J Optim Theory Appl (2009) 140: 513–535
DOI 10.1007/s10957-008-9458-3

Block-Coordinate Gradient Descent Method for Linearly Constrained Nonsmooth Separable Optimization

P. Tseng · S. Yun

Published online: 30 September 2008

Abstract We consider the problem of minimizing the weighted sum of a smooth function f and a convex function P of n real variables subject to m linear equality constraints. We propose a block-coordinate gradient descent method for solving this problem, with the coordinate block chosen by a Gauss-Southwell-q rule based on sufficient predicted descent. We establish global convergence to first-order stationarity for this method and, under a local error bound assumption, linear rate of convergence. If f is convex with Lipschitz continuous gradient, then the method terminates in $O(n^2/\epsilon)$ iterations with an ϵ-optimal solution. If P is separable, then the Gauss-Southwell-q rule is implementable in $O(n)$ operations when $m = 1$ and in $O(n^2)$ operations when $m > 1$. In the special case of support vector machines training, for which f is convex quadratic, P is separable, and $m = 1$, this complexity bound is comparable to the best known bound for decomposition methods. If f is convex, then, by gradually reducing the weight on P to zero, the method can be adapted to solve the bilevel problem of minimizing P over the set of minima of $f + \delta_X$, where X denotes the closure of the feasible set. This has application in the least 1-norm solution of maximum-likelihood estimation.

Keywords Nonsmooth optimization · Linear constraints · Support vector machines · Bilevel optimization · ℓ_1-regularization · Coordinate gradient descent · Global convergence · Linear convergence rate · Complexity bound

Communicated by A. Miele.

This research was supported by the National Science Foundation, Grant No. DMS-0511283.

P. Tseng (✉)
Department of Mathematics, University of Washington, Seattle, WA, USA
e-mail: tseng@math.washington.edu

S. Yun
Department of Mathematics, National University of Singapore, Singapore, Republic of Singapore
e-mail: matys@nus.edu.sg

1 Introduction

We consider a class of constrained nonsmooth optimization problems of the form

$$\min_{x\in\Re^n} F_c(x) \stackrel{\text{def}}{=} f(x) + cQ(x), \tag{1}$$

where $c > 0$,

$$Q(x) \stackrel{\text{def}}{=} \begin{cases} P(x), & \text{if } Ax = b, \\ \infty, & \text{else,} \end{cases}$$

$P : \Re^n \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous (lsc) function [1], $A \in \Re^{m\times n}$, $b \in \Re^m$, and f is real-valued and smooth (i.e., continuously differentiable) on an open subset of $\Re^n$ containing $\text{dom}\, Q = \{x \mid Q(x) < \infty\}$. We assume $\text{dom}\, Q \neq \emptyset$. Then Q is proper, convex, lsc, and Q is polyhedral whenever P is polyhedral. The objective function F_c is in general nonsmooth and nonconvex. Of particular interest is the case when m is small, n is large, and P is separable, i.e.,

$$P(x) = \sum_{j=1}^{n} P_j(x_j), \tag{2}$$

for some proper, convex, lsc functions $P_j : \Re \to (-\infty, \infty]$. However, Q is not separable due to the constraints $Ax = b$ (unless $m = 0$).

The problem (1) with P separable is quite general and includes as special cases problems of box-constrained smooth optimization and, more generally, nonsmooth separable optimization ($m = 0$) [2–6], as well as monotropic optimization ($f \equiv 0$) [7], and linearly constrained smooth optimization (P is the indicator function for a box) [8–11]. In applications arising in signal denoising, image processing, and data classification, the problem is often large scale ($n \geq 10000$) and P may be nonsmooth to induce solution sparsity; see [12–19] and references therein. Such applications include Basis Pursuit/Lasso (f is convex quadratic, P is the 1-norm, $m = 0$) [13, 14, 16] and support vector machine (SVM) training (f is quadratic, P is the indicator function for a box, $m = 1$) [20, 21]. Methods that update x one coordinate block at a time are well suited to solve these problems, due to their low computational cost per iteration and ease of implementation and parallelization. Such methods include (block) SOR methods for finding sparse representation of signals and decomposition methods for SVM training; see [8, 16, 17, 19–28] and references therein. Recently, block-coordinate gradient descent (CGD) methods were proposed in [6] for solving the case of $m = 0$ and then extended in [29] for linearly constrained smooth optimization. These methods approximate f by a quadratic at the current iterate x, apply block-coordinate descent to generate a feasible descent direction d, and then update x by performing an inexact line search along d. Numerical experiences in [6, 29, 30] suggest that the CGD methods can be effective in practice.

In this paper, we extend the CGD methods in [6, 29] to solve the general problem (1). As in [6, 29], we choose the coordinate block according to a Gauss-Southwell-q rule and choose the stepsize according to an Armijo-like rule; see (7) and (10). (In [6], a Gauss-Seidel rule and a Gauss-Southwell-r rule for choosing the coordinate block are also considered. We do not consider them here for reasons to be explained in Sect. 8.) Our main contributions are three-fold. First, we show that, in the case where P is separable and piecewise-linear/quadratic with $O(1)$ pieces, the Gauss-Southwell-q rule is implementable in $O(n)$ operations when $m=1$ and in $O(n^2)$ operations when $m>1$; see Sect. 6. This is based on conformal realization [7, Sect. 10B], [31] of a diagonally scaled gradient "projection" direction, and extends the procedure in [29, Sect. 6] for linearly constrained smooth optimization. The resulting method uses only $O(n)$ operations per iteration when $m=1$, P is separable and piecewise-linear/quadratic with $O(1)$ pieces (e.g., 1-norm), and f is quadratic or has a partially separable structure; see the end of Sect. 6. Second, we show that, for any $\epsilon>0$, the CGD method terminates in $O(n^2/\epsilon)$ iterations with an ϵ-optimal solution, assuming f is convex with Lipschitz continuous gradient; see Theorem 5.1. This is the first complexity bound for a CGD method. When specialized to the training of SVM ($m=1$, P is the indicator function for a box $\prod_{j=1}^n [l_j,u_j]$, and f is quadratic), the resulting complexity bound of

$$O\left(\frac{n^3\Lambda b_{\max}^2}{\epsilon}+n^2\Lambda\max\left\{0,\log\left(\frac{(F_c(x^{\mathrm{init}})-\min_x F_c(x))}{nb_{\max}}\right)\right\}\right)$$

operations for achieving ϵ-optimality, where $b_{\max}=\max_{1\le j\le n}(u_j-l_j)$ and Λ is the maximum norm of the 2×2 principal submatrices of $\nabla^2 f(x)$, is comparable to the currently best bounds for decomposition methods [26, 32, 33]; see Sect. 6 for details. In addition, the method achieves global convergence to first-order stationarity and, under a local error bound assumption, linear rate of convergence; see Theorems 4.1 and 4.2. This generalizes [6, Theorem 3] and [29, Theorem 5.1] for the cases of $m=0$ and linearly constrained smooth optimization. Third, when f is convex and under a mild assumption on Q, we show that, by gradually decreasing c towards zero at a suitable rate during the execution of the CGD method, we can solve the following bilevel problem:

$$\min_{x\in S_f} Q(x), \tag{3}$$

where S_f denotes the set of minima of f over X, where X denotes the closure of $\operatorname{dom} Q$. This problem arises, for example, in the least 1-norm solution of a least square problem or a maximum likelihood estimation problem [16, 17].

In our notation, $\Re^n$ denotes the space of n-dimensional real column vectors, ^{T} denotes transpose. For any $x\in\Re^n$, x_j denotes the jth component of x, $x_{\mathcal{J}}$ denotes the subvector of x comprising x_j, $j\in\mathcal{J}$, and $\|x\|_p=(\sum_{j=1}^n |x_j|^p)^{1/p}$ for $1\le p<\infty$ and $\|x\|_\infty=\max_j |x_j|$. For simplicity, we write $\|x\|=\|x\|_2$. For any nonempty $\mathcal{J}\subseteq\mathcal{N}=\{1,\dots,n\}$, $|\mathcal{J}|$ denotes the cardinality of $\mathcal{J}$. For any symmetric matrices $H,D\in\Re^{n\times n}$, we write $H\succeq D$ (respectively, $H\succ D$) to mean that $H-D$ is positive semidefinite (respectively, positive definite). $H_{\mathcal{J}\mathcal{J}}=[H_{ij}]_{i,j\in\mathcal{J}}$ denotes the principal submatrix of H indexed by $\mathcal{J}$. $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ denote the minimum and maximum eigenvalues of H. We denote by I the identity matrix and by 0 the matrix of zero entries. Unless otherwise specified, $\{x^k\}$ denotes the sequence $x^0,x^1,\dots$.

2 (Block) Coordinate Gradient Descent Method

In this section, we describe our method for solving (1). As in [6, 29], we use $\nabla f(x)$ to build a quadratic approximation of f at x and apply coordinate descent to generate a feasible descent direction d at x. More precisely, we choose a nonempty subset $\mathcal{J} \subseteq \mathcal{N}$, a symmetric matrix $H \in \Re^{n\times n}$, and move x along the direction

$$d_H(x;\mathcal{J}) \overset{\text{def}}{=} \underset{d\in\Re^n}{\arg\min}\left\{\nabla f(x)^T d + \frac{1}{2}d^T Hd + cQ(x+d) \mid d_j = 0\ \forall j \notin \mathcal{J}\right\}. \quad (4)$$

Here $d_H(x;\mathcal{J})$ depends on H through $H_{\mathcal{J}\mathcal{J}}$ only. To ensure that $d_H(x;\mathcal{J})$ is well defined, we assume that $H_{\mathcal{J}\mathcal{J}}$ is positive definite on $\text{Null}(A_{\mathcal{J}})$ (the null space of $A_{\mathcal{J}}$) or, equivalently, $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0$, where $A_{\mathcal{J}}$ denotes the submatrix of A comprising columns indexed by $\mathcal{J}$ and $B_{\mathcal{J}}$ is a matrix whose columns form an orthonormal basis for $\text{Null}(A_{\mathcal{J}})$. The direction (4) reduces to those used in [6, 29] when $m = 0$ or P is the indicator function for a box.

First, we have the following generalization of [6, Lemma 1] and [29, Lemma 2.1], showing that a nonzero $d_H(x;\mathcal{J})$ is a descent direction of F_c at x. We include its proof for completeness.

Lemma 2.1 *For any $x \in \text{dom}\, Q$, nonempty $\mathcal{J} \subseteq \mathcal{N}$ and symmetric $H \in \Re^{n\times n}$ with $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0$, let $d = d_H(x;\mathcal{J})$ and $g = \nabla f(x)$. Then,*

$$F_c(x+\alpha d) \le F_c(x) + \alpha(g^T d + cQ(x+d) - cQ(x)) + o(\alpha), \quad \forall \alpha \in (0,1], \quad (5)$$

$$g^T d + cQ(x+d) - cQ(x) \le -d^T Hd \le -\lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}})\|d\|^2. \quad (6)$$

Proof Inequality (5) and the first inequality in (6) follow from [6, Lemma 1]. Since $d_{\mathcal{J}} \in \text{Null}(A_{\mathcal{J}})$, so that $d_{\mathcal{J}} = B_{\mathcal{J}} y$ for some vector y, we have

$$d^T Hd = y^T B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} y \ge \|y\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}) = \|d\|^2 \lambda_{\min}(B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}),$$

where the second equality uses $B_{\mathcal{J}}^T B_{\mathcal{J}} = I$. This proves the second inequality in (6). □

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

CGD Method

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

Step 1. Choose a nonempty $\mathcal{J}^k \subseteq \mathcal{N}$ and a symmetric $H^k \in \Re^{n\times n}$ with $B_{\mathcal{J}^k}^T H_{\mathcal{J}^k\mathcal{J}^k}^k B_{\mathcal{J}^k} \succ 0$.
Step 2. Solve (4) with $x = x^k$, $\mathcal{J} = \mathcal{J}^k$, $H = H^k$ to obtain $d^k = d_{H^k}(x^k;\mathcal{J}^k)$.
Step 3. Choose a stepsize $\alpha^k > 0$ and set $x^{k+1} = x^k + \alpha^k d^k$.

Various stepsize rules for smooth optimization [8, 9, 11] can be adapted to our setting. The following Armijo rule, used in [6, 29], is simple, requires only function evaluations, and seems effective in theory and practice.

Armijo Rule

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

$$F_c(x^k + \alpha^k d^k) \le F_c(x^k) + \alpha^k \sigma \Delta^k, \tag{7}$$

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

$$\Delta^k \stackrel{\text{def}}{=} \nabla f(x^k)^T d^k + \gamma {d^k}^T H^k d^k + cQ(x^k + d^k) - cQ(x^k). \tag{8}$$

Since $B^T_{\mathcal{J}^k} H^k_{\mathcal{J}^k\mathcal{J}^k} B_{\mathcal{J}^k} \succ 0$ and $0 \le \gamma < 1$, we see from Lemma 2.1 that

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

and $\Delta^k \le (\gamma - 1){d^k}^T H^k d^k < 0$, whenever $d^k \ne 0$. Since $0 < \sigma < 1$, this shows that α^k given by the Armijo rule is well defined and positive. 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, so larger stepsizes will be accepted if we choose either σ near 0 or γ near 1.

For convergence, the index subset $\mathcal{J}^k$ must be chosen judiciously. We will choose $\mathcal{J}^k$ according to the *Gauss-Southwell-q* rule, which was introduced in [6] for the case of $m = 0$ and was shown in [6], [29] to be effective in theory and practice. Specifically, let

$$q_H(x; \mathcal{J}) \stackrel{\text{def}}{=} \left\{ \nabla f(x)^T d + \frac{1}{2} d^T H d + cQ(x + d) - cQ(x) \right\}_{d = d_H(x;\mathcal{J})}, \tag{9}$$

which is the predicted descent when x is moved along the direction $d_H(x; \mathcal{J})$. The Gauss-Southwell-q rule chooses the index subset $\mathcal{J}^k$ to achieve sufficient predicted descent, i.e.,

$$q_{D^k}(x^k; \mathcal{J}^k) \le \upsilon q_{D^k}(x^k; \mathcal{N}), \tag{10}$$

where $D^k \succ 0$ (typically diagonal) and $0 < \upsilon \le 1$. In fact, it suffices that $B^T_{\mathcal{N}} D^k B_{\mathcal{N}} \succ 0$ for our analysis. We will discuss in Sect. 6 how to efficiently implement this rule when P is separable and piecewise-linear/quadratic.

3 Properties of Search Direction

In this section we derive various properties of the search direction $d_H(x; \mathcal{J})$ and the corresponding predicted descent $q_H(x; \mathcal{J})$. These properties will be used in later sections to analyze the convergence rate and the complexity of the CGD method.

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

Lemma 3.1 *For any symmetric matrix* $H \in \Re^{n\times n}$ *satisfying* $B_{\mathcal{N}}^T H_{\mathcal{N}\mathcal{N}} B_{\mathcal{N}} \succ 0$, *an* $x \in \mathrm{dom}\, Q$ *is a stationary point of* F_c *if and only if* $d_H(x;\mathcal{N}) = 0$.

Proof Let C be a matrix whose columns form an orthonormal basis for the column span of A^T. Then, $d_H(x;\mathcal{N})$ is unchanged when H is replaced by $H + \theta CC^T$ for any $\theta \in \Re$. Moreover, $H + \theta CC^T \succ 0$ for all θ sufficiently large. Then, we apply Lemma 2 in [6] to (1) to obtain the desired result. □

The following lemma shows that $\|d_H(x;\mathcal{J})\|$ changes not too fast with the quadratic coefficients H. It will be used to prove Theorem 4.2. We give its proof for completeness, which is similar to those of [6, Lemma 3] and [29, Lemma 3.1].

Lemma 3.2 *Fix any* $x \in \mathrm{dom}\, Q$, *nonempty* $\mathcal{J} \subseteq \mathcal{N}$, *and symmetric matrices* $H, \tilde{H} \in \Re^{n\times n}$ *satisfying* $U \succ 0$ *and* $\tilde{U} \succ 0$, *where* $U = B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$ *and* $\tilde{U} = B_{\mathcal{J}}^T \tilde{H}_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$. *Let* $d = d_H(x;\mathcal{J})$ *and* $\tilde{d} = d_{\tilde{H}}(x;\mathcal{J})$. *Then,*

$$\|\tilde{d}\| \le \frac{1 + \lambda_{\max}(S) + \sqrt{1 - 2\lambda_{\min}(S) + \lambda_{\max}(S)^2}}{2} \frac{\lambda_{\max}(U)}{\lambda_{\min}(\tilde{U})} \|d\|, \tag{11}$$

where $S = U^{-1/2}\tilde{U}U^{-1/2}$.

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 applying [34, Theorem 10.1] to (1), we have

$$d \in \arg\min_u (g + Hd)^T u + cQ(x+u) - cQ(x),$$

$$\tilde{d} \in \arg\min_u (g + \tilde{H}\tilde{d})^T u + cQ(x+u) - cQ(x).$$

Thus,

$$(g + Hd)^T d + cQ(x+d) - cQ(x) \le (g + Hd)^T \tilde{d} + cQ(x+\tilde{d}) - cQ(x),$$

$$(g + \tilde{H}\tilde{d})^T \tilde{d} + cQ(x+\tilde{d}) - cQ(x) \le (g + \tilde{H}\tilde{d})^T d + cQ(x+d) - cQ(x).$$

Adding the above two inequalities and rearranging terms yield

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

Since $d, \tilde{d} \in \mathrm{Null}(A)$, we have $d = B_{\mathcal{N}} y$ and $\tilde{d} = B_{\mathcal{N}} \tilde{y}$ for some vectors $y, \tilde{y}$. Substituting these into the above inequality and using the definitions of $U, \tilde{U}$ yield

$$y^T U y - y^T (U + \tilde{U})\tilde{y} + \tilde{y}^T \tilde{U}\tilde{y} \le 0.$$

Then proceeding as in the proof of [6, Lemma 3] and using $\|d\| = \|y\|$, $\|\tilde{d}\| = \|\tilde{y}\|$ (since $B_{\mathcal{N}}^T B_{\mathcal{N}} = I$), we obtain (11). □

The next lemma bounds $\nabla f(x)^T(x' - \bar{x}) + cQ(x') - cQ(\bar{x})$ from above by a weighted sum of $\|x - \bar{x}\|^2$ and $-q_D(x; \mathcal{J})$, where $x' = x + \alpha d$, $d = d_H(x; \mathcal{J})$, and $\mathcal{J}$ satisfies a condition analogous to (10). This lemma, which extends [29, Lemma 3.3] for the case of linearly constrained smooth optimization, will be used to prove Theorem 4.2.

Lemma 3.3 *Fix any* $x \in \operatorname{dom} Q$, *nonempty* $\mathcal{J} \subseteq \mathcal{N}$, *symmetric matrices* $H, D \in \Re^{n \times n}$ *satisfying* $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0$, $\bar{\delta} I \succeq D \succ 0$, *and*

$$q_D(x; \mathcal{J}) \leq \upsilon\, q_D(x; \mathcal{N}), \tag{12}$$

with $\bar{\delta} > 0$, $0 < \upsilon \leq 1$. *Then, for any* $\bar{x} \in \operatorname{dom} Q$, $0 \leq \alpha \leq 1$, *we have*

$$g^T(x' - \bar{x}) + cQ(x') - cQ(\bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 - \frac{1}{\upsilon} q_D(x; \mathcal{J}), \tag{13}$$

where $g = \nabla f(x)$, $x' = x + \alpha d$, $d = d_H(x; \mathcal{J})$.

Proof Since $\bar{x} - x$ is a feasible solution of the minimization subproblem (4) corresponding to $\mathcal{N}$ and D, we have

$$q_D(x; \mathcal{N}) \leq g^T(\bar{x} - x) + \frac{1}{2}(\bar{x} - x)^T D(\bar{x} - x) + cQ(\bar{x}) - cQ(x).$$

Since $\bar{\delta} I \succeq D$, we have $(\bar{x} - x)^T D(\bar{x} - x) \leq \bar{\delta}\|\bar{x} - x\|^2$. This together with (12) yields

$$\frac{1}{\upsilon} q_D(x; \mathcal{J}) \leq g^T(\bar{x} - x) + \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 + cQ(\bar{x}) - cQ(x).$$

Rearranging terms, we have

$$g^T(x - \bar{x}) + cQ(x) - cQ(\bar{x}) \leq \frac{\bar{\delta}}{2}\|\bar{x} - x\|^2 - \frac{1}{\upsilon} q_D(x; \mathcal{J}). \tag{14}$$

Also, by the definition of d and (6) in Lemma 2.1, for any $\alpha \geq 0$ we have

$$\alpha(g^T d + cQ(x + d) - cQ(x)) \leq 0.$$

Since Q is convex so that $cQ(x + \alpha d) - cQ(x) \leq \alpha(cQ(x + d) - cQ(x))$, this implies

$$\alpha g^T d + cQ(x + \alpha d) - cQ(x) \leq 0.$$

Adding this to (14) yields (13). □

The next lemma shows that Δ is bounded above by a constant multiple of $q_H(x; \mathcal{J})$. It also bounds $q_H(x; \mathcal{J})$ from above by a constant multiple of $q_D(x; \mathcal{J})$. This lemma is new and will be used to analyze the complexity of the CGD method when f is convex; see Theorem 5.1.

Lemma 3.4 *For any* $x \in \operatorname{dom} Q$, *nonempty* $\mathcal{J} \subseteq \mathcal{N}$, *and symmetric matrix* $H \in \Re^{n\times n}$ *satisfying* $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succ 0$, *the following results hold with* $d = d_H(x;\mathcal{J})$ *and* $g = \nabla f(x)$.

(a) *For any* $0 \le \gamma < 1$,

$$\Delta \le \min\{1, 2-2\gamma\} q_H(x;\mathcal{J}),$$

where $\Delta = g^T d + \gamma d^T H d + cQ(x+d) - cQ(x)$.

(b) *For any symmetric matrix* $D \in \Re^{n\times n}$ *satisfying* $B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succeq B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$ *and any* $0 < \omega \le 1$,

$$q_H(x;\mathcal{J}) \le q_D(x;\mathcal{J}) \le \omega q_{\omega D}(x;\mathcal{J}).$$

Proof (a) If $\gamma \le 1/2$, then $d^T H d \ge 0$ by (6) in Lemma 2.1, so that

$$\Delta = q_H(x;\mathcal{J}) + \left(\gamma - \frac{1}{2}\right) d^T H d \le q_H(x;\mathcal{J}).$$

Otherwise, $1/2 < \gamma < 1$ and we have from (6) in Lemma 2.1 that

$$\begin{aligned}\Delta &= g^T d + cQ(x+d) - cQ(x) + (2\gamma - 1) d^T H d + (1-\gamma) d^T H d \\ &\le g^T d + cQ(x+d) - cQ(x) + (2\gamma-1)(-g^T d - cQ(x+d) + cQ(x)) \\ &\quad + (1-\gamma) d^T H d \\ &= (2-2\gamma) q_H(x;\mathcal{J}).\end{aligned}$$

Thus $\Delta \le \min\{1, 2-2\gamma\} q_H(x;\mathcal{J})$.

(b) Let $\bar{d} = d_D(x;\mathcal{J})$. Then

$$\begin{aligned}q_H(x;\mathcal{J}) &= g^T d + \frac{1}{2} d^T H d + cQ(x+d) - cQ(x) \\ &\le g^T \bar{d} + \frac{1}{2}\bar{d}^T H \bar{d} + cQ(x+\bar{d}) - cQ(x) \\ &\le g^T \bar{d} + \frac{1}{2}\bar{d}^T D \bar{d} + cQ(x+\bar{d}) - cQ(x) \\ &= q_D(x;\mathcal{J}),\end{aligned}$$

where the third step uses $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$ and $A_{\mathcal{J}} \bar{d}_{\mathcal{J}} = 0$. This proves the first inequality. To prove the second inequality, we note that

$$\begin{aligned}q_{\omega D}(x;\mathcal{J}) &= \min_{u_j = 0\ \forall j \notin \mathcal{J}} \left\{ g^T u + \frac{\omega}{2} u^T D u + cQ(x+u) - cQ(x) \right\} \\ &= \frac{1}{\omega} \min_{u_j = 0\ \forall j \notin \mathcal{J}} \left\{ g^T(\omega u) + \frac{1}{2}(\omega u)^T D(\omega u) + \omega(cQ(x+u) - cQ(x)) \right\} \\ &\ge \frac{1}{\omega} \min_{u_j = 0\ \forall j \notin \mathcal{J}} \left\{ g^T(\omega u) + \frac{1}{2}(\omega u)^T D(\omega u) + cQ(x+\omega u) - cQ(x) \right\}\end{aligned}$$

$$= \frac{1}{\omega} q_D(x; \mathcal{J}),$$

where the inequality uses the convexity of Q. □

Corollary 3.1 *For any $x \in \operatorname{dom} Q$, nonempty $\mathcal{J} \subseteq \mathcal{N}$, and symmetric matrices $H, D \in \Re^{n\times n}$ satisfying $0 \prec B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \bar{\lambda} I$ and $B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \succeq \underline{\delta} I$, we have*

$$q_H(x; \mathcal{J}) \le \min\left\{1, \frac{\underline{\delta}}{\bar{\lambda}}\right\} q_D(x; \mathcal{J}).$$

Proof We have $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \frac{\bar{\lambda}}{\underline{\delta}} B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$. If $\frac{\bar{\lambda}}{\underline{\delta}} \le 1$, then $B_{\mathcal{J}}^T H_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq \frac{\bar{\lambda}}{\underline{\delta}} B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}} \preceq B_{\mathcal{J}}^T D_{\mathcal{J}\mathcal{J}} B_{\mathcal{J}}$, so Lemma 3.4(b) yields $q_H(x; \mathcal{J}) \le q_D(x; \mathcal{J})$. If $\frac{\bar{\lambda}}{\underline{\delta}} > 1$, then Lemma 3.4(b) again yields

$$q_H(x; \mathcal{J}) \le q_{\frac{\bar{\lambda}}{\underline{\delta}} D}(x; \mathcal{J}) \le \frac{\underline{\delta}}{\bar{\lambda}} q_D(x; \mathcal{J}).$$

This proves the desired result. □

4 Global Convergence and Convergence Rate Analysis

In this section, we analyze the global convergence and asymptotic convergence rate of the CGD method using the Gauss-Southwell-q rule, analogous to those obtained for the cases of $m = 0$ [6, Theorems 1 and 3] and linearly constrained smooth optimization [29, Theorems 4.1 and 5.1]. Analogous to [29], we make the following assumption on $\{H^k\}$ in the CGD method.

Assumption 4.1 $\bar{\lambda} I \succeq B_{\mathcal{J}^k}^T H_{\mathcal{J}^k \mathcal{J}^k}^k B_{\mathcal{J}^k} \succeq \underline{\lambda} I$ for all k, where $0 < \underline{\lambda} \le \bar{\lambda}$.

Assumption 4.1 allows H^k to closely approximate $\nabla^2 f(x^k)$ provided $\nabla^2 f(x^k)_{\mathcal{J}^k \mathcal{J}^k}$ is positive definite over $\mathrm{Null}(A_{\mathcal{J}^k})$. The following theorem states the global convergence properties of the CGD method. Its proof is omitted since it is nearly identical to that of [6, Theorem 1(a), (b), (d), (f)] for the case of $m = 0$, with minor modification to account for [6, Assumption 1] being relaxed to Assumption 4.1.

Theorem 4.1 *Let $\{x^k\}$, $\{\mathcal{J}^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method, where $\{H^k\}$ satisfies Assumption 4.1 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha_{\text{init}}^k > 0$. Then, the following results hold.*

(a) *$\{F_c(x^k)\}$ is nonincreasing and Δ^k given by (8) satisfies*

$$-\Delta^k \ge (1-\gamma) {d^k}^T H^k d^k \ge (1-\gamma)\underline{\lambda} \|d^k\|^2, \quad \forall k, \tag{15}$$

$$F_c(x^{k+1}) - F_c(x^k) \le \sigma \alpha^k \Delta^k \le 0, \quad \forall k. \tag{16}$$

(b) *If* $\{\mathcal{J}^k\}$ *satisfies* (10), $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ *for all* k, *where* $0 < \underline{\delta} \leq \bar{\delta}$, *and either* (1) Q *is continuous on* $\operatorname{dom} Q$ *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.

(c) *If, for any* $\ell \in \{1, \ldots, n\}$, *there exists* $L_\ell \geq 0$ *such that*

$$\|\nabla f(y) - \nabla f(z)\| \leq L_\ell \|y - z\|,$$
$$\forall y, z \in \operatorname{dom} Q \text{ with } y_j = z_j \ \forall j \notin \mathcal{J}, \ \forall \mathcal{J} \subseteq \mathcal{N} \text{ with } |\mathcal{J}| \leq \ell, \qquad (17)$$

then $\alpha^k \geq \min\{\alpha^k_{\text{init}}, \beta \min\{1, 2\underline{\lambda}(1-\sigma+\sigma\gamma)/L_\ell\}$ *for all* k. *If* $\lim_{k\to\infty} F_c(x^k) > -\infty$ *also, then* $\{\Delta^k\} \to 0$ *and* $\{d^k\} \to 0$.

If P is separable, then Q is automatically continuous on $\operatorname{dom} Q$ [34, Corollary 2.37]. The next theorem establishes the convergence rate of the CGD method under Assumption 4.1 and the following assumption that is analogous to [6, Assumption 2]. In what follows, $\bar{X}$ denotes the set of stationary points of F_c and

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

Assumption 4.2

(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

$$\operatorname{dist}(x, \bar{X}) \leq \tau \|d_I(x; \mathcal{N})\| \quad \text{whenever } F_c(x) \leq \zeta, \ \|d_I(x; \mathcal{N})\| \leq \epsilon.$$

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

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

Assumption 4.2(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; see [27, 35, 36] and references therein. Assumption 4.2(b) says that the isocost surfaces of F_c restricted to the solution set $\bar{X}$ are "properly separated." Assumption 4.2(b) holds automatically if f is convex or f is quadratic and P is polyhedral; see [6, 36] for further discussions. Upon applying [6, Theorem 4] to the problem (1), we obtain the following sufficient conditions for Assumption 4.2(a) to hold.

Proposition 4.1 *Suppose that* $\bar{X} \neq \emptyset$ *and any of the following conditions hold.*

(C1) f *is strongly convex and satisfies* (17) *with* $\ell = n$ *for some* $L_n \geq 0$.
(C2) f *is quadratic*. P *is polyhedral*.
(C3) $f(x) = g(Ex) + q^T x$ *for all* $x \in \Re^n$, *where* $E \in \Re^{p\times n}$, $q \in \Re^n$, *and* g *is a strongly convex differentiable function on* $\Re^p$ *with* ∇g *Lipschitz continuous on* $\Re^p$. P *is polyhedral*.
(C4) $f(x) = \max_{y\in Y}\{(Ex)^T y - g(y)\} + q^T x$ *for all* $x \in \Re^n$, *where* Y *is a polyhedral set in* $\Re^p$, $E \in \Re^{p\times n}$, $q \in \Re^n$, *and* g *is a strongly convex differentiable function on* $\Re^p$ *with* ∇g *Lipschitz continuous on* $\Re^p$. P *is polyhedral*.

Then Assumption 4.2(*a*) *holds*.

The next theorem establishes, under Assumptions 4.1 and 4.2, the linear rate of convergence of the CGD method using (10) to choose $\{\mathcal{J}^k\}$. Its proof, which uses Theorem 4.1 and Lemmas 2.1, 3.2, 3.3, is similar to the proof of [29, Theorem 5.1], except that "f" is replaced by "F_c" and "cQ" is added in some places. The full proof can be found in the Appendix of the original report: http://www.math.washington.edu/~tseng/papers/cgd_cnobi.pdf. In what follows, by Q-linear and R-linear convergence, we mean linear convergence in the quotient and the root sense, respectively [37, Chapt. 9].

Theorem 4.2 *Assume that f satisfies (17) with $\ell = n$ for some $L_n \geq 0$. Let $\{x^k\}$, $\{H^k\}$, $\{d^k\}$ be sequences generated by the CGD method, where $\{H^k\}$ satisfies Assumption 4.1, $\{\mathcal{J}^k\}$ satisfies (10) with $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ for all k $(0 < \underline{\delta} \leq \bar{\delta})$. If F_c satisfies Assumption 4.2 and $\{\alpha^k\}$ is chosen by the Armijo rule with $\sup_k \alpha^k_{\text{init}} \leq 1$ and $\inf_k \alpha^k_{\text{init}} > 0$, 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 to a point in $\bar{X}$.*

Theorem 4.2 generalizes [6, Theorem 3] by relaxing [6, Assumption 1] to Assumption 4.1 and, more significantly, not assuming Q is block-separable. The assumption (17) with $\ell = n$ in Theorem 4.2 can be relaxed to ∇f being Lipschitz continuous on $\operatorname{dom} Q \cap (X^0 + \varrho B)$ 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 $\{x \mid F_c(x) \leq F_c(x^0)\}$. For simplicity, we do not consider this more relaxed assumption here.

5 Complexity Analysis for f Convex

The following theorem is the main result of this section, giving an upper bound on the number of iterations for the CGD method to achieve ϵ-optimality when f is convex with Lipschitz continuous gradient. Its proof uses Lemmas 2.1, 3.4(a), Corollary 3.1, and Theorem 4.1(c). In what follows, $\lceil \cdot \rceil$ denotes the ceiling function.

Theorem 5.1 *Suppose that f is convex and satisfies (17) for some $L_\ell \geq 0$ $(\ell \geq 1)$. Suppose that $\inf_x F_c(x) > -\infty$. Let $\{x^k\}$, $\{\mathcal{J}^k\}$, $\{H^k\}$ be sequences generated by the CGD method, where $\{H^k\}$ satisfies Assumption 4.1, $\{\mathcal{J}^k\}$ satisfies (10) with $\bar{\delta} I \succeq D^k \succeq \underline{\delta} I$ and $|\mathcal{J}^k| \leq \ell$ for all k $(0 < \underline{\delta} \leq \bar{\delta}$, $\ell \geq 1)$, and $\{\alpha^k\}$ is chosen by the Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$. Let $e^k = F_c(x^k) - \inf_x F_c(x)$ for all k. Then, $e^k \leq \epsilon$ whenever*

$$k \geq \begin{cases} \max\{0, \lceil \frac{2}{C\sigma\underline{\alpha}} \log(\frac{e^0}{\epsilon}) \rceil\}, & \text{if } \epsilon > \bar{\delta} r^0; \\ \max\{0, \lceil \frac{2}{C\sigma\underline{\alpha}} \log(\frac{e^0}{\bar{\delta} r^0}) \rceil\} + \lceil \frac{\bar{\delta} r^0}{C\sigma\underline{\alpha}\epsilon} \rceil, & \text{else}, \end{cases}$$

where $r^0 = \max_x\{\operatorname{dist}(x, \bar{X})^2 \mid F_c(x) \leq F_c(x^0)\}$, $\bar{X} = \arg\min_x F_c(x)$, $C = \min\{1, 2 - 2\gamma\} \min\{1, \underline{\delta}/\bar{\lambda}\}\upsilon$, $\underline{\alpha} = \min\{\inf_k \alpha^k_{\text{init}}, \beta \min\{1, 2\underline{\lambda}(1 - \sigma + \sigma\gamma)/L_\ell\}$.

Proof For each $k = 0, 1, \ldots$, by (7), we have

$$\begin{aligned} e^{k+1} - e^k &= F_c(x^{k+1}) - F_c(x^k) \\ &\le \sigma\alpha^k \Delta^k \\ &\le \sigma\alpha^k \min\{1, 2-2\gamma\} q_{H^k}(x^k; \mathcal{J}^k) \\ &\le C\sigma\alpha^k q_{D^k}(x^k; \mathcal{N}) \\ &\le C\sigma\underline{\alpha} q_{D^k}(x^k; \mathcal{N}), \end{aligned} \tag{18}$$

where the second inequality uses Assumption 4.1 and Lemma 3.4(a), the third inequality uses Corollary 3.1 and (10), and the last inequality uses Theorem 4.1(c), implying that $\alpha^k \ge \underline{\alpha}$.

For each $k = 0, 1, \ldots$, and $t \in [0, 1]$, let $g^k = \nabla f(x^k)$ and let $\bar{x}^k \in \bar{X}$ satisfy $\|x^k - \bar{x}^k\| = \mathrm{dist}(x^k, \bar{X})$. Then,

$$\begin{aligned} &q_{D^k}(x^k; \mathcal{N}) \\ &= \min_{d\in\Re^n} {g^k}^T d + \frac{1}{2} d^T D^k d + cQ(x^k + d) - cQ(x^k) \\ &\le {g^k}^T t(\bar{x}^k - x^k) + \frac{t^2}{2}(\bar{x}^k - x^k)^T D^k(\bar{x}^k - x^k) + cQ(x^k + t(\bar{x}^k - x^k)) - cQ(x^k) \\ &\le {g^k}^T t(\bar{x}^k - x^k) + \frac{t^2}{2}(\bar{x}^k - x^k)^T D^k(\bar{x}^k - x^k) + tcQ(\bar{x}^k) - tcQ(x^k) \\ &\le t(f(\bar{x}^k) - f(x^k)) + tcQ(\bar{x}^k) - tcQ(x^k) + \frac{t^2}{2}\bar{\delta}\,\mathrm{dist}(x^k, \bar{X})^2 \\ &= -te^k + \frac{t^2}{2}\bar{\delta}\,\mathrm{dist}(x^k, \bar{X})^2 \\ &\le -te^k + \frac{t^2}{2}\bar{\delta} r^0, \end{aligned}$$

where the second inequality uses the convexity of Q and the third inequality uses the convexity of f. This holds for all $t \in [0, 1]$. Minimizing the right-hand side with respect to t yields

$$q_{D^k}(x^k; \mathcal{N}) \le -\frac{(e^k)^2}{2\bar{\delta} r^0}$$

if $e^k \le \bar{\delta} r^0$; else,

$$q_{D^k}(x^k; \mathcal{N}) \le -e^k + \frac{1}{2}\bar{\delta} r^0 < -\frac{1}{2} e^k.$$

This together with (18) yields

$$e^{k+1} \le e^k - \frac{C\sigma\underline{\alpha}}{\bar{\delta} r^0}(e^k)^2 = e^k\left(1 - \frac{C\sigma\underline{\alpha}}{\bar{\delta} r^0}(e^k)\right) \tag{19}$$

if $e^k \le \bar{\delta} r^0$; else,

$$e^{k+1} \le e^k - \frac{C\sigma\underline{\alpha}}{2} e^k. \tag{20}$$

Case (1): If $\epsilon > \bar{\delta} r^0$, then (20) implies $e^k \le \epsilon$ whenever

$$e^0 \left(1 - \frac{C\sigma\underline{\alpha}}{2}\right)^k < e^0 \exp(-kC\sigma\underline{\alpha}/2) \le \epsilon,$$

or equivalently,

$$k \ge \max\left\{0, \left\lceil \frac{2}{C\sigma\underline{\alpha}} \log\left(\frac{e^0}{\epsilon}\right) \right\rceil\right\}.$$

Case (2): If $\epsilon \le \bar{\delta} r^0$, then (20) implies $e^k \le \bar{\delta} r^0$ whenever

$$e^0 \left(1 - \frac{C\sigma\underline{\alpha}}{2}\right)^k < e^0 \exp(-k_0 C\sigma\underline{\alpha}/2) \le \bar{\delta} r^0,$$

or equivalently,

$$k \ge k_0 \stackrel{\text{def}}{=} \max\left\{0, \left\lceil \frac{2}{C\sigma\underline{\alpha}} \log\left(\frac{e^0}{\bar{\delta} r^0}\right) \right\rceil\right\}.$$

For each $k \ge k_0$, $e^k \le \bar{\delta} r^0$. If $e^k = 0$, then $e^k \le \epsilon$. Otherwise $e^k > 0$. Then, $e^j > 0$ for $j = k_0, k_0 + 1, \ldots, k$ and we consider the reciprocals $\xi_j = 1/e^j$. By (19) and $e^k > 0$, we have $0 \le C_1 e^j < 1$ for $j = k_0, k_0 + 1, \ldots, k - 1$, where $C_1 = C\sigma\underline{\alpha}/(\bar{\delta} r^0)$. Thus, (19) yields

$$\xi_{j+1} - \xi_j \ge \frac{1}{e^j(1 - C_1 e^j)} - \frac{1}{e^j} = \frac{C_1}{1 - C_1 e^j} \ge C_1, \quad j = 0, 1, \ldots, k - 1.$$

Therefore $\xi_k = \xi_{k_0} + \sum_{j=k_0}^{k-1} (\xi_{j+1} - \xi_j) \ge C_1(k - k_0)$ and consequently

$$e^k = \frac{1}{\xi_k} \le \frac{1}{C_1(k - k_0)}.$$

It follows that $e^k \le \epsilon$ whenever

$$k \ge k_0 + \left\lceil \frac{1}{C_1 \epsilon} \right\rceil = \max\left\{0, \left\lceil \frac{2}{C\sigma\underline{\alpha}} \log\left(\frac{e^0}{\bar{\delta} r^0}\right) \right\rceil\right\} + \left\lceil \frac{\bar{\delta} r^0}{C\sigma\underline{\alpha}\epsilon} \right\rceil. \qquad \square$$

If we take $\gamma = 1/2$, $D^k = H^k = I$ and $\alpha^k_{\text{init}} = 1$ for all k, then $\underline{\delta} = \bar{\delta} = \underline{\lambda} = \bar{\lambda} = 1$ and $C = \upsilon$, and the iteration bounds in Theorem 5.1 reduce to

$$\begin{aligned} &O\left(\frac{L_\ell}{\upsilon} \max\left\{0, \log\left(\frac{e^0}{\epsilon}\right)\right\}\right), && \text{if } \epsilon > r^0; \\ &O\left(\frac{L_\ell}{\upsilon} \max\left\{0, \log\left(\frac{e^0}{r^0}\right)\right\} + \frac{L_\ell r^0}{\upsilon \epsilon}\right), && \text{else.} \end{aligned} \tag{21}$$

Notice that $r^0 = 0$ whenever $x^0 \in \bar{X}$. If $\bar{X}$ is bounded, then it can be seen that $r^0 \to 0$ as $\operatorname{dist}(x^0, \bar{X}) \to 0$.

6 Index Subset Selection for P Separable

In this section, we study efficient ways to find an index subset $\mathcal{J}^k$ satisfying (10) for some constant $0 < \upsilon \leq 1$. One obvious choice is $\mathcal{J}^k = \mathcal{N}$, which satisfies (10) with $\upsilon = 1$. However, the corresponding search direction (4) may be expensive to compute and, for SVM applications, the gradient ∇f would be expensive to update. We will extend the procedure developed in [29] for linearly constrained smooth optimization to use a conformal realization of $d_{D^k}(x^k; \mathcal{N})$ [7, Sect. 10B], [31] to find $\mathcal{J}^k$ of small size when P is separable. Our main result is Proposition 6.1, showing the existence of such $\mathcal{J}^k$ by construction.

For any $d \in \Re^n$, the support of d is $\operatorname{supp}(d) \stackrel{\text{def}}{=} \{j \in \mathcal{N} \mid d_j \neq 0\}$. We say $d' \in \Re^n$ is *conformal* to $d \in \Re^n$ if

$$\operatorname{supp}(d') \subseteq \operatorname{supp}(d), \quad d'_j d_j \geq 0 \ \forall j \in \mathcal{N}, \tag{22}$$

i.e., the nonzero components of d' have the same signs as the corresponding components of d. A nonzero $d \in \Re^n$ is an *elementary vector* of $\operatorname{Null}(A)$ if $d \in \operatorname{Null}(A)$ and there is no nonzero $d' \in \operatorname{Null}(A)$ that is conformal to d and $\operatorname{supp}(d') \neq \operatorname{supp}(d)$. Each elementary vector d satisfies $|\operatorname{supp}(d)| \leq \operatorname{rank}(A) + 1$ (since any subset of $\operatorname{rank}(A) + 1$ columns of A are linearly dependent) [7, Exercise 10.6].

First, we derive a lower bound on $P(x + d) - P(x)$, based on a conformal realization of d, for the case when P is separable. This bound will be used to prove Proposition 6.1.

Lemma 6.1 *Suppose P is separable, i.e., has the form (2). For any $x, x + d \in \operatorname{dom} P$, let d be expressed as $d = d^1 + \cdots + d^r$, for some $r \geq 1$ and some nonzero $d^t \in \Re^n$ conformal to d for $t = 1, \ldots, r$. Then*

$$P(x + d) - P(x) \geq \sum_{t=1}^{r} \bigl(P(x + d^t) - P(x)\bigr).$$

Proof Since P is separable, it suffices to prove that, for each $j \in \mathcal{N}$,

$$P_j(x_j + d_j^1 + \cdots + d_j^r) - P_j(x_j) \geq \sum_{t=1}^{r} \bigl(P_j(x_j + d_j^t) - P_j(x_j)\bigr). \tag{23}$$

We prove this by induction on r. This clearly holds for $r = 1$. Suppose (23) holds for $r < s$, where $s \geq 2$. We show below that (23) holds for $r = s$. If $d_j^1 + \cdots + d_j^{s-1} = 0$, then (23) reduces to the case of $r = 1$ and hence holds. If $d_j^s = 0$, then (23) reduces to the case of $r < s$ and hence holds. Thus it remains to consider the case of $d_j^1 + \cdots + d_j^{s-1} \neq 0$ and $d_j^s \neq 0$. Since $d_j^1, d_j^2, \ldots, d_j^s$ are conformal to d_j, either

(i) $d_j^1+\cdots+d_j^{s-1}>0$ and $d_j^s>0$ or (ii) $d_j^1+\cdots+d_j^{s-1}<0$ and $d_j^s<0$. In case (i), we have $x_j+d_j^1+\cdots+d_j^{s-1}<x_j+d_j$ and $x_j+d_j^s<x_j+d_j$, so the convexity of P_j [34, Lemma 2.12] implies

$$\frac{P_j(x_j+d_j^1+\cdots+d_j^{s-1})-P_j(x_j)}{d_j^1+\cdots+d_j^{s-1}}\le\frac{P_j(x_j+d_j)-P_j(x_j)}{d_j},$$
$$\frac{P_j(x_j+d_j^s)-P_j(x_j)}{d_j^s}\le\frac{P_j(x_j+d_j)-P_j(x_j)}{d_j}.$$

Multiplying the above two inequalities by, respectively, $d_j^1+\cdots+d_j^{s-1}>0$ and $d_j^s>0$ and summing, we have

$$\begin{aligned}&P_j(x_j+d_j^1+\cdots+d_j^{s-1})-P_j(x_j)+P_j(x_j+d_j^s)-P_j(x_j)\\&\quad\le P_j(x_j+d_j)-P_j(x_j).\end{aligned}\tag{24}$$

In case (ii), we have $x_j+d_j^1+\cdots+d_j^{s-1}>x_j+d_j$ and $x_j+d_j^s>x_j+d_j$, so the convexity of P_j implies

$$\frac{P_j(x_j+d_j^1+\cdots+d_j^{s-1})-P_j(x_j)}{d_j^1+\cdots+d_j^{s-1}}\ge\frac{P_j(x_j+d_j)-P_j(x_j)}{d_j},$$
$$\frac{P_j(x_j+d_j^s)-P_j(x_j)}{d_j^s}\ge\frac{P_j(x_j+d_j)-P_j(x_j)}{d_j}.$$

Multiplying the above two inequalities by, respectively, $d_j^1+\cdots+d_j^{s-1}<0$ and $d_j^s<0$ and summing, we again obtain (24). Since (23) holds for $r<s$, we also have

$$P_j(x_j+d_j^1+\cdots+d_j^{s-1})-P_j(x_j)\ge\sum_{t=1}^{s-1}\big(P_j(x_j+d_j^t)-P_j(x_j)\big).$$

Combining this with (24) proves that (23) holds for $r=s$. □

Lemma 6.1 is false if we drop the assumption that P is separable. For example, take $P(x)=\|x\|$, $x=0$, and $d=(1,1,-2)^T$. Then, d can be expressed as $d=d^1+d^2=(1,0,-1)^T+(0,1,-1)^T$, but $P(x+d)-P(x)=\sqrt{6}<2\sqrt{2}=\sum_{t=1}^2(P(x+d^t)-P(x))$.

By using Lemma 6.1 and generalizing the proof of [29, Proposition 6.1], we obtain the following main result of this section.

Proposition 6.1 *For any $x\in\operatorname{dom}Q$, $\ell\in\{\operatorname{rank}(A)+1,\dots,n\}$, and diagonal $D\succ0$, if P is separable, then there exists a nonempty $\mathcal{J}\subseteq\mathcal{N}$ satisfying $|\mathcal{J}|\le\ell$ and*

$$q_D(x;\mathcal{J})\le\frac{1}{n-\ell+1}q_D(x;\mathcal{N}).\tag{25}$$

Proof Let $d = d_D(x; \mathcal{N})$. We divide our argument into three cases.

Case (i) $d = 0$: Then $q_D(x; \mathcal{N}) = 0$. Thus, for any nonempty $\mathcal{J} \subseteq \mathcal{N}$ with $|\mathcal{J}| \le \ell$, we have from (9) and Lemma 2.1 with $H = D$ that $q_D(x; \mathcal{J}) \le 0 = q_D(x; \mathcal{N})$, so (25) holds.

Case (ii) $d \ne 0$ and $|\operatorname{supp}(d)| \le \ell$: Then $\mathcal{J} = \operatorname{supp}(d)$ satisfies $q_D(x; \mathcal{J}) = q_D(x; \mathcal{N})$ and hence (25), as well as $|\mathcal{J}| \le \ell$.

Case (iii) $d \ne 0$ and $|\operatorname{supp}(d)| > \ell$: Since $d \in \operatorname{Null}(A)$, it has a conformal realization [7, Sect. 10B], [31], namely,

$$d = v^1 + \cdots + v^s,$$

for some $s \ge 1$ and some nonzero elementary vectors $v^t \in \operatorname{Null}(A)$, $t = 1, \dots, s$, conformal to d. Then, for some $\alpha > 0$, $\operatorname{supp}(d')$ is a proper subset of $\operatorname{supp}(d)$ and $d' \in \operatorname{Null}(A)$, where $d' = d - \alpha v^1$. (Note that αv^1 is an elementary vector of $\operatorname{Null}(A)$, so that $|\operatorname{supp}(\alpha v^1)| \le \operatorname{rank}(A) + 1 \le \ell$.) We repeat the above reduction step with d' in place of d. Since $|\operatorname{supp}(d')| \le |\operatorname{supp}(d)| - 1$, after at most $|\operatorname{supp}(d)| - \ell$ reduction steps, we obtain

$$d = d^1 + \cdots + d^r, \tag{26}$$

for some $r \le |\operatorname{supp}(d)| - \ell + 1$ and some nonzero $d^t \in \operatorname{Null}(A)$ conformal to d with $|\operatorname{supp}(d^t)| \le \ell$, $t = 1, \dots, r$. Since $|\operatorname{supp}(d)| \le n$, we have $r \le n - \ell + 1$.

Since $Ad^t = 0$, this implies $A(x + d^t) = b$, $t = 1, \dots, r$. Also (9) and (26) imply that

$$\begin{aligned}
q_D(x; \mathcal{N}) &= g^T d + \frac{1}{2} d^T D d + cQ(x+d) - cQ(x) \\
&= g^T d + \frac{1}{2} d^T D d + cP(x+d) - cP(x) \\
&= \sum_{t=1}^r g^T d^t + \frac{1}{2} \sum_{s=1}^r \sum_{t=1}^r (d^s)^T D d^t + cP(x+d) - cP(x) \\
&\ge \sum_{t=1}^r g^T d^t + \frac{1}{2} \sum_{t=1}^r (d^t)^T D d^t + cP(x+d) - cP(x) \\
&\ge \sum_{t=1}^r g^T d^t + \frac{1}{2} \sum_{t=1}^r (d^t)^T D d^t + \sum_{t=1}^r (cP(x+d^t) - cP(x)) \\
&\ge r \min_{t=1,\dots,r} \left\{ g^T d^t + \frac{1}{2} (d^t)^T D d^t + cP(x+d^t) - cP(x) \right\} \\
&= r \min_{t=1,\dots,r} \left\{ g^T d^t + \frac{1}{2} (d^t)^T D d^t + cQ(x+d^t) - cQ(x) \right\},
\end{aligned}$$

where $g = \nabla f(x)$ and the first inequality uses (22) and $D \succ 0$ being diagonal, so that $(d^s)^T D d^t \ge 0$ for all s, t; the second inequality uses Lemma 6.1. Thus, if we let $\bar{t}$ be an index t attaining the above minimum and let $\mathcal{J} = \operatorname{supp}(d^{\bar{t}})$, then $|\mathcal{J}| \le \ell$ and

$$\frac{1}{r} q_D(x; \mathcal{N}) \ge g^T d^{\bar{t}} + \frac{1}{2}(d^{\bar{t}})^T D d^{\bar{t}} + cQ(x + d^{\bar{t}}) - cQ(x) \ge q_D(x; \mathcal{J}),$$

where the second inequality uses $A(x + d^{\bar{t}}) = b$ and $d_j^{\bar{t}} = 0$ for $j \notin \mathcal{J}$. □

It can be seen from its proof that Proposition 6.1 still holds if the diagonal matrix D is only positive semidefinite, provided that $q_D(x; \mathcal{N}) > -\infty$ (such as when dom Q is bounded). However, Proposition 6.1 is false if we drop the assumption that P is separable. Take

$$m = 1,\ n = 3, \quad f(x) = x_1 + x_2 + x_3, \quad P(x) = \sqrt{x_1^2 + x_2^2} + |x_3|,$$
$$A = [1\ 1\ -1], \quad b = 0.$$

Then, $x = 0$ is not a stationary point ($d = (-1, -1, -2)^T$ is a feasible descent direction), so $q_D(x; \mathcal{N}) < 0$ for any $D \succ 0$. However, it is straightforward to check that $q_D(x; \mathcal{J}) \ge 0$ whenever $|\mathcal{J}| \le 2$.

The proof of Proposition 6.1 suggests, for any $\ell \in \{\operatorname{rank}(A) + 1, \dots, n\}$, an $O(n - \ell)$-step reduction procedure for finding a conformal realization (26) of $d_D(x; \mathcal{N})$ with $r \le n - \ell + 1$ and a corresponding $\mathcal{J}$ satisfying $|\mathcal{J}| \le \ell$ and (25). In the case of $m = 1$ and $\ell = 2$, such a conformal realization can be found in $O(n)$ operations, as is discussed in [29, Sect. 7]. In the case of $m = 2$ and $\ell = 3$, such a conformal realization can be found in $O(n \log n)$ operations. For $m \ge 3$, the currently best time complexity of finding such a conformal realization is $O(m^3(n - \ell)^2)$ operations. See [29, Sect. 7] for more detailed discussions.

There remains the question of how to find $d_D(x; \mathcal{N})$ with $D \succ 0$ diagonal. In the linearly constrained case of $P_j = \delta_{[l_j, u_j]}$ for all j, as is considered in [29], this reduces to a quadratic program with separable convex objective function of the form

$$\min_d \left\{ \nabla f(x)^T d + \frac{1}{2} d^T D d \mid Ad = 0,\ l - x \le d \le u - x \right\}, \tag{27}$$

which is solvable in $O(n)$ operations for m fixed [38, 39]; also see [40, 41] and references therein for the special case of $m = 1$. For general P_j, finding $d_D(x; \mathcal{N})$ reduces to a monotropic optimization problem which can be solved using various methods; see [7, 42] and references therein. However, these methods in general do not run in linear time. If each P_j is polyhedral or, more generally, piecewise-linear/quadratic with ν_j pieces, then, as we show below, finding $d_D(x; \mathcal{N})$ is reducible to a problem of the form (27) with $\nu_1 + \dots + \nu_n$ variables, and hence is solvable in $O(\nu_1 + \dots + \nu_n)$ operations for m fixed. Here we assume without loss of generality that dom P_j is not a singleton, so that $\nu_j \ge 1$. In particular, since D is diagonal, $d_D(x; \mathcal{N})$ is the optimal solution of a problem of the form

$$\min_d \left\{ \sum_{j=1}^n \Pi_j(d_j) \,\middle|\, Ad = 0 \right\}, \tag{28}$$

where each Π_j is strictly convex, piecewise-linear/quadratic with ν_j pieces. Let the breakpoints of Π_j be denoted by $-\infty \le a_j^0 < a_j^1 < \cdots < a_j^{\nu_j} \le \infty$ (so a_j^0 and $a_j^{\nu_j}$ are the endpoints of $\operatorname{dom}\Pi_j$). Let

$$\Pi_j^1(d_j^1) \stackrel{\text{def}}{=} \begin{cases} \Pi_j(a_j^1 + d_j^1), & \text{if } 0 \ge d_j^1 \ge a_j^0 - a_j^1; \\ \infty, & \text{else,} \end{cases}$$

$$\Pi_j^\ell(d_j^\ell) \stackrel{\text{def}}{=} \begin{cases} \Pi_j(a_j^{\ell-1} + d_j^\ell), & \text{if } 0 \le d_j^\ell \le a_j^\ell - a_j^{\ell-1}, \\ \infty, & \text{else,} \end{cases} \qquad \ell = 2, \ldots, \nu_j.$$

We consider the following problem

$$\min_{d_j^\ell} \left\{ \sum_{j=1}^{n} \sum_{\ell=1}^{\nu_j} \Pi_j^\ell(d_j^\ell) \;\middle|\; \sum_{j=1}^{n} \sum_{\ell=1}^{\nu_j} A_j d_j^\ell = 0 \right\}. \tag{29}$$

This problem, with $\nu_1 + \cdots + \nu_n$ variables, has the same form as (27) since the objective function is separable and each component function is strictly convex quadratic over its domain. Moreover, the optimal solution of (29) must satisfy

$$d_j^1 d_j^2 = 0, \qquad d_j^{\ell+1} > 0 \;\Rightarrow\; d_j^\ell = a_j^\ell - a_j^{\ell-1}, \quad \ell = 2, \ldots, \nu_j,\; j = 1, \ldots, n. \tag{30}$$

If $d_j^1 d_j^2 \ne 0$, then $d_j^1 < 0$, $d_j^2 > 0$, and the strict convexity of Π_j would imply

$$\frac{\Pi_j(a_j^1 + d_j^1 + d_j^2) - \Pi_j(a_j^1 + d_j^1)}{d_j^2} < \frac{\Pi_j(a_j^1 + d_j^2) - \Pi_j(a_j^1)}{d_j^2}$$

and hence $\Pi_j(a_j^1 + d_j^1 + d_j^2) + \Pi_j(a_j^1) < \Pi_j(a_j^1 + d_j^1) + \Pi_j(a_j^1 + d_j^2)$. Then replacing d_j^1, d_j^2 by $d_j^1 + d_j^2$, 0 when $d_j^1 + d_j^2 < 0$ (and replacing d_j^1, d_j^2 by 0, $d_j^1 + d_j^2$ when $d_j^1 + d_j^2 \ge 0$) would yield another feasible solution of (29) with a lower objective value. The second condition in (30) can be argued similarly. Hence, by using

$$d_j = a_j^1 + d_j^1 + \cdots + d_j^{\nu_j}, \quad j = 1, \ldots, n, \tag{31}$$

we can construct from the optimal solution of (29) a feasible solution of (28) with the same objective value. Conversely, we can construct from the optimal solution of (28) a feasible solution of (29) that has the same objective value and satisfies (30), (31).

By combining the above observations, we can conclude the following about finding an index subset $\mathcal{J}$ satisfying $|\mathcal{J}| \le \ell$ and (25) when each P_j is piecewise-linear/quadratic with $O(1)$ pieces: For $m = 1$ and $\ell = 2$, $\mathcal{J}$ can be found in $O(n)$ operations and, for $m \ge 2$ and $\ell \in \{\operatorname{rank}(A) + 1, \ldots, n\}$, $\mathcal{J}$ can be found in $O(n^2)$ operations, where the constant in $O(\cdot)$ depends on m. It is an open question whether this can be improved to $O(n)$ operations.

Note that $r^0 \le n b_{\max}^2$, where $b_{\max} = \max_{1 \le j \le n}(u_j - l_j)$ and $l_j \le u_j$ denote the endpoints of $\operatorname{dom} P_j$, which we assume to be bounded. Thus, if f is convex and satisfies (17) for some ℓ, then it follows from (21) that, for $m = 1$ and $\ell = 2$, the

CGD method can be implemented to achieve ϵ-optimality in

$$O\left(\frac{n^2 L_2 b_{\max}^2}{\epsilon} + nL_2 \max\left\{0, \log\left(\frac{e^0}{nb_{\max}}\right)\right\}\right) \cdot O(n + N_f)$$

operations, where N_f is the number of operations for evaluating f and ∇f at the current iterate. If in addition f is quadratic or has the partially separable form

$$f(x) = g(Ex) + q^T x,$$

where $g : \Re^p \to (-\infty, \infty]$ is convex block-separable with $O(1)$ size blocks, $q \in \Re^n$, and each column of $E \in \Re^{p\times n}$ has $O(1)$ nonzeros, then $N_f = O(n)$. When specialized to the training of SVM, for which $P_j = \delta_{[l_j,u_j]}$, $A = [1 \ \cdots \ 1]$, and f is quadratic, the preceding complexity bound reduces to

$$O\left(\frac{n^3 \Lambda b_{\max}^2}{\epsilon} + n^2 \Lambda \max\left\{0, \log\left(\frac{e^0}{nb_{\max}}\right)\right\}\right)$$

operations, where $\Lambda = \max_{i\neq j} \sqrt{(H_{ii} - H_{ij})^2 + (H_{jj} - H_{ij})^2}/\sqrt{2}$ and $H = \nabla^2 f(x)$. For this same problem, Hush and Scovel [32] proposed a decomposition method, based on block-coordinate descent, and proved that, for any $\epsilon > 0$, the method finds an ϵ-optimal solution in $O(b_{\max}^2 n^3 \log n(e^0 + n^2\Lambda)/\epsilon)$ operations. This method was extended by List and Simon [26] to problems with general linear constraints, and the overall complexity bound was improved to $O(\frac{n^3\Lambda b_{\max}^2}{\epsilon} + n^2 \max\{0, \log(\frac{e^0}{n\Lambda b_{\max}})\})$ operations. Hush et al. [33] later proposed a more practical decomposition method that achieves the same complexity bounds as in [26]. Our complexity bound for the CGD method on this problem is comparable to the above bound when $m = 1$ (which covers SVM), and is off by a factor of $\log n$ when $m = 2$ and by a factor of n when $m \geq 3$, due to the extra cost of finding a conformal realization of $d_D(x; \mathcal{N})$. This extra cost is the price for achieving linear convergence shown in Theorem 4.2.

7 Bilevel Optimization

In this section, we show that when f is convex, we can apply the CGD method to solve the bilevel problem (3) by decreasing c towards zero whenever the current iterate x^k is an approximate stationary point of (1). In particular, by Lemma 3.1, $\|d_{D^k}(x^k; \mathcal{N})\|$ acts as a "residual" function, measuring how close x^k comes to being stationary for (1). We will use the following measure of approximate stationarity:

$$\|d_{D^k}(x^k; \mathcal{N})\| \leq \epsilon^k, \qquad \|D^k d_{D^k}(x^k; \mathcal{N})\| \leq \epsilon^k, \tag{32}$$

$$-(D^k x^k + \nabla f(x^k))^T d_{D^k}(x^k; \mathcal{N}) \leq \epsilon^k, \tag{33}$$

with $\epsilon^k > 0$ to be specified. Notice that if $\mathcal{J}^k$ is chosen as described in Sect. 6, then $d_{D^k}(x^k; \mathcal{N})$ would be available as a byproduct and need not be computed additionally.

Our method for solving (3) uses similar idea as in [17] for a primal-dual interior-point method. At each outer iteration k ($k=0,1,2,\ldots$), a regularization parameter $c^k>0$ and an accuracy tolerance ϵ^k are chosen, and the CGD method is applied to solve (1) with $c=c^k$ until it finds an approximate solution x^k satisfying the conditions (32) and (33). Since the idea of decreasing c is reminiscent of homotopy methods for equation solving, we call this the CGD-homotopy method.

CGD-Homotopy Method

Choose $x^0 \in \operatorname{dom} Q$. For $k=1,2,\ldots,$ generate x^k from x^{k-1} according to the following outer iteration:

Step 1. Choose $c^k>0$ and $\epsilon^k>0$.

Step 2. Compute an $x^k \in \operatorname{dom} Q$ satisfying (32) and (33) for some $D^k \succ 0$ by applying the CGD method to (1) with $c=c^k$ and initial iterate $x=x^{k-1}$.

The following theorem shows that, by letting $c^k \to 0$ and $\epsilon^k \to 0$ at suitable rates in the CGD-homotopy method, every cluster point of the approximate solutions $\{x^k\}$ solves (3).

Theorem 7.1 *Suppose f is convex, $S_f \cap \operatorname{dom} Q \neq \emptyset$, and (3) has an optimal solution. Consider any c^k and ϵ^k, $k=1,2,\ldots,$ satisfying*

$$\lim_{k\to\infty} c^k = 0, \qquad \lim_{k\to\infty} \frac{\epsilon^k}{c^k} = 0. \tag{34}$$

Consider any x^k satisfying (32) and (33) with $c=c^k$ for $k=1,2,\ldots$. Then every cluster point of $\{x^k\}$ is an optimal solution of (3). If Q is level-bounded, then $\{x^k\}$ has a cluster point.

Proof Let x^* be any optimal solution of (3), i.e., $x^* \in \arg\min_{x\in S_f} Q(x)$. By Fermat's rule [34, Theorem 10.1],

$$\hat{d}^k \in \arg\min_d (g^k + D^k\hat{d}^k)^T d + c^k Q(x^k+d) - c^k Q(x^k),$$

where we let $\hat{d}^k = d_{D^k}(x^k;\mathcal{N})$ and $g^k = \nabla f(x^k)$. Hence,

$$\begin{aligned}&(g^k + D^k\hat{d}^k)^T\hat{d}^k + c^k Q(x^k+\hat{d}^k) - c^k Q(x^k)\\ &\quad\le (g^k + D^k\hat{d}^k)^T(x^*-x^k) + c^k Q(x^*) - c^k Q(x^k).\end{aligned}$$

Using $(\hat{d}^k)^T D^k \hat{d}^k \ge 0$ and rearranging and canceling terms, we obtain

$$\begin{aligned}&f(x^k) + c^k Q(x^k+\hat{d}^k)\\ &\quad\le f(x^k) + (g^k + D^k\hat{d}^k)^T(x^*-x^k) + c^k Q(x^*) - (g^k)^T\hat{d}^k\\ &\quad\le f(x^*) + (D^k\hat{d}^k)^T(x^*-x^k) + c^k Q(x^*) - (g^k)^T\hat{d}^k\\ &\quad\le f(x^*) + \|D^k\hat{d}^k\|\|x^*\| - (D^kx^k+g^k)^T\hat{d}^k + c^k Q(x^*)\\ &\quad\le f(x^*) + \epsilon^k\|x^*\| + \epsilon^k + c^k Q(x^*),\end{aligned} \tag{35}$$

where the second inequality follows from f being convex, so that $f(x^k) + (g^k)^T(x^* - x^k) \le f(x^*)$, and the last inequality uses (32) and (33). Since $x^* \in S_f$ and $x^k \in \operatorname{dom} Q$, $f(x^*) \le f(x^k)$. This together with (35) implies

$$c^k Q(x^k + \hat{d}^k) \le \epsilon^k \|x^*\| + \epsilon^k + c^k Q(x^*).$$

Dividing both sides by c^k yields

$$Q(x^k + \hat{d}^k) \le \frac{\epsilon^k}{c^k}\|x^*\| + \frac{\epsilon^k}{c^k} + Q(x^*). \tag{36}$$

By (34), $\{c^k\} \to 0$ and $\{\epsilon^k\} \to 0$, so that, by (32), $\{\hat{d}^k\} \to 0$. This, together with (35) and Q being convex (so Q is bounded below on any compact set), implies that any cluster point $\bar{x}$ of $\{x^k\}$ satisfies $f(\bar{x}) \le f(x^*)$. Since $x^k \in \operatorname{dom} Q$ for all k, $\bar{x} \in X$. Moreover, (34), (36), $\{\hat{d}^k\} \to 0$, and the lsc property of Q imply $Q(\bar{x}) \le Q(x^*)$. Thus $\bar{x} \in S_f$ and $\bar{x}$ is an optimal solution of (3).

Suppose Q is level-bounded. By (34) and (36), $\{x^k + \hat{d}^k\}$ is bounded. This together with $\{\hat{d}^k\} \to 0$ implies that $\{x^k\}$ has cluster points. □

It is not known if Theorem 7.1 still holds if we replace "$d_{D^k}(x^k; \mathcal{N})$" in (32) and (33) by "$d_{H^k}(x^k; \mathcal{J}^k)$" with $\mathcal{J}^k$ satisfying (10), even though the latter is also available as a byproduct of the CGD method. Thus the notion of approximate stationarity for (1) must be chosen with care. The following lemma shows that the bilevel problem (3) has an optimal solution under a mild assumption on Q.

Lemma 7.1 *Suppose $S_f \cap \operatorname{dom} Q \ne \emptyset$ and Q is level-bounded over S_f. Then the minimum of Q over S_f is finite and attained on a nonempty compact subset of S_f.*

Proof Let $\tilde{Q} \equiv \delta_{S_f} + Q$, where δ_{S_f} is the indicator function of the set S_f. Then $\tilde{Q}$ is proper because $S_f \cap \operatorname{dom} Q \ne \emptyset$, and it is lsc since its level sets $S_f \cap \{x \mid Q(x) \le \xi\}$, with $\xi < \infty$, are closed (due to S_f being closed and Q being lsc). Since Q is level-bounded over S_f, these level sets are bounded. Then the minimum of $\tilde{Q}$ is finite and attained on a nonempty compact set. □

8 Conclusions and Extensions

We have extended a block-coordinate gradient descent method to linearly constrained nonsmooth separable minimization, and have analyzed its global convergence and asymptotic convergence rate. In the case where f is convex, we also analyzed its computational complexity and presented a homotopy strategy to solve a bilevel version of the problem.

There are many directions for extensions. Can the complexity bound in Sect. 5 be sharpened? Can the homotopy strategy be extended to handle nonconvex f? The Gauss-Southwell-r rule for choosing $\mathcal{J}^k$, studied in [6] for the case of $m = 0$, can also be extended to the case of $m \ge 1$. We did not consider it here because (i) we do not have a convergence rate analysis analogous to Theorem 4.2 and (ii) our numerical experience in [6] suggests that this rule is not better than the Gauss-Southwell-q rule in practice. The classical Gauss-Seidel rule for choosing $\mathcal{J}^k$, studied in [6] for the case of $m = 0$, can also be extended to the case of $m \geq 1$ provided P is separable. However, this rule seems impractical since it would require cycling through $\binom{n}{m+1}$ coordinate blocks of size $m + 1$ each.

Suppose P is not separable but block-separable of the form

$$P(x) = \sum_{\mathcal{J} \in \mathcal{C}} P_{\mathcal{J}}(x_{\mathcal{J}}),$$

where $\mathcal{J} \in \mathcal{C}$ form a partition of $\mathcal{N}$. Then Lemma 6.1 and Proposition 6.1 are no longer applicable as we saw in Sect. 6. This case is of practical interest as it arises in group Lasso, for which $P_{\mathcal{J}}(x_{\mathcal{J}}) = \|x_{\mathcal{J}}\|$; see [30]. Can we still efficiently find a small $\mathcal{J}^k$ satisfying (10)? Can the Gauss-Seidel rule, used in [6, 30] for the case of $m = 0$, be extended to the case of $m \geq 1$? This is open even when $m = 1$ and $P_{\mathcal{J}}(x_{\mathcal{J}}) = \|x_{\mathcal{J}}\|$.

Problem (1) can be generalized to the following problem:

$$\min_{x \in \Re^n} \{f(x) + cP(x) \mid f_1(x) = 0, \ldots, f_m(x) = 0\},$$

where $f_1, \ldots, f_m$ are twice continuously differentiable functions. Can the CGD method be extended to solve this more general problem?

References

1. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
2. Fukushima, M.: A successive quadratic programming method for a class of constrained nonsmooth optimization problems. Math. Program. **49**, 231–251 (1990–1991)
3. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. **12**, 989–1000 (1981)
4. 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)
5. Mine, H., 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)
6. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. B **117**, 387–423 (2009)
7. Rockafellar, R.T.: Network Flows and Monotropic Optimization. Wiley, New York (1984). Republished by Athena Scientific, Belmont (1998)
8. Bertsekas, D.P.: Nonlinear Programming. 2nd edn. Athena Scientific, Belmont (1999)
9. Fletcher, R.: Practical Methods of Optimization. 2nd edn. Wiley, New York (1987)
10. Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. Academic Press, New York (1981)
11. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (1999)
12. Candés, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
13. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. SIAM Rev. **43**, 129–159 (2001)
14. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994)
15. Friedlander, M.P., Tseng, P.: Exact regularization of convex programs. SIAM J. Optim. **18**, 1326–1350 (2007)
16. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**, 302–332 (2007)

17. Sardy, S., Tseng, P.: AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. J. Comput. Graph. Stat. **13**, 283–309 (2004)
18. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
19. Tseng, P.: Convergence of block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 473–492 (2001)
20. Joachims, T.: Making large-scale SVM learning practical. In: Schólkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods—Support Vector Learning, pp. 169–184. MIT Press, Cambridge (1999)
21. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. In: Schólkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods—Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
22. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, London (1997)
23. Chen, P.-H., Fan, R.-E., Lin, C.-J.: A study on SMO-type decomposition methods for support vector machines. IEEE Trans. Neural Netw. **17**, 893–908 (2006)
24. Grippo, L., Sciandrone, M.: On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. Oper. Res. Lett. **26**, 127–136 (2000)
25. Lin, C.-J., Lucidi, S., Palagi, L., Risi, A., Sciandrone, M.: A decomposition algorithm model for singly linearly constrained problems subject to lower and upper bounds. Technical Report, DIS-Università di Roma "La Sapienza", Rome (2007). J. Optim. Theory Appl. (to appear)
26. List, N., Simon, H.U.: General Polynomial Time Decomposition Algorithms. Lecture Notes in Computer Science, vol. 3559, pp. 308–322. Springer, Berlin (2005)
27. Luo, Z.Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**, 408–425 (1992)
28. Mangasarian, O.L., Musicant, D.R.: Successive overrelaxation for support vector machines. IEEE Trans. Neural Netw. **10**, 1032–1037 (1999)
29. Tseng, P., Yun, S.: A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Report, Department of Mathematics, University of Washington, Seattle (2007). Comput. Optim. Appl. (to appear)
30. Meier, L., van de Geer, S., Bühlmann, P.: The group Lasso for logistic regression. J. R. Stat. Soc. B **70**, 53–71 (2008)
31. Rockafellar, R.T.: The elementary vectors of a subspace of R^N. In: Bose, R.C., Dowling, T.A. (eds.) Combinatorial Mathematics and its Applications. Proc. of the Chapel Hill Conference 1967, pp. 104–127. Univ. North Carolina Press, Chapel Hill (1969)
32. Hush, D., Scovel, C.: Polynomial-time decomposition algorithms for support vector machines. Mach. Learn. **51**, 51–71 (2003)
33. Hush, D., Kelly, P., Scovel, C., Steinwart, I.: QP algorithms with guaranteed accuracy and run time for support vector machines. J. Mach. Learn. Res. **7**, 733–769 (2006)
34. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, New York (1998)
35. Luo, Z.Q., Tseng, P.: On the convergence rate of dual ascent methods for linearly constrained convex minimization. Math. Oper. Res. **18**, 846–867 (1993)
36. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**, 157–178 (1993)
37. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. SIAM, Philadelphia (2000)
38. Berman, P., Kovoor, N., Pardalos, P.M.: Algorithms for the least distance problem. In: Pardalos, P.M. (ed.) Complexity in Numerical Optimization, pp. 33–56. World Scientific, Singapore (1993)
39. Megiddo, N., Tamir, A.: Linear time algorithms for some separable quadratic programming problems. Oper. Res. Lett. **13**, 203–211 (1993)
40. Brucker, P.: An $O(n)$ algorithm for quadratic knapsack problems. Oper. Res. Lett. **3**, 163–166 (1984)
41. Kiwiel, K.C.: On linear time algorithms for the continuous quadratic knapsack problem. J. Optim. Theory Appl. **134**, 549–554 (2007)
42. Tseng, P.: An ϵ-out-of-kilter method for monotropic programming problems. Math. Oper. Res. **26**, 221–233 (2001)