JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 109, No. 3, pp. 475–494, JUNE 2001

Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization[1]

P. TSENG[2]

Communicated by O. L. Mangasarian

Abstract. We study the convergence properties of a (block) coordinate descent method applied to minimize a nondifferentiable (nonconvex) function $f(x_1, \ldots, x_N)$ with certain separability and regularity properties. Assuming that f is continuous on a compact level set, the subsequence convergence of the iterates to a stationary point is shown when either f is pseudoconvex in every pair of coordinate blocks from among $N-1$ coordinate blocks or f has at most one minimum in each of $N-2$ coordinate blocks. If f is quasiconvex and hemivariate in every coordinate block, then the assumptions of continuity of f and compactness of the level set may be relaxed further. These results are applied to derive new (and old) convergence results for the proximal minimization algorithm, an algorithm of Arimoto and Blahut, and an algorithm of Han. They are applied also to a problem of blind source separation.

Key Words. Block coordinate descent, nondifferentiable minimization, stationary point, Gauss–Seidel method, convergence, quasiconvex functions, pseudoconvex functions.

1. Introduction

A popular method for minimizing a real-valued continuously differentiable function f of n real variables, subject to bound constraints, is the (block) coordinate descent method. In this method, the coordinates are partitioned into N blocks and, at each iteration, f is minimized with respect to one of the coordinate blocks while the other coordinates are held fixed. This method, which is related closely to the Gauss–Seidel and SOR methods for equation solving (Ref. 1), was studied early by Hildreth (Ref. 2) and Warga (Ref. 3), and is described in various books on optimization (Refs. 1 and 4–

[1]This work was partially supported by the National Science Foundation Grant CCR-9731273.
[2]Professor, Department of Mathematics, University of Washington, Seattle, Washington.

0022-3239/01/0600-0475$19.50/0

10). Its applications include channel capacity computation (Refs. 11–12), image reconstruction (Ref. 7), dynamic programming (Refs. 13–15), and flow routing (Ref. 16). It may be applied also to the dual of a linearly constrained, strictly convex program to obtain various decomposition methods (see Refs. 6–7, 17–22, and references therein) and parallel SOR methods (Ref. 23).

Convergence of the (block) coordinate descent method requires typically that f be strictly convex (or quasiconvex or hemivariate) differentiable and, taking into account the bound constraints, has bounded level sets (e.g., Refs. 3–4 and 24–25). Zadeh (Ref. 26; also see Ref. 27) relaxed the strict convexity assumption to pseudoconvexity, which allows f to have a nonunique minimum along coordinate directions. For certain classes of convex functions, the level sets need not be bounded (see Refs. 2, 6–7, 17, 19–22, and references therein). If f is not (pseudo)convex, then an example of Powell (Ref. 28) shows that the method may cycle without approaching any stationary point of f. Nonetheless, convergence can be shown for special cases of non(pseudo)convex f, as when f is quadratic (Ref. 29), or f is strictly pseudoconvex in each of $N-2$ coordinate blocks (Ref. 27), or f has unique minimum in each coordinate block (Ref. 8, p. 159). If f is not differentiable, the coordinate descent method may get stuck at a nonstationary point even when f is convex (e.g., Ref. 4, p. 94). For this reason, it is perceived generally that the method is unsuitable when f is nondifferentiable. However, an exception occurs when the nondifferentiable part of f is separable. Such a structure for f was considered first by Auslender (Ref. 4, p. 94) in the case where f is strongly convex. This structure is implicit in a decomposition method and projection method of Han (Refs. 18, 30), for which f is the convex dual functional associated with a certain linearly constrained convex program (see Ref. 22 for detailed discussions). This structure arises also in least-square problems where an l_1-penalty is placed on a subset of the parameters in order to minimize the support (see Refs. 31–33 and references therein).

Motivated by the preceding works, we consider in this paper the nondifferentiable (nonconvex) case where the nondifferentiable part of f is separable. Specifically, we assume that f has the following special form:

$$f(x_1, \ldots, x_N) = f_0(x_1, \ldots, x_N) + \sum_{k=1}^{N} f_k(x_k), \tag{1}$$

for some $f_0: \Re^{n_1 + \cdots + n_N} \mapsto \Re \cup \{\infty\}$ and some $f_k: \Re^{n_k} \mapsto \Re \cup \{\infty\}$, $k = 1, \ldots, N$. Here, $N, n_1, \ldots, n_N$ are positive integers. We assume that f is proper, i.e., $f \not\equiv \infty$. We will refer to each x_k, $k = 1, \ldots, N$, as a coordinate block of $x = (x_1, \ldots, x_N)$. We will show that each cluster point of the iterates generated by the (block) coordinate descent method is a stationary point of

f, provided that f_0 has a certain smoothness property (see Lemma 3.1), f is continuous on a compact level set, and either f is pseudoconvex in every pair of coordinate blocks from among $N-1$ coordinate blocks, or f has at most one minimum in each of $N-2$ coordinate blocks (see Theorem 4.1). If f is quasiconvex and hemivariate in every coordinate block, then the assumptions of continuity of f and compactness of the level set may be relaxed further (see Proposition 5.1). These results unify and extend some previous results in Refs. 4, 6, 8, 26–27. For example, previous results assumed that f is pseudoconvex and that $f_1, \ldots, f_N$ are indicator functions for closed convex sets, whereas we assume only that f is pseudoconvex in every pair of coordinate blocks from among $N-1$ coordinate blocks, with no additional assumption made on $f_1, \ldots, f_N$. Previous results also did not consider the case where f is not continuous on its effective domain. Lastly, we apply our results to derive new (and old) convergence results for the proximal minimization algorithm, an algorithm of Arimoto and Blahut (Refs. 11–12), and an algorithm of Han (Ref. 30); see Examples 6.1–6.3. We also apply them to a problem of blind source separation described in Refs. 31, 33; see Example 6.4.

In our notation, $\Re^m$ denotes the space of m-dimensional real column vector. For any x, $y \in \Re^m$, we denote by $\langle x, y\rangle$ the Euclidean inner product of x, y and by $\|x\|$ the Euclidean norm of x, i.e.,

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

For any set $S \subseteq \Re^m$, we denote by $\text{int}(S)$ the interior of S and denote

$$\text{bdry}(S) = S \backslash \text{int}(S).$$

For any $h: \Re^m \mapsto \Re \cup \{\infty\}$, we denote by $\text{dom}\, h$ the effective domain of h, i.e.,

$$\text{dom}\, h = \{x \in \Re^m \mid h(x) < \infty\}.$$

For any $x \in \text{dom}\, h$ and any $d \in \Re^m$, we denote the (lower) directional derivative of h at x in the direction d by

$$h'(x; d) = \liminf_{\lambda \downarrow 0} [h(x + \lambda d) - h(x)]/\lambda.$$

We say that h is quasiconvex if

$$h(x + \lambda d) \leq \max\{h(x), h(x + d)\}, \qquad \text{for all } x, d \text{ and } \lambda \in [0, 1];$$

h is pseudoconvex if

$$h(x + d) \geq h(x), \qquad \text{whenever } x \in \text{dom}\, h \text{ and } h'(x; d) \geq 0;$$

see Ref. 34, p. 146; and h is hemivariate if h is not constant on any line segment belonging to $\text{dom}\, h$ (Ref. 1). For any nonempty $I \subseteq \{1, \ldots, m\}$, we

say that $h(x_1, \ldots, x_m)$ is pseudoconvex [respectively, has at most one minimum point].

2. Block Coordinate Descent Method

We describe formally the block coordinate descent (BCD) method below.

BCD Method.

Initialization. Choose any $x^0 = (x_1^0, \ldots, x_N^0) \in \operatorname{dom} f$.

Iteration $r+1$, $r \geq 0$. Given $x^r = (x_1^r, \ldots, x_N^r) \in \operatorname{dom} f$, choose an index $s \in \{1, \ldots, N\}$ and compute a new iterate

$$x^{r+1} = (x_1^{r+1}, \ldots, x_N^{r+1}) \in \operatorname{dom} f$$

satisfying

$$x_s^{r+1} \in \arg\min_{x_s} f(x_1^r, \ldots, x_{s-1}^r, x_s, x_{s+1}^r, \ldots, x_N^r), \tag{2}$$

$$x_j^{r+1} = x_j^r, \qquad \forall j \neq s. \tag{3}$$

We note that the minimization in (2) is attained if

$$X^0 = \{x: f(x) \leq f(x^0)\}$$

is bounded and f is lower semicontinuous (lsc) on X^0, so X^0 is compact (Ref. 35). Alternatively, this minimization is attained if f is convex, has a minimum point, and is hemivariate in each coordinate block (but the level sets of f need not be bounded). To ensure convergence, we need further that each coordinate block is chosen sufficiently often in the method. In particular, we will choose the coordinate blocks according to the following rule (see, e.g., Refs. 7–8, 21, 25).

Essentially Cyclic Rule. There exists a constant $T \geq N$ such that every index $s \in \{1, \ldots, N\}$ is chosen at least once between the rth iteration and the $(r+T-1)$th iteration, for all r.

A well-known special case of this rule, for which $T = N$, is given below.

Cyclic Rule. Choose $s = k$ at iterations k, $k+N$, $k+2N, \ldots$, for $k = 1, \ldots, N$.

3. Stationary Points of f

We say that z is a stationary point of f if $z \in \operatorname{dom} f$ and

$$f'(z; d) \geq 0, \qquad \forall d.$$

We say that z is a coordinatewise minimum point of f if $z \in \operatorname{dom} f$ and

$$f(z + (0, \ldots, d_k, \ldots, 0)) \geq f(z), \qquad \forall d_k \in \Re^{n_k}, \tag{4}$$

for all $k = 1, \ldots, N$. Here and throughout, we denote by $(0, \ldots, d_k, \ldots, 0)$ the vector in $\Re^{n_1 + \cdots + n_N}$ whose kth coordinate block is d_k and whose other coordinates are zero. We say that f is regular at $z \in \operatorname{dom} f$ if

$$f'(z; d) \geq 0, \qquad \forall d = (d_1, \ldots, d_N),$$
$$\text{such that } f'(z; (0, \ldots, d_k, \ldots, 0)) \geq 0, \qquad k = 1, \ldots, N. \tag{5}$$

This notion of regularity is weaker than that used by Auslender (Ref. 4, p. 93), which entails

$$f'(z; d) = \sum_{k=1}^{N} f'(z; (0, \ldots, d_k, \ldots, 0)), \qquad \text{for all } d = (d_1, \ldots, d_N).$$

For example, the function

$$f(x_1, x_2) = \phi(x_1, x_2) + \phi(-x_1, x_2) + \phi(x_1, -x_2) + \phi(-x_1, -x_2),$$

where

$$\phi(a, b) = \max\{0, a + b - \sqrt{a^2 + b^2}\},$$

is regular at $z = (0, 0)$ in the sense of (5), but is not regular in the sense of Ref. 4, p. 93.

Since (4) implies

$$f'(z; (0, \ldots, d_k, \ldots, 0)) \geq 0, \qquad \text{for all } d_k,$$

it follows that a coordinatewise minimum point z of f is a stationary point of f whenever f is regular at z. To ensure regularity of f at z, we consider one of the following smoothness assumptions on f_0:

(A1) $\operatorname{dom} f_0$ is open and f_0 is Gâteaux-differentiable on $\operatorname{dom} f_0$.

(A2) f_0 is Gâteaux-differentiable on $\operatorname{int}(\operatorname{dom} f_0)$ and, for every $z \in \operatorname{dom} f \cap \operatorname{bdry}(\operatorname{dom} f_0)$, there exist $k \in \{1, \ldots, N\}$ and $d_k \in \Re^{n_k}$ such that $f(z + (0, \ldots, d_k, \ldots, 0)) < f(z)$.

Assumption A1 was considered essentially by Auslender (Ref. 4, Example 2 on p. 94). In contrast to Assumption A1, Assumption A2 allows $\operatorname{dom} f_0$ to include boundary points. We will see an application (Example 6.2) where A2 holds but not A1.

Lemma 3.1. Under A1, f is regular at each $z \in \operatorname{dom} f$. Under A2, f is regular at each coordinatewise minimum point z of f.

Proof. Under A1, if $z = (z_1, \ldots, z_N) \in \operatorname{dom} f$, then $z \in \operatorname{dom} f_0$. Under A2, if $z = (z_1, \ldots, z_N)$ is a coordinatewise minimum point of f, then $z \notin \operatorname{bdry}(\operatorname{dom} f_0)$, so $z \in \operatorname{int}(\operatorname{dom} f_0)$. Thus, under either A1 or A2, f_0 is Gâteaux-differentiable at z. Fix any $d = (d_1, \ldots, d_N)$ such that

$$f'(z; (0, \ldots, d_k, \ldots, 0)) \geq 0, \qquad k = 1, \ldots, N.$$

Then,

$$\begin{aligned}
f'(z; d) &= \langle \nabla f_0(z), d \rangle + \liminf_{\lambda \downarrow 0} \sum_{k=1}^{N} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \\
&\geq \langle \nabla f_0(z), d \rangle + \sum_{k=1}^{N} \liminf_{\lambda \downarrow 0} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \\
&= \langle \nabla f_0(z), d \rangle + \sum_{k=1}^{N} f_k'(z_k; d_k) \\
&= \sum_{k=1}^{N} f'(z; (0, \ldots, d_k, \ldots, 0)) \\
&\geq 0. \qquad \square
\end{aligned}$$

4. Convergence Analysis: I

Our first convergence result unifies and extends a result of Auslender (Ref. 4, p. 95) for the nondifferentiable convex case and some results of Grippo and Sciandrone (Ref. 27), Luenberger (Ref. 8, p. 159), and Zadeh (Ref. 26) for the differentiable case. In what follows, $r \equiv (N-1) \bmod N$ means $r = N-1, 2N-1, 3N-1, \ldots$.

Theorem 4.1. Assume that the level set $X^0 = \{x : f(x) \leq f(x^0)\}$ is compact and that f is continuous on X^0. Then, the sequence $\{x^r = (x_1^r, \ldots, x_N^r)\}_{r = 0, 1, \ldots}$ generated by the BCD method using the essentially cyclic rule is defined and bounded. Moreover, the following statements

hold:

(a) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N\}$, and if f is regular at every $x \in X^0$, then every cluster point of $\{x^r\}$ is a stationary point of f.

(b) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N-1\}$, if f is regular at every $x \in X^0$, and if the cyclic rule is used, then every cluster point of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a stationary point of f.

(c) If $f(x_1, \ldots, x_N)$ has at most one minimum in x_k for $k = 2, \ldots, N-1$, and if the cyclic rule is used, then every cluster point z of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a coordinatewise minimum point of f. In addition, if f is regular at z, then z is a stationary point of f.

Proof. Since X^0 is compact, an induction argument on r shows that x^{r+1} is defined, $f(x^{r+1}) \leq f(x^r)$, and $x^{r+1} \in X^0$ for all $r = 0, 1, \ldots$. Thus, $\{x^r\}$ is bounded. Consider any subsequence $\{x^r\}_{r \in \mathscr{R}}$, with $\mathscr{R} \subseteq \{0, 1, \ldots\}$, converging to some z. For each $j \in \{1, \ldots, T\}$, $\{x^{r-T+1+j}\}_{r \in \mathscr{R}}$ is bounded, so by passing to a subsequence, if necessary, we can assume that

$$\{x^{r-T+1+j}\}_{r \in \mathscr{R}} \text{converges to some } z^j = (z^j_1, \ldots, z^j_N), \qquad j = 1, \ldots, T.$$

Thus,

$$z^{T-1} = z.$$

Since $\{f(x^r)\}$ converges monotonically and f is continuous on X^0, we obtain that

$$f(x^0) \geq \lim_{r \to \infty} f(x^r) = f(z^1) = \cdots = f(z^T). \tag{6}$$

By further passing to a subsequence, if necessary, we can assume that the index s chosen at iteration $r - T + 1 + j$, $j \in \{1, \ldots, T\}$, is the same for all $r \in \mathscr{R}$, which we denote by s^j.

For each $j \in \{1, \ldots, T\}$, since s^j is chosen at iteration $r - T + 1 + j$ for $r \in \mathscr{R}$, then (2) and (3) yield

$$\begin{aligned} &f(x^{r-T+1+j}) \leq f(x^{r-T+1+j} + (0, \ldots, d_{s^j}, \ldots, 0)), && \forall d_{s^j}, j = 1, \ldots, T, \\ &x^{r-T+1+j}_k = x^{r-T+j}_k, && \forall k \neq s^j, j = 2, \ldots, T. \end{aligned}$$

Then, the continuity of f on X^0 yields in the limit that

$$\begin{aligned} &f(z^j) \leq f(z^j + (0, \ldots, d_{s^j}, \ldots, 0)), && \forall d_{s^j}, j = 1, \ldots, T, \\ &z^j_k = z^{j-1}_k, \qquad \forall k \neq s^j, && j = 2, \ldots, T. \end{aligned} \tag{7}$$

Then, (6) and (7) yield

$$f(z^{j-1}) \leq f(z^{j-1} + (0, \ldots, d_{s^j}, \ldots, 0)), \qquad \forall d_{s^j}, j = 2, \ldots, T. \tag{8}$$

(a), (b) Suppose that f is regular at every $x \in X^0$ and that $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{s^1\} \cup \cdots \cup \{s^{T-1}\}$. This holds under the assumption (a) or under the assumption (b), with $\{x^r\}_{r \in \mathcal{R}}$ being any convergent subsequence of $\{x^r\}_{r \equiv (N-1) \bmod N}$. We claim that, for $j = 1, \ldots, T-1$,

$$f(z^j) \leq f(z^j + (0, \ldots, d_k, \ldots, 0)), \qquad \forall d_k, \forall k = s^1, \ldots, s^j. \tag{9}$$

By (7), (9) holds for $j = 1$. Suppose that (9) holds for $j = 1, \ldots, l-1$ for some $l \in \{2, \ldots, T-1\}$. We show that (9) holds for $j = l$. From (8), we have that

$$f(z^{l-1}) \leq f(z^{l-1} + (0, \ldots, d_{s^l}, \ldots, 0)), \qquad \forall d_{s^l},$$

implying

$$f'(z^{l-1}; (0, \ldots, z^l_{s^l} - z^{l-1}_{s^l}, \ldots, 0)) \geq 0.$$

Also, since (9) holds for $j = l-1$, we have that, for each $k = s^1, \ldots, s^{l-1}$,

$$f'(z^{l-1}; (0, \ldots, d_k, \ldots, 0)) \geq 0, \qquad \forall d_k.$$

Since by (6) $z^{l-1} \in X^0$, so f is regular at z^{l-1}, the above two relations imply

$$f'(z^{l-1}; (0, \ldots, d_k, \ldots, 0) + (0, \ldots, z^l_{s^l} - z^{l-1}_{s^l}, \ldots, 0)) \geq 0, \qquad \forall d_k.$$

Since f is pseudoconvex in (x_k, x_{s^l}), this yields [also using $z^l = z^{l-1} + (0, \ldots, z^l_{s^l} - z^{l-1}_{s^l}, \ldots, 0)$] for $k = s^1, \ldots, s^{l-1}$ that

$$f(z^l + (0, \ldots, d_k, \ldots, 0)) \geq f(z^{l-1}) = f(z^l), \qquad \forall d_k.$$

Since we have also that (7) holds with $j = l$, we see that (9) holds for $j = l$. By induction, (9) holds for all $j = 1, \ldots, T-1$.

Since $z^{T-1} = z$ and (9) holds for $j = T-1$, then (4) holds for $k = s^1, \ldots, s^{T-1}$. Since $z^{T-1} = z$ and (8) holds (in particular, for $j = T$), then (4) holds for $k = s^T$ also. Since

$$\{1, \ldots, N\} = \{s^1\} \cup \cdots \cup \{s^T\},$$

this implies that z is a coordinatewise minimum point of f. Since f is regular at z, then z is in fact a stationary point of f.

(c) Suppose that $f(x_1, \ldots, x_N)$ has at most one minimum in x_k for $k = s^2, \ldots, s^{T-1}$. This holds under the assumption (c), with $\{x^r\}_{r \in \mathcal{R}}$ being any convergent subsequence of $\{x^r\}_{r \equiv (N-1) \bmod N}$. For each $j = 2, \ldots, T-1$, since

(7) and (8) hold, then the function

$$d_{s^j} \mapsto f(z^j + (0, \ldots, d_{s^j}, \ldots, 0))$$

attains its minimum at both $d_{s^j} = 0$ and $d_{s^j} = z^{j-1}_{s^j} - z^j_{s^j}$. By assumption, the minimum point is unique, implying $0 = z^{j-1}_{s^j} - z^j_{s^j}$, or equivalently, $z^{j-1} = z^j$. Thus, $z^1 = z^2 = \cdots = z^{T-1} = z$ and (7) yields that (4) holds for $k = s^1, \ldots, s^{T-1}$. Since $z^{T-1} = z$ and (8) holds (in particular, for $j = T$), then (4) holds for $k = s^T$ also. Since

$$\{1, \ldots, N\} = \{s^1\} \cup \cdots \cup \{s^T\},$$

this implies that z is a coordinatewise minimum point of f. If f is regular at z, then z is also a stationary point of f. □

Notice that, if f is pseudoconvex, then f is pseudoconvex in (x_k, x_i) for every i, $k \in \{1, \ldots, N\}$; if f is quasiconvex and hemivariate in x_k, then f has at most one minimum in x_k. The converses do not hold. For example, the 2-variable Rosenbrock function has a unique minimum point but is not quasiconvex. The following 3-variable quadratic function

$$f(x_1, x_2, x_3) = (1/2)x_1^2 + (1/2)x_2^2 + (1/2)x_3^2 + x_1x_3 + x_2x_3 - x_1x_2$$

is convex in every pair of variables, but is not pseudoconvex. In particular, for $x = (0, 0, 1/2)$ and $d = (1, 1, -1)$, we have $f'(x; d) = 1/2 \geq 0$, while $f(x + d) = -7/8 < f(x) = 1/8$. This example generalizes to any quadratic function

$$f(x) = \langle x, Qx \rangle.$$

where $Q \in \mathscr{R}^{N \times N}$ is symmetric, not positive semidefinite, but whose 2×2 principal submatrices are positie semidefinite. Then, for any d satisfying $\langle d, Qd \rangle < 0$ and any x satisfying

$$0 \leq \langle x, Qd \rangle < -(1/2)\langle d, Qd \rangle,$$

we have that

$$f'(x; d) \geq 0, \qquad \text{while } f(x + d) < f(x).$$

Thus, parts (a) and (c) of Theorem 4.1 may be viewed as extensions of two results of Grippo and Sciandrone (Ref. 27, Propositions 5.2, 5.3) for the case of f_0 being continuously differentiable and each f_k being the indicator function of some closed convex set. In turn, the first of these results extended a result of Zadeh (Ref. 26) for which $f_k \equiv 0$ for all k. Part (b) makes a less restrictive assumption on f than part (a), though its assumption on the BCD method is more restrictive. Part (b) is sharp in the sense that it is false if instead we assume that f is convex in every coordinate block. This

is because the Powell 3-variable example (Ref. 28) is convex in each variable; see Ref. 27, Section 6 for further discussions of the example. We will see an application (Example 6.4) in which part (b) applies but not part (a) nor (c).

5. Convergence Analysis: II

The convergence analysis of the previous section assumes f to be continuous on a bounded level set and makes no use of the special structure (1) of f. In this section, we show that this assumption can be relaxed by exploiting the special structure (1), provided that f is quasiconvex and hemivariate in each coordinate block. More precisely, we will make the following assumptions on $f, f_0, f_1, \ldots, f_N$:

(B1) f_0 is continuous on $\operatorname{dom} f_0$.

(B2) For each $k \in \{1, \ldots, N\}$ and $(x_j)_{j \neq k}$, the function $x_k \mapsto f(x_1, \ldots, x_N)$ is quasiconvex and hemivariate.

(B3) $f_0, f_1, \ldots, f_N$ are lsc.

We will see some applications (Ref. 6, Section 3.4.3 and Examples 6.1–6.3) for which f satisfies this weaker assumption although it is not strictly convex. In addition, we will make one of the following technical assumptions on f_0:

(C1) $\operatorname{dom} f_0$ is open and f_0 tends to ∞ at every boundary point of $\operatorname{dom} f_0$.

(C2) $\operatorname{dom} f_0 = Y_1 \times \cdots \times Y_N$, for some $Y_k \subseteq \mathscr{R}^{n_k}$, $k = 1, \ldots, N$.

In contrast to Assumption C1, Assumption C2 allows f_0 to have a finite value on $\operatorname{bdry}(\operatorname{dom} f)$. We will see in Example 6.2 a nonseparable function f_0 that satisfies Assumptions B1–B3 and C2, but not C1. We show below that Assumptions B1–B3, together with either Assumption C1 or C2, ensure that every cluster point of the iterates generated by the BCD method is a coordinate minimum point of f. The proof of this result is patterned after an argument given by Bertsekas and Tsitsiklis (Ref. 6, pp. 220–221; also see Ref. 27), but is complicated by the fact that f is not necessarily differentiable (or even continuous) on its effective domain.

Proposition 5.1. Suppose that $f, f_0, f_1, \ldots, f_N$ satisfy Assumptions B1–B3 and that f_0 satisfies either Assumption C1 or C2. Also, assume that the sequence $\{x^r = (x_1^r, \ldots, x_N^r)\}_{r=0,1,\ldots}$ generated by the BCD method using the essentially cyclic rule is defined. Then, either $\{f(x^r)\} \downarrow -\infty$, or else every cluster point $z = (z_1, \ldots, z_N)$ is a coordinatewise minimum point of f.

Proof. Since $f(x^0) < \infty$ and $f(x^{r+1}) \leq f(x^r)$ for all r, then either $\{f(x^r)\} \downarrow -\infty$, or else $\{f(x^r)\}$ converges to some limit and $\{f(x^{r+1}) - f(x^r)\} \to 0$. Consider the latter case and let z be any cluster point of $\{x^r\}$. Since f is lsc by Assumption B3, we have

$$f(z) \leq \lim_{r \to \infty} f(x^r) < \infty,$$

so $z \in \operatorname{dom} f$. We show below that z satisfies (4) for $k = 1, \ldots, N$.

First, we claim that, for any infinite subsequence

$$\{x^r\}_{r \in \mathscr{R}} \to z, \tag{10}$$

with $\mathscr{R} \subseteq \{0, 1, \ldots\}$, there holds that

$$(x^{r+1}\}_{r \in \mathscr{R}} \to z. \tag{11}$$

We prove this by contradiction. Suppose that this were not true. Then, there exists an infinite subsequence $\mathscr{R}'$ of $\mathscr{R}$ and a scalar $\epsilon > 0$ such that

$$\|x^{r+1} - x^r\| \geq \epsilon, \qquad \text{for all } r \in \mathscr{R}'.$$

By further passing to a subsequence, if necessary, we can assume that there is some nonzero vector d for which

$$\{(x^{r+1} - x^r)/\|x^{r+1} - x^r\|\}_{r \in \mathscr{R}'} \to d, \tag{12}$$

and that the same coordinate block, say x_s, is chosen t the $(r+1)$st iteration for all $r \in \mathscr{R}'$. Moreover, (10) implies that $\{f_0(x^r)\}_{r \in \mathscr{R}}$ and $\{f_k(x_k^r)\}_{r \in \mathscr{R}}$, $k = 1, \ldots, N$, are bounded from below, which together with the convergence of $\{f(x^r)\} = \{f_0(x^r) + \sum_{k=1}^N f_k(x_k^r)\}$ implies that $\{f_0(x^r)\}_{r \in \mathscr{R}}$ and $\{f_k(x_k^r)\}_{r \in \mathscr{R}}$, $k = 1, \ldots, N$, are bounded. Hence, by further passing to a subsequence, if necessary, we can assume that there is some scalar θ for which

$$\{f_0(x^r) + f_s(x_s^r)\}_{r \in \mathscr{R}'} \to \theta. \tag{13}$$

Fix any $\lambda \in [0, \epsilon]$. Let

$$\hat{z} = z + \lambda d, \tag{14}$$

and for each $r \in \mathscr{R}'$, let

$$\hat{x}^r = x^r + \lambda(x^{r+1} - x^r)/\|x^{r+1} - x^r\|. \tag{15}$$

Then, by (10), (12), and (14),

$$\{\hat{x}^r\}_{r \in \mathscr{R}'} \to \hat{z}. \tag{16}$$

For each $r \in \mathscr{R}'$, we see from (2) that x^{r+1} is obtained from x^r by minimizing f with respect to x_s, while the other coordinates are held fixed. Since

$$\lambda/\|x^{r+1} - x^r\| \leq \lambda/\epsilon \leq 1,$$

so $\hat{x}^r$ lies on the line segment joining x^r with x^{r+1}, this together with $f(x^{r+1}) \leq f(x^r)$ and the quasiconvexity of $x_s \mapsto f(x_1^r, \ldots, x_{s-1}^r, x_s, x_{s+1}^r, \ldots, x_N^r)$ implies

$$f(\hat{x}^r) \leq f(x^r), \qquad \forall r \in \mathscr{R}'.$$

Since f is lsc, this and (16) imply $\hat{z} \in \operatorname{dom} f$. Also, this and (1) and the observation that x^r and $\hat{x}^r$ differ only in their sth coordinate block imply

$$f_0(\hat{x}^r) + f_s(\hat{x}_s^r) \leq f_0(x^r) + f_s(x_s^r), \qquad \forall r \in \mathscr{R}'.$$

This combined with (13) yields

$$\lim_{r \to \infty,\, r \in \mathscr{R}'} \sup\{ f_0(\hat{x}^r) + f_s(\hat{x}_s^r)\} \leq \theta. \tag{17}$$

Also, since

$$\{ f(x^{r+1}) - f(x^r)\}_{r \in \mathscr{R}'} \to 0,$$

we have equivalently that

$$\{ f_0(x^{r+1}) + f_s(x_s^{r+1}) - f_0(x^r) - f_s(x_s^r)\}_{r \in \mathscr{R}'} \to 0,$$

so (13) implies

$$\{ f_0(x^{r+1}) + f_s(x_s^{r+1})\}_{r \in \mathscr{R}'} \to \theta. \tag{18}$$

Let

$$\delta = f_0(\hat{z}) + f_s(\hat{z}_s) - \theta.$$

Since f_0 and f_s are lsc, we have from (16), (17) that $\delta \leq 0$. We claim that in fact $\delta = 0$. Suppose that this were not true, so that $\delta > 0$. By (16) and the observation that, for all $r \in \mathscr{R}'$, $\hat{x}^r$ and x^r differ in only their sth coordinate block, we have

$$\{(x_1^r, \ldots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \ldots, x_N^r)\}_{r \in \mathscr{R}'} \to \hat{z}. \tag{19}$$

Moreover, the vector on the left-hand side of (19) is in $\operatorname{dom} f_0$ for all $r \in \mathscr{R}'$ sufficiently large. Since $\hat{z} \in \operatorname{dom} f_0$, this is certainly true under Assumption C1; under Assumption C2, this is also true because $x^r \in \operatorname{dom} f_0$ for all r and $\operatorname{dom} f_0$ has a product structure corresponding to the coordinate blocks. Then, (18) together with (19) and the continuity of f_0 on $\operatorname{dom} f_0$ implies that, for all $r \in \mathscr{R}'$ sufficiently large, there holds that

$$\begin{aligned} &f_0(x_1^r, \ldots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \ldots, x_N^r) + f_s(\hat{z}_s) \\ &\leq f_0(x^{r+1}) + f_s(x_s^{r+1}) + \delta/2, \end{aligned}$$

or equivalently [via (1) and the observation that x^r and x^{r+1} differ in only their sth coordinate block],

$$f(x_1^r, \ldots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \ldots, x_N^r) \leq f(x^{r+1}) + \delta/2,$$

a contradiction to the fact that x^{r+1} is obtained from x^r by minimizing f with respect to the sth coordinate block, while the other coordinates are held fixed. Hence, $\delta = 0$ and therefore

$$f_0(\hat{z}) + f_s(\hat{z}_s) = \theta.$$

Since the choice of λ was arbitrary, we obtain [also using (14)]

$$f_0(z + \lambda d) + f_s(z_s + \lambda d_s) = \theta, \qquad \forall \lambda \in [0, \epsilon],$$

where d_s denotes the sth coordinate block of d. Since x^r and x^{r+1} differ in only their sth coordinate block for all $r \in \mathscr{R}'$, then all coordinate blocks of d, except d_s, are zero [see (12)], and the above relation, together with (1), shows that $f(z + \lambda d)$ is constant (and finite) for all $\lambda \in [0, \epsilon]$, a contradiction to Assumption B2, namely, that f is hemivariate in the sth coordinate block. Hence, (11) holds.

Since (11) holds for any subsequence $\{x^r\}_{r \in \mathscr{R}}$ of $\{x^r\}$ converging to z, we can apply (11) to the subsequence $\{x^{r+1}\}_{r \in \mathscr{R}}$ to conclude that $\{x^{r+2}\}_{r \in \mathscr{R}} \to z$ and so on, yielding

$$\{x^{r+j}\}_{r \in \mathscr{R}} \to z, \qquad \forall j = 0, 1, \ldots, T, \tag{20}$$

where T is the bound specified in the essentially cyclic rule.

We claim that (20), together with Assumption C1 or C2, implies

$$f_0(z) + f_k(z_k) \leq f_0(z_1, \ldots, z_{k-1}, x_k, z_{k+1}, \ldots, z_N) + f_k(x_k), \tag{21}$$

for all x_k and all $k \in \{1, \ldots, N\}$. To see this, fix any $k \in \{1, \ldots, N\}$. Since the coordinate blocks are chosen according to the essentially cyclic rule, there exists some $j \in \{1, \ldots, T\}$ and an infinite subsequence $\mathscr{R}' \subseteq \mathscr{R}$ such that the coordinate block x_k is chosen at the $(r+j)$th iteration for all $r \in \mathscr{R}'$. Then, for each $r \in \mathscr{R}'$, $x^{r+j}{}_k$ minimizes $f_0(x^{r+j}{}_1, \ldots, x^{r+j}{}_{k-1}, x_k, x^{r+j}{}_{k+1}, \ldots, x^{r+j}{}_N) + f_k(x_k)$ over all x_k [see (1), (2), (3)], so that

$$\begin{aligned} &f_0(x^{r+j}) + f_k(x^{r+j}{}_k) \\ &\leq f_0(x^{r+j}{}_1, \ldots, x^{r+j}{}_{k-1}, x_k, x^{r+j}{}_{k+1}, \ldots, x^{r+j}{}_N) + f_k(x_k), \qquad \forall x_k. \end{aligned} \tag{22}$$

Fix any $x_k \in \operatorname{dom} f_k$ such that $(z_1, \ldots, z_{k-1}, x_k, z_{k+1}, \ldots, z_N) \in \operatorname{dom} f_0$. Suppose that Assumption C1 holds, so $\operatorname{dom} f_0$ is open. Since $z \in \operatorname{dom} f_0$,

then (20) implies that

$$(x^{r+j}{}_1, \ldots, x^{r+j}{}_{k-1}, x_k, x^{r+j}{}_{k+1}, \ldots, x^{r+j}{}_N) \in \operatorname{dom} f_0,$$

$$\text{for all } r \in \mathscr{R}' \text{ sufficiently large.}$$

Passing to the limit as $r \to \infty$, $r \in \mathscr{R}'$, and using the lsc property of f_k and the continuity of f_0 on the open set $\operatorname{dom} f_0$, we obtain from (20) and (22) that (21) holds. Suppose instead that Assumption C2 holds, so

$$\operatorname{dom} f_0 = Y_1 \times \cdots \times Y_N, \qquad \text{for some } Y_1 \subseteq \Re^{n_1}, \ldots, Y_N \subseteq \Re^{n_N}.$$

Then, the first quantity on the right-hand side of (22) is finite for all $r \in \Re'$. Passing to the limit as $r \to \infty$, $r \in \Re'$, and using the lsc property of f_k and the continuity of f_0 on $\operatorname{dom} f_0$, we obtain from (20) and (22) that (21) holds. If $x_k \notin \operatorname{dom} f_k$ or $(z_1, \ldots, z_{k-1}, x_k, z_{k+1}, \ldots, z_N) \notin \operatorname{dom} f_0$, then the right-hand side of (21) has the extended value ∞, so (21) holds trivially. Since the above choice of k was arbitrary, this shows that (21) holds for all x_k and all $k \in \{1, \ldots, N\}$. Then, it follows from (1) that (4) holds for all $k = 1, \ldots, N$. □

Proposition 5.1 extends a result of Grippo and Sciandrone (Ref. 27, Proposition 5.1) for the special case where each f_k is the indicator function for some closed convex set and f_0 is continuously differentiable and (block) coordinatewise strictly pseudoconvex. In turn, the latter result is an extension of a result of Bertsekas and Tsitsiklis (Ref. 6, Proposition 3.9 in Section 3.3.5), which assumes further f_0 to be convex. As a corollary of Proposition 5.1, we obtain the following convergence result for the BCD method.

Theorem 5.1. Suppose that $f, f_0, f_1, \ldots, f_N$ satisfy Assumptions B1–B3 and that f_0 satisfies either Assumption C1 or C2. Also, assume that $\{x: f(x) \leq f(x^0)\}$ is bounded. Then, the sequence $\{x^r\}$ generated by the BCD method using the essentially cyclic rule is defined, bounded, and every cluster point is a coordinatewise minimum point of f.

Theorem 5.1 extends a result of Auslender [see Theorem 1.2(a) in Ref. 4, p. 95] for the special case where f_k is convex for all k, $\operatorname{dom} f_0 = Y_1 \times \cdots \times Y_N$ for some closed convex sets $Y_k \subseteq \Re^{n_k}$, $k = 1, \ldots, N$, and f_0 is strongly convex and continuous on $\operatorname{dom} f_0$.

6. Applications

We describe four interesting applications of the BCD method below. In all applications, the objective function f is not necessarily strictly convex nor differentiable everywhere on its effective domain.

Example 6.1. Proximal Minimization Algorithm. Let $\psi: \Re^n \mapsto \Re \cup \{\infty\}$ be a proper (i.e., $\psi \not\equiv \infty$) lsc function. Fix any scalar $c > 0$, and consider the proper lsc function f defined by

$$f(x, y) = c\|x - y\|^2 + \psi(x).$$

Clearly, this function has the form (1) with

$$f_0(x, y) = c\|x - y\|^2, \qquad f_1 = \psi, \qquad f_2 \equiv 0.$$

Applying the BCD method to f yields a method whereby $f(x, y)$ is alternately minimized with respect to x and y. This method has the form

$$x^{r+1} = \arg\min_x c\|x - x^r\|^2 + \psi(x), \qquad r = 0, 1, \ldots,$$

which is the proximal minimization algorithm with fixed parameter c for minimizing ψ; see Ref. 6, Section 3.4.3 and Refs. 36–37 and references therein.

It is easily seen that f, f_0, f_1, f_2 satisfy Assumptions B1–B3 and that f_0 satisfies Assumptions A1 and C1. Moreover, f is regular everywhere on dom f. Then, by Proposition 5.1, if ψ is bounded below (so, f is bounded below), then every cluster point z of the iterates generated by the above proximal minimization algorithm is a stationary point of ψ, i.e.,

$$\psi'(z; d) \geq 0, \qquad \text{for all } d.$$

Notice that Theorem 4.1 is not applicable here, since f need not be continuous on its level sets.

Example 6.2. Arimoto–Blahut Algorithm. Let P_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, be given nonnegative scalars satisfying

$$\sum_j P_{ij} = 1, \qquad \text{for all } i.$$

The P_{ij} may be viewed as probabilities. Consider the proper lsc function f defined by

$$f(x, y) = f_0(x, y) + f_1(x) + f_2(y),$$

where

$$f_0(x,y)=\begin{cases}\sum_{j=1}^{m}\sum_{i=1}^{n}P_{ij}x_i\phi(y_{ij}/x_i), & \text{if } x\geq 0,\ y>0,\\ \infty, & \text{otherwise,}\end{cases}$$

$$f_1(x)=\begin{cases}0, & \text{if } \sum_{i=1}^{n}x_i=1,\\ \infty, & \text{otherwise,}\end{cases}$$

$$f_2(y)=\begin{cases}0, & \text{if } \sum_{i=1}^{n}y_{ij}=1, \quad \forall j=1,\ldots,m.\\ \infty, & \text{otherwise,}\end{cases}$$

with $\phi(t)=-\log(t)$. In our notation, x is a vector in $\Re^n$ whose ith coordinate is x_i, and y is a vector in $\Re^{nm}$ whose $((i-1)m+j)$th coordinate is y_{ij}. Applying the BCD method to f yields a method whereby $f(x,y)$ is alternately minimized with respect to x and y. This in turn can be seen to be the Arimoto–Blahut algorithm for computing the capacity of a discrete memoryless communication channel (Refs. 11–12).

It can be verified that f, f_0, f_1, f_2 are convex and satisfy Assumptions B1–B3. Convexity of f_0 follows from observing that $(a,b)\mapsto a\phi(b/a)$ is convex. Moreover, f has compact level sets and is continuous on each level set, and f_0 satisfies Assumptions A2 and C2. Notice that f is not strictly convex and f_0 does not satisfy Assumption A1 or C1. Thus, by Lemma 3.1 and Theorem 5.1 or Theorem 4.1(c), the sequence of iterates generated by the Arimoto–Blahut algorithm is bounded and each cluster point is a stationary point of f. By the convexity of f, this is in fact a minimum point of f. This result matches those obtained in Refs. 11–12. Analogous convergence results are obtained for variants of the Arimoto–Blahut algorithm, whereby we use, for example,

$$\phi(t)=t\log(t) \quad \text{or} \quad \phi(t)=1/t.$$

Example 6.3. Han Algorithm. Let f be the proper lsc convex function studied by Han [Ref. 30, (D′)],

$$f(x_1,\ldots,x_N)=(1/2)\|x_1+\cdots+x_N-d\|^2+\sum_{k=1}^{N}f_k(x_k),$$

where d is a given vector in $\Re^m$ and each $f_k:\Re^m\mapsto\Re\cup\{\infty\}$ is a proper lsc convex function. Also see Ref. 18 for a special case where f_k is the support

function of a closed convex set. Clearly, f is of the form (1) with

$$f_0(x_1, \ldots, x_N) = (1/2)\|x_1 + \cdots + x_N - d\|^2.$$

Han proposed in Ref. 30 an algorithm for minimizing f, which may be viewed as an instance of the BCD method using the cyclic rule, as was shown in Ref. 22.

It is seen easily that $f, f_0, f_1, \ldots, f_N$ satisfy Assumptions B1–B3 and that f_0 satisfies Assumptions A1 and C1. Thus, by Lemma 3.1 and Proposition 5.1 [also see the remark following (3)], if f has a minimum point, then the iterates generated by the Han algorithm are defined and every cluster point is a minimum point of f. This result matches Proposition 4.3 in Ref. 30. On the other hand, by using the convexity of the functions, stronger convergence results can be obtained; see Refs. 22, 38.

Example 6.4. Blind Source Separation. In Ref. 33, Zibulevsky and Pearlmutter studied an optimization formulation of the blind source separation, whereby an error term of the form

$$(1/2\sigma^2)\|AS - X\|_F^2 + \sum_{j,t} f_j^t(s_j^t),$$

is minimized with respect to

$$A \in \Re^{m \times n} \quad \text{and} \quad S = [s_j^t]_{j=1,\ldots,n,\, t=1,\ldots,T} \in \Re^{n \times T}.$$

Here, $X \in \Re^{m \times T}$ are the given data; $\|\cdot\|_F$ denotes the Frobenious norm; $\sigma > 0$; and each $f_j^t: \Re \mapsto [0, \infty]$ is a proper convex function that is continuous on its effective domain and has bounded level sets. In Ref. 31, the particular choice of $f_j^t(\cdot) = |\cdot|$ is used. To ensure the existence of an optimal solution, it was suggested in Ref. 33 that constraints such as

$$\|A_i\| \leq 1, \qquad i = 1, \ldots, m, \tag{23}$$

be imposed, where A_i denotes the ith row of A. The objective function of this problem has the form (1) with $N = 1 + nT$,

$$f_0(A, s_1^1, \ldots, s_n^T) = (1/2\sigma^2)\|AS - X\|_F^2,$$

$$f_1(A) = \begin{cases} 0, & \text{if } \|A_i\| \leq 1, \quad i = 1, \ldots, m, \\ \infty, & \text{else,} \end{cases}$$

and f_j^t, $j = 1, \ldots, n$, $t = 1, \ldots, T$, as given. Notice that minimizing f with respect to A entails minimizing a convex quadratic function over the Cartesian product of m Euclidean balls, while minimizing f with respect to each s_j^t entails minimizing the sum of a convex quadratic function of one

variable with a convex function of one variable. Thus, the BCD method applied to this f can be implemented fairly inexpensively. If we replace (23) by the single ball constraint

$$\|A\|_F \leq \rho,$$

for some fixed $\rho > 0$, then minimizing f with respect to A can be solved efficiently using e.g. the Moré–Sorenson method.

It is not difficult to see that f is continuous on its effective domain and has compact level sets. Moreover, f is convex in $(s_1^1, \ldots, s_n^T)$, f_1 is convex, and f_0 satisfies Assumption A1. Thus, by Lemma 3.1 and Theorem 4.1(b), the iterates generated by the BCD method using the cyclic rule are defined and every cluster point is a stationary point of f. Notice that f is not pseudo-convex in every pair of coordinate blocks and that f need not have at most one minimum in each s_j^t, so neither Theorem 5.1, nor part (a) of Theorem 4.1, nor part (c) of Theorem 4.1 is applicable here.

Instead of treating each s_j^t as a coordinate block, we can treat alternatively $S = [s_j^t]_{j,t}$ as a coordinate block. However, minimizing f with respect to S is more difficult. In the case of $f_j^t(\cdot) = |\cdot|$, this would require solving a large convex quadratic programming problem. A comparison of a primal–dual interior-point method and the BCD method for solving such a problem is given in Ref. 32.

References

1. Ortega, J. M., and Rheinboldt, W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.
2. Hildreth, C., *A Quadratic Programming Procedure*, Naval Research Logistics Quarterly, Vol. 4, pp. 79–85, 1957; see also Erratum, Naval Research Logistics Quarterly, Vol. 4, p. 361, 1957.
3. Warga, J., *Minimizing Certain Convex Functions*, SIAM Journal on Applied Mathematics, Vol. 11, pp. 588–593, 1963.
4. Auslender, A., *Optimisation Méthodes Numériques*, Masson, Paris, France, 1976.
5. Bertsekas, D. P., *Nonlinear Programming*, 2nd Edition, Athena Scientific, Belmont, Massachusetts, 1999.
6. Bertsekas, D. P., and Tsitsiklis, J. N., *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
7. Censor, Y., and Zenios, S. A., *Parallel Optimization: Theory, Algorithms, and Applications,* Oxford University Press, Oxford, United Kingdom, 1997.
8. Luenberger, D. G., *Linear and Nonlinear Programming*, Addison–Wesley, Reading, Massachusetts, 1973.

9. Polak, E., *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, NY, 1971.
10. Zangwill, W. I., *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
11. Arimoto, S., *An Algorithm for Computing the Capacity of Arbitrary DMCs*, IEEE Transactions on Information Theory, Vol. 18, pp. 14–20, 1972.
12. Blahut, R., *Computation of Channel Capacity and Rate Distortion Functions*, IEEE Transactions on Information Theory, Vol. 18, pp. 460–473, 1972.
13. Howson, H. R., and Sancho, N. G. F., *A New Algorithm for the Solution of Multistate Dynamic Programming Problems*, Mathematical Programming, Vol. 8, pp. 104–116, 1975.
14. Korsak, A. J., and Larson, R. E., *A Dynamic Programming Successive Approximations Technique with Convergence Proofs*, Automatica, Vol. 6, pp. 253–260, 1970.
15. Zuo, Z. Q., and Wu, C. P., *Successive Approximation Technique for a Class of Large-Scale NLP Problems and Its Application to Dynamic Programming*, Journal of Optimization Theory and Applications, Vol. 62, pp. 515–527, 1989.
16. Stern, T. E., *A Class of Decentralized Routing Algorithms Using Relaxation*, IEEE Transactions on Communications, Vol. 25, pp. 1092–1102, 1977.
17. Bregman, L. M., *The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming*, USSR Computational Mathematics and Mathematical Physics, Vol. 7, pp. 200–217, 1967.
18. Han, S. P., *A Successive Projection Method*, Mathematical Programming, Vol. 40, pp. 1–14, 1988.
19. Kiwiel, K. C., *Free-Steering Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints*, Mathematics of Operations Research, Vol. 22, pp. 326–349, 1997.
20. Luo, Z. Q., and Tseng, P., *On the Convergence Rate of Dual Ascent Methods for Strictly Convex Minimization*, Mathematics of Operations Research, Vol. 18, pp. 846–867, 1993.
21. Tseng, P., *Dual Ascent Methods for Problems with Strictly Convex Costs and Linear Constraints: A Unified Approach*, SIAM Journal on Control and Optimization, Vol. 28, pp. 214–242, 1990.
22. Tseng, P., *Dual Coordinate Ascent Methods for Nonstrictly Convex Minimization*, Mathematical Programming, Vol. 59, pp. 231–247, 1993.
23. Mangasarian, O. L., and De Leone, R., *Parallel Successive Overrelaxation Methods for Symmetric Linear Complementarity Problems and Linear Programs*, Journal of Optimization Theory and Applications, Vol. 54, pp. 437–446, 1987.
24. Cea, J., and Glowinski, R., *Sur des Methodes d'Optimisation par Relaxation*, Revue Française d'Automatique, Informatique et Recherche Opérationnelle, Vol. R3, pp. 5–32, 1973.
25. Sargent, R. W. H., and Sebastian, D. J., *On the Convergence of Sequential Minimization Algorithms*, Journal of Optimization Theory and Applications, Vol. 12, pp. 567–575, 1973.

26. ZADEH, N., *A Note on the Cyclic Coordinate Ascent Method*, Management Science, Vol. 16, pp. 642–644, 1970.
27. GRIPPO, L., and SCIANDRONE, M., *On the Convergence of the Block Nonlinear Gauss–Seidel Method under Convex Constraints*, Operations Research Letters, Vol. 26, pp. 127–136, 2000.
28. POWELL, M. J. D., *On Search Directions for Minimization Algorithms*, Mathematical Programming, Vol. 4, pp. 193–201, 1973.
29. LUO, Z. Q., and TSENG, P., *Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach*, Annals of Operations Research, Vol. 46, pp. 157–178, 1993.
30. HAN, S. P., *A Decomposition Method and Its Application to Convex Programming*, Mathematics of Operations Research, Vol. 14, pp. 237–248, 1989.
31. BOFILL, P., and ZIBULEVSKY, M., *Sparse Undetermined ICA: Estimating the Mixing Matrix and the Sources Separately*, Technical Report UPC-DAC-2000-7, Universitat Politècnica de Catalunya, Barcelona, Spain, 1999.
32. SARDY, S., BRUCE, A., and TSENG, P., *Block Coordinate Relaxation Methods for Nonparametric Wavelet Denoising*, Journal of Computational and Graphical Statistics, Vol. 9, pp. 361–379, 2000.
33. ZIBULEVSKY, M., and PEARLMUTTER, B., *Blind Source Separation by Sparse Decomposition*, Technical Report CS99-1, Computer Science Department, University of New Mexico, Albuquerque, New Mexico, 1999.
34. MANGASARIAN, O. L., *Nonlinear Programming*, McGraw-Hill, New York, NY, 1969.
35. ROCKAFELLAR, R. T., *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.
36. MARTINET, B., *Determination Approchée d'un Point Fixe d'une Application Pseudo-Contractante: Cas de l'Application Prox*, Comptes Rendus des Séances de l'Académie des Sciences, Vol. 274A, pp. 163–165, 1972.
37. ROCKAFELLAR, R. T., *Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming*, Mathematics of Operations Research, Vol. 1, pp. 97–116, 1976.
38. BAUSCHKE, H. H., and LEWIS, A. S., *Dykstra's Algorithm with Bregman Projections: A Convergence Proof*, Optimization (to appear).