Approximation Accuracy, Gradient Methods, and Error Bound for Structured Convex Optimization[1]

May 14, 2009 (revised July 7, 2009)

Paul Tseng[2]

Abstract

Convex optimization problems arising in applications, possibly as approximations of intractable problems, are often structured and large scale. When the data are noisy, it is of interest to bound the solution error relative to the (unknown) solution of the original noiseless problem. Related to this is an error bound for the linear convergence analysis of first-order gradient methods for solving these problems. Example applications include compressed sensing, variable selection in regression, TV-regularized image denoising, and sensor network localization.

Key words. Convex optimization, compressed sensing, ℓ_1-regularization, nuclear/trace norm, regression, variable selection, sensor network localization, approximation accuracy, proximal gradient method, error bound, linear convergence.

1 Introduction

Optimization problems arising in application areas such as signal/image denoising, compressed sensing, regression, multi-task learning, classification, sensor network localization, are often large scale, possibly nonconvex or NP-hard, and the data may be noisy. Such problems may be approximated by convex relaxations that are highly structured.

Question 1: How accurate approximations are the convex relaxations?

Specifically, can we bound the error between a solution of the convex relaxation and an (unknown) solution of the original noiseless problem in terms of knowable quantities such as the noise level? This question is meaningful when a solution of the original problem changes (semi)continuously with small perturbations in the data, so the problem has both discrete and continuous nature. Examples include compressed sensing and sensor network localization; see Section 2. A certain noise-aware property of the convex relaxation appears key.

Question 2: How fast can the convex relaxations be solved?

Due to the large problem size, first-order gradient methods seem better suited to exploit structures such as sparsity, (partial) separability, and simple nonsmoothness in the convex relaxations;

[1]This work is supported by National Science Foundation grant DMS-0511283.

[2]Department of Mathematics, University of Washington, Seattle, WA 98195, U.S.A. (tseng@math.washington.edu)

see Section 3. The asymptotic convergence rate of these methods depend on an error bound on the distance to the solution set of the convex relaxation in terms of a certain residual function. We prove such an error bound for a class of 2-norm-regularized problems that includes the group lasso for linear and logistic regression; see Section 4. Thus our aims are threefold: exposit on existing results, present a new result, and suggest future research.

We begin with a problem that has received much attention recently: compressed sensing. In the basic version of this problem, we wish to find a sparse representation of a given noiseless signal $b^0 \in \Re^m$ from a dictionary of n elementary signals. This may be formulated as

$$\min_{x|Ax=b^0} \sharp(x), \tag{1}$$

where $A \in \Re^{m\times n}$ comprises the elementary signals for its columns and $\sharp(x)$ counts the number of nonzero components in $x \in \Re^n$. In typical applications, m and n are large ($m, n \geq 2000$). This problem is known to be difficult (NP-hard) and a popular solution approach is to approximate it by a convex relaxation, with $\sharp(\cdot)$ replaced by the 1-norm $\|\cdot\|_1$.[3] This results in a linear program:

$$\min_{x|Ax=b^0} \|x\|_1 \tag{2}$$

that can be efficiently solved by simplex or interior point methods [28, 103, 105, 150, 154]. Moreover, when the optimal value of (1) is sufficiently small and the columns of A are "approximately orthogonal," which occur with high probability when A is, say, a Gaussian random matrix, the solution of (2) also solves (1), i.e., the relaxation is exact [26, 27, 28, 35, 40, 42, 56, 57, 59, 62, 83, 111, 133, 134]. For a noisy signal b, $Ax = b$ may be inconsistent and we seek a sparse solution x whose residual $Ax - b$ is small in the least square sense, either in the primal form of "lasso" [131]

$$\min_{x|\ \|Ax-b\|_2\leq\rho} \|x\|_1 \tag{3}$$

($\rho > 0$) or in the dual form of "basis pursuit" [30, 43, 44, 121]

$$\min_x \|Ax - b\|_2^2 + \tau\|x\|_1 \tag{4}$$

($\tau > 0$). Partial results on the accuracy of this relaxation are known [41, 133, 143]. Various methods have been proposed to solve (3) and (4) for fixed ρ and τ, including interior point method [30], coordinate minimization [121, 122], proximal gradient (with or without smoothing) [13, 14, 34], proximal minimization [155]. Homotopy approaches have been proposed to follow the solution as τ varies between 0 and ∞ [54, 63, 110]. The efficiency of these methods depend on the structure of A. The polyhedral and separable structure of $\|\cdot\|_1$ is key.

A problem related to (4) is image denoising using total variation (TV) regularization:

$$\min_{x\in\mathcal{B}_2^{\frac{n}{2}}} \|Ax - b\|_2^2, \tag{5}$$

where n is even, $\mathcal{B}_2$ denotes the unit 2-norm (Euclidea) ball in $\Re^2$, $A \in \Re^{m\times n}$ is the adjoint of the discrete (via finite difference) gradient mapping, and $b \in \Re^m$. Interior-point, proximal

[3] We are using the term "relaxation" loosely since $\sharp(\cdot)$ majorizes $\|\cdot\|_1$ only on the unit ∞-norm ball.

minimization, coordinate minimization, and gradient projection methods have been proposed for its solution [60, 108, 147, 152, 160, 161].

A second problem related to (1) and of growing interest is matrix rank minimization. The basic problem is

$$\min_{x|Ax=b^0} \operatorname{rank}(x), \tag{6}$$

where $\operatorname{rank}(x)$ is the rank of a matrix $x \in \Re^{p\times n}$, and A is a linear mapping from $\Re^{p\times n}$ to $\Re^m$, and $b^0 \in \Re^m$. In the case of matrix completion, we have $Ax = (x_{ij})_{(i,j)\in\mathcal{A}}$, where $\mathcal{A}$ indexes the known entries of x. This problem is also NP-hard, and a convex relaxation has been proposed whereby $\operatorname{rank}(x)$ is replaced by the nuclear/trace norm $\|x\|_{\text{nuc}}$ (the 1-norm of the singular values of x) [24, 25, 50, 73, 82, 115, 156]. The resulting problem

$$\min_{x|Ax=b^0} \|x\|_{\text{nuc}} \tag{7}$$

has a more complex structure than (2), and only recently have solution methods, including interior point method and dual gradient method, been developed [24, 73, 82] and exactness results been obtained [25, 115]. For noisy b, we may consider, analogous to (4), the relaxation

$$\min_x \|Ax - b\|_F^2 + \tau\|x\|_{\text{nuc}}, \tag{8}$$

where $\tau > 0$ and $\|\cdot\|_F$ denotes the Frobenious-norm. This problem has applications to dimension reduction in multivariate linear regression [75] and multi-task learning [1, 3, 106]. Recently, proximal/gradient methods have been applied to its solution [75, 82, 132]. How well does (8) approximate (6)?

A third problem related to (1) and of recent interest is that of sparse inverse covariance estimation, where we seek a positive definite $x \in \mathcal{S}^n$ that is sparse and whose inverse approximates a given sample covariance matrix $s \in \mathcal{S}^n$. This may be formulated as the nonconvex problem

$$\min_{\underline{\lambda}I\preceq x\preceq\bar{\lambda}I} -\log\det(x) + \langle s, x\rangle + \tau\sharp(x) \tag{9}$$

with $0 \le \underline{\lambda} < \bar{\lambda} \le \infty$ and $\tau > 0$, where $\langle s, x\rangle = \operatorname{trace}[sx]$, $\sharp(x)$ counts the number of nonzeros in x, I denotes the $n\times n$ identity matrix, and $\preceq$ denotes the partial ordering with respect to the cone $\mathcal{S}_+^n$ of positive semidefinite matrices [5, 9, 53, 74, 158]. Replacing $\sharp(x)$ by $\|x\|_1 := \sum_{i,j=1}^n |x_{ij}|$ yields the convex relaxation

$$\min_{\underline{\lambda}I\preceq x\preceq\bar{\lambda}I} -\log\det(x) + \langle s, x\rangle + \tau\|x\|_1. \tag{10}$$

Interior-point method appears unsuited for solving (10) owing to the large size of the Newton equation to be solved. Block-coordinate minimization methods (with each block corresponding to a row/column of x) and Nesterov's accelerated gradient methods have been applied to its solution [5, 9, 53, 74]. How well does (10) approximate (9)?

A fourth problem of much recent interest is that of ad hoc wireless sensor network localization [4, 17, 18, 20, 21, 29, 36, 37, 47, 70, 72, 127]. In the basic version of this problem, we have n points $z_1^0, \ldots, z_n^0$ in $\Re^d$ ($d \ge 1$). We know the last $n - m$ points ("anchors") and an estimate $d_{ij} \ge 0$ of the Euclidean distance $d_{ij}^0 = \|z_i^0 - z_j^0\|_2$ between "neighboring" points i and j for all

$(i,j) \in \mathcal{A}$, where $\mathcal{A} \subseteq (\{1,\ldots,m\}\times\{1,\ldots,n\}) \cup (\{1,\ldots,n\}\times\{1,\ldots,m\})$. We wish to estimate the first m points ("sensors"). This problem may be formulated as

$$\min_{z_1,\ldots,z_m} \sum_{(i,j)\in\mathcal{A}^s} |\|z_i - z_j\|_2^2 - d_{ij}^2| + \sum_{(i,j)\in\mathcal{A}^a} |\|z_i - z_j^0\|_2^2 - d_{ij}^2|, \tag{11}$$

where $\mathcal{A}^s := \{(i,j) \in \mathcal{A} \mid i < j \leq m\}$ and $\mathcal{A}^a := \{(i,j) \in \mathcal{A} \mid i \leq m < j\}$ are the sets of, respectively, sensor-to-sensor and sensor-to-anchor neighboring pairs. Typically, $d = 2$ and two points are neighbors if the distance between them is below some threshold, e.g., the radio range. In variants of this problem, constraints such as bounds on distances and angles-of-arrival are also present [16, 29, 37, 70]. This problem is NP-hard for any $d \geq 1$. It is closely related to distance geometry problems arising in molecular conformation [19, 90], graph rigidity/realization [2, 36, 47, 127], and max-min/avg dispersion [33, 114, 149]. Letting $z := (\, z_1 \quad \cdots \quad z_m \,)$ and I denote the $d \times d$ identity matrix, Biswas and Ye [20, 21] proposed the following convex relaxation of (11):

$$\begin{array}{ll} \min\limits_{x} & \sum\limits_{(i,j)\in\mathcal{A}} \left|\ell_{ij}(x) - d_{ij}^2\right| \\ \text{subject to} & x = \begin{pmatrix} y & z^T \\ z & I \end{pmatrix} \succeq 0, \end{array} \tag{12}$$

where $y = (\, y_{ij} \,)_{1\leq i,j\leq m}$, and

$$\ell_{ij}(x) := \begin{cases} y_{ii} - 2y_{ij} + y_{jj} & \text{if } (i,j) \in \mathcal{A}^s, \\ y_{ii} - 2z_i^T z_j^0 + \|z_j^0\|^2 & \text{if } (i,j) \in \mathcal{A}^a. \end{cases} \tag{13}$$

The relaxation (12) is a semidefinite program (SDP) and is exact if it has a solution of rank d. On the other hand, (12) is still difficult to solve by existing methods for SDP when $m > 500$, and decomposition methods have been proposed [21, 29]. Recently, Wang, Zheng, Boyd, and Ye [148] proposed a further relaxation of (12), called edge-based SDP (ESDP) relaxation, which is solved much faster by an interior point method than (12), and yields solution comparable in approximation accuracy as (12). The ESDP relaxation is obtained by relaxing the constraint $x \succeq 0$ in (12) to require only those principal submatrices of x associated with $\mathcal{A}$ to be positive semidefinite. Specifically, the ESDP relaxation is

$$\begin{array}{ll} \min\limits_{x} & \sum\limits_{(i,j)\in\mathcal{A}} \left|\ell_{ij}(x) - d_{ij}^2\right| \\ \text{subject to} & x = \begin{pmatrix} y & z^T \\ z & I \end{pmatrix}, \\ & \begin{pmatrix} y_{ii} & y_{ij} & z_i^T \\ y_{ij} & y_{jj} & z_j^T \\ z_i & z_j & I \end{pmatrix} \succeq 0 \quad \forall (i,j) \in \mathcal{A}^s, \\ & \begin{pmatrix} y_{ii} & z_i^T \\ z_i & I_d \end{pmatrix} \succeq 0 \qquad\quad \forall i \leq m. \end{array} \tag{14}$$

Notice that the objective function and the positive semidefinite constraints in (14) do not depend on y_{ij}, $(i,j) \notin \mathcal{A}$. How well does (12) approximate (11)? Only partial results are known in the noiseless case, i.e., (11) has zero optimal value [127, 141]. How well does (14) approximate (11)?

The above convex problems share a common structure, namely, they entail minimizing the sum of a smooth (i.e., continuously differentiable) convex function and a "simple" nonsmooth

convex function. More specifically, they have the form

$$\min_{x\in\mathcal{E}} F(x) := f(x) + \tau P(x), \tag{15}$$

where $\mathcal{E}$ is a finite-dimensional real linear space endowed with a norm $\|\cdot\|$, $\tau > 0$, $P : \mathcal{E} \to (-\infty, \infty]$ is lower semicontinuous (lsc), convex, with $\mathrm{dom}P = \{x \mid P(x) < \infty\}$ closed, and $f : \mathcal{E} \to (-\infty, \infty]$ is convex and smooth on $\mathrm{dom}f$, assumed open, and $f(x) \to \infty$ whenever x approaches a boundary point of $\mathrm{dom}f$ [116]. A well-known special case of (15) is smooth constrained convex optimization, for which P is the indicator function for a nonempty closed convex set $\mathcal{X} \subseteq \mathcal{E}$, i.e.,

$$P(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{else.} \end{cases} \tag{16}$$

The class of problems (15) was studied in [6, 89] and by others; see [144] and references therein. For example, (4) corresponds to

$$\mathcal{E} = \Re^n, \quad \|\cdot\| = \|\cdot\|_2, \quad f(x) = \|Ax - b\|_2^2, \quad P(x) = \|x\|_1, \tag{17}$$

(5) corresponds to

$$\mathcal{E} = \Re^n, \quad \|\cdot\| = \|\cdot\|_2, \quad f(x) = \|Ax - b\|_2^2, \quad P(x) = \begin{cases} 0 & \text{if } x \in \mathcal{B}_2^{\frac{n}{2}}, \\ \infty & \text{else,} \end{cases} \tag{18}$$

(8) corresponds to

$$\mathcal{E} = \Re^{p\times n}, \quad \|\cdot\| = \|\cdot\|_F, \quad f(x) = \|Ax - b\|_F^2, \quad P(x) = \|x\|_{\mathrm{nuc}}, \tag{19}$$

and (10) corresponds to

$$\mathcal{E} = \mathcal{S}^n, \quad \|\cdot\| = \|\cdot\|_F, \quad f(x) = -\log\det(x) + \langle s, x\rangle, \quad P(x) = \|x\|_1. \tag{20}$$

In the case of $0 < \underline{\lambda} < \bar{\lambda} < \infty$, Lu [74] proposed a reformulation of (10) via Fenchel duality [116, 117], corresponding to

$$f(x) = \sup_{\underline{\lambda}I \preceq y \preceq \bar{\lambda}I} \langle x, y\rangle - \log\det(y), \quad P(x) = \begin{cases} 0 & \text{if } \|x - s\|_\infty \le \tau, \\ \infty & \text{else,} \end{cases} \tag{21}$$

where $\|x\|_\infty = \max_{i,j} |x_{ij}|$. (Note that P in (21) depends on τ.) Importantly, for (17), (18), (20), (21), P is block-separable, i.e.,

$$P(x) = \sum_{J\in\mathcal{J}} P_J(x_J), \tag{22}$$

where $\mathcal{J}$ is some partition of the coordinate indices of x. Then (15) is amenable to solution by (block) coordinatewise methods. Another special case of interest is variable selection in logistic regression [88, 126, 157, 159], corresponding to

$$\mathcal{E} = \Re^n, \quad \|\cdot\| = \|\cdot\|_2, \quad f(x) = \sum_{i=1}^m \log\left(1 + e^{A_i x}\right) - b_i A_i x, \quad P(x) = \|x\|_1, \tag{23}$$

with A_i the ith row of $A \in \Re^{m\times n}$ and $b_i \in \{0,1\}$; also see [55, 123, 124, 125] for variants. A closely related problem is group variable selection ("group lasso"), which uses instead

$$P(x) = \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2, \tag{24}$$

with $\omega_J \geq 0$, in (17) and (23) [88, 157]. The TV image reconstruction model in [147, Eq. (1.3)] and the TV-L_1 image deblurring model in [152, Eq. (1.5)] are special cases of (15) with f quadratic and P of the form (24).

How accurate approximations are the convex relaxations (7), (8), (10), (12), (14), and other related convex relaxations such as that for max-min dispersion, allowing for noisy data? What are the iteration complexities and asymptotic convergence rates of first-order gradient methods for solving these and related problems? First-order methods are attractive since they can exploit sparse or partial separable structure of f and block-separable structure of P.

Throughout, $\Re^n$ denotes the space of n-dimensional real column vectors, $\mathcal{S}^n = \{x \in \Re^{n\times n} \mid x = x^T\}$, and ^{T} denotes transpose. For any $x \in \Re^n$ and nonempty $J \subseteq \{1,...,n\}$, x_j denotes jth coordinate of x, x_J denotes subvector of x comprising x_j, $j \in J$, and $\|x\|_\rho = \left(\sum_{j=1}^n |x_j|^\rho\right)^{1/\rho}$ for $0 < \rho < \infty$, and $\|x\|_\infty = \max_j |x_j|$. For any $x \in \Re^{p\times n}$, x_{ij} denotes the (i,j)th entry of x.

2 Approximation Accuracy of the Convex Relaxations

Let a_j denote column j of A in (1), normalized so that $\|a_j\|_2 = 1$, for all j. It is easily seen that the relaxation (2) is exact (i.e., its solution also solves (1)) if $a_1, \ldots, a_n$ are pairwise orthogonal. This hints that (2) may remain exact if these columns are approximately orthogonal. Mallat and Zhang [83] introduced the following measure of approximate orthogonality, called "mutual coherence", in their study of matching pursuit:

$$\mu := \max_{1\leq i\neq j\leq n} \left|a_i^T a_j\right|. \tag{25}$$

There exist overcomplete sets with $n \approx m^2$ and $\mu \approx 1/\sqrt{m}$; see [129, pages 265-266] and references therein. This μ is central to the analysis in [40, 41, 42, 56, 57, 59, 62, 83, 133, 134, 143]. In particular, (2) is exact whenever $N^0 < \frac{1}{2}(\mu^{-1}+1) = O(n^{\frac{1}{4}})$, where N^0 denotes the optimal value of (1) [40, Theorem 7], [62, Theorem 1], [56], [133, Theorems A and B]. When b is noisy, it can be shown that the solution x^ρ of the noise-aware model (3) is close to the solution x^0 of the original noiseless problem (1):

$$\|x^\rho - x^0\|_2^2 \leq \frac{(\delta+\rho)^2}{1-\mu(4N^0-1)} \tag{26}$$

whenever $\rho \geq \delta := \|b - b^0\|_2$ and $N^0 < (\frac{1}{2} - O(\mu))\mu^{-1} + 1$; see [41, Theorem 3.1], [143, Theorem 1]. The bound (26) also extends to $\rho < \delta$ with some limitations [143, Theorem 1]. In addition to convex relaxation, greedy methods can also be shown to recover or approximate x^0 and identify the support of x^0 under similar conditions on N^0, δ, and ρ; see [41, Theorem 5.1(a)], [133, Theorems A and B], [143, Theorems 3 and 4].

A different measure of approximate orthogonality, introduced by Candès and Tao [28], is the "restricted isometry" constant μ_N, defined as the smallest scalar satisfying

$$(1-\mu_N)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\mu_N)\|x\|_2^2 \quad \forall x \text{ with } \sharp(x) \le N. \tag{27}$$

In fact, $\mu_2 = \mu$.[4] It is known that $\|x^\rho - x^0\|_2 = O(\rho)$ whenever $\rho \ge \delta$ and $\mu_{3N^0} + 3\mu_{4N^0} < 2$ [27, Theorem 1]; also see [27, Theorem 2] for a relaxation of the latter condition. For A randomly generated from certain classes of distributions (e.g., Gaussian), this condition holds with high probability (for N^0 in the order of n to within log factors); see [26, 28, 39, 45, 111, 135] and references therein for the noiseless case and [27, 38] for the noisy case. The approximation bounds in [27, 38] for (3) require $\rho \ge \delta$, as well as $n = O(m)$ in [38]. Is extension to $\rho < \delta$ possible, as in [143, Theorem 1]?

Can the aforementioned exactness results and error bounds be extended to matrix rank minimization (6) and its convex relaxations (7) and (8)? Recent progress has been made in the noiseless case [25, 115]. The nuclear norm has a more complex structure than the 1-norm.

For the sensor network localization problem (11), its SDP relaxation (12) is exact if the distances d_{ij}, $(i,j) \in \mathcal{A}$, are exact (i.e., $d_{ij} = d_{ij}^0$ for all $(i,j) \in \mathcal{A}$) and any relative-interior solution (i.e., a point in the relative interior of the solution set) of (12) has rank d [127, Theorem 2]. However, this assumption is quite strong. What can we say in general? Remarkably, a kind of partial exactness still holds. Biswas and Ye [20, Section 4] introduced the notion of individual traces for a feasible solution x of (12), defined as

$$\mathrm{tr}_i(x) := y_{ii} - \|z_i\|^2, \quad i = 1, \ldots, m,$$

or, equivalently, the diagonals of the Schur complement $y - z^T z$. It can be shown that, for any relative-interior solution x of (12), $\mathrm{tr}_i(x) = 0$ implies z_i is invariant over the solution set of (12) and hence equals z_i^0, the true position of sensor i, when distances are exact [141, Proposition 4.1]. An analogous result holds for the ESDP relaxation (14) [148, Theorem 2]. Thus, upon finding a relative-interior solution x of (12) or (14), we know that every sensor i with $\mathrm{tr}_i(x) = 0$ (within numerical accuracy) has z_i as its true position. Is this result sharp? Yes, at least when the distances are exact. In this case, the converse holds for the ESDP relaxation; see [113, Theorem 1]. It is not known if the converse holds for the SDP relaxation. On the other hand, an example in [113, Example 2] shows that, in the noisy case where the distances are inexact, $\mathrm{tr}_i(x) = 0$ is not a reliable indicator of sensor i position accuracy even when x is the unique solution of the SDP/ESDP relaxation. The reason is that the solution set of the SDP/ESDP relaxation can change abruptly under arbitrarily small data perturbation. This contrasts with a second-order cone (SOCP) relaxation of the same problem, whose solution set changes gradually with data perturbation [141, Section 7]. To overcome this difficulty, a noise-aware version of the ESDP relaxation, analogous to (3) for compressed sensing, was proposed in [113, Section 5]. Specifically, for any $\rho = (\rho_{ij})_{(i,j)\in\mathcal{A}} \ge 0$, let

$$\mathcal{S}^\rho_{\text{resdp}} := \left\{ x \mid x \text{ is feasible for (14) and } |\ell_{ij}(x) - d_{ij}^2| \le \rho_{ij} \ \forall (i,j) \in \mathcal{A} \right\}. \tag{28}$$

[4] Why? Since $\|a_j\|_2 = 1$ for all j, (27) with $N = 2$ reduces to $1 - \mu_2 \le 1 + 2a_i^T a_j x_i x_j \le 1 + \mu_2$ for all $i \ne j$ and all $x_i, x_j \in \Re$ with $x_i^2 + x_j^2 = 1$. Since $(x_i, x_j) \mapsto 2x_i x_j$ has minimum value of -1 and maximum value of 1 on the unit sphere, the smallest μ_2 for which this holds is precisely μ given by (25).

Then $\mathcal{S}^{\rho}_{\text{resdp}}$ contains the noiseless ESDP solutions whenever $\rho \geq \delta := (|d_{ij}^2 - (d_{ij}^0)^2|)_{(i,j)\in\mathcal{A}}$. Moreover, defining its "analytic center" as

$$x^{\rho} := \underset{x\in\mathcal{S}^{\rho}_{\text{resdp}}}{\arg\min} - \sum_{(i,j)\in\mathcal{A}^s} \log\det\left(\begin{pmatrix} y_{ii} & y_{ij} & z_i^T \\ y_{ij} & y_{jj} & z_j^T \\ z_i & z_j & I \end{pmatrix}\right) - \sum_{i=1}^{m} \log \text{tr}_i(x), \tag{29}$$

it can be shown that, for $\rho > \delta$ sufficiently small, $\text{tr}_i(x^{\rho}) = 0$ is a reliable indicator of sensor i position accuracy; see [113, Theorem 4]. Moreover, for any $\rho > \delta$, we have the following computable bound on the *individual* sensor position accuracy:

$$\|z_i^{\rho} - z_i^0\| \leq \sqrt{2|\mathcal{A}^s| + m} \cdot \text{tr}_i(x^{\rho})^{\frac{1}{2}} \quad \forall i.$$

Can these results be extended to the SDP relaxation (12) or SOS relaxations [66, 104] or to handle additional constraints such as bounds on distances and angles-of-arrival?

Closely related to sensor network localization is the continuous *max-min dispersion* problem, whereby, given existing points $z_{m+1}^0, \ldots, z_n^0$ in $\Re^d$ ($d \geq 1$), we wish to locate new points $z_1, \ldots, z_m$ inside, say, a box $[0,1]^d$ that are furthest from each other and existing points [33, 149]:

$$\max_{z_1,\ldots,z_m\in[0,1]^d} \min\left\{\min_{i<j\leq m} \omega_{ij}\|z_i - z_j\|_2, \min_{i\leq m<j} \omega_{ij}\|z_i - z_j^0\|_2\right\}, \tag{30}$$

with $\omega_{ij} > 0$. Replacing "max" by "sum" yields the max-avg dispersion problem [114, Section 4]. It can be shown (by reduction from 0/1-integer program feasibility) that (30) is NP-hard if d is a part of the input, even when $m = 1$. How accurate approximations are their convex (e.g., SDP, SOCP) relaxations? Another related problem arises in protein structure prediction, whereby the distances between neighboring atoms and their bond angles are known, and we wish to find positions of the atoms that minimize a certain energy function [119]. Although the energy function is complicated and highly nonlinear, one can focus on the most nonlinear terms, such as the Lennard-Jones interactions, in seeking approximate solutions.

3 Gradient Methods for Solving the Convex Relaxations

How to solve (15)? We will assume that ∇f is Lipschitz continuous on a closed convex set $\mathcal{X} \supseteq \text{dom}P$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in \mathcal{X}, \tag{31}$$

for some $L > 0$, where $\mathcal{E}^*$ is the vector space of continuous linear functionals on $\mathcal{E}$, endowed with the dual norm $\|x^*\|_* = \sup_{\|x\|\leq 1}\langle x^*, x\rangle$ and $\langle x^*, x\rangle$ is the value of $x^* \in \mathcal{E}^*$ at $x \in \mathcal{E}$. This assumption, which is satisfied by (17), (18), (19), (21), (23), can be relaxed to hold for (20) as well. Owing to its size and structure, (15) is suited for solution by first-order gradient methods, whereby at each iteration f is approximated by a linear function plus a "simple" proximal term. We describe such methods below. To simplify notation, we denote the linearization of f in F at $y \in \mathcal{X}$ by

$$\ell_F(x; y) := f(y) + \langle \nabla f(y), x - y\rangle + \tau P(x) \quad \forall x. \tag{32}$$

3.1 Proximal Gradient Methods

Choose a strictly convex function $\eta : \mathcal{E} \to (-\infty, \infty]$ that is differentiable on an open set containing $\mathcal{X}$,[5] Then the function

$$D(x, y) := \eta(x) - \eta(y) - \langle \nabla\eta(y), x - y\rangle \quad \forall y \in \mathcal{X},\ \forall x \in \mathcal{E},$$

is nonnegative and zero if and only if $x = y$, so D acts as a proximity measure. This function was studied by Bregman [23] and many others; see [8, 10, 32, 46, 67, 130] and references therein. By scaling η if necessary, we assume that

$$D(x, y) \geq \frac{1}{2}\|x - y\|^2 \qquad \forall\ x, y \in \mathcal{X}. \tag{33}$$

The classical gradient projection method of Goldstein and Levitin, Polyak (see [15, 112]) naturally generalizes to solve (15) using the Bregman function D, with constant stepsize $1/L$:

$$x^{k+1} = \arg\min_x \left\{ \ell_F(x; x^k) + \frac{L}{2} D(x, x^k) \right\}, \quad k = 0, 1, \ldots, \tag{34}$$

where $x^0 \in \mathrm{dom} P$. This method, which we call the *proximal gradient* (PG) method, is closely related to the mirror descent method of Nemirovski and Yudin [93], as is discussed in [8, 11]; also see [89] for the case of $\eta(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. When P is the 1-norm or has the block-separable form (22), the new point x^{k+1} can be found in closed form, which is a key advantage of this method for large-scale optimization; see [144, 151] and references therein. When P is given by (16) and $\mathcal{X}$ is the unit simplex, x^{k+1} can be found in closed form in $O(n)$ floating point operations (flops) by taking $\eta(x)$ to be the $x \log x$-entropy function [11, Section 5], [98, Lemma 4]. Moreover, the corresponding D satisfies (33) with $\|\cdot\|$ being the 1-norm [11, Proposition 5.1], [98, Lemma 3]. If $\eta(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ is used instead, then x^{k+1} can still be found in $O(n)$ flops, but this requires using a more complicated algorithm; see [69] and references therein. It can be shown that

$$F(x^k) - \inf F \leq O\left(\frac{L}{k}\right) \quad \forall k,$$

and hence $O(\frac{L}{\epsilon})$ iterations suffice to come within $\epsilon > 0$ of $\inf F$; see, e.g., [13, Theorem 3.1], [97, Theorem 2.1.14], [112, page 166], [146, Theorem 5.1].

In a series of work [94, 95, 98] (also see [112, page 171]), Nesterov proposed three methods for solving the smooth constrained case (16) that, at each iteration, use either one or two projection steps together with extrapolation to accelerate convergence. These accelerated gradient projection methods generate points $\{x^k\}$ that achieve

$$F(x^k) - \inf F \leq O\left(\frac{L}{k^2}\right) \quad \forall k,$$

so that $O(\sqrt{\frac{L}{\epsilon}})$ iterations suffice to come within $\epsilon > 0$ of $\inf F$. In [98], it is shown that various large convex-concave optimization problems can be efficiently solved by applying these methods to a smooth approximation with Lipschitz constant $L = O(1/\epsilon)$. These methods have inspired

[5] This assumption can be relaxed to η being differentiable on the interior of $\mathcal{X}$ only.

various extensions and variants [8, Section 5], [13, 61, 71], [97, Section 2.2], [98, 101, 142], as well as applications to compressed sensing, sparse covariance selection, matrix completion, etc. [14, 5, 64, 74, 75, 82, 96, 132]. In particular, all three methods can be extended to solve (15) in a unified way and achieve $O(\sqrt{\frac{L}{\epsilon}})$ iteration complexity; see [142] and discussion below. The work per iteration is between $O(n)$ and $O(n^3)$ flops for the applications of Section 1. In contrast, the number of iterations for interior point methods is at best $O(\sqrt{n}\log\left(\frac{1}{\epsilon}\right))$ and the work per iteration is typically between $O(n^3)$ and $O(n^4)$ ops. Thus, for moderate ϵ (say, $\epsilon = .001$), moderate L (which may depend on n), and large n ($n \geq 10000$), a proximal gradient method can outperform interior point methods.

The first accelerated proximal gradient (APG) method for solving (15) is the simplest, but requires $\mathcal{E}$ to be a Hilbert space (i.e., $\mathcal{E}^* = \mathcal{E}$, $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$) and $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$, so that $D(x,y) = \frac{1}{2}\|x-y\|^2$. For any $x^0 = x^{-1} \in \mathrm{dom}P$ and $\theta_0 = \theta_{-1} = 1$, it generates (for $k = 0, 1, \ldots$)

$$\begin{aligned} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), && (35)\\ x^{k+1} &= \arg\min_x \left\{ \ell_F(x; y^k) + \frac{L}{2}\|x - y^k\|^2 \right\}, && (36)\\ \theta_{k+1} &= \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}. && (37) \end{aligned}$$

An inductive argument shows that $\theta_k \leq \frac{2}{k+2}$ for all k. As $k \to \infty$, we have $\frac{\theta_k}{\theta_{k-1}} = \sqrt{1-\theta_k} \to 1$, so that, by (35), y^k is asymptotically an isometric extrapolation from x^{k-1} to x^k. In particular, y^k may lie outside of $\mathrm{dom}P$. However, since $x^k, x^{k-1} \in \mathrm{dom}P$, it is readily seen that $y^k \in \{2x - w \mid x, w \in \mathrm{dom}P\}$ (since $x + \alpha(x - x^{-1}) = 2x - w$ with $w = (1-\alpha)x + \alpha x^{-1}$). This method was recently proposed by Beck and Teboulle [13] as an extension of Nesterov's first method [94]; also see [61] for refinements in the unconstrained case of $P \equiv 0$.

The second APG method imposes no requirement on $\mathcal{E}$ or D, and maintains $y^k \in \mathrm{dom}P$, so it is less restrictive than the first method. For any $x^0, z^0 \in \mathrm{dom}P$ and $\theta_0 = 1$, it generates (for $k = 0, 1, \ldots$)

$$\begin{aligned} y^k &= (1-\theta_k)x^k + \theta_k z^k, && (38)\\ z^{k+1} &= \arg\min_x \left\{ \ell_F(x; y^k) + \theta_k L D(x, z^k) \right\}, && (39)\\ x^{k+1} &= (1-\theta_k)x^k + \theta_k z^{k+1}, && (40) \end{aligned}$$

with θ_{k+1} given by (37). Since $0 < \theta_k \leq 1$, we have from $x^k, z^k \in \mathrm{dom}P$ that $y^k \in \mathrm{dom}P$. In the smooth constrained case (16), this method corresponds to Auslender and Teboulle's extension [8, Section 5] of Nesterov's second method [95]; also see [97, page 90]. A variant proposed by Lan, Lu, and Monteiro [71, Section 3] replaces (40) by a PG step from y^k.

The third APG method differs from the second method mainly in the computation of z^{k+1}. For any $x^0 \in \mathrm{dom}P$ and $z^0 = \arg\min_{x \in \mathrm{dom}P} \eta(x)$, $\theta_0 = 1$, it generates (for $k = 0, 1, \ldots$) y^k by (38),

$$z^{k+1} = \arg\min_x \left\{ \sum_{i=0}^{k} \frac{\ell_F(x; y^i)}{\theta_i} + L\eta(x) \right\}, \quad (41)$$

and x^{k+1}, θ_{k+1} by (40), (37). Thus, (41) replaces $\ell_F(x; y^k)/\theta_k$ in (39) by its cumulative sum and replaces $D(\cdot, z^k)$ by $\eta(\cdot)$. In the case of (16), this method is similar to Nesterov's third method [98] but with only one projection instead of two. In the case of $\mathcal{E}$ being a Hilbert space and $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$, this method bears some resemblance to an accelerated dual gradient method in [101, Section 4].

The preceding accelerated methods may look unintuitive, but they arise "naturally" from refining the analysis of the PG method, as is discussed in Appendix A. Moreover, these methods are *equivalent* (i.e., generate the same sequences) when $P \equiv 0$, $\mathcal{E}$ is a Hilbert space, and the Bregman function D is quadratic (i.e., $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$). Some extensions of these methods, including cutting planes, estimating L, are discussed in [142]. In particular, L can be estimated using backtracking: increase L and repeat the iteration whenever a suitable sufficient descent condition (e.g., (53) or (57)) is violated; see [13, 94, 142]. Below we summarize the iteration complexity of the PG and APG methods.

Theorem 1 **(a)** *Let $\{x^k\}$ be generated by the PG method* (34)*. For any $x \in \mathrm{dom}P$, we have*

$$F(x^k) \le F(x) + \frac{1}{k} L D(x, x^0) \quad \forall k \ge 1.$$

(b) *Assume $\mathcal{E}$ is a Hilbert space, $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$, and $\mathcal{X} \supseteq \{2x - w \mid x, w \in \mathrm{dom}P\}$. Let $\{x^k\}$ be generated by the first APG method* (35)–(37)*. For any $x \in \mathrm{dom}P$, we have*

$$F(x^k) \le F(x) + \theta_{k-1}^2 L D(x, x^0) \quad \forall k \ge 1.$$

(c) *Let $\{x^k\}$ be generated by the second APG method* (37)–(40)*. For any $x \in \mathrm{dom}P$, we have*

$$F(x^k) \le F(x) + \theta_{k-1}^2 L D(x, z^0) \quad \forall k \ge 1.$$

(d) *Let $\{x^k\}$ be generated by the third APG method* (37), (38), (40), (41)*. For any $x \in \mathrm{dom}P$, we have*

$$F(x^k) \le F(x) + \theta_{k-1}^2 L(\eta(x) - \eta(z^0)) \quad \forall k \ge 1.$$

A proof of Theorem 1(a)-(c) is given in Appendix A. A proof of part (d) can be found in [142, Corollary 3(a)]. Taking any x satisfying $F(x) \le \inf F + \frac{\epsilon}{2}$ in Theorem 1 yields $F(x^k) \le \inf F + \epsilon$ after $k = O(\frac{L}{\epsilon})$ iterations for the PG method and after $k = O(\sqrt{\frac{L}{\epsilon}})$ iterations for the APG methods.

How can we terminate the PG and APG methods in practice with a guaranteed optimality gap? The bounds in Theorem 1 requires estimating the distance to an $\frac{\epsilon}{2}$-minimizer of F and are rather conservative. In the case where f has the form

$$f(x) = \max_{v \in V} \phi(x, v),$$

for some saddle function ϕ and convex set V in a suitable space, duality gap can be used to terminate the methods [92, 98, 99]. The dual problem is $\max_v Q(v)$, with dual function

$$Q(v) := \min_x \{\phi(x, v) + \tau P(x)\}.$$

Then we compute (say, every 5 or 10 iterations) a candidate dual solution

$$v^k \;=\; \arg\max_v \phi(x^k, v),$$

and check that $F(x^k) - Q(v^k) \leq \epsilon$. In fact, assuming furthermore that $\mathrm{dom}P$ is bounded, it can be shown using an idea of Nesterov [98, Theorem 3] that

$$0 \;\leq\; F(x^{k+1}) - Q(\bar{v}^k) \;\leq\; \theta_k^2 L \max_{x \in \mathrm{dom}P} (\eta(x) - \eta(z^0)) \quad \forall k \geq 0,$$

where x^{k+1}, y^k, θ_k are generated by the third APG method, and we let

$$v^k \;=\; \arg\max_v \phi(y^k, v), \qquad \bar{v}^k \;=\; (1-\theta_k)\bar{v}^{k-1} + \theta_k v^k.$$

with $\bar{v}^{-1} = 0$ [142, Corollary 3(c)]; also see [74, Theorem 2.2] and [99] for related results in the constrained case (16). Analogous bounds hold for the first two APG methods; see [142, Corollaries 1(b) and 2].

When applied to (4), the first APG method yields an accelerated version of the iterative thresholding method of Daubechie et al. [13, 34]. What about the other two methods? How efficiently can these methods be applied to solve (5), (7), (8), (10), (23), and related problems such as those in [48, 123, 124, 125, 153]? When applied to (7) and (8), singular value decomposition is needed at each iteration, and efficiency depends on the cost for this decomposition. However, only the largest singular values and their associated singular vectors are needed [24]. Can these be efficiently computed or updated? Some progress on this have recently been made [82, 132].

Can the iteration complexity be further improved? The proofs suggest that the convergence rate can be improved to $O(\frac{1}{k^p})$ $(p > 2)$ if we can replace $\|\cdot\|^2$ in the proofs by $\|\cdot\|^p$; see Appendix A. However, this may require using a higher-order approximation of f in $\ell_F(\cdot; y)$, so the improvement would not come "for free".

3.2 Block-Coordinate Gradient Methods

When $\mathcal{E}$ is a Hilbert space and P is block-separable (22), we can apply (34) block-coordinatewise, possibly with L and D dynamically adjusted, resulting in a *block-coordinate gradient* (BCG) method. More precisely, given $x^k \in \mathrm{dom}P$, we choose J_k as the union of some subcollection of indices in $\mathcal{J}$ and choose a self-adjoint positive definite linear mapping $H^k : \mathcal{E} \to \mathcal{E}$ (ideally $H^k \approx \nabla^2 f(x^k)$), compute

$$d^k \;=\; \arg\min_d \left\{ \ell_F(x^k + d; x^k) + \frac{1}{2}\langle H^k d, d\rangle \;\mid d_j = 0 \;\forall j \notin J_k \right\}, \tag{42}$$

and update

$$x^{k+1} = x^k + \alpha_k d^k, \tag{43}$$

with stepsize $\alpha_k > 0$ [144, 146]. This method may be viewed as a coordinate/SOR version of a sequential quadratic programming (SQP) method, and it is related to the variable/gradient

distribution methods for unconstrained smooth optimization [51, 58, 86, 120] and (block) coordinate minimization methods [15, 84, 87, 105, 121, 136, 140]. In the case of $H^k d = Ld$ and J_k comprising all coordinate indices of x, (42)–(43) with $\alpha_k = 1$ reduces to the PG method (34) with $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$.

How to choose α_k and J_k? Various stepsize rules for smooth optimization [15, 52, 105] can be adapted to this nonsmooth setting. One that works well in theory and practice is an Armijo-type rule adapted from SQP methods:

$$\alpha_k \;=\; \max\left\{\alpha \in \{1, \beta, \beta^2, \ldots\} \mid F(x^k + \alpha d^k) \leq F(x^k) + \alpha\sigma\Delta_k\right\}, \tag{44}$$

where $0 < \beta, \sigma < 1$ and Δ_k is the difference between the optimal value of (42) and $F(x^k)$. This rule requires only function evaluations, and Δ_k predicts the descent from x^k along d^k. For global convergence, the index subset J_k is chosen either in a *Gauss-Seidel* manner, i.e., $J_k \cup \cdots \cup J_{k+K}$ covers all subsets of $\mathcal{J}$ for some constant $K \geq 0$ [31, 88, 107, 140, 144] or J_k is chosen in a *Gauss-Southwell* manner to satisfy

$$\Delta_k \leq \omega\Delta_k^{\text{all}},$$

where $0 < \omega < 1$ and Δ_k^{all} denotes the analog of Δ_k when J_k is replaced by the entire coordinate index set [144, 145, 146]. Moreover, assuming (31) with $\mathcal{X} = \text{dom}P$, the BCG method using the Gauss-Southwell choice of J_k finds an ϵ-minimizer of F in $O(\frac{L}{\epsilon})$ iterations [146, Theorem 5.1].

The above BCG method, which has been successful for compressed sensing and variable selection in regression [88, 126, 159] and can be extended to handle linear constraints as in support vector machine (SVM) training [145], may also be suited for solving (5), (10), (14), (21), and related problems. When applied to (10) with J_k indexing a row/column of x, d^k and α_k are computable in $O(n^2)$ flops using Schur complement and properties of determinant. This method may be similarly applied to Lu's reformulation (21). This contrasts with the block-coordinate minimization method in [9] which uses $O(n^3)$ flops per iteration. This method can also be applied to solve (29) by using a smooth convex penalty $\frac{1}{2\theta}\max\{0, |\cdot| - \rho_{ij}\}^2$ ($\theta > 0$) for each constraint in (28) of the form $|\cdot| \leq \rho_{ij}$. By choosing each coordinate block to comprise z_i, y_{ii}, and $(y_{ij})_{j|(i,j)\in\mathcal{A}}$, the computation distributes, with sensor i needing to communicate only with its neighbors when updating its position – an important consideration for practical implementation. The resulting method can accurately position up to 9000 sensors in little over a minute; see [113, Section 7.3]. Can Nesterov's extrapolation techniques of Section 3.1 be adapted to the BCG method?

3.3 Incremental Gradient Methods

A problem that frequently arises in machine learning and neural network training has the form (15) with

$$f(x) = \sum_{i=1}^{m} f_i(x), \tag{45}$$

where each $f_i : \mathcal{E} \to (-\infty, \infty]$ is smooth, convex on an open subset of $\mathcal{E}$ containing dom$P$. (The convexity assumption can be relaxed depending on the context.) For example, $f_i(x)$ may be the output error for an input-out system (e.g., a classifier), parameterized by x, on the ith training

example. P given by (16) would confer constraints on x and $P(\cdot) = \|\cdot\|_1$ would induce a sparse representation. m may be large. A popular approach to minimizing f of the form (45) is by an incremental gradient method ("on-line back-propagation"):

$$x^{k+1} = x^k - \alpha_k \nabla f_{i_k}(x^k), \tag{46}$$

with i_k chosen cyclically from $\{1, \ldots, m\}$ and stepsize $\alpha_k > 0$ either constant or adjusted dynamically; see [15, 68, 77, 85, 91, 128, 137] and references therein. When some the ∇f_i's are "similar", this incremental method is more efficient than (pure) gradient method since it does not wait for all component gradients $\nabla f_1, ..., \nabla f_m$ to be evaluated before updating x. However, global convergence (i.e., $\nabla f(x^k) \to 0$) generally requires $\alpha_k \to 0$, which slows convergence. Recently, Blatt, Hero and Gauchman [22] proposed an aggregate version of (46) that approximates $\nabla f(x^k)$ by

$$g^k = \sum_{\ell=k-m+1}^{k} \nabla f_{i_\ell}(x^\ell).$$

This method achieves global convergence for any sufficiently small constant stepsize α_k, but requires $O(mn)$ storage. We can reduce the storage to $O(n)$ by updating a cumulative average of the component gradients:

$$g^k_{\text{sum}} = g^{k-1}_{\text{sum}} + \nabla f_{i_k}(x^k), \qquad\qquad g^k = \frac{m}{k} g^k_{\text{sum}},$$

with $g^{-1}_{\text{sum}} = 0$. We then use g^k in the PG, APG, or BCG method. The resulting incremental methods share features with recently proposed averaging gradient methods [65, 102]. What are their convergence properties, iteration complexities, and practical performances?

4 Error Bound and Linear Convergence of Gradient Methods

Analogous to superlinear convergence for second-order methods, linear convergence is a good indicator of efficiency for first-order methods. Key to a linear convergence analysis is the following Lipschizian error bound on $\text{dist}(x, \bar{\mathcal{X}}) := \min_{s \in \bar{\mathcal{X}}} \|x - s\|$ in terms of the norm of the residual

$$R(x) \;:=\; \arg\min_d \left\{ \ell_F(x + d; x) + \frac{1}{2}\|d\|^2 \right\}, \tag{47}$$

where $\bar{\mathcal{X}}$ denotes the set of minimizers of F, which we assume to be nonempty.

EB condition: For any $\zeta \geq \min F$, there exist scalars $\kappa > 0$ and $\epsilon > 0$ such that

$$\text{dist}(x, \bar{\mathcal{X}}) \leq \kappa \|R(x)\| \quad \text{whenever} \qquad F(x) \leq \zeta,\ \|R(x)\| \leq \epsilon. \tag{48}$$

Under the EB condition, asymptotic linear and even superlinear convergence can be established for various methods, including interior point, gradient projection, proximal minimization, coordinate minimization, and coordinate gradient methods – even if $\bar{\mathcal{X}}$ is unbounded; see [12, 49, 80, 81, 139, 144, 145, 146] and references therein. Moreover, the EB condition holds under any of the following conditions; see [144, Theorem 4] as well as [109, Theorem 3.1], [79, Theorem 2.1] for the constrained case (16).

C1. $\mathcal{E} = \Re^n$, $f(x) = h(Ax) + \langle c, x\rangle$ for all $x \in \Re^n$, where $A \in \Re^{m\times n}$, $c \in \Re^n$, and h is a strongly convex differentiable function on $\Re^m$ with ∇h Lipschitz continuous on $\Re^m$. P is polyhedral.

C2. $\mathcal{E} = \Re^n$, $f(x) = \max_{y\in Y}\{\langle y, Ax\rangle - h(y)\} + \langle c, x\rangle$ for all $x \in \Re^n$, where Y is a polyhedral set in $\Re^m$, $A \in \Re^{m\times n}$, $c \in \Re^n$, and h is a strongly convex differentiable function on $\Re^m$ with ∇h Lipschitz continuous on $\Re^m$. P is polyhedral.

C3. f is strongly convex and satisfies (31) for some $L > 0$.

What if f is not strongly convex and P is non-polyhedral? In particular, we are interested in the group lasso for linear and logistic regression (see (17), (23), (24)), for which f is not strongly convex (unless A has full column rank) and P is non-polyhedral (unless J is a singleton for all $J \in \mathcal{J}$). The following new result shows that the error bound (48) holds for the group lasso. The proof, given in Appendix B, exploits the structure of the weighted sum of 2-norms (24). To our knowledge, this is the first Lipschitzian error bound result for f not strongly convex and P non-polyhedral.

Theorem 2 *Suppose that $\mathcal{E} = \Re^n$, P has the form (24) with $\omega_J > 0$ for all $J \in \mathcal{J}$, and f has the form*

$$f(x) = h(Ax), \tag{49}$$

where $A \in \Re^{m\times n}$, and $h : \Re^m \to (-\infty, \infty]$ is differentiable on dom$h$*, which is assumed to be convex and open. Also suppose that (a) h is strongly convex and ∇h is Lipschitz continuous on any compact convex subset of* dom$h$*, and (b) $h(y) \to \infty$ whenever y approaches a boundary point of* dom$h$*. If $\bar{\mathcal{X}} \neq \emptyset$, then*

$$\{x \mid F(x) \leq \zeta\} \text{ is bounded} \quad \forall \zeta \in \Re, \tag{50}$$

and the EB condition (48) holds.

The assumptions on h in Theorem 2 are satisfied by

$$h(y) = \frac{1}{2}\|y - b\|^2,$$

corresponding to linear regression, and

$$h(y) = \sum_{i=1}^{m} \log(1 + e^{y_i}) - \langle b, y\rangle \quad \textit{with} \quad b \in \{0, 1\}^m,$$

corresponding to logistic regression (23). The assumptions are also satisfied by

$$h(y) = -\sum_{i=1}^{m} \log(y_i) + \langle b, y\rangle \quad \textit{with} \quad b \geq 0,$$

which arises in likelihood estimation under Poisson noise [123]. In the first two examples, h is bounded from below by zero. In the third example, h is unbounded from below but tends to $-\infty$ sublinearly. Since P given by (24) is homogeneous of degree 1, it is readily seen that $\bar{\mathcal{X}} \neq \emptyset$

for all three examples. (An example of $\bar{\mathcal{X}} = \emptyset$ is $\min_x e^x + 2x + |x|$.) However, $\bar{\mathcal{X}}$ need not be a singleton. An example is

$$\min_x \frac{1}{2}|x_1 + x_2 - 2|^2 + |x_1| + |x_2|,$$

for which $\bar{\mathcal{X}} = \{(1-t, t) \mid t \in [0,1]\}$. Can Theorem 2 be extended to f satisfying C2 and P given by (24) or to (18) or (19)? Can the constant κ in (48), which determines the convergence ratio, be estimated for compressed sensing (17) in terms of the restricted isometry constant μ_N?

Corollary 1 *Under the assumptions of Theorem 2, let $\{(x^k, H^k, J_k, \alpha_k)\}$ be generated by the BCG method (42)–(43) with (i) $\underline{\lambda} I \preceq H^k \preceq \bar{\lambda} I$ for all k ($0 < \underline{\lambda} \leq \bar{\lambda}$), (ii) $\{J_k\}$ cycling through $J \in \mathcal{J}$, and (iii) $\{\alpha^k\}$ chosen by the Armijo rule (44). If $\bar{\mathcal{X}} \neq \emptyset$, then $\{F(x^k)\}_{k\in\mathcal{T}}$ converges at least Q-linearly and $\{x^k\}_{k\in\mathcal{T}}$ converges at least R-linearly, where $\mathcal{T} = \{0, K, 2K, \ldots\}$ and K is the cardinality of $\mathcal{J}$.*

Proof. Theorem 2 shows that [144, Assumption 2(a)] is satisfied. By [144, Theorem 1(a)], $\{F(x^k)\} \downarrow$, so, by (50), $\{x^k\}$ is bounded. Since F is convex, [144, Assumption 2(b)] is automatically satisfied. Also, $\{x^k\}$ is lies in a compact convex subset of $\mathrm{dom} f$, over which ∇f is Lipschitz continuous (since ∇h is Lipschitz continuous over the the image of this subset under A). Conditions (i)–(iii) imply that the remaining assumptions in [144, Theorem 2(b)] are satisfied. In particular, the restricted Gauss-Seidel rule in [144] holds with the given $\mathcal{T}$. Since $\{x^k\}$ is bounded, [144, Theorem 2(b)] implies that $\{F(x^k)\}_{k\in\mathcal{T}}$ converges at least Q-linearly and $\{x^k\}_{k\in\mathcal{T}}$ converges at least R-linearly [107]. ■

By Corollary 1, the method in [157], [88, Section 2.2.1] for linear group lasso (corresponding to $H^k = A^T A$) and the method in [88, Section 2.2.2] for logistic group lasso (corresponding to $H^k = \nu_k I$ with $\nu_k > 0$) attain linear convergence. Linear convergence of the block-coordinate minimization methods used in [147] for TV-regularized image reconstruction, with $f(w,u) = \frac{1}{2}\|w - Bu\|_2^2 + \frac{\gamma}{2}\|Au - b\|_2^2$, $P(w,u) = \sum_J \|w_J\|_2$, and in [152] for TV-regularized image deblurring, with $f(w,u,z) = \frac{1}{2}\|w - Bu\|_2^2 + \frac{\gamma}{2}\|Au - b - z\|_2^2$, $P(w,u,z) = \sum_J \omega_J \|w_J\|_2 + \|z\|_1$, can be analyzed similarly using a generalization of Theorem 2 to allow $\omega_J = 0$ for some J. Can the linear convergence analysis be extended to the APG methods or their variants?

5 Appendix A: Analyzing the PG and APG Methods

Since f is convex and (31) holds with $\mathcal{X} \supseteq \mathrm{dom} P$, we have from (32) that

$$F(x) \;\geq\; \ell_F(x; y) \;\geq\; F(x) - \frac{L}{2}\|x - y\|^2 \quad \forall x \in \mathrm{dom} P,\ y \in \mathcal{X}. \tag{51}$$

The following "3-point" property is also key; see [32, Lemma 3.2], [71, Lemma 6], [142, Section 2].

3-Point Property: For any proper lsc convex function $\psi : \mathcal{E} \to (-\infty, \infty]$ and any $z \in \mathcal{X}$, if η is differentiable at $z_+ = \arg\min_x \{\psi(x) + D(x, z)\}$, then

$$\psi(x) + D(x, z) \geq \psi(z_+) + D(z_+, z) + D(x, z_+) \quad \forall x \in \mathrm{dom} P.$$

The APG methods can be motivated by analyzing the PG method (34). Let $\{x^k\}$ be generated by the PG method. For any $x \in \mathrm{dom}P$ and any $k \in \{0, 1, \ldots\}$, we have

$$\begin{aligned} F(x^{k+1}) &\leq \ell_F(x^{k+1}; x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &\leq \ell_F(x^{k+1}; x^k) + LD(x^{k+1}, x^k) \\ &\leq \ell_F(x; x^k) + LD(x, x^k) - LD(x, x^{k+1}) \\ &\leq F(x) + LD(x, x^k) - LD(x, x^{k+1}) \qquad \forall x \in \mathrm{dom}P, \end{aligned} \tag{52}$$

where the first and fourth inequalities use (51), the second inequality uses (33), and the third inequality uses (34) and the 3-Point Property with $\psi(x) = \ell_F(x; x^k)/L$. Letting $e_k = F(x^k) - F(x)$ and $\Delta_k = LD(x, x^k)$, this simplifies to the recursion

$$e_{k+1} \leq \Delta_k - \Delta_{k+1} \quad \forall k \geq 0.$$

It follows from $e_{k+1} \leq e_k$ that $(k+1)e_{k+1} \leq \Delta_0 - \Delta_{k+1} \leq \Delta_0$, which proves Theorem 1(a).

The above proof suggests that, for faster convergence, we should find a similar recursion as (52) but with L scaled by something tending to zero with k. To do this, we need to modify (34). Suppose for simplicity $\mathcal{E}$ is a Hilbert space and $\eta(\cdot) = \frac{1}{2}\|\cdot\|^2$ (so that $D(x, y) = \frac{1}{2}\|x - y\|^2$). One such modification is to replace x^k in (34) by some $y^k \in \mathcal{X}$ to be determined, yielding (36). Then, as in the above proof for the PG method, we have

$$\begin{aligned} F(x^{k+1}) &\leq \ell_F(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 \qquad (53) \\ &\leq \ell_F(y; x^k) + \frac{L}{2}\|y - y^k\|^2 - \frac{L}{2}\|y - x^{k+1}\|^2 \\ &\leq F(y) + \frac{L}{2}\|y - y^k\|^2 - \frac{L}{2}\|y - x^{k+1}\|^2 \qquad \forall y \in \mathrm{dom}P, \qquad (54) \end{aligned}$$

where the first and third inequalities use (51), and the second inequality uses (36) and the 3-Point Property. To get L to be scaled by something tending to zero, set $y = (1 - \theta_k)x^k + \theta_k x$ in the above inequality, with $x \in \mathrm{dom}P$ arbitrary and $0 < \theta_k \leq 1$ to be determined. We can then factor θ_k out of x to scale L, yielding

$$\begin{aligned} & F(x^{k+1}) \\ \leq\ & F((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - y^k\|^2 - \frac{L}{2}\|(1-\theta_k)x^k + \theta_k x - x^{k+1}\|^2 \\ =\ & F((1-\theta_k)x^k + \theta_k x) + \frac{L}{2}\theta_k^2\|x + (\theta_k^{-1} - 1)x^k - \theta_k^{-1}y^k\|^2 - \frac{L}{2}\theta_k^2\|x + (\theta_k^{-1} - 1)x^k - \theta_k^{-1}x^{k+1}\|^2, \end{aligned}$$

where we have rearranged the terms to look like the recursion (52). We want the two terms inside $\|\ \|^2$ to have the form "$x - z^k$" and "$x - z^{k+1}$", which we get by setting

$$z^k = -(\theta_k^{-1} - 1)x^k + \theta_k^{-1}y^k$$

and y^k by (35). Using also the convexity of F, we then obtain that

$$F(x^{k+1}) \ \leq\ (1-\theta_k)F(x^k) + \theta_k F(x) + \theta_k^2\frac{L}{2}\|x - z^k\|^2 - \theta_k^2\frac{L}{2}\|x - z^{k+1}\|^2 \quad \forall k. \tag{55}$$

Letting $e_k = F(x^k) - F(x)$ and $\Delta_k = \frac{L}{2}\|x - z^k\|^2$, this simplifies to

$$e_{k+1} \leq (1-\theta_k)e_k + \theta_k^2\Delta_k - \theta_k^2\Delta_{k+1}.$$

Upon dividing both sides by θ_k^2, we see that, by choosing θ_{k+1} so that $\frac{1}{\theta_k^2} = \frac{1-\theta_{k+1}}{\theta_{k+1}^2}$ (which, upon solving for θ_{k+1}, yields (37)), this rewrites as the recursion

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2}e_{k+1} + \Delta_{k+1} \leq \frac{1-\theta_k}{\theta_k^2}e_k + \Delta_k,$$

which propagates backwards to yield

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2}e_{k+1} + \Delta_{k+1} \;\leq\; \frac{1-\theta_0}{\theta_0^2}e_0 + \Delta_0 \quad \forall k.$$

Since $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}$, setting $\theta_0 = 1$ simplifies this to $\frac{1}{\theta_k^2}e_{k+1} \leq \Delta_0 - \Delta_{k+1} \leq \Delta_0$. Also, we have from (35) and taking $\theta_{-1} = 1$ that $z^0 = y^0 = x^0$. This proves Theorem 1(b).

The preceding proof depends crucially on rearranging terms inside $\|\ \|^2$, so it cannot be directly extended to other Bregman functions D. However, (55) suggests we seek a recursion of the form

$$F(x^{k+1}) \;\leq\; (1-\theta_k)F(x^k) + \theta_k F(x) + \theta_k^2 LD(x, z^k) - \theta_k^2 LD(x, z^{k+1}) \quad \forall k, \tag{56}$$

which, as our derivation of (52) and (55) suggests, can be achieved by setting z^{k+1} by (39) and using the 3-Point Property, setting x^{k+1} by (40) (analogous to our setting of y in (54)), and then choosing y^k so that $x^{k+1} - y^k = \theta_k(z^{k+1} - z^k)$ (which works out to be (38)). Specifically, for any $k \in \{0, 1, \ldots\}$, we have

$$\begin{aligned}
F(x^{k+1}) &\leq \ell_F(x^{k+1}; y^k) + \frac{L}{2}\|x^{k+1} - y^k\|^2 \\
&= \ell_F((1-\theta_k)x^k + \theta_k z^{k+1}; y^k) + \frac{L\theta_k^2}{2}\|z^{k+1} - z^k\|^2 \\
&\leq (1-\theta_k)\ell_F(x^k; y^k) + \theta_k \ell_F(z^{k+1}; y^k) + \theta_k^2 LD(z^{k+1}, z^k) \qquad (57) \\
&\leq (1-\theta_k)\ell_F(x^k; y^k) + \theta_k\left(\ell_F(x; y^k) + \theta_k LD(x, z^k) - \theta_k LD(x, z^{k+1})\right) \quad \forall x \in \mathrm{dom} h,
\end{aligned}$$

where the first inequality uses (51), the second inequality uses the convexity of $\ell_F(\cdot; y^k)$ and (33), the last inequality uses the 3-Point Property with $\psi(x) = \ell_F(x; y^k)/(\theta_k L)$. Using (51) yields (56), from which Theorem 1(c) follows.

The third APG method (see (41)) can be derived and analyzed similarly using an analog of the 3-Point Property for η; see [142, Property 2]. For brevity we omit the details.

6 Appendix B: A Proof of Theorem 2

Througout we assume that the assumptions of Theorem 2 are satisfied. By scaling f by $1/\tau$, we will assume without loss of generality that $\tau = 1$.

Lemma 1 *For any $x \in \Re^n$, letting $g = \nabla f(x)$, we have for all $J \in \mathcal{J}$ that*

$$R(x)_J = \begin{cases} -x_J & \text{if } \|g_J - x_J\|_2 \leq \omega_J, \\ -\left(1 - \frac{\omega_J}{\|g_J - x_J\|_2}\right) g_J - \left(\frac{\omega_J}{\|g_J - x_J\|_2}\right) x_J & \text{if } \|g_J - x_J\|_2 \geq \omega_J. \end{cases} \tag{58}$$

Moreover, R is continuous on $\mathrm{dom} f$.

Proof. Let $r = R(x)$. Then (24), (32), (47), and $\tau = 1$ yield

$$r \;=\; \arg\min_d \left\{ \langle g, d\rangle + \frac{1}{2}\|d\|^2 + \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2 \right\},$$

whose necessary and sufficient optimality condition is

$$0 \in g_J + r_J + \omega_J \partial \|x_J + r_J\|_2 \quad \forall J \in \mathcal{J}.$$

Fix any $J \in \mathcal{J}$. Since $\partial\|0\|_2$ is the unit 2-norm ball and $\partial\|x_J\|_2 = \{x_J/\|x_J\|_2\}$ if $x_J \neq 0$, we have that $r_J = -x_J$ if $\|g_J - x_J\|_2 \leq \omega_J$ and otherwise

$$g_J + r_J + \omega_J \frac{x_J + r_J}{\|x_J + r_J\|_2} = 0. \tag{59}$$

Letting $\alpha = \|x_J + r_J\|_2$, we solve for r_J to obtain

$$r_J = -\frac{\alpha g_J + \omega_J x_J}{\alpha + \omega_J}.$$

Hence $x_J + r_J = \frac{\alpha}{\alpha+\omega_J}(x_J - g_J)$ so that

$$\alpha = \|x_J + r_J\|_2 = \frac{\alpha}{\alpha + \omega_J}\|x_J - g_J\|_2.$$

Solving for α yields $\alpha = \|x_J - g_J\|_2 - \omega_J$, which when plugged into the above formula for r_J yields (58).

The continuity of R follows from the continuity of ∇f and the continuity of the right-hand side of (58) in g. In particular, the two formulas in (58) yield $-x_J$ at the boundary $\|g_J - x_J\|_2 = \omega_J$. ∎

Since h is strictly convex, $x \mapsto Ax$ is invariant over $\bar{\mathcal{X}}$, i.e., there exists $\bar{y} \in \mathrm{dom} h$ such that

$$Ax = \bar{y} \quad \forall x \in \bar{\mathcal{X}}. \tag{60}$$

Since P is given by (24), it follows from (49) and (60) that

$$\bar{\mathcal{X}} = \left\{ x \mid \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_2 = \min F - h(\bar{y}),\ Ax = \bar{y} \right\},$$

so that $\bar{\mathcal{X}}$ is bounded (since $\omega_J > 0$ for all $J \in \mathcal{J}$), as well as being closed convex. Since F is convex and $\bar{\mathcal{X}}$ is bounded, it follows from [116, Theorem 8.7] that (50) holds. By using these observations and Lemma 1, we prove below the EB condition (48).

We argue by contradiction. Suppose there exists a $\zeta \geq \min F$ such that (48) fails to hold for all $\kappa > 0$ and $\epsilon > 0$. Then there exists a sequence $x^1, x^2, \ldots$ in $\Re^n \setminus \bar{\mathcal{X}}$ satisfying

$$F(x^k) \leq \zeta \ \forall k, \qquad \{r^k\} \to 0, \qquad \left\{\frac{r^k}{\delta_k}\right\} \to 0, \tag{61}$$

where for simplicity we let $r^k := R(x^k)$, $\delta_k := \|x^k - \bar{x}^k\|_2$, and $\bar{x}^k := \arg\min_{s \in \bar{\mathcal{X}}} \|x^k - s\|_2$. Let

$$g^k := \nabla f(x^k) = A^T \nabla h(Ax^k), \qquad \bar{g} := A^T \nabla h(\bar{y}). \tag{62}$$

By (60) and (62), $A\bar{x}^k = \bar{y}$ and $\nabla f(\bar{x}^k) = \bar{g}$ for all k.

By (50) and (61), $\{x^k\}$ is bounded. By further passing to a subsequence if necessary, we can assume that $\{x^k\} \to$ some $\bar{x}$. Since $\{R(x^k)\} = \{r^k\} \to 0$ and R is continuous by Lemma 1, this implies $R(\bar{x}) = 0$, so $\bar{x} \in \bar{\mathcal{X}}$. Hence $\delta_k \leq \|x^k - \bar{x}\|_2 \to 0$ as $k \to \infty$ so that $\{\bar{x}^k\} \to \bar{x}$. Also, by (60) and (62), $\{g^k\} \to \nabla f(\bar{x}) = \bar{g}$. Since $P(x^k) \geq 0$, (49) implies $h(Ax^k) = F(x^k) - P(x^k) \leq F(x^k) \leq \zeta$ for all k. Since $\{Ax^k\}$ is bounded and $h(y) \to \infty$ whenever y approaches a boundary point of dom$h$, this implies that $\{Ax^k\}$ and $\bar{y}$ lie in some compact convex subset Y of the open convex set dom$h$. By our assumption on h, h is strongly convex and ∇h is Lipschitz continuous on Y, so, in particular,

$$\sigma\|y - \bar{y}\|_2^2 \leq \langle \nabla h(y) - \nabla h(\bar{y}), y - \bar{y}\rangle \quad \text{and} \quad \|\nabla h(y) - \nabla h(\bar{y})\|_2 \leq L\|y - \bar{y}\|_2 \quad \forall y \in Y, \tag{63}$$

for some $0 < \sigma \leq L$.

We claim that there exists $\kappa > 0$ such that

$$\|x^k - \bar{x}^k\|_2 \leq \kappa \|Ax^k - \bar{y}\|_2 \quad \forall k, \tag{64}$$

We argue this by contradiction. Suppose this is false. Then, by passing to a subsequence if necessary, we can assume that

$$\left\{\frac{\|Ax^k - \bar{y}\|_2}{\|x^k - \bar{x}^k\|_2}\right\} \to 0.$$

Since $\bar{y} = A\bar{x}^k$, this is equivalent to $\{Au^k\} \to 0$, where we let

$$u^k := \frac{x^k - \bar{x}^k}{\delta_k} \quad \forall k. \tag{65}$$

Then $\|u^k\|_2 = 1$ for all k. By further passing to a subsequence if necessary, we will assume that $\{u^k\} \to$ some $\bar{u}$. Then $A\bar{u} = 0$ and $\|\bar{u}\|_2 = 1$. Moreover,

$$Ax^k = A(\bar{x}^k + \delta_k u^k) = \bar{y} + \delta_k A u^k = \bar{y} + o(\delta_k).$$

Since Ax^k and $\bar{y}$ are in Y, the Lipschitz continuity of ∇h on Y (see (63)) and (62) yield

$$g^k = \bar{g} + o(\delta_k). \tag{66}$$

By further passing to a subsequence if necessary, we can assume that, for each $J \in \mathcal{J}$, either (a) $\|g_J^k - x_J^k\|_2 \leq \omega_J$ for all k or (b) $\|g_J^k - x_J^k\|_2 > \omega_J$ and $\bar{x}_J^k \neq 0$ for all k or (c) $\|g_J^k - x_J^k\|_2 > \omega_J$ and $\bar{x}_J^k = 0$ for all k.

Case (a). In this case, Lemma 1 implies that $r_J^k = -x_J^k$ for all k. Since $\{r^k\} \to 0$ and $\{x^k\} \to \bar{x}$, this implies $\bar{x}_J = 0$. Also, by (65) and (61),

$$u_J^k = \frac{-r_J^k - \bar{x}_J^k}{\delta_k} = \frac{o(\delta_k) - \bar{x}_J^k}{\delta_k}. \tag{67}$$

Thus $\bar{u}_J = -\lim_{k\to\infty} \bar{x}_J^k/\delta_k$. Suppose $\bar{u}_J \neq 0$. Then $\bar{x}_J^k \neq 0$ for all k sufficiently large, so $\bar{x}^k \in \bar{\mathcal{X}}$ and the optimality condition for (15) with $\tau = 1$, (24), and $\nabla f(\bar{x}^k) = \bar{g}$ imply

$$\bar{g}_J + \omega_J \frac{\bar{x}_J^k}{\|\bar{x}_J^k\|_2} = 0, \tag{68}$$

so $\bar{u}_J$ is a positive multiple of $\bar{g}_J$.

Case (b). Since $\bar{x}^k \in \bar{\mathcal{X}}$ and $\bar{x}_J^k \neq 0$, the optimality condition for (15) with $\tau = 1$, (24), and $\nabla f(\bar{x}^k) = \bar{g}$ imply (68) holds for all k. Then Lemma 1 implies

$$\begin{aligned}
-r_J^k &= \left(1 - \frac{\omega_J}{\|g_J^k - x_J^k\|_2}\right) g_J^k + \frac{\omega_J x_J^k}{\|g_J^k - x_J^k\|_2} \\
&= \left(1 - \frac{\omega_J}{\|g_J^k - x_J^k\|_2}\right) (\bar{g}_J + o(\delta_k)) + \frac{\omega_J(\bar{x}_J^k + \delta_k u_J^k)}{\|g_J^k - x_J^k\|_2} \\
&= \bar{g}_J + \frac{\omega_J(\bar{x}_J^k - \bar{g}_J)}{\|g_J^k - x_J^k\|_2} + o(\delta_k) + \frac{\omega_J \delta_k u_J^k}{\|g_J^k - x_J^k\|_2} \\
&= \left(\frac{\omega_J}{\|\bar{g}_J - \bar{x}_J^k\|_2} - \frac{\omega_J}{\|g_J^k - x_J^k\|_2}\right) (\bar{g}_J - \bar{x}_J^k) + o(\delta_k) + \frac{\omega_J \delta_k u_J^k}{\|g_J^k - x_J^k\|_2} \\
&= \left(\frac{\omega_J}{\|\bar{g}_J - \bar{x}_J^k\|_2} - \frac{\omega_J}{\|\bar{g}_J - \bar{x}_J^k - \delta_k u_J^k + o(\delta_k)\|_2}\right) (\bar{g}_J - \bar{x}_J^k) \\
&\quad + o(\delta_k) + \frac{\omega_J \delta_k u_J^k}{\|\bar{g}_J - \bar{x}_J^k - \delta_k u_J^k + o(\delta_k)\|_2} \\
&= \frac{\omega_J \langle \bar{g}_J - \bar{x}_J^k, -\delta_k u_J^k \rangle}{\|\bar{g}_J - \bar{x}_J^k\|_2^3} (\bar{g}_J - \bar{x}_J^k) + o(\delta_k) + \frac{\omega_J \delta_k u_J^k}{\|\bar{g}_J - \bar{x}_J^k\|_2} \\
&= \frac{\omega_J \delta_k}{\|\bar{g}_J - \bar{x}_J^k\|_2} \left(\frac{\langle \bar{g}_J, -u_J^k \rangle}{\omega_J^2} \bar{g}_J + u_J^k\right) + o(\delta_k),
\end{aligned}$$

where the second and fifth equalities use (65) and (66); the fourth and last equalities use (68) so that $\bar{g}_J = \omega_J \frac{\bar{g}_J - \bar{x}_J^k}{\|\bar{g}_J - \bar{x}_J^k\|_2}$; the sixth equality uses $\nabla_x \|x\|_2^{-1} = -\frac{x}{\|x\|_2^3}$. Multiplying both sides by $\|\bar{g}_J - \bar{x}_J^k\|_2/(\omega_J \delta_k)$ and using (61) and $\|\bar{g}_J\|_2 = \omega_J$ (by (68)) yields in the limit

$$0 = -\frac{\langle \bar{g}_J, \bar{u}_J \rangle}{\|\bar{g}_J\|_2^2} \bar{g}_J + \bar{u}_J. \tag{69}$$

Thus $\bar{u}_J$ is a nonzero multiple of $\bar{g}_J$.

Case (c). In this case, it follows from $\{\bar{x}^k\} \to \bar{x}$ that $\bar{x}_J = 0$. Since $\|g_J^k - x_J^k\|_2 > \omega_J$ for all k, this implies $\|\bar{g}_J\|_2 \geq \omega_J$. Since $\bar{x} \in \bar{\mathcal{X}}$, the optimality condition $0 \in \bar{g}_J + \omega_J \partial \|0\|_2$ implies

$\|\bar{g}_J\|_2 \leq \omega_J$. Hence $\|\bar{g}_J\|_2 = \omega_J$. Then Lemma 1 implies

$$\begin{aligned} -r_J^k &= \left(\frac{\omega_J}{\|\bar{g}_J\|_2} - \frac{\omega_J}{\|g_J^k - x_J^k\|_2}\right) g_J^k + \frac{\omega_J x_J^k}{\|g_J^k - x_J^k\|_2} \\ &= \frac{\omega_J \langle \bar{g}_J, g_J^k - x_J^k - \bar{g}_J\rangle}{\|\bar{g}_J\|_2^3} g_J^k + o(\|g_J^k - x_J^k - \bar{g}_J\|_2) + \frac{\omega_J x_J^k}{\|g_J^k - x_J^k\|_2} \\ &= -\frac{\omega_J \langle \bar{g}_J, x_J^k\rangle}{\|\bar{g}_J\|_2^3} g_J^k + o(\delta_k) + \frac{\omega_J x_J^k}{\|g_J^k - x_J^k\|_2}, \end{aligned}$$

where the second equality uses $\nabla_x \|x\|_2^{-1} = -\frac{x}{\|x\|_2^3}$, and the third equality uses (66) and $\bar{x}_J^k = 0$. Since $\bar{x}_J^k = 0$ for all k, (65) implies $\{x_J^k/\delta_k\} = \{u_J^k\} \to \bar{u}_J$. Thus, dividing both sides by δ_k and using (61), $\{x_J^k\} \to 0$, and $\|\bar{g}_J\|_2 = \omega_J$ yield in the limit (69). Since $\|g_J^k - x_J^k\|_2 > \omega_J$ for all k, (59) implies

$$g_J^k + r_J^k + \omega_J \frac{x_J^k + r_J^k}{\|x_J^k + r_J^k\|_2} = 0 \quad \forall k.$$

Suppose $\bar{u}_J \neq 0$. Then $u_J^k = x_J^k/\delta_k \neq 0$ for all k sufficiently large, so that

$$\langle g_J^k, u_J^k\rangle = \frac{\langle g_J^k, x_J^k\rangle}{\delta_k} = -\frac{\langle r_J^k, x_J^k\rangle}{\delta_k} - \omega_J \frac{\|x_J^k\|_2^2 + \langle r_J^k, x_J^k\rangle}{\delta_k \|x_J^k + r_J^k\|_2} = -\frac{\langle r_J^k, x_J^k\rangle}{\delta_k} - \omega_J \frac{\|u_J^k\|_2 + \langle \frac{r_J^k}{\delta_k}, \frac{x_J^k}{\|x_J^k\|_2}\rangle}{\left\| \frac{x_J^k}{\|x_J^k\|_2} + \frac{r_J^k}{\|x_J^k\|_2}\right\|_2}.$$

Then (61) and $\{\|x_J^k\|_2/\delta_k\} \to \|\bar{u}_J\| > 0$ yield in the limit that $\langle \bar{g}_J, \bar{u}_J\rangle = -\omega_J \|\bar{u}_J\|_2 < 0$. This together with (69) implies $\bar{u}_J$ is a negative multiple of $\bar{g}_J$.

Since $\{(x^k - \bar{x}^k)/\delta_k\} = \{u^k\} \to \bar{u} \neq 0$, we have $\langle x^k - \bar{x}^k, \bar{u}\rangle > 0$ for all k sufficiently large. Fix any such k and let

$$\hat{x} := \bar{x}^k + \epsilon \bar{u}$$

with $\epsilon > 0$. Since $A\bar{u} = 0$, we have $\nabla f(\hat{x}) = \nabla f(\bar{x}^k) = \bar{g}$. We show below that, for $\epsilon > 0$ sufficiently small, $\hat{x}$ satisfies

$$0 \in \bar{g}_J + \omega_J \partial \|\hat{x}_J\|_2 \tag{70}$$

for all $J \in \mathcal{J}$, and hence $\hat{x} \in \bar{\mathcal{X}}$. Then $\langle x^k - \bar{x}^k, \bar{u}\rangle > 0$ and $\|\bar{u}\|_2 = 1$ yield

$$\|x^k - \hat{x}\|_2^2 = \|x^k - \bar{x}^k - \epsilon\bar{u}\|_2^2 = \|x^k - \bar{x}^k\|_2^2 - 2\epsilon\langle x^k - \bar{x}^k, \bar{u}\rangle + \epsilon^2 < \|x^k - \bar{x}^k\|_2^2$$

for all $\epsilon > 0$ sufficiently small, which contradicts $\bar{x}^k$ being the point in $\bar{\mathcal{X}}$ nearest to x^k and thus proves (64). For each $J \in \mathcal{J}$, if $\bar{u}_J = 0$, then $\hat{x}_J = \bar{x}_J^k$ and (70) holds automatically (since $\bar{x}^k \in \bar{\mathcal{X}}$). Suppose that $\bar{u}_J \neq 0$. We prove (70) below by considering the three aforementioned cases (a), (b), and (c).

Case (a). Since $\bar{u}_J \neq 0$, we have that $\bar{u}_J$ is a positive multiple of $\bar{g}_J$. Also, by (68), $\bar{x}_J^k$ is a negative multiple of $\bar{g}_J$. Hence $\hat{x}_J$ is a negative multiple of $\bar{g}_J$ for all $\epsilon > 0$ sufficiently small, so, by (68), it satisfies

$$\bar{g}_J + \omega_J \frac{\hat{x}_J}{\|\hat{x}_J\|_2} = 0. \tag{71}$$

Case (b). Since (68) holds, $\bar{x}_J^k$ is a negative multiple of $\bar{g}_J$. Also, $\bar{u}_J$ is a nonzero multiple of $\bar{g}_J$. A similar argument as in case (a) shows that $\hat{x}_J$ satisfies (71) for all $\epsilon > 0$ sufficiently small.

Case (c). We have $\bar{x}_J^k = 0$ and $\bar{u}_J$ is a negative multiple of $\bar{g}_J$. Hence $\hat{x}_J$ is a negative multiple of $\bar{g}_J$ for all $\epsilon > 0$, so it satisfies (71).

The remaining argument is similar to the proof of [144, Theorem 4], but using (64) and the strong convexity of h in place of the strong convexity of f. For each k, since r^k is a solution of the subproblem (47), by Fermat's rule [118, Theorem 10.1],

$$r^k \in \arg\min_d \langle g^k + r^k, d\rangle + P(x^k + d).$$

Hence

$$\langle g^k + r^k, r^k\rangle + P(x^k + r^k) \leq \langle g^k + r^k, \bar{x}^k - x^k\rangle + P(\bar{x}^k).$$

Since $\bar{x}^k \in \bar{\mathcal{X}}$ and $\nabla f(\bar{x}^k) = \bar{g}$, we have similarly that

$$P(\bar{x}^k) \leq \langle \bar{g}, x^k + r^k - \bar{x}^k\rangle + P(x^k + r^k).$$

Adding the above two inequalities and simplifying yield

$$\langle g^k - \bar{g}, x^k - \bar{x}^k\rangle + \|r^k\|_2^2 \leq \langle \bar{g} - g^k, r^k\rangle + \langle r^k, \bar{x}^k - x^k\rangle.$$

Since Ax^k and $A\bar{x}^k = \bar{y}$ are in Y, we also have from (62), (63), and (64) that

$$\langle g^k - \bar{g}, x^k - \bar{x}^k\rangle = \langle \nabla h(Ax^k) - \nabla h(\bar{y}), Ax^k - \bar{y}\rangle \geq \sigma\|Ax^k - \bar{y}\|_2^2 \geq \frac{\sigma}{\kappa^2}\|x^k - \bar{x}^k\|_2^2.$$

Combining these two inequalities and using (63) yield

$$\begin{aligned}\frac{\sigma}{\kappa^2}\|x^k - \bar{x}^k\|_2^2 + \|r^k\|_2^2 &\leq \langle \nabla h(\bar{y}) - \nabla h(Ax^k), Ar^k\rangle + \langle r^k, \bar{x}^k - x^k\rangle \\ &\leq L\|A\|_2^2\|x^k - \bar{x}^k\|_2\|r^k\|_2 + \|x^k - \bar{x}^k\|_2\|r^k\|_2,\end{aligned}$$

where $\|A\|_2 := \max_{\|d\|_2=1} \|Ad\|_2$. Thus

$$\frac{\sigma}{\kappa^2}\|x^k - \bar{x}^k\|_2^2 \leq (L\|A\|_2^2 + 1)\|x^k - \bar{x}^k\|_2\|r^k\|_2 \quad \forall k.$$

Dividing both sides by $\|x - \bar{x}^k\|_2$ yields a contradiction to (61).

The key to the above proof is the bound (64). An analogous bound was used in the proof of [78, Lemma 2.6(b)] for the case of (16) with $\mathcal{X}$ polyhedral and f satisfying C1. The above proof is more complex due to P being non-polyhedral.

References

[1] Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P., A new approach to collaborative filtering: operator estimation with spectral regularization, to appear in J. Machine Learn. Res. 2009.

[2] Alfakih, A. Y., Graph rigidity via Euclidean distance matrices, Lin. Algeb. Appl., 310 (2000), 149-165.

[3] Argyriou, A., Evgeniou, T., and Pontil, M., Convex multi-task feature learning, Machine Learn. 73 (2008), 243-272.

[4] Aspnes, J., Goldenberg, D., and Yang, Y. R., On the computational complexity of sensor network localization, in Lecture Notes in Computer Science 3121, Springer-Verlag, 2004, pp. 32-44.

[5] D'Aspremont, A., Banerjee, O., and Ghaoui, L. E., First-order methods for sparse covariance selection, SIAM J. Matrix Anal. Appl. 30 (2008), 56-66.

[6] Auslender, A., Minimisation de fonctions localement lipschitziennes: applications à la programmation mi-convexe, mi-différentiable, in O. L. Mangasarian, R. R. Meyer and S. M. Robinson, editors, Nonlinear Programming, 3, Academic Press, New York (1978), 429-460.

[7] Auslender, A. and Teboulle, M., Interior projection-like methods for monotone variational inequalities, Math. Program. 104 (2005), 39-68.

[8] Auslender, A. and Teboulle, M., Interior gradient and proximal methods for convex and conic optimization, SIAM J. Optim. 16 (2006), 697-725.

[9] Banerjee, O., Ghaoui, L. E., and D'Aspremont, A., Model selection through sparse maximum likelihood estimation, J. Mach. Learn. Res. 9 (2008), 485-516.

[10] Bauschke, H. H., Borwein, J. M., and Combettes, P. L., Bregman monotone optimization algorithms, SIAM J. Control Optim. 42 (2003), 596-636.

[11] Beck, A. and Teboulle, M., Mirror descent and nonlinear projected subgradient methods for convex optimization, Oper. Res. Letters 31 (2003), 167-175.

[12] Beck, A. and Teboulle, M., A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization, Math. Oper. Res. 31 (2006), 398-417.

[13] Beck, A. and Teboulle, M., A fast iterative shrinkage-threshold algorithm for linear inverse problems, Report, Department of Industrial Engineering and Management, Technion, Haifa, 2008.

[14] Becker, S., Bobin, J., and Candès, E. J., NESTA: A Fast and accurate first-order method for sparse recovery, Report, California Institute of Technology, Pasadena, April 2009.

[15] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999.

[16] Biswas, P., Aghajan, H. and Ye, Y., Semidefinite programming algorithms for sensor network localization using angle of arrival information, in Proc. 39th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, 2005.

[17] Biswas, P., Liang, T.-C., Toh, K.-C., Wang, T.-C., and Ye, Y., Semidefinite programming approaches for sensor network localization with noisy distance measurements, IEEE Trans. Auto. Sci. Eng. 3 (2006), 360-371.

[18] Biswas, P., Liang, T.-C., Wang, T.-C. and Ye, Y., Semidefinite programming based algorithms for sensor network localization, ACM Trans. Sensor Networks 2 (2006), 188-220.

[19] Biswas, P., Toh, K.-C. and Ye, Y., A distributed SDP approach for large-scale noisy anchor-free graph realization with applications to molecular conformation, SIAM J. Sci. Comput. 30 (2008), 1251-1277.

[20] Biswas, P. and Ye, Y., Semidefinite programming for ad hoc wireless sensor network localization, Proc. 3rd IPSN, 2004, 46-54.

[21] Biswas, P. and Ye, Y., A distributed method for solving semidefinite programs arising from ad hoc wireless sensor network localization, in Multiscale Optimization Methods and Applications, Vol. 82 of Nonconvex Optim. Appl., 2003.

[22] Blatt, D. Hero, A. O., and Gauchman, H., A convergent incremental gradient method with a constant step size, SIAM J. Optim. 18 (2007), 29-51.

[23] Bregman, L. M., The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967), 200-217.

[24] Cai, J.-F., Candès, E. J. and Shen, Z., Singular value thresholding algorithm for matrix completion, Report, Applied and Computational Mathematics, California Institute of Technology, Pasadena, September 2008.

[25] Candès, E. J. and Recht, B., Exact matrix completion via convex optimization, Report, Applied and Computational Mathematics, California Institute of Technology, Pasadena, May 2008.

[26] Candès, E. J., Romberg, J., and Tao, T., Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Inf. Theory 52 (2006) 489-509.

[27] Candès, E. J., Romberg, J., and Tao, T., Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (2006), 1207-1223.

[28] Candès, E. J. and Tao, T., Decoding by linear programming, IEEE Inf. Theory 51 (2005), 4203-4215.

[29] Carter, W., Jin, H. H., Saunders, M. A., and Ye, Y., An adaptive rule-based algorithm is proposed to solve localization problems for scalable wireless sensor network localization, SIAM J. Optim. 17 (2006), 1102-1128.

[30] Chen, S., Donoho, D., and Saunders, M., Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1999), 33-61.

[31] Censor, Y. and Zenios, S. A., Parallel Optimization: Theory, Algorithms, and Applications, Oxford Univ. Press, New York, 1997.

[32] Chen, G. and Teboulle, M., Convergence analysis of a proximal-like minimization algorithm using Bregman functions, SIAM J. Optim. 3 (1993), 538-543.

[33] Dasarathy, B. and White, L. J., A maxmin location problem, Oper. Res. 28 (1980), 1385-1401.

[34] Daubechies, I., Defrise, M., and De Mol, C., An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 57 (2004), 1413-1457.

[35] Davies, M. E. and Gribonval, R., Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$, Report, IRISA No 1899, Rennes, July 2008.

[36] Ding, Y., Krislock, N., Qian, J. and Wolkowicz, H., Sensor network localization, Euclidean distance matrix completions, and graph realization, Report, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, February 2008.

[37] Doherty, L., Pister, K. S. J., and El Ghaoui, L., Convex position estimation in wireless sensor networks, Proc. 20th INFOCOM 3 (2001), 1655-1663.

[38] Donoho, D. L., For most large underdetermined systems of linear equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution, Report, Department of Statistics, Stanford University, Stanford, 2006.

[39] Donoho, D. L., For most large underdetermined systems of linear equations, the minimal ℓ^1-norm solution is also the sparsest solution, Comm. Pure Appl. Math. 59 (2006) 797-929.

[40] Donoho, D. L. and Elad, M., Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization, Proc. Nat. Acad. Sci. 100 (2003) 2197-2202.

[41] Donoho, D. L., Elad, M., and Temlyakov, V. N., Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inf. Theory 52 (2006) 6-18.

[42] Donoho, D. L. and Huo, X., Uncertainty principles and ideal atomic decomposition, IEEE Trans. Inf. Theory 47 (2001) 2845-2862.

[43] Donoho, D. L. and Johnstone, I. M., Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81 (1994), 425-455.

[44] Donoho, D. L. and Johnstone, I. M., Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (1995), 1200-1224.

[45] Donoho, D. L., Tsaig, Y., Drori, I., and Starck, J.-L., Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit, Report, Department of Statistics, Stanford University, Stanford, 2006.

[46] Eckstein, J., Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming, Math. Oper. Res. 18 (1993) 202-226.

[47] Eren, T., Goldenberg, D. K., Whiteley, W., Yang, Y. R., Morse, A. S., Anderson, B. D. O., and Belhumeur, P. N., Rigidity, computation, and randomization in network localization, Proc. 23rd INFOCOM, (2004).

[48] Evgeniou, T. and Pontil, M., Regularized multi-task learning, Proc. 17th SIGKDD Conf. on Knowledge, Discovery and Data Mining, Seattle, 2004, 109-117.

[49] Facchinei, F. and Pang, J.-S., Finite-Dimensional Variational Inequalities and Complementarity Problems, Vols. I and II, Springer-Verlag, New York, 2003.

[50] Fazel, M., Hindi, H., and Boyd, S. P., A rank minimization heuristic with application to minimum order system approximation, in Proc. American Control Conf., Arlington, 2001, 4734-4739.

[51] Ferris, M. C. and Mangasarian, O. L., Parallel variable distribution, SIAM J. Optim. 4 (1994), 815-832.

[52] Fletcher, R., An overview of unconstrained optimization, in Algorithms for Continuous Optimization, edited by E. Spedicato, Kluwer Academic, Dordrecht, 1994, 109-143.

[53] Friedman, J., Hastie, T., and Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics (2007), 1-10.

[54] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R., Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2007), 302-332.

[55] Friedman, J., Hastie, T., and Tibshirani, Regularization paths for generalized linear models via coordinate descent, Report, Department of Statistics, Stanford University, Stanford, July 2008.

[56] Fuchs, J.-J., On sparse representations in arbitrary redundant bases, IEEE Trans. Inf. Theory 50 (2004), 1341-1344.

[57] Fuchs, J.-J., Recovery of exact sparse representations in the presence of bounded noise, IEEE Trans. Inf. Theory, 51 (2005), 3601-3608.

[58] Fukushima, M., Parallel variable transformation in unconstrained optimization, SIAM J. Optim. 8 (1998), 658-672.

[59] Gilbert, A. C., Muthukrishnan, S., and Strauss, M. J., Approximation of functions over redundant dictionaries using coherence, Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2003, 243-252.

[60] Goldfarb, D. and Yin, W., Second-order cone programming methods for total variation based image restoration, SIAM J. Sci. Comput. 27 (2005), 622-645.

[61] Gonzaga, C. C. and Karas, E. W., Optimal steepest descent algorithms for unconstrained convex problems: fine tuning Nesterov's method, Report, Department of Mathematics, Federal University of Santa Catarina, Florianópolis, August 2008.

[62] Gribonval, R. and Nielsen, M., Sparse representations in unions of bases, IEEE Trans. Inf. Theory 49 (2003) 3320-3325.

[63] Hale, E., Yin, W., and Zhang, Y., Fixed-point continuation for ℓ_1-minimization: methodology and convergence, Report, Department of Computational and Applied Mathematics, Rice University, Houston, revised 2008.

[64] Hoda, S., Gilpin, A., and Peña, J., Smoothing techniques for computing Nash equilibria of sequential games, Report, Carnegie Mellon University, Pittsburg, March 2008.

[65] Juditsky, A., Lan, G., Nemirovski, A., and Shapiro, A., Stochastic approximation approach to stochastic programming, Report, 2007; to appear in SIAM J. Optim.

[66] Kim, S., Kojima, M. and Waki, H., Exploiting sparsity in SDP relaxation for sensor network localization, Research report B-447, Tokyo Institute of Technology, Tokyo, October 2008.

[67] Kiwiel, K. C., Proximal minimization methods with generalized Bregman functions, SIAM J. Control Optim. 35 (1997), 1142-1168.

[68] Kiwiel, K. C., Convergence of approximate and incremental subgradient methods for convex optimization, SIAM J. Optim. 14 (2003), 807-840.

[69] Kiwiel, K. C., On linear time algorithms for the continuous quadratic knapsack problem, J. Optim. Theory Appl. 134 (2007), 549-554.

[70] Krislock, N., Piccialli, V., and Wolkowicz, H., Robust semidefinite programming approaches for sensor network localization with anchors, Report, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, May 2006.

[71] Lan, G., Lu, Z., and Monteiro, R. D. C., Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming, Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, December 2006.

[72] Liang, T.-C., Wang, T.-C. and Ye, Y., A gradient search method to round the semidefinite programming relaxation solution for ad hoc wireless sensor network localization, Report, Electrical Engineering, Stanford University, Stanford, October 2004.

[73] Liu, Z. and Vandenberghe, L., Interior-point method for nuclear norm approximation with application to system identification, Report, Electrical Engineering Department, UCLA, Los Angeles, 2008.

[74] Lu, Z., Smooth optimization approach for sparse covariance selection, Report, Department of Mathematics, Simon Fraser University, Burnaby, January 2008; submitted to SIAM J. Optim.

[75] Lu, Z., Monteiro, R. D. C., and Yuan, M., Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, January 2008; revised March 2009.

[76] Lu, Z. Nemirovski, A. S., and Monteiro, R. D. C., Large-scale semidefinite programming via saddle point mirror-prox algorithm, Math. Program. 109 (2007), 211-237.

[77] Luo, Z.-Q., On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks, Neural Comput. 3 (1991), 226-245.

[78] Luo, Z.-Q. and Tseng, P., On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. 30 (1992), 408-425.

[79] Luo, Z.-Q. and Tseng, P., On the convergence rate of dual ascent methods for linearly constrained convex minimization. Math. Oper. Res. 18 (1993), 846-867.

[80] Luo, Z.-Q. and Tseng, P., Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993), 157-178.

[81] Luo, Z.-Q. and Tseng, P., On the rate of convergence of a distributed asynchronous routing algorithm, IEEE Trans. Automat. Control 39 (1994), 1123-1129.

[82] Ma, S., Goldfarb, D., and Chen, L., Fixed point and Bregman iterative methods for matrix rank minimization, Report 08-78, UCLA Computational and Applied Mathematics, 2008.

[83] Mallat, S. and Zhang, Z., Matching pursuit in a time-frequency dictionary, IEEE Trans. Signal Process. 41 (1993), 3397-3415.

[84] Mangasarian, O. L, Sparsity-preserving SOR algorithms for separable quadratic and linear programming, Comput. Oper. Res. 11 (1984), 105-112.

[85] Mangasarian, O. L., Mathematical programming in neural networks, ORSA J. Comput. 5 (1993), 349-360.

[86] Mangasarian, O. L., Parallel gradient distribution in unconstrained optimization, SIAM J. Control Optim. 33 (1995), 1916-1925.

[87] Mangasarian, O. L. and Musicant, D. R., Successive overrelaxation for support vector machines, IEEE Trans. Neural Networks, 10 (1999), 1032-1037.

[88] Meier, L., van de Geer, S., and Bühlmann, P., The group Lasso for logistic regression, J. Royal Statist. Soc. B. 70 (2008), 53-71.

[89] 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 (1981), 9-23.

[90] Moré, J. J. and Wu, Z., Global continuation for distance geometry problems, SIAM J. Optim., 7 (1997), 814-836.

[91] Nedić, A. and Bertsekas, D. P., Incremental subgradient methods for nondifferentiable optimization, SIAM J. Optim. 12 (2001), 109-138.

[92] Nemirovski, A., Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, SIAM J. Optim. 15 (2005), 229-251.

[93] Nemirovski, A. and Yudin, D., Problem Complexity and Method Efficiency in Optimization, Wiley, New York, 1983.

[94] Nesterov, Y., A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, Doklady AN SSSR 269 (1983), 543-547; translated as Soviet Math. Dokl.

[95] Nesterov, Y., On an approach to the construction of optimal methods of minimization of smooth convex functions, Èkonom. i. Mat. Metody 24 (1988), 509-517.

[96] Nesterov, Y., Smoothing technique and its applications in semidefinite optimization, Report, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, October 2004.

[97] Nesterov, Y., Introductory Lectures on Convex Optimization, Kluwer Academic Publisher, Dordrecht, The Netherlands, 2004.

[98] Nesterov, Y., Smooth minimization of nonsmooth functions, Math. Program. 103 (2005), 127-152.

[99] Nesterov, Y., Excessive gap technique in nonsmooth convex minimization, SIAM J. Optim. 16 (2005), 235-249.

[100] Nesterov, Y., Dual extrapolation and its applications to solving variational inequalities and related problems, Math. Program. 109 (2007), 319-344.

[101] Nesterov, Y., Gradient methods for minimizing composite objective function, Report, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, September 2007.

[102] Nesterov Y., Primal-dual subgradient methods for convex problems, Math. Program. 2007, DOI 10.1007/s10107-007-0149-x.

[103] Nesterov, Y. and Nemirovskii, A., Interior Point Polynomial Algorithms in Convex Programming, SIAM, Philadelphia, 1994.

[104] Nie, J., Sum of squares method for sensor network localization, Report, Department of Mathematics, University of California, Berkeley, June 2006; to appear in Comput. Optim. Appl.

[105] Nocedal, J. and Wright S. J., Numerical Optimization, Springer-Verlag, New York, 1999.

[106] Obozinski, G., Taskar, B., and Jordan, M. I., Joint covariate selection and joint subspace selection for multiple classification problems, to appear in Stat. Comput. 2009.

[107] Ortega, J. M. and Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables, reprinted by SIAM, Philadelphia, 2000.

[108] Osher, S., Burger, M., Goldfarb, D., and Xu, J., An iterative regularization method for total variation-based image restoration, SIAM J. Multiscale Modeling Simulation 4 (2005), 460-489.

[109] Pang, J.-S., A posteriori error bounds for the linearly–constrained variational inequality problem, Math. Oper. Res. 12 (1987), 474–484.

[110] Park, M.-Y. and Hastie, T., An L_1 regularization-path algorithm for generalized linear models, J. Roy. Soc. Stat. B 69 (2007), 659-677.

[111] Pfander, G. E., Rauhut, H., and Tanner, J., Identification of matrices having a sparse representation, Report, 2008, submited to IEEE Trans. Signal Proc.

[112] Polyak, B. T., Introduction to Optimization, Optimization Software, New York, 1987.

[113] Pong, T. K. and Tseng, P., (Robust) edge-based semidefinite programming relaxation of sensor network localization, Report, Department of Mathematics, University of Washington, Seattle, January 2009; submitted to Math. Program.

[114] Ravi, S. S., Rosenkrantz, D. J. , and Tayi, G. K., Heuristic and special case algorithms for dispersion problems, Oper. Res. 42 (1994), 299-310.

[115] Recht, F., Fazel, M., and Parrilo, P., Guaranteed minimum rank solutions of matrix equation via nuclear norm minimization, 2007, submitted to SIAM Review.

[116] Rockafellar, R. T., Convex Analysis, Princeton Univ. Press, Princeton, 1970.

[117] Rockafellar, R. T., Conjugate Duality and Optimization, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[118] Rockafellar, R.T. and Wets R.J.-B., Variational Analysis, Springer-Verlag, New York, 1998.

[119] Rohl C. A., Strauss C. E. M., Misura K., and Baker D. Protein structure prediction using Rosetta, Methods Enzym. 383 (2004), 66-93.

[120] Sagastizábal, C. A. and Solodov, M. V., Parallel variable distribution for constrained optimization, Comput. Optim. Appl. 22 (2002), 111-131.

[121] Sardy, S., Bruce, A., and Tseng, P., Block coordinate relaxation methods for nonparametric wavelet denoising, J. Comput. Graph. Stat. 9 (2000), 361-379.

[122] Sardy, S., Bruce, A., and Tseng, P., Robust wavelet denoising, IEEE Trans. Signal Proc. 49 (2001), 1146-1152.

[123] Sardy, S., Antoniadis, A., and Tseng, P., Automatic smoothing with wavelets for a wide class of distributions, J. Comput. Graph. Statist. 13 (2004), 399-421.

[124] 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 (2004), 283-309.

[125] Sardy, S. and Tseng, P., On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions, J. Amer. Statist. Assoc. 99 (2004), 191-204.

[126] Shi, J., Yin, W., Osher, S., and Sajda, P., A fast algorithm for large scale ℓ_1-regularized logistic regression, Report, Department of Computational and Applied Mathematics, Rice University, Houston, 2008.

[127] So, A. M.-C. and Ye, Y., Theory of semidefinite programming for sensor network localization, Math. Program. 109 (2007), 367-384.

[128] Solodov, M. V., Incremental gradient algorithms with stepsizes bounded away from zero, Comput. Optim. Appl., 11 (1998), 23-35.

[129] Strohmer, T. and Heath Jr., R., Grassmannian frames with applications to coding and communications, Appl. Comp. Harm. Anal. 14 (2003), 257-275.

[130] Teboulle, M., Convergence of proximal-like algorithms, SIAM J. Optim. 7 (1997), 1069-1083.

[131] Tibshirani, R., Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. B. 58 (1996), 267-288.

[132] Toh, K.-C. and Yun, S., An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, Report, Department of Mathematics, National University of Singapore, Singapore, 2009.

[133] Tropp, J. A., Greed is good: algorithmic results for sparse approximation, IEEE Trans. Inf. Theory 50 (2004), 2231-2242.

[134] Tropp, J. A., Just relax: Convex programming methods for identifying sparse signals in noise, IEEE Trans. Inf. Theory 52 (2006), 1030-1051.

[135] Tropp, J. A. and Gilbert, A. C., Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (2007), 4655-4666.

[136] Tseng, P., Dual coordinate ascent methods for non-strictly convex minimization, Math. Program. 59 (1993), 231-247.

[137] Tseng, P., An incremental gradient(-projection) method with momentum term and adaptive stepsize rule, SIAM J. Optim. 8 (1998), 506-531.

[138] Tseng, P., On the rate of convergence of a partially asynchronous gradient projection algorithm. SIAM J. Optim. 1 (1991), 603-619.

[139] Tseng, P., Error bounds and superlinear convergence analysis of some Newton-type methods in optimization, in Nonlinear Optimization and Related Topics, Kluwer, Dordrecht, 2000, 445-462.

[140] Tseng, P., Convergence of block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (2001), 473-492.

[141] Tseng, P., Second-order cone programming relaxation of sensor network localizations, SIAM J. Optim. 18 (2007), 156-185.

[142] Tseng, P., On accelerated proximal gradient methods for convex-concave optimization Report, Department of Mathematics, University of Washington, Seattle, May 2008, submitted to SIAM J. Optim.

[143] Tseng, P., Further results on a stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inf. Theory. 55 (2009), 888-899.

[144] Tseng, P. and Yun S., A coordinate gradient descent method for nonsmooth separable minimization, Math. Program. 117 (2009), 387-423.

[145] Tseng, P. and Yun S., A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training, to appear in Comput. Optim. Appl.

[146] Tseng, P. and Yun S., A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization, to appear in J. Optim. Theory Appl. 140 (2009), 513-535.

[147] Wang, Y., Yang, J., Yin, W., and Zhang, Y., A new alternating minimization algorithm for total variation image reconstruction, Report, Department of Computational and Applied Mathematics, Rice University, Houston, 2007; to appear in SIAM Imaging Sci.

[148] Wang, Z., Zheng, S., Ye, Y., and Boyd, S., Further relaxations of the semidefinite programming approach to sensor network localization, SIAM J. Optim. 19 (2008), 655-673.

[149] White, D. J., A heuristic approach to a weighted maxmin dispersion problem, IMA J. Math. Appl. Business Ind. 9 (1996), 219-231.

[150] Wright, S. J., Primal-Dual Interior-Point Methods, SIAM, Philadelphia, 1997.

[151] Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T., Sparse reconstruction by separable approximation, Report, Computer Sciences Department, University of Wisconsin, Madison, October 2007.

[152] Yang, J., Zhang, Y., and Yin, W., An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise, Report, Department of Computational and Applied Mathematics, Rice University, Houston, 2008.

[153] Ye, J., Ji, S., and Chen, J., Multi-class discriminant kernel learning via convex programming, J. Machine Learning Res. 9 (2008), 719-758.

[154] Ye, Y., Interior Point Algorithms: Theory and Analysis, John Wiley & Sons, New York, 1997.

[155] Yin, W., Osher, S., Goldfarb, D., and Darbon, J., Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing, SIAM J. Imaging Sci. 1 (2008), 143-168.

[156] Yuan, M., Ekici, A., Lu, Z., and Monteiro, R., Dimension reduction and coefficient estimation in multivariate linear regression, J. Royal Stat. Soc. B 9 (2007), 329-346.

[157] Yuan, M. and Lin, Y., Model selection and estimation in regression with grouped variables, J. Royal Stat. Soc. B 68 (2006), 49-67.

[158] Yuan, M. and Lin, Y., Model selection and estimation in the Gaussian graphical model, Biometrika 94 (2007), 19-35.

[159] Yun, S. and Toh, K.-C., A coordinate gradient descent method for ℓ_1-regularized convex minimization, Report, Department of Mathematics, National University of Singapore, Singapore, 2008; submitted to Comput. Optim. Appl.

[160] Zhu, M. and Chan, T. F., An efficient primal-dual hybrid gradient algorithm for total variation image restoration, Report, Department of Mathematics, UCLA, Los Angeles, 2008.

[161] Zhu, M., Wright, S. J., and Chan, T. F., Duality-based algoritms for total-variation regularized image restoration, Report, Department of Mathematics, UCLA, Los Angeles, 2008.