Mathematical Programming manuscript No.
(will be inserted by the editor)

Paul Tseng · Sangwoon Yun

A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization

in honor of Steve Robinson, for his many contributions to mathematical programming, including the subject of this paper: nonsmooth optimization and error bound

Received: xxxx / Revised version: xxxx

Abstract. We consider the problem of minimizing the sum of a smooth function and a *separable* convex function. This problem includes as special cases bound-constrained optimization and smooth optimization with ℓ_1-regularization. We propose a (block) coordinate gradient descent method for solving this class of nonsmooth separable problems. We establish global convergence and, under a local Lipschitzian error bound assumption, linear convergence for this method. The local Lipschitzian error bound holds under assumptions analogous to those for constrained smooth optimization, e.g., the convex function is polyhedral and the smooth function is (nonconvex) quadratic or is the composition of a strongly convex function with a linear mapping. We report numerical experience with solving the ℓ_1-regularization of unconstrained optimization problems from Moré et al. [37] and from the CUTEr set [22]. Comparison with L-BFGS-B and MINOS, applied to a reformulation of the ℓ_1-regularized problem as a bound-constrained optimization problem, is also reported.

Key words. Error bound – Global convergence – Linear convergence rate – Nonsmooth optimization – Coordinate descent

1. Introduction

A type of nonconvex nonsmooth optimization problem that arises in many applications has the following form:

$$\min_x \quad F_c(x) \stackrel{\text{def}}{=} f(x) + cP(x), \tag{1}$$

where $c > 0$, $P : \Re^n \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous (lsc) function [47], and f is smooth (i.e., continuously differentiable) on an open subset of $\Re^n$ containing $\mathrm{dom}P = \{x \mid P(x) < \infty\}$. Of particular interest is when P has a block-separable structure.

A special case of (1) is the bound-constrained optimization problem, where

$$P(x) = \begin{cases} 0 & \text{if } l \le x \le u; \\ \infty & \text{else,} \end{cases} \tag{2}$$

with $l \le u$ (possibly with $-\infty$ or ∞ components). Such a model arises, e.g., in signal denoising based on Markov random field prior [52]. Another special case

Paul Tseng: Department of Mathematics, University of Washington, Seattle, WA 98195, USA e-mail: tseng@math.washington.edu

Sangwoon Yun: Department of Mathematics, University of Washington, Seattle, WA 98195, USA e-mail: sangwoon@math.washington.edu

of (1) that has attracted much interest in signal/image denoising and data mining/classification is the following optimization problem with ℓ_1-regularization:

$$\min_x \quad f(x) + c\|x\|_1, \tag{3}$$

where the 1-norm is introduced to induce sparsity in the final solution. This is useful for finding a sparse representation of a noisy signal or for smoothing a signal/image to have a sparse number of jumps, etc. [5,7,10,11,34,50,51]. In the above cases, P is separable, i.e., a sum of univariate convex lsc functions. Moreover, P enjoys a polyhedral property, i.e, its epigraph epi$P = \{(x,\xi) \mid P(x) \le \xi\}$ is a polyhedral set. Using duality, the "support vector regression" model [5,12,57] can be shown to be a special case of (1) with P separable and piecewise-linear/quadratic. The group Lasso model for regression [35,58] is a special case of (1) with

$$P(x) = \|x_{\mathcal{J}_1}\|_2 + \cdots + \|x_{\mathcal{J}_N}\|_2, \tag{4}$$

where $\mathcal{J}_1, \ldots, \mathcal{J}_N$ is a partition of $\{1, \ldots, n\}$. Here P is block-separable.

The problem (1) has previously been studied in [1,19,21,25,36]. The work most closely related to ours is that of Fukushima and Mine [21], who proposed a proximal gradient descent method with stepsize chosen by an Armijo-type rule. Further extensions of this method were studied by Kiwiel [25] and Fukushima [19]. Mine and Fukushima [36] studied a related Frank-Wolfe-type method. If dom$P = \Re^n$, then (1) is a special case of a composite nonsmooth optimization problem studied in [1,6,16]. The methods proposed in [1,6,16] may be viewed as trust-region versions of the aforementioned method of Fukushima and Mine [21]. If f is convex, then an ϵ-subgradient method can also be applied [3,4,45]. However, the above studies did not present numerical results, so the practical performance of these methods cannot be judged. In the special case of bound-constrained optimization, gradient-projection methods [3,4,24,32,38] or coordinate descent methods [8,23,30,33,42] can be effective. Other methods based on trust region or active set, possibly in conjunction with gradient projection to do active-set identification, have also been much studied; see [4,9,41,59] and references therein. In the special case of (3) where $f(x) = \|Ax - b\|_2^2$, the columns of $A \in \Re^{m\times n}$ are wavelet functions, and $b \in \Re^m$, primal-dual interior-point method [7] and block-coordinate minimization method [49] have been effective; see [50–52] for further discussions and special cases. We can reformulate (1) as a smooth optimization problem over a closed convex set:

$$\min_{x,\xi} \{\ f(x) + c\xi \ \mid \ P(x) - \xi \le 0\ \}. \tag{5}$$

However, existing methods cannot easily exploit the (block) separable structure of P. The special case of (3) can be reformulated as a bound-constrained optimization problem, but the dimension doubles; see Section 7.3.

Thus, even in the special case of (3), there appears to be no existing method that can efficiently solve this problem when f is nonquadratic and n is large. We propose a new method that can efficiently solve (1) and (3) on a large scale.

Our idea is simple: Since coordinate-wise minimization is expensive when f is nonquadratic, we will replace f in F_c by a quadratic approximation. Specifically, we propose a (block) coordinate gradient descent (abbreviated as CGD) method for solving (1) with P having a block-separable structure. At each iteration, we approximate f by a strictly convex quadratic and apply block coordinate descent to generate a descent direction. Then we do an inexact line search along this direction to ensure sufficient descent and re-iterate. This method is simple, highly parallelizable, and is suited for solving large-scale problems. We show that each cluster point of the iterates generated by this method is a stationary point of F_c, provided that the coordinates are updated in either a Gauss-Seidel manner or a Gauss-Southwell manner; see Theorem 1. Thus, coordinate gradient descent not only has cheaper iterations than exact coordinate descent, it also has stronger global convergence properties, able to avoid cycling [43]. We next show that if a local Lipschitzian error bound on the distance to the set of stationary points $\bar{X}$ holds and the isocost surfaces of F_c restricted to $\bar{X}$ are properly separated, then the iterates generated by the CGD method converge at least linearly to a stationary point of F_c; see Theorems 2, 3. This result is analogous to those obtained for feasible descent methods for constrained smooth optimization [26–29, 53]. This local error bound holds if either (i) f is strongly convex with Lipschitz continuous gradient or (ii) P is polyhedral (not necessarily separable) and f is quadratic or the dual of certain strictly convex function or the composition of a strongly convex function with Lipschitz continuous gradient and an affine mapping; see Theorem 4. While error bound for constrained smooth optimization problem has been much studied [14,26–28,44], to our knowledge, error bound for the nonsmooth problem (1) has not been studied previously, and the convergence rate analysis involves new proof ideas to handle the nonsmoothness of F_c. The CGD method may be viewed roughly as a block coordinate version of the method in [21] using a general proximal term, though we also use a different stepsize rule (similar to one in [6]) which is needed for the convergence rate analysis. Our global convergence and convergence rate analyses require weaker assumptions than those in [21]. In Section 7, we describe an implementation of the CGD method, along with convergence acceleration techniques, and we report our numerical experience with solving ℓ_1-regularization of nonlinear least square problems from [37] and unconstrained smooth optimization problems from the CUTEr set [22]. We compare the CGD method with L-BFGS-B [59] and MINOS [39], applied to a reformulation of the ℓ_1-regularized problem as a bound-constrained optimization problem. Our comparison suggests that the CGD method can be effective in practice. In fact, our method is now used for generalized linear model fitting and logistic regression [2,35]. We discuss conclusions and extensions in Section 8.

In our notation, $\Re^n$ denotes the space of n-dimensional real column vectors, and ^{T} denotes transpose. For any $x \in \Re^n$ and nonempty $\mathcal{J} \subseteq \mathcal{N} \stackrel{\text{def}}{=} \{1, \ldots, n\}$, x_j denotes the jth component of x, $x_{\mathcal{J}}$ denotes the subvector of x comprising x_j, $j \in \mathcal{J}$, and $\|x\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}$ for $1 \leq p < \infty$ and $\|x\|_\infty = \max_j |x_j|$. For simplicity, we write $\|x\| = \|x\|_2$. Also, $\mathcal{J}_{\mathrm{C}} = \mathcal{N} \backslash \mathcal{J}$. For $n \times n$ real symmetric

matrices A and B, we write $A \succeq B$ (respectively, $A \succ B$) to mean that $A - B$ is positive semidefinite (respectively, positive definite). $A_{\mathcal{J}\mathcal{J}} = [A_{ij}]_{i,j\in\mathcal{J}}$ denotes the principal submatrix of A indexed by $\mathcal{J}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A. We denote by I the identity matrix and by 0_n the $n \times n$ matrix of zero entries. Unless otherwise specified, $\{x^k\}$ denotes the sequence $x^0, x^1, ...$ and, for any integer $\ell \geq 0$, $\{x^{k+\ell}\}_{\mathcal{K}}$ denotes a subsequence $\{x^{k+\ell}\}_{k\in\mathcal{K}}$ with $\mathcal{K} \subseteq \{0, 1, ...\}$.

2. (Block) Coordinate Gradient Descent Method

In our method, we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate an improving direction d at x. More precisely, we choose a nonempty index subset $\mathcal{J} \subseteq \mathcal{N}$ and a symmetric matrix $H \succ 0_n$ (approximating the Hessian $\nabla^2 f(x)$), and move x along the direction $d = d_H(x; \mathcal{J})$, where

$$d_H(x;\mathcal{J}) \overset{\text{def}}{=} \arg\min_d \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) \;\mid\; d_j = 0 \;\forall j \notin \mathcal{J} \right\}. \tag{6}$$

Notice that $d_H(x;\mathcal{J})$ depends on H only through $H_{\mathcal{J}\mathcal{J}}$. This coordinate gradient descent approach may be viewed as a hybrid of gradient-projection and coordinate descent; also see [15,20,31]. In particular,

- if $\mathcal{J} = \mathcal{N}$ and P is given by (2), then d is a scaled gradient-projection direction for bound-constrained minimization [4,24,38,41];
- if f is quadratic and we choose $H = \nabla^2 f(x)$, then d is a (block) coordinate descent direction [4,41,49,54,55].

If H is block-diagonal and P is accordingly block-separable, then (6) decomposes into subproblems that can be solved in parallel.

Using the convexity of P, we have the following descent lemma.

Lemma 1. *For any $x \in \mathrm{dom} P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and $H \succ 0_n$, let $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$. Then*

$$F_c(x + \alpha d) \leq F_c(x) + \alpha \left(g^T d + cP(x+d) - cP(x) \right) + o(\alpha) \quad \forall \alpha \in (0, 1], \tag{7}$$

$$g^T d + cP(x+d) - cP(x) \leq -d^T H d. \tag{8}$$

Proof. For any $\alpha \in (0, 1]$, we have from the convexity of P and $H \succ 0_n$ that

$$\begin{aligned} F_c(x+\alpha d) - F_c(x) &= f(x+\alpha d) - f(x) + cP(\alpha(x+d) + (1-\alpha)x) - cP(x) \\ &\leq f(x+\alpha d) - f(x) + \alpha cP(x+d) + (1-\alpha)cP(x) - cP(x) \\ &= \alpha g^T d + o(\alpha) + \alpha(cP(x+d) - cP(x)), \end{aligned}$$

which proves (7).

For any $\alpha \in (0,1)$, we have from (6) and the convexity of P that

$$g^T d + \frac{1}{2} d^T H d + cP(x+d) \leq g^T(\alpha d) + \frac{1}{2}(\alpha d)^T H(\alpha d) + cP(x+\alpha d)$$
$$\leq \alpha g^T d + \frac{1}{2}\alpha^2 d^T H d + \alpha cP(x+d) + (1-\alpha)cP(x).$$

Rearranging terms yields

$$(1-\alpha)g^T d + (1-\alpha)(cP(x+d) - cP(x)) + \frac{1}{2}(1-\alpha^2)d^T H d \leq 0.$$

Since $1-\alpha^2 = (1-\alpha)(1+\alpha)$, dividing both sides by $1-\alpha > 0$ and then taking $\alpha \uparrow 1$ prove (8).

We now describe formally the block coordinate gradient descent (abbreviated as CGD) method.

CGD method:

Choose $x^0 \in \mathrm{dom} P$. For $k = 0, 1, 2, ...$, generate x^{k+1} from x^k according to the iteration:

1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and an $H^k \succ 0_n$.
2. Solve (6) with $x = x^k$, $\mathcal{J} = \mathcal{J}^k$, $H = H^k$ to obtain $d^k = d_{H^k}(x^k; \mathcal{J}^k)$.
3. Choose a stepsize $\alpha^k > 0$ and set $x^{k+1} = x^k + \alpha^k d^k$.

Various stepsize rules for smooth optimization [4,17,18,41] can be adapted to our nonsmooth setting to choose α^k. In this paper, we adapt the Armijo rule, based on Lemma 1 and [6, Subsections 4.2, 4.3], which is simple and effective.

Armijo rule:

Choose $\alpha^k_{\mathrm{init}} > 0$ and let α^k be the largest element of $\{\alpha^k_{\mathrm{init}}\beta^j\}_{j=0,1,...}$ satisfying

$$F_c(x^k + \alpha^k d^k) \leq F_c(x^k) + \alpha^k \sigma \Delta^k, \tag{9}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 \leq \gamma < 1$, and

$$\Delta^k \stackrel{\mathrm{def}}{=} \nabla f(x^k)^T d^k + \gamma {d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k). \tag{10}$$

Since $H^k \succ 0_n$ and $0 \leq \gamma < 1$, we see from Lemma 1 that

$$F_c(x^k + \alpha d^k) \leq F_c(x^k) + \alpha \Delta^k + o(\alpha) \quad \forall \alpha \in (0,1],$$

and $\Delta^k \leq (\gamma - 1){d^k}^T H^k d^k < 0$ whenever $d^k \neq 0$. Since $0 < \sigma < 1$, this shows that α^k given by the Armijo rule is well defined and positive. This rule, like that for sequential quadratic programming methods [4,6,9,17,19,41], requires only function evaluations. And, by choosing α^k_{init} based on the previous stepsize α^{k-1}, the number of function evaluations can be kept small in practice. Notice that Δ^k increases with γ. Thus, larger stepsizes will be accepted if we choose either σ near 0 or γ near 1. The descent condition (9) is similar to those used in [1,6] and the term Δ^k therein seems essential to our convergence rate analysis; see Section 8.

For convergence, the index subset $\mathcal{J}^k$ must be chosen judiciously. For smooth optimization, $\mathcal{J}^k$ is often chosen in a *Gauss-Seidel* manner, e.g., $\mathcal{J}^k$ cycles through $\{1\}, \{2\}, ..., \{n\}$ or, more generally, $\mathcal{J}^0, \mathcal{J}^1, ...$ collectively cover $1, 2, ..., n$ for every T consecutive iterations, where $T \geq 1$ [8,23,29,42,55], i.e.,

$$\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+T-1} = \mathcal{N}, \quad k = 0, 1, ... \tag{11}$$

As we shall see, this generalized Gauss-Seidel rule can also be applied to our nonsmooth (block) separable problem to achieve global convergence. However, to establish linear convergence, we need a more restrictive choice of $\mathcal{J}^k$, specifically, there exists a subsequence $\mathcal{T} \subseteq \{0, 1, ...\}$ such that

$$0 \in \mathcal{T}, \qquad \mathcal{N} = \big(\text{disjoint union of } \mathcal{J}^k, \mathcal{J}^{k+1}, ..., \mathcal{J}^{\tau(k)-1}\big) \quad \forall k \in \mathcal{T}, \tag{12}$$

where $\tau(k) \stackrel{\text{def}}{=} \min\{k' \in \mathcal{T} \mid k' > k\}$. In particular, (12) is a special case of (11) with $T \leq n$. This choice seems most effective when P is block-separable with large blocks, as in the case of group Lasso (4); see [35].

For smooth optimization, $\mathcal{J}^k$ can also be chosen in a *Gauss-Southwell* manner, indexing partial derivatives of the objective function that are within a multiplicative factor of being maximum in magnitude [20,42,49]. This can be extended to our nonsmooth separable problem as follows. For any $x \in \text{dom}P$ and $H \succ 0_n$, let

$$d_H(x) \stackrel{\text{def}}{=} d_H(x; \mathcal{N}). \tag{13}$$

We will see in Lemma 2 that an $x \in \text{dom}P$ is a stationary point of F_c if and only if $d_H(x) = 0$. Thus, $\|d_H(x)\|_\infty$ acts as a scaled "residual" function (with scaling matrix H), measuring how close x comes to being stationary for F_c. Moreover, if H is diagonal and P is separable, then $d_H(x)_j$, the jth components of $d_H(x)$, depends on x_j only and is easily computable.

- If $P \equiv 0$, then $d_H(x)_j = -\nabla f(x)_j / H_{jj}$.
- If P is given by (2), then $d_H(x)_j = \text{mid}\{l_j - x_j, -\nabla f(x)_j / H_{jj}, u_j - x_j\}$.
- If P is the 1-norm, then $d_H(x)_j = -\text{mid}\{(\nabla f(x)_j - c)/H_{jj}, x_j, (\nabla f(x)_j + c)/H_{jj}\}$.

[$\text{mid}\{a, b, c\}$ denotes the median (mid-point) of a, b, c.] Accordingly, we choose $\mathcal{J}^k$ to satisfy

$$\|d_{D^k}(x^k; \mathcal{J}^k)\|_\infty \geq \upsilon \|d_{D^k}(x^k)\|_\infty, \tag{14}$$

where $0 < \upsilon \leq 1$ and $D^k \succ 0_n$ is diagonal (e.g., $D^k = I$ or $D^k = \text{diag}(H^k)$). Other norms beside ∞-norm can also be used. We will call (14) the *Gauss-Southwell-r* rule. Notice that $\mathcal{J}^k = \mathcal{N}$ is a valid choice. If P is the indicator function for a closed convex set $X \subseteq \Re^n$, then $d_I(x) = [x - \nabla f(x)]^+_X - x$, where $[x]^+_X$ denotes the orthogonal projection of x onto X. Thus, $d_H(x)$ is a generalization of the projection residual for constrained smooth optimization.

We will see that the above Gauss-Southwell-r rule yields global convergence of the CGD method. However, we have been unable to establish linear convergence due to the nonsmoothness of F_c, whose local growth rates along different coordinate directions are not adequately captured by the residual $d_{D^k}(x^k)$; see

Section 5 for more discussions. This motivated us to consider a (new) Gauss-Southwell rule based on the optimal objective value of (6) rather than the norm of its optimal solution. For any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $H \succ 0_n$,

$$q_H(x;\mathcal{J}) \stackrel{\mathrm{def}}{=} \left(\nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d)\right)_{d=d_H(x;\mathcal{J})} - cP(x). \quad (15)$$

Thus $q_H(x;\mathcal{J})$ estimates the descent in F_c from x to $x+d_H(x;\mathcal{J})$. We have from (8) in Lemma 1 that $q_H(x;\mathcal{N}) \le -\frac{1}{2} d_H(x)^T H d_H(x) \le 0$, so that $q_H(x;\mathcal{N}) = 0$ if and only if $d_H(x) = 0$. If P is separable and H is diagonal, then $q_H(x;\mathcal{J})$ is separable in the sense that $q_H(x;\mathcal{J}) = \sum_{j\in\mathcal{J}} q_H(x;j)$. Accordingly, we choose $\mathcal{J}^k$ to satisfy

$$q_{D^k}(x^k;\mathcal{J}^k) \le \upsilon\, q_{D^k}(x^k;\mathcal{N}), \quad (16)$$

where $0 < \upsilon \le 1$, $D^k \succ 0_n$ is diagonal. We call this the *Gauss-Southwell-q* rule. These Gauss-Southwell rules seem most effective when P is separable; see Section 7.

3. Properties of Search Direction

In this section we study properties of the search direction $d_H(x,\mathcal{J})$ and the residual $d_H(x)$ which will be useful for analyzing the global convergence and linear convergence rate of the CGD method.

Formally, we say that $x \in \Re^n$ is *a stationary point* of F_c if $x \in \mathrm{dom}F_c$ and $F_c'(x;d) \ge 0$ for all $d \in \Re^n$. The following lemma gives a useful alternative characterization of stationarity.

Lemma 2. *For any $H \succ 0_n$, an $x \in \mathrm{dom}P$ is a stationary point of F_c if and only if $d_H(x) = 0$.*

Proof. Fix any $x \in \mathrm{dom}P$ and $H \succ 0_n$. If $d_H(x) \ne 0$, then (7) and (8) show that $d_H(x)$ is a descent direction for F_c at x, implying that x is not a stationary point of F_c. Conversely, if $d_H(x) = 0$, then

$$g^T u + \frac{1}{2} u^T H u + cP(x+u) \ge cP(x) \quad \forall u \in \Re^n,$$

where $g = \nabla f(x)$. For any $d \in \Re^n$, letting $u = \alpha d$ for $\alpha > 0$ yields

$$\alpha g^T d + \frac{1}{2}\alpha^2 d^T H d + cP(x+\alpha d) \ge cP(x) \quad \forall \alpha > 0. \quad (17)$$

Since $f(x+\alpha d) - f(x) = \alpha g^T d + o(\alpha)$, this together with (17) yields

$$\begin{aligned} F_c'(x;d) &= \lim_{\alpha\downarrow 0} \frac{f(x+\alpha d) - f(x) + cP(x+\alpha d) - cP(x)}{\alpha} \\ &\ge \lim_{\alpha\downarrow 0} \frac{o(\alpha) - \frac{1}{2}\alpha^2 d^T H d}{\alpha} = 0 \quad \forall d \in \Re^n. \end{aligned}$$

Hence $F_c'(x;d) \ge 0$ for all d, implying that x is a stationary point of F_c.

The next lemma shows that $\|d_H(x;\mathcal{J})\|$ changes not too fast with the quadratic coefficients H. It will be used to prove Theorems 1 and 2.

Lemma 3. *For any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $H \succ 0_n$, $\tilde{H} \succ 0_n$, let $d = d_H(x;\mathcal{J})$ and $\tilde{d} = d_{\tilde{H}}(x;\mathcal{J})$. Then*

$$\|\tilde{d}\| \le \frac{1+\lambda_{\max}(Q)+\sqrt{1-2\lambda_{\min}(Q)+\lambda_{\max}(Q)^2}}{2}\ \frac{\lambda_{\max}(H_{\mathcal{J}\mathcal{J}})}{\lambda_{\min}(\tilde{H}_{\mathcal{J}\mathcal{J}})}\ \|d\|, \qquad (18)$$

where $Q = H_{\mathcal{J}\mathcal{J}}^{-1/2}\tilde{H}_{\mathcal{J}\mathcal{J}}H_{\mathcal{J}\mathcal{J}}^{-1/2}$. If $H_{\mathcal{J}\mathcal{J}} \succ \tilde{H}_{\mathcal{J}\mathcal{J}}$, then also

$$\|d\| \le \sqrt{\frac{\lambda_{\max}(H_{\mathcal{J}\mathcal{J}}-\tilde{H}_{\mathcal{J}\mathcal{J}})}{\lambda_{\min}(H_{\mathcal{J}\mathcal{J}}-\tilde{H}_{\mathcal{J}\mathcal{J}})}}\ \|\tilde{d}\|. \qquad (19)$$

Proof. Since $d_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$, it suffices to prove the lemma for the case of $\mathcal{J} = \mathcal{N}$. Let $g = \nabla f(x)$. By the definition of d and $\tilde{d}$ and Fermat's rule [48, Theorem 10.1],

$$d \in \arg\min_u\ (g+Hd)^T u + cP(x+u),$$
$$\tilde{d} \in \arg\min_u\ (g+\tilde{H}\tilde{d})^T u + cP(x+u).$$

Thus

$$(g+Hd)^T d + cP(x+d) \le (g+Hd)^T\tilde{d} + cP(x+\tilde{d}),$$
$$(g+\tilde{H}\tilde{d})^T\tilde{d} + cP(x+\tilde{d}) \le (g+\tilde{H}\tilde{d})^T d + cP(x+d).$$

Adding the above two inequalities and rearranging terms yield

$$d^T H d - d^T(H+\tilde{H})\tilde{d} + \tilde{d}^T\tilde{H}\tilde{d} \le 0.$$

Then, by completing the square on the first two terms, we have

$$\|H^{1/2}d - H^{-1/2}(H+\tilde{H})\tilde{d}/2\|^2 \le \|H^{-1/2}(H+\tilde{H})\tilde{d}\|^2/4 - \tilde{d}^T\tilde{H}\tilde{d}.$$

By making the substitution $u = H^{1/2}d$, $\tilde{u} = H^{1/2}\tilde{d}$, this can be rewritten as

$$\|u - (I+Q)\tilde{u}/2\|^2 \le \|(I+Q)\tilde{u}\|^2/4 - \tilde{u}^T Q\tilde{u}.$$

The right-hand side simplifies to $\|(I-Q)\tilde{u}\|^2/4$, so taking square root of both sides yields

$$\|u - (I+Q)\tilde{u}/2\| \le \|(I-Q)\tilde{u}\|/2.$$

We apply the triangular inequality to the left-hand side and rearrange terms to obtain

$$\|(I+Q)\tilde{u}\|/2 - \|(I-Q)\tilde{u}\|/2 \le \|u\|.$$

Multiplying both sides by $2\|(I+Q)\tilde{u}\| + 2\|(I-Q)\tilde{u}\|$ and simplifying yields

$$4\tilde{u}^T Q\tilde{u} \le 2\|u\|(\|(I+Q)\tilde{u}\| + \|(I-Q)\tilde{u}\|).$$

Since $Q \succ 0_n$, this together with $\|(I+Q)\tilde{u}\| \leq (1+\lambda_{\max}(Q))\|\tilde{u}\|$ and $\|(I-Q)\tilde{u}\| \leq \sqrt{1-2\lambda_{\min}(Q)+\lambda_{\max}(Q)^2}\|\tilde{u}\|$ yields

$$2\tilde{u}^T Q\tilde{u} \leq \|u\|(1+\lambda_{\max}(Q)+\sqrt{1-2\lambda_{\min}(Q)+\lambda_{\max}(Q)^2})\|\tilde{u}\|.$$

Since $\tilde{u}^T Q\tilde{u} = \tilde{d}^T\tilde{H}\tilde{d} \geq \lambda_{\min}(\tilde{H})\|\tilde{d}\|^2$ and $\|u\| \leq \sqrt{\lambda_{\max}(H)}\|d\|$, $\|\tilde{u}\| \leq \sqrt{\lambda_{\max}(H)}\|\tilde{d}\|$, this yields (18).

Suppose $H \succ \tilde{H}$. From the definition of d and $\tilde{d}$, we have

$$g^T d + \frac{1}{2}d^T H d + cP(x+d) \leq g^T\tilde{d} + \frac{1}{2}\tilde{d}^T H\tilde{d} + cP(x+\tilde{d}),$$
$$g^T\tilde{d} + \frac{1}{2}\tilde{d}^T\tilde{H}\tilde{d} + cP(x+\tilde{d}) \leq g^T d + \frac{1}{2}d^T\tilde{H}d + cP(x+d).$$

Adding the above two inequalities and rearranging terms yields

$$d^T(H-\tilde{H})d \leq \tilde{d}^T(H-\tilde{H})\tilde{d}.$$

Hence

$$\lambda_{\min}(H-\tilde{H})\|d\|^2 \leq \lambda_{\max}(H-\tilde{H})\|\tilde{d}\|^2,$$

which proves (19).

If $H = \gamma I$ and $\tilde{H} = \tilde{\gamma}I$ with $\gamma \geq \tilde{\gamma} > 0$, Lemma 3 yields that

$$\|d\| \leq \|\tilde{d}\| \leq \frac{\gamma}{\tilde{\gamma}}\|d\|.$$

By switching the roles of H and $\tilde{H}$, (18) also yields $\|d\| = O(\|\tilde{d}\|)$. However, this bound seems not as sharp as (19).

The next lemma shows that $d_H(x;\mathcal{J})$ changes not too fast with the linear coefficients $\nabla f(x)$. It will be used to prove Theorem 2 on the linear convergence of the CGD method.

Lemma 4. *Let $h:\Re^n \to \Re$ be a smooth function satisfying $(\nabla h(u)-\nabla h(v))^T(u-v) \geq \rho\|u-v\|_p^p$ for all $u,v \in \Re^n$, for some $\rho > 0$ and $p > 1$. Let q satisfy $\frac{1}{p}+\frac{1}{q} = 1$. Then, for any $x \in \mathrm{dom}P$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and $\bar{g}, \tilde{g} \in \Re^n$,*

$$\|\bar{d}-\tilde{d}\|_p \leq \rho^{-q/p}\|\bar{g}_{\mathcal{J}} - \tilde{g}_{\mathcal{J}}\|_q^{q/p},$$

where $\bar{d} = \operatorname{arg\,min}_{d|d_j=0\ \forall j\notin\mathcal{J}} \bar{g}^T d + h(d) + cP(x+d)$ and $\tilde{d} = \operatorname{arg\,min}_{d|d_j=0\ \forall j\notin\mathcal{J}} \tilde{g}^T d + h(d) + cP(x+d)$.

Proof. By assumption, h is strictly convex and coercive, so $\bar{d}$ and $\tilde{d}$ are well defined. By Fermat's rule [48, Theorem 10.1],

$$\bar{d} \in \operatorname*{arg\,min}_{d|d_j=0\ \forall j\notin\mathcal{J}} (\bar{g}+\nabla h(\bar{d}))^T d + cP(x+d), \quad \tilde{d} \in \operatorname*{arg\,min}_{d|d_j=0\ \forall j\notin\mathcal{J}} (\tilde{g}+\nabla h(\tilde{d}))^T d + cP(x+d).$$

Hence

$$(\bar{g}+\nabla h(\bar{d}))^T\bar{d} + cP(x+\bar{d}) \leq (\bar{g}+\nabla h(\bar{d}))^T\tilde{d} + cP(x+\tilde{d}),$$

$$(\tilde{g} + \nabla h(\tilde{d}))^T \tilde{d} + cP(x + \tilde{d}) \leq (\tilde{g} + \nabla h(\tilde{d}))^T \bar{d} + cP(x + \bar{d}).$$

Summing the above two inequalities and rearranging terms, we have

$$(\tilde{g} - \bar{g})^T (\bar{d} - \tilde{d}) \geq (\nabla h(\bar{d}) - \nabla h(\tilde{d}))^T (\bar{d} - \tilde{d}) \geq \rho \|\bar{d} - \tilde{d}\|_p^p.$$

Since $\bar{d}_j = \tilde{d}_j = 0$ for all $j \notin \mathcal{J}$ and $\|u\|_q \|v\|_p \geq u^T v$ for any $u, v \in \Re^n$, this yields

$$\|\tilde{g}_{\mathcal{J}} - \bar{g}_{\mathcal{J}}\|_q \|\bar{d} - \tilde{d}\|_p \geq \rho \|\bar{d} - \tilde{d}\|_p^p,$$

which, upon simplification, proves the desired result.

It can be shown that $h(d) = \frac{1}{p}\|d\|_p^p$, with $p \geq 2$, satisfies the assumption of Lemma 4 with $\rho = 1/2^{p-2}$.

We say that P is *block-separable* with respect to nonempty $\mathcal{J} \subseteq \mathcal{N}$ if

$$P(x) = P_{\mathcal{J}}(x_{\mathcal{J}}) + P_{\mathcal{J}_C}(x_{\mathcal{J}_C}) \quad \forall x \in \Re^n, \tag{20}$$

for some proper, convex, lsc functions $P_{\mathcal{J}}$ and $P_{\mathcal{J}_C}$. In this case, the subproblem (6) reduces to the following subproblem:

$$\min_{d_{\mathcal{J}}} \ \nabla f(x)_{\mathcal{J}}^T d_{\mathcal{J}} + \frac{1}{2} d_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} d_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) \tag{21}$$

where $H_{\mathcal{J}\mathcal{J}}$ is the principal submatrix of H indexed by $\mathcal{J}$. Using this observation, we have the next lemma concerning stepsizes satisfying the Armijo descent condition (9). This lemma will be used to prove Theorems 1(f) and 2.

Lemma 5. *For any $x \in \mathrm{dom} P$, $H \succ 0_n$, and nonempty $\mathcal{J} \subseteq \mathcal{N}$, let $d = d_H(x; \mathcal{J})$ and $g = \nabla f(x)$. For any $\gamma \in [0, 1)$, the following results hold with $\Delta = g^T d + \gamma d^T H d + cP(x + d) - cP(x)$.*

(a) If P is block-separable with respect to $\mathcal{J}$, then, for any $\bar{x} \in \Re^n$, $\alpha \in (0, 1]$, and $x' = x + \alpha d$,

$$(g + Hd)_{\mathcal{J}}^T (x' - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \leq (\alpha - 1) \left[(1 - \gamma) d^T H d + \Delta \right].$$

(b) If f satisfies

$$\|\nabla f(y) - \nabla f(z)\| \leq L \|y - z\| \quad \forall y, z \in \mathrm{dom} P, \tag{22}$$

for some $L \geq 0$, and $H \succeq \underline{\lambda} I$, where $\underline{\lambda} > 0$, then the descent condition

$$F_c(x + \alpha d) - F_c(x) \leq \sigma \alpha \Delta, \tag{23}$$

is satisfied for any $\sigma \in (0, 1)$ whenever $0 \leq \alpha \leq \min\{1, 2\underline{\lambda}(1 - \sigma + \sigma\gamma)/L\}$.

Proof. (a) Since $d = d_H(x; \mathcal{J})$, by (21) and Fermat's rule [48, Theorem 10.1],

$$d_{\mathcal{J}} \in \arg\min_{u_{\mathcal{J}}} (g + Hd)_{\mathcal{J}}^T u_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + u_{\mathcal{J}}).$$

Thus,

$$(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) \le (g + Hd)_{\mathcal{J}}^T (\bar{x} - x)_{\mathcal{J}} + cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}). \tag{24}$$

Since $x' = x + \alpha d$, we have

$$\begin{aligned}
&(g + Hd)_{\mathcal{J}}^T (x' - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \\
&= (\alpha - 1)(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) + (g + Hd)_{\mathcal{J}}^T (x + d - \bar{x})_{\mathcal{J}} - cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \\
&\le (\alpha - 1)(g + Hd)_{\mathcal{J}}^T d_{\mathcal{J}} + cP_{\mathcal{J}}(x'_{\mathcal{J}}) - cP_{\mathcal{J}}(x_{\mathcal{J}} + d_{\mathcal{J}}) \\
&= (\alpha - 1)(g + Hd)^T d + cP(x') - cP(x + d) \\
&\le (\alpha - 1)(g + Hd)^T d + (1 - \alpha)cP(x) + \alpha cP(x + d) - cP(x + d) \\
&= (\alpha - 1)(g^T d + d^T H d + cP(x + d) - cP(x)) \\
&= (\alpha - 1)(1 - \gamma) d^T H d + (\alpha - 1)\Delta,
\end{aligned}$$

where the second step uses (24), the third step uses $d_j = 0$ for all $j \notin \mathcal{J}$, and the fourth step uses the convexity of P and $0 < \alpha \le 1$. This proves the desired result.

(b) For any $\alpha \in [0, 1]$, we have from the convexity of P and the Cauchy-Schwarz inequality that

$$\begin{aligned}
&F_c(x + \alpha d) - F_c(x) \\
&= f(x + \alpha d) - f(x) + cP(x + \alpha d) - cP(x) \\
&= \alpha \nabla f(x)^T d + cP(x + \alpha d) - cP(x) + \int_0^1 (\nabla f(x + t\alpha d) - \nabla f(x))^T (\alpha d)\, dt \\
&\le \alpha \nabla f(x)^T d + \alpha(cP(x + d) - cP(x)) + \alpha \int_0^1 \|\nabla f(x + t\alpha d) - \nabla f(x)\| \|d\|\, dt \\
&\le \alpha(\nabla f(x)^T d + cP(x + d) - cP(x)) + \alpha^2 \frac{L}{2} \|d\|^2 \\
&= \alpha(g^T d + \gamma d^T H d + cP(x + d) - cP(x)) - \alpha\gamma d^T H d + \alpha^2 \frac{L}{2} \|d\|^2,
\end{aligned} \tag{25}$$

where the third step uses the convexity of P; the fourth step uses (22) and the convexity of $\mathrm{dom} P$, in which x and $x + d$ lie. If $\alpha \le 2\underline{\lambda}(1 - \sigma + \sigma\gamma)/L$, then $d^T H d \ge \underline{\lambda}\|d\|^2$ implies

$$\begin{aligned}
\alpha \frac{L}{2} \|d\|^2 - \gamma d^T H d &\le (1 - \sigma + \sigma\gamma) d^T H d - \gamma d^T H d \\
&= (1 - \sigma)(1 - \gamma) d^T H d \\
&\le -(1 - \sigma)(g^T d + \gamma d^T H d + cP(x + d) - cP(x)),
\end{aligned}$$

where the third step uses (8) in Lemma 1. This together with (25) proves (23).

If P is separable, then P is block-separable with respect to every nonempty $\mathcal{J} \subseteq \mathcal{N}$, with $P_{\mathcal{J}}(x_{\mathcal{J}}) = \sum_{j \in \mathcal{J}} P_j(x_j)$. The converse also holds, since if P is block-separable with respect to $\mathcal{J}, \mathcal{K} \subseteq \mathcal{N}$ such that $\mathcal{J} \cap \mathcal{K} \neq \emptyset$, then P is block-separable with respect to $\mathcal{J} \cap \mathcal{K}$.[1]

4. Global Convergence Analysis

In this section we analyze the global convergence of the CGD method under the following reasonable assumption on the choice of H^k. The proof uses Lemmas 1, 2, 3, and 5(b).

Assumption 1 *$\bar{\lambda} I \succeq H^k \succeq \underline{\lambda} I$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.*

Theorem 1. *Let $\{x^k\}$, $\{d^k\}$, $\{H^k\}$ be sequences generated by the CGD method under Assumption 1, where $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$. Then the following results hold.*

(a) $\{F_c(x^k)\}$ is nonincreasing and Δ^k given by (10) satisfies

$$-\Delta^k \geq (1-\gamma) {d^k}^T H^k d^k \geq (1-\gamma)\underline{\lambda} \|d^k\|^2 \quad \forall k, \tag{26}$$

$$F_c(x^{k+1}) - F_c(x^k) \leq \sigma \alpha^k \Delta^k \leq 0 \quad \forall k. \tag{27}$$

(b) If $\{x^k\}_{\mathcal{K}}$ is a convergent subsequence of $\{x^k\}$, then $\{\alpha^k \Delta^k\} \to 0$ and $\{d^k\}_{\mathcal{K}} \to 0$. If in addition $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \leq \bar{\delta}$, then $\{d_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}} \to 0$.

(c) If $\{\mathcal{J}^k\}$ is chosen by the Gauss-Southwell-r rule (14) and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \leq \bar{\delta}$, then every cluster point of $\{x^k\}$ is a stationary point of F_c.

(d) If $\{\mathcal{J}^k\}$ is chosen by the Gauss-Southwell-q rule (16), $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k, where $0 < \underline{\delta} \leq \bar{\delta}$, and either (1) P is continuous on $\text{dom} P$ or (2) $\inf_k \alpha^k > 0$ or (3) $\alpha^k_{\text{init}} = 1$ for all k, then every cluster point of $\{x^k\}$ is a stationary point of F_c.

(e) If $\{\mathcal{J}^k\}$ is chosen by the generalized Gauss-Seidel rule (11), P is block-separable with respect to $\mathcal{J}^k$ for all k, and $\sup_k \alpha^k < \infty$, then every cluster point of $\{x^k\}$ is a stationary point of F_c.

(f) If f satisfies (22) for some $L \geq 0$, then $\inf_k \alpha^k > 0$. If $\lim_{k\to\infty} F_c(x^k) > -\infty$ also, then $\{\Delta^k\} \to 0$ and $\{d^k\} \to 0$.

Proof. (a) The inequalities (26) follow from (10), (8) in Lemma 1, $0 \leq \gamma < 1$, and $H^k \succeq \underline{\lambda} I$. Since $x^{k+1} = x^k + \alpha^k d^k$ and α^k is chosen by the Armijo rule (9), we have (27) and hence $\{F_c(x^k)\}$ is nonincreasing.

[1] Why? Fix $x_{\mathcal{J}\setminus\mathcal{K}} = \bar{x}_{\mathcal{J}\setminus\mathcal{K}}$ for some $\bar{x}_{\mathcal{J}\setminus\mathcal{K}} \in \text{dom} P_{\mathcal{J}\setminus\mathcal{K}}$ and vary $x_{\mathcal{J}}$. Since $P_{\mathcal{J}}(x_{\mathcal{J}}) + P_{\mathcal{J}\setminus\mathcal{K}}(\bar{x}_{\mathcal{J}\setminus\mathcal{K}}) = P_{\mathcal{K}}(x_{\mathcal{K}}) + P_{\mathcal{K}\setminus\mathcal{J}}(x_{\mathcal{K}\setminus\mathcal{J}})$, $P_{\mathcal{J}}(x_{\mathcal{J}})$ is a sum of two functions, one of $x_{\mathcal{J}\cap\mathcal{K}}$ only and the other of $x_{\mathcal{J}\setminus\mathcal{K}}$ only.

(b) Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to some $\bar{x}$. Since F_c is lsc, $F_c(\bar{x}) \leq \liminf_{\substack{k\to\infty \\ k\in\mathcal{K}}} F_c(x^k)$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $\{F_c(x^k)\}$ converges to a finite limit. Hence, $\{F_c(x^k) - F_c(x^{k+1})\} \to 0$. Then, by (27),

$$\{\alpha^k \Delta^k\} \to 0. \tag{28}$$

Suppose that $\{d^k\}_{\mathcal{K}} \not\to 0$. By passing to a subsequence if necessary, we can assume that, for some $\delta > 0$, $\|d^k\| \geq \delta$ for all $k \in \mathcal{K}$. Then, by (28), $\{\alpha^k\}_{\mathcal{K}} \to 0$. Since $\inf_k \alpha^k_{\text{init}} > 0$, there exists some index $\bar{k} \geq 0$ such that $\alpha^k < \alpha^k_{\text{init}}$ and $\alpha^k \leq \beta$ for all $k \in \mathcal{K}$ with $k \geq \bar{k}$. Since α^k is chosen by the Armijo rule, this implies that

$$F_c(x^k + (\alpha^k/\beta)d^k) - F_c(x^k) > \sigma(\alpha^k/\beta)\Delta^k \quad \forall k \in \mathcal{K},\ k \geq \bar{k}.$$

Thus

$$\begin{aligned}
&\sigma\Delta^k \\
&= \sigma\left(g^{k\,T}d^k + \gamma d^{k\,T}H^k d^k + cP(x^k + d^k) - cP(x^k)\right) \\
&< \frac{f(x^k + (\alpha^k/\beta)d^k) - f(x^k) + cP(x^k + (\alpha^k/\beta)d^k) - cP(x^k)}{\alpha^k/\beta} \\
&\leq \frac{f(x^k + (\alpha^k/\beta)d^k) - f(x^k) + (\alpha^k/\beta)cP(x^k + d^k) + (1 - \alpha^k/\beta)cP(x^k) - cP(x^k)}{\alpha^k/\beta} \\
&= \frac{f(x^k + (\alpha^k/\beta)d^k) - f(x^k)}{\alpha^k/\beta} + cP(x^k + d^k) - cP(x^k) \quad \forall k \in \mathcal{K},\ k \geq \bar{k},
\end{aligned}$$

where the second inequality uses $0 < \alpha^k/\beta \leq 1$ and the convexity of P. Using the definition of Δ^k, we can rewrite this as

$$-(1-\sigma)\Delta^k + \gamma d^{k\,T}H^k d^k \leq \frac{f(x^k + (\alpha^k/\beta)d^k) - f(x^k)}{\alpha^k/\beta} - g^{k\,T}d^k.$$

Since, by (26), the left-hand side is greater than or equal to $((1-\sigma)(1-\gamma) + \gamma)\underline{\lambda}\|d^k\|^2$, dividing both sides by $\|d^k\|$ yields

$$((1-\sigma)(1-\gamma)+\gamma)\underline{\lambda}\|d^k\| \leq \frac{f(x^k + \hat{\alpha}^k d^k/\|d^k\|) - f(x^k)}{\hat{\alpha}^k} - \frac{g^{k\,T}d^k}{\|d^k\|} \quad \forall k \in \mathcal{K},\ k \geq \bar{k}, \tag{29}$$

where we let $\hat{\alpha}^k = \alpha^k\|d^k\|/\beta$. By (26), $-\alpha^k\Delta^k \geq (1-\gamma)\underline{\lambda}\alpha^k\|d^k\|^2 \geq (1-\gamma)\underline{\lambda}\alpha^k\|d^k\|\delta$ for all $k \in \mathcal{K}$, so (28) and $(1-\gamma)\underline{\lambda} > 0$ imply $\{\alpha^k\|d^k\|\}_{\mathcal{K}} \to 0$ and hence $\{\hat{\alpha}^k\}_{\mathcal{K}} \to 0$. Also, since $\{d^k/\|d^k\|\}_{\mathcal{K}}$ is bounded, by passing to a subsequence if necessary, we can assume that $\{d^k/\|d^k\|\}_{\mathcal{K}} \to$ some $\bar{d}$. Taking the limit as $k \in \mathcal{K}, k \to \infty$ in the inequality (29) and using the smoothness of f, we obtain

$$0 < ((1-\sigma)(1-\gamma) + \gamma)\underline{\lambda}\delta \leq \nabla f(\bar{x})^T\bar{d} - \nabla f(\bar{x})^T\bar{d} = 0,$$

a clear contradiction. Thus $\{d^k\}_{\mathcal{K}} \to 0$.

Suppose that, in addition, $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Then, for each k,

$$\frac{\bar{\delta}}{\underline{\lambda}} I \succeq \bar{\delta}(H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1} \succeq (H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1/2} D^k_{\mathcal{J}^k\mathcal{J}^k} (H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1/2} \succeq \underline{\delta}(H^k_{\mathcal{J}^k\mathcal{J}^k})^{-1} \succeq \frac{\underline{\delta}}{\bar{\lambda}} I.$$

Then (18) in Lemma 3 yields

$$\|d_{D^k}(x^k;\mathcal{J}^k)\| \le \frac{1+\bar{\delta}/\underline{\lambda}+\sqrt{1-2\underline{\delta}/\bar{\lambda}+(\bar{\delta}/\underline{\lambda})^2}}{2} \, \frac{\bar{\lambda}}{\underline{\delta}} \, \|d^k\|. \tag{30}$$

Since $\{d^k\}_{\mathcal{K}} \to 0$, this implies $\{d_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0$.

(c) Suppose that $\mathcal{J}^k$ is chosen by the Gauss-Southwell-r rule and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. Then, by (b), $\{d_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0$. By the Gauss-Southwell-r rule (14), this in turn implies $\{r^k\}_{\mathcal{K}} \to 0$, where we denote for simplicity $r^k = d_{D^k}(x^k)$. By (6) and (13), we have

$$g^{k^T} r^k + \frac{1}{2} r^{k^T} D^k r^k + cP(x^k+r^k) \le g^{k^T}(x-x^k) + \frac{1}{2}(x-x^k)^T D^k (x-x^k) + cP(x) \quad \forall x \in \Re^n,$$

so passing to the limit as $k \in \mathcal{K}, k \to \infty$ and using the smoothness of f and lsc of P yields

$$cP(\bar{x}) \le \nabla f(\bar{x})^T (x-\bar{x}) + \frac{1}{2}(x-\bar{x})^T \bar{D}(x-\bar{x}) + cP(x) \quad \forall x \in \Re^n,$$

where $\bar{D}$ is any cluster point of $\{D^k\}_{\mathcal{K}}$. Since $D^k \succeq \underline{\delta} I$ for all $k \in \mathcal{K}$, $\bar{D} \succ 0_n$. This shows that $d_{\bar{D}}(\bar{x}) = 0$ so that, by Lemma 2, $\bar{x}$ is a stationary point of F_c.

(d) Suppose that $\mathcal{J}^k$ is chosen by the Gauss-Southwell-q rule, and $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. By (b), $\{d^k\}_{\mathcal{K}} \to 0$ and $\{\tilde{d}^k\}_{\mathcal{K}} \to 0$, where we denote $\tilde{d}^k = d_{D^k}(x^k;\mathcal{J}^k)$.

Suppose furthermore that either P is continuous on $\mathrm{dom} P$ or $\alpha^k_{\mathrm{init}} = 1$ for all k or $\inf_k \alpha^k > 0$. We will show that

$$\{q_{D^k}(x^k;\mathcal{J}^k)\}_{\mathcal{K}} \to 0. \tag{31}$$

Then, by (16), $\{q_{D^k}(x^k;\mathcal{N})\}_{\mathcal{K}} \to 0$. Since, by (15), (8) in Lemma 1, and $D^k \succeq \underline{\delta} I$, we also have

$$q_{D^k}(x^k;\mathcal{N}) \le -\frac{1}{2} d_{D^k}(x^k)^T D^k d_{D^k}(x^k) \le -\frac{\underline{\delta}}{2} \|d_{D^k}(x^k)\|^2 \quad \forall k, \tag{32}$$

this implies that $\{d_{D^k}(x^k)\}_{\mathcal{K}} \to 0$. Then, arguing as in the proof of (c), we obtain that $\bar{x}$ is a stationary point of F_c.

We prove (31) by contradiction. Suppose that (31) is false, i.e.,

$$q_{D^k}(x^k;\mathcal{J}^k) \le -\delta \quad \forall k \in \mathcal{K}', \tag{33}$$

for some $\delta > 0$ and $\mathcal{K}' \subseteq \mathcal{K}$ with infinitely many elements. We show below that

$$\{P(x^k + \tilde{d}^k) - P(x^k)\}_{\mathcal{K}'} \to 0. \tag{34}$$

Case (1): Suppose P is continuous on $\mathrm{dom}P$. Since $x^k, x^k+\tilde{d}^k \in \mathrm{dom}P$, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, and $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$, (34) readily follows.

Case (2): Suppose $\inf_k \alpha^k > 0$. By (b), $\{\Delta^k\}_{\mathcal{K}'} \to 0$. We also have from $d^k = d_{H^k}(x^k; \mathcal{J}^k)$ and $\tilde{d}^k = d_{D^k}(x^k; \mathcal{J}^k)$ for all k that

$$\begin{aligned}\Delta^k + (\frac{1}{2} - \gamma){d^k}^T H^k d^k &= {g^k}^T d^k + \frac{1}{2}{d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k)\\ &\le {g^k}^T \tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^T H^k \tilde{d}^k + cP(x^k + \tilde{d}^k) - cP(x^k)\\ &\le \frac{1}{2}(\tilde{d}^k)^T H^k \tilde{d}^k - (\tilde{d}^k)^T D^k \tilde{d}^k,\end{aligned}$$

where the last step uses (8) in Lemma 1. Since $\{d^k\}_{\mathcal{K}'} \to 0$ and $\{H^k\}$ is bounded, the left-hand side tends to zero as $k \in \mathcal{K}', k \to \infty$. Since $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$ and $\{D^k\}$ is bounded, the right-hand side tends to zero as $k \in \mathcal{K}', k \to \infty$. Thus the quantity between them also tends to zero as $k \in \mathcal{K}', k \to \infty$. Since f is smooth so that $\{g^k\}_{\mathcal{K}'} \to \nabla f(\bar{x})$, (34) follows.

Case (3): Suppose $\alpha^k_{\mathrm{init}} = 1$ for all k. By further passing to a subsequence if necessary, we can assume that either $\alpha^k = 1$ for all $k \in \mathcal{K}'$ or $\alpha^k < 1$ for all $k \in \mathcal{K}'$. In the first subcase, the same argument as in Case (2) proves (34). In the second subcase, we have from the Armijo rule that $F_c(x^k + d^k) > F_c(x^k) + \sigma\Delta^k$ or, equivalently,

$$f(x^k + d^k) - f(x^k) + (1-\sigma)c(P(x^k + d^k) - P(x^k)) > \sigma({g^k}^T d^k + \gamma {d^k}^T H^k d^k)$$

for all $k \in \mathcal{K}'$. Since $\sigma < 1$, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, $\{d^k\}_{\mathcal{K}'} \to 0$, and $\{H^k\}_{\mathcal{K}'}$ is bounded, this shows that $\liminf_{\substack{k \in \mathcal{K}' \\ k \to \infty}} (P(x^k + d^k) - P(x^k)) \ge 0$. Since

$$0 \ge \Delta^k = {g^k}^T d^k + \gamma {d^k}^T H^k d^k + cP(x^k + d^k) - cP(x^k) \quad \forall k,$$

this in turn yields that $\{\Delta^k\}_{\mathcal{K}'} \to 0$. Then, the same argument as in Case (2) proves (34) also.

We have from (15) that

$$q_{D^k}(x^k; \mathcal{J}^k) = {g^k}^T \tilde{d}^k + \frac{1}{2}(\tilde{d}^k)^T D^k \tilde{d}^k + cP(x^k + \tilde{d}^k) - cP(x^k) \quad \forall k \in \mathcal{K}'.$$

Since f is smooth, $\{x^k\}_{\mathcal{K}'} \to \bar{x}$, $\{\tilde{d}^k\}_{\mathcal{K}'} \to 0$, and $\{D^k\}_{\mathcal{K}'}$ is bounded, this together with (34) yields $\{q_{D^k}(x^k; \mathcal{J}^k)\}_{\mathcal{K}'} \to 0$, contradicting (33).

(e) Suppose that $\{\mathcal{J}^k\}$ is chosen by the generalized Gauss-Seidel rule (11), P is block-separable with respect to $\mathcal{J}^k$ for all k, and $\sup_k \alpha^k < \infty$. The latter implies $\{\alpha^k\}$ is bounded. Suppose that $\bar{x}$ is a cluster point of $\{x^k\}$. Let $\{x^k\}_{\mathcal{K}}$ be a subsequence of $\{x^k\}$ converging to $\bar{x}$. By further passing to a subsequence if necessary, we can assume that $\{H^k\}_{\mathcal{K}} \to$ some $\bar{H}$ and $\mathcal{J}^k = \mathcal{J}$ for all $k \in \mathcal{K}$.

Since $H^k \succeq \underline{\lambda} I$ for all k, we have $\bar{H} \succeq \underline{\lambda} I \succ 0_n$. By the definition of d^k and $\mathcal{J}^k = \mathcal{J}$, we have from (20) that

$$g^k_{\mathcal{J}}{}^T d^k_{\mathcal{J}} + \frac{1}{2}(d^k_{\mathcal{J}})^T H^k_{\mathcal{J}\mathcal{J}} d^k_{\mathcal{J}} + cP_{\mathcal{J}}(x^k_{\mathcal{J}} + d^k_{\mathcal{J}})$$
$$\leq g^k_{\mathcal{J}}{}^T (x - x^k)_{\mathcal{J}} + \frac{1}{2}(x - x^k)^T_{\mathcal{J}} H^k_{\mathcal{J}\mathcal{J}} (x - x^k)_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}) \quad \forall x \in \Re^n.$$

Since $\{d^k\}_{\mathcal{K}} \to 0$ by (b), passing to the limit as $k \in \mathcal{K}, k \to \infty$ and using the smoothness of f and lsc of $P_{\mathcal{J}}$ yields

$$cP_{\mathcal{J}}(\bar{x}_{\mathcal{J}}) \leq \nabla f(\bar{x})^T_{\mathcal{J}}(x - \bar{x})_{\mathcal{J}} + \frac{1}{2}(x - \bar{x})^T_{\mathcal{J}} \bar{H}_{\mathcal{J}\mathcal{J}}(x - \bar{x})_{\mathcal{J}} + cP_{\mathcal{J}}(x_{\mathcal{J}}) \quad \forall x \in \Re^n.$$

This shows that $d_{\bar{H}}(\bar{x}; \mathcal{J}) = 0$ so that, by Lemma 2, $\bar{x}$ is a stationary point of F_c with respect to the components indexed by $\mathcal{J}$, i.e., $F'_c(\bar{x}; d) \geq 0$ for all $d \in \Re^n$ with $d_j = 0$ for $j \notin \mathcal{J}$.

Since $\{d^k\}_{\mathcal{K}} \to 0$, the boundedness of $\{\alpha^k\}_{\mathcal{K}}$ implies $\{x^{k+1}\}_{\mathcal{K}} \to \bar{x}$. This in turn implies $\{d^{k+1}\}_{\mathcal{K}} \to 0$ by (b), and so $\{x^{k+2}\}_{\mathcal{K}} \to \bar{x}$. Continuing in this manner, we obtain that $\{x^{k+\ell}\}_{\mathcal{K}} \to \bar{x}$, for $\ell = 1, ..., T-1$. Thus, we can apply the above argument to $\{x^{k+\ell}\}_{\mathcal{K}}$ to obtain

$$F'_c(\bar{x}; d) \geq 0 \quad \forall d \in \Re^n \text{ with } d_j = 0 \; \forall j \notin \mathcal{J}_\ell, \quad \ell = 0, 1, ..., T-1,$$

where $\mathcal{J}_0, \mathcal{J}_1, ..., \mathcal{J}_{T-1}$ are nonempty subsets of $\mathcal{N}$ whose union equals $\mathcal{N}$; see (11). Since f is differentiable and P is block-separable with respect to $\mathcal{J}_0, \mathcal{J}_1, ..., \mathcal{J}_{T-1}$, this in turn implies that $F'_c(\bar{x}; d) \geq 0$ for all $d \in \Re^n$, so $\bar{x}$ is a stationary point of F_c.

(f) Since α^k is chosen by the Armijo rule, either $\alpha^k = \alpha^k_{\text{init}}$ or else, by Lemma 5(b), $\alpha^k/\beta > \min\{1, 2\underline{\lambda}(1-\sigma+\sigma\gamma)/L\}$. Since $\inf_k \alpha^k_{\text{init}} > 0$, this implies $\inf_k \alpha^k > 0$. If $\lim_{k\to\infty} F_c(x^k) > -\infty$ also, then this and (27) imply $\{\Delta^k\} \to 0$, which together with (26) imply $\{d^k\} \to 0$.

Notice that the assumption $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ in Theorem 1(b), (c), (d) is automatically satisfied if we choose $D^k = I$ or $D^k = \text{diag}(H^k)$ under Assumption 1. Also, the assumption $\sup_k \alpha^k < \infty$ in Theorem 1(e) is automatically satisfied if we choose $\sup_k \alpha^k_{\text{init}} < \infty$. In the case where P is separable, in addition to being proper convex lsc, P is automatically continuous on $\text{dom} P$ [48, Corollary 2.37].

To our knowledge, Theorem 1 is new even in the unconstrained smooth case (i.e., $P \equiv 0$). Theorem 1(e) shows that the CGD method has stronger global convergence properties than the coordinate minimization method when both update coordinates in a Gauss-Seidel manner. In particular, the CGD method cannot cycle on Powell's example [43].

If we choose $\mathcal{J}^k = \mathcal{N}$ and $H^k = \lambda^k I$ for all k with $\bar{\lambda} \geq \lambda^k \geq \underline{\lambda} > 0$, then the CGD method is closely related to the method of Fukushima and Mine [21]. Since $\mathcal{J}^k$ satisfies (14), Theorem 1(c) implies that every cluster point of $\{x^k\}$ is a stationary point of F_c. In contrast, the convergence result in [21, Theorem 4.1] further assumes that ∇f has a Lipschitz property and $P'(x; \cdot)$ has a continuity property.

5. Convergence Rate Analysis

In this section we analyze the asymptotic convergence rate of the CGD method under the following assumption, analogous to that made for constrained smooth optimization [29]. In what follows, $\bar{X}$ denotes the set of stationary points of F_c and

$$\mathrm{dist}(x, \bar{X}) = \min_{\bar{x} \in \bar{X}} \|x - \bar{x}\| \qquad \forall x \in \Re^n.$$

Assumption 2 *(a) $\bar{X} \neq \emptyset$ and, for any $\zeta \geq \min_x F_c(x)$, there exist scalars $\tau > 0$ and $\epsilon > 0$ such that*

$$\mathrm{dist}(x, \bar{X}) \leq \tau \|d_I(x)\| \quad \textit{whenever} \quad F_c(x) \leq \zeta,\ \|d_I(x)\| \leq \epsilon. \tag{35}$$

(b) There exists a scalar $\delta > 0$ such that

$$\|x - y\| \geq \delta \quad \textit{whenever} \quad x \in \bar{X},\ y \in \bar{X},\ F_c(x) \neq F_c(y).$$

Assumption 2 is a generalization of Assumptions A and B in [29] for constrained smooth problems. Assumption 2(a) is a local Lipschitzian error bound assumption, saying that the distance from x to $\bar{X}$ is locally in the order of the norm of the residual at x. Error bounds of this kind have been extensively studied. Assumption 2(b) says that the isocost surfaces of F_c restricted to the solution set $\bar{X}$ are "properly separated." Assumption 2(b) holds automatically if f is a convex function. It also holds if f is quadratic and P is polyhedral, as can be seen by applying [26, Lemma 3.1] to (5).

Our analysis will use ideas from the proof in [29, Appendix] for smooth constrained problems. However, the nonsmooth nature of the objective function F_c requires new proof ideas. In particular, the proof in [29, Appendix] relies on using the error bound (35) to derive an inequality like

$$F_c(x^{k+1}) - \bar{v} \leq \tau' \|x^{k+1} - x^k\|^2$$

for all k sufficiently large, where $\tau' > 0$ and $\bar{v} = \lim_{k \to \infty} F_c(x^k)$; see [29, page 175]. For the nonsmooth case, we cannot derive this same inequality but instead work with a weaker inequality whereby the quadratic term $\|x^{k+1} - x^k\|^2$ is replaced by $-\Delta^k$.

The next two theorems establish, under Assumptions 1–2 and (22), the linear convergence of the CGD method using either the restricted Gauss-Seidel rule or the Gauss-Southwell-q rule to choose $\{\mathcal{J}^k\}$. Their proofs use Theorem 1 and Lemmas 1, 3, 4, 5(a). In what follows, by Q-linear and R-linear convergence, we mean linear convergence in the quotient and the root sense, respectively [42, Chapter 9].

Theorem 2. *Assume that f satisfies (22) for some $L \geq 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 1, where $\{\mathcal{J}^k\}$ is chosen by the restricted Gauss-Seidel rule (12) with $\mathcal{T} \subseteq \{0, 1, ...\}$. Then the following results hold.*

(a) $\|d_I(x^k)\| \le \sup_j \alpha^j\ Cr^k$ *for all* $k \in \mathcal{T}$, *where* $r^k = \sum_{\ell=k}^{\tau(k)-1} \|d^\ell\|$ *and* $C > 0$ *depends on* $n, L, \underline{\lambda}, \bar{\lambda}$.

(b) If F_c *satisfies Assumption 2,* P *is block-separable with respect to* $\mathcal{J}^k$ *for all* k, *and* $\{\alpha^k\}$ *is chosen by the Armijo rule with* $\sup_k \alpha^k_{\rm init} \le 1$ *and* $\inf_k \alpha^k_{\rm init} > 0$, *then either* $\{F_c(x^k)\} \downarrow -\infty$ *or* $\{F_c(x^k)\}_{\mathcal{T}}$ *converges at least Q-linearly and* $\{x^k\}_{\mathcal{T}}$ *converges at least R-linearly.*

Proof. (a) Let $g^k = \nabla f(x^k)$ for all k. For each $k \in \mathcal{T}$, we have from (12) and (13) that

$$\|d_I(x^k)\| = \sqrt{\sum_{\ell=k}^{\tau(k)-1} \|d_I(x^k; \mathcal{J}^\ell)\|^2} \ \le \sum_{\ell=k}^{\tau(k)-1} \|d_I(x^k; \mathcal{J}^\ell)\|.$$

Since $x^\ell_{\mathcal{J}^\ell} = x^k_{\mathcal{J}^\ell}$, we obtain from Lemma 4 with $h(u) = \|u\|^2/2$, $p = 2$, $\rho = 1$, $\mathcal{J} = \mathcal{J}^\ell$, $\bar{d} = d_I(x^k; \mathcal{J}^\ell)$, $\tilde{d} = d_I(x^\ell; \mathcal{J}^\ell)$, $\bar{g} = g^k$, $\tilde{g} = g^\ell$ that

$$\|d_I(x^\ell; \mathcal{J}^\ell) - d_I(x^k; \mathcal{J}^\ell)\| \le \|g^\ell_{\mathcal{J}^\ell} - g^k_{\mathcal{J}^\ell}\| \le L\|x^\ell - x^k\|,$$

where the second inequality uses (22) and $x^\ell, x^k \in {\rm dom} P$. Combining the above two relations and using triangle inequality yield

$$\begin{aligned}
\|d_I(x^k)\| &\le \sum_{\ell=k}^{\tau(k)-1} \left(\|d_I(x^\ell; \mathcal{J}^\ell)\| + L\|x^\ell - x^k\|\right) \\
&\le \sum_{\ell=k}^{\tau(k)-1} \left(\theta\|d_{H^\ell}(x^\ell; \mathcal{J}^\ell)\| + L\|x^\ell - x^k\|\right) \\
&\le \sum_{\ell=k}^{\tau(k)-1} \left(\theta\|d^\ell\| + L\sum_{j=k}^{\ell-1} \alpha^j \|d^j\|\right),
\end{aligned}$$

where the second step uses Lemma 3 with $H = H^\ell$ and $\tilde{H} = I$, and we denote $\theta = (1 + 1/\underline{\lambda} + \sqrt{1 - 2/\bar{\lambda} + 1/\underline{\lambda}^2})\bar{\lambda}/2$; the last step uses $\|x^\ell - x^k\| = \|\sum_{j=k}^{\ell-1} \alpha^j d^j\| \le \sum_{j=k}^{\ell-1} \alpha^j \|d^j\|$. Since $\tau(k) - k \le n$, this yields the desired result.

(b) By Theorem 1(a), $\{F_c(x^k)\}$ is nonincreasing. Thus either $\{F_c(x^k)\} \downarrow -\infty$ or $\lim_{k\to\infty} F_c(x^k) > -\infty$. Suppose the latter. Since α^k is chosen by the Armijo rule with $\inf_k \alpha^k_{\rm init} > 0$, Theorem 1(f) implies $\{d^k\} \to 0$. Since $\tau(k) - k \le n$ for all $k \in \mathcal{T}$, this implies that $\{r^k\}_{\mathcal{T}} \to 0$ and hence, by (a), $\{d_I(x^k)\}_{\mathcal{T}} \to 0$. Since $\{F_c(x^k)\}$ is nonincreasing, this implies that $F_c(x^k) \le F_c(x^0)$ and $\|d_I(x^k)\| \le \epsilon$ for all $k \in \mathcal{T}$ with $k \ge$ some $\bar{k}$. Then, by (a) and Assumption 2(a), we have

$$\|x^k - \bar{x}^k\| \le \tau' r^k \quad \forall k \in \mathcal{T},\ k \ge \bar{k}, \tag{36}$$

where $\tau' > 0$ and $\bar{x}^k \in \bar{X}$ satisfies $\|x^k - \bar{x}^k\| = {\rm dist}(x^k, \bar{X})$. Since $\{r^k\}_{\mathcal{T}} \to 0$, this implies $\{x^k - \bar{x}^k\}_{\mathcal{T}} \to 0$. Since $\{x^{k+1} - x^k\} = \{\alpha^k d^k\} \to 0$, this and Assumption

2(b) imply that $\{\bar{x}^k\}_{\mathcal{T}}$ eventually settles down at some isocost surface of F_c, i.e., there exist an index $\hat{k} \geq \bar{k}$ and $\bar{v} \in \Re$ such that

$$F_c(\bar{x}^k) = \bar{v} \quad \forall k \in \mathcal{T},\ k \geq \hat{k}. \tag{37}$$

Fix any index $k \in \mathcal{T}$ with $k \geq \hat{k}$. Since $\bar{x}^k$ is a stationary point of F_c, we have

$$\nabla f(\bar{x}^k)^T(x^k - \bar{x}^k) + cP(x^k) - cP(\bar{x}^k) \geq 0.$$

We also have from the Mean Value Theorem that

$$f(x^k) - f(\bar{x}^k) = \nabla f(\psi^k)^T(x^k - \bar{x}^k),$$

for some ψ^k lying on the line segment joining x^k with $\bar{x}^k$. Since $x^k, \bar{x}^k$ lie in the convex set $\mathrm{dom}P$, so does ψ^k. Combining these two relations and using (37), we obtain

$$\begin{aligned} \bar{v} - F_c(x^k) &\leq (\nabla f(\bar{x}^k) - \nabla f(\psi^k))^T(x^k - \bar{x}^k) \\ &\leq \|\nabla f(\bar{x}^k) - \nabla f(\psi^k)\|\|x^k - \bar{x}^k\| \\ &\leq L\|x^k - \bar{x}^k\|^2, \end{aligned}$$

where the last inequality uses (22), the convexity of $\mathrm{dom}P$, and $\|\psi^k - \bar{x}^k\| \leq \|x^k - \bar{x}^k\|$. This together with $\{x^k - \bar{x}^k\}_{\mathcal{T}} \to 0$ proves that

$$\liminf_{\substack{k \in \mathcal{T} \\ k \to \infty}} F_c(x^k) \geq \bar{v}. \tag{38}$$

Fix any $k \in \mathcal{T}$. For $\ell \in \{k, k+1, ..., \tau(k)-1\}$, we have from the Armijo rule (9) that

$$F_c(x^{\ell+1}) - F_c(x^\ell) \leq \sigma\alpha^\ell\Delta^\ell.$$

Summing this over $\ell = k, k+1, ..., \tau(k)-1$ yields that

$$F_c(x^{\tau(k)}) - F_c(x^k) \leq \sum_{\ell=k}^{\tau(k)-1} \sigma\alpha^\ell\Delta^\ell. \tag{39}$$

Also, using (20) and letting $\xi^\ell = P_{\mathcal{J}^\ell}(x^{\tau(k)}_{\mathcal{J}^\ell})$, $\bar{\xi}^\ell = P_{\mathcal{J}^\ell}(\bar{x}^k_{\mathcal{J}^\ell})$, we have that, for $k \geq \hat{k}$,

$$\begin{aligned} &F_c(x^{\tau(k)}) - \bar{v} \\ &= f(x^{\tau(k)}) + cP(x^{\tau(k)}) - f(\bar{x}^k) - cP(\bar{x}^k) \\ &= \nabla f(\tilde{x}^k)^T(x^{\tau(k)} - \bar{x}^k) + \sum_{\ell=k}^{\tau(k)-1} \left[c\xi^\ell - c\bar{\xi}^\ell\right] \\ &= (\nabla f(\tilde{x}^k) - g^k)^T(x^{\tau(k)} - \bar{x}^k) + \sum_{\ell=k}^{\tau(k)-1} \left[(g^k - g^\ell)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell}\right] \end{aligned}$$

$$
\begin{aligned}
&- \sum_{\ell=k}^{\tau(k)-1} (H^\ell d^\ell)^T_{\mathcal{J}^\ell}(x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + \sum_{\ell=k}^{\tau(k)-1} \left[\left(g^\ell + H^\ell d^\ell\right)^T_{\mathcal{J}^\ell} (x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + c\xi^\ell - c\bar{\xi}^\ell \right] \\
&\le L\|\tilde{x}^k - x^k\| \|x^{\tau(k)} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} L\|x^k - x^\ell\| \|x^{\ell+1} - \bar{x}^k\| \\
&\quad + \sum_{\ell=k}^{\tau(k)-1} \bar{\lambda}\|d^\ell\| \|x^{\ell+1} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} \left[\left(g^\ell + H^\ell d^\ell\right)^T_{\mathcal{J}^\ell} (x^{\ell+1} - \bar{x}^k)_{\mathcal{J}^\ell} + c\xi^\ell - c\bar{\xi}^\ell \right] \\
&\le L\|\tilde{x}^k - x^k\| \|x^{\tau(k)} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} L\|x^\ell - x^k\| \|x^{\ell+1} - \bar{x}^k\| \\
&\quad + \sum_{\ell=k}^{\tau(k)-1} \bar{\lambda}\|d^\ell\| \|x^{\ell+1} - \bar{x}^k\| + \sum_{\ell=k}^{\tau(k)-1} (\alpha^\ell - 1) \left[(1-\gamma) {d^\ell}^T H^\ell d^\ell + \Delta^\ell \right], \qquad (40)
\end{aligned}
$$

where the second step uses the Mean Value Theorem with $\tilde{x}^k$ a point lying on the segment joining $x^{\tau(k)}$ with $\bar{x}^k$; the third step uses (12) and $x^{\tau(k)}_{\mathcal{J}^\ell} = x^{\ell+1}_{\mathcal{J}^\ell}$ for $k \le \ell < \tau(k)$; the fourth step uses $\bar{\lambda} I \succeq H^\ell \succ 0_n$, (22), and the convexity of $\mathrm{dom} P$; and the last step uses $\xi^\ell = P_{\mathcal{J}^\ell}(x^{\ell+1}_{\mathcal{J}^\ell})$, $\alpha^k \le \alpha^k_{\text{init}} \le 1$, and Lemma 5(a).

Fix any $k \in \mathcal{T},\ k \ge \hat{k}$. Using the inequalities $\|\tilde{x}^k - x^k\| \le \|x^{\tau(k)} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{\ell+1} - \bar{x}^k\| \le \|x^{\ell+1} - x^k\| + \|x^k - \bar{x}^k\|$, $\|x^{\ell+1} - x^k\| \le \sum_{j=k}^{\ell} \alpha^j \|d^j\|$ for $k \le \ell < \tau(k)$, we see from (36) and $\alpha^j \le 1$ that the right-hand side of (40) is bounded above by

$$
C_1 \sum_{\ell=k}^{\tau(k)-1} \|d^\ell\|^2 + \sum_{\ell=k}^{\tau(k)-1} (\alpha^\ell - 1) \left[(1-\gamma) {d^\ell}^T H^\ell d^\ell + \Delta^\ell \right]
$$

for some constant $C_1 > 0$ depending on $L, \tau', n, \bar{\lambda}$ only. Since, by (26), we have $-\Delta^\ell \ge (1-\gamma) {d^\ell}^T H^\ell d^\ell \ge (1-\gamma)\underline{\lambda}\|d^\ell\|^2$, the above quantity is bounded above by

$$
-C_2 \sum_{\ell=k}^{\tau(k)-1} \Delta^\ell
$$

for some constant $C_2 > 0$ depending on $L, \tau', n, \bar{\lambda}, \underline{\lambda}, \gamma$ only. Combining this with (39), (40), and $\inf_k \alpha^k > 0$ (see Theorem 1(f)) yields

$$
F_c(x^{\tau(k)}) - \bar{v} \le C_3 (F_c(x^k) - F_c(x^{\tau(k)})) \quad \forall k \in \mathcal{T},\ k \ge \hat{k},
$$

where $C_3 = C_2/(\sigma \inf_k \alpha^k)$. Upon rearranging terms and using (38), we have

$$
0 \le F_c(x^{\tau(k)}) - \bar{v} \le \frac{C_3}{1+C_3} (F_c(x^k) - \bar{v}) \quad \forall k \in \mathcal{T},\ k \ge \hat{k},
$$

so $\{F_c(x^k)\}_{\mathcal{T}}$ converges to $\bar{v}$ at least Q-linearly.

Finally, (26) implies $\Delta^\ell \le (\gamma-1)\underline{\lambda}\|d^\ell\|^2$, so that (39) and $x^{\ell+1}-x^\ell=\alpha^\ell d^\ell$ yield

$$\sigma(1-\gamma)\underline{\lambda}\sum_{\ell=k}^{\tau(k)-1}\frac{\|x^{\ell+1}-x^\ell\|^2}{\alpha^\ell}\le F_c(x^k)-F_c(x^{\tau(k)})\quad\forall k\in\mathcal{T},\ k\ge\hat{k}.$$

This implies

$$\begin{aligned}\|x^{\tau(k)}-x^k\|&\le\sqrt{(\tau(k)-k)\sum_{\ell=k}^{\tau(k)-1}\|x^{\ell+1}-x^\ell\|^2}\\&\le\sqrt{n\frac{(\sup_\ell\alpha^\ell)}{\sigma(1-\gamma)\underline{\lambda}}(F_c(x^k)-F_c(x^{\tau(k)}))}\quad\forall k\in\mathcal{T},\ k\ge\hat{k}.\end{aligned}$$

Since $\{F_c(x^k)-F_c(x^{\tau(k)})\}_{\mathcal{T}}\to 0$ at least R-linearly and $\sup_\ell\alpha^\ell\le 1$, this implies that $\{x^k\}_{\mathcal{T}}$ converges at least R-linearly.[2]

Theorem 3. *Assume that f satisfies* (22) *for some $L\ge 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method satisfying Assumption 1, where $\{\mathcal{J}^k\}$ is chosen by Gauss-Southwell-q rule* (16) *with P block-separable with respect to $\mathcal{J}^k$ and $\bar{\delta}I\succeq D^k\succeq\underline{\delta}I$ for all k ($0<\underline{\delta}\le\bar{\delta}$). If F_c satisfies Assumption 2 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k\alpha^k_{\rm init}\le 1$ and $\inf_k\alpha^k_{\rm init}>0$, then either $\{F_c(x^k)\}\downarrow-\infty$ or $\{F_c(x^k)\}$ converges at least Q-linearly and $\{x^k\}$ converges at least R-linearly.*

Proof. For brevity, the proof is omitted. See the unpublished report [56].

The assumption (22) in Theorems 2 and 3 can be relaxed to ∇f being Lipschitz continuous on $\mathrm{dom}P\cap(X^0+\varrho B)$ for some $\varrho>0$, where B denotes the unit Euclidean ball in $\Re^n$ and X^0 denotes the convex hull of the level set $\{x\mid F_c(x)\le F_c(x^0)\}$. For simplicity, we did not consider this more relaxed assumption.

As we noted in Section 2, we have been unable to establish the linear convergence of the CGD method using the Gauss-Southwell-r rule to choose $\{\mathcal{J}^k\}$. Only in the simple case where f and P are separable, in addition to the assumptions of Theorem 3, have we been able to prove linear convergence. In fact, even in this case our proof is nontrivial, even though the problem decomposes into n univariate problems. This is because different coordinates can converge at different rates, which needs to be explicitly taken into account in the proof.

[2] More precisely, writing $\mathcal{T}=\{k_1,k_2,\ldots\}$, we have $\|x^{k_t+1}-x^{k_t}\|=O\left(\sqrt{F_c(x^{k_t})-F_c(x^{k_t+1})}\right)=O(\vartheta^t)$ for $t=1,2,\ldots$, where $\vartheta=\sqrt{\frac{C_3}{1+C_3}}$. Thus $\{x^{k_t}\}_{t=1,2,\ldots}$ satisfies Cauchy's criterion for convergence, implying it has a unique limit $\bar{x}$. Moreover, for any $t'>t$, we have $\|x^{k_t}-x^{k_{t'}}\|\le\sum_{j=t}^{t'-1}\|x^{k_j+1}-x^{k_j}\|\le O(\sum_{j=t}^{t'-1}\vartheta^j)=O(\vartheta^t)$. Taking $t'\to\infty$ yields $\|x^{k_t}-\bar{x}\|=O(\vartheta^t)$ for any t, so $\limsup_{t\to\infty}\|x^{k_t}-\bar{x}\|^{1/t}\le\vartheta<1$.

6. Error Bound

In this section we show that Assumption 2(a) is satisfied under problem assumptions analogous to those for constrained smooth optimization. In fact, we will show that error bound for (3) is closely related to that for constrained smooth optimization problems.

By using $\mathrm{epi}P = \{(x,\xi)|P(x) \le \xi\}$, we can reformulate (1) as the constrained smooth optimization problem (see (5)):

$$\min_{(x,\xi)} \{ f(x) + c\xi \mid (x,\xi) \in \mathrm{epi}P \}. \tag{41}$$

For any $(x,\xi) \in \mathrm{epi}P$, the corresponding projection residual is the optimal solution of the subproblem:

$$\min_{(d,\delta)} \left\{ \nabla f(x)^T d + \frac{1}{2}\|d\|^2 + \frac{1}{2}\delta^2 + c\delta \mid (x+d, \xi+\delta) \in \mathrm{epi}P \right\}. \tag{42}$$

The following lemma shows that if P is Lipschitz continuous on $\mathrm{dom}P$, then the norm of this projection residual is bounded above by a multiple of $\|d_I(x)\|$ whenever $\xi = P(x)$.

Lemma 6. *Suppose that P is Lipschitz continuous on* $\mathrm{dom}P$. *There exists a scalar $\kappa > 0$ (depending only on the Lipschitz constant of P) such that, for any $x \in \mathrm{dom}P$ and $\xi = P(x)$,*

$$\|(\tilde{d}, \tilde{\delta})\| \le \kappa \|d_I(x)\|,$$

where $(\tilde{d}, \tilde{\delta})$ is an optimal solution of the subproblem (42).

Proof. Fix any $x \in \mathrm{dom}P$ and $\xi = P(x)$. By (6) and (13), $(d_I(x), \bar{\delta})$ is the optimal solution of the subproblem:

$$\min_{(d,\delta)} \left\{ \nabla f(x)^T d + \frac{1}{2}\|d\|^2 + c\delta \mid (x+d, \xi+\delta) \in \mathrm{epi}P \right\},$$

where we let $\bar{\delta} = P(x + d_I(x)) - P(x)$. By Fermat's rule [48, Theorem 10.1],

$$(d_I(x), \bar{\delta}) \in \arg\min_{(d,\delta)} \{ (\nabla f(x) + d_I(x))^T d + c\delta \mid (x+d, \xi+\delta) \in \mathrm{epi}P \}.$$

Hence

$$(\nabla f(x) + d_I(x))^T d_I(x) + c\bar{\delta} \le (\nabla f(x) + d_I(x))^T \tilde{d} + c\tilde{\delta}.$$

Also, since $(\tilde{d}, \tilde{\delta})$ is the optimal solution of the subproblem (42), we have

$$\nabla f(x)^T \tilde{d} + \frac{1}{2}\|\tilde{d}\|^2 + \frac{1}{2}\tilde{\delta}^2 + c\tilde{\delta} \le \nabla f(x)^T d_I(x) + \frac{1}{2}\|d_I(x)\|^2 + \frac{1}{2}\bar{\delta}^2 + c\bar{\delta}.$$

Adding the above two inequalities and simplifying yield

$$\frac{1}{2}\|d_I(x)\|^2 - d_I(x)^T \tilde{d} + \frac{1}{2}\|\tilde{d}\|^2 + \frac{1}{2}\tilde{\delta}^2 \le \frac{1}{2}\bar{\delta}^2.$$

Multiplying both sides by 2 and rewriting the first three terms into a square, we have

$$\|d_I(x) - \tilde{d}\|^2 + \tilde{\delta}^2 \le \bar{\delta}^2.$$

Thus $\tilde{\delta}^2 \le \bar{\delta}^2$ and $\|d_I(x) - \tilde{d}\|^2 \le \bar{\delta}^2$. Taking square root of both sides and using the triangle inequality yield

$$|\tilde{\delta}| \le |\bar{\delta}|, \qquad \|\tilde{d}\| - \|d_I(x)\| \le |\bar{\delta}|. \tag{43}$$

Now, the Lipschitz continuity of P on dom$P$ implies that $|\bar{\delta}| = |P(x + d_I(x)) - P(x)| \le K\|d_I(x)\|$, where K is the Lipschitz constant. Then (43) yields that

$$|\tilde{\delta}| \le K\|d_I(x)\|, \qquad \|\tilde{d}\| \le (K+1)\|d_I(x)\|,$$

which proves the desired result.

The following local error bound results from [26–28,44] show that, for all x sufficiently close to $\bar{X}$, dist$(x, \bar{X})$ can be bounded from above by the norm of the solution of the subproblem (42) under certain problem assumptions.

Lemma 7. *Assume that $\bar{X} \ne \emptyset$ and any of the following conditions hold.*

C1 f is quadratic. P is polyhedral.

C2 $f(x) = g(Ex) + q^T x$ for all $x \in \Re^n$, where $E \in \Re^{m\times n}$, $q \in \Re^n$, and g is a strongly convex differentiable function on $\Re^m$ with ∇g Lipschitz continuous on $\Re^m$. P is polyhedral.

C3 $f(x) = \max_{y\in Y}\{(Ex)^T y - g(y)\} + q^T x$ for all $x \in \Re^n$, where Y is a polyhedral set in $\Re^m$, $E \in \Re^{m\times n}$, $q \in \Re^n$, and g is a strongly convex differentiable function on $\Re^m$ with ∇g Lipschitz continuous on $\Re^m$. P is polyhedral.

Then, for any $\zeta \in \Re$, there exist scalars $\tau' > 0$ and $\epsilon' > 0$ such that

$$\text{dist}(x, \bar{X}) \le \tau'\|(\tilde{d}, \tilde{\delta})\| \quad \textit{whenever} \quad F_c(x) \le \zeta,\ \|(\tilde{d}, \tilde{\delta})\| \le \epsilon', \tag{44}$$

where $(\tilde{d}, \tilde{\delta})$ is the optimal solution of the subproblem (42) with $\xi = P(x)$.

Proof. Since epi$P$ is convex, each stationary point $(\bar{x}, \bar{\xi})$ of (41) satisfies

$$\nabla f(\bar{x})^T(x - \bar{x}) + c(\xi - \bar{\xi}) \ge 0 \quad \forall (x, \xi) \in \text{epi}P,$$

from which it readily follows that $\bar{\xi} = P(\bar{x})$ and $\bar{x} \in \bar{X}$. Under C1, the objective function of (41) is quadratic and epi$P$ is a polyhedral set. Fix any $\zeta \in \Re$. By applying [26, Theorem 2.3] (also see [44]) to (41), there exist scalars $\tau' > 0$ and $\epsilon' > 0$ such that

$$\min_{\bar{x}\in\bar{X}} \|(x, P(x)) - (\bar{x}, P(\bar{x}))\| \ \le\ \tau'\|(\tilde{d}, \tilde{\delta})\| \quad \text{whenever} \quad F_c(x) \le \zeta,\ \|(\tilde{d}, \tilde{\delta})\| \le \epsilon',$$

where $(\tilde{d}, \tilde{\delta})$ is the optimal solution of (42) with $\xi = P(x)$. Since $\|x - \bar{x}\| \le \|(x, P(x)) - (\bar{x}, P(\bar{x}))\|$ for all $\bar{x} \in \bar{X}$, this proves (44).

Under C2, the objective function of (41) has the form $g\left([E\ 0]\begin{bmatrix} x \\ \xi \end{bmatrix}\right) + [q^T\ c]\begin{bmatrix} x \\ \xi \end{bmatrix}$ and epi$P$ is a polyhedral set. Then, by applying [27, Theorem 2.1] to (41) and arguing similarly as above, (44) can be proved. Under C3, a similar argument using [28, Theorem 4.1] (also see [29, Theorem 2.1]) proves (44).

By using Lemmas 6 and 7, we obtain the main result of this section.

Theorem 4. *Assumption 2(a) is satisfied if $\bar{X} \neq \emptyset$ and any of the conditions C1, C2, C3 in Lemma 7 holds or if the following condition holds.*

C4 f is strongly convex and satisfies (22) *for some $L \geq 0$.*

Proof. Under C1 or C2 or C3, P is polyhedral so, by Example 9.35 in [48], P is Lipschitz continuous on dom$P$. Then Lemmas 6 and 7 yield that Assumption 2(a) holds.

Under C4, for any $x \in \mathrm{dom}P$, since $d_I(x)$ is a solution of the subproblem (6) with $\mathcal{J} = \mathcal{N}$ and $H = I$, by Fermat's rule [48, Theorem 10.1],

$$d_I(x) \in \arg\min_d \ (\nabla f(x) + d_I(x))^T d + cP(x+d).$$

Hence, for any $\bar{x} \in \bar{X}$ (in fact, $\bar{X}$ is a singleton), we have

$$(\nabla f(x) + d_I(x))^T d_I(x) + cP(x + d_I(x)) \leq (\nabla f(x) + d_I(x))^T (\bar{x} - x) + cP(\bar{x}).$$

Since $\bar{x}$ is a stationary point of F_c, we also have

$$cP(\bar{x}) \leq \nabla f(\bar{x})^T (x + d_I(x) - \bar{x}) + cP(x + d_I(x)).$$

Adding the above two inequalities and simplifying yield

$$(\nabla f(x) - \nabla f(\bar{x}))^T (x - \bar{x}) + \|d_I(x)\|^2 \leq (\nabla f(\bar{x}) - \nabla f(x))^T d_I(x) + d_I(x)^T (\bar{x} - x).$$

It follows from the strong convexity of f and (22) that

$$\lambda \|x - \bar{x}\|^2 + \|d_I(x)\|^2 \leq L\|x - \bar{x}\| \|d_I(x)\| + \|x - \bar{x}\| \|d_I(x)\|,$$

for some scalar constants $0 < \lambda \leq L$. Thus

$$\lambda \|x - \bar{x}\|^2 \leq (L+1)\|x - \bar{x}\| \|d_I(x)\|.$$

Dividing both sides by $\lambda \|x - \bar{x}\|$ whenever $x \neq \bar{x}$ shows that Assumption 2(a) is satisfied with $\tau = (L+1)/\lambda$ and $\epsilon = \infty$ (independent of ζ).

Notice that the objective function of (41) is not strongly convex under C4. Thus existing error bound results for strongly convex objective function (e.g., [14, Proposition 6.3.1]) cannot be applied to (41).

7. Implementation and Numerical Experience

In order to better understand its practical performance, we have implemented the CGD method in Matlab, using Matlab's vector operations, to solve the ℓ_1-regularized problem (3). In this section, we describe the implementation, together with convergence acceleration techniques, and report our numerical experience on test problems with $n = 1000$ from Moré et al. [37] and the CUTEr set [22]. Comparison with MINOS [39] and L-BFGS-B [59], applied to a reformulation of (3) as a bound constrained optimization problem, is also reported.

7.1. Test functions

For the function f in (3), we chose 10 test functions with $n = 1000$ from the set of nonlinear least square functions used by Moré et al. [37]. These functions, listed in Table 1 with the numbering from [37, pages 26-28] shown in parentheses, were chosen for their diverse characteristics. The functions ER, DBV, BT are nonconvex, with sparse Hessian; TRIG, BAL are nonconvex, with dense Hessian; EPS is convex, with sparse Hessian; VD, LFR are strongly convex with dense Hessian; LR1, LR1Z are convex quadratic with dense Hessian of rank 1. Both ER and EPS have block-diagonal Hessians, while VD, LR1, LR1Z have ill-conditioned Hessians. Since we wish to see how solution sparsity (i.e., number of nonzeros) changes with c, we modified the Extended Powell singular function slightly, replacing "$5^{1/2}(x_{4i-1} - x_{4i})$" with "$5^{1/2}(x_{4i-1} - x_{4i} - 1)$" so that the solution is not always at the origin. We coded the function, gradient, and Hessian diagonals in Matlab using vector operations.

We also chose 10 functions with $n = 1000$ from the unconstrained problems in the CUTEr set [22]. These functions, listed in Table 3, were similarly chosen for their diverse characteristics, as well as Hessian availability. The function, gradient, and (sparse) Hessian are called within Matlab using the CUTEr tools "ufr", "ugr" and "ush".

7.2. Implementation of the CGD method

In our implementation of the CGD method, we choose a diagonal Hessian approximation

$$H^k = \operatorname{diag}\left[\min\{\max\{\nabla^2 f(x^k)_{jj}, 10^{-2}\}, 10^9\}\right]_{j=1,\ldots,n},$$

which has the advantage that d^k has a closed form and can be computed efficiently in Matlab using vector operations. We tested the alternative choice of $H^k = I$, which does not require Hessian evaluation, but its overall performance was worse. If Hessian computation is expensive, a compromise would be to recompute the Hessian diagonal once every few iterations. We choose the index subset $\mathcal{J}^k$ by either (i) the restricted Gauss-Seidel rule (12), whereby $\mathcal{N}$ is partitioned into $n_b \in \{5, 10, n\}$ subsets with $\lceil n/n_b \rceil$ elements each (except possibly the last subset which may have fewer elements) and $\mathcal{J}^k$ cycles through these subsets, or (ii) the Gauss-Southwell-r rule (14) with $D^k = H^k$,

$$\mathcal{J}^k = \left\{j \mid |d_{D^k}(x^k; j)| \geq v^k \|d_{D^k}(x^k)\|_\infty\right\},$$
$$v^{k+1} = \begin{cases} \max\{10^{-4}, v^k/10\} & \text{if } \alpha^k > 10^{-3} \\ \min\{.9, 50v^k\} & \text{if } \alpha^k < 10^{-6} \\ v^k & \text{else} \end{cases}$$

(initially $v^0 = .5$) or (iii) the Gauss-Southwell-q rule (16) with $D^k = H^k$,

$$\mathcal{J}^k = \left\{j \mid q_{D^k}(x^k; j) \leq v^k \min_i q_{D^k}(x^k; i)\right\},$$

and v^k chosen as in (ii). The above updating formula for v^k is guided by the observation that smaller v^k results in more coordinates being updated but a smaller stepsize α^k, while a larger v^k has the opposite effect. Thus if α^k is large, we decrease v^k and if α^k is small, we increase v^k. The thresholds 10^{-3} and 10^{-6} were found after some experimentation to work well on our test problems. The stepsize α^k is chosen by the Armijo rule (9) with

$$\sigma = .1, \quad \beta = .5, \quad \gamma = 0, \qquad \alpha^0_{\text{init}} = 1, \quad \alpha^k_{\text{init}} = \min\left\{\frac{\alpha^{k-1}}{\beta}, 1\right\} \ \forall k \geq 1.$$

We experimented with other values of $0 \leq \gamma < 1$, but the cpu times and the number of iterations did not change appreciably in our tests.

Each CGD iteration requires 1 gradient evaluation and 1 Hessian diagonal evaluation to find the direction d^k, and at least 1 function evaluation to find the stepsize α^k. These are the dominant computations. For the CUTEr test functions, Hessian evaluation is the most dominant computation when the Hessian is dense. (CUTEr does not offer the option of evaluating only the Hessian diagonals.)

Since H^k is diagonal, the CGD method resembles a block coordinate version of a diagonally scaled steepest descent method [4, page 71] when $c = 0$. As such, the convergence rate of the method is likely slow when the Hessian $\nabla^2 f(x^k)$ is far from being diagonally dominant, as was observed on some of the functions from Table 1, such as LR1, LR1Z, and VD. This motivated us to introduce two techniques to accelerate the convergence, which we describe below.

The first technique uses an active-set identification strategy of Facchinei, Fischer, and Kanzow [13] (also see [14, Section 6.7]) to estimate which components of x would be nonzero at a solution and then uses a fast method for unconstrained smooth optimization to update these components. The method we chose is the limited-memory BFGS (L-BFGS) method of Nocedal [40,41]. In particular, we store the m ($m \geq 1$) most recent pairs of Δx and Δg that make sufficiently acute angles. More precisely, we store $\Delta x^k = x^k - x^{k-1}$ and $\Delta g^k = g^k - g^{k-1}$ (with $g^k = \nabla f(x^k)$) whenever

$$\|\Delta g^k\| > 10^{-20}, \qquad \frac{\Delta x^{k\,T} \Delta g^k}{\|\Delta g^k\|^2} > \frac{10^{-10}}{\max_j H^k_{jj}}.$$

In an acceleration step at x^k, we use the L-BFGS formula (with $m = 5$) to construct a positive definite Hessian inverse approximation B^k and set

$$d^k_{\mathcal{J}^k} = -B^k_{\mathcal{J}^k \mathcal{J}^k} \nabla_{x_{\mathcal{J}^k}} F_c(x^k), \qquad d^k_j = 0 \ \forall j \notin \mathcal{J}^k,$$

where $\mathcal{J}^k = \left\{ j \mid |x^k_j| > \rho\left(\|d_{H^k}(x^k)\|_\infty\right)\right\}$ with identification function $\rho(t) = \dfrac{-.0001}{\ln\left(\min\{.1, .01t\}\right)}$. We then update $x^{k+1} = x^k + \alpha^k d^k$ with α^k chosen by the Armijo rule with $\sigma = .1$, $\beta = .5$, $\gamma = 0$, and $\alpha^k_{\text{init}} = 1$. This acceleration step is invoked at iteration k whenever $k \geq 10$ and $k < 50 \pmod{100}$. We choose 50

since L-BFGS typically terminates in less than 50 iterations on the test functions when $c = 0$.

The second technique is motivated by the rank-1 Hessian for the functions LR1 and LR1Z. In an acceleration step at x^k, we choose h^k to satisfy the rank-1 secant equation

$$(h^k h^{k^T}) s^k = y^k,$$

where s^k and y^k are the most recently stored pair of Δx and Δg. This yields $h^k = y^k / \sqrt{s^{k^T} y^k}$. We next solve the subproblem with rank-1 Hessian

$$\min_d \; g^{k^T} d + \frac{1}{2}(h^{k^T} d)^2 + c\|x^k + d\|_1.$$

This subproblem need not have an optimal solution (e.g., when $h_j^k = 0$ and $|g_j^k| > c$ for some j), but if it has an optimal solution, then there exists an optimal solution d^k with at most one nonzero component, which can be computed efficiently using Matlab's vector operations. (In general, if the subproblem with rank-p Hessian has an optimal solution, then there exists an optimal solution with at most p nonzero components.) We then update $x^{k+1} = x^k + \alpha^k d^k$ with α^k chosen as in the L-BFGS acceleration step. This second acceleration step is invoked once every 10 consecutive CGD iterations.

We terminate the CGD method when

$$\|H^k d_{H^k}(x^k)\|_\infty \leq 10^{-4}. \tag{45}$$

Here we scale $d_{H^k}(x^k)$ by H^k to reduce its sensitivity to H^k. We can alternatively use the criterion $\|d_I(x^k)\|_\infty \leq 10^{-4}$. The advantage of (45) is that $d_{H^k}(x^k)$ is already computed by the CGD method, unlike $d_I(x^k)$. In a few cases where $\nabla^2 f$ is ill-conditioned, the Armijo descent condition (9) eventually cannot be satisfied by any $\alpha^k > 0$ due to cancellation error in the function evaluations. (In Matlab Version 7.0, floating point subtraction is accurate up to 15 digits only.) In these cases, no further progress is possible so we exit the method when (9) remains unsatisfied after α^k reaches 10^{-30}.

7.3. L-BFGS-B and MINOS

The ℓ_1-regularized problem (3) can be formulated as a bound-constrained smooth optimization problem:

$$\min_{y \geq 0, z \geq 0} \; f(y - z) + c\, e^T (y + z),$$

where e is the vector of 1s, to which many methods can be applied for its solution. Thus, it is of interest to compare the CGD method with such methods. We considered two such methods. One is L-BFGS-B, a Fortran implementation of a limited memory BFGS algorithm for large-scale bound-constrained optimization [59]. A second is MINOS (Version 5.5.1), which has a Fortran implementation

Name	c	CGD-GSeidel	CGD-GS-q	CGD-GS-q-acc
		♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu(CGD/L-BFGS/R1)
BAL	1	>5h	[a]1000/1000.00/.1	1000/1000.00/.2(10/29/1)
(27)	10	>5h	[a]1000/9999.98/.2	1000/9999.97/.1(10/21/1)
	100	>5h	>5h	[b]1000/99997.5/.1(10/18/1)
BT	.1	1000/70.3320/.8	1000/70.3320/.1	1000/70.3320/.1(10/14/1)
(30)	1	1000/671.819/1.0	1000/671.819/.2	1000/671.819/.1(10/19/1)
	10	0/1000.00/.07	0/1000.00/.02	0/1000.00/.03(8/0/1)
DBV	.1	0/0.00000/.9	2/0.00000/.04	0/0.00000/.02(2/0/1)
(28)	1	0/0.00000/.03	2/0.00000/.02	0/0.00000/.02(2/0/1)
	10	0/0.00000/.03	0/0.00000/.02	0/0.00000/.02(2/0/1)
ER	1	1000/436.250/4.9	1000/436.250/.8	1000/436.250/.2(10/38/1)
(21)	10	0/500.000/.4	0/500.000/.1	0/500.000/.3(11/40/1)
	100	0/500.000/.07	0/500.000/.01	0/500.000/.04(8/0/1)
TRIG	.1	0/0.00000/.6	0/0.00000/.1	0/0.00000/.1(12/9/0)
(26)	1	0/0.00000/.08	0/0.00000/.02	0/0.00000/.02(6/0/0)
	10	0/0.00000/.02	0/0.00000/.01	0/0.00000/.01(1/0/0)
EPS	1	1000/351.146/1.4	1000/351.146/.3	1000/351.146/.2(10/30/1)
(22)	10	200/1250.00/.2	250/1250.00/.04	250/1250.00/.05(10/0/1)
	100	0/1250.00/.07	0/1250.00/.01	0/1250.00/.02(2/0/1)
LR1	.1	>5h	>5h	1/249.625/.1(10/0/2)
(33)	1	>5h	>5h	1/249.625/.1(10/0/2)
	10	>5h	>5h	1/249.625/.05(8/0/1)
LR1Z	.1	>5h	>5h	1/251.125/.1(10/0/2)
(34)	1	>5h	>5h	1/251.125/.1(10/0/1)
	10	>5h	>5h	1/251.125/.1(10/0/1)
LFR	.1	[a]1000/98.5000/.04	1000/98.5000/.01	1000/98.5000/.01(1/0/0)
(32)	1	1000/751.000/.02	1000/751.000/.01	1000/751.000/.01(1/0/0)
	10	0/1001.00/.02	0/1001.00/.01	0/1001.00/.01(1/0/0)
VD	1	>5h	>5h	1000/937.594/.6(56/80/5)
(25)	10	>5h	>5h	[b]1000/6726.81/42.6(3791/4199/420)
	100	>5h	>5h	[b]1000/55043.1/106.2(8291/9198/920)

Table 1. Comparing the CGD method using the restricted Gauss-Seidel rule and the Gauss-Southwell-q rule, with or without acceleration steps, on 10 test functions from [37], with x^0 given as in [37].
[a]CGD exited due to the Armijo stepsize reaching 10^{-30}.
[b]CGD exited due to the Armijo stepsize in an L-BFGS acceleration step reaching 10^{-30}.

of an active-set method for linearly constrained smooth optimization [39]. To accommodate problems with $n = 1000$, we set Superbasics limit to $2n + 1$ and Workspace to 5,000,000 in MINOS. The objective function and its gradient are coded in Fortran, with f taken from Table 1. For a given starting point x^0 for (3), we accordingly initialize $y^0 = \max\{x^0, 0\}$ and $z^0 = \max\{-x^0, 0\}$, with the "max" taken componentwise.

7.4. Numerical Results

We now report the performance of the CGD method using either the restricted Gauss-Seidel (GSeidel) rule or the Gauss-Southwell-r (GS-r) rule or the Gauss-Southwell-q (GS-q) rule, with or without the aforementioned acceleration techniques, and we compare it with the performances of L-BFGS-B and MINOS. All runs are performed on an HP DL360 workstation, running Red Hat Linux 3.5 and Matlab (Version 7.0). All Fortran codes are compiled using the Gnu F-77 compiler (Version 3.2.57). Tables 1–3 show the final objective value, the cpu time

Name	c	L-BFGS-B	MINOS	CGD-GS-q-acc
		♯nz/obj/cpu	♯nz/obj/cpu	♯nz/obj/cpu(CGD/L-BFGS/R1)
BAL	1	[c]1000/1000.00/.02	1000/1000.00/49.9	1000/1000.00/.1(10/10/1)
	10	[c]1000/9999.98/.03	1000/9999.97/48.4	1000/9999.98/.2(10/9/1)
	100	[c]1000/99997.5/.1	1000/99997.5/48.9	[b]1000/99997.5/.1(10/15/1)
BT	.1	[c]1000/84.0033/.02	1000/71.725/100.6	1000/71.7481/.9(111/97/0)
	1	[c]981/668.724/.2	997/672.418/94.7	1000/626.670/42.4(4156/4154/5)
	10	0/1000.00/.00	0/1000.00/56.0	0/1000.00/.01(1/0/0)
DBV	.1	[c]999/83.4557/.01	0/0.00000/51.5	0/0.00000/.5(11/40/2)
	1	0/0.00000/.01	0/0.00000/50.8	2/0.00000/.03(3/0/1)
	10	0/0.00000/.00	0/0.00000/52.5	0/0.00000/.01(1/0/0)
ER	1	1000/436.250/.1	1000/436.250/71.5	1000/436.250/.1(10/24/1)
	10	[c]500/1721.15/.00	0/500.000/50.2	0/500.000/.3(11/40/1)
	100	0/500.000/.00	0/500.000/52.4	0/500.000/.03(7/0/1)
TRIG	.1	[c]1000/14.1282/.1	0/0.00000/58.5	1/.626211/.6(29/40/4)
	1	0/0.00000/.1	1/6.21995/62.0	1/6.21364/.5(47/40/6)
	10	0/0.00000/.1	0/0.00000/61.9	1/61.2209/.5(38/40/5)
EPS	1	[c]999/352.526/.05	1000/351.146/60.3	1000/351.146/.3(13/40/2)
	10	1/1250.00/.01	243/1250.00/44.2	249/1250.00/.1(8/0/1)
	100	0/1250.00/.01	0/1250.00/51.5	0/1250.00/.01(2/0/1)
LR1	.1	[c]1000/424.663/.00	[d]2/249.625/59.7	1/249.625/.1(10/0/2)
	1	[c]1000/2000.00/.01	[d]1/249.625/57.2	1/249.625/.1(10/0/2)
	10	[c]1000/17753.4/.01	1/249.625/58.0	1/249.625/.05(8/0/1)
LR1Z	.1	[c]1000/426.087/.00	[d]4/251.125/59.2	1/251.125/.1(10/0/2)
	1	[c]1000/2000.75/.01	[d]3/251.125/58.4	1/251.125/.1(10/0/1)
	10	[c]1000/17747.3/.00	1/251.125/59.7	1/251.125/.1(10/0/1)
LFR	.1	1000/98.5000/.00	1000/98.5000/77.2	1000/98.5000/.01(1/0/0)
	1	1000/751.000/.01	1000/751.000/73.8	1000/751.000/.01(1/0/0)
	10	0/1001.00/.00	0/1001.00/53.3	0/1001.00/.01(1/0/0)
VD	1	[c]1000/1000.00/.00	1000/937.594/43.0	1000/937.594/.5(55/59/6)
	10	[c]974/$5.18\cdot10^{12}$/2.3	413/6726.81/56.9	[b]1000/6726.81/60.3(5140/5698/571)
	100	[c]996/75135.5/.2	136/55043.1/57.4	[b]1000/55043.1/88.1(7030/7804/781)

Table 2. Comparing the CGD method using the Gauss-Southwell-q rule and acceleration steps with L-BFGS-B and MINOS on test functions from Table 1, with $x^0 = (1, 1, ..., 1)^T$.
[b]CGD exited due to the Armijo stepsize in an L-BFGS acceleration step reaching 10^{-30}.
[c]L-BFGS-B exited due to the objective value cannot be improved upon.
[d]MINOS exited due to the current point cannot be improved upon.

(in seconds), and the number of nonzero components (♯nz) in the final solution found. (A component is considered to be nonzero if its absolute value exceeds 10^{-15}.) The number of L-BFGS acceleration steps and rank-1 acceleration steps are also shown. For each function, three different values of c are chosen to track changes in the solution sparsity ♯nz. In Tables 1 and 2, different starting points are used. In our experience, CGD-GS-r and CGD-GS-q have comparable performances, so for brevity we report only the performance of CGD-GS-q in Tables 1 and 2. Also, we found CGD-GSeidel to have better performance with $n_b = 5$ than with $n_b = 10$ or $n_b = n$. Thus we set $n_b = 5$ in our tests.

From Table 1, we see that CGD-GS-q is typically faster than CGD-GSeidel. But CGD-GS-q is still too slow (more than 5 hours of cpu time) on functions whose Hessian are far from being diagonally dominant, namely, BAL, LR1, LR1Z, and VD. By using the acceleration steps, CGD-GS-q-acc shows significantly better performance on these functions. We also tested CGD-GSeidel with acceleration steps, but its performance is not better than CGD-GS-q-acc and so we do not report it here.

From Table 2, we see that CGD-GS-q-acc is competitive with MINOS in terms of solution accuracy (as measured by the final objective value), and is generally faster in terms of cpu time (except on VD). L-BFGS-B is fast, but often exits when still far from a solution with a large projected gradient. This is due to the relative improvement in objective value being below $factr \cdot epsmch$, where $factr = 10^7$ and $epsmch$ is the machine precision generated by the code (about 10^{-19} in our tests). We experimented with $factr$ set to zero but it did not change significantly the results. Similar comparative performances were observed when the starting point was changed to $x^0 = (-1, -1, ..., -1)^T$; see [56, Table 6].

Thus MINOS seems more robust than L-BFGS-B, though it is slower (possibly due to the many active bounds at a solution). For the nonconvex functions BT and TRIG, multiple local minima exist and, depending on the starting point, the methods can converge to different local minima with different objective value.

Table 3 reports the performance of CGD-GS-r-acc and CGD-GS-q-acc on the CUTEr test functions. Both are able to meet the termination criterion (45) in typically under a second, except on NONCVXU2 and PENALTY1. On PENALYT1, the termination tolerance 10^{-4} in (45) was too loose, with the final objective value accurate up to only 1 or 2 significant digits, so we tightened it to 10^{-9}. The final objective value for other functions appear to be accurate up to 5 significant digits, as tightening the tolerance to 10^{-6} did not change them. Notice that, on INDEF, LIARWHD, NONCVXU2, PENALTY1, WOODS, for which f is nonconvex, CGD-GS-r-acc and CGD-GS-q-acc can terminate at different solutions, depending on the starting point x^0.

8. Conclusions and Extensions

We have presented a block coordinate gradient descent method for minimizing the sum of a smooth function and a convex separable function. The method may be viewed as a hybrid of gradient-projection and coordinate descent methods, or as a block coordinate version of descent methods in [6,21]. We analyzed the global convergence and asymptotic convergence rate of the method. We also presented numerical results to verify the practical efficiency of the method.

We can relax the Armijo descent condition (9) by replacing Δ^k with its upper bound $(\gamma - 1){d^k}^T H^k d^k$ (see (26)), i.e.,

$$F_c(x^k + \alpha^k d^k) \leq F_c(x^k) + \alpha^k \sigma(\gamma - 1){d^k}^T H^k d^k. \tag{46}$$

The global convergence analysis in Theorem 1 (except (d)) can be extended accordingly. The convergence rate analysis in Theorem 2 can be similarly extended, provided that $\alpha^k = 1$ for all k sufficiently large (so that the last term in (40) equals zero). Using Lemma 1 and the fact that, under assumption (22), $f(x + d) - f(x) \leq \nabla f(x)^T d + L\|d\|^2/2$ for all $x, x + d \in \mathrm{dom} P$ (see [4, page 667] or the proof of Lemma 5(b)), it is readily seen that the latter holds if we choose $\alpha^k_{\text{init}} = 1$ and $\gamma \geq L/(2\underline{\lambda})$. A similar convergence rate result was shown by

Name	c	CGD-GS-r-acc	CGD-GS-q-acc
		♯nz/obj/cpu(CGD/L-BFGS/R1)	♯nz/obj/cpu(CGD/L-BFGS/R1)
EG2	.1	1/-998.890/.02(2/0/1)	1/-998.890/.02(2/0/1)
	1	1/-998.377/.02(2/0/1)	1/-998.377/.02(2/0/1)
	10	1/-993.290/.02(2/0/1)	1/-993.290/.02(2/0/1)
EXTROSNB	.1	5/.235809/.8(61/42/2)	5/.235809/.8(59/50/2)
	1	3/.873442/.2(14/13/1)	2/.873441/.8(59/48/2)
	10	0/1.00000/.04(4/0/1)	0/1.00000/.5(12/40/1)
INDEF	1	[b]1000/-499.000/1.5(58/41/3)	1000/-499.000/.9(30/40/1)
	10	2/-301.161/.2(10/5/1)	2/-18.4175/.2(11/5/0)
	100	2/-197.836/.2(10/4/1)	3/499.605/.1(5/0/0)
LIARWHD	.1	1000/101.025/.2(10/19/1)	1000/97.5328/.1(10/8/1)
	1	1000/750.203/.3(10/26/1)	1000/750.203/.1(10/5/1)
	10	0/1000.00/.5(11/40/2)	0/1000.00/.04(4/0/1)
NONCVXU2	.1	948/2390.60/7.0(375/440/40)	957/2710.90/12.3(625/690/40)
	1	683/3120.28/13.2(687/712/24)	677/3124.66/8.7(451/452/10)
	10	0/4000.00/1.9(91/90/9)	5/4000.00/1.9(92/90/8)
PENALTY1	.01	[f]1/.0149673/37.7(11/15/1)	[f]1/.0149673/88.8(25/40/2)
	.1	[f]1/.0571739/14.9(10/0/1)	[f]0/.072500/14.9(10/0/0)
	1	[f]0/.072500/12.1(8/0/1)	[f]0/.072500/14.7(10/0/0)
WOODS	1	1000/985.710/2.1(149/160/12)	1000/985.710/2.1(149/157/12)
	10	750/8655.68/.8(59/56/2)	1000/8655.70/.2(11/25/0)
	100	249/10500.0/.5(11/40/1)	750/10500.7/.5(12/40/0)
QUARTC	.1	1000/50028.1/.2(11/18/0)	1000/50028.1/.2(11/25/0)
	1	1000/500028/.1(11/15/0)	1000/500028/.2(11/22/0)
	10	$999/4.99482{\cdot}10^6$/.1(11/13/0)	$1000/4.99482{\cdot}10^6$/.2(11/26/0)
DIXON3DQ	.1	6/.470417/.6(52/40/0)	6/.470417/.6(46/40/0)
	1	2/1.62500/.02(3/0/1)	2/1.62500/.05(7/0/0)
	10	0/2.00000/.01(1/0/0)	0/2.00000/.02(4/0/0)
TRIDIA	.1	8/.185656/.5(51/40/6)	8/.185656/.6(58/48/3)
	1	2/.911765/.5(40/40/3)	2/.911765/.5(43/40/2)
	10	0/1.00000/.3(11/40/2)	0/1.00000/.3(12/40/1)

Table 3. Comparing the CGD method using the Gauss-Southwell rules and acceleration steps on 10 CUTEr test functions, with x^0 as given.
[b]CGD exited due to the Armijo stepsize in an L-BFGS acceleration step reaching 10^{-30}.
[f]CGD is terminated using tolerance 10^{-7}.

Fukushima and Mine for their method [21, Theorem 5.1] under the additional assumption that f is (locally) strongly convex. On the other hand, Theorem 3 does not seem amenable to a similar extension. If the Lipschitz constant L is unknown, we can still ensure that $\alpha^k = 1$ by adaptively scaling H^k when generating d^k, analogous to the Armijo rule along the projection arc for constrained smooth optimization [4, page 236]. In particular, we choose s^k to be the largest element of $\{s\beta^j\}_{j=0,1,\ldots}$ $(s > 0)$ such that

$$d^k = d_{H^k/s^k}(x^k; \mathcal{J}^k)$$

satisfies the relaxed Armijo descent condition (46) with $\alpha^k = 1$. This adaptive scaling strategy is more expensive computationally since d^k needs to be recomputed each time s^k is changed. Still, if P is separable and we choose H^k to be diagonal, then d^k is relatively cheap to recompute.

There are many directions for future research. For example, in our current implementation of the CGD method, we used diagonal H^k. How about block-diagonal H^k? (For efficiency, this may need to be coded in Fortran since Matlab's vector operations might not be usable.) Can other acceleration techniques be developed? How would the CGD method perform on bound-constrained problems? Can the CGD method be extended to handle linear equality constraints (as arises in support vector machine applications) or, more generally, smooth equality constraints? Can the assumption on P in Theorem 1(d) be dropped? Can a linear convergence rate result similar to Theorem 3 be proved when $\{\mathcal{J}^k\}$ is chosen by the Gauss-Southwell-r rule?

Acknowledgments

We thank Michael Saunders for providing us with MINOS Version 5.5.1. We thank two anonymous referees for their detailed comments and suggestions to shorten the paper. This research is supported by the National Science Foundation, Grant No. DMS-0511283 and DMS-0104055.

References

1. Auslender, A.: Minimisation de fonctions localement lipschitziennes: applications à la programmation mi-convexe, mi-différentiable. In Mangasarian, O.L., Meyer, R.R., and Robinson, S.M., eds. Nonlinear Programming, 3, pp. 429–460. Academic Press, New York, 1978
2. Balakrishnan, S.: Private communication, October 2006.
3. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York, 1982
4. Bertsekas, D.P.: Nonlinear Programming, 2nd edition. Athena Scientific, Belmont, 1999.
5. Bradley, P.S., Fayyad, U.M., and Mangasarian, O.L.: Mathematical programming for data mining: formulations and challenges. INFORMS J. Comput. **11**, 217–238 (1999)
6. Burke, J.V.: Descent methods for composite nondifferentiable optimization problems. Math. Program. **33**, 260–279 (1985)
7. Chen, S., Donoho, D., and Saunders, M.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1999)
8. Censor, Y. and Zenios, S.A.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, New York, 1997
9. Conn, A.R., Gould, N.I.M., and Toint, Ph.L.: Trust-Region Methods, SIAM, Philadelphia, 2000
10. Donoho, D.L. and Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994)
11. Donoho, D.L. and Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Amer. Statist. Assoc. **90**, 1200–1224 (1995)
12. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., and Vapnik, V.: Support vector regression machines. In Mozer, M.C., Jordan, M.I., and Petsche, T., eds. Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, 1997
13. Facchinei, F., Fischer, A., and Kanzow, C.: On the accurate identification of active constraints. SIAM J. Optim. **9**, 14–32 (1998)
14. Facchinei, F. and Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I and II. Springer-Verlag, New York, 2003
15. Ferris, M.C. and Mangasarian, O.L.: Parallel variable distribution. SIAM J. Optim. **4**, 815–832 (1994)
16. Fletcher, R.: A model algorithm for composite nondifferentiable optimization problems. Math. Program. Study **17**, 67–76 (1982)

17. Fletcher, R.: Practical Methods of Optimization, 2nd edition. John Wiley & Sons, Chichester, 1987
18. Fletcher, R.: An overview of unconstrained optimization. In Spedicato, E., ed. Algorithms for Continuous Optimization, pp. 109–143. Kluwer Academic, Dordrecht, 1994
19. Fukushima, M.: A successive quadratic programming method for a class of constrained nonsmooth optimization problems. Math. Program. **49**, 231–251 (1990/91)
20. Fukushima, M.: Parallel variable transformation in unconstrained optimization. SIAM J. Optim. **8**, 658–672 (1998)
21. Fukushima, M. and Mine, H.: A generalized proximal point algorithm for certain nonconvex minimization problems. Int. J. Systems Sci. **12**, 989–1000 (1981)
22. Gould, N.I.M., Orban, D., and Toint, Ph.L.: CUTEr, a constrained and unconstrained testing environment, revisited. ACM Trans. Math. Software **29**, 373–394 (2003)
23. Grippo, L. and Sciandrone, M.: On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. Oper. Res. Letters **26**, 127-136 (2000)
24. Kelley, C.T.: Iterative Methods for Optimization. SIAM, Philadelphia, 1999
25. Kiwiel, K.C.: A method for minimizing the sum of a convex function and a continuously differentiable function. J. Optim. Theory Appl. **48**, 437–449 (1986)
26. Luo, Z.-Q. and Tseng, P.: Error bounds and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM J. Optim. **2**, 43–54 (1992)
27. Luo, Z.-Q. and Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**, 408–425 (1992)
28. Luo, Z.-Q. and Tseng, P.: On the convergence rate of dual ascent methods for linearly constrained convex minimization. Math. Oper. Res. **18**, 846–867 (1993)
29. Luo, Z.-Q. and Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
30. Mangasarian, O.L.: Sparsity-preserving SOR algorithms for separable quadratic and linear programming. Comput. Oper. Res. **11**, 105–112 (1984)
31. Mangasarian, O.L.: Parallel gradient distribution in unconstrained optimization. SIAM J. Control Optim. **33**, 1916–1925 (1995)
32. Mangasarian, O.L. and De Leone, R.: Parallel gradient projection successive overrelaxation for symmetric linear complementarity problems and linear programs. Ann. Oper. Res. **14**, 41–59 (1988)
33. Mangasarian, O.L. and Musicant, D.R.: Successive overrelaxation for support vector machines. IEEE Trans. Neural Networks **10**, 1032–1037 (1999)
34. Mangasarian, O.L. and Musicant, D.R.: Large scale kernel regression via linear programming. Machine Learning **46**, 255–269 (2002)
35. Meier, L., van de Geer, S., and Bühlmann, P.: The group Lasso for logistic regression. Report, Seminar für Statistik, ETH Zürich, Zürich, March 2006.
36. Mine, H. and Fukushima, M.: A minimization method for the sum of a convex function and a continuously differentiable function. J. Optim. Theory Appl. **33**, 9–23 (1981)
37. Moré, J.J., Garbow, B.S., and Hillstrom, K.E.: Testing unconstrained optimization software. ACM Trans. Math. Software **7**, 17–41 (1981)
38. Moré, J.J. and Toraldo, G.: On the solution of large quadratic programming problems with bound constraints. SIAM J. Optim. **1**, 93–113 (1991)
39. Murtagh, B.A. and Saunders, M.A.: MINOS 5.5 user's guide. Report SOL 83-20R. Department of Operations Research, Stanford University, Stanford, July 1998
40. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comp. **35**, 773-782 (1980)
41. Nocedal, J. and Wright S.J.: Numerical Optimization. Springer-Verlag, New York, 1999
42. Ortega, J.M. and Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Reprinted by SIAM, Philadelphia, 2000
43. Powell, M.J.D.: On search directions for minimization algorithms. Math. Program. **4**, 193–201 (1973)
44. Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Math. Program. Study **14**, 206–214 (1981)
45. Robinson, S.M.: Linear convergence of ϵ-subgradient descent methods for a class of convex functions. Math. Program. **86**, 41–50 (1999)
46. Robinson, S.M.: Calmness and Lipschitz continuity for multifunctions. Report, Department of Industrial Engineering, University of Wisconsin, Madison, 2006.
47. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, 1970
48. Rockafellar, R.T. and Wets R.J.-B.: Variational Analysis. Springer-Verlag, New York, 1998

49. Sardy, S., Bruce, A., and Tseng, P.: Block coordinate relaxation methods for nonparametric wavelet denoising. J. Comput. Graph. Stat. **9**, 361–379 (2000)
50. Sardy, S., Bruce, A., and Tseng, P.: Robust wavelet denoising. IEEE Trans. Signal Proc. **49**, 1146–1152 (2001)
51. Sardy, S. and Tseng, P.: AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. J. Comput. Graph. Statist. **13**, 283–309 (2004)
52. Sardy, S. and Tseng, P.: On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions. J. Amer. Statist. Assoc. **99**, 191–204 (2004)
53. Tseng, P.: On the rate of convergence of a partially asynchronous gradient projection algorithm. SIAM J. Optim. **1**, 603–619 (1991)
54. Tseng, P.: Dual coordinate ascent methods for non-strictly convex minimization. Math. Program. **59**, 231–247 (1993)
55. Tseng, P.: Convergence of block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 473–492 (2001)
56. Tseng, P. and Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Report, Department of Mathematics, University of Washington, Seattle; June 2006; revised February 2007. `http://www.math.washington.edu/~tseng/papers.html`
57. Vapnik, V., Golowich, S.E., and Smola, A.: Support vector method for function approximation, regression estimation, and signal processing. in Mozer, M.C., Jordan, M.I., and Petsche, T., eds. Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, 1997
58. Yuan, L. and Lin, Y.: Model selection and estimation in regression with grouped variables. J. Royal Stat. Soc. **68**, 49–67 (2006)
59. Zhu, C., Byrd, R.H., and Nocedal, J.: L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. ACM Trans. Math. Software **23**, 550–560 (1997)