Department of Computer Science, University of British Columbia
Technical Report TR-2010-02, January 2010

SPARSE OPTIMIZATION WITH LEAST-SQUARES CONSTRAINTS

EWOUT VAN DEN BERG AND MICHAEL P. FRIEDLANDER*

Abstract. The use of convex optimization for the recovery of sparse signals from incomplete or compressed data is now common practice. Motivated by the success of basis pursuit in recovering sparse vectors, new formulations have been proposed that take advantage of different types of sparsity. In this paper we propose an efficient algorithm for solving a general class of sparsifying formulations. For several common types of sparsity we provide applications, along with details on how to apply the algorithm, and experimental results.

Key words. basis pursuit, compressed sensing, convex program, duality, group sparsity, matrix completion, Newton's method, root-finding, sparse solutions

AMS subject classifications. 49M29, 65K05, 90C25, 90C06

1. Introduction. Many signal- and image-processing applications aim to approximate an object as a superposition of only a few elementary atoms from a basis or dictionary. Although the problem of finding the smallest subset of atoms that gives the "best" representation is generally intractable [48], relaxations that involve convex optimization often perform remarkably well, and can, under certain conditions [16, 25, 26], recover the sparsest solution. However, the resulting optimization problems are typically large scale and involve nonsmooth objective functions. This paper offers an algorithmic framework for solving a class of optimization problems that has wide applicability to the reconstruction of signals and images with sparse representations.

Compressed sensing [12, 24] is a particular application of sparse approximation. In this case, a signal y is measured by applying an m-by-n linear operator M, typically with $m \ll n$; this yields the compressed observation $b := My$. Direct reconstruction of y from b is generally impossible because, faced with an underdetermined linear system, there are infinitely many solutions. This situation changes when we know that y has a sparse representation x in terms of a basis B. By finding a solution x such that

$$Ax \approx b \quad \text{with} \quad x \text{ sparse}, \tag{1.1}$$

where $A := MB$, we can then reconstruct y as $\widehat{y} := Bx$.

The compressed sensing concept has been successfully applied in areas ranging from magnetic resonance imaging [42] to seismic data interpolation [38], and led to the development of specialized optimization algorithms; see, e.g., [5, 32, 35, 40].

Several extensions to (1.1) lead to other interesting applications. For example, in applications where multiple measurements of a signal are possible (such as measurements over time or over various channels), we wish to find a set of solutions such that

$$Ax^1 \approx b^1, \ldots, Ax^r \approx b^r \quad \text{with} \quad x^1, \ldots, x^r \text{ jointly sparse}; \tag{1.2}$$

*Department of Computer Science, University of British Columbia, Vancouver V6T 1Z4, B.C., Canada ({ewout78,mpf}@cs.ubc.ca). This work was partially supported by a Discovery Grant from the Natural Sciences and Engineering Research Council. January 30, 2010

in other words, a set of solutions with a shared sparsity pattern is required. In the problem of rank minimization, rank is analogous to sparsity, and in that case we seek a matrix X such that

$$\mathcal{A}X \approx b \quad \text{with} \quad X \text{ low rank}; \tag{1.3}$$

here, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear operator.

1.1. Convex relaxations. Each of the these sparse recovery problems have convex optimization relaxations. The basis pursuit denoising approach [21] for (1.1) balances the 2-norm of the residual against the 1-norm of the solution, and solves the problem

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma, \tag{1.4}$$

where σ is a measure of the noise level; setting $\sigma = 0$ yields the basis pursuit problem.

One of the most promising approaches for (1.2) is based on solving the sum-of-norms formulation

$$\underset{X}{\text{minimize}} \quad \|X\|_{1,2} \quad \text{subject to} \quad \|AX - B\|_F \leq \sigma, \tag{1.5}$$

where the mixed p, q-norm is defined as

$$\|X\|_{p,q} := \left(\sum_{i=1}^{m} \|X_i^T\|_q^p \right)^{1/p} \tag{1.6}$$

with X_i denoting the ith row of X, and $\|\cdot\|_q$ the conventional q-norm [6, 30, 60]. The more general group-sparsity problem can be similarly formulated [30, 56].

The rank-minimization problem (1.3) can be approached via

$$\underset{X}{\text{minimize}} \quad \|X\|_n \quad \text{subject to} \quad \|\mathcal{A}X - b\|_2 \leq \sigma, \tag{1.7}$$

where $\|X\|_n$ is the nuclear norm of X, defined by the sum of singular values [31, 51].

1.2. Approach. The problems of interest in this paper can all be expressed as

$$\underset{x}{\text{minimize}} \quad \kappa(x) \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma, \tag{P_σ}$$

where κ is convex. We are particularly interested in nonsmooth functions κ that promote sparsity, and we assume throughout that κ is a gauge function—i.e., a convex, nonnegative, positively homogeneous function such that $\kappa(0) = 0$ [52, §15]. This class of functions, which includes norms, subsumes the formulations described in section 1.1, and is sufficiently general to accommodate other problems of practical interest. Functions more general than gauges do not fit into our theoretical framework; this limitation is discussed further in section 2.3.

The applications and implementations that we discuss revolve around the 2-norm of the misfit $\|Ax-b\|_2$, which is the case that most often arises in practice; however, our theoretical development allows for more general measures of misfit; see [3, Chapter 5].

A variation of (P_σ) that swaps the role of κ and the norm of the misfit is given by

$$\phi(\tau) := \underset{x}{\text{minimize}} \quad \|Ax - b\|_2 \quad \text{subject to} \quad \kappa(x) \leq \tau, \tag{L_τ}$$

and plays an essential role in our approach. The function $\phi(\tau)$ gives the optimal objective value as a function of the parameter τ. Problems (P_σ) and (L_τ) are equivalent

Algorithm 1: *Newton root-finding framework*

Input: A, b, σ
$x_0 \leftarrow 0$; $r_0 \leftarrow b$; $\tau_0 \leftarrow 0$; $k \leftarrow 0$
while $|\|r_k\|_2 - \sigma| > \epsilon$ **do**
  Solve (L_τ) for x_k
  $r_k \leftarrow Ax_k - b$
1 $\tau_{k+1} \leftarrow \tau_k - (\phi(\tau_k) - \sigma)/\phi'(\tau_k)$
  $k \leftarrow k + 1$
return $x^* \leftarrow x_k$; $\tau^* \leftarrow \tau_k$

in the sense that if there exists a solution x^* of (P_σ) for a given σ, then there exists a corresponding $\tau := \kappa(x^*)$ that causes x^* to also be a solution of (L_τ); a symmetric relation holds for $\sigma := \|Ax^* - b\|_2$. Thus, $\phi(\tau)$ gives the best possible trade-off between τ and σ, and its graph defines the Pareto curve. When $\kappa(x) = \|x\|_1$, (L_τ) is known as the Lasso problem [58].

Our approach to solving (P_σ) for a given σ by is based on iteratively refining an estimate of the smallest parameter τ_σ that satisfies the nonlinear equation

$$\phi(\tau) = \sigma; \tag{1.8}$$

this procedure requires computing successively more accurate solutions of (L_τ). When ϕ is differentiable—in section 2 we give conditions under which this holds—we can apply Newton's method to obtain a root of (1.8); Algorithm 1 provides a sketch of the overall approach.

The effectiveness of this approach hinges on efficiently solving (L_τ), efficiently evaluating the derivative $\phi'(\tau)$, and stability under approximations of $\phi(\tau)$ and $\phi'(\tau)$.

1.3. Related work. The root-finding framework for the particular case where $\kappa(x) = \|x\|_1$ was first described in [5], and implemented in the SPGL1 [4] software package. The success of the SPGL1 package in practice—see, e.g., [19, 36, 37, 39, 46, 62, 64]—motives us to provide a unified algorithm that applies to a wider class of problems, including sign-restricted basis pursuit denoise, sum-of-norms, and matrix completion problems.

The optimization problems that we discuss can all be formulated as second-order cone or semi-definite programs and solved by interior-point implementations such as SeDuMi [57] or SDPT3 [61]. However, these solvers depend on explicit matrix representations of A and cannot benefit from the fast implicit operators that typically arise in sparse-approximation applications. Other solvers suitable for general sparse-approximation problems, either apply to Lagrangian version of the problems that we consider (e.g., SpaRSA [63]), or make restrictive assumptions on A (e.g., NESTA [2]).

1.4. Outline. In section 2 we give conditions under which the Pareto curve is differentiable, and in section 3 describe practical aspects of the algorithm required for efficient implementation, including inexact subproblem solves. Section 4 describes two solvers for the subproblem (L_τ) which depend on the orthogonal projection of iterates onto the feasible set $\kappa(x) \leq \tau$. In the remaining four sections we develop the tools required for solving various incarnations of (P_σ), including weighted basis pursuit denoise (section 5), sum-of-norms (section 6), nonnegative basis pursuit (section 7), maxtrix completion (section 8), and sparse/low-rank matrix decomposition (section 9). Each section contains concrete applications of each of these problems and the results of numerical experiements.

1.5. Reproducible research. Following the discipline of reproducible research, the source code and data files required to reproduce all of the experimental results of this paper, including the figures and tables, and an extensive appendix of additional numerical experiments, can be downloaded from `http://www.cs.ubc.ca/~mpf/10vdBergFriedlander`.

2. The Pareto curve. We prove differentiability of the Pareto curve using two results. The first result [52, Theorem 25.1] states that a convex function is differentiable at a point x if and only if the subgradient at that point is unique; naturally, the gradient at x is then given by the unique subgradient. The second result [9, Propositions 6.1.2b, 6.5.8a] asserts that for a convex program of the form (L_τ), λ is a Lagrange multiplier if and only if $-\lambda \in \partial\phi(\tau)$, provided that $\phi(\tau)$ is finite, which is clearly the case. The key is then to derive an expression for the Lagrange multiplier and establish its uniqueness. In the following sections we derive the required dual problem and subsequently prove convexity and differentiability of ϕ.

2.1. The dual subproblem. As a first step in the derivation of the dual, we rewrite (L_τ) in terms of x and an explicit residual term r:

$$\underset{x,r}{\text{minimize}} \quad \|r\|_2 \quad \text{subject to} \quad Ax + r = b, \quad \kappa(x) \leq \tau. \tag{2.1}$$

The dual to this equivalent problem is given by

$$\underset{y,\lambda}{\text{maximize}} \quad \mathcal{L}(y,\lambda) \quad \text{subject to} \quad \lambda \geq 0, \tag{2.2}$$

where $y \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$ are dual variables, and $\mathcal{L}$ is the Lagrange dual function

$$\mathcal{L}(y,\lambda) := \inf_{x,r} \ \|r\|_2 - y^T(Ax + r - b) + \lambda(\kappa(x) - \tau).$$

By separability of the infimum over x and r we can rewrite $\mathcal{L}$ in terms of two separate suprema, giving

$$\mathcal{L}(y,\lambda) = b^T y - \tau\lambda - \sup_r \ \{y^T r - \|r\|_2\} - \sup_x \ \{y^T A x - \lambda\kappa(x)\}. \tag{2.3}$$

We recognize the first supremum as the conjugate function of $\|\cdot\|_2$, and the second supremum as the conjugate function of $\lambda\kappa(x)$. Because κ is a gauge function, its conjugate can be conveniently expressed as

$$\kappa^*(u) := \sup_w \ w^T u - \kappa(w) = \begin{cases} 0 & \text{if } \kappa^\circ(u) \leq 1 \\ \infty & \text{otherwise,} \end{cases} \tag{2.4}$$

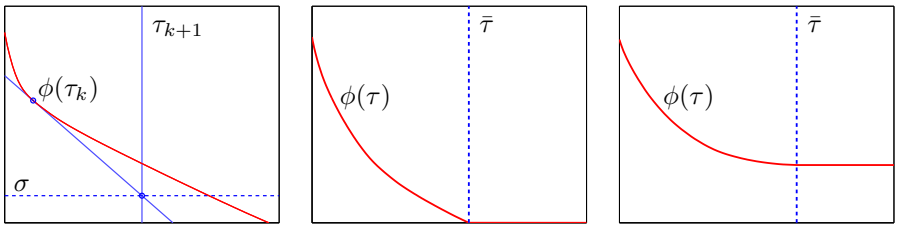
where the polar of κ is defined by

$$\kappa^\circ(u) = \sup_w \ \{w^T u \mid \kappa(w) \leq 1\}, \tag{2.5}$$

which is itself a gauge function [52, Theorem 15.1]. If κ is a norm, the polar reduces to the dual norm. It follows from substitution of (2.4) in (2.3) that the dual of (L_τ) is

$$\underset{y,\lambda}{\text{maximize}} \quad b^T y - \tau\lambda \quad \text{subject to} \quad \|y\|_2 \leq 1, \quad \kappa^\circ(A^T y) \leq \lambda. \tag{2.6}$$

Note that the constraint $\lambda \geq 0$ in (2.2) is implied by the nonnegativity of κ°.



(a) (b) (c)

Fig. 2.1: *(a) One Newton iteration on a typical Pareto curve for* (1.4)*; Pareto curves for (b)* $b \in \text{Range}(A)$ *and (c)* $b \notin \text{Range}(A)$.

2.2. Convexity and differentiability. Expressions for the optimal dual pair (y, λ) can be easily computed from the optimal primal solution. To derive y, first note from (2.4) that

$$\sup_r \; y^T r - \|r\|_2 = 0 \quad \text{if} \quad \|y\|_2 \le 1.$$

Therefore, $y = r/\|r\|_2$. In case $r = 0$, we can without loss of generality take $\|y\|_2 = 1$ in (2.6). To derive the optimal λ, note that as long as $\tau > 0$, λ must be at its lower bound $\kappa^\circ(A^T y)$, for otherwise we can increase the objective. Consequently,

$$\lambda = \kappa^\circ(A^T y) = \kappa^\circ(A^T r)/\|r\|_2, \tag{2.7}$$

where the last equality follows from positive homogeneity of κ°.

The following theorem describes the behavior of the Pareto curve over the interval on which it is decreasing. The quantity

$$\bar{\tau} = \min\{\tau \ge 0 \mid \phi(\tau) = \min_x \|Ax - b\|_2\}$$

gives the largest τ of interest; see Fig. 2.1(b,c).

THEOREM 2.1.
(a) *The function* ϕ *is convex and nonincreasing for all* $\tau \ge 0$.
(b) *For all* $\tau \in (0, \bar{\tau})$, ϕ *is continuously differentiable and*

$$\phi'(\tau) = -\kappa^\circ(A^T r_\tau)/\|r_\tau\|_2, \tag{2.8}$$

where $r_\tau := b - Ax_\tau$, *and* x_τ *is the optimal primal solution of* (L_τ).

Proof. (Part a) The fact that $\phi(\tau)$ is nonincreasing follows directly from the observation that the feasible set enlarges as τ increases. Next, consider any nonnegative scalars τ_1 and τ_2, and let x_1 and x_2 be the corresponding minimizers of (L_τ). For any $\beta \in [0, 1]$ define $x_\beta = \beta x_1 + (1 - \beta)x_2$, and note that by convexity of κ,

$$\kappa(x_\beta) = \kappa(\beta x_1 + (1 - \beta)x_2) \le \beta\kappa(x_1) + (1 - \beta)\kappa(x_2) = \beta\tau_1 + (1 - \beta)\tau_2.$$

With $\tau_\beta := \beta\tau_1 + (1 - \beta)\tau_2$ this gives $\kappa(x_\beta) \le \tau_\beta$, thus showing that x_β is a feasible point for (L_τ) with $\tau = \tau_\beta$. By the definition of ϕ,

$$\begin{aligned}
\phi(\tau_\beta) &\le \|Ax_\beta - b\|_2 \\
&= \|\beta(Ax_1 - b) + (1 - \beta)(Ax_2 - b)\|_2 \\
&\le \beta\|Ax_1 - b\|_2 + (1 - \beta)\|Ax_2 - b\|_2 \\
&= \beta\phi(\tau_1) + (1 - \beta)\phi(\tau_2),
\end{aligned}$$

as required for convexity of ϕ.

(Part b) Recall that differentiability of $\phi(\tau)$ corresponds to uniqueness of λ. This follows immediately from (2.7) combined with the uniqueness of the optimal r_τ in (2.1). Moreover, $\phi'(\tau) = -\lambda_\tau$, which yields (2.8). ∎

As the following theorem asserts, differentiability also holds when $\|r\|_2$ in (P_σ) and (L_τ) is replaced with a more general function $\rho(r)$. For a proof, see [3, Chapter 5].

THEOREM 2.2. *Let ρ and κ be gauge functions such that $\rho(r)$ is differentiable whenever $r \neq 0$. Then for all $\tau \in (0, \bar{\tau})$, ϕ is continuously differentiable and*

$$\phi'(\tau) = -\kappa^\circ(A^T y_\tau),$$

where $y_\tau := \arg\max_y \{b^T y - \tau\kappa^\circ(A^T y) \text{ subject to } \rho^\circ(y) \leq 1\}$.

2.3. Rationale for the gauge restriction. In Theorems 2.1 and 2.2 we assume κ to be a gauge function. This gives a sufficient condition for differentiability of ϕ, as well as a convenient expression for the dual problem. However, as the following example shows, this assumption is not necessary for differentiability. Consider the problem

$$\underset{x,r\in\mathbb{R}}{\text{minimize}} \quad \|r\|_2 \quad \text{subject to} \quad x + r = \beta,\ f(x) \leq \tau,$$

where $f : \mathbb{R} \to \mathbb{R}_+$ is a nonnegative convex function that vanishes at the origin and for which $0 \in \partial f(x)$ implies $x = 0$. Fix $\beta > 0$, and note that the solutions $x_\tau \geq 0$, $r_\tau = \beta - x_\tau$, and $\phi(\tau) := \|r\|_2 = \beta - x_\tau$. Because of the subdifferential assumption, the inverse mapping $f_+^{-1}(\tau) := \{x \geq 0 \mid f(x) = \tau\}$ for $\tau \geq 0$ is a well defined scalar, and on the interval $0 \leq \tau \leq \bar{\tau} := f_+^{-1}(\beta)$, we have $x_\tau = f_+^{-1}(\tau)$, and $\phi(\tau) = \beta - f_+^{-1}(\tau)$. Clearly, $\phi(\tau)$ is differentiable on the given interval if and only if $f_+^{-1}(\tau)$ is differentiable on the same interval. There is nothing in the definition of f that restricts it to be a gauge function, and we could for example choose $f(x) = x^2$ to obtain a differentiable ϕ. On the other hand, it is also easy to construct a function that is not differentiable on the chosen interval, leading to a non-differentiable ϕ.

3. Root finding with approximate functions and gradients. The Newton root-finding step in Algorithm 1 requires the computation of $\phi(\tau_k)$ and $\phi'(\tau_k)$. In practice these quantities may be too expensive to compute to high accuracy. In this section we quantify the local convergence properties of Algorithm 1 based on inaccurate solves of the subproblem (L_τ), which implies that the Newton iteration in Step 1 is replaced by

$$\tau_{k+1} \leftarrow \tau_k + \Delta_k, \quad \text{where} \quad \Delta_k := -(\bar{\phi}(\tau_k) - \sigma)/\bar{\phi}'(\tau_k), \tag{3.1}$$

and $\bar{\phi}$ and $\bar{\phi}'$ are approximations to ϕ and ϕ'.

3.1. Constructing function and gradient approximations. The algorithms for solving (L_τ) that we discuss in section 4 maintain feasibility at all iterations. As a result, an approximate solution $\bar{x}_\tau$ and its corresponding residual $\bar{r}_\tau := b - A\bar{x}_\tau$ satisfy

$$\kappa(\bar{x}_\tau) \leq \tau \quad \text{and} \quad \|\bar{r}_\tau\|_2 \geq \|r_\tau\|_2 > 0, \tag{3.2}$$

where the second set of inequalities holds because $\bar{x}_\tau$ is suboptimal and $\tau < \bar{\tau}$. We use the estimates $\bar{x}_\tau$ and $\bar{r}_\tau$ to construct dual feasible variables

$$\bar{y}_\tau := \bar{r}_\tau / \|\bar{r}_\tau\|_2 \quad \text{and} \quad \bar{\lambda}_\tau := \kappa^\circ(A^T \bar{y}_\tau);$$

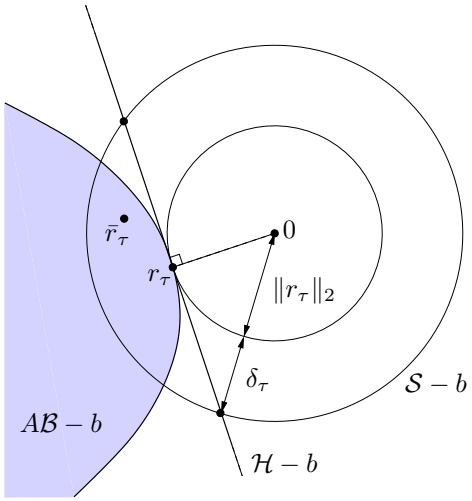


Fig. 3.1: *Illustration for proof of Lemma 3.2, coordinates shifted by $-b$.*

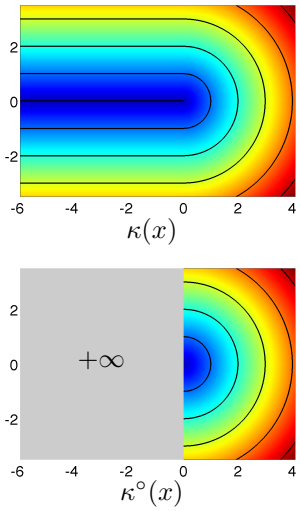


Fig. 3.2: *Gauge function with partially finite polar.*

c.f. (2.7). The objective of the dual problem (2.6), evaluated at any feasible point, gives a lower bound on the optimal value $\|r_\tau\|_2$. Therefore,

$$b^T\bar{y}_\tau - \tau\bar{\lambda}_\tau \le \|r_\tau\|_2 \le \|\bar{r}_\tau\|_2.$$

With the duality gap defined as

$$\delta_\tau := \|\bar{r}_\tau\|_2 - (b^T\bar{y}_\tau - \tau\bar{\lambda}_\tau), \tag{3.3}$$

we can bound the difference between the exact and approximate objective values by

$$0 \le \|\bar{r}_\tau\|_2 - \|r_\tau\|_2 \le \delta_\tau. \tag{3.4}$$

The approximate minimization of (L_τ) can then be used to construct approximations of ϕ and ϕ', which we define by

$$\bar{\phi}(\tau) = \|\bar{r}_\tau\|_2 \quad \text{and} \quad \bar{\phi}'(\tau) = -\bar{\lambda}_\tau. \tag{3.5}$$

Care must be taken to ensure that the multiplier estimate $\bar{\lambda}_\tau$ is always well defined. In particular, even though $\kappa^\circ(A^Ty_\tau)$ may be finite, arbitrarily close approximations $\bar{y}_\tau$ can result in $-\bar{\lambda}_\tau = \kappa^\circ(A^T\bar{y}_\tau) = \infty$. For example, consider the following gauge function $\kappa : \mathbb{R}^2 \to \mathbb{R}_+$ and its polar, which is illustrated in Fig. 3.2:

$$\kappa(w) = \begin{cases} |w_2| & \text{if } w_1 \le 0 \\ \|w\|_2 & \text{otherwise;} \end{cases} \qquad \kappa^\circ(u) = \begin{cases} \infty & \text{if } u_1 < 0 \\ \|u\|_2 & \text{otherwise.} \end{cases}$$

With $u = (0, 1)$, $\kappa^\circ(u) = 1$, but for $\bar{u} = (-\epsilon, 1)$ for any $\epsilon > 0$, $\kappa^\circ(\bar{u}) = \infty$. Hence it can be impossible to approximate the corresponding gradient when there is a direction of recession [8, section 1.4.1]. Such situations can arise whenever $\kappa(x) = 0$ does not imply $x = 0$. Similarly, it is possible for $\bar{\lambda}_\tau = 0$ for $\tau < \bar{\tau}$, even though $\lambda_\tau \neq 0$. This situation may arise when A is not full rank and $A^Tr = 0$, or when κ is not finite

everywhere, causing the polar to have a direction of recession. In the remainder of this section we preclude such cases.

ASSUMPTION 3.1. *(a) A has full rank; (b) κ is finite everywhere; and (c) κ has no directions of recession.*

Part (b) of the assumption implies that $\kappa^\circ(u) > 0$ for all $u \neq 0$, and combined with part (a), implies that $\bar{\phi}'(\tau) := -\bar{\lambda}_\tau \neq 0$ for all $\tau < \bar{\tau}$. It follows from part (c) that $\kappa^\circ(u)$, and therefore $\bar{\phi}'(\tau) := -\bar{\lambda}_\tau$, is finite for all u. These assumptions also imply the following useful bound:

$$L_{\min}\sigma_{\min}(A)\|q\| \leq \kappa^\circ(A^T q) \leq L_{\max}\sigma_{\max}(A)\|q\|_2, \tag{3.6}$$

where $L_{\min}$ and $L_{\max}$ are constants and $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ are the smallest and largest singular values of A.

3.2. Accuracy of the gradient. The duality gap δ_τ at $\bar{x}_\tau$ provides a bound on the difference between $\phi(\tau)$ and $\bar{\phi}(\tau)$. In this section we derive a similar bound on the difference between $\phi'(\tau)$ and $\bar{\phi}'(\tau)$ based on the relative duality gap

$$\eta_\tau = \delta_\tau / \|\bar{r}_\tau\|_2. \tag{3.7}$$

We first need the following bound.

LEMMA 3.2. *Let (x_τ, r_τ) be the optimal solution for (2.1). Then*

$$\frac{\|\bar{r}_\tau - r_\tau\|_2}{\|\bar{r}_\tau\|} \leq \sqrt{\eta_\tau^2 + 2\eta_\tau}.$$

Proof. Let the feasible set of (2.1) over x be denoted by $\mathcal{B} := \{x \mid \kappa(x) \leq \tau\}$. The range of Ax over all feasible x is given by $A\mathcal{B}$, which is convex because convexity is preserved under linear maps. It follows that a unique separating hyperplane $\mathcal{H}$, with normal r_τ, exists between $A\mathcal{B}$ and the Euclidean ball of radius $\|r_\tau\|_2$ centered at b; see Fig. 3.1. Let $\mathcal{S} = \{b + r \mid \|r\|_2 = \|r_\tau\|_2 + \delta_\tau\}$ denote a sphere centered around b. The distance $\|r_\tau - \bar{r}_\tau\|_2$ can be seen to be bounded by the maximum distance between Ax_τ and the points in $\mathcal{S} \cap A\mathcal{B}$, which is itself bounded by the norm of the difference d between Ax_τ and any point on the intersection between $\mathcal{S}$ and the separating hyperplane $\mathcal{H}$. For the norm of d we have $(\|r_\tau\|_2 + \delta_\tau)^2 = \|r_\tau\|_2^2 + \|d\|_2^2$ and the stated result follows. Thus,

$$\|r_\tau - \bar{r}_\tau\|_2 \leq \|d\|_2 \leq \sqrt{\delta_\tau^2 + 2\|r_\tau\|_2\delta_\tau} \leq \sqrt{\delta_\tau^2 + 2\|\bar{r}_\tau\|_2\delta_\tau},$$

where the last inequality follows from $\|r_\tau\|_2 \leq \|\bar{r}_\tau\|_2$. We obtain the required inequality by dividing both sides by $\|\bar{r}_\tau\|$ and using the definition of η_τ. ∎

We use this result to derive the required bound on the difference between the exact and approximate gradients.

LEMMA 3.3. *If Assumption 3.1 holds, there exists a constant $c > 0$, independent of τ, such that*

$$|\bar{\phi}'(\tau) - \phi'(\tau)| \leq c \cdot \sqrt{\eta_\tau^2 + 2\eta_\tau}.$$

Proof. We consider two cases. In the first, suppose that $\bar{\phi}'(\tau) \leq \phi'(\tau)$. With the definition of $-\phi'(\tau)$ this gives

$$\begin{aligned}\phi'(\tau) - \bar{\phi}'(\tau) &= \frac{\kappa^\circ(A^T\bar{r}_\tau)}{\|\bar{r}_\tau\|_2} - \frac{\kappa^\circ(A^T r_\tau)}{\|r_\tau\|_2} \leq \frac{\kappa^\circ(A^T\bar{r}_\tau) - \kappa^\circ(A^T r_\tau)}{\|\bar{r}_\tau\|_2} \leq \frac{\kappa^\circ(A^T[\bar{r}_\tau - r_\tau])}{\|\bar{r}_\tau\|} \\ &\leq L_{\max}\sigma_{\max}(A)\frac{\|\bar{r} - r\|_2}{\|\bar{r}\|_2} \leq L_{\max}\sigma_{\max}(A)\sqrt{\eta_\tau^2 + 2\eta_\tau},\end{aligned}$$

where the second-to-last inequality follows from (3.6), and the last inequality follows from Lemma 3.2. For the second case we consider $\phi'(\tau) \leq \bar{\phi}'(\tau)$. This gives

$$\begin{aligned}\bar{\phi}'(\tau) - \phi'(\tau) &= \frac{\kappa^\circ(A^T r_\tau)}{\|r_\tau\|_2} - \frac{\kappa^\circ(A^T\bar{r}_\tau)}{\|\bar{r}_\tau\|_2} \leq \frac{\kappa^\circ(A^T r_\tau)}{\|r_\tau\|_2} - \frac{\kappa^\circ(A^T r_\tau)}{\|\bar{r}_\tau\|_2} + \frac{\kappa^\circ(A^T[r_\tau - \bar{r}_\tau])}{\|\bar{r}_\tau\|_2} \\ &\leq \frac{\kappa^\circ(A^T r_\tau)}{\|r_\tau\|_2} \cdot \frac{\|\bar{r}_\tau\|_2 - \|r_\tau\|_2}{\|\bar{r}_\tau\|_2} + \frac{\kappa^\circ(A^T[r_\tau - \bar{r}_\tau])}{\|\bar{r}_\tau\|_2} \\ &\leq L_{\max}\sigma_{\max}(A)\left[\frac{\|\bar{r}_\tau - r_\tau\|}{\|\bar{r}_\tau\|_2} + \frac{\|\bar{r}_\tau - r_\tau\|}{\|\bar{r}_\tau\|_2}\right] \leq 2L_{\max}\sigma_{\max}(A)\sqrt{\eta_\tau^2 + 2\eta_\tau}.\end{aligned}$$

The first inequality follows from the reverse triangle inequality. We then use (3.6) and Lemma 3.2. ∎

3.3. Local convergence rate. With the aid of Lemma (3.3) we establish the following local convergence result.

THEOREM 3.4. *Suppose Assumption 3.1 holds and that the iteration defined by* (3.1) *and* (3.5) *converges to* $\tau^* \in (0, \bar{\tau})$. *Then* $\tau_k \to \tau^*$ *superlinearly if* $\lim_{k\to\infty} \eta_k = 0$.

Proof. Define $\delta_k = \delta_{\tau_k}$, $\eta_k = \eta_{\tau_k}$, $\phi_k = \phi(\tau_k)$, etc., and recall that $\phi_k = \|r_{\tau_k}\|_2$ and $\bar{\phi}_k = \|\bar{r}_{\tau_k}\|_2$ (c.f. (3.5).) Because Δ_k is an approximate Newton step, there is a corresponding error r_k in the exact Newton equation, i.e., ϵ_k satisfies

$$\phi_k' \Delta_k = -(\phi_k - \sigma) + \epsilon_k.$$

The superlinear convergence rate follows from Corollary 3.5 of Dembo et al. [22], where it is required that $|\epsilon_k|/|\phi_k| \leq \gamma_k$, for some sequence $\gamma_k \to 0$. To this end, note that

$$\begin{aligned}|\epsilon_k| &= \left|(\phi_k - \sigma) - \phi_k' \frac{\bar{\phi}_k - \sigma}{\bar{\phi}_k'}\right| \\ &= \left|(\phi_k - \sigma) - \left(\frac{\bar{\phi}_k'}{\bar{\phi}_k'} + \frac{\phi_k' - \bar{\phi}_k'}{\bar{\phi}_k'}\right)(\bar{\phi}_k - \sigma)\right| \\ &\leq |\phi_k - \bar{\phi}_k| + |\phi_k' - \bar{\phi}_k'| \cdot \frac{|\bar{\phi}_k - \sigma|}{|\bar{\phi}_k'|}.\end{aligned}$$

For all k sufficiently large, $\frac{1}{2} > \eta_k \equiv \delta_k/\bar{\phi}_k$, and therefore $\frac{1}{2}\bar{\phi}_k > \delta_k$. It then follows from (3.4) that $\phi_k > \frac{1}{2}\bar{\phi}_k$. Finally, because $\tau_k \to \tau^*$ and $\phi(\tau_*) = \sigma$, continuity of ϕ implies that there exists constants $0 < c_1 \leq c_2$ such that $c_1 \leq \phi_k \leq c_2$ for all sufficiently large k. Hence, using Lemma 3.3,

$$\frac{|\epsilon_k|}{\phi_k} < \frac{2|\phi_k - \bar{\phi}_k|}{\bar{\phi}_k} + \frac{c\sqrt{\eta_k^2 + 2\eta_k}}{\phi_k} \cdot \frac{|\bar{\phi}_k - \sigma|}{|\bar{\phi}_k'|} < 2\eta_k + \frac{c\sqrt{\eta_k^2 + 2\eta_k}}{c_1} \cdot \frac{2c_2 + \sigma}{L_{\min}\sigma_{\min}(A)}.$$

The theorem then follows by setting γ_k equal to the right-most term above inequality, which clearly goes to zero as $\eta_k \to 0$. ∎

3.4. Practical aspects. Assumption 3.1 excludes functions κ that takes infinite values. This unfortunately excludes a useful class of functions—such as those described in section 7—from our analysis. A similar result could be obtained, however, by modifying the iterations to require η_k to be sufficiently small before taking a Newton step. In particular, Lemma 3.2 guarantees that the approximation $\bar{\phi}'(\tau)$ will be nonzero if η_k is chosen sufficiently small. This could be easily enforced in practice.

Whenever κ has no direction of recession, $\phi(0)$ and $\phi'(0)$ are available in closed form. This implies that a natural choice for the initial value of τ is $\tau_0 = 0$. In particular, the constraint $\kappa(x) \leq \tau_0 = 0$ implies the trivial solution $x_\tau = 0$ of (L_τ). It follows from the definition of ϕ that $\phi(0) = \|Ax_\tau - b\|_2 = \|b\|_2$. Moreover, (2.8) in Theorem 2.1 implies that $\phi'(0) = -\kappa^\circ(A^T b)/\|b\|_2$. Thus, a first root-finding step can be obtained without solving (L_τ), and requires only a single matrix-vector product with A^T and one evaluation of κ°.

4. Implementation. Algorithm 1 for solving (P_σ) requires the solution of a sequence of problems (L_τ). Although the method does not prescribe a particular subproblem solver, the performance of the overall approach crucially depends on the efficiency of the method used to solve the subproblems. We have implemented two different methods for doing so: the spectral projected-gradient algorithm [10], which is implemented as part of the SPGL1 software package [5], and the projected limited-memory quasi-Newton approach developed in [53]. A third approach, recently proposed by Gu et al. [34], applies Nesterov's accelerated proximal-gradient method [50] to the special case where $\kappa(x) = \|x\|_1$.

The root-finding framework has been implemented in the software package SPOR. Two different approaches to solving the subproblems are included: SPOR-SPG and SPOR-PQN use the spectral projected-gradient and quasi-Newton algorithms, respectively, for the subproblems. For comparison, SPOR-SPG-H is a modified version of SPOR-SPG that implements looser subproblem stopping criteria. These algorithms are used in our numerical experiments in sections 5–8.

Each of these algorithms requires a method to compute the orthogonal projection of an iterate x_k onto the feasible region, i.e.,

$$\arg\min_x \ \|x_k - x\|_2 \ \text{ subject to } \ \kappa(x) \leq \tau, \tag{4.1}$$

which is parameterized by the current value of τ. The exact implementation of the projection is again irrelevant to the algorithms, as long as it can be done efficiently. For each of the applications discussed in sections 5–9, we provide efficient algorithms for solving the projection problems that correspond to a particular function κ; each is based on solving the 1-norm projection problem

$$\mathcal{P}_\tau(\bar{x}) := \arg\min_x \ \|\bar{x} - x\|_2 \ \text{ subject to } \ \|x\|_1 \leq \tau.$$

This projection can be obtained with an $\mathcal{O}(n \log n)$ algorithm. A randomized $\mathcal{O}(n)$ algorithm is also available [7, 29].

5. Weighted basis pursuit denoise. In this section we apply the root-finding framework to the weighted basis pursuit denoise problem

$$\underset{x}{\text{minimize}} \quad \kappa(x) := \|Wx\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \sigma, \tag{5.1}$$

where W is an $n \times n$ diagonal matrix with nonnegative elements; this corresponds to (P_σ) with $\kappa(x) = \|Wx\|_1$. The most prominent use of this formulation (with $\sigma = 0$)

is the reweighted ℓ_1 algorithm [18] which, compared to basis pursuit, yields a higher probability of recovering x_0 from measurement $b = Ax_0$. In the following sections we describe the required ingredients: the dual norm, and a method for projection onto the weighted 1-norm ball. Numerical experiments are given in section 5.3.

5.1. Polar function. For the dual of the weighted norm $\|Wx\|_1$ we use the following, more general, result:

THEOREM 5.1. *Let κ be a gauge function, and $\kappa_\Phi(w) := \kappa(\Phi w)$ where the linear operator Φ is invertible. Then $\kappa_\Phi^\circ(u) = \kappa^\circ(\Phi^{-T}u)$.*

Proof. Applying the definition of gauge polars (2.5) to κ_Φ gives

$$\kappa_\Phi^\circ(u) = \sup_w \{w^T u \mid \kappa_\Phi(w) \le 1\} = \sup_w \{w^T u \mid \kappa(\Phi w) \le 1\}. \tag{5.2}$$

Now define $v = \Phi w$. Invertibility of Φ implies

$$\kappa_\Phi^\circ(u) = \sup_v \{(\Phi^{-1}v)^T u \mid \kappa(v) \le 1\} = \sup_v \{v^T \Phi^{-T} u \mid \kappa(v) \le 1\} = \kappa^\circ(\Phi^{-T}u),$$

as required. ∎

The following result follows immediately.

COROLLARY 5.2. *Let W be any nonsingular diagonal matrix. Then the dual of $\|Wx\|_1$ is given by $\|W^{-1}x\|_\infty$.*

5.1.1. Dealing with zero weights. Whenever W has a zero diagonal entry, the supremum in (5.2) is infinite whenever the corresponding entry in u is nonzero. As a result, the dual of (5.1), which is given by (2.6) with the constraint $\kappa^\circ(A^T y) \le \tau$ replaced with $\|W^{-T}A^T y\|_\infty \le \tau$, is not well defined.

Instead of (5.1), we therefore consider the problem

$$\underset{r,x}{\text{minimize}} \quad \|r\|_2 \quad \text{subject to} \quad A_1 x_1 + A_2 x_2 + r = b, \quad \|W_1 x_1\|_1 \le \tau,$$

where $x = (x_1, x_2)$ and $A = [A_1, A_2]$ are (possibly reordered) partitions of x and A, and W_1 corresponds to the nonzero weights in the original W. The corresponding Lagrange-dual function is given by

$$\mathcal{L}(y, x) = b^T y - \tau\lambda - \sup_r \{y^T r - \|r\|_2\} - \sup_{x_1} \{y^T A_1 x_1 - \lambda \|W_1 x_1\|_1\} - \sup_{x_2} \{y^T A_2 x_2\}.$$

We can immediately see that in order for the supremum over x_2 to be finite (zero in fact), we require $A_2^T y = 0$. Theorem 2.2 continues to hold, except that the derivative of the Pareto curve is given by $\phi'(\tau) = -\|W^{-T}A_1^T y\|_\infty$.

Care must be taken when computing the initial iterate of the root-finding method, as described in section 3.4. In particular, if zero weights are present, the solution of (L_τ) with $\tau = 0$ does not trivially yield a solution $x = 0$, and in order to compute $\phi(0)$ and $\phi'(0)$ it is necessary to solve the least-squares problem

$$\underset{x_2}{\text{minimize}} \quad \|r\|_2 \quad \text{subject to} \quad A_2 x_2 + r = b.$$

5.2. Projection. The projection onto a diagonally weighted one-norm ball, $\{x \mid \|Wx\|_1 \le \tau\}$, with $W = \operatorname{diag}(w)$, is defined by (4.1), with $\kappa(x) = \|Wx\|_1$. It is easily seen that for each i with $w_i = 0$, the corresponding component of the projection is given by $\bar{x}_i$. Therefore, without loss of generality, we can assume that $w_i \neq 0$. We only need to consider the case where $\tau > 0$. (Otherwise, the projection is trivially zero.)

We derive an efficient projection algorithm starting with the following result on the Lagrangian formulation of the projection problem.

THEOREM 5.3. *For fixed $\bar{x}$, $\lambda \ge 0$, and diagonal W,*

$$x(\lambda) := \arg\min_x \ \tfrac{1}{2}\|\bar{x} - x\|_2^2 + \lambda\|Wx\|_1 = \operatorname{sgn}(\bar{x}) \cdot \max\{0, |\bar{x}| - \lambda|w|\}, \qquad (5.3)$$

where the operators in the last expression are taken componentwise.

Proof. For $x(\lambda)$ to be a solution of optimization problem in (5.3), we require that the subgradient of the objective contains zero. Since the objective is separable, this reduces to

$$0 \in \bar{x} - x(\lambda) + \lambda \cdot |w| \cdot \operatorname{sgn}(\bar{x}).$$

It can be verified that this is satisfied by setting $x(\lambda) := \operatorname{sgn}(\bar{x}) \cdot \max\{0, |\bar{x}| - \lambda|w|\}$. ∎

Using this closed-form solution, we can derive an algorithm for computing the Lagrange multiplier λ of (4.1) which yields $x(\lambda)$, i.e., the projection of $\bar{x}$ on the weighted 1-norm ball with radius τ. Note that λ is the solution of the equation

$$f(\lambda) = \tau \quad \text{with} \quad f(\lambda) := \|Wx(\lambda)\|_1, \qquad (5.4)$$

where f is piecewise linear. From (5.3) and (5.4) we see that (a) $f(\lambda)$ is strictly decreasing in λ, until $f(\lambda) = 0$; (b) whenever $\lambda \ge |\bar{x}_i/w_i|$, we have $x(\lambda)_i = 0$, and that we no longer need to consider $\bar{x}_i$ for those values; (c) the function $f(\lambda)$ is linear on intervals $|\bar{x}_{[i]}/w_{[i]}| \le \lambda \le |\bar{x}_{[i+1]}/w_{[i+1]}|$, where the bracketed subscripts denote an ordering on the entries of $\bar{x}$ and w such that $\{|\bar{x}_{[i]}/w_{[i]}|\}_i$ is a non-decreasing sequence.

Let $\lambda_i := |\bar{x}_{[i]}/w_{[i]}|$. The projection algorithm proceeds by first finding the smallest index i such that $f(\lambda_{i-1}) > \tau \ge f(\lambda_i)$, with $\lambda_0 \equiv 0$. Within this interval, $f(\lambda)$ is linear with slope $-\sum_{j=1}^{i} w_{[j]}^2$. Solving $f(\lambda) = \tau$ then gives $\lambda^* = \lambda_i - (\tau - f(\lambda_i))/\sum_{j=1}^{i} w_{[j]}^2$. We then apply (5.3) with λ^* to obtain $x(\lambda)$.

The resulting projection algorithm generalizes the usual 1-norm projection $\mathcal{P}_\tau$, and has the same computationaly complexity.

5.3. Experiments. We now turn to the performance evaluation of the SPOR-SPG and SPOR-PQN algorithms on a series of weighted basis pursuit denoise problem (5.1). Each problem consists of a randomly generated $20j \times 40j$ matrix A, with integer scaling factor j, and random W and b. For the weight matrix we have $W_{i,i} \sim 1 + |\mathcal{N}(0,1)|$, where $\mathcal{N}(0,1)$ denotes the standard normal distribution. The vector b is defined elementwise by $b_i \sim [Ax_0]_i + 0.1 \cdot \mathcal{N}(0,1)$, where x_0 is an s-sparse vector with non-zero entries randomly drawn from the normal distribution.

We compare the performance against SDPT3 [61] though the CVX interface [33]. Preliminary tests show that it proves difficult to get extremely accurate solutions with either SPOR-SPG or SPOR-PQN. Therefore, to make a fair comparison with the potentially highly accurate SDPT3, we set the CVX parameter `cvx_precision low`, which asks for a relative accuracy of 10^{-4}.

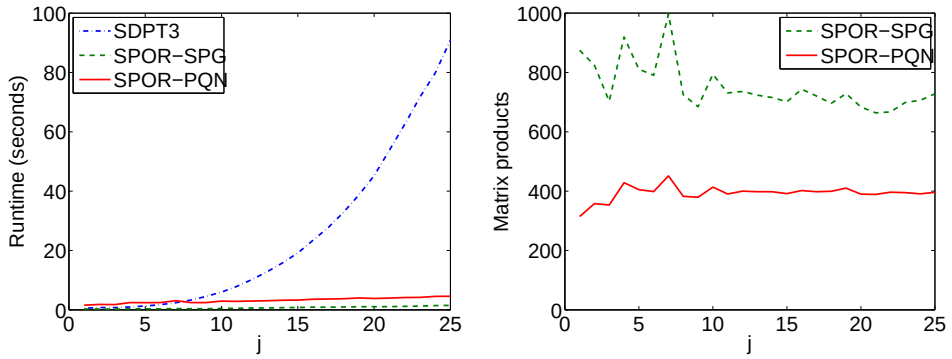


Fig. 5.1: *Performance of SDPT3, SPOR-SPG, and SPOR-PQN on weighted basis pursuit problems with random* $20j \times 40j$ A *and* $4j$*-sparse* x_0*; (a) runtime versus scale, and (b) number of matrix-vector products with* A *or* A^T *versus scale.*

For all scaling factors $j = 1, \ldots, 25$, sparsity level $4j$, and $\sigma = 0.02 \cdot \|b\|_2$ we plot the resulting runtime in Fig. 5.1(a). From this plot it is apparent that SDPT3 does not scale well with problem size. SPOR-SPG is about twice as fast as SPOR-PQN, even though it requires approximately twice as many matrix-vector products with A and A^T; see Fig. 5.1(b). However, the runtime of SPOR-PQN increases at a slightly lower rate than that of SPOR-SPG, and therefore SPOR-PQN should prove more efficient for problems where multiplications with A are computationally expensive.

Closer inspection of the numerical results (see online appendix, section 1.5) reveals that the performance of SPOR-SPG and SPOR-PQN depends largely on σ, and whether or not b has a sparse represention x_0. When there is no sparse representation and σ is small, SPOR-SPG and SPOR-PQN generally require much more time to complete than SDPT3, which is insensitive to the values of σ and b.

6. Group sparsity. In this section we develop the theory required to apply the root-finding framework to solve the sum-of-norms problem (1.5). This corresponds to (P_σ) with $\kappa(x) = \|X\|_{1,2}$. We first study a practical application of this problem to source localization. In the signal-processing context, (1.5) is known as the multiple measurement vector (MMV) problem [6].

6.1. Application – source localization. When a radio telescope is aimed at a source emitting plane waves, signals reflected on the dish are all focussed on the receiver in phase thus leading to an amplified signal; see Fig. 6.1. A similar amplification can be accomplished with an array of omnidirectional sensors by summing the sensor outputs by either delaying the outputs received with each sensor, or applying appropriate phase shifts. Let $S_{i,j}$ represent a set of narrowband signals arriving from angles θ_i at time t_j, where $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Under the narrowband assumption, and provided the spacing between sensors is sufficiently small, the sensor output can be formulated as

$$B = A(\theta)S + N,$$

where N is a matrix of random additive noise, and $A(\theta)$ is the phase shift/gain matrix with each column corresponding to a single arrival angle θ_i, and the number of rows equal to the number of sensors, say m.

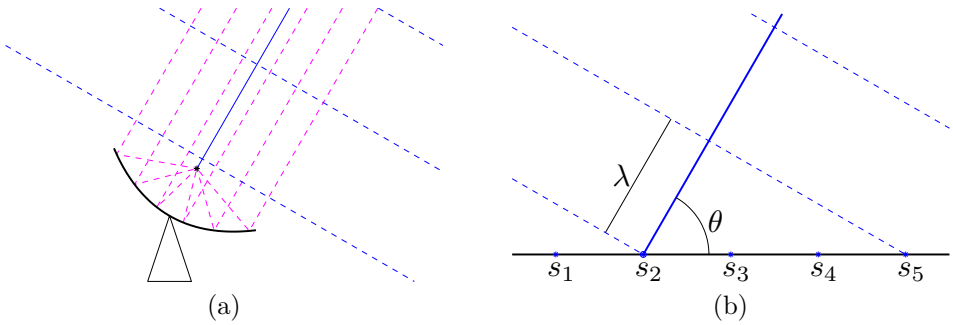


Fig. 6.1: *Focussing planar waves from a given direction using (a) a radio telescope, (b) an array of omnidirectional sensors.*

In the two-dimensional case, with sensor positions given by p_i, $i = 1, \ldots, m$, the complex phase shift for sensor i, relative to the origin, for angle θ_j, is given by

$$A_{i,j} = \exp\{2\imath\pi \cos(\theta_j) p_i/\lambda\},$$

with $\imath = \sqrt{-1}$ and wavelength λ.

In source localization problem, the angles θ, and thus the matrix $A(\theta)$, are unknown. When the number of sources is small, we can discretize the space into a set of angles ψ and find a sparse approximate solution to the linear system

$$A(\psi)X = B.$$

Assuming sources are stationary or moving slowly with respect to the observation time, we would like the nonzero entries in X (corresponding to different angles of arrival) to be restricted to a small number of rows. The approach taken by Malioutov et al. [44] amounts exactly to the MMV problem (1.5). Note that the misfit between $A(\psi)X$ and B in this case is due not only to the signal noise N, but also to error in the angular discretization ψ.

For a concrete example, consider an array of twenty omnidirectional sensors spaced at half the wavelength of interest. Impinging on this array are five far-field sources, located at angles 60°, 65°, 80°, 100.5°, and 160° relative to the horizon. The array records twenty samples at a signal-to-noise ratio of 10dB.

We recover the directions of arrival by discretizing the space at an angular resolution of 1° and compare the results obtained via MMV (1.5) and BPDN (1.4) against those obtained via beamforming, Capon, and MUSIC (see [45] for more information). The resulting powers from each possible direction of arrival are shown in Fig. 6.2. Note that the MMV formulation results in the best predictions.

For a more realistic example we next consider a three-dimensional source localization problem. Because of the added dimension, discretizing the space of all possible directions and positioning of sensors becomes somewhat harder. We want to locate the sensors within a unit circle on a two-dimensional plane, so that they are near-uniformly spread. This is conveniently done using existing circle-packing results [55]. For the discretization of arrival directions we associate with each direction a set of points p_i on the unit sphere and approximately minimize the potential energy $\sum_{i \neq j} 1/\|p_i - p_j\|_2$. The final result is then obtained by discarding the points in one halfspace. The discretizations obtained for 80 sensors and 100 directions are shown in Fig. 6.3, along with a signal coming from eight directions. Given a set of such

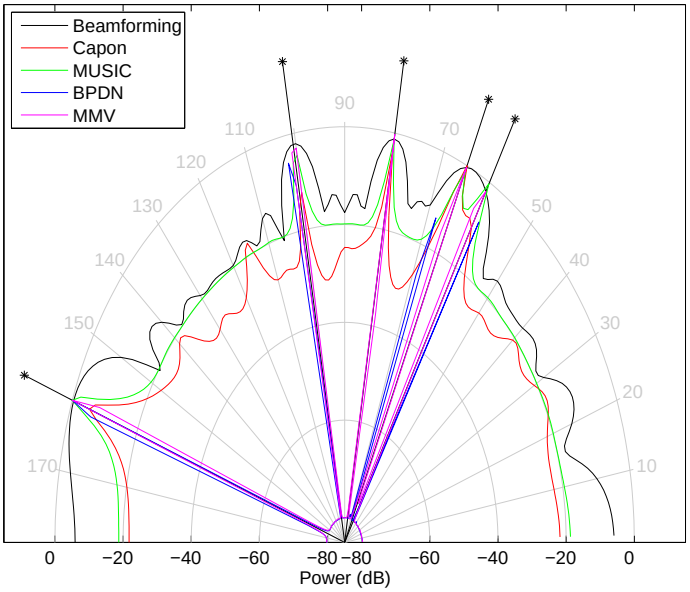


Fig. 6.2: *Angular spectra obtained using beamforming, Capon, MUSIC, BPDN, and MMV for five uncorrelated far-field sources at angles 60°, 65°, 80°, 100.5°, and 160°. Direction of arrival discretized at 1° resolution.*

signals we then run MMV and obtain an approximate solution on the coarse grid (see Fig. 6.4(a,e)). Because the grid is unlikely to exactly coincide with the exact directions there is some uncertainty in the direction. This can be reduced by locally refining the grid and repeating reconstruction until the desired results it reached. This process is shown in Fig. 6.4(b–d,f–h).

6.2. Polar function. The $\|\cdot\|_{p,q}$ norm defined in (1.6) is a special case of the general sum-of-mixed norms. The following results gives the duals for the general case.

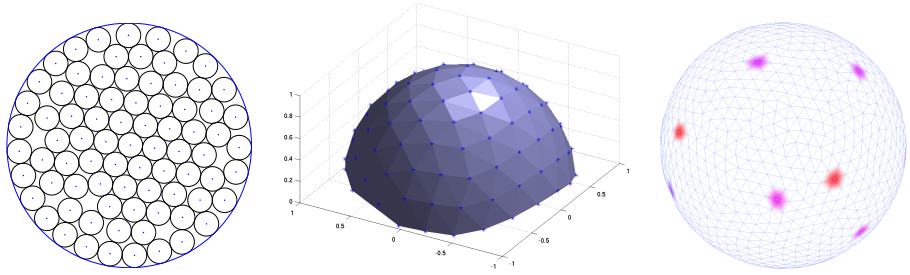
THEOREM 6.1. *Let σ_i, $i = 1, \ldots, k$ represent disjoint index sets such that $\bigcup_i \sigma_i = \{1, \ldots, n\}$. Associate with each group i a primal norm $\|\cdot\|_{p_i}$ and dual norm $\|\cdot\|_{d_i}$, and denote $v_i(x) = \|x_{\sigma_i}\|_{p_i}$ and $w_i(x) = \|x_{\sigma_i}\|_{d}$. Let $\|\cdot\|_p$ be a norm such that for all vectors $s, t \geq 0$, $\|s\|_p \leq \|s+t\|_p$, and let $\|\cdot\|_d$ denote its dual norm. Then the dual norm of $\|\!|\cdot|\!\|_p := \|v(\cdot)\|_p$ is given by $\|\!|\cdot|\!\|_d := \|w(\cdot)\|_d$.*

Proof. First we need to show that $\|\!|x|\!\|_p$ is indeed a norm. It is easily seen that the requirements $\|\!|x|\!\|_p \geq 0$, $\|\!|\alpha x|\!\|_p = |\alpha| \cdot \|\!|x|\!\|_p$, and $\|\!|x|\!\|_p = 0 \iff x = 0$ hold. For the triangle inequality we need to show that $\|\!|s+t|\!\|_p \leq \|\!|s|\!\|_p + \|\!|t|\!\|_p$. Using the triangle inequality on the outer norms $\|\cdot\|_p$, we have that $0 \leq v(s+t) \leq v(s) + v(t)$, componentwise. The assumption on the outer norm $\|\cdot\|_p$, then allows us to write

$$\|\!|s+t|\!\|_p = \|v(s+t)\|_p \leq \|v(s) + v(t)\|_p \leq \|v(s)\|_p + \|v(t)\|_p = \|\!|s|\!\|_p + \|\!|t|\!\|_p,$$

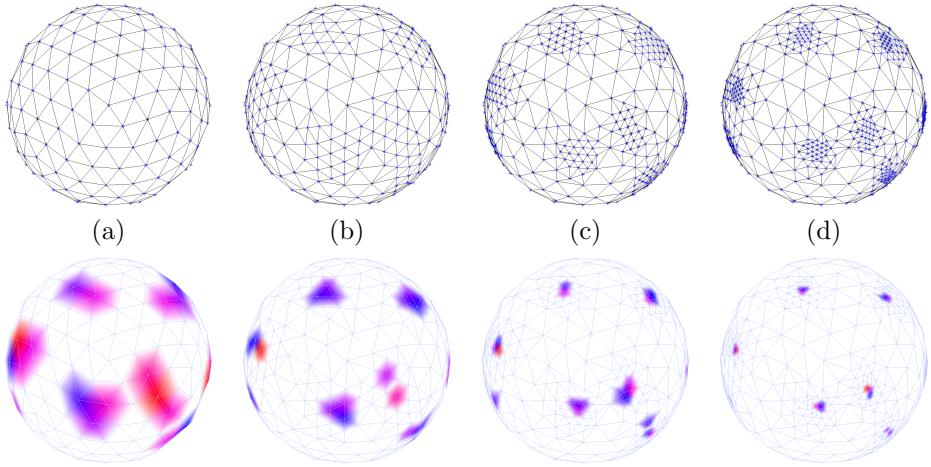
as desired. Next, to derive the dual norm we note that the dual of any norm is defined implicitly by the following equation:

$$\|x\|_d := \sup_z \{x^T z \mid \|z\|_p \leq 1\}.$$



(a) (b) (c)

Fig. 6.3: *Configuration of (a) 80 sensors on the plane, (b) coarse grid of 100 arrival directions on the half sphere, and (c) top view of actual arrival directions on a fine grid.*



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6.4: *Grid of (a) initial arrival directions, and (b–d) after up to three refinement stepts, along with corresponding solutions (e–f).*

For each given subvector x_{σ_i}, the supremum of $x_{\sigma_i}^T z_{\sigma_i}$ with $\|x_{\sigma_i}\|_{p_i} \leq 1$ is given by $w_i(x) := \|x_{\sigma_i}\|_d$. This quantity scales linearly with the bound we impose on the primal norm, i.e., under the condition $\|x_{\sigma_i}\|_{p_i} \leq t_i$, the supremum becomes $t_i w_i(x)$. Writing $w = \{w_i(x)\}_i$, we can write the supremum over the entire vector as

$$\begin{aligned}\sup_z \{x^T z \mid \|z\|_p \leq 1\} &= \sup_{t,z} \left\{ \textstyle\sum_i t_i x_{\sigma_i}^T z_{\sigma_i} \mid \|t\|_p \leq 1,\ \|z_{\sigma_i}\|_{p_i} \leq 1 \right\} \\ &= \sup_t \{t^T w \mid \|t\|_p \leq 1\}.\end{aligned}$$

But this is exactly the definition of $\|w\|_d$, as desired. ∎

Note that the requirement on the outer primal norm $\|\cdot\|_p$ is essential in deriving the triangle inequality, but does not hold for all norms. For example, the norm $\|x\| := \|\Phi x\|_2$, with any non-diagonal, invertible matrix Φ does not satisfy the requirement. Importantly though, the requirement is satisfied for the common norms $\|x\|_\gamma$, $1 \leq \gamma \leq \infty$;

By repeated application of the above theorem we can derive the dual of arbitrarily nested norms. For example, the function $\|x_{(1)}\|_2 + \max\{\|x_{(2)}\|_1, \|x_{(3)}\|_2\}$

applied to a vector consisting of $x_{(1)}, x_{(2)}, x_{(3)}$ is a norm whose dual is given by $\max\{\|x_{(1)}\|_2, \|x_{(2)}\|_\infty + \|x_{(3)}\|_2\}$. Likewise, by vectorizing X and imposing appropriate groups, we can use Theorem 6.1 to obtain the dual of $\|X\|_{p,q}$:

COROLLARY 6.2. *The dual of $\|X\|_{p,q}$ with $p, q \geq 1$, is given by $\|X\|_{p',q'}$, where p' and q' are such that $1/p + 1/p' = 1$ and $1/q + 1/q' = 1$.*

6.3. Projection. Orthogonal projection onto general (p, q)-norm balls requires the solution of an optimization problem with quadratic objective over the feasible set. In the special case of Euclidean projection onto the balls induced by the $\|X\|_{1,2}$ and $\sum_i \|x_{(i)}\|_2$ norms, we can use the following theorem.

THEOREM 6.3. *Let $c_{(i)}$ be a set of vectors, possibly of different length. Then the solution $x^* = (x^*_{(1)}, \ldots, x^*_{(k)})$ of*

$$\underset{x}{\text{minimize}} \quad \sum_i \tfrac{1}{2}\|c_{(i)} - x_{(i)}\|_2^2 \quad \text{subject to} \quad \textstyle\sum_i \|x_{(i)}\|_2 \leq \tau, \tag{6.1}$$

is given by

$$x^*_{(i)} = \begin{cases} (u^*_i/v_i) \cdot c_{(i)} & \text{if } v_i \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $v_i = \|c_{(i)}\|_2$ and $u^* = \mathcal{P}_\tau(v)$ is the 1-norm projection of v.

Proof. We first treat the special case where $v_i = 0$. The projection for these groups is trivially given by $x_{(i)} = c_{(i)} = 0$, thus allowing us to exclude these groups. Next, rewrite (6.1) as

$$\underset{x,u}{\text{minimize}} \quad \sum_i \tfrac{1}{2}\|c_{(i)} - x_{(i)}\|_2^2 \quad \text{subject to} \quad \|x_{(i)}\|_2 \leq u_i, \ \|u\|_1 \leq \tau. \tag{6.2}$$

Fixing $u = u^*$ makes the problem separable, reducing the problem for each i to

$$\underset{x_{(i)}}{\text{minimize}} \quad \tfrac{1}{2}\|c_{(i)} - x_{(i)}\|_2^2 \quad \text{subject to} \quad \|x_{(i)}\|_2^2 \leq u_i^2.$$

For $u_i = 0$ this immediately gives $x_{(i)} = 0$. Otherwise the first-order optimality conditions on x require that the gradient of the Lagrangian,

$$\mathcal{L}(\lambda_i) = \tfrac{1}{2}\|c_{(i)} - x_{(i)}\|_2^2 + \lambda(\|x_{(i)}\|_2^2 - u_i^2),$$

with $\lambda \geq 0$, be equal to zero; that is, $\nabla\mathcal{L}(x_{(i)}) = x_{(i)} - c_{(i)} + 2\lambda_i x_{(i)} = 0$. It follows that $x_{(i)} = c_{(i)}/(1 + 2\lambda_i) = \gamma_i c_{(i)}$, such that $\|x_{(i)}\|_2 = \gamma\|c_{(i)}\|_2 = u_i$ (which also holds for $u_i = 0$). Using the definition $v_i = \|c_{(i)}\|_2$, and the fact that $x_{(i)} = \gamma_i c_{(i)}$, we can rewrite each term of the objective of (6.1) as

$$\begin{aligned} \|c_{(i)} - x_{(i)}\|_2^2 &= c_{(i)}^T c_{(i)} - 2\gamma_i c_{(i)}^T c_{(i)} + \gamma_i^2 c_{(i)}^T c_{(i)} \\ &= \|c_{(i)}\|_2^2 - 2\gamma_i\|c_{(i)}\|_2^2 + \gamma_i^2\|c_{(i)}\|_2^2 \\ &= v_i^2 - 2\gamma_i v_i^2 + \gamma_i^2 v_i^2 = (v_i - \gamma_i v_i)^2 = (v_i - u_i)^2. \end{aligned}$$

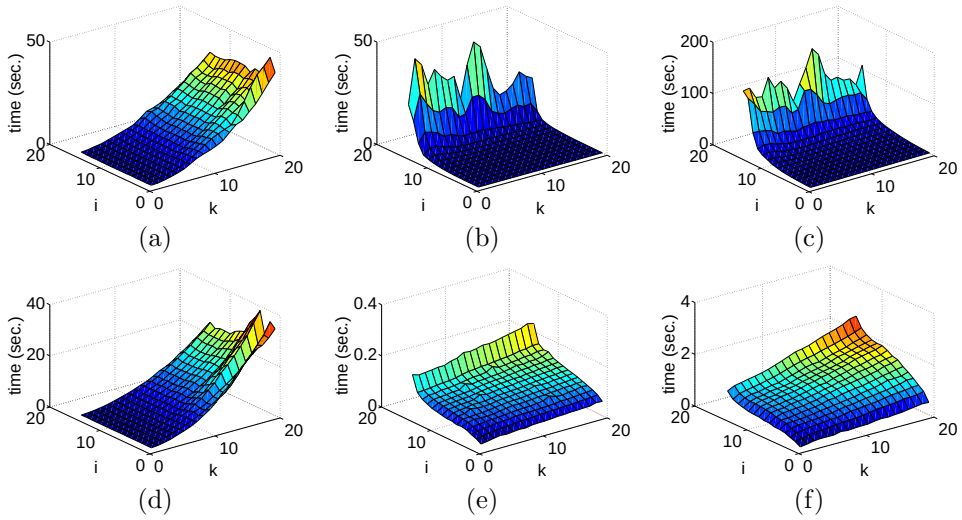


Fig. 6.5: *Average runtime over 10 problems for (a,d) SDPT3, (b,e) SPOR-SPG, and (c,f) SPOR-PQN, with* 50×200 *A,* $50 \times k$ *matrix B, and* $\sigma = \sigma_i$*. The results in the top row are for sparsity* $s = 20$*, those at the bottom are for* $s = 5$*.*

Finally, substituting this expression into (6.2) yields $\mathcal{P}_\tau(v)$ because the constraint $\|x_{(i)}\|_2 = u_i$ is automatically satisfied by setting $x_{(i)} = \gamma_i c_{(i)} = (u_i/v_i) \cdot c_{(i)}$. ∎

This proof can be extended to deal with weighted group projection. In that case the problem reduces to projection onto the weighted one-norm ball.

6.4. Experiments. With the results from sections 6.2 and 6.3 we can implement the root-finding algorithm for the MMV problem. To evaluate the performance we generate a set of random test problems with an increasing number k of columns in B, and compare the runtime with SDPT3. Throughout we use random 50×200 matrices A with entries drawn i.i.d. from the standard normal distribution. We set $B = AX_0$, where X_0 is a $n \times k$ matrix with s nonzero rows. For the experiments we use $k = 1, \ldots, 20$, and $\sigma_i = ((i-1)/15)^3 \cdot \|B\|_F$ for $i = 1, \ldots, 15$.

From Figs. 6.5(a) and (d) it is clear that while the runtime of SDPT3 is insensitive to the value of σ, but slows down quickly as the number of observations k increases; see also Fig. 6.6(a). SPOR-SPG and SPOR-PQN, on the other hand, are sensitive to σ, especially when there is no strictly sparse solution to the MMV problem; see Figs. 6.6(b,c). In particular, with $s = 20$ (i.e., relatively low sparsity level), these solvers require large runtimes for low values of σ. By contrast, the solution times are insensitive to σ when $s = 5$ (i.e., very sparse solutions). Regardless of σ, both solvers scale very well with increasing number of observations. Although SPOR-PQN requires more time to solve the problem it typically does so using fewer function evaluations; compare Fig. 6.6(a,b), and (c,d). Consequently, SPOR-PQN is expected to outperform SPOR-SPG when function evaluations are relatively expensive.

7. Sign-constrained basis pursuit. In this section we consider the generalized sign-restricted basis pursuit denoise. Without loss of generality, we consider the nonnegative formulation

$$\underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \le \sigma, \quad x \ge 0. \tag{7.1}$$

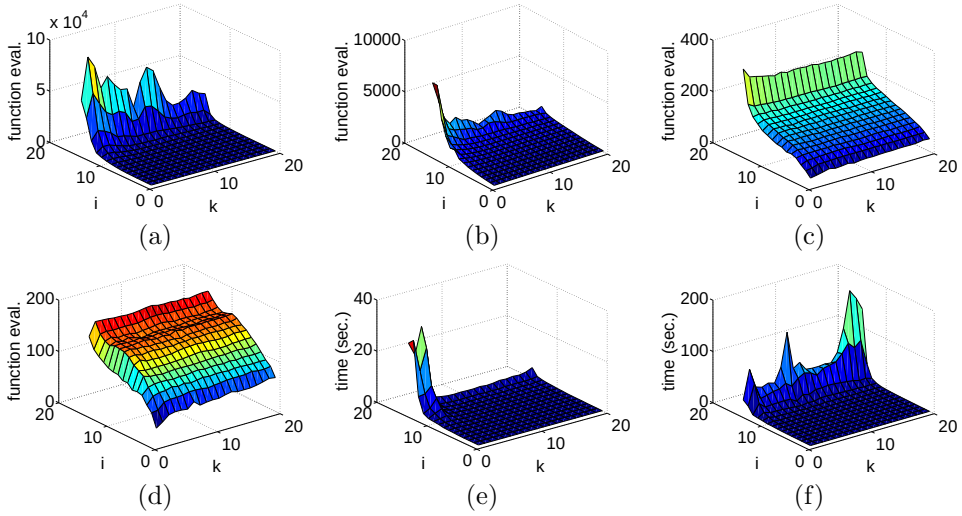


Fig. 6.6: *Average number of function evaluations for SPOR-SPG and SPOR-PQN for (a,b)* $s = 20$*, and (c,d)* $s = 5$*. Plots (e,f) show the SPOR-SPG runtime for two individual problems with* $s = 20$*.*

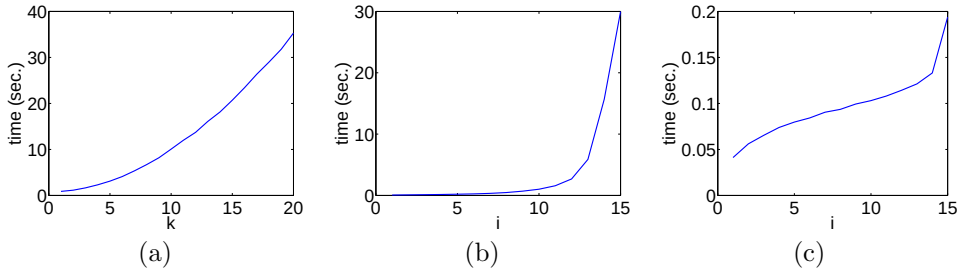


Fig. 6.7: *Runtime averaged over (a)* σ_i *for SDPT3 with* $s = 20$*, (b,c)* k *for SPOR-SPG with* $s = 20$ *and* $s = 5$*, respectively.*

(More general sign restrictions on x can be easily accommodated by redefining A.) The use of sign information as a prior can greatly help with the recovery of sparse signals, as shown by Donoho and Tanner [23, 27] for nonnegative basis pursuit (NNBP). Indeed, Fig. 7.2 shows that NNBP clearly outperforms general BP in the fraction of randomly chosen sparse x_0 that can be recovered from $b = Ax_0$. We next describe a problem in analytical chemistry that can conveniently be expressed as an NNBP.

7.1. Application – mass spectrometry. Mass spectrometry is a powerful method used to identify the chemical composition of a material sample. There exist several different approaches, but we restrict ourselves here to electron ionization (EI) mass spectrometry in which analyte molecules are ionized using high-energy electrons. Such ionization is often followed by fragmentation of the molecule with bonds breaking and forming in a manner characteristic of the original molecule. Mass spectrometers register the relative abundance of ions for a range of mass-to-charge (m/z) ratios, which can be used to deduce the chemical make-up of the compound [47, 54]. Once analyzed, the mass spectrum can be used as a signature to recognize the corresponding compound.

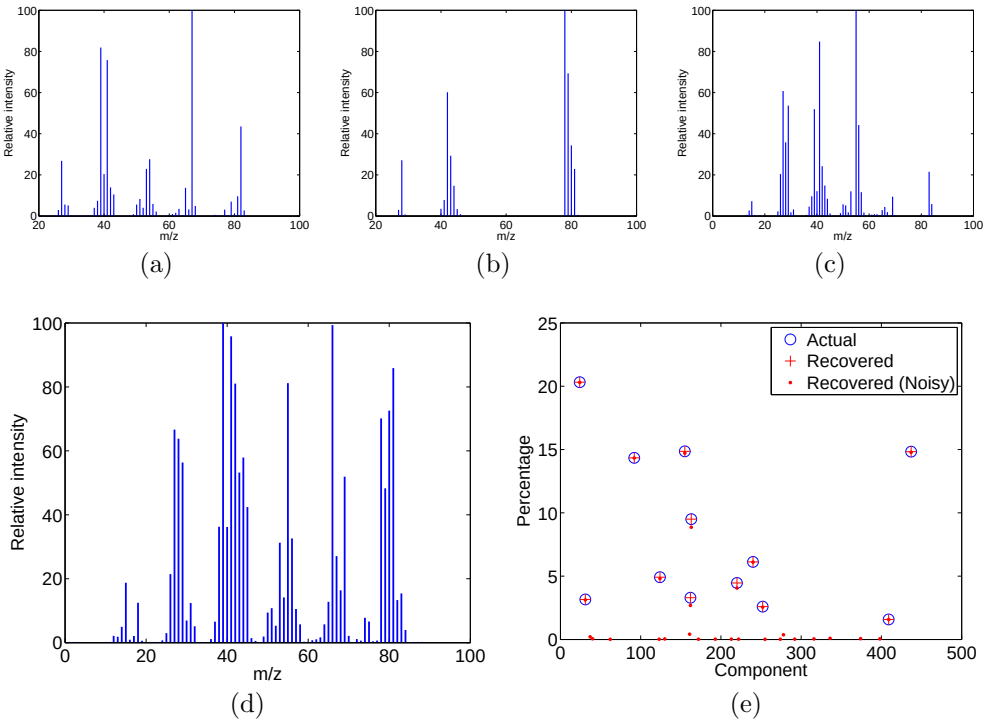


Fig. 7.1: *Mass spectra of (a)–(c) three of the twelve components in the mixture (d), along with (e) the actual and recovered relative abundance. The dictionary contains the mass spectra of 438 different molecules.*

As an example, consider the mass spectrum of propane (C_3H_8), illustrated in Fig. 7.3. The molecular ion generally has a mass of 44 u (unified atomic mass units) consisting of three ^{12}C atoms and eight ^{1}H atoms (the small peak at 45 m/z is due to prescence of ^{13}C isotopes). The peaks directly preceding 44 m/z are ions with increasingly many hydrogen atoms missing. The most intense peak at 29 m/z corresponds to ethyl ($C_2H_5{}^+$) ions, which is again preceded by ions with fewer hydrogen atoms. Finally, there are the peaks around the methyl ($CH_3{}^+$) ion at 15 m/z.

When analyzing mixtures, the components contribute independently to the measured spectrum. (In practice mixtures are generally separated using one of several types of chromatograph before introduction into the mass spectrometer.) In case of electron ionization, this superposition is linear [47] and the spectrum can thus be written as a nonnegative combination of the individual spectra. When presented with a mixed mass spectrum b we can identify the components by forming a dictionary of spectra A from possible substances and find a sparse nonnegative solution x satisfying $Ax \approx b$. This can be formulated as (7.1). Du and Angeletti [28] recently proposed a similar formulation.

To evaluate the approach we create a dictionary A containing the mass spectra of 438 compounds obtained from the NIST Chemistry WebBook [49]. Each spectrum is normalized and expanded to contain the intensities for 82 m/z values. The spectrum b of a synthetic mixture is created by adding the spectra of twelve compounds with randomly selected ratios; see Figs. 7.1(a–d). We then solve (7.1) with appropriate σ, and repeat the experiment with additive noise. The results of these simulations are

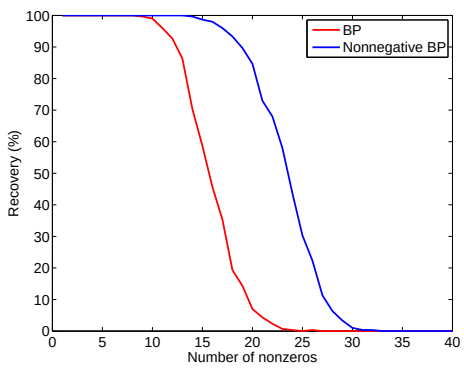


Fig. 7.2: *Equivalence breakdown curve for* 40×80 *random Gaussian matrix averaged over 300 random nonnegative* x_0.

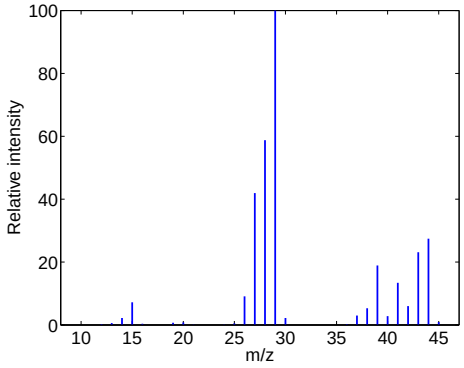


Fig. 7.3: *Mass spectrum of propane using electron ionization [49].*

shown in Fig. 7.1(e).

7.2. Polar function. We impose sign-restrictions on x by extending κ to be an extended real function with infinit values for all x violating the desired sign pattern. In this section we show how this affects the root-finding approach.

THEOREM 7.1. *Let* $\mathcal{C}$ *be the intersection of (half)spaces* $\mathcal{S}_i$ *with*

$$\mathcal{S}_i = \{x \in \mathbb{R}^n \mid x_i \geq 0\}, \quad \text{or} \quad \mathcal{S}_i = \{x \in \mathbb{R}^n \mid x_i \leq 0\}, \quad \text{or} \quad \mathcal{S}_i = \mathbb{R}^n.$$

Further, let $\|x\|_p$ *be any norm that is invariant under sign changes in* x*, and let*

$$\kappa(x) = \begin{cases} \|x\|_p & x \in \mathcal{C} \\ \infty & \text{otherwise.} \end{cases}$$

Then, with $\|\cdot\|_d$ *the dual of* $\|\cdot\|_p$*, and* $\mathcal{P}_{\mathcal{C}}(x)$ *the Euclidean projection onto* $\mathcal{C}$*,*

$$\kappa^\circ(x) = \|\mathcal{P}_{\mathcal{C}}(x)\|_d.$$

Proof. We consider three different cases: (i) $x \in \mathcal{C}$, (ii) $x \in \mathcal{C}^\circ$ (the polar of $\mathcal{C}$), and (iii) $x \notin \mathcal{C} \cup \mathcal{C}^\circ$, which cover all $x \in \mathbb{R}^n$. In the first case, we have $\mathcal{P}_{\mathcal{C}}(x) = x$ and therefore only need to show that the polar $\kappa^\circ(x)$ in (2.5) is attained by some $w \in \mathcal{C}$. This implies that for those x it does not matter whether we use $\kappa(x)$ or $\|x\|_p$, hence giving $\kappa^\circ(x) = \|x\|_d$. It suffices to show that w lies in the same orthant as x. Assuming the contrary, it is easily seen that $x^T w$ is increased by flipping the sign of the violating component while $\kappa(w)$ remains the same, giving a contradiction.

For the second case, $x \in \mathcal{C}^\circ$, it can be seen that $w^T x \leq 0$ for all $w \in \mathcal{C}$, and we therefore have $\kappa^\circ(x) = 0$. The results then follows from the fact that $\mathcal{P}_{\mathcal{C}}(x) = 0$, and therefore that $\kappa^\circ(x) = \|0\|_d = 0$.

For third case we define $u = \mathcal{P}_{\mathcal{C}}(x)$, and $v = x - u$, where $u^T v = 0$ due to projection. Now let w give the supremum in $\kappa^\circ(u)$. It follows from properties of $\|\cdot\|_p$ and $\mathcal{C}$ that w has the same support at u, and consequently $w^T v = 0$. Therefore

$$\kappa^\circ(x) = x^T w/\kappa(w) = (u^T w + v^T w)/\kappa(w) = u^T w/\kappa(w) = \kappa^\circ(u) = \|u\|_d,$$

which yields the required conclusion. ∎

Problem	Size of A	Sparsity	Noise level (ν)	$\sigma/\|r\|_2$	$\|x\|_1$
nonnegn01	100×256	10	0.01	1.02	6.1484e+0
nonnegn02	500×8192	200	0.001	0.9	1.5039e+2
nonnegn03	1500×8192	200	0.001	0.9	1.7499e+2
nonnegn04	500×8192	100	0.001	0.9	8.6661e+1
nonnegn05	500×8192	10	0.001	0.9	8.6205e+0
nonnegn06	500×8192	100	0.05	0.9	8.2346e+1
massspec	82×438	12	0.012	1.0	9.9505e−1

Table 7.1: *Test problem settings for nonnegative basis pursuit denoise experiments.*

Sign invariance is clearly satisfied for all p-norms, and hence is also satisfied for all nested norms based on them. In the latter case, the outer norms do not need to satisfy invariance under sign changes. It is therefore possible, for example, to impose independent sign restrictions on the real or imaginary parts of complex-valued vectors.

7.3. Projection. The projection for the sign-restricted formulation consists of setting to zero all components of x that violate the restriction and projecting the remainder onto the ball induced by the underlying norm.

7.4. Experiments. We apply the nonnegative basis pursuit denoise framework on a set of randomly restricted discrete cosine transformation (DCT) operators with noise scaled to $\nu\|Ax_0\|_2$, and σ set close to $\|r\|_2$. In most cases we underestimate σ, which makes the problem harder to solve for SPOR. The parameters for the different problems in the test set, including the noisy mass spectrometry setting described in Section 7.1, are given in Table 7.1.

Unfortunately, there are not many solvers that are specific for the nonnegative basis pursuit problem. Although in principle is should be possible to modify, e.g., GPSR [32] to forego variable splitting and work directly with the nonnegative part only, no such attempt has been made. As a result, this leaves us with CVX/SDPT3 [33, 61] for smaller problems, and FISTA [1] for the penalized formulation.

From the runtime and number of matrix-vector products reported in Table 7.2, it is apparent that both SPOR and FISTA require more effort to solve problems where the number of nonzero entries in x_0 is large compared to the length of b. The hardest amongst these problems is `nonnegn02`, which has few measurements per nonzero entry in x_0. Comparing `nonnegn04` to `nonnegn06` shows, as expected, that an increased σ makes the problem easier to solve. Likewise, `nonnegn03` is solved faster and more accurately than `nonnegn02` because of the larger number of measurements. Finally, note that CVX/SDPT3 does well on the mass spectrometry problem.

8. Nuclear-norm minimization. The matrix completion problem is a special case of the low-rank matrix recovery problem

$$\underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad \text{rank}(X) \quad \text{subject to} \quad \|\mathcal{A}X - b\|_2 \le \sigma,$$

where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^k$ is a linear operator; see (1.3). This rank minimization problem is generally intractable; Fazel [31] and Recht et al. [51] suggest the convex relaxation (1.7), which minimizes the sum of singular values.

Conditions for exact recovery of X_0 from $b := \mathcal{A}X_0$ using (1.7) with $\sigma = 0$ were recently studied by Recht et al. [51], who leverage the restricted isometry technique developed for exact vector recovery using the 1-norm. They derive necessary and

	Solver	Time (s)	Aprod	$\|x\|_1$	$\frac{\|\|r\|_2-\sigma\|}{\sigma}$	$\|x-x_0\|_\infty$
nonnegn01	CVX/SDPT3	1.3e+0	n.a.	6.1484e+0	2.1e−10	5.6e−8
	SPOR-SPG-H	3.4e−1	71	6.1476e+0	5.6e−3	1.2e−4
	SPOR-SPG	6.4e−1	153	6.1483e+0	4.2e−4	7.1e−6
	SPOR-PQN	7.8e−1	120	6.1484e+0	1.3e−4	4.5e−6
	FISTA	1.4e+0	335	6.1484e+0	9.1e−9	3.8e−9
nonnegn02	SPOR-SPG-H	8.9e+0	1052	1.5044e+2	2.9e−2	1.3e−1
	SPOR-SPG	3.1e+1	4158	1.5041e+2	9.9e−3	1.0e−1
	SPOR-PQN	1.1e+2	1350	1.5044e+2	7.8e−4	1.1e−1
	FISTA	1.5e+2	20011	1.5039e+2	9.0e−5	1.4e−3
nonnegn03	SPOR-SPG-H	1.7e+0	190	1.7499e+2	1.0e−2	2.6e−5
	SPOR-SPG	1.9e+0	246	1.7499e+2	1.2e−3	2.1e−5
	SPOR-PQN	5.5e+0	190	1.7499e+2	1.1e−2	2.9e−4
	FISTA	1.7e+1	2283	1.7499e+2	2.9e−9	3.4e−9
nonnegn04	SPOR-SPG-H	1.1e+1	1273	8.6662e+1	4.1e−2	9.7e−3
	SPOR-SPG	1.7e+1	2287	8.6665e+1	1.7e−3	7.9e−3
	SPOR-PQN	7.0e+1	900	8.6674e+1	5.2e−3	1.2e−2
	FISTA	1.5e+2	20011	8.6661e+1	5.2e−5	1.2e−5
nonnegn05	SPOR-SPG-H	4.6e−1	51	8.6202e+0	1.7e−2	1.2e−4
	SPOR-SPG	8.0e−1	101	8.6163e+0	2.2e−1	5.1e−4
	SPOR-PQN	6.7e+0	470	8.6205e+0	1.3e−3	4.0e−5
	FISTA	1.3e+1	1789	8.6205e+0	1.0e−7	1.3e−9
nonnegn06	SPOR-SPG-H	2.1e+0	241	8.2343e+1	7.9e−4	8.4e−3
	SPOR-SPG	4.7e+0	624	8.2346e+1	7.0e−5	6.0e−4
	SPOR-PQN	2.9e+1	676	8.2345e+1	1.9e−4	2.5e−3
	FISTA	8.3e+1	10963	8.2346e+1	1.8e−7	1.8e−6
massspec	CVX/SDPT3	1.1e+0	n.a.	9.9505e−1	2.2e−9	1.1e−8
	SPOR-SPG-H	3.2e+0	2172	9.7853e−1	4.4e+0	3.8e−2
	SPOR-SPG	3.8e+1	30070	9.9504e−1	2.4e−3	1.2e−5
	SPOR-PQN	2.3e+2	30975	9.9581e−1	2.4e−4	2.4e−3
	FISTA	2.5e+1	20023	9.9506e−1	2.9e−3	4.2e−5

Table 7.2: *Solver performance on the set nonnegative basis pursuit denoise test problems.*

sufficient conditions for recovery, and use them to compute recovery bounds for linear operators $\mathcal{A}$ whose matrix representation has independent random Gaussian entries.

Choosing $\mathcal{A}$ in (1.7) to be an operator that restricts elements of a matrix to the set Ω gives the nuclear-norm formulation for noisy matrix completion:

$$\underset{X}{\text{minimize}} \quad \kappa(X) := \|X\|_n \quad \text{subject to} \quad \|X_\Omega - B_\Omega\|_2 \le \sigma. \tag{8.1}$$

Conditions for exact recovery, with $\sigma = 0$, were studied by Candès and Recht [15] and Candès and Tao [17]. Candès and Plan [14] consider the noisy case, with $\sigma > 0$.

8.1. Application – distance matrix completion. As an illustration of nuclear norm minimization for matrix completion consider the following scenario; see also [15, 51]. Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ denote the coordinates of n sensors in $\mathbb{R}^d$. Given the squared pairwise distance $D_{i,j} := (x_i - x_j)^T(x_i - x_j)$ for a limited number of pairs $(i, j) \in \Omega$, we want to determine the distance between any pair of sensors. That is, we

Problem	Size of M	Rank	Observed entries	Noise level (ν)
nucnorm01	10×10	2	80%	0.05
nucnorm02	50×50	4	60%	0.01
nucnorm03	100×100	7	50%	0.02
nucnorm04	100×100	7	30%	0.03
nucnorm05	200×200	12	20%	0.01
nucnorm06	200×200	2	20%	0.01

Table 8.1: *Test problems for matrix completion experiments. The parameter ν applies to the noisy problem versions.*

want to find the Euclidean (squared) distance matrix D given by

$$D = ev^T + ve^T - 2X^TX,$$

where e denotes the vector of all ones, and each $v_i := \|x_i\|_2$. Because D is a low-rank matrix (it has rank at most $d+2$), we can try to apply (1.7) to recover D from D_Ω.

8.2. Polar function. The polar (i.e., the dual) of the nuclear norm $\|X\|_n$ is well-known to be given by the operator norm $\|X\|_2$, which corresponds to the largest singular value of X.

8.3. Projection. From existing results it follows that projection onto the nuclear norm ball with radius τ can be obtained by computing the singular value decomposition (SVD) and projecting the singular values onto the usual 1-norm ball with radius τ.

THEOREM 8.1. *Let C be an $m \times n$ matrix with the singular value decomposition $C = USV^T$, where U and V orthonormal and $S = \operatorname{diag}(s)$. Then the solution of the nuclear-norm projection problem*

$$\underset{X}{\text{minimize}} \quad \|C - X\|_F \quad \text{subject to} \quad \|X\|_n \le \tau, \tag{8.2}$$

is $X^* = U \operatorname{diag}(\bar{s})V$, where $\bar{s} = \mathcal{P}_\tau(s)$ is the 1-norm projection of s.

Proof. This follows directly from [11, Theorem 2.1] or [43, Theorem 3]. ∎

8.4. Experiments. A number of solvers for matrix completion based on nuclear norm minimization have been recently introduced. Ma et al. [43] propose FPCA, which combines approximate SVD with fixed-point iterations to solve a penalized formulation of (1.7); Cai et al. [11] derive a singular value soft-thresholding algorithm, called SVT, for solving a slight relaxation to the exact matrix completion problem (1.7). Two more solvers were proposed by Toh and Yun [59], and Liu et al. [41], but their implementations are not yet publicly available.

Our first set of experiments is for matrix completion without noise, i.e., (8.1) with $\sigma = 0$. Both FPCA and SVT approximate the solution to this problem by choosing a small penalty parameter in their respective formulations. From Table 8.2 we see that SVT reaches its default maximum number of 500 iterations on problems `nucnorm02`, `nucnorm04`, and `nucnorm05`. Despite the fact that SVT computes fewer SVDs, it is still slower and less accurate than FPCA. The reason for this difference lies predominantly in the way the SVDs are computed. Note that SVT is designed for large-scale problems, and hence the results reported here may not show its true potential. FPCA does well

	Solver	Time (s)	#SVD	$\|X\|_n$	$\frac{\|\|r\|_2-\sigma\|}{\max(1,\sigma)}$	$\|X-X^*\|$
nucnorm02	CVX/SDPT3	1.0e+1	n.a.	2.0179e+2	0.0e+0	—
	SPOR-SPG-H	9.1e+0	540	2.0180e+2	9.7e−5	3.5e−1
	SPOR-SPG	4.4e+0	524	2.0179e+2	5.9e−4	3.2e−1
	SPOR-PQN	1.5e+1	2554	2.0180e+2	9.5e−4	3.3e−1
	FPCA	5.9e+0	4517	2.0179e+2	5.8e−5	2.5e−4
	SVT	3.2e+0	500	2.0184e+2	4.6e−2	9.4e−1
nucnorm04	CVX/SDPT3	6.6e+1	n.a.	7.1411e+2	0.0e+0	—
	SPOR-SPG-H	7.8e+1	1074	7.1470e+2	8.3e−5	1.8e+1
	SPOR-SPG	9.7e+1	2778	7.1415e+2	8.4e−4	2.8e+0
	SPOR-PQN	4.8e+2	20033	7.1435e+2	2.5e−13	7.9e+0
	FPCA	1.1e+1	4546	7.1433e+2	2.4e−4	8.6e+0
	SVT	1.6e+2	500	8.7067e+2	4.4e+1	1.1e+2
nucnorm05	CVX/SDPT3	1.0e+3	n.a.	2.2578e+3	0.0e+0	—
	SPOR-SPG-H	5.2e+2	1259	2.2583e+3	1.9e−5	1.9e+1
	SPOR-SPG	3.9e+2	2106	2.2579e+3	4.1e−3	5.5e+0
	SPOR-PQN	1.8e+3	13132	2.2590e+3	8.0e−3	8.8e+0
	FPCA	3.3e+1	4666	2.3208e+3	7.8e−4	2.8e+2
	SVT	6.4e+2	500	3.6454e+3	2.1e+2	5.0e+2

Table 8.2: *Solver performance on a set of matrix completion problems without noise.*

	Solver	Time (s)	#SVD	$\|X\|_n$	$\frac{\|\|r\|_2-\sigma\|}{\max(1,\sigma)}$	$\|X-X^*\|$
nucnorm02	CVX/SDPT3	1.3e+1	n.a.	2.0061e+2	3.1e−9	4.3e−5
	SPOR-SPG-H	2.1e+0	110	2.0063e+2	3.8e−5	3.0e−1
	SPOR-SPG	4.6e+0	519	2.0061e+2	5.2e−7	5.5e−3
	SPOR-PQN	1.2e+1	2058	2.0061e+2	1.3e−6	1.1e−2
	FPCA	3.4e+0	2500	2.0045e+2	7.7e−2	5.0e−1
	FISTA	6.3e+0	538	2.0061e+2	3.7e−1	8.9e−6
nucnorm04	CVX/SDPT3	8.9e+1	n.a.	6.9528e+2	2.1e−11	4.2e−4
	SPOR-SPG-H	1.2e+1	138	6.9531e+2	9.2e−5	3.8e+0
	SPOR-SPG	2.6e+1	679	6.9528e+2	3.7e−5	6.1e−2
	SPOR-PQN	1.5e+2	5490	6.9528e+2	6.5e−5	3.2e−1
	FPCA	5.2e+0	2000	6.9383e+2	3.2e−1	1.7e+1
	FISTA	7.4e+1	1319	6.9528e+2	8.5e−1	1.8e−4
nucnorm06	CVX/SDPT3	9.8e+2	n.a.	3.9998e+2	1.9e−8	1.3e−4
	SPOR-SPG-H	2.7e+1	51	3.9999e+2	6.9e−5	1.3e+0
	SPOR-SPG	3.0e+1	134	3.9998e+2	3.1e−7	4.7e−3
	SPOR-PQN	1.4e+2	833	3.9998e+2	4.1e−7	6.1e−2
	FPCA	1.2e+1	2000	3.9974e+2	5.0e−2	2.9e+0
	FISTA	7.6e+1	235	3.9998e+2	7.3e−1	1.9e−3

Table 8.3: *Solver performance on a set of matrix completion problems with noise.*

on the first two problems, but less so on `nucnorm05`. This is most likely due to the fixed value of the penalty parameter, which here also leads to a higher misfit and objective. CVX/SDPT3 does not scale well with the problem size and, as noted in [41], is not designed for matrices with more than 100 rows and columns. Similarly, as a consequence of the way projection is currently implemented, SPOR-SPG and

SPOR-PQN do not scale very well either mainly because it computes the full SVD for each projection step, although in principle only a few singluar values are required. Interestingly, SPOR-SPG is more accurate, and nearly twice as fast as SPOR-SPG-H on `nucnorm02` and `nucnorm05`.

For the noisy problem instances, shown in Table 8.3, SPOR-SPG takes about twice as long as SPOR-SPG-H, but does obtain substantially more accurate solutions. FISTA also has a good performance, but does not scale well because, as with SPOR, we compute the full SVD. Finally, SPOR-PQN is not very well suited to the nuclear norm minimization problem because it requires many projection iterations, which are expensive in this context.

9. Sparse and low-rank matrix decomposition. Chandrasekaran et al. [20] and Candès et al. [13] consider the problem of decomposing a matrix C into as $C = A + B$, where A is a low-rank matrix A and B is sparse. This can be accomplished by solving

$$\underset{A,B}{\text{minimize}} \quad \gamma\|\operatorname{vec}(A)\|_1 + \|B\|_n \quad \text{subject to} \quad A + B = C, \tag{9.1}$$

where $\operatorname{vec}(A)$ denotes the vectorization of matrix A. In this section we briefly look at the more general problem

$$\underset{A,B}{\text{minimize}} \quad \gamma\|\operatorname{vec}(A)\|_1 + \|B\|_n \quad \text{subject to} \quad \left\| R \cdot \begin{bmatrix} \operatorname{vec}(A) \\ \operatorname{vec}(B) \end{bmatrix} - c \right\|_2 \leq \sigma, \tag{9.2}$$

which nicely ties together results from the previous sections. Formulation (9.2) reduces to (9.1) by choosing $R = [I, I]$, $c = \operatorname{vec}(C)$, and $\sigma = 0$, and is itself a special case of (P_σ), with $\kappa(A, B) := \gamma\|\operatorname{vec}(A)\|_1 + \|B\|_n$. Proceeding as before we derive the polar κ°, and present an efficient projection algorithm.

9.1. Polar function. For the derivation of the polar, we combine A and B into a single vector x with disjoint parts $x_{\sigma_1} = \operatorname{vec}(A)$, and $x_{\sigma_2} = \operatorname{vec}(B)$. It then follows from Theorems 6.1 and 5.1 that $\kappa^\circ(A, B) = \max\{\|\operatorname{vec}(A)\|_\infty/\gamma, \|B\|\}$.

9.2. Projection. The projection problem corresponding to (4.1) is given by

$$\underset{X,Y}{\text{minimize}} \quad \tfrac{1}{2}\left\| \begin{bmatrix} A \\ B \end{bmatrix} - \begin{bmatrix} X \\ Y \end{bmatrix} \right\|_F^2 \quad \text{subject to} \quad \kappa(X, Y) \leq \tau.$$

Denoting $\operatorname{vec}(A)$ and $\operatorname{vec}(X)$ by a and x respectively, we can rewrite this as

$$\underset{x,Y}{\text{minimize}} \quad \tfrac{1}{2}\|a - x\|_2^2 + \tfrac{1}{2}\|B - Y\|_F^2 \quad \text{subject to} \quad \gamma\|x\|_1 + \|Y\|_F \leq \tau.$$

We derive the solution via the dual problem. To this end, note that the corresponding Lagrange function is given by

$$\mathcal{L}(x, Y, \lambda) = \tfrac{1}{2}\|a - x\|_2^2 + \tfrac{1}{2}\|B - Y\|_F^2 + \lambda(\gamma\|x\|_1 + \|Y\|_n - \tau),$$

which leads to the Lagrange dual function

$$\inf_{x,Y} \mathcal{L}(x, Y, \lambda) = \inf_x \left\{ \tfrac{1}{2}\|a - x\|_2^2 + \lambda\gamma\|x\|_1 \right\} + \inf_Y \left\{ \tfrac{1}{2}\|B - Y\|_F^2 + \lambda\|Y\|_n \right\} - \lambda\tau.$$

The first infimum on the right-hand side has a closed-form solution corresponding to soft-thresholding of the vector a. The second infimum, likewise, corresponds to

soft-thresholding of the singular values of B. For a given γ, the values of x_λ and Y_λ are uniquely determined by λ. The value of κ is non-increasing in λ, and λ should therefore be the smallest nonnegative value such that $\kappa(\mathrm{mat}(x_\lambda), Y_\lambda) \le \tau$; here, $\mathrm{mat}(\cdot)$ is the inverse of $\mathrm{vec}(\cdot)$.

Let $B = USV^T$ with $S = \mathrm{diag}(s)$. We reduce the mixed-norm projection to weighted projection onto the 1-norm ball (see section 5):

$$\underset{x,\bar{s}}{\text{minimize}} \quad \frac{1}{2}\left\| \begin{bmatrix} a \\ s \end{bmatrix} - \begin{bmatrix} x \\ \bar{s} \end{bmatrix} \right\|_2^2 \quad \text{subject to} \quad \left\| \begin{bmatrix} \gamma I & \\ & I \end{bmatrix} \begin{bmatrix} x \\ \bar{s} \end{bmatrix} \right\|_1 \le \tau.$$

The final projection is obtained by setting $X = \mathrm{mat}(x)$, and $Y = U\,\mathrm{diag}(\bar{s})V^T$.

10. Future work. The root-finding algorithm that we propose is able to solve a wide variety of sparse recovery problems, and is one of the few solvers that can handle an explicit least-squares misfit constraint.

The overall performance of the algorithm ultimately depends on being able to efficiently minimize a linear-least squares problem over a convex set. The two subproblem solvers that we have experimented with have proven adequate on an interesting range of problems, but further improvement may be possible.

Acknowledgments. We would like the thank Dmitri Malioutov for making available to us his source-localization code, and Yun Ling for clarifying certain aspects of mass spectography.

References.

[1] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sciences, 2 (2009), pp. 183–202.

[2] S. Becker, J. Bobin, and E. Candès, *Nesta: A fast and accurate first-order method for sparse recovery*, tech. rep., California Institute of Technology, April 2009.

[3] E. van den Berg, *Convex optimization for generalized sparse recovery*, PhD thesis, University of British Columbia, December 2009.

[4] E. van den Berg and M. P. Friedlander, *SPGL1: A solver for large-scale sparse reconstruction*. Available at `http://www.cs.ubc.ca/labs/scl/index.php/Main/Spgl1`, June 2007.

[5] ——, *Probing the Pareto frontier for basis pursuit solutions*, Tech. rep. TR-2008-01, Department of Computer Science, University of British Columbia, Vancouver, January 2008. To appear in *SIAM J. Sci. Comp.*

[6] ——, *Theoretical and empirical results for recovery from multiple measurements*, Tech. Rep. TR-2009-7, Department of Computer Science, University of British Columbia, Vancouver, September 2009. To appear in *IEEE Trans. Inform. Theory.*

[7] E. van den Berg, M. Schmidt, M. P. Friedlander, and K. Murphy, *Group sparsity via linear-time projection*, Tech. Rep. TR-2008-09, Dept. of Computer Science, UBC, June 2008.

[8] D. Bertsekas, *Convex optimization theory*, Athena Scientific, Massachusetts, 2009.

[9] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex analysis and optimization*, Athena Scientific, 2003.

[10] E. G. Birgin, J. M. Martínez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.

[11] J.-F. Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*. Submitted, September 2008.
[12] E. J. Candès, *Compressive sampling*, in Proceedings of the International Congress of Mathematicians, 2006.
[13] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Arxiv preprint math/0912.3599, (2009).
[14] E. J. Candès and Y. Plan, *Matrix completion with noise*. Submitted to Proceedings of the IEEE, 2009.
[15] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*. To appear in Found. of Comput. Math., 2008.
[16] E. J. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.
[17] E. J. Candès and T. Tao, *The power of convex relaxation: near-optimal matrix completion*. Submitted, March 2009.
[18] E. J. Candès, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted L1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.
[19] W. L. Chan, M. L. Moravec, R. G. Baraniuk, and D. M. Mittleman, *Terahertz imaging with compressed sensing and phase retrieval*, Opt. Lett., 33 (2008), pp. 974–976.
[20] V. Chandrasekaran, S. Sanghavi, P. P. Parrilo, and A. S. Willsky, *Rank-sparsity incoherence for matrix decomposition*, Arxiv preprint math/0906.2220, (2009).
[21] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.
[22] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
[23] D. L. Dohono and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9452–9457.
[24] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[25] D. L. Donoho, *For most large underdetermined systems of equations the minimal ℓ_1-norm near-solution approximates the sparsest near-solution*, Comm. Pure Appl. Math., 59 (2006), pp. 907–934.
[26] ——, *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), pp. 797–829.
[27] D. L. Donoho and J. Tanner, *Sparse nonnegative solution of underdetermined linear equations by linear programming*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9446–9451.
[28] P.-C. Du and R. H. Angeletti, *Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution*, Analytical Chemistry, 78 (2006), pp. 3385–3392.
[29] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the ℓ_1-ball for learning in high dimensions*, in Proceedings of the 25th International Converence on Machine Learning, 2008, pp. 272–279.
[30] Y. C. Eldar and M. Mishali, *Robust recovery of signals from a union of subspaces*. arXiv 0807.4581, July 2008.
[31] M. Fazel, *Matrix Rank Minimization with Applications*, PhD thesis, Stanford

University, 2002.

[32] M. Figueiredo, R. Nowak, and S. J. Wright, *Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems*, Sel. Top. in Signal Process., IEEE J., 1 (2007), pp. 586–597.

[33] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming (web page and software)*. `http://stanford.edu/~boyd/cvx`, February 2009.

[34] M. Gu, L.-H. Lim, and C. J. Wu, *PARNES: A rapidly convergent algorithm for accurate recovery of sparse and approximately sparse signals*. arXiv 0911.0492, November 2009.

[35] E. Hale, W. Yin, and Y. Zhang, *A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing*, Tech. Rep. TR07-07, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2007.

[36] X. Hang and F. Wu, *L1 Least Square for Cancer Diagnosis using Gene Expression Data*, J. Comput. Sci. Syst. Biol., 2 (2009), pp. 167–173.

[37] F. Herrmann, Y. Erlangga, and T. Lin, *Compressive simultaneous full-waveform simulation*, Geophysics, 74 (2009), p. A35.

[38] F. J. Herrmann and G. Hennenfent, *Non-parametric seismic data recovery with curvelet frames*, Geophys. J. Int., 173 (2008), pp. 233–248.

[39] A. HesamMohseni, M. Babaie-Zadeh, and C. Jutten, *Inflating compressed samples: A joint source-channel coding approach for noise-resistant compressed sensing*, Acoustics, Speech, and Signal Processing, IEEE International Conference on, 0 (2009), pp. 2957–2960.

[40] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *An interior-point method for large-scale L1-regularized least squares*, IEEE J. Sel. Top. Signal Process., 1 (2007), pp. 606–617.

[41] Y.-J. Liu, D. Sun, and K.-C. Toh, *An implementable proximal point algorithmic framework for nuclear norm minimization*. Preprint, July 2009.

[42] M. Lustig, D. L. Donoho, and J. M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Mag. Resonance Med., 58 (2007), pp. 1182–1195.

[43] S. Ma, D. Goldfarb, and L. Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*. arXiv 0905.1643, May 2009.

[44] D. Malioutov, M. Çetin, and A. S. Willsky, *A sparse signal reconstruction perspective for source localization with sensor arrays*, IEEE Trans. Sig. Proc., 53 (2005), pp. 3010–3022.

[45] D. M. Malioutov, *A sparse signal reconstruction perspective for source localization with sensor arrays*, Master's thesis, Dept. Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, July 2003.

[46] S. Marchesini, *Ab initio compressive phase retrieval*, Arxiv preprint arXiv:0809.2006, (2008).

[47] F. W. McLafferty and F. Tureček, *Interpretation of Mass Spectra*, University Science Books, Mill Valley, CA, fourth ed., 1993.

[48] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comp., 24 (1995), pp. 227–234.

[49] National Institute of Standards and Technology, *NIST Chemistry WebBook*. `http://webbook.nist.gov/chemistry/`, 2009.

[50] Y. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Program.,

103 (2005).

[51] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.* arXiv 0706.4138, June 2007.

[52] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[53] M. Schmidt, E. van den Berg, M. P. Friedlander, and K. Murphy, *Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm*, in Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, D. van Dyk and M. Welling, eds., vol. 5, Clearwater Beach, Florida, April 2009, pp. 456–463.

[54] R. M. Smith, *Understanding mass spectra: A basic approach*, John Wiley and Sons, Hoboken, NJ, second ed., 2004.

[55] E. Specht, *Packing of circles in the unit circle.* `http://hydra.nat.uni-magdeburg.de/packing/cci/cci.html`, 2009.

[56] M. Stojnic, F. Parvaresh, and B. Hassibi, *On the reconstruction of block-sparse signals with an optimal number of measurements.* Available at arXiv 0804.0041, March 2008.

[57] J. F. Sturm, *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones (updated for Version 1.05)*, tech. rep., Department of Econometrics, Tilburg University, Tilburg, The Netherlands, August 1998 – October 2001.

[58] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. R. Statist. Soc. B., 58 (1996), pp. 267–288.

[59] K.-C. Toh and S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems.* Preprint, April 2009.

[60] J. A. Tropp, *Algorithms for simultaneous sparse approximation: Part ii: Convex relaxation*, Signal Processing, 86 (2006), pp. 589–602.

[61] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Math. Program., Ser. B, 95 (2003), pp. 189–217.

[62] Y. Wiaux, L. Jacques, G. Puy, A. Scaife, and P. Vandergheynst, *Compressed sensing imaging techniques for radio interferometry*, Monthly Notices of the Royal Astronomical Society, 395 (2009), pp. 1733–1742.

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

[64] J. Zheng, E. Jacobs, et al., *Video compressive sensing using spatial domain sparsity*, Optical Engineering, 48 (2009), pp. 087006–1–087006–10.