The Convex Geometry of Linear Inverse Problems

**Venkat Chandrasekaran · Benjamin Recht ·
Pablo A. Parrilo · Alan S. Willsky**

Abstract In applications throughout science and engineering one is often faced with
the challenge of solving an ill-posed inverse problem, where the number of available
measurements is smaller than the dimension of the model to be estimated. However in
many practical situations of interest, models are constrained structurally so that they
only have a few degrees of freedom relative to their ambient dimension. This paper
provides a general framework to convert notions of simplicity into convex penalty
functions, resulting in convex optimization solutions to linear, underdetermined in-
verse problems. The class of simple models considered includes those formed as
the sum of a few atoms from some (possibly infinite) elementary atomic set; exam-

V. Chandrasekaran (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena,
CA 91125, USA
e-mail: venkatc@caltech.edu

B. Recht
Computer Sciences Department, University of Wisconsin, Madison, WI 53706, USA
e-mail: brecht@cs.wisc.edu

P.A. Parrilo · A.S. Willsky
Laboratory for Information and Decision Systems, Department of Electrical Engineering and
Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

P.A. Parrilo
e-mail: parrilo@mit.edu

A.S. Willsky
e-mail: willsky@mit.edu

ples include well-studied cases from many technical fields such as sparse vectors (signal processing, statistics) and low-rank matrices (control, statistics), as well as several others including sums of a few permutation matrices (ranked elections, multiobject tracking), low-rank tensors (computer vision, neuroscience), orthogonal matrices (machine learning), and atomic measures (system identification). The convex programming formulation is based on minimizing the norm induced by the convex hull of the atomic set; this norm is referred to as the *atomic norm*. The facial structure of the atomic norm ball carries a number of favorable properties that are useful for recovering simple models, and an analysis of the underlying convex geometry provides sharp estimates of the number of generic measurements required for exact and robust recovery of models from partial information. These estimates are based on computing the Gaussian widths of tangent cones to the atomic norm ball. When the atomic set has algebraic structure the resulting optimization problems can be solved or approximated via semidefinite programming. The quality of these approximations affects the number of measurements required for recovery, and this tradeoff is characterized via some examples. Thus this work extends the catalog of simple models (beyond sparse vectors and low-rank matrices) that can be recovered from limited linear information via tractable convex programming.

Keywords Convex optimization · Semidefinite programming · Atomic norms · Real algebraic geometry · Gaussian width · Symmetry

Mathematics Subject Classification 52A41 · 90C25 · 90C22 · 60D05 · 41A45

1 Introduction

Deducing the state or structure of a system from partial, noisy measurements is a fundamental task throughout the sciences and engineering. A commonly encountered difficulty that arises in such inverse problems is the limited availability of data relative to the ambient dimension of the signal to be estimated. However many interesting signals or models in practice contain few degrees of freedom relative to their ambient dimension. For instance a small number of genes may constitute a signature for disease, very few parameters may be required to specify the correlation structure in a time series, or a sparse collection of geometric constraints might completely specify a molecular configuration. Such low-dimensional structure plays an important role in making inverse problems well-posed. In this paper we propose a unified approach to transform notions of simplicity into convex penalty functions, thus obtaining convex optimization formulations for inverse problems.

   We describe a model as simple if it can be written as a nonnegative combination of a few elements from an atomic set. Concretely let $\mathbf{x} \in \mathbb{R}^p$ be formed as follows:

$$\mathbf{x} = \sum_{i=1}^{k} c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A}, c_i \geq 0, \tag{1}$$

where $\mathcal{A}$ is a set of atoms that constitute simple building blocks of general signals. Here we assume that $\mathbf{x}$ is *simple* so that k is relatively small. For example $\mathcal{A}$ could

be the finite set of unit-norm one-sparse vectors, in which case **x** is a sparse vector, or $\mathcal{A}$ could be the infinite set of unit-norm rank-one matrices, in which case **x** is a low-rank matrix. These two cases arise in many applications, and have received a tremendous amount of attention recently as several authors have shown that sparse vectors and low-rank matrices can be recovered from highly incomplete information [16, 17, 27, 28, 63]. However a number of other structured mathematical objects also fit the notion of simplicity described in (1). The set $\mathcal{A}$ could be the collection of unit-norm rank-one tensors, in which case **x** is a low-rank tensor and we are faced with the familiar challenge of low-rank tensor decomposition. Such problems arise in numerous applications in computer vision and image processing [1], and in neuroscience [5]. Alternatively $\mathcal{A}$ could be the set of permutation matrices; sums of a few permutation matrices are objects of interest in ranking [43] and multiobject tracking. As yet another example, $\mathcal{A}$ could consist of measures supported at a single point so that **x** is an atomic measure supported at just a few points. This notion of simplicity arises in problems in system identification and statistics.

In each of these examples as well as several others, a fundamental problem of interest is to recover **x** given limited *linear* measurements. For instance the question of recovering a sparse function over the group of permutations (i.e., the sum of a few permutation matrices) given linear measurements in the form of partial Fourier information was investigated in the context of ranked election problems [43]. Similar linear inverse problems arise with atomic measures in system identification, with orthogonal matrices in machine learning, and with simple models formed from several other atomic sets (see Sect. 2.2 for more examples). Hence we seek tractable computational tools to solve such problems. When $\mathcal{A}$ is the collection of one-sparse vectors, a method of choice is to use the ℓ_1 norm to induce sparse solutions. This method has seen a surge in interest in the last few years as it provides a tractable convex optimization formulation to exactly recover sparse vectors under various conditions [16, 27, 28]. More recently the nuclear norm has been proposed as an effective convex surrogate for solving rank minimization problems subject to various affine constraints [17, 63].

Motivated by the success of these methods we propose a general convex optimization framework in Sect. 2 in order to recover objects with structure of the form (1) from limited linear measurements. The guiding question behind our framework is: How do we take a concept of simplicity such as sparsity and derive the ℓ_1 norm as a convex heuristic? In other words what is the natural procedure to go from the set of one-sparse vectors $\mathcal{A}$ to the ℓ_1 norm? We observe that the convex hull of (unit-Euclidean-norm) one-sparse vectors is the unit ball of the ℓ_1 norm, or the cross-polytope. Similarly the convex hull of the (unit-Euclidean-norm) rank-one matrices is the nuclear norm ball; see Fig. 1 for illustrations. These constructions suggest a natural generalization to other settings. Under suitable conditions the convex hull conv($\mathcal{A}$) defines the unit ball of a norm, which is called the *atomic norm* induced by the atomic set $\mathcal{A}$. We can then minimize the atomic norm subject to measurement constraints, which results in a convex programming heuristic for recovering simple models given linear measurements. As an example suppose we wish to recover the sum of a few permutation matrices given linear measurements. The convex hull of the set of permutation matrices is the *Birkhoff polytope* of doubly stochastic matrices [76], and our proposal is to solve a convex program that minimizes the
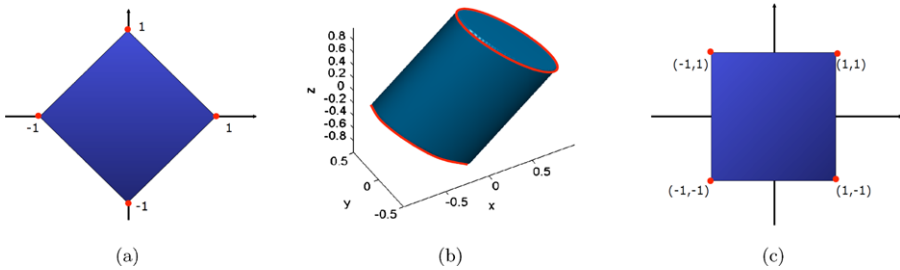
Fig. 1 Unit balls of some atomic norms: In each figure, the set of atoms is graphed in *red*, and the unit ball of the associated atomic norm is graphed in *blue*. In (**a**), the atoms are the unit-Euclidean-norm one-sparse vectors, and the atomic norm is the ℓ_1 norm. In (**b**), the atoms are the 2×2 symmetric unit-Euclidean-norm rank-one matrices, and the atomic norm is the nuclear norm. In (**c**), the atoms are the vectors $\{-1, +1\}^2$, and the atomic norm is the ℓ_∞ norm (Color figure online)

norm induced by this polytope. Similarly if we wish to recover an orthogonal matrix from linear measurements we would solve a *spectral norm* minimization problem, as the spectral norm ball is the convex hull of all orthogonal matrices. As discussed in Sect. 2.5 the atomic norm minimization problem is, in some sense, the best convex heuristic for recovering simple models with respect to a given atomic set.

We give general conditions for exact and robust recovery using the atomic norm heuristic. In Sect. 3 we provide concrete bounds on the number of generic linear measurements required for the atomic norm heuristic to succeed. This analysis is based on computing certain *Gaussian widths* of tangent cones with respect to the unit balls of the atomic norm [38]. Arguments based on Gaussian width have been fruitfully applied to obtain bounds on the number of Gaussian measurements for the special case of recovering sparse vectors via ℓ_1 norm minimization [66, 69], but computing Gaussian widths of general cones is not easy. Therefore it is important to exploit the special structure in atomic norms, while still obtaining sufficiently general results that are broadly applicable. An important theme in this paper is the connection between Gaussian widths and various notions of *symmetry*. Specifically by exploiting symmetry structure in certain atomic norms as well as convex duality properties, we give bounds on the number of measurements required for recovery using very general atomic norm heuristics. For example we provide precise estimates of the number of generic measurements required for exact recovery of an orthogonal matrix via spectral norm minimization, and the number of generic measurements required for exact recovery of a permutation matrix by minimizing the norm induced by the Birkhoff polytope. While these results correspond to the recovery of individual atoms from random measurements, our techniques are more generally applicable to the recovery of models formed as sums of a few atoms as well. We also give tighter bounds than those previously obtained on the number of measurements required to robustly recover sparse vectors and low-rank matrices via ℓ_1 norm and nuclear norm minimization. In all of the cases we investigate, we find that the number of measurements required to reconstruct an object is proportional to its intrinsic dimension rather than the ambient dimension, thus confirming prior folklore. See Table 1 for a summary of these results.

Table 1 A summary of the recovery bounds obtained using Gaussian width arguments

Underlying model	Convex heuristic	No. of Gaussian measurements
s-Sparse vector in $\mathbb{R}^p$	ℓ_1 norm	$2s\log(p/s) + 5s/4$
$m \times m$ rank-r matrix	Nuclear norm	$3r(2m - r)$
Sign-vector $\{-1, +1\}^p$	ℓ_∞ norm	$p/2$
$m \times m$ permutation matrix	Norm induced by Birkhoff polytope	$9m\log(m)$
$m \times m$ orthogonal matrix	Spectral norm	$(3m^2 - m)/4$

Although our conditions for recovery and bounds on the number of measurements hold generally, we note that it may not be possible to obtain a computable representation for the convex hull conv($\mathcal{A}$) of an arbitrary set of atoms $\mathcal{A}$. This leads us to another important theme of this paper, which we discuss in Sect. 4, on the connection between algebraic structure in $\mathcal{A}$ and the semidefinite representability of the convex hull conv($\mathcal{A}$). In particular when $\mathcal{A}$ is an algebraic variety the convex hull conv($\mathcal{A}$) can be approximated as (the projection of) a set defined by linear matrix inequalities. Thus the resulting atomic norm minimization heuristic can be solved via semidefinite programming. A second issue that arises in practice is that even with algebraic structure in $\mathcal{A}$ the semidefinite representation of conv($\mathcal{A}$) may not be computable in polynomial time, which makes the atomic norm minimization problem intractable to solve. A prominent example here is the tensor nuclear norm ball, obtained by taking the convex hull of the rank-one tensors. In order to address this problem we give a hierarchy of semidefinite relaxations using *theta bodies* that approximate the original (intractable) atomic norm minimization problem [39]. We also highlight that while these semidefinite relaxations are more tractable to solve, we require more measurements for exact recovery of the underlying model than if we solve the original intractable atomic norm minimization problem. Hence there is a tradeoff between the complexity of the recovery algorithm and the number of measurements required for recovery. We illustrate this tradeoff with the cut polytope and its relaxations.

Outline   Section 2 describes the construction of the atomic norm, gives several examples of applications in which these norms may be useful to recover simple models, and provides general conditions for recovery by minimizing the atomic norm. In Sect. 3 we investigate the number of generic measurements for exact or robust recovery using atomic norm minimization, and give estimates in a number of settings by analyzing the Gaussian widths of certain tangent cones. We address the problem of semidefinite representability and tractable relaxations of the atomic norm in Sect. 4. Section 5 describes some algorithmic issues as well as a few simulation results, and we conclude with a discussion and open questions in Sect. 6.

2 Atomic Norms and Convex Geometry

In this section we describe the construction of an atomic norm from a collection of simple atoms. In addition we give several examples of atomic norms, and discuss

their properties in the context of solving ill-posed linear inverse problems. We denote the Euclidean norm by $\| \cdot \|$.

2.1 Definition

Let $\mathcal{A}$ be a collection of atoms that is a compact subset of $\mathbb{R}^p$. We will assume throughout this paper that no element $\mathbf{a} \in \mathcal{A}$ lies in the convex hull of the other elements $\mathrm{conv}(\mathcal{A} \backslash \mathbf{a})$, i.e., the elements of $\mathcal{A}$ are the extreme points of $\mathrm{conv}(\mathcal{A})$. Let $\|\mathbf{x}\|_{\mathcal{A}}$ denote the gauge of $\mathcal{A}$ [65]:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t > 0 : x \in t \, \mathrm{conv}(\mathcal{A})\}. \tag{2}$$

Note that the gauge is always a convex, extended real-valued function for any set $\mathcal{A}$. We will assume without loss of generality that the centroid of $\mathrm{conv}(\mathcal{A})$ is at the origin, as this can be achieved by appropriate recentering. With this assumption the gauge function evaluates to $+\infty$ if $\mathbf{x}$ does not lie in the affine hull of $\mathrm{conv}(\mathcal{A})$. Further the gauge function can be rewritten as [10]:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\left\{\sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} : \mathbf{x} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a}, \ c_{\mathbf{a}} \geq 0 \ \forall \mathbf{a} \in \mathcal{A}\right\}.$$

If $\mathcal{A}$ is centrally symmetric about the origin (i.e., $\mathbf{a} \in \mathcal{A}$ if and only if $-\mathbf{a} \in \mathcal{A}$) we have that $\| \cdot \|_{\mathcal{A}}$ is a norm, which we call the *atomic norm* induced by $\mathcal{A}$. The support function of $\mathcal{A}$ is given as:

$$\|\mathbf{x}\|_{\mathcal{A}}^* = \sup\{\langle \mathbf{x}, \mathbf{a} \rangle : \mathbf{a} \in \mathcal{A}\}. \tag{3}$$

If $\| \cdot \|_{\mathcal{A}}$ is a norm the support function $\| \cdot \|_{\mathcal{A}}^*$ is the dual norm of this atomic norm. From this definition we see that the unit ball of $\| \cdot \|_{\mathcal{A}}$ is equal to $\mathrm{conv}(\mathcal{A})$. In many examples of interest the set $\mathcal{A}$ is not centrally symmetric, so that the gauge function does not define a norm. However our analysis is based on the underlying convex geometry of $\mathrm{conv}(\mathcal{A})$, and our results are applicable even if $\| \cdot \|_{\mathcal{A}}$ does not define a norm. Therefore, with an abuse of terminology, we generally refer to $\| \cdot \|_{\mathcal{A}}$ as the atomic norm of the set $\mathcal{A}$ even if $\| \cdot \|_{\mathcal{A}}$ is not a norm. We note that the duality characterization between (2) and (3) when $\| \cdot \|_{\mathcal{A}}$ is a norm is in fact applicable even in infinite-dimensional Banach spaces by Bonsall's atomic decomposition theorem [10], but our focus is on the finite-dimensional case in this work. We investigate in greater detail the issues of representability and efficient approximation of these atomic norms in Sect. 4.

Equipped with a convex penalty function given a set of atoms, we propose a convex optimization method to recover a "simple" model given limited linear measurements. Specifically suppose that $\mathbf{x}^{\star}$ is formed according to (1) from a set of atoms $\mathcal{A}$. Further suppose that we have a known linear map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$, and we have linear information about $\mathbf{x}^{\star}$ as follows:

$$\mathbf{y} = \Phi \mathbf{x}^{\star}. \tag{4}$$

The goal is to reconstruct $\mathbf{x}^\star$ given $\mathbf{y}$. We consider the following convex formulation to accomplish this task:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_{\mathcal{A}} \quad \text{s.t. } \mathbf{y} = \Phi\mathbf{x}. \tag{5}$$

When $\mathcal{A}$ is the set of one-sparse atoms this problem reduces to standard ℓ_1 norm minimization. Similarly when $\mathcal{A}$ is the set of rank-one matrices this problem reduces to nuclear norm minimization. More generally if the atomic norm $\|\cdot\|_{\mathcal{A}}$ is tractable to evaluate, then (5) potentially offers an efficient convex programming formulation for reconstructing $\mathbf{x}^\star$ from the limited information $\mathbf{y}$. The *dual problem* of (5) is given as follows:

$$\begin{aligned} \max_{\mathbf{z}} \ & \mathbf{y}^T \mathbf{z} \\ \text{s.t. } & \left\| \Phi^\dagger \mathbf{z} \right\|_{\mathcal{A}}^* \leq 1. \end{aligned} \tag{6}$$

Here $\Phi^\dagger$ denotes the adjoint (or transpose) of the linear measurement map Φ.

The convex formulation (5) can be suitably modified in case we only have access to inaccurate, noisy information. Specifically suppose that we have noisy measurements $\mathbf{y} = \Phi\mathbf{x}^\star + \omega$ where ω represents the noise term. A natural convex formulation is one in which the constraint $\mathbf{y} = \Phi\mathbf{x}$ of (5) is replaced by the relaxed constraint $\|\mathbf{y} - \Phi\mathbf{x}\| \leq \delta$, where δ is an upper bound on the size of the noise ω:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_{\mathcal{A}} \quad \text{s.t. } \|\mathbf{y} - \Phi\mathbf{x}\| \leq \delta. \tag{7}$$

We say that we have *exact recovery* in the noise-free case if $\hat{\mathbf{x}} = \mathbf{x}^\star$ in (5), and *robust recovery* in the noisy case if the error $\|\hat{\mathbf{x}} - \mathbf{x}^\star\|$ is small in (7). In Sects. 2.4 and 3 we give conditions under which the atomic norm heuristics (5) and (7) recover $\mathbf{x}^\star$ exactly or approximately. Atomic norms have found fruitful applications in problems in approximation theory of various function classes [3, 24, 44, 59]. However this prior body of work was concerned with infinite-dimensional Banach spaces, and none of these references considers or provides recovery guarantees that are applicable in our setting.

2.2 Examples

Next we provide several examples of atomic norms that can be viewed as special cases of the construction above. These norms are obtained by convexifying atomic sets that are of interest in various applications.

Sparse Vectors The problem of recovering sparse vectors from limited measurements has received a great deal of attention, with applications in many problem domains. In this case the atomic set $\mathcal{A} \subset \mathbb{R}^p$ can be viewed as the set of unit-norm one-sparse vectors $\{\pm\mathbf{e}_i\}_{i=1}^p$, and k-sparse vectors in $\mathbb{R}^p$ can be constructed using a linear combination of k elements of the atomic set. In this case it is easily seen that the convex hull $\mathrm{conv}(\mathcal{A})$ is given by the *cross-polytope* (i.e., the unit ball of the ℓ_1 norm), and the atomic norm $\|\cdot\|_{\mathcal{A}}$ corresponds to the ℓ_1 norm in $\mathbb{R}^p$.

Low-Rank Matrices   Recovering low-rank matrices from limited information is also a problem that has received considerable attention as it finds applications in problems in statistics, control, and machine learning. The atomic set $\mathcal{A}$ here can be viewed as the set of rank-one matrices of unit-Euclidean-norm. The convex hull conv($\mathcal{A}$) is the *nuclear norm ball* of matrices in which the sum of the singular values is less than or equal to one.

Sparse and Low-Rank Matrices   The problem of recovering a sparse matrix and a low-rank matrix given information about their sum arises in a number of model selection and system identification settings. The corresponding atomic norm is constructed by taking the convex hull of an atomic set obtained via the union of rank-one matrices and (suitably scaled) one-sparse matrices. This norm can also be viewed as the *infimal convolution* of the ℓ_1 norm and the nuclear norm, and its properties have been explored in [14, 19].

Permutation Matrices   A problem of interest in a ranking context [43] or an object tracking context is that of recovering permutation matrices from partial information. Suppose that a small number k of rankings of m candidates is preferred by a population. Such preferences can be modeled as the sum of a few $m \times m$ permutation matrices, with each permutation corresponding to a particular ranking. By conducting surveys of the population one can obtain partial linear information of these preferred rankings. The set $\mathcal{A}$ here is the collection of permutation matrices (consisting of $m!$ elements), and the convex hull conv($\mathcal{A}$) is the *Birkhoff polytope* or the set of doubly stochastic matrices [76]. The centroid of the Birkhoff polytope is the matrix $\mathbf{1}\mathbf{1}^T/m$, so it needs to be recentered appropriately. We mention here recent work by Jagabathula and Shah [43] on recovering a sparse function over the symmetric group (i.e., the sum of a few permutation matrices) given partial Fourier information; although the algorithm proposed in [43] is tractable, it is not based on convex optimization.

Binary Vectors   In integer programming one is often interested in recovering vectors in which the entries take on values of ± 1. Suppose that there exists such a sign-vector, and we wish to recover this vector given linear measurements. This corresponds to a version of the multiple knapsack problem [52]. In this case $\mathcal{A}$ is the set of all sign-vectors, and the convex hull conv($\mathcal{A}$) is the *hypercube* or the unit ball of the ℓ_∞ norm. The image of this hypercube under a linear map is also referred to as a zonotope [76].

Vectors from Lists   Suppose there is an unknown vector $\mathbf{x} \in \mathbb{R}^p$, and that we are given the entries of this vector without any information about the locations of these entries. For example if $\mathbf{x} = [3\ 1\ 2\ 2\ 4]'$, then we are only given the list of numbers $\{1, 2, 2, 3, 4\}$ without their positions in $\mathbf{x}$. Further suppose that we have access to a few linear measurements of $\mathbf{x}$. Can we recover $\mathbf{x}$ by solving a convex program? Such a problem is of interest in recovering partial rankings of elements of a set. An extreme case is one in which we only have two preferences for rankings, i.e., a vector in $\{1, 2\}^p$ composed only of one's and two's, which reduces to a special case of the problem above of recovering binary vectors (in which the number of entries of each sign is fixed). For this problem the set $\mathcal{A}$ is the set of all permutations of $\mathbf{x}$ (which we

know since we have the list of numbers that compose **x**), and the convex hull conv($\mathcal{A}$) is the *permutahedron* [67, 76]. As with the Birkhoff polytope, the permutahedron also needs to be recentered about the point $\mathbf{1}^T \mathbf{x}/p$.

Matrices Constrained by Eigenvalues   This problem is in a sense the noncommutative analog of the one above. Suppose that we are given the eigenvalues λ of a symmetric matrix, but no information about the eigenvectors. Can we recover such a matrix given some additional linear measurements? In this case the set $\mathcal{A}$ is the set of all symmetric matrices with eigenvalues λ, and the convex hull conv($\mathcal{A}$) is given by the *Schur–Horn orbitope* [67].

Orthogonal Matrices   In many applications matrix variables are constrained to be orthogonal, which is a nonconvex constraint and may lead to computational difficulties. We consider one such simple setting in which we wish to recover an orthogonal matrix given limited information in the form of linear measurements. In this example the set $\mathcal{A}$ is the set of $m \times m$ orthogonal matrices, and conv($\mathcal{A}$) is the *spectral norm ball*.

Measures   Recovering a measure given its moments is another question of interest that arises in system identification and statistics. Suppose one is given access to a linear combination of moments of an atomically supported measure. How can we reconstruct the support of the measure? The set $\mathcal{A}$ here is the moment curve, and its convex hull conv($\mathcal{A}$) goes by several names including the *Caratheodory orbitope* [67]. Discretized versions of this problem correspond to the set $\mathcal{A}$ being a finite number of points on the moment curve; the convex hull conv($\mathcal{A}$) is then a *cyclic polytope* [76].

Cut Matrices   In some problems one may wish to recover low-rank matrices in which the entries are constrained to take on values of ± 1. Such matrices can be used to model basic user preferences, and are of interest in problems such as collaborative filtering [68]. The set of atoms $\mathcal{A}$ could be the set of rank-one signed matrices, i.e., matrices of the form $\mathbf{z}\mathbf{z}^T$ with the entries of **z** being ± 1. The convex hull conv($\mathcal{A}$) of such matrices is the *cut polytope* [25]. An interesting issue that arises here is that the cut polytope is in general intractable to characterize. However there exist several well-known tractable semidefinite relaxations to this polytope [25, 37], and one can employ these in constructing efficient convex programs for recovering cut matrices. We discuss this point in greater detail in Sect. 4.3.

Low-Rank Tensors   Low-rank tensor decompositions play an important role in numerous applications throughout signal processing and machine learning [47]. Developing computational tools to recover low-rank tensors is therefore of great interest. In principle we could solve a tensor nuclear norm minimization problem, in which the tensor nuclear norm ball is obtained by taking the convex hull of rank-one tensors. A computational challenge here is that the tensor nuclear norm is in general intractable to compute; in order to address this problem we discuss further convex relaxations to the tensor nuclear norm using theta bodies in Sect. 4. A number of additional technical issues also arise with low-rank tensors, including the nonexistence in general of a singular value decomposition analogous to that for matrices [46], and the difference between the rank of a tensor and its border rank [23].

Nonorthogonal Factor Analysis   Suppose that a data matrix admits a factorization $X = AB$. The matrix nuclear norm heuristic will find a factorization into *orthogonal* factors in which the columns of A and rows of B are mutually orthogonal. However if a priori information is available about the factors, precision and recall could be improved by enforcing such priors. These priors may sacrifice orthogonality, but the factors might better conform with assumptions about how the data are generated. For instance in some applications one might know in advance that the factors should only take on a discrete set of values [68]. In this case, we might try to fit a sum of rank-one matrices that are bounded in ℓ_∞ norm rather than in ℓ_2 norm. Another prior that commonly arises in practice is that the factors are nonnegative (i.e., in nonnegative matrix factorization). These and other priors on the basic rank-one summands induce different norms on low-rank models than the standard nuclear norm [34], and may be better suited to specific applications.

2.3 Background on Tangent and Normal Cones

In order to properly state our results, we recall some basic concepts from convex analysis. A convex set $\mathcal{C}$ is a *cone* if it is closed under positive linear combinations. The polar $\mathcal{C}^*$ of a cone $\mathcal{C}$ is the cone

$$\mathcal{C}^* = \left\{ x \in \mathbb{R}^p : \langle x, z \rangle \leq 0 \; \forall z \in \mathcal{C} \right\}.$$

Given some nonzero $\mathbf{x} \in \mathbb{R}^p$ we define the *tangent cone* at $\mathbf{x}$ with respect to the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}} \text{conv}(\mathcal{A})$ as

$$T_{\mathcal{A}}(\mathbf{x}) = \text{cone}\left\{ \mathbf{z} - \mathbf{x} : \|\mathbf{z}\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}} \right\}. \tag{8}$$

The cone $T_{\mathcal{A}}(\mathbf{x})$ is equal to the set of *descent directions* of the atomic norm $\| \cdot \|_{\mathcal{A}}$ at the point $\mathbf{x}$, i.e., the set of all directions $\mathbf{d}$ such that the directional derivative is negative.

The *normal cone* $N_{\mathcal{A}}(\mathbf{x})$ at $\mathbf{x}$ with respect to the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}} \text{conv}(\mathcal{A})$ is defined to be the set of all directions $\mathbf{s}$ that form obtuse angles with every descent direction of the atomic norm $\| \cdot \|_{\mathcal{A}}$ at the point $\mathbf{x}$:

$$N_{\mathcal{A}}(\mathbf{x}) = \left\{ \mathbf{s} : \langle \mathbf{s}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \; \forall \mathbf{z} \text{ s.t. } \|\mathbf{z}\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}} \right\}. \tag{9}$$

The normal cone is equal to the set of all normals of hyperplanes given by normal vectors $\mathbf{s}$ that support the scaled unit ball $\|\mathbf{x}\|_{\mathcal{A}} \text{conv}(\mathcal{A})$ at $\mathbf{x}$. Observe that the polar cone of the tangent cone $T_{\mathcal{A}}(\mathbf{x})$ is the normal cone $N_{\mathcal{A}}(\mathbf{x})$ and vice versa. Moreover we have the basic characterization that the normal cone $N_{\mathcal{A}}(\mathbf{x})$ is the conic hull of the subdifferential of the atomic norm at $\mathbf{x}$.

2.4 Recovery Condition

The following result gives a characterization of the favorable underlying geometry required for exact recovery. Let $\text{null}(\Phi)$ denote the nullspace of the operator Φ.

Proposition 2.1 *We have that* $\hat{\mathbf{x}} = \mathbf{x}^\star$ *is the unique optimal solution of* (5) *if and only if* $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \{0\}$.

Proof Eliminating the equality constraints in (5) we have the equivalent optimization problem

$$\min_{\mathbf{d}} \|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} \quad \text{s.t. } \mathbf{d} \in \text{null}(\Phi).$$

Suppose $\text{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \{0\}$. Since $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} \leq \|\mathbf{x}^\star\|$ implies $\mathbf{d} \in T_{\mathcal{A}}(\mathbf{x}^\star)$, we have that $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} > \|\mathbf{x}^\star\|_{\mathcal{A}}$ for all $\mathbf{d} \in \text{null}(\Phi) \setminus \{0\}$. Conversely $\mathbf{x}^\star$ is the unique optimal solution of (5) if $\|\mathbf{x}^\star + \mathbf{d}\|_{\mathcal{A}} > \|\mathbf{x}^\star\|_{\mathcal{A}}$ for all $\mathbf{d} \in \text{null}(\Phi) \setminus \{0\}$, which implies that $\mathbf{d} \notin T_{\mathcal{A}}(\mathbf{x}^\star)$. ☐

Proposition 2.1 asserts that the atomic norm heuristic succeeds if the nullspace of the sampling operator does not intersect the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ at $\mathbf{x}^\star$. In Sect. 3 we provide a characterization of tangent cones that determines the number of Gaussian measurements required to guarantee such an empty intersection.

A tightening of this empty intersection condition can also be used to address the noisy approximation problem. The following proposition characterizes when $\mathbf{x}^\star$ can be *well approximated* using the convex program (7).

Proposition 2.2 *Suppose that we are given n noisy measurements* $\mathbf{y} = \Phi\mathbf{x}^\star + \omega$ *where* $\|\omega\| \leq \delta$ *and* $\Phi : \mathbb{R}^p \to \mathbb{R}^n$. *Let* $\hat{\mathbf{x}}$ *denote an optimal solution of* (7). *Further suppose for all* $\mathbf{z} \in T_{\mathcal{A}}(\mathbf{x}^\star)$ *that we have* $\|\Phi\mathbf{z}\| \geq \epsilon\|\mathbf{z}\|$. *Then* $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$.

Proof The set of descent directions at $\mathbf{x}^\star$ with respect to the atomic norm ball is given by the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. The error vector $\hat{\mathbf{x}} - \mathbf{x}^\star$ lies in $T_{\mathcal{A}}(\mathbf{x}^\star)$ because $\hat{\mathbf{x}}$ is a minimal atomic norm solution, and hence $\|\hat{\mathbf{x}}\|_{\mathcal{A}} \leq \|\mathbf{x}^\star\|_{\mathcal{A}}$. It follows by the triangle inequality that

$$\|\Phi(\hat{\mathbf{x}} - \mathbf{x}^\star)\| \leq \|\Phi\hat{\mathbf{x}} - \mathbf{y}\| + \|\Phi\mathbf{x}^\star - \mathbf{y}\| \leq 2\delta. \tag{10}$$

By assumption we have that

$$\|\Phi(\hat{\mathbf{x}} - \mathbf{x}^\star)\| \geq \epsilon\|\hat{\mathbf{x}} - \mathbf{x}^\star\|, \tag{11}$$

which allows us to conclude that $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$. ☐

Therefore, we need only concern ourselves with estimating the minimum value of $\frac{\|\Phi\mathbf{z}\|}{\|\mathbf{z}\|}$ for nonzero $\mathbf{z} \in T_{\mathcal{A}}(\mathbf{x}^\star)$. We denote this quantity as the *minimum gain* of the measurement operator Φ restricted to the cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. In particular if this minimum gain is bounded away from zero, then the atomic norm heuristic also provides robust recovery when we have access to noisy linear measurements of $\mathbf{x}^\star$. Minimum gain conditions have been employed in recent recovery results via ℓ_1 norm minimization, block-sparse vector recovery, and low-rank matrix reconstruction [8, 18, 54, 72]. All of these results rely heavily on strong decomposability conditions of the ℓ_1 norm and the matrix nuclear norm. However there are several examples of atomic norms (for

instance the ℓ_∞ norm and the norm induced by the Birkhoff polytope) as specified in Sect. 2.2 that do not satisfy such decomposability conditions. As we will see in the sequel, the more geometric viewpoint adopted in this paper provides a fruitful framework in which to analyze the recovery properties of general atomic norms.

2.5 Why Atomic Norm?

The atomic norm induced by a set $\mathcal{A}$ possesses a number of favorable properties that are useful for recovering "simple" models from limited linear measurements. The key point to note from Sect. 2.4 is that the smaller the tangent cone at a point $\mathbf{x}^\star$ with respect to conv($\mathcal{A}$), the easier it is to satisfy the empty intersection condition of Proposition 2.1.

Based on this observation it is desirable that points in conv($\mathcal{A}$) with smaller tangent cones correspond to simpler models, while points in conv($\mathcal{A}$) with larger tangent cones generally correspond to more complicated models. The construction of conv($\mathcal{A}$) by taking the convex hull of $\mathcal{A}$ ensures that this is the case. The extreme points of conv($\mathcal{A}$) correspond to the simplest models, i.e., those models formed from a single element of $\mathcal{A}$. Further the low-dimensional faces of conv($\mathcal{A}$) consist of those elements that are obtained by taking linear combinations of a few basic atoms from $\mathcal{A}$. These are precisely the properties desired, as points lying in these low-dimensional faces of conv($\mathcal{A}$) have smaller tangent cones than those lying on larger faces.

We also note that the atomic norm is, in some sense, the best possible convex heuristic for recovering simple models. Any reasonable heuristic penalty function should be constant on the set of atoms $\mathcal{A}$. This ensures that no atom is preferred over any other. Under this assumption, we must have that, for any $\mathbf{a} \in \mathcal{A}$, $\mathbf{a}' - \mathbf{a}$ must be a descent direction for all $\mathbf{a}' \in \mathcal{A}$. The best convex penalty function is one in which the cones of descent directions at $\mathbf{a} \in \mathcal{A}$ are as small as possible. This is because, as described above, smaller cones are more likely to satisfy the empty intersection condition required for exact recovery. Since the tangent cone at $\mathbf{a} \in \mathcal{A}$ with respect to conv($\mathcal{A}$) is precisely the conic hull of $\mathbf{a}' - \mathbf{a}$ for $\mathbf{a}' \in \mathcal{A}$, the atomic norm is the best convex heuristic for recovering models where simplicity is dictated by the set $\mathcal{A}$.

Our reasons for proposing the atomic norm as a useful convex heuristic are quite different from previous justifications of the ℓ_1 norm and the nuclear norm. In particular let $f : \mathbb{R}^p \to \mathbb{R}$ denote the cardinality function that counts the number of nonzero entries of a vector. Then the ℓ_1 norm is the *convex envelope* of f restricted to the unit ball of the ℓ_∞ norm, i.e., the best convex underestimator of f restricted to vectors in the ℓ_∞ norm ball. This view of the ℓ_1 norm in relation to the function f is often given as a justification for its effectiveness in recovering sparse vectors. However if we consider the convex envelope of f restricted to the Euclidean norm ball, then we obtain a very different convex function than the ℓ_1 norm! With more general atomic sets, it may not be clear a priori what the bounding set should be in deriving the convex envelope. In contrast the viewpoint adopted in this paper leads to a natural, unambiguous construction of the ℓ_1 norm and other general atomic norms. Further, as explained above, it is the favorable *facial structure* of the atomic norm ball that makes the atomic norm a suitable convex heuristic to recover simple models, and this connection is transparent in the definition of the atomic norm.

3 Recovery from Generic Measurements

We consider the question of using the convex program (5) to recover "simple" models formed according to (1) from a *generic* measurement operator or map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$. Specifically, we wish to compute estimates on the number of measurements n so that we have exact recovery using (5) for *most* operators comprising n measurements. That is, the measure of n-measurement operators for which recovery fails using (5) must be exponentially small. In order to conduct such an analysis we study random *Gaussian* maps whose entries are independent and identically distributed (i.i.d.) Gaussians. These measurement operators have a nullspace that is uniformly distributed among the set of all $(p-n)$-dimensional subspaces in $\mathbb{R}^p$. In particular we analyze when such operators satisfy the conditions of Proposition 2.1 and Proposition 2.2 for exact recovery.

3.1 Recovery Conditions Based on Gaussian Width

Proposition 2.1 requires that the nullspace of the measurement operator Φ must miss the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$. Gordon [38] gave a solution to the problem of characterizing the probability that a random subspace (of some fixed dimension) distributed uniformly misses a cone. We begin by defining the Gaussian width of a set, which plays a key role in Gordon's analysis.

Definition 3.1 The *Gaussian width* of a set $S \subset \mathbb{R}^p$ is defined as:

$$w(S) := \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{z} \in S} \mathbf{g}^T \mathbf{z}\right],$$

where $\mathbf{g} \sim \mathcal{N}(0, I)$ is a vector of independent zero-mean unit-variance Gaussians.

Gordon characterized the likelihood that a random subspace misses a cone $\mathcal{C}$ purely in terms of the dimension of the subspace and the Gaussian width $w(\mathcal{C} \cap \mathbb{S}^{p-1})$, where $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ is the unit sphere. Before describing Gordon's result formally, we introduce some notation. Let λ_k denote the expected length of a k-dimensional Gaussian random vector. By elementary integration, we have that $\lambda_k = \sqrt{2}\Gamma(\frac{k+1}{2})/\Gamma(\frac{k}{2})$. Further, by induction, one can show that λ_k is tightly bounded as $\frac{k}{\sqrt{k+1}} \le \lambda_k \le \sqrt{k}$.

The main idea underlying Gordon's theorem is a bound on the minimum gain of an operator restricted to a set. Specifically, recall that $\mathrm{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \{0\}$ is the condition required for recovery by Proposition 2.1. Thus if we have that the minimum gain of Φ restricted to vectors in the set $T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ is bounded away from zero, then it is clear that $\mathrm{null}(\Phi) \cap T_{\mathcal{A}}(\mathbf{x}^\star) = \emptyset$. We refer to such minimum gains restricted to a subset of the sphere as *restricted minimum singular values*, and the following theorem of Gordon gives a bound on these quantities.

Theorem 3.2 (Corollary 1.2, [38]) *Let Ω be a closed subset of $\mathbb{S}^{p-1}$. Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ be a random map with i.i.d. zero-mean Gaussian entries having variance one. Then*

$$\mathbb{E}\left[\min_{\mathbf{z} \in \Omega} \|\Phi \mathbf{z}\|_2\right] \ge \lambda_n - w(\Omega). \tag{12}$$

Theorem 3.2 allows us to characterize exact recovery in the noise-free case using the convex program (5), and robust recovery in the noisy case using the convex program (7). Specifically we consider the number of measurements required for exact or robust recovery when the measurement map $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ consists of i.i.d. zero-mean Gaussian entries having variance $1/n$. The normalization of the variance ensures that the columns of Φ are approximately unit-norm, and is necessary in order to properly define a signal-to-noise ratio. The following corollary summarizes the main results of interest in our setting.

Corollary 3.3 *Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ be a random map with i.i.d. zero-mean Gaussian entries having variance $1/n$. Further let $\Omega = T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ denote the spherical part of the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$.*

1. *Suppose that we have measurements $\mathbf{y} = \Phi \mathbf{x}^\star$ and solve the convex program (5). Then $\mathbf{x}^\star$ is the unique optimum of (5) with probability at least $1 - \exp(-\frac{1}{2}[\lambda_n - w(\Omega)]^2)$ provided*

$$n \geq w(\Omega)^2 + 1.$$

2. *Suppose that we have noisy measurements $\mathbf{y} = \Phi \mathbf{x}^\star + \omega$, with the noise ω bounded as $\|\omega\| \leq \delta$, and that we solve the convex program (7). Letting $\hat{\mathbf{x}}$ denote the optimal solution of (7), we have that $\|\mathbf{x}^\star - \hat{\mathbf{x}}\| \leq \frac{2\delta}{\epsilon}$ with probability at least $1 - \exp(-\frac{1}{2}[\lambda_n - w(\Omega) - \sqrt{n}\epsilon]^2)$ provided*

$$n \geq \frac{w(\Omega)^2 + 3/2}{(1 - \epsilon)^2}.$$

Proof The two results are simple consequences of Theorem 3.2 and a concentration of measure argument. Recall that for a function $f : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant L and a random Gaussian vector, $\mathbf{g} \in \mathbb{R}^d$, with mean zero and identity variance,

$$\mathbb{P}\big[f(\mathbf{g}) \geq \mathbb{E}[f] - t\big] \geq 1 - \exp\left(-\frac{t^2}{2L^2}\right) \tag{13}$$

(see, for example, [49, 60]). For any set $\Omega \subset \mathbb{S}^{p-1}$, the function

$$\Phi \mapsto \min_{\mathbf{z} \in \Omega} \|\Phi \mathbf{z}\|_2$$

is Lipschitz with respect to the Frobenius norm with constant 1. Thus, applying Theorem 3.2 and (13), we find that

$$\mathbb{P}\Big[\min_{\mathbf{z} \in \Omega} \|\Phi \mathbf{z}\|_2 \geq \epsilon\Big] \geq 1 - \exp\left(-\frac{1}{2}\big(\lambda_n - w(\Omega) - \sqrt{n}\epsilon\big)^2\right) \tag{14}$$

provided that $\lambda_n - w(\Omega) - \sqrt{n}\epsilon \geq 0$.

The first part now follows by setting $\epsilon = 0$ in (14). The concentration inequality is valid provided that $\lambda_n \geq w(\Omega)$. To verify this, note that

$$\lambda_n \geq \frac{n}{\sqrt{n+1}} \geq \sqrt{\frac{w(\Omega)^2 + 1}{1 + 1/n}} \geq \sqrt{\frac{w(\Omega)^2 + w(\Omega)^2/n}{1 + 1/n}} = w(\Omega).$$

Here, both inequalities use the fact that $n \geq w(\Omega)^2 + 1$.

For the second part, we have from (14) that

$$\left\| \Phi(\mathbf{z}) \right\| = \|\mathbf{z}\| \left\| \Phi\left(\frac{\mathbf{z}}{\|\mathbf{z}\|} \right) \right\| \geq \epsilon \|\mathbf{z}\|$$

for all $\mathbf{z} \in \mathcal{T}_{\mathcal{A}}(\mathbf{x}^\star)$ with high probability if $\lambda_n \geq w(\Omega) + \sqrt{n}\epsilon$. In this case, we can apply Proposition 2.2 to conclude that $\|\hat{\mathbf{x}} - \mathbf{x}^\star\| \leq \frac{2\delta}{\epsilon}$. To verify that concentration of measure can be applied is more or less the same procedure as in the proof of Part 1. First note that, under the assumptions of the theorem,

$$w(\Omega)^2 + 1 \leq n(1-\epsilon)^2 - 1/2 \leq n(1-\epsilon)^2 - 2\epsilon(1-\epsilon) + \frac{\epsilon^2}{n} = \left(\sqrt{n}(1-\epsilon) - \frac{\epsilon}{\sqrt{n}} \right)^2$$

as $\epsilon(1-\epsilon) \leq 1/4$ for $\epsilon \in (0,1)$. Using this fact, we then have

$$\lambda_n - \sqrt{n}\epsilon \geq \frac{n - (n+1)\epsilon}{\sqrt{n+1}} \geq \sqrt{\frac{w(\Omega)^2 + 1}{1 + 1/n}} \geq w(\Omega)$$

as desired. $\qquad\qquad\square$

Gordon's theorem thus provides a simple characterization of the number of measurements required for reconstruction with the atomic norm. Indeed the Gaussian width of $\Omega = T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}$ is the only quantity that we need to compute in order to obtain bounds for both exact and robust recovery. Unfortunately it is generally not easy to compute Gaussian widths. Rudelson and Vershynin [66] have worked out Gaussian widths for the special case of tangent cones at sparse vectors on the boundary of the ℓ_1 ball, and derived results for sparse vector recovery using ℓ_1 minimization that improve upon previous results. In the next section we give various well-known properties of the Gaussian width that are useful in computations. In Sect. 3.3 we discuss a new approach to width computations that gives near-optimal recovery bounds in a variety of settings.

3.2 Properties of Gaussian Width

The Gaussian width has deep connections to convex geometry. Since the length and direction of a Gaussian random vector are independent, one can verify that for $S \subset \mathbb{R}^p$

$$w(S) = \frac{\lambda_p}{2} \int_{\mathbb{S}^{p-1}} \left(\max_{\mathbf{z} \in S} \mathbf{u}^T \mathbf{z} - \min_{\mathbf{z} \in S} \mathbf{u}^T \mathbf{z} \right) d\mathbf{u} = \frac{\lambda_p}{2} b(S)$$

where the integral is with respect to the Haar measure on $\mathbb{S}^{p-1}$ and $b(S)$ is known as the *mean width* of S. The mean width measures the average length of S along unit directions in $\mathbb{R}^p$ and is one of the fundamental *intrinsic volumes* of a body studied in combinatorial geometry [45]. Any continuous valuation that is invariant under rigid motions and homogeneous of degree 1 is a multiple of the mean width and hence a

multiple of the Gaussian width. We can use this connection with convex geometry to underscore several properties of the Gaussian width that are useful for computation.

The Gaussian width of a body is invariant under translations and unitary transformations. Moreover it is homogeneous in the sense that $w(tK) = tw(K)$ for $t > 0$. The width is also monotonic. If $S_1 \subseteq S_2 \subseteq \mathbb{R}^p$, then it is clear from the definition of the Gaussian width that

$$w(S_1) \leq w(S_2).$$

Less obviously, the width is modular in the sense that if S_1 and S_2 are convex bodies with $S_1 \cup S_2$ convex, we also have

$$w(S_1 \cup S_2) + w(S_1 \cap S_2) = w(S_1) + w(S_2).$$

This equality follows from the fact that w is a valuation [4]. Also note that if we have a set $S \subseteq \mathbb{R}^p$, then the Gaussian width of S is equal to the Gaussian width of the convex hull of S:

$$w(S) = w(\text{conv}(S)).$$

This result follows from the basic fact in convex analysis that the maximum of a convex function over a convex set is achieved at an extreme point of the convex set.

If $V \subset \mathbb{R}^p$ is a subspace in $\mathbb{R}^p$, then we have that

$$w(V \cap \mathbb{S}^{p-1}) = \sqrt{\dim(V)},$$

which follows from standard results on random Gaussians. This result also agrees with the intuition that a random Gaussian map Φ misses a k-dimensional subspace with high probability as long as $\dim(\text{null}(\Phi)) \geq k + 1$. Finally, if a cone $S \subset \mathbb{R}^p$ is such that $S = S_1 \oplus S_2$, where $S_1 \subset \mathbb{R}^p$ is a k-dimensional cone, $S_2 \subset \mathbb{R}^p$ is a $(p-k)$-dimensional cone that is orthogonal to S_1, and $\oplus$ denotes the direct sum operation, then the width can be decomposed as follows:

$$w(S \cap \mathbb{S}^{p-1})^2 \leq \mathbb{E}_{\mathbf{g}}\left[\left(\sup_{\mathbf{z} \in S_1 \cap \mathbb{S}^{p-1}} \mathbf{g}^T \mathbf{z}\right)^2\right] + \mathbb{E}_{\mathbf{g}}\left[\left(\sup_{\mathbf{z} \in S_2 \cap \mathbb{S}^{p-1}} \mathbf{g}^T \mathbf{z}\right)^2\right].$$

Here $\mathbf{g} \sim \mathcal{N}(0, I)$ is as usual a vector of independent zero-mean unit-variance Gaussians. These observations are useful in a variety of situations. For example a width computation that frequently arises is one in which $S = S_1 \oplus S_2$ as described above, with S_1 being a k-dimensional subspace. It follows that the width of $S \cap \mathbb{S}^{p-1}$ is bounded as

$$w(S \cap \mathbb{S}^{p-1})^2 \leq k + \mathbb{E}_{\mathbf{g}}\left[\left(\sup_{\mathbf{z} \in S_2 \cap \mathbb{S}^{p-1}} \mathbf{g}^T \mathbf{z}\right)^2\right]. \tag{15}$$

Another tool for computing Gaussian widths is based on Dudley's inequality [32, 49], which bounds the width of a set in terms of the covering number of the set at all scales.

Definition 3.4 Let S be an arbitrary compact subset of $\mathbb{R}^p$. The *covering number* of S in the Euclidean norm at resolution ϵ is the smallest number, $\mathfrak{N}(S, \epsilon)$, such that $\mathfrak{N}(S, \epsilon)$ Euclidean balls of radius ϵ cover S.

Theorem 3.5 (Dudley's Inequality) *Let S be an arbitrary compact subset of $\mathbb{R}^p$, and let $\mathbf{g}$ be a random vector with i.i.d. zero-mean, unit-variance Gaussian entries. Then*

$$w(S) \leq 24 \int_0^\infty \sqrt{\log\big(\mathfrak{N}(S, \epsilon)\big)} \, d\epsilon. \tag{16}$$

We note here that a weak converse to Dudley's inequality can be obtained via Sudakov's minoration [49] by using the covering number for just a single scale. Specifically we have the following *lower bound* on the Gaussian width of a compact subset $S \subset \mathbb{R}^p$ for any $\epsilon > 0$:

$$w(S) \geq c\epsilon \sqrt{\log\big(\mathfrak{N}(S, \epsilon)\big)}.$$

Here $c > 0$ is some universal constant.

Although Dudley's inequality can be applied quite generally, estimating covering numbers is difficult in most instances. There are a few simple characterizations available for spheres and Sobolev spaces, and some tractable arguments based on Maurey's empirical method [49]. However it is not evident how to compute these numbers for general convex cones. Also, in order to apply Dudley's inequality we need to estimate the covering number at all scales. Dudley's inequality can also be quite loose in its estimates, and it often introduces extraneous polylogarithmic factors. In the next section we describe a new mechanism for estimating Gaussian widths, which provides near-optimal guarantees for recovery of sparse vectors and low-rank matrices, as well as for several of the recovery problems discussed in Sect. 3.4.

3.3 New Results on Gaussian Width

We now present a framework for computing Gaussian widths by bounding the Gaussian width of a cone via the distance to the dual cone. To be fully general let $\mathcal{C}$ be a nonempty convex cone in $\mathbb{R}^p$, and let $\mathcal{C}^*$ denote the polar of $\mathcal{C}$. We can then upperbound the Gaussian width of any cone $\mathcal{C}$ in terms of the polar cone $\mathcal{C}^*$.

Proposition 3.6 *Let $\mathcal{C}$ be any nonempty convex cone in $\mathbb{R}^p$, and let $\mathbf{g} \sim \mathcal{N}(0, I)$ be a random Gaussian vector. Then we have the following bound*:

$$w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big) \leq \mathbb{E}_{\mathbf{g}}\big[\text{dist}\big(\mathbf{g}, \mathcal{C}^*\big)\big],$$

where dist *here denotes the Euclidean distance between a point and a set.*

The proof is given in Appendix A, and it follows from an appeal to convex duality. Proposition 3.6 is more or less a restatement of the fact that the support function of a convex cone is equal to the distance to its polar cone. As it is the square of the

Gaussian width that is of interest to us (see Corollary 3.3), it is often useful to apply Jensen's inequality to make the following approximation:

$$\mathbb{E}_{\mathbf{g}}\big[\text{dist}\big(\mathbf{g}, \mathcal{C}^*\big)\big]^2 \leq \mathbb{E}_{\mathbf{g}}\big[\text{dist}\big(\mathbf{g}, \mathcal{C}^*\big)^2\big]. \tag{17}$$

The inspiration for our characterization in Proposition 3.6 of the width of a cone in terms of the expected distance to its dual came from the work of Stojnic [69], who used linear programming duality to construct Gaussian-width-based estimates for analyzing recovery in sparse reconstruction problems. Specifically, Stojnic's relatively simple approach recovered well-known phase transitions in sparse signal recovery [29], and also generalized to block-sparse signals and other forms of structured sparsity.

This new dual characterization yields a number of useful bounds on the Gaussian width, which we describe here. In the following section we use these bounds to derive new recovery results. The first result is a bound on the Gaussian width of a cone in terms of the Gaussian width of its polar.

Lemma 3.7 *Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a nonempty closed, convex cone. Then we have that*

$$w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big)^2 + w\big(\mathcal{C}^* \cap \mathbb{S}^{p-1}\big)^2 \leq p.$$

Proof Combining Proposition 3.6 and (17), we have that

$$w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big)^2 \leq \mathbb{E}_{\mathbf{g}}\big[\text{dist}\big(\mathbf{g}, \mathcal{C}^*\big)^2\big],$$

where as before $\mathbf{g} \sim \mathcal{N}(0, I)$. For any $\mathbf{z} \in \mathbb{R}^p$ we let $\Pi_{\mathcal{C}}(\mathbf{z}) = \arg\inf_{\mathbf{u} \in \mathcal{C}} \|\mathbf{z} - \mathbf{u}\|$ denote the projection of $\mathbf{z}$ onto $\mathcal{C}$. From standard results in convex analysis [65], we note that one can decompose any $\mathbf{z} \in \mathbb{R}^p$ into orthogonal components as follows:

$$\mathbf{z} = \Pi_{\mathcal{C}}(\mathbf{z}) + \Pi_{\mathcal{C}^*}(\mathbf{z}), \qquad \langle \Pi_{\mathcal{C}}(\mathbf{z}), \Pi_{\mathcal{C}^*}(\mathbf{z})\rangle = 0.$$

Therefore we have the following sequence of bounds:

$$\begin{aligned}
w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big)^2 &\leq \mathbb{E}_{\mathbf{g}}\big[\text{dist}\big(\mathbf{g}, \mathcal{C}^*\big)^2\big] \\
&= \mathbb{E}_{\mathbf{g}}\big[\big\|\Pi_{\mathcal{C}}(\mathbf{g})\big\|^2\big] \\
&= \mathbb{E}_{\mathbf{g}}\big[\|\mathbf{g}\|^2 - \big\|\Pi_{\mathcal{C}^*}(\mathbf{g})\big\|^2\big] \\
&= p - \mathbb{E}_{\mathbf{g}}\big[\big\|\Pi_{\mathcal{C}^*}(\mathbf{g})\big\|^2\big] \\
&= p - \mathbb{E}_{\mathbf{g}}\big[\text{dist}(\mathbf{g}, \mathcal{C})^2\big] \\
&\leq p - w\big(\mathcal{C}^* \cap \mathbb{S}^{p-1}\big)^2. \qquad \square
\end{aligned}$$

In many recovery problems one is interested in computing the width of a self-dual cone. For such cones the following corollary to Lemma 3.7 gives a simple solution.

Corollary 3.8 *Let* $\mathcal{C} \subset \mathbb{R}^p$ *be a self-dual cone, i.e.,* $\mathcal{C} = -\mathcal{C}^*$. *Then we have that*

$$w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big)^2 \leq \frac{p}{2}.$$

Proof The proof follows directly from Lemma 3.7 as $w(\mathcal{C} \cap \mathbb{S}^{p-1})^2 = w(\mathcal{C}^* \cap \mathbb{S}^{p-1})^2$. □

Our next bound for the width of a cone $\mathcal{C}$ is based on the volume of its polar $\mathcal{C}^* \cap \mathbb{S}^{p-1}$. The *volume* of a measurable subset of the sphere is the fraction of the sphere $\mathbb{S}^{p-1}$ covered by the subset. Thus it is a quantity between zero and one.

Theorem 3.9 (Gaussian Width from Volume of the Polar) *Let* $\mathcal{C} \subseteq \mathbb{R}^p$ *be any closed, convex, solid cone, and suppose that its polar* $\mathcal{C}^*$ *is such that* $\mathcal{C}^* \cap \mathbb{S}^{p-1}$ *has a volume of* $\Theta \in [0, 1]$. *Then for* $p \geq 9$ *we have that*

$$w\big(\mathcal{C} \cap \mathbb{S}^{p-1}\big) \leq 3\sqrt{\log\left(\frac{4}{\Theta}\right)}.$$

The proof of this theorem is given in Appendix B. The main property that we appeal to in the proof is *Gaussian isoperimetry*. In particular there is a formal sense in which a spherical cap[1] is the "extremal case" among all subsets of the sphere with a given volume Θ. Other than this observation the proof mainly involves a sequence of integral calculations.

Note that if we are given a specification of a cone $\mathcal{C} \subset \mathbb{R}^p$ in terms of a membership oracle, it is possible to efficiently obtain good numerical estimates of the volume of $\mathcal{C} \cap \mathbb{S}^{p-1}$ [33]. Moreover, simple symmetry arguments often give relatively accurate estimates of these volumes. Such estimates can then be put into Theorem 3.9 to yield bounds on the width.

3.4 New Recovery Bounds

We use the bounds derived in the last section to obtain new recovery results. First, using the dual characterization of the Gaussian width in Proposition 3.6, we are able to obtain sharp bounds on the number of measurements required for recovering sparse vectors and low-rank matrices from random Gaussian measurements using convex optimization (i.e., ℓ_1 norm and nuclear norm minimization).

Proposition 3.10 *Let* $\mathbf{x}^\star \in \mathbb{R}^p$ *be an s-sparse vector. Letting* $\mathcal{A}$ *denote the set of unit-Euclidean-norm one-sparse vectors, we have that*

$$w\big(T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}\big)^2 \leq 2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s.$$

Thus, $2s \log(\frac{p}{s}) + \frac{5}{4}s + 1$ *random Gaussian measurements suffice to recover* $\mathbf{x}^\star$ *via* ℓ_1 *norm minimization with high probability.*

---

[1]A *spherical cap* is a subset of the sphere obtained by intersecting the sphere $\mathbb{S}^{p-1}$ with a halfspace.

Proposition 3.11 *Let $\mathbf{x}^\star$ be an $m_1 \times m_2$ rank-r matrix with $m_1 \leq m_2$. Letting $\mathcal{A}$ denote the set of unit-Euclidean-norm rank-one matrices, we have that*

$$w\left(T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{m_1 m_2 - 1}\right)^2 \leq 3r(m_1 + m_2 - r).$$

Thus $3r(m_1 + m_2 - r) + 1$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via nuclear norm minimization with high probability.

The proofs of these propositions are given in Appendix C. The number of measurements required by these bounds is on the same order as previously known results. For sparse vectors, previous results obtaining $2s \log(p/s)$ were asymptotic [26, 30, 74]. Our bounds, in contrast, hold with high probability in finite dimensions. For low-rank matrices, our bound provides sharper constants than those previously derived (for example [15]) and is also applicable over a wider range of matrix ranks and number of measurements than in previous work [64]. We also note that we have robust recovery at these thresholds. Further these results do not require explicit recourse to any type of restricted isometry property [15], and the proofs are simple and based on elementary integrals.

Next we obtain a set of recovery results by appealing to Corollary 3.8 on the width of a self-dual cone. These examples correspond to the recovery of individual atoms (i.e., the extreme points of the set $\mathrm{conv}(\mathcal{A})$), although the same machinery is applicable in principle to estimate the number of measurements required to recover models formed as sums of a few atoms (i.e., points lying on low-dimensional faces of $\mathrm{conv}(\mathcal{A})$). We first obtain a well-known result on the number of measurements required for recovering sign-vectors via ℓ_∞ norm minimization.

Proposition 3.12 *Let $\mathcal{A} \in \{-1, +1\}^p$ be the set of sign-vectors in $\mathbb{R}^p$. Suppose $\mathbf{x}^\star \in \mathbb{R}^p$ is a vector formed as a convex combination of k sign-vectors in $\mathcal{A}$ such that $\mathbf{x}^\star$ lies on a k-face of the ℓ_∞ norm unit ball. Then we have that*

$$w\left(T_{\mathcal{A}}(\mathbf{x}^\star) \cap \mathbb{S}^{p-1}\right)^2 \leq \frac{p+k}{2}.$$

Thus $\frac{p+k}{2}$ random Gaussian measurements suffice to recover $\mathbf{x}^\star$ via ℓ_∞ norm minimization with high probability.

Proof The tangent cone at $\mathbf{x}^\star$ with respect to the ℓ_∞ norm ball is the direct sum of a k-dimensional subspace and a (rotated) $(p-k)$-dimensional nonnegative orthant. As the orthant is self-dual, we obtain the required bound by combining Corollary 3.8 and (15). □

This result agrees with previously computed bounds in [31, 52], which relied on a more complicated combinatorial argument. Next we compute the number of measurements required to recover orthogonal matrices via spectral norm minimization (see Sect. 2.2). Let $\mathbb{O}(m)$ denote the group of $m \times m$ orthogonal matrices, viewed as a subgroup of the set of nonsingular matrices in $\mathbb{R}^{m \times m}$.

Proposition 3.13 *Let* $\mathbf{x}^\star \in \mathbb{R}^{m \times m}$ *be an orthogonal matrix, and let* $\mathcal{A}$ *be the set of all orthogonal matrices. Then we have that*

$$w\big(T_\mathcal{A}(\mathbf{x}^\star) \cap \mathbb{S}^{m^2-1}\big)^2 \leq \frac{3m^2 - m}{4}.$$

Thus $\frac{3m^2-m}{4}$ *random Gaussian measurements suffice to recover* $\mathbf{x}^\star$ *via spectral norm minimization with high probability.*

Proof Due to the symmetry of the orthogonal group, it suffices to consider the tangent cone at the identity matrix I with respect to the spectral norm ball. Recall that the spectral norm ball is the convex hull of the orthogonal matrices. Therefore the *tangent space* at the identity matrix with respect to the orthogonal group $\mathbb{O}(m)$ is a subset of the tangent cone $T_\mathcal{A}(I)$. It is well known that this tangent space is the Lie Algebra of all $m \times m$ skew-symmetric matrices. Thus we only need to compute the component S of $T_\mathcal{A}(I)$ that lies in the subspace of symmetric matrices:

$$\begin{aligned}
S &= \text{cone}\{M - I : \|M\|_\mathcal{A} \leq 1, \ M \text{ symmetric}\} \\
&= \text{cone}\big\{UDU^T - UU^T : \|D\|_\mathcal{A} \leq 1, \ D \text{ diagonal}, \ U \in \mathbb{O}(m)\big\} \\
&= \text{cone}\big\{U(D - I)U^T : \|D\|_\mathcal{A} \leq 1, \ D \text{ diagonal}, \ U \in \mathbb{O}(m)\big\} \\
&= -\text{PSD}_m.
\end{aligned}$$

Here PSD_m denotes the set of $m \times m$ symmetric positive semidefinite matrices. As this cone is self-dual, we can apply Corollary 3.8 in conjunction with the observations in Sect. 3.2 to conclude that

$$w\big(T_\mathcal{A}(I) \cap \mathbb{S}^{m^2-1}\big)^2 \leq \binom{m}{2} + \frac{1}{2}\binom{m+1}{2} = \frac{3m^2 - m}{4}.$$

$\square$

We note that the number of degrees of freedom in an $m \times m$ orthogonal matrix (i.e., the dimension of the manifold of orthogonal matrices) is $\frac{m(m-1)}{2}$. Propositions 3.12 and 3.13 point to the importance of obtaining recovery bounds with sharp constants. Larger constants in either result would imply that the number of measurements required exceeds the ambient dimension of the underlying $\mathbf{x}^\star$. In these and many other cases of interest Gaussian width arguments not only give order-optimal recovery results, but also provide precise constants that result in sharp recovery thresholds.

Finally we give a third set of recovery results that appeal to the Gaussian width bound of Theorem 3.9. The following measurement bound applies to cases when $\text{conv}(\mathcal{A})$ is a *symmetric polytope* (roughly speaking, all the vertices are "equivalent"), and is a simple corollary of Theorem 3.9.

Corollary 3.14 *Suppose that the set* $\mathcal{A}$ *is a finite collection of* m *points, with the convex hull* $\text{conv}(\mathcal{A})$ *being a vertex-transitive polytope* [76] *whose vertices are the points in* $\mathcal{A}$. *Using the convex program* (5) *we have that* $9\log(m)$ *random Gaussian*

measurements suffice, *with high probability*, *for exact recovery of a point in* $\mathcal{A}$, *i.e.*, *a vertex of* conv($\mathcal{A}$).

Proof We recall the basic fact from convex analysis that the normal cones at the vertices of a convex polytope in $\mathbb{R}^p$ provide a partitioning of $\mathbb{R}^p$. As conv($\mathcal{A}$) is a vertex-transitive polytope, the normal cone at a vertex covers $\frac{1}{m}$ fraction of $\mathbb{R}^p$. Applying Theorem 3.9, we have the desired result. □

Clearly we require the number of vertices to be bounded as $m \leq \exp\{\frac{p}{9}\}$, so that the estimate of the number of measurements is not vacuously true. This result has useful consequences in settings in which conv($\mathcal{A}$) is a *combinatorial polytope*, as such polytopes are often vertex-transitive. We have the following example on the number of measurements required to recover permutation matrices.[2]

Proposition 3.15 *Let* $\mathbf{x}^\star \in \mathbb{R}^{m \times m}$ *be a permutation matrix, and let* $\mathcal{A}$ *be the set of all* $m \times m$ *permutation matrices. Then* $9m \log(m)$ *random Gaussian measurements suffice*, *with high probability*, *to recover* $\mathbf{x}^\star$ *by solving the optimization problem* (5), *which minimizes the norm induced by the Birkhoff polytope of doubly stochastic matrices.*

Proof This result follows from Corollary 3.14 by noting that there are $m!$ permutation matrices of size $m \times m$. □

4 Representability and Algebraic Geometry of Atomic Norms

All of our discussion thus far has focused on arbitrary atomic sets $\mathcal{A}$. As seen in Sect. 2 the geometry of the convex hull conv($\mathcal{A}$) completely determines conditions under which exact recovery is possible using the convex program (5). In this section we address the question of computing atomic norms for general sets of atoms. These issues are critical in order to be able to solve the convex optimization problem (5). Although the convex hull conv($\mathcal{A}$) is always a mathematically well-defined object, testing membership in this set is generally undecidable (for example, if $\mathcal{A}$ is a fractal). Further, even if these convex hulls are computable they may not admit efficient representations. For example if $\mathcal{A}$ is the set of rank-one signed matrices (see Sect. 2.2), the corresponding convex hull conv($\mathcal{A}$) is the cut polytope for which there is no known tractable characterization. Consequently one may have to resort to efficiently computable approximations of conv($\mathcal{A}$). The tradeoff in using such approximations in our atomic norm minimization framework is that we require more measurements for robust recovery. This section is devoted to providing a better understanding of these issues.

---

[2]While Proposition 3.15 follows as a consequence of the general result in Corollary 3.14, one can remove the constant factor 9 in the statement of Proposition 3.15 by carrying out a more refined analysis of the Birkhoff polytope.

4.1 Role of Algebraic Structure

In order to obtain exact or approximate representations (analogous to the cases of the ℓ_1 norm and the nuclear norm) it is important to identify properties of the atomic set $\mathcal{A}$ that can be exploited computationally. We focus on cases in which the set $\mathcal{A}$ has algebraic structure. Specifically let the ring of multivariate polynomials in p variables be denoted by $\mathbb{R}[\mathbf{x}] = \mathbb{R}[\mathbf{x}_1, \ldots, \mathbf{x}_p]$. We then consider real algebraic varieties [9].

Definition 4.1 A *real algebraic variety* $S \subseteq \mathbb{R}^p$ is the set of real solutions of a system of polynomial equations:

$$S = \big\{ \mathbf{x} : g_j(\mathbf{x}) = 0, \ \forall j \big\},$$

where $\{g_j\}$ is a finite collection of polynomials in $\mathbb{R}[\mathbf{x}]$.

Indeed all of the atomic sets $\mathcal{A}$ considered in this paper are examples of algebraic varieties. Algebraic varieties have the remarkable property that (the closure of) their convex hull can be arbitrarily well-approximated in a constructive manner as (the projection of) a set defined by linear matrix inequality constraints [39, 58]. A potential complication may arise, however, if these semidefinite representations are intractable to compute in polynomial time. In such cases it is possible to approximate the convex hulls via a hierarchy of tractable semidefinite relaxations. We describe these results in more detail in Sect. 4.2. Therefore the atomic norm minimization problems such as (7) arising in such situations can be solved exactly or approximately via semidefinite programming.

Algebraic structure also plays a second important role in atomic norm minimization problems. If an atomic norm $\| \cdot \|_{\mathcal{A}}$ is intractable to compute, we may approximate it via a more tractable norm $\| \cdot \|_{\text{app}}$. However not every approximation of the atomic norm is equally good for solving inverse problems. As illustrated in Fig. 2 we can construct approximations of the ℓ_1 ball that are tight in a *metric* sense, with $(1 - \epsilon) \| \cdot \|_{\text{app}} \le \| \cdot \|_{\ell_1} \le (1 + \epsilon) \| \cdot \|_{\text{app}}$, but where the tangent cones at sparse vectors in the new norm are halfspaces. In such a case, the number of measurements required to recover the sparse vector ends up being on the same order as the ambient dimension. (Note that the ℓ_1 norm is in fact tractable to compute; we simply use it here for illustrative purposes.) The key property that we seek in approximations

Fig. 2 The convex body given by the *dotted line* is a good metric approximation to the ℓ_1 ball. However as its "corners" are "smoothed out," the tangent cone at $\mathbf{x}^\star$ goes from being a proper cone (with respect to the ℓ_1 ball) to a halfspace (with respect to the approximation)
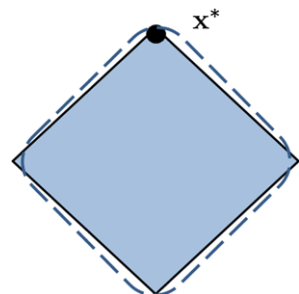
to an atomic norm $\|\cdot\|_{\mathcal{A}}$ is that they *preserve algebraic structure* such as the vertices/extreme points and more generally the low-dimensional faces of the conv($\mathcal{A}$). As discussed in Sect. 2.5, points on such low-dimensional faces correspond to simple models, and algebraic-structure-preserving approximations ensure that the tangent cones at simple models with respect to the approximations are not too much larger than the corresponding tangent cones with respect to the original atomic norms (see Sect. 4.3 for a concrete example).

4.2 Semidefinite Relaxations Using Theta Bodies

In this section we give a family of semidefinite relaxations to the atomic norm minimization problem whenever the atomic set has algebraic structure. To begin with if we approximate the atomic norm $\|\cdot\|_{\mathcal{A}}$ by another atomic norm $\|\cdot\|_{\tilde{\mathcal{A}}}$ defined using a *larger* collection of atoms $\mathcal{A} \subseteq \tilde{\mathcal{A}}$, it is clear that

$$\|\cdot\|_{\tilde{\mathcal{A}}} \leq \|\cdot\|_{\mathcal{A}}.$$

Consequently outer approximations of the atomic set give rise to approximate norms that provide lower bounds on the optimal value of the problem (5).

In order to provide such lower bounds on the optimal value of (5), we discuss semidefinite relaxations of the convex hull conv($\mathcal{A}$). All of our discussion here is based on results described in [39] for semidefinite relaxations of convex hulls of algebraic varieties using theta bodies. We only give a brief review of the relevant constructions, and refer the reader to the vast literature on this subject for more details (see [39, 58] and the references therein). For subsequent reference in this section, we recall the definition of a polynomial ideal [9, 41].

Definition 4.2 A *polynomial ideal* $I \subset \mathbb{R}[\mathbf{x}]$ is a subset of the ring of polynomials that contains the zero polynomial (the polynomial that is identically zero), is closed under addition, and has the property that $f \in I, g \in \mathbb{R}[\mathbf{x}]$ implies that $f \cdot g \in I$.

To begin with we note that a *sum-of-squares* (SOS) polynomial in $\mathbb{R}[\mathbf{x}]$ is a polynomial that can be written as the (finite) sum of squares of other polynomials in $\mathbb{R}[\mathbf{x}]$. Verifying the nonnegativity of a multivariate polynomial is intractable in general; therefore SOS polynomials play an important role in real algebraic geometry, as an SOS polynomial is easily seen to be nonnegative everywhere. Further checking whether a polynomial is an SOS polynomial can be accomplished efficiently via semidefinite programming [58].

Turning our attention to the description of the convex hull of an algebraic variety, we will assume for simplicity that the convex hull is closed. Let $I \subseteq \mathbb{R}[\mathbf{x}]$ be a polynomial ideal, and let $V_{\mathbb{R}}(I) \in \mathbb{R}^p$ be its real algebraic variety:

$$V_{\mathbb{R}}(I) = \big\{ \mathbf{x} : f(\mathbf{x}) = 0, \ \forall f \in I \big\}.$$

One can then show that the convex hull conv($V_{\mathbb{R}}(I)$) is given as:

$$\mathrm{conv}\big(V_{\mathbb{R}}(I) \big) = \big\{ \mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear and nonnegative on } V_{\mathbb{R}}(I) \big\}$$

$$= \bigl\{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear s.t. } f = h + g, \ \forall h \text{ nonnegative}, \forall g \in I \bigr\}$$

$$= \bigl\{\mathbf{x} : f(\mathbf{x}) \geq 0, \forall f \text{ linear s.t. } f \text{ nonnegative modulo } I \bigr\}.$$

A linear polynomial here is one that has a maximum degree of one, and the meaning of "modulo an ideal" is clear. As nonnegativity modulo an ideal may be intractable to check, we can consider a relaxation to a polynomial being SOS modulo an ideal, i.e., a polynomial that can be written as $\sum_{i=1}^{q} h_i^2 + g$ for g in the ideal. Since it is tractable to check via semidefinite programming whether bounded-degree polynomials are SOS, the k-th theta body of an ideal I is defined as follows in [39]:

$$\mathrm{TH}_k(I) = \bigl\{\mathbf{x} : f(\mathbf{x}) \geq 0, \ \forall f \text{ linear s.t. } f \text{ is } k\text{-sos modulo } I \bigr\}.$$

Here k-sos refers to an SOS polynomial in which the components in the SOS decomposition have degree at most k. The k-th theta body $\mathrm{TH}_k(I)$ is a convex relaxation of $\mathrm{conv}(V_{\mathbb{R}}(I))$, and one can verify that

$$\mathrm{conv}\bigl(V_{\mathbb{R}}(I)\bigr) \subseteq \cdots \subseteq \mathrm{TH}_{k+1}(I) \subseteq \mathrm{TH}_k\bigl(V_{\mathbb{R}}(I)\bigr).$$

By the arguments given above (see also [39]) these theta bodies can be described using semidefinite programs of size polynomial in k. Hence by considering theta bodies $\mathrm{TH}_k(I)$ with increasingly larger k, one can obtain a hierarchy of tighter semidefinite relaxations of $\mathrm{conv}(V_{\mathbb{R}}(I))$. We also note that in many cases of interest such semidefinite relaxations preserve low-dimensional faces of the convex hull of a variety, although these properties are not known in general. We will use some of these properties below when discussing approximations of the cut polytope.

Approximating Tensor Norms   We conclude this section with an example application of these relaxations to the problem of approximating the tensor nuclear norm. For notational simplicity we focus on the case of tensors of order three that lie in $\mathbb{R}^{m \times m \times m}$, i.e., tensors indexed by three numbers, although our discussion is applicable more generally. In particular the atomic set $\mathcal{A}$ is the set of unit-Euclidean-norm rank-one tensors:

$$\mathcal{A} = \bigl\{\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} : \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^m, \ \|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1 \bigr\}$$

$$= \bigl\{N \in \mathbb{R}^{m^3} : N = \mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}, \ \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^m, \|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1 \bigr\},$$

where $\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}$ is the tensor product of three vectors. Note that the second description is written as the projection onto $\mathbb{R}^{m^3}$ of a variety defined in $\mathbb{R}^{m^3+3m}$. The nuclear norm is then given by (2), and is intractable to compute in general. Now let $I_{\mathcal{A}}$ denote a polynomial ideal of polynomial maps from $\mathbb{R}^{m^3+3m}$ to $\mathbb{R}$:

$$I_{\mathcal{A}} = \Biggl\{ g : g = \sum_{i,j,k=1}^{m} g_{ijk}(N_{ijk} - \mathbf{u}_i \mathbf{v}_j \mathbf{w}_k) + g_u\bigl(\mathbf{u}^T \mathbf{u} - 1\bigr) + g_v\bigl(\mathbf{v}^T \mathbf{v} - 1\bigr)$$

$$+ g_w\bigl(\mathbf{w}^T \mathbf{w} - 1\bigr), \forall g_{ijk}, g_u, g_v, g_w \Biggr\}.$$

Here $g_u, g_v, g_w, \{g_{ijk}\}_{i,j,k}$ are polynomials in the variables N, $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$. Following the program described above for constructing approximations, a family of semidefinite relaxations to the tensor nuclear norm ball can be prescribed in this manner via the theta bodies $\mathrm{TH}_k(I_\mathcal{A})$.

4.3 Tradeoff Between Relaxation and Number of Measurements

As discussed in Sect. 2.5 the atomic norm is the best convex heuristic for solving ill-posed linear inverse problems of the type considered here. However we may wish to approximate the atomic norm in cases when it is intractable to compute exactly, and the discussion in the preceding section provides one approach for constructing a family of relaxations. As one might expect, the tradeoff for using such approximations, i.e., a *weaker* convex heuristic than the atomic norm, is an increase in the number of measurements required for exact or robust recovery. The reason is that the approximate norms have *larger* tangent cones at their extreme points, which makes it harder to satisfy the empty intersection condition of Proposition 2.1. We highlight this tradeoff here with an illustrative example involving the cut polytope.

The cut polytope is defined as the convex hull of all cut matrices:

$$\mathcal{P} = \mathrm{conv}\big\{\mathbf{z}\mathbf{z}^T : \mathbf{z} \in \{-1, +1\}^m\big\}.$$

As described in Sect. 2.2 low-rank matrices that are composed of ± 1's as entries are of interest in collaborative filtering [68], and the norm induced by the cut polytope is a potential convex heuristic for recovering such matrices from limited measurements. However it is well known that the cut polytope is intractable to characterize [25], and therefore we need to use tractable relaxations instead. We consider the following two relaxations of the cut polytope. The first is the popular relaxation that is used in semidefinite approximations of the MAX-CUT problem:

$$\mathcal{P}_1 = \{M : M \text{ symmetric}, \ M \succeq 0, \ M_{ii} = 1, \ \forall i = 1, \ldots, p\}.$$

This is the well-studied elliptope [25], and it can be interpreted as the second theta body relaxation (see Sect. 4.2) of the cut polytope $\mathcal{P}$ [39]. We also investigate the performance of a second, weaker relaxation:

$$\mathcal{P}_2 = \big\{M : M \text{ symmetric}, \ M_{ii} = 1, \ \forall i, \ |M_{ij}| \leq 1, \ \forall i \neq j\big\}.$$

This polytope is simply the convex hull of symmetric matrices with ± 1's in the off-diagonal entries, and 1's on the diagonal. We note that $\mathcal{P}_2$ is an extremely weak relaxation of $\mathcal{P}$, but we use it here only for illustrative purposes. It is easily seen that

$$\mathcal{P} \subset \mathcal{P}_1 \subset \mathcal{P}_2,$$

with all the inclusions being strict. Figure 3 gives a toy sketch that highlights all the main geometric aspects of these relaxations. In particular $\mathcal{P}_1$ has many more extreme points than $\mathcal{P}$, although the set of vertices of $\mathcal{P}_1$, i.e., points that have full-dimensional normal cones, is precisely the set of cut matrices (which are the vertices of $\mathcal{P}$) [25]. The convex polytope $\mathcal{P}_2$ contains many more vertices compared to $\mathcal{P}$ as shown in

Fig. 3 A toy sketch illustrating the cut polytope $\mathcal{P}$ and the two approximations $\mathcal{P}_1$ and $\mathcal{P}_2$. Note that $\mathcal{P}_1$ is a sketch of the standard semidefinite relaxation that has the *same* vertices as $\mathcal{P}$. On the other hand $\mathcal{P}_2$ is a polyhedral approximation to $\mathcal{P}$ that has many more vertices
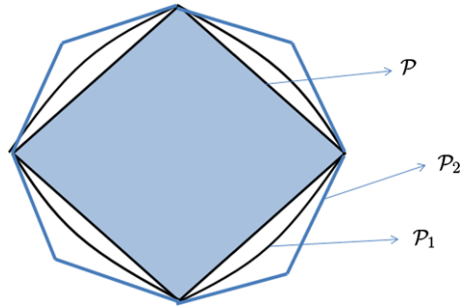

Fig. 3. As expected the tangent cones at vertices of $\mathcal{P}$ become increasingly larger as we use successively weaker relaxations. The following result summarizes the number of random measurements required for recovering a cut matrix, i.e., a rank-one sign matrix, using the norms induced by each of these convex bodies.

Proposition 4.3 *Suppose $\mathbf{x}^\star \in \mathbb{R}^{m \times m}$ is a rank-one sign matrix, i.e., a cut matrix, and we are given n random Gaussian measurements of $\mathbf{x}^\star$. We wish to recover $\mathbf{x}^\star$ by solving a convex program based on the norms induced by each of $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2$. We have exact recovery of $\mathbf{x}^\star$ in each of these cases with high probability under the following conditions on the number of measurements*:

1. *Using $\mathcal{P}$: $n = \mathcal{O}(m)$.*
2. *Using $\mathcal{P}_1$: $n = \mathcal{O}(m)$.*
3. *Using $\mathcal{P}_2$: $n = \frac{m^2 - m}{4}$.*

Proof For the first part, we note that $\mathcal{P}$ is a symmetric polytope with 2^{m-1} vertices. Therefore we can apply Corollary 3.14 to conclude that $n = \mathcal{O}(m)$ measurements suffices for exact recovery.

For the second part we note that the tangent cone at $\mathbf{x}^\star$ with respect to the nuclear norm ball of $m \times m$ matrices contains within it the tangent cone at $\mathbf{x}^\star$ with respect to the polytope $\mathcal{P}_1$. Hence we appeal to Proposition 3.11 to conclude that $n = \mathcal{O}(m)$ measurements suffices for exact recovery.

Finally, we note that $\mathcal{P}_2$ is essentially the hypercube in $\binom{m}{2}$ dimensions. Appealing to Proposition 3.12, we conclude that $n = \frac{m^2 - m}{4}$ measurements suffices for exact recovery. □

It is not too hard to show that these bounds are order-optimal, and that they cannot be improved. Thus this particular instance rigorously demonstrates that the number of measurements required for exact recovery increases as the relaxations get weaker (and as the tangent cones get larger). The principle underlying this illustration holds more generally, namely that there exists a tradeoff between the complexity of the convex heuristic and the number of measurements required for exact or robust recovery. It would be of interest to quantify this tradeoff in other settings, for example, in problems in which we use increasingly tighter relaxations of the atomic norm via theta bodies.

We also note that the tractable relaxation based on $\mathcal{P}_1$ is only off by a constant factor with respect to the optimal heuristic based on the cut polytope $\mathcal{P}$. This suggests the potential for tractable heuristics to approximate hard atomic norms with provable approximation ratios, akin to methods developed in the literature on approximation algorithms for hard combinatorial optimization problems.

4.4 Terracini's Lemma and Lower Bounds on Recovery

Algebraic structure in the atomic set $\mathcal{A}$ also provides a means for computing *lower bounds* on the number of measurements required for exact recovery. The recovery condition of Proposition 2.1 states that the nullspace null(Φ) of the measurement operator $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^n$ must miss the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ at the point of interest $\mathbf{x}^\star$. Suppose that this tangent cone contains a q-dimensional subspace. It is then clear from straightforward linear algebra arguments that the number of measurements n must exceed q. Indeed this bound must hold for *any* linear measurement scheme. Thus the dimension of the subspace contained inside the tangent cone (i.e., the dimension of the lineality space) provides a simple lower bound on the number of linear measurements.

In this section we discuss a method to obtain estimates of the dimension of a subspace component of the tangent cone. We focus again on the setting in which $\mathcal{A}$ is an algebraic variety. Indeed in all of the examples of Sect. 2.2, the atomic set $\mathcal{A}$ is an algebraic variety. In such cases simple models $\mathbf{x}^\star$ formed according to (1) can be viewed as elements of *secant varieties*.

Definition 4.4 Let $\mathcal{A} \in \mathbb{R}^p$ be an algebraic variety. Then the k'th *secant variety* $\mathcal{A}^k$ is defined as the union of all affine spaces passing through any $k + 1$ points of $\mathcal{A}$.

Secant varieties and their tangent spaces have been extensively studied in algebraic geometry [41]. A particular question of interest is to characterize the dimensions of secant varieties and tangent spaces. In our context, estimates of these dimensions are useful in giving lower bounds on the number of measurements required for recovery. Specifically we have the following result, which states that certain linear spaces must lie in the tangent cone at $\mathbf{x}^\star$ with respect to conv($\mathcal{A}$).

Proposition 4.5 *Let $\mathcal{A} \subset \mathbb{R}^p$ be a smooth variety, and let $\mathcal{T}(\mathbf{u}, \mathcal{A})$ denote the tangent space at any $\mathbf{u} \in \mathcal{A}$ with respect to $\mathcal{A}$. Suppose $\mathbf{x} = \sum_{i=1}^{k} c_i \mathbf{a}_i$, $\forall \mathbf{a}_i \in \mathcal{A}$, $c_i \geq 0$, such that*

$$\|\mathbf{x}\|_{\mathcal{A}} = \sum_{i=1}^{k} c_i.$$

Then the tangent cone $T_{\mathcal{A}}(\mathbf{x}^\star)$ *contains the following linear space*:

$$\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A}) \subset T_{\mathcal{A}}(\mathbf{x}^\star),$$

where $\oplus$ denotes the direct sum of subspaces.

Proof We note that if we perturb $\mathbf{a}_1$ slightly to *any* neighboring $\mathbf{a}'_1$ so that $\mathbf{a}'_1 \in \mathcal{A}$, then the resulting $\mathbf{x}' = c_1 \mathbf{a}'_1 + \sum_{i=2}^k c_2 \mathbf{a}_i$ is such that $\|\mathbf{x}'\|_{\mathcal{A}} \leq \|\mathbf{x}\|_{\mathcal{A}}$. The proposition follows directly from this observation.   □

This result is applicable, for example, when $\mathcal{A}$ is the variety of rank-one matrices or the variety of rank-one tensors, as these are smooth varieties. By Terracini's lemma [41] from algebraic geometry the subspace $\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A})$ is in fact the estimate for the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ at $\mathbf{x}$ with respect to the $(k-1)$'th secant variety $\mathcal{A}^{k-1}$:

Proposition 4.6 (Terracini's Lemma) *Let $\mathcal{A} \subset \mathbb{R}^p$ be a smooth affine variety, and let $\mathcal{T}(\mathbf{u}, \mathcal{A})$ denote the tangent space at any $\mathbf{u} \in \mathcal{A}$ with respect to $\mathcal{A}$. Suppose $\mathbf{x} \in \mathcal{A}^{k-1}$ is a generic point such that $\mathbf{x} = \sum_{i=1}^k c_i \mathbf{a}_i$, $\forall \mathbf{a}_i \in \mathcal{A}, c_i \geq 0$. Then the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ at $\mathbf{x}$ with respect to the secant variety $\mathcal{A}^{k-1}$ is given by $\mathcal{T}(\mathbf{a}_1, \mathcal{A}) \oplus \cdots \oplus \mathcal{T}(\mathbf{a}_k, \mathcal{A})$. Moreover the dimension of $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ is at most (and is expected to be) $\min\{p, (k+1)\dim(\mathcal{A}) + k\}$.*

Combining these results we have that estimates of the dimension of the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ lead directly to lower bounds on the number of measurements required for recovery. The intuition here is clear as the number of measurements required must be bounded below by the number of "degrees of freedom," which is captured by the dimension of the tangent space $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$. However Terracini's lemma provides us with general estimates of the dimension of $\mathcal{T}(\mathbf{x}, \mathcal{A}^{k-1})$ for generic points $\mathbf{x}$. Therefore we can directly obtain lower bounds on the number of measurements, purely by considering the dimension of the variety $\mathcal{A}$ and the number of elements from $\mathcal{A}$ used to construct $\mathbf{x}$ (i.e., the order of the secant variety in which $\mathbf{x}$ lies). As an example the dimension of the base variety of normalized order-three tensors in $\mathbb{R}^{m \times m \times m}$ is $3(m-1)$. Consequently, in principle, if we were to solve the tensor nuclear norm minimization problem, we should expect to require at least $\mathcal{O}(km)$ measurements to recover a rank-k tensor.

5 Computational Experiments

5.1 Algorithmic Considerations

While a variety of atomic norms can be represented or approximated by linear matrix inequalities, these representations do not necessarily translate into practical implementations. Semidefinite programming can be technically solved in polynomial time, but general interior point solvers typically only scale to problems with a few hundred variables. For larger scale problems, it is often preferable to exploit structure in the atomic set $\mathcal{A}$ to develop fast, first-order algorithms.

A starting point for first-order algorithm design lies in determining the structure of the proximity operator (or Moreau envelope) associated with the atomic norm,

$$\Pi_{\mathcal{A}}(\mathbf{x}; \mu) := \arg\min_{\mathbf{z}} \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|^2 + \mu \|\mathbf{z}\|_{\mathcal{A}}. \tag{18}$$

Here μ is some positive parameter. Proximity operators have already been harnessed for fast algorithms involving the ℓ_1 norm [20, 21, 35, 40, 73] and the nuclear norm [12, 51, 71] where these maps can be quickly computed in closed form. For the ℓ_1 norm, the ith component of $\Pi_{\mathcal{A}}(\mathbf{x}; \mu)$ is given by

$$\Pi_{\mathcal{A}}(\mathbf{x}; \mu)_i = \begin{cases} \mathbf{x}_i + \mu, & \mathbf{x}_i < -\mu, \\ 0, & -\mu \le \mathbf{x}_i \le \mu, \\ \mathbf{x}_i - \mu, & \mathbf{x}_i > \mu. \end{cases} \tag{19}$$

This is called the *soft thresholding* operator. For the nuclear norm, $\Pi_{\mathcal{A}}$ soft thresholds the singular values. In either case, the only structure necessary for the cited algorithms to converge is the convexity of the norm. Indeed, essentially any algorithm developed for ℓ_1 or nuclear norm minimization can in principle be adapted for atomic norm minimization. One simply needs to apply the operator $\Pi_{\mathcal{A}}$ wherever a shrinkage operation was previously applied.

For a concrete example, suppose f is a smooth function, and consider the optimization problem

$$\min_{\mathbf{x}} \ f(\mathbf{x}) + \mu \|\mathbf{x}\|_{\mathcal{A}}. \tag{20}$$

The classical projected gradient method for this problem alternates between taking steps along the gradient of f and then applying the proximity operator associated with the atomic norm. Explicitly, the algorithm consists of the iterative procedure

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{A}}\big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k); \alpha_k \lambda\big) \tag{21}$$

where $\{\alpha_k\}$ is a sequence of positive stepsizes. Under very mild assumptions, this iteration can be shown to converge to a stationary point of (20) [36]. When f is convex, the returned stationary point is a globally optimal solution. Recently, Nesterov has described a particular variant of this algorithm that is guaranteed to converge at a rate no worse than $O(k^{-1})$, where k is the iteration counter [57]. Moreover, he proposes simple enhancements of the standard iteration to achieve an $O(k^{-2})$ convergence rate for convex f and a linear rate of convergence for strongly convex f.

If we apply the projected gradient method to the regularized inverse problem

$$\min_{\mathbf{x}} \ \|\Phi \mathbf{x} - \mathbf{y}\|^2 + \lambda \|\mathbf{x}\|_{\mathcal{A}} \tag{22}$$

then the algorithm reduces to the straightforward iteration

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{A}}\big(\mathbf{x}_k + \alpha_k \Phi^{\dagger}(\mathbf{y} - \Phi \mathbf{x}_k); \alpha_k \lambda\big). \tag{23}$$

Here (22) is equivalent to (7) for an appropriately chosen $\lambda > 0$ and is useful for estimation from noisy measurements.

The basic (noiseless) atomic norm minimization problem (5) can be solved by minimizing a sequence of instances of (22) with monotonically decreasing values of λ. Each subsequent minimization is initialized from the point returned by the previous step. Such an approach corresponds to the classic *method of multipliers* [6] and has proven effective for solving problems regularized by the ℓ_1 norm and for total variation denoising [13, 75].

This discussion demonstrates that when the proximity operator associated with some atomic set $\mathcal{A}$ can be easily computed, then efficient first-order algorithms are immediate. For novel atomic norm applications, one can thus focus on algorithms and techniques to compute the associated proximity operators. From a computational perspective, it may be easier to compute the proximity operator via dual atomic norm. Associated to each proximity operator is the dual operator

$$\Lambda_{\mathcal{A}}(\mathbf{x}; \mu) = \arg\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 \quad \text{s.t. } \|\mathbf{y}\|_{\mathcal{A}}^* \le \mu. \tag{24}$$

By an appropriate change of variables, $\Lambda_{\mathcal{A}}$ is nothing more than the projection of $\mu^{-1}\mathbf{x}$ onto the unit ball in the dual atomic norm:

$$\Lambda_{\mathcal{A}}(\mathbf{x}; \mu) = \arg\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \mu^{-1}\mathbf{x}\|^2 \quad \text{s.t. } \|\mathbf{y}\|_{\mathcal{A}}^* \le 1. \tag{25}$$

From convex programming duality, we have $\mathbf{x} = \Pi_{\mathcal{A}}(\mathbf{x}; \mu) + \Lambda_{\mathcal{A}}(\mathbf{x}; \mu)$. This can be seen by observing

$$\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \mu\|\mathbf{z}\|_{\mathcal{A}} = \min_{\mathbf{z}} \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \le \mu} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \langle \mathbf{y}, \mathbf{z} \rangle \tag{26}$$

$$= \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \le \mu} \min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \langle \mathbf{y}, \mathbf{z} \rangle \tag{27}$$

$$= \max_{\|\mathbf{y}\|_{\mathcal{A}}^* \le \mu} -\frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{x}\|^2. \tag{28}$$

In particular, $\Pi_{\mathcal{A}}(\mathbf{x}; \mu)$ and $\Lambda_{\mathcal{A}}(\mathbf{x}; \mu)$ form a complementary primal-dual pair for this optimization problem. Hence, we only need to be able to efficiently compute the Euclidean projection onto the dual norm ball to compute the proximity operator associated with the atomic norm.

Finally, though the proximity operator provides an elegant framework for algorithm generation, many other possible algorithmic approaches may be employed to take advantage of the particular structure of an atomic set $\mathcal{A}$. For instance, we can rewrite (24) as

$$\Lambda_{\mathcal{A}}(\mathbf{x}; \mu) = \arg\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \mu^{-1}\mathbf{x}\|^2 \quad \text{s.t. } \langle \mathbf{y}, \mathbf{a} \rangle \le 1 \quad \forall \mathbf{a} \in \mathcal{A}. \tag{29}$$

Suppose we have access to a procedure that, given $\mathbf{z} \in \mathbb{R}^n$, can decide whether $\langle \mathbf{z}, \mathbf{a} \rangle \le 1$ for all $\mathbf{a} \in \mathcal{A}$, or can find a violated constraint where $\langle \mathbf{z}, \hat{\mathbf{a}} \rangle > 1$. In this case, we can apply a cutting plane method or an ellipsoid method to solve (24) or (6) [56, 61]. Similarly, if it is simpler to compute a subgradient of the atomic norm than it is to compute a proximity operator, then the standard subgradient method [7, 56] can be applied to solve problems of the form (22). Each computational scheme will have different advantages and drawbacks for specific atomic sets, and relative effectiveness needs to be evaluated on a case-by-case basis.

5.2 Simulation Results

We describe the results of numerical experiments in recovering orthogonal matrices, permutation matrices, and rank-one sign matrices (i.e., cut matrices) from random linear measurements by solving convex optimization problems. All the atomic norm minimization problems in these experiments are solved using a combination of the SDPT3 package [70] and the YALMIP parser [50].

Orthogonal Matrices We consider the recovery of 20×20 orthogonal matrices from random Gaussian measurements via *spectral norm minimization*. Specifically we solve the convex program (5), with the atomic norm being the spectral norm. Figure 4 gives a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

Permutation Matrices We consider the recovery of 20×20 permutation matrices from random Gaussian measurements. We solve the convex program (5), with the atomic norm being the norm induced by the Birkhoff polytope of 20×20 doubly stochastic matrices. Figure 4 gives a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

Cut Matrices We consider the recovery of 20×20 cut matrices from random Gaussian measurements. As the cut polytope is intractable to characterize, we solve the convex program (5) with the atomic norm being approximated by the norm induced by the semidefinite relaxation $\mathcal{P}_1$ described in Sect. 4.3. Recall that this is the second theta body associated with the convex hull of cut matrices, and so this experiment verifies that objects can be recovered from theta body approximations. Figure 4 gives
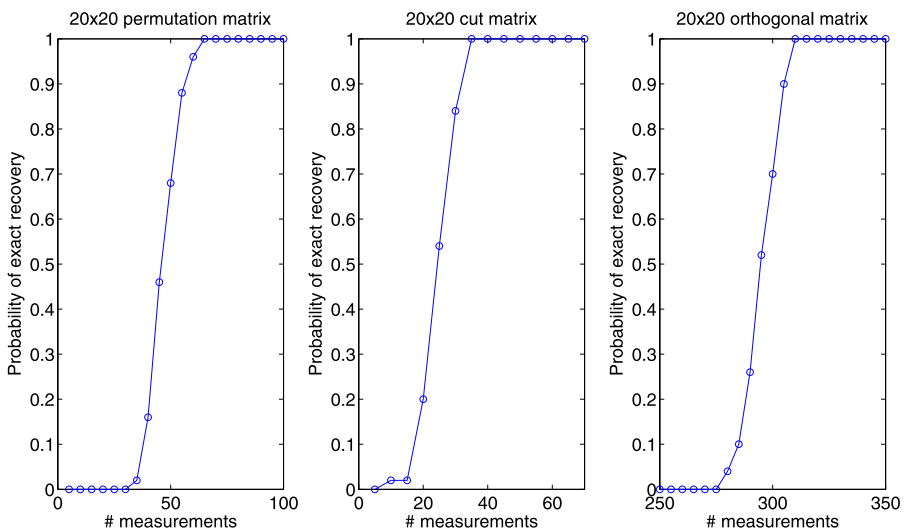

Fig. 4 Plots of the number of measurements available versus the probability of exact recovery (computed over 50 trials) for various models

a plot of the probability of exact recovery (computed over 50 random trials) versus the number of measurements required.

In each of these experiments we see agreement between the observed phase transitions, and the theoretical predictions (Propositions 3.13, 3.15, and 4.3) of the number of measurements required for exact recovery. In particular the phase transition in Fig. 4 for the number of measurements required to recover an orthogonal matrix is very close to the prediction $n \approx \frac{3m^2-m}{4} = 295$ of Proposition 3.13. We refer the reader to [29, 52, 63] for similar phase transition plots for recovering sparse vectors, low-rank matrices, and signed vectors from random measurements via convex optimization.

6 Conclusions and Future Directions

This manuscript has illustrated that, for a fixed set of base atoms, the atomic norm is the best choice of a convex regularizer for solving ill-posed inverse problems with the prescribed priors. With this in mind, our results in Sects. 3 and 4 outline methods for computing hard limits on the number of measurements required for recovery from *any* convex heuristic. Using the calculus of Gaussian widths, such bounds can be computed in a relatively straightforward fashion, especially if one can appeal to notions of convex duality and symmetry. This computational machinery of widths and dimension counting is surprisingly powerful: near-optimal bounds on estimating sparse vectors and low-rank matrices from partial information follow from elementary integration. Thus we expect that our new bounds concerning symmetric, vertex-transitive polytopes are also nearly tight. Moreover algebraic reasoning allowed us to explore the inherent tradeoffs between computational efficiency and measurement demands. More complicated algorithms for atomic norm regularization might extract structure from less information, but approximation algorithms are often sufficient for near-optimal reconstructions.

This report serves as a foundation for many new exciting directions in inverse problems, and we close our discussion with a description of several natural possibilities for future work.

Width Calculations for More Atomic Sets The calculus of Gaussian widths described in Sect. 3 provides the building blocks for computing the Gaussian widths for the application examples discussed in Sect. 2. We have not yet exhaustively estimated the widths in all of these examples, and a thorough cataloging of the measurement demands associated with different prior information would provide a more complete understanding of the fundamental limits of solving underdetermined inverse problems. Moreover our list of examples is by no means exhaustive. The framework developed in this paper provides a compact and efficient methodology for constructing regularizers from very general prior information, and new regularizers can be easily created by translating grounded expert knowledge into new atomic norms.

Recovery Bounds for Structured Measurements Our recovery results focus on generic measurements because, for a general set $\mathcal{A}$, it does not make sense to delve

into specific measurement ensembles. Particular structures of the measurement matrix Φ will depend on the application and the atomic set $\mathcal{A}$. For instance, in compressed sensing, much work focuses on randomly sampled Fourier coefficients [16] and random Toeplitz and circulant matrices [42, 62]. With low-rank matrices, several authors have investigated reconstruction from a small collection of entries [17]. In all of these cases, some notion of *incoherence* plays a crucial role, quantifying the amount of information garnered from each row of Φ. It would be interesting to explore how to appropriately generalize notions of incoherence to new applications. Is there a particular definition that is general enough to encompass most applications? Or do we need a specialized concept to match the specifics of each atomic norm?

Quantifying the Loss Due to Relaxation Section 4.3 illustrates how the choice of approximation of a particular atomic norm can dramatically alter the number of measurements required for recovery. However, as was the case for vertices of the cut polytope, some relaxations incur only a very modest increase in measurement demands. Using techniques similar to those employed in the study of semidefinite relaxations of hard combinatorial problems, is it possible to provide a more systematic method to estimate the number of measurements required to recover points from polynomial-time computable norms?

Atomic Norm Decompositions While the techniques of Sects. 3 and 4 provide bounds on the estimation of points in low-dimensional secant varieties of atomic sets, they do not provide a procedure for actually constructing decompositions. That is, we have provided bounds on the number of measurements required to recover points $\mathbf{x}$ of the form

$$\mathbf{x} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a}$$

when the coefficient sequence $\{c_{\mathbf{a}}\}$ is sparse, but we do not provide any methods for actually recovering c itself. These decompositions are useful, for instance, in actually computing the rank-one binary vectors optimized in semidefinite relaxations of combinatorial algorithms [2, 37, 55], or in the computation of tensor decompositions from incomplete data [47]. Is it possible to use algebraic structure to generate deterministic or randomized algorithms for reconstructing the atoms that underlie a vector $\mathbf{x}$, especially when approximate norms are used?

Large-Scale Algorithms Finally, we think that the most fruitful extensions of this work lie in a thorough exploration of the empirical performance and efficacy of atomic norms on large-scale inverse problems. The proposed algorithms in Sect. 5 require only the knowledge of the proximity operator of an atomic norm, or a Euclidean projection operator onto the dual norm ball. Using these design principles and the geometry of particular atomic norms should enable the scaling of atomic norm techniques to massive data sets.

Appendix A: Proof of Proposition 3.6

Proof First note that the Gaussian width can be upper-bounded as follows:

$$w\left(\mathcal{C} \cap \mathbb{S}^{p-1}\right) \leq \mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{z} \in \mathcal{C} \cap \mathcal{B}(0,1)} \mathbf{g}^T \mathbf{z}\right], \tag{30}$$

where $\mathcal{B}(0, 1)$ denotes the unit Euclidean ball. The expression on the right-hand side inside the expected value can be expressed as the optimal value of the following convex optimization problem for each $\mathbf{g} \in \mathbb{R}^p$:

$$\begin{aligned} \max_{\mathbf{z}} \ & \mathbf{g}^T \mathbf{z} \\ \text{s.t.} \ & \mathbf{z} \in \mathcal{C}, \\ & \|\mathbf{z}\|^2 \leq 1. \end{aligned} \tag{31}$$

We now proceed to form the dual problem of (31) by first introducing the Lagrangian

$$\mathcal{L}(\mathbf{z}, \mathbf{u}, \gamma) = \mathbf{g}^T \mathbf{z} + \gamma\left(1 - \mathbf{z}^T \mathbf{z}\right) - \mathbf{u}^T \mathbf{z}$$

where $\mathbf{u} \in \mathcal{C}^*$ and $\gamma \geq 0$ is a scalar. To obtain the dual problem we maximize the Lagrangian with respect to $\mathbf{z}$, which amounts to setting

$$\mathbf{z} = \frac{1}{2\gamma}(\mathbf{g} - \mathbf{u}).$$

Putting this into the Lagrangian above gives the dual problem

$$\begin{aligned} \min \ & \gamma + \frac{1}{4\gamma}\|\mathbf{g} - \mathbf{u}\|^2 \\ \text{s.t.} \ & \mathbf{u} \in \mathcal{C}^*, \\ & \gamma \geq 0. \end{aligned}$$

Solving this optimization problem with respect to γ we find that $\gamma = \frac{1}{2}\|\mathbf{g} - \mathbf{u}\|$, which gives the dual problem to (31)

$$\begin{aligned} \min \ & \|\mathbf{g} - \mathbf{u}\| \\ \text{s.t.} \ & \mathbf{u} \in \mathcal{C}^*. \end{aligned} \tag{32}$$

Under very mild assumptions about $\mathcal{C}$, the optimal value of (32) is equal to that of (31) (for example as long as $\mathcal{C}$ has a nonempty relative interior, strong duality holds).

Hence we have derived

$$\mathbb{E}_{\mathbf{g}}\left[\sup_{\mathbf{z}\in\mathcal{C}\cap\mathcal{B}(0,1)}\mathbf{g}^T\mathbf{z}\right] = \mathbb{E}_{\mathbf{g}}\big[\mathrm{dist}\big(\mathbf{g},\mathcal{C}^*\big)\big]. \tag{33}$$

This equation combined with the bound (30) gives us the desired result.   □

Appendix B: Proof of Theorem 3.9

Proof We set $\beta = \frac{1}{\Theta}$. First note that if $\beta \geq \frac{1}{4}\exp\{\frac{p}{9}\}$ then the width bound exceeds $\sqrt{p}$, which is the maximal possible value for the width of $\mathcal{C}$. Thus, we will assume throughout that $\beta \leq \frac{1}{4}\exp\{\frac{p}{9}\}$.

Using Proposition 3.6 we need to upper-bound the expected distance to the polar cone. Let $\mathbf{g} \sim \mathcal{N}(0, I)$ be a normally distributed random vector. Then the norm of $\mathbf{g}$ is independent from the angle of $\mathbf{g}$. That is, $\|\mathbf{g}\|$ is independent from $\mathbf{g}/\|\mathbf{g}\|$. Moreover $\mathbf{g}/\|\mathbf{g}\|$ is distributed as a uniform sample on $\mathbb{S}^{p-1}$, and $\mathbb{E}_{\mathbf{g}}[\|\mathbf{g}\|] \leq \sqrt{p}$. Thus we have

$$\mathbb{E}_{\mathbf{g}}\big[\mathrm{dist}\big(\mathbf{g},\mathcal{C}^*\big)\big] \leq \mathbb{E}_{\mathbf{g}}\big[\|\mathbf{g}\| \cdot \mathrm{dist}\big(\mathbf{g}/\|\mathbf{g}\|,\mathcal{C}^*\cap\mathbb{S}^{p-1}\big)\big] \leq \sqrt{p}\,\mathbb{E}_{\mathbf{u}}\big[\mathrm{dist}\big(\mathbf{u},\mathcal{C}^*\cap\mathbb{S}^{p-1}\big)\big] \tag{34}$$

where $\mathbf{u}$ is sampled uniformly on $\mathbb{S}^{p-1}$.

To bound the latter quantity, we will use isoperimetry. Suppose A is a subset of $\mathbb{S}^{p-1}$ and B is a spherical cap with the same volume as A. Let $N(A, r)$ denote the locus of all points in the sphere of Euclidean distance at most r from the set A. Let μ denote the Haar measure on $\mathbb{S}^{p-1}$ and let $\mu(A; r)$ denote the measure of $N(A, r)$. Then spherical isoperimetry states that $\mu(A; r) \geq \mu(B; r)$ for all $r \geq 0$ (see, for example, [48, 53]).

Let B now denote a spherical cap with $\mu(B) = \mu(\mathcal{C}^*\cap\mathbb{S}^{p-1})$. Then we have

$$\mathbb{E}_{\mathbf{u}}\big[\mathrm{dist}\big(\mathbf{u},\mathcal{C}^*\cap\mathbb{S}^{p-1}\big)\big] = \int_0^\infty \mathbb{P}\big[\mathrm{dist}\big(\mathbf{u},\mathcal{C}^*\cap\mathbb{S}^{p-1}\big) > t\big]\,\mathrm{d}t \tag{35}$$

$$= \int_0^\infty \big(1 - \mu\big(\mathcal{C}^*\cap\mathbb{S}^{p-1}; t\big)\big)\,\mathrm{d}t \tag{36}$$

$$\leq \int_0^\infty \big(1 - \mu(B; t)\big)\,\mathrm{d}t \tag{37}$$

where the first equality is the integral form of the expected value and the last inequality follows by isoperimetry. Hence we can bound the expected distance to the polar cone intersecting the sphere using only knowledge of the volume of spherical caps on $\mathbb{S}^{p-1}$.

To proceed let $v(\varphi)$ denote the volume of a spherical cap subtending a solid angle φ. An explicit formula for $v(\varphi)$ is

$$v(\varphi) = z_p^{-1}\int_0^\varphi \sin^{p-1}(\vartheta)\,\mathrm{d}\vartheta \tag{38}$$

where $z_p = \int_0^\pi \sin^{p-1}(\vartheta)\,d\vartheta$ [45]. Let $\varphi(\beta)$ denote the minimal solid angle of a cap such that β copies of that cap cover $\mathbb{S}^{p-1}$. Since the geodesic distance on the sphere is always greater than or equal to the Euclidean distance, if K is a spherical cap subtending ψ radians, $\mu(K;t) \geq \upsilon(\psi + t)$. Therefore

$$\int_0^\infty \big(1 - \mu(B;t)\big)\,dt \leq \int_0^\infty \big(1 - \upsilon(\varphi(\beta) + t)\big)\,dt. \tag{39}$$

We can proceed to simplify the right-hand side integral:

$$\int_0^\infty \big(1 - \upsilon(\varphi(\beta) + t)\big)\,dt$$

$$= \int_0^{\pi - \varphi(\beta)} \big(1 - \upsilon(\varphi(\beta) + t)\big)\,dt \tag{40}$$

$$= \pi - \varphi(\beta) - \int_0^{\pi - \varphi(\beta)} \upsilon(\varphi(\beta) + t)\,dt \tag{41}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^{\pi - \varphi(\beta)} \int_0^{\varphi(\beta) + t} \sin^{p-1}\vartheta\,d\vartheta\,dt \tag{42}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^\pi \int_{\max(\vartheta - \varphi(\beta), 0)}^{\pi - \varphi(\beta)} \sin^{p-1}\vartheta\,dt\,d\vartheta \tag{43}$$

$$= \pi - \varphi(\beta) - z_p^{-1} \int_0^\pi \big\{\pi - \varphi(\beta) - \max(\vartheta - \varphi(\beta), 0)\big\}\sin^{p-1}\vartheta\,d\vartheta \tag{44}$$

$$= z_p^{-1} \int_0^\pi \max(\vartheta - \varphi(\beta), 0)\sin^{p-1}\vartheta\,d\vartheta \tag{45}$$

$$= z_p^{-1} \int_{\varphi(\beta)}^\pi \big(\vartheta - \varphi(\beta)\big)\sin^{p-1}\vartheta\,d\vartheta \tag{46}$$

(43) follows by switching the order of integration, and the rest of these equalities follow by straightforward integration and some algebra.

Using the inequalities that $z_p \geq \frac{2}{\sqrt{p-1}}$ (see [48]) and $\sin(x) \leq \exp(-(x - \pi/2)^2/2)$ for $x \in [0, \pi]$, we can bound the last integral as

$$z_p^{-1} \int_{\varphi(\beta)}^\pi \big(\vartheta - \varphi(\beta)\big)\sin^{p-1}\vartheta\,d\vartheta$$

$$\leq \frac{\sqrt{p-1}}{2} \int_{\varphi(\beta)}^\pi \big(\vartheta - \varphi(\beta)\big)\exp\left(-\frac{p-1}{2}\left(\vartheta - \frac{\pi}{2}\right)^2\right)d\vartheta. \tag{47}$$

Performing the change of variables $a = \sqrt{p-1}(\vartheta - \frac{\pi}{2})$, we are left with the integral

$$\frac{1}{2} \int_{\sqrt{p-1}(\varphi(\beta) - \pi/2)}^{\sqrt{p-1}\,\pi/2} \left\{\frac{a}{\sqrt{p-1}} + \left(\frac{\pi}{2} - \varphi(\beta)\right)\right\}\exp\left(-\frac{a^2}{2}\right)da \tag{48}$$

$$= -\frac{1}{2\sqrt{p-1}} \exp\left(-\frac{a^2}{2}\right) \Bigg|_{\sqrt{p-1}(\varphi(\beta)-\pi/2)}^{\sqrt{p-1}\pi/2}$$

$$+ \frac{\frac{\pi}{2} - \varphi(\beta)}{2} \int_{\sqrt{p-1}(\varphi(\beta)-\pi/2)}^{\sqrt{p-1}\pi/2} \exp\left(-\frac{a^2}{2}\right) da \qquad (49)$$

$$\leq \frac{1}{2\sqrt{p-1}} \exp\left(-\frac{p-1}{2}(\pi/2 - \varphi(\beta))^2\right) + \sqrt{\frac{\pi}{2}}\left(\frac{\pi}{2} - \varphi(\beta)\right). \qquad (50)$$

In this final bound, we bounded the first term by dropping the upper integrand, and for the second term we used the fact that

$$\int_{-\infty}^{\infty} \exp\left(-x^2/2\right) dx = \sqrt{2\pi}. \qquad (51)$$

We are now left with the task of computing a lower bound for $\varphi(\beta)$. We need to first reparameterize the problem. Let K be a spherical cap. Without loss of generality, we may assume that

$$K = \left\{ x \in \mathbb{S}^{p-1} : x_1 \geq h \right\} \qquad (52)$$

for some $h \in [0, 1]$. Here h is the *height* of the cap over the equator. Via elementary trigonometry, the solid angle that K subtends is given by $\pi/2 - \sin^{-1}(h)$. Hence, if $h(\beta)$ is the largest number such that β caps of height h cover $\mathbb{S}^{p-1}$, then $h(\beta) = \sin(\pi/2 - \varphi(\beta))$.

The quantity $h(\beta)$ may be estimated using the following estimate from [11]. For $h \in [0, 1]$, let $\gamma(p, h)$ denote the volume of a spherical cap of $\mathbb{S}^{p-1}$ of height h.

Lemma B.1 (See [11]) *For* $1 \geq h \geq \frac{2}{\sqrt{p}}$,

$$\frac{1}{10h\sqrt{p}}\left(1 - h^2\right)^{\frac{p-1}{2}} \leq \gamma(p, h) \leq \frac{1}{2h\sqrt{p}}\left(1 - h^2\right)^{\frac{p-1}{2}}. \qquad (53)$$

Note that for $h \geq \frac{2}{\sqrt{p}}$,

$$\frac{1}{2h\sqrt{p}}\left(1 - h^2\right)^{\frac{p-1}{2}} \leq \frac{1}{4}\left(1 - h^2\right)^{\frac{p-1}{2}} \leq \frac{1}{4}\exp\left(-\frac{p-1}{2}h^2\right). \qquad (54)$$

So if

$$h = \sqrt{\frac{2\log(4\beta)}{p-1}} \qquad (55)$$

then $h \leq 1$ because we have assumed $\beta \leq \frac{1}{4}\exp\{\frac{p}{9}\}$ and $p \geq 9$. Moreover, $h \geq \frac{2}{\sqrt{p}}$ and the volume of the cap with height h is less than or equal to $1/\beta$. That is

$$\varphi(\beta) \geq \pi/2 - \sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right). \qquad (56)$$

Combining the estimate (50) with Proposition 3.6, and using our estimate for $\varphi(\beta)$, we get the bound

$$w(\mathcal{C}) \leq \frac{1}{2}\sqrt{\frac{p}{p-1}} \exp\left(-\frac{p-1}{2}\sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right)^2\right)$$

$$+ \sqrt{\frac{\pi p}{2}}\sin^{-1}\left(\sqrt{\frac{2\log(4\beta)}{p-1}}\right). \tag{57}$$

This expression can be simplified by using the following bounds. First, $\sin^{-1}(x) \geq x$ lets us upper-bound the first term by $\sqrt{\frac{p}{p-1}}\frac{1}{8\beta}$. For the second term, using the inequality $\sin^{-1}(x) \leq \frac{\pi}{2}x$ results in the upper bound

$$w(\mathcal{C}) \leq \sqrt{\frac{p}{p-1}}\left(\frac{1}{8\beta} + \frac{\pi^{3/2}}{2}\sqrt{\log(4\beta)}\right). \tag{58}$$

For $p \geq 9$ the upper bound can be expressed simply as $w(\mathcal{C}) \leq 3\sqrt{\log(4\beta)}$. We recall that $\beta = \frac{1}{\Theta}$, which completes the proof of the theorem. $\qquad\square$

Appendix C: Direct Width Calculations

We first give the proof of Proposition 3.10.

Proof Let $\mathbf{x}^\star$ be an s-sparse vector in $\mathbb{R}^p$ with ℓ_1 norm equal to 1, and let $\mathcal{A}$ denote the set of unit-Euclidean-norm one-sparse vectors. Let Δ denote the set of coordinates where $\mathbf{x}^\star$ is nonzero. The normal cone at $\mathbf{x}^\star$ with respect to the ℓ_1 ball is given by

$$N_{\mathcal{A}}(\mathbf{x}^\star) = \text{cone}\left\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}_i = \text{sgn}(\mathbf{x}_i^\star) \text{ for } i \in \Delta, |\mathbf{z}_i| \leq 1 \text{ for } i \in \Delta^c\right\} \tag{59}$$

$$= \left\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}_i = t\,\text{sgn}(\mathbf{x}_i^\star) \text{ for } i \in \Delta, |\mathbf{z}_i| \leq t \text{ for } i \in \Delta^c \text{ for some } t > 0\right\}. \tag{60}$$

Here Δ^c represents the zero entries of $\mathbf{x}^\star$. The minimum squared distance to the normal cone at $\mathbf{x}^\star$ can be formulated as a one-dimensional convex optimization problem for arbitrary $\mathbf{z} \in \mathbb{R}^p$

$$\inf_{\mathbf{u} \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{z} - \mathbf{u}\|_2^2 = \inf_{\substack{t \geq 0 \\ |\mathbf{u}_i| < t, i \in \Delta^c}} \sum_{i \in \Delta}(\mathbf{z}_i - t\,\text{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j \in \Delta^c}(\mathbf{z}_j - \mathbf{u}_j)^2 \tag{61}$$

$$= \inf_{t \geq 0} \sum_{i \in \Delta}(\mathbf{z}_i - t\,\text{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j \in \Delta^c}\text{shrink}(\mathbf{z}_j, t)^2 \tag{62}$$

where

$$\text{shrink}(z, t) = \begin{cases} z + t & z < -t, \\ 0 & -t \leq z \leq t, \\ z - t & z > t \end{cases} \tag{63}$$

is the ℓ_1-shrinkage function. Hence, for any fixed $t \geq 0$ independent of $\mathbf{g}$, we have

$$\mathbb{E}\left[\inf_{\mathbf{u} \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g} - \mathbf{u}\|_2^2\right] \leq \mathbb{E}\left[\sum_{i \in \Delta} (\mathbf{g}_i - t \operatorname{sgn}(\mathbf{x}_i^\star))^2 + \sum_{j \in \Delta^c} \operatorname{shrink}(\mathbf{g}_j, t)^2\right] \quad (64)$$

$$= s(1 + t^2) + \mathbb{E}\left[\sum_{j \in \Delta^c} \operatorname{shrink}(\mathbf{g}_j, t)^2\right]. \quad (65)$$

Now we directly integrate the second term, treating each summand individually. For a zero-mean, unit-variance normal random variable g,

$$\mathbb{E}\left[\operatorname{shrink}(g, t)^2\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-t} (g + t)^2 \exp(-g^2/2) \, dg$$

$$+ \frac{1}{\sqrt{2\pi}} \int_t^\infty (g - t)^2 \exp(-g^2/2) \, dg \quad (66)$$

$$= \frac{2}{\sqrt{2\pi}} \int_t^\infty (g - t)^2 \exp(-g^2/2) \, dg \quad (67)$$

$$= -\frac{2}{\sqrt{2\pi}} t \exp(-t^2/2) + \frac{2(1 + t^2)}{\sqrt{2\pi}} \int_t^\infty \exp(-g^2/2) \, dg \quad (68)$$

$$\leq \frac{2}{\sqrt{2\pi}} \left(-t + \frac{1 + t^2}{t}\right) \exp(-t^2/2) \quad (69)$$

$$= \frac{2}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2). \quad (70)$$

The first simplification follows because the shrink function and Gaussian distributions are symmetric about the origin. The second equality follows by integrating by parts. The inequality follows by a tight bound on the Gaussian Q-function:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-g^2/2) \, dg \leq \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp(-x^2/2) \quad \text{for } x > 0. \quad (71)$$

Using this bound, we get

$$\mathbb{E}\left[\inf_{\mathbf{u} \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g} - \mathbf{u}\|_2^2\right] \leq s(1 + t^2) + (p - s) \frac{2}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2). \quad (72)$$

Setting $t = \sqrt{2 \log(p/s)}$ gives

$$\mathbb{E}\left[\inf_{\mathbf{z} \in N_{\mathcal{A}}(\mathbf{x}^\star)} \|\mathbf{g} - \mathbf{z}\|_2^2\right] \leq s\left(1 + 2 \log\left(\frac{p}{s}\right)\right) + \frac{s(1 - s/p)}{\pi \sqrt{\log(p/s)}} \leq 2s \log(p/s) + \frac{5}{4} s. \quad (73)$$

The last inequality follows because

$$\frac{(1 - s/p)}{\pi\sqrt{\log(p/s)}} \leq 0.204 < 1/4 \tag{74}$$

whenever $0 \leq s \leq p$.   □

Next we give the proof of Proposition 3.11.

Proof Let $\mathbf{x}^\star$ be an $m_1 \times m_2$ matrix of rank r with singular value decomposition $U\Sigma V^*$, and let $\mathcal{A}$ denote the set of rank-one unit-Euclidean-norm matrices of size $m_1 \times m_2$. Without loss of generality, impose the conventions $m_1 \leq m_2$, Σ is $r \times r$, U is $m_1 \times r$, and V is $m_2 \times r$, and assume the nuclear norm of $\mathbf{x}^\star$ is equal to 1.

Let $\mathbf{u}_k$ (respectively $\mathbf{v}_k$) denote the k'th column of U (respectively V). It is convenient to introduce the orthogonal decomposition $\mathbb{R}^{m_1 \times m_2} = \Delta \oplus \Delta^\perp$ where Δ is the linear space spanned by elements of the form $\mathbf{u}_k \mathbf{z}^T$ and $\mathbf{y}\mathbf{v}_k^T$, $1 \leq k \leq r$, where $\mathbf{z}$ and $\mathbf{y}$ are arbitrary, and $\Delta^\perp$ is the orthogonal complement of Δ. The space $\Delta^\perp$ is the subspace of matrices spanned by the family $(\mathbf{y}\mathbf{z}^T)$, where $\mathbf{y}$ (respectively $\mathbf{z}$) is any vector orthogonal to all the columns of U (respectively V). The normal cone of the nuclear norm ball at $\mathbf{x}^\star$ is given by the cone generated by the subdifferential at $\mathbf{x}^\star$:

$$N_{\mathcal{A}}(\mathbf{x}^\star) = \text{cone}\{UV^T + W \in \mathbb{R}^{m_1 \times m_2} : W^T U = 0, \ WV = 0, \ \|W\|_{\mathcal{A}}^* \leq 1\} \tag{75}$$

$$= \{tUV^* + W \in \mathbb{R}^{m_1 \times m_2} : W^T U = 0, \ WV = 0, \ \|W\|_{\mathcal{A}}^* \leq t, \ t \geq 0\}. \tag{76}$$

Note that here $\|Z\|_{\mathcal{A}}^*$ is the operator norm, equal to the maximum singular value of Z [63].

Let G be a Gaussian random matrix with i.i.d. entries, each with mean zero and unit variance. Then the matrix

$$Z(G) = \|\mathcal{P}_{\Delta^\perp}(G)\|UV^* + \mathcal{P}_{\Delta^\perp}(G) \tag{77}$$

is in the normal cone at $\mathbf{x}^\star$. We can then compute

$$\mathbb{E}\big[\|G - Z(G)\|_F^2\big] = \mathbb{E}\big[\|\mathcal{P}_{\Delta}(G) - \mathcal{P}_{\Delta}(Z(G))\|_F^2\big] \tag{78}$$

$$= \mathbb{E}\big[\|\mathcal{P}_{\Delta}(G)\|_F^2\big] + \mathbb{E}\big[\|\mathcal{P}_{\Delta}(Z(G))\|_F^2\big] \tag{79}$$

$$= r(m_1 + m_2 - r) + r\,\mathbb{E}\big[\|\mathcal{P}_{\Delta^\perp}(G)\|^2\big]. \tag{80}$$

Here (79) follows because $\mathcal{P}_{\Delta}(G)$ and $\mathcal{P}_{\Delta^\perp}(G)$ are independent. The final line follows because $\dim(T) = r(m_1 + m_2 - r)$ and the Frobenius (i.e., Euclidean) norm of UV^* is $\|UV^*\|_F = \sqrt{r}$. Due to the isotropy of Gaussian random matrices, $\mathcal{P}_{\Delta^\perp}(G)$ is identically distributed as an $(m_1 - r) \times (m_2 - r)$ matrix with i.i.d. Gaussian entries each with mean zero and variance one. We thus know that

$$\mathbb{P}\big[\|\mathcal{P}_{\Delta^\perp}(G)\| \geq \sqrt{m_1 - r} + \sqrt{m_2 - r} + s\big] \leq \exp(-s^2/2) \tag{81}$$

(see, for example, [22]). To bound the latter expectation, we again use the integral form of the expected value. Letting $\mu_{T^\perp}$ denote the quantity $\sqrt{m_1 - r} + \sqrt{m_2 - r}$,

we have

$$\mathbb{E}\big[\big\|\mathcal{P}_{\Delta^\perp}(G)\big\|^2\big] = \int_0^\infty \mathbb{P}\big[\big\|\mathcal{P}_{\Delta^\perp}(G)\big\|^2 > h\big]\,dh \tag{82}$$

$$\leq \mu_{T^\perp}^2 + \int_{\mu_{T^\perp}^2}^\infty \mathbb{P}\big[\big\|\mathcal{P}_{\Delta^\perp}(G)\big\|^2 > h\big]\,dh \tag{83}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \mathbb{P}\big[\big\|\mathcal{P}_{\Delta^\perp}(G)\big\|^2 > \mu_{T^\perp}^2 + t\big]\,dt \tag{84}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \mathbb{P}\big[\big\|\mathcal{P}_{\Delta^\perp}(G)\big\| > \mu_{T^\perp} + \sqrt{t}\big]\,dt \tag{85}$$

$$\leq \mu_{T^\perp}^2 + \int_0^\infty \exp(-t/2)\,dt \tag{86}$$

$$= \mu_{T^\perp}^2 + 2. \tag{87}$$

Using this bound in (80), we obtain

$$\mathbb{E}\Big[\inf_{Z\in N_{\mathcal{A}}(\mathbf{x}^\star)}\|G-Z\|_F^2\Big] \leq r(m_1+m_2-r)+r(\sqrt{m_1-r}+\sqrt{m_2-r})^2+2r \tag{88}$$

$$\leq r(m_1+m_2-r)+2r(m_1+m_2-2r)+2r \tag{89}$$

$$\leq 3r(m_1+m_2-r) \tag{90}$$

where the second inequality follows from the fact that $(a+b)^2 \leq 2a^2 + 2b^2$. We conclude that $3r(m_1+m_2-r)$ random measurements are sufficient to recover a rank r, $m_1 \times m_2$ matrix using the nuclear norm heuristic. $\qquad\square$

References

1. S. Aja-Fernandez, R. Garcia, D. Tao, X. Li, *Tensors in Image Processing and Computer Vision*. Advances in Pattern Recognition (Springer, Berlin, 2009).
2. N. Alon, A. Naor, Approximating the cut-norm via Grothendieck's inequality, *SIAM J. Comput.* **35**, 787–803 (2006).
3. A. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inf. Theory* **39**, 930–945 (1993).
4. A. Barvinok, *A Course in Convexity* (American Mathematical Society, Providence, 2002).
5. C. Beckmann, S. Smith, Tensorial extensions of independent component analysis for multisubject FMRI analysis, *NeuroImage* **25**, 294–311 (2005).
6. D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods* (Athena Scientific, Nashua, 2007).
7. D. Bertsekas, A. Nedic, A. Ozdaglar, *Convex Analysis and Optimization* (Athena Scientific, Nashua, 2003).
8. P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, *Ann. Stat.* **37**, 1705–1732 (2009).
9. J. Bochnak, M. Coste, M. Roy, *Real Algebraic Geometry* (Springer, Berlin, 1988).
10. F.F. Bonsall, A general atomic decomposition theorem and Banach's closed range theorem, *Q. J. Math.* **42**, 9–14 (1991).

11. A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovasz, M. Simonovits, Approximation of diameters: randomization doesn't help, in *Proceedings of the 39th Annual Symposium on Foundations of Computer Science* (1998), pp. 244–251.

12. J. Cai, E. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim.* **20**, 1956–1982 (2008).

13. J. Cai, S. Osher, Z. Shen, Linearized Bregman iterations for compressed sensing, *Math. Comput.* **78**, 1515–1536 (2009).

14. E. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? *J. ACM* **58**, 1–37 (2011).

15. E. Candès, Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements, *IEEE Trans. Inf. Theory* **57**, 2342–2359 (2011).

16. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).

17. E.J. Candès, B. Recht, Exact matrix completion via convex optimization, *Found. Comput. Math.* **9**, 717–772 (2009).

18. E. Candès, T. Tao, Decoding by linear programming, *IEEE Trans. Inf. Theory* **51**, 4203–4215 (2005).

19. V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, A.S. Willsky, Rank-sparsity incoherence for matrix decomposition, *SIAM J. Optim.* **21**, 572–596 (2011).

20. P. Combettes, V. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.* **4**, 1168–1200 (2005).

21. I. Daubechies, M. Defriese, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Commun. Pure Appl. Math.* **LVII**, 1413–1457 (2004).

22. K.R. Davidson, S.J. Szarek, Local operator theory, random matrices and Banach spaces, in *Handbook of the Geometry of Banach Spaces*, vol. I (2001), pp. 317–366.

23. V. de Silva, L. Lim, Tensor rank and the ill-posedness of the best low-rank approximation problem, *SIAM J. Matrix Anal. Appl.* **30**, 1084–1127 (2008).

24. R. DeVore, V. Temlyakov, Some remarks on greedy algorithms, *Adv. Comput. Math.* **5**, 173–187 (1996).

25. M. Deza, M. Laurent, *Geometry of Cuts and Metrics* (Springer, Berlin, 1997).

26. D.L. Donoho, High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension, *Discrete Comput. Geom.* (online) (2005).

27. D.L. Donoho, For most underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution, *Commun. Pure Appl. Math.* **59**, 797–829 (2006).

28. D.L. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).

29. D. Donoho, J. Tanner, Sparse nonnegative solution of underdetermined linear equations by linear programming, *Proc. Natl. Acad. Sci. USA* **102**, 9446–9451 (2005).

30. D. Donoho, J. Tanner, Counting faces of randomly-projected polytopes when the projection radically lowers dimension, *J. Am. Math. Soc.* **22**, 1–53 (2009).

31. D. Donoho, J. Tanner, Counting the faces of randomly-projected hypercubes and orthants with applications, *Discrete Comput. Geom.* **43**, 522–541 (2010).

32. R.M. Dudley, The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, *J. Funct. Anal.* **1**, 290–330 (1967).

33. M. Dyer, A. Frieze, R. Kannan, A random polynomial-time algorithm for approximating the volume of convex bodies, *J. ACM* **38**, 1–17 (1991).

34. M. Fazel, Matrix rank minimization with applications, Ph.D. thesis, Department of Electrical Engineering, Stanford University (2002).

35. M. Figueiredo, R. Nowak, An EM algorithm for wavelet-based image restoration, *IEEE Trans. Image Process.* **12**, 906–916 (2003).

36. M. Fukushima, H. Mine, A generalized proximal point algorithm for certain non-convex minimization problems, *Int. J. Inf. Syst. Sci.* **12**, 989–1000 (1981).

37. M. Goemans, D. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, *J. ACM* **42**, 1115–1145 (1995).

38. Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, in *Geometric Aspects of Functional Analysis, Israel Seminar 1986–1987*. Lecture Notes in Mathematics, vol. 1317 (1988), pp. 84–106.

39. J. Gouveia, P. Parrilo, R. Thomas, Theta bodies for polynomial ideals, *SIAM J. Optim.* **20**, 2097–2118 (2010).

40. T. Hale, W. Yin, Y. Zhang, A fixed-point continuation method for ℓ_1-regularized minimization: methodology and convergence, *SIAM J. Optim.* **19**, 1107–1130 (2008).

41. J. Harris, *Algebraic Geometry: A First Course* (Springer, Berlin).
42. J. Haupt, W.U. Bajwa, G. Raz, R. Nowak, Toeplitz compressed sensing matrices with applications to sparse channel estimation, *IEEE Trans. Inform. Theory* **56**(11), 5862–5875 (2010).
43. S. Jagabathula, D. Shah, Inferring rankings using constrained sensing, *IEEE Trans. Inf. Theory* **57**, 7288–7306 (2011).
44. L. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Stat.* **20**, 608–613 (1992).
45. D. Klain, G. Rota, *Introduction to Geometric Probability* (Cambridge University Press, Cambridge, 1997).
46. T. Kolda, Orthogonal tensor decompositions, *SIAM J. Matrix Anal. Appl.* **23**, 243–255 (2001).
47. T. Kolda, B. Bader, Tensor decompositions and applications, *SIAM Rev.* **51**, 455–500 (2009).
48. M. Ledoux, *The Concentration of Measure Phenomenon* (American Mathematical Society, Providence, 2000).
49. M. Ledoux, M. Talagrand, *Probability in Banach Spaces* (Springer, Berlin, 1991).
50. J. Löfberg, YALMIP: A toolbox for modeling and optimization in MATLAB, in *Proceedings of the CACSD Conference*, Taiwan (2004). Available from http://control.ee.ethz.ch/~joloef/yalmip.php.
51. S. Ma, D. Goldfarb, L. Chen, Fixed point and Bregman iterative methods for matrix rank minimization, *Math. Program.* **128**, 321–353 (2011).
52. O. Mangasarian, B. Recht, Probability of unique integer solution to a system of linear equations, *Eur. J. Oper. Res.* **214**, 27–30 (2011).
53. J. Matoušek, *Lectures on Discrete Geometry* (Springer, Berlin, 2002).
54. S. Negahban, P. Ravikumar, M. Wainwright, B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Preprint (2010).
55. Y. Nesterov, Quality of semidefinite relaxation for nonconvex quadratic optimization. Technical report (1997).
56. Y. Nesterov, *Introductory Lectures on Convex Optimization* (Kluwer Academic, Amsterdam, 2004).
57. Y. Nesterov, Gradient methods for minimizing composite functions, CORE discussion paper 76 (2007).
58. P.A. Parrilo, Semidefinite programming relaxations for semialgebraic problems, *Math. Program.* **96**, 293–320 (2003).
59. G. Pisier, *Remarques sur un résultat non publié de B. Maurey*. Séminaire d'analyse fonctionnelle (Ecole Polytechnique Centre de Mathematiques, Palaiseau, 1981).
60. G. Pisier, Probabilistic methods in the geometry of Banach spaces, in *Probability and Analysis*, pp. 167–241 (1986).
61. E. Polak, *Optimization: Algorithms and Consistent Approximations* (Springer, Berlin, 1997).
62. H. Rauhut, Circulant and Toeplitz matrices in compressed sensing, in *Proceedings of SPARS'09*, (2009).
63. B. Recht, M. Fazel, P.A. Parrilo, Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization, *SIAM Rev.* **52**, 471–501 (2010).
64. B. Recht, W. Xu, B. Hassibi, Null space conditions and thresholds for rank minimization, *Math. Program., Ser. B* **127**, 175–211 (2011).
65. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970).
66. M. Rudelson, R. Vershynin, Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements, in *CISS 2006 (40th Annual Conference on Information Sciences and Systems)* (2006).
67. R. Sanyal, F. Sottile, B. Sturmfels, Orbitopes, Preprint, arXiv:0911.5436 (2009).
68. N. Srebro, A. Shraibman, Rank, trace-norm and max-norm in *18th Annual Conference on Learning Theory (COLT)* (2005).
69. M. Stojnic, Various thresholds for ℓ_1-optimization in compressed sensing, Preprint, arXiv:0907.3666 (2009).
70. K. Toh, M. Todd, R. Tutuncu, SDPT3—a MATLAB software package for semidefinite-quadratic-linear programming. Available from http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.
71. K. Toh, S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, *Pac. J. Optim.* **6**, 615–640 (2010).
72. S. van de Geer, P. Bühlmann, On the conditions used to prove oracle results for the Lasso, *Electron. J. Stat.* **3**, 1360–1392 (2009).
73. S. Wright, R. Nowak, M. Figueiredo, Sparse reconstruction by separable approximation, *IEEE Trans. Signal Process.* **57**, 2479–2493 (2009).

74. W. Xu, B. Hassibi, Compressive sensing over the Grassmann manifold: a unified geometric framework, *IEEE Trans. Inform. Theory* **57**(10), 6894–6919 (2011).
75. W. Yin, S. Osher, J. Darbon, D. Goldfarb, Bregman iterative algorithms for compressed sensing and related problems, *SIAM J. Imaging Sci.* **1**, 143–168 (2008).
76. G. Ziegler, *Lectures on Polytopes* (Springer, Berlin, 1995).