho(1-\lambda)\left(1-\frac{L}{2\rho}\right)^2\right)(f(x^i)-\bar{f}).$$

Hence the sequence $\{f(x^i)\}$ converges $Q$-linearly. By (20), the sequence $\{\nabla f(x^i)\}$ converges $R$-linearly to zero. Also, by (19), the sequence $\{d(x^i, X_s)\}$ converges $R$-linearly to zero. □

*Remark.* At this time, it is an open question whether the sequence $\{x^i\}$ itself converges linearly under the assumptions of Theorem 3. Note that if we had a serial gradient descent method where

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

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 of $\{\nabla f(x^i)\}$. The difficulty with the parallel algorithm is that we cannot explicitly relate $\{\nabla f(x^i)\}$ to $\{x^{i+1} - x^i\}$.

Careful re-examination of the proof of Theorem 3 shows that at the $(i+1)$-st iteration every parallel processor decreases the objective function $f(\cdot)$ of the original problem by a factor of $\|\nabla f(x^i)\|^2$ (this at least is true under our assumptions on the directions $d_t^i,\ t = 1, \ldots, p$). Hence if the processors were to proceed with updating their base points completely independently without using any information from the other processors, we could still guarantee the same convergence results for each of the $p$ sequences of iterates generated. Of course, this approach essentially yields $p$ serial processes and therefore is a theoretical extreme. This observation is however of significance because it implies that we are allowed a lot of flexibility in devising PVD algorithms and, in particular, in defining the points of synchronization.

*Remark.* A practically important example of weak sharp minima of order 2 is provided by the implicit Lagrangian reformulation [12] of the nonlinear complementarity problem.

Consider the following nonlinear complementarity problem [3, 15] (NCP) 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 $F : \Re^n \to \Re^n$ is a continuously differentiable mapping. In [12] it was established that the NCP can be solved via (smooth) unconstrained minimization of the following implicit Lagrangian function:

$$\begin{aligned} M(x,\alpha) := {} & 2\alpha\langle x, F(x)\rangle + \|[x - \alpha F(x)]^+\|^2 - \|x\|^2 \\ & + \|[F(x) - \alpha x]^+\|^2 - \|F(x)\|^2, \end{aligned}$$

where $\alpha > 1$ 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.

In [8] it was established that

$$2(\alpha - 1)\|r(x)\|^2 \leq M(x,\alpha) \leq 2\alpha(\alpha - 1)\|r(x)\|^2, \quad \forall x \in \Re^n,$$

where $r(x) := x - [x - F(x)]^+$. Therefore the set of solutions $X_s$ of the NCP is a set of weak sharp minima of order 2 for the implicit Lagrangian whenever the projection-type error bound holds:

$$d(x, X_s) \leq \rho\|r(x)\| \quad \forall x \text{ with } \|r(x)\| \leq \epsilon,$$

where $\rho$ and $\epsilon$ are positive constants (independent of $x$). This error bound is known to hold when $F(\cdot)$ is affine (see [9, 17]) or $F(\cdot)$ has certain strong monotonicity structure (see [21], Theorem 2). Moreover, under additional assumptions on $F(\cdot)$, this condition holds globally with $\epsilon = \infty$ (see [7, 8, 11, 14]).

Therefore our analysis shows that certain unconstrained minimization techniques applied to minimizing the implicit Lagrangian attain linear rate of convergence (under certain conditions). This is an interesting result given that the implicit Lagrangian is not known to be strongly convex in any neighborhood of its zero minima.

## 4. Concluding remarks

New parallel variable distribution algorithms with inexact subproblem solution and with a certain natural sufficient descent condition imposed on the parallel subproblems were proposed and analyzed. The modified algorithms present a flexible framework and make it easier to achieve good load balancing among the parallel processors. New and improved linear convergence results were derived for strongly convex problems and problems with weak sharp minima of order 2. A study of partially asynchronous distributed algorithms [1] that make use of PVD approach can be an interesting subject of future research.

## References

1. D.P. Bertsekas and J.N. Tsitsiklis, Parallel and Distributed Computation, Prentice-Hall, Inc.: Englewood Cliffs, New Jersey, 1989.
2. J.V. Burke and M.C. Ferris, "Weak sharp minima in mathematical programming," SIAM Journal on Control and Optimization, vol. 31, no. 5, pp. 1340–1359, 1993.
3. R.W. Cottle, F. Giannessi, and J.-L. Lions (eds.), Variational Inequalities and Complementarity Problems: Theory and Applications, Wiley: New York, 1980.
4. M.C. Ferris and O.L. Mangasarian, "Parallel variable distribution," SIAM Journal on Optimization, vol. 4, no. 4, pp. 815–832, 1994.
5. L. Grippo, F. Lampariello, and S. Lucidi, "A nonmonotone line search technique for Newton's method," SIAM Journal of Numerical Analysis, vol. 23, pp. 707–716, 1986.
6. L. Grippo, F. Lampariello, and S. Lucidi, "A class of nonmonotone stabilization methods in unconstrained optimization," Numerische Mathematik, vol. 59, pp. 779–805, 1991.
7. X.-D. Luo and P. Tseng, "On global projection-type error bound for the linear complementarity problem," Linear Algebra and Its Applications (to appear).
8. Z.-Q. Luo, O.L. Mangasarian, J. Ren, and M.V. Solodov, "New error bounds for the linear complementarity problem," Mathematics of Operations Research, vol. 19, pp. 880–892, 1994.
9. Z.-Q. Luo and P. Tseng, "Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem," SIAM Journal on Optimization, vol. 2, pp. 43–54, 1992.
10. O.L. Mangasarian, "Parallel gradient distribution in unconstrained optimization," SIAM Journal on Control and Optimization, vol. 33, no. 6, pp. 1916–1925, 1995.
11. O.L. Mangasarian and J. Ren, "New improved error bounds for the linear complementarity problem," Mathematical Programming, vol. 66, pp. 241–255, 1994.
12. O.L. Mangasarian and M.V. Solodov, "Nonlinear complementarity as unconstrained and constrained minimization," Mathematical Programming, vol. 62, pp. 277–297, 1993.
13. J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, 1970.

14. J.-S. Pang, "A posteriori error bounds for the linearly-constrained variational inequality problem," Mathematics of Operations Research, vol. 12, pp. 474–484, 1987.
15. J.-S. Pang, "Complementarity problems," in R. Horst and P. Pardalos (eds.), Handbook of Global Optimization, Kluwer Academic Publishers: Boston, Massachusetts, 1995.
16. B.T. Polyak, "Introduction to Optimization," Optimization Software, Inc.: Publications Division, New York, 1987.
17. S.M. Robinson, "Some continuity properties of polyhedral multifunctions," Mathematical Programming Study, vol. 14, pp. 206–214, 1981.
18. M.V. Solodov and S.K. Zavriev, "Error-stability properties of generalized gradient-type algorithms," Mathematical Programming Technical Report 94-05, Computer Science Department, University of Wisconsin, 1210 West Dayton Street, Madison, Wisconsin 53706, U.S.A., June 1994 (revised July 1995).
19. M.V. Solodov, "On the convergence of constrained parallel variable distribution algorithms," Technical Report B-094, Instituto de Matematica Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botanico, Rio de Janeiro, RJ, CEP 22460, Brazil, Oct. 1995. SIAM Journal on Optimization, accepted for publication.
20. P. Tseng, "Dual coordinate ascent methods for non-strictly convex minimization," Mathematical Programming, vol. 59, pp. 231–248, 1993.
21. P. Tseng, On linear convergence of iterative methods for the variational inequality problem," Journal of Computational and Applied Mathematics, vol. 60, pp. 237–252, 1995.