# Sharp recovery bounds for convex demixing, with applications

Michael B. McCoy[*][†]        Joel A. Tropp[†]

**Abstract**

Demixing refers to the challenge of identifying two structured signals given only the sum of the two signals and prior information about their structures. Examples include the problem of separating a signal that is sparse with respect to one basis from a signal that is sparse with respect to a second basis, and the problem of decomposing an observed matrix into a low-rank matrix plus a sparse matrix. This paper describes and analyzes a framework, based on convex optimization, for solving these demixing problems, and many others. This work introduces a randomized signal model which ensures that the two structures are incoherent, i.e., generically oriented. For an observation from this model, this approach identifies a summary statistic that reflects the complexity of a particular signal. The difficulty of separating two structured, incoherent signals depends only on the total complexity of the two structures. Some applications include (i) demixing two signals that are sparse in mutually incoherent bases; (ii) decoding spread-spectrum transmissions in the presence of impulsive errors; and (iii) removing sparse corruptions from a low-rank matrix. In each case, the theoretical analysis of the convex demixing method closely matches its empirical behavior.

## 1  Introduction

In modern data-intensive science, it is common to observe a superposition of multiple information-bearing signals. *Demixing* refers to the challenge of separating out the constituent signals from the observation. A fundamental computational question is to understand when a tractable algorithm can successfully complete the demixing. Problems of this sort arise in fields as diverse as acoustics [1], astronomy [70], communications [7], [8], geophysics [77], image processing [71], [35], machine learning [14], and statistics [12]. Some well-known examples of convex methods for demixing include morphological component analysis [70], robust principal component analysis [17], [12], and inpainting [35].

This work presents a general framework for demixing based on convex optimization. We study the geometry of the optimization problem, and we develop conditions that describe precisely when our method succeeds. Let us illustrate the major aspects of our approach through a concrete example.

---

## 1.1 A first application: Morphological component analysis

Starck et al. use demixing to model the problem of distinguishing stars from galaxies in an astronomical image [70]. This task requires hypotheses on the two types of objects. First, we must assume that stars and galaxies exhibit different kinds of structure: stars appear as localized bright points, while galaxies are wispy or filamented. Second, we must insist that the image is not so full of stars, nor of galaxies, that they obscure one another. These two properties are modeled by the notions of *incoherence* and *sparsity*. With these hypotheses, we can solve the demixing problem using a method known as morphological component analysis (MCA) [70], [71], [35], [7], [84].

### 1.1.1 The MCA signal model

We model the observation $z_0 \in \mathbb{R}^d$ as the superposition of two structured signals:

$$z_0 = Ax_0 + By_0 \in \mathbb{R}^d.$$

The matrices $A$ and $B$ are known, while the vectors $x_0$ and $y_0$ are unknown. Each column of $A$ contains an elementary structure that might appear in the first signal; the columns of $B$ reflect the structures in the second signal. The vector $x_0$ selects the columns of $A$ that appear in the first signal, e.g., stars in different locations, while $y_0$ selects the columns of $B$ that generate the second signal, e.g., galaxies in different locations. *Incoherence* demands that the columns of $A$ and $B$ are weakly correlated, and *sparsity* requires that $x_0$ and $y_0$ have few nonzero elements.

For simplicity, we assume that $A$ and $B$ are orthonormal bases. By changing coordinates, we may take $A = I$, the identity matrix. The observation then has the form

$$z_0 = x_0 + Qy_0$$

for a known orthogonal matrix $Q$. The specialization to orthonormal bases is standard [26], [35], [71], [46].

Instead of restricting our attention to specific choices of $Q$ that are incoherent with the identity matrix, we consider an idealized model for incoherence where $Q$ is a uniformly random orthogonal matrix. This formulation ensures that the structures in the two signals are oriented generically with respect to each other. Other authors have also used this approach to study incoherence [26], [35].

We quantify the sparsity of the two constituent signals by fixing parameters $\tau_x$ and $\tau_y$ in the interval $[0, 1]$ such that the unknown signals $x_0$ and $y_0$ satisfy

$$\mathrm{nnz}(x_0) = \lceil \tau_x d \rceil \quad \text{and} \quad \mathrm{nnz}(y_0) = \lceil \tau_y d \rceil,$$

where $\mathrm{nnz}(x)$ denotes the *n*umber of *n*on*z*ero elements of $x$.* In other words, $\tau_x$ and $\tau_y$ measure the proportion of nonzero entries in $x_0$ and $y_0$. These sparsity parameters emerge as the major factor that determines how hard it is to extract $x_0$ and $y_0$ from the observation $z_0$.

### 1.1.2 The constrained MCA demixing procedure

The goal of morphological component analysis is to identify the pair $(x_0, y_0)$ of sparse vectors given the observation $z_0$ and the matrix $Q$. A natural technique for finding a sparse vector that satisfies certain conditions is to minimize the $\ell_1$ norm subject to these constraints [18], where the $\ell_1$ norm is defined as $\|x\|_{\ell_1} := \sum_{i=1}^d |x_i|$.

Assume that we have access to side information $\alpha = \|y_0\|_{\ell_1}$. Then the intuition above leads us to frame the following convex optimization problem for demixing:

$$
\begin{aligned}
\text{minimize} \quad & \|x\|_{\ell_1} \\
\text{subject to} \quad & \|y\|_{\ell_1} \leq \alpha \text{ and } x + Qy = z_0,
\end{aligned}
\tag{1.1}
$$

---

*We prefer the notation $\mathrm{nnz}(\cdot)$ over $\|\cdot\|_{\ell_0}$ because the number of nonzero elements in a vector is not a norm.
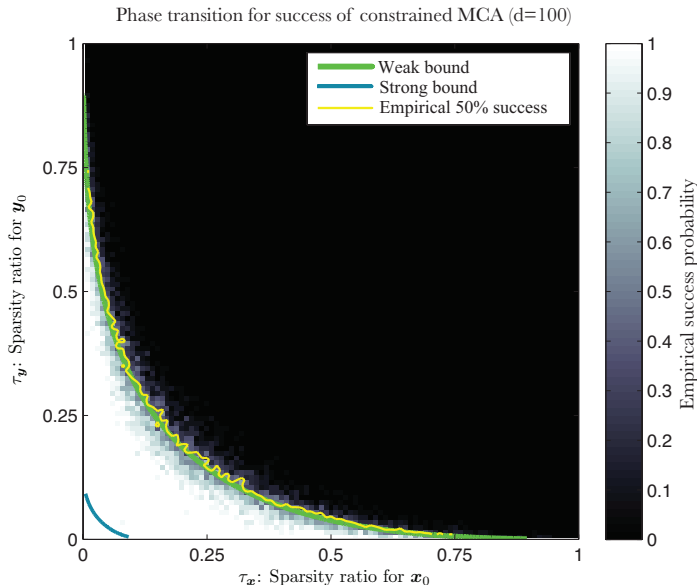
**Figure 1** *Performance of constrained MCA.* The variables $\tau_x$ and $\tau_y$ on the axes represent the fraction of components in $x_0$ and $y_0$ that are nonzero. The background shading indicates the empirical probability that the constrained MCA problem (1.1) identifies the pair $(x_0, y_0)$ from the observation $z_0 = x_0 + Q y_0$, where $Q$ is a random orthogonal basis. The yellow curve marks the empirical 50% success threshold. The green curve locates the theoretical phase transition for demixing a single pair $(x_0, y_0)$. For sparsity levels below the blue curve, constrained MCA (1.1) provably recovers all $(\tau_x, \tau_y)$-sparse pairs with high probability in the dimension. Further details are available in Section 1.1.3 and Section 6.1.

where the decision variables are $x, y \in \mathbb{R}^d$. We call this optimization problem *constrained MCA*, and we say that it *succeeds* if $(x_0, y_0)$ is the unique optimal point of (1.1). Since (1.1) can be written as a linear program, constrained MCA offers a tractable procedure for attempting to identify the underlying components $(x_0, y_0)$, provided the observation $z_0$, the orthogonal matrix $Q$, and the side information $\alpha = \|y_0\|_{\ell_1}$.

Constrained MCA is closely related to the standard MCA procedure, which is a Lagrangian formulation of (1.1) that does not require the side information $\alpha$ [**71**, Eq. (4)]. The constrained problem (1.1) is more powerful than the standard MCA procedure, so it provides hard limits on the effectiveness of the usual approach. In most cases, the two methods are equivalent, provided that we can choose the Lagrange multiplier correctly—a nontrivial task in itself. See Section 1.2.4 for more details.

### 1.1.3 Numerical and theoretical results for constrained MCA

Figure 1 displays the result of a numerical experiment on constrained MCA. We fix the dimension $d = 100$. For sparsity levels $(\tau_x, \tau_y)$ varying over the unit square $[0, 1]^2$, we form vectors $x_0$ and $y_0$ with sparsity levels $\mathrm{nnz}(x_0) = \lceil \tau_x d \rceil$ and $\mathrm{nnz}(y_0) = \lceil \tau_y d \rceil$. (The manner in which we choose the nonzero entries is irrelevant.) We draw a random orthogonal matrix $Q$ and construct the observation $z_0 = x_0 + Q y_0$. Then we solve the constrained MCA problem (1.1) to identify the pair $(x_0, y_0)$. The background of the figure shows the empirical probability of success over the randomness in $Q$; dark areas denote low probability of success, while light areas denote high success rates. The yellow curve marks the 50% success threshold.

This work establishes two theoretical results for constrained MCA. The first result provides a phase transition curve, parameterized by the sparsity $(\tau_x, \tau_y)$, for the probability that constrained MCA will demix a single pair $(x_0, y_0)$ from the associated observation $z_0$. This *weak bound* is marked by the green line in
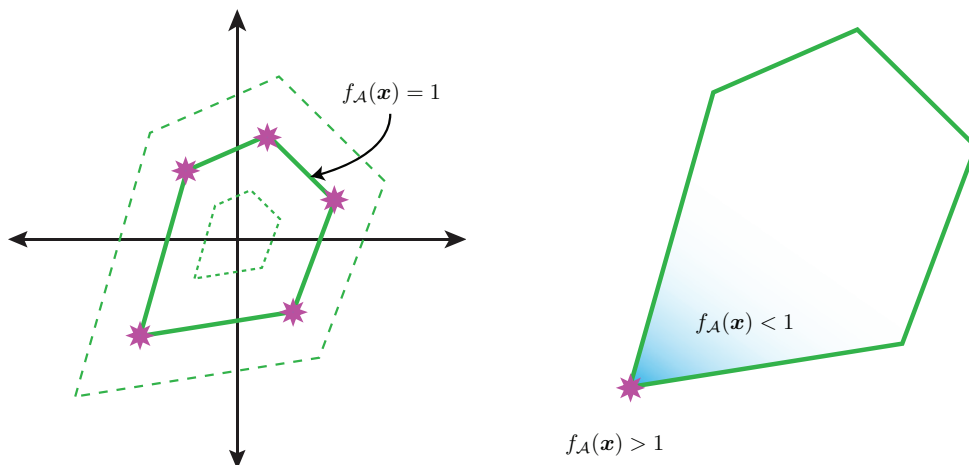
**Figure 2** *Atomic gauge.* [Left] Let $\mathscr{A}$ be an atomic set consisting of five atoms (stars). The "unit ball" of the atomic gauge $f_{\mathscr{A}}$ is the closed convex hull of $\mathscr{A}$ (heavy line). Other level sets (dashed lines) of $f_{\mathscr{A}}$ are dilations of the unit ball. [Right] At an atom (star), the unit ball of $f_{\mathscr{A}}$ tends to have sharp corners. Most perturbations away from this atom increase the value of $f_{\mathscr{A}}$, so the atomic gauge is an effective measure of the complexity of an atomic signal.

Figure 1. Observe that the green line coincides almost perfectly with the empirical phase transition.

Second, we establish a *strong bound.* For a fixed instantiation of the random orthogonal basis $\boldsymbol{Q}$, with high probability, constrained MCA (1.1) can identify *every* sufficiently sparse pair $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ from the associated observation $\boldsymbol{z}_0$. The blue curve in the bottom left corner of Figure 1 is a lower estimate for the sparsity pairs $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$ where this uniform guarantee holds. Section 6.1 provides the details regarding the computation of the weak and strong bounds as well as a fully detailed description of our numerical experiment.

## 1.2 A recipe for demixing

This work is not primarily about MCA. We are interested in developing methods that apply to a whole spectrum of demixing problems. The following two sections describe how to construct a convex program that can separate two structured signals.

### 1.2.1 Structured signals and atomic gauges

The $\ell_1$ norm is a convex complexity measure that tends to be small near sparse vectors, so we can minimize the $\ell_1$ norm to promote sparsity. We now describe a method for building complexity measures that are appropriate for other types of structure. This construction was originally introduced in the nonlinear approximation literature [**21**], [**78**]. The recent paper [**15**] explains how to apply these ideas to solve signal processing problems.

In practice, we often encounter signals that are formed as a positive linear combination of a few elementary structures, called *atoms*, drawn from a fixed collection. For example, a sparse vector in $\mathbb{R}^d$ is a conic combination of a small number elements from the set $\{\pm \boldsymbol{e}_i : i = 1, \ldots, d\}$ of signed standard basis vectors. We want to construct a function that reflects the complexity of an atomic signal.

We define the *atomic gauge* of a vector $\boldsymbol{x} \in \mathbb{R}^d$ with respect to a set $\mathscr{A} \subset \mathbb{R}^d$ of atoms by

$$f_{\mathscr{A}}(\boldsymbol{x}) := \inf\left\{\lambda \geq 0 : \boldsymbol{x} \in \lambda \cdot \mathrm{conv}(\mathscr{A})\right\},$$

with the convention that $f_{\mathscr{A}}(\boldsymbol{x}) = +\infty$ if the set is empty. Then $f_{\mathscr{A}}$ is a homogeneous convex function. The "unit ball" of $f_{\mathscr{A}}$ is $\overline{\mathrm{conv}}(\mathscr{A})$, and the level sets of $f_{\mathscr{A}}$ are dilations of this unit ball. The atomic gauge $f_{\mathscr{A}}$ is a norm if and only if $\overline{\mathrm{conv}}(\mathscr{A})$ is a bounded, symmetric set that contains zero in its interior.

The convex hull of an atomic set $\mathscr{A}$ tends to have sharp corners at atoms; see Figure 2. At these sharp points, most perturbations of the objective increase the value of the gauge, so the atomic gauge tends to take small values at atoms. Similar behavior occurs at a signal comprised of a relatively small number of atoms. This observation is a key reason that atomic gauges make good complexity measures for atomic signals [15]. Some common atomic gauges include

- *The $\ell_1$ norm.* The $\ell_1$ norm on $\mathbb{R}^d$ is the atomic gauge generated by the set $\mathscr{A} = \{\pm \boldsymbol{e}_i : i = 1, \ldots, d\}$ of signed standard basis vectors. This norm is widely used to promote sparsity [18], [13], [24]. The $\ell_1$ norm may also be defined for matrices via the formula $\|\boldsymbol{X}\|_{\ell_1} = \sum_{i,j} |X_{ij}|$. In this context, the $\ell_1$ norm reflects the sparsity of matrices [16], [14], [17].

- *The $\ell_\infty$ norm.* The $\ell_\infty$ norm on $\mathbb{R}^d$, given by $\|\boldsymbol{x}\|_{\ell_\infty} = \max_{i=1,\ldots,d} |x_i|$, is the atomic gauge generated by the set $\mathscr{A} = \{\pm 1\}^d \subset \mathbb{R}^d$ of all $2^d$ sign vectors. We use this norm to demix binary codewords. See also [31], [15], [51]. For matrices, the $\ell_\infty$ norm returns $\|\boldsymbol{X}\|_{\ell_\infty} = \max_{i,j} |X_{ij}|$; this function is the atomic gauge generated by the set of sign matrices.

- *The Schatten 1-norm.* The Schatten 1-norm on $\mathbb{R}^{m \times n}$ is the sum of the singular values of a matrix. It is the atomic gauge generated by the set of rank-one matrices in $\mathbb{R}^{m \times n}$ with unit Frobenius norm. Minimizing the Schatten 1-norm promotes low rank [36], [64].

- *The operator norm.* The operator norm returns the maximum singular value of a matrix. On the space $\mathbb{R}^{n \times n}$ of square matrices, the operator norm is the atomic gauge generated by the set $\mathsf{O}_n$ of orthogonal matrices. This norm can be used to search for orthogonal matrices [15, Prop. 3.13]

Our applications focus on these four instances, but a dizzying variety other structure-promoting atomic gauges are available. For example, there are atomic gauges for vectors that are sparse in a dictionary (also known as analysis-sparsity) [34], [11], block- and group-sparse vectors [20], [72], and low-rank tensors and probability measures [15, Sec. 2].

### 1.2.2 A generic model for incoherence

Demixing is hopeless when the structures in the constituent signals are too strongly aligned. As an extreme example, suppose we observe $\boldsymbol{z}_0 = \boldsymbol{x}_0 + \boldsymbol{y}_0$, where both $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ are sparse. There is clearly no principled way to assign the nonzero elements of $\boldsymbol{z}_0$ correctly to $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$. In contrast, if we observe $\boldsymbol{z}_0 = \boldsymbol{x}_0 + \mathbf{H}\boldsymbol{y}_0$, where $\mathbf{H}$ is a normalized Walsh–Hadamard transform and both $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ are sparse, then the pair $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ is typically identifiable [79]. The latter situation is more favorable than the former because the Walsh–Hadamard matrix and the identity matrix are incoherent; that is, their columns are weakly correlated.

In order to avoid restricting our attention to special cases, such as the Walsh–Hadamard matrix, we model incoherence by assuming that the basis $\boldsymbol{Q}$ is drawn randomly from the invariant Haar measure on the set of all orthogonal matrices $\mathsf{O}_d$. We call this the *random basis model.* This idealized approach to incoherence guarantees that the structures in the two constituent signals are generically oriented. This model has precedents in the literature on sparse approximation [26], [35], and it is analogous to the assumption of a measurement operator with a uniformly random nullspace that appears in the context of compressed sensing [25], [73].

We expect that the random basis model also sheds light on other highly incoherent problems, such as the case where $\boldsymbol{Q} = \mathbf{H}$ is the Walsh–Hadamard transform or $\boldsymbol{Q} = \mathbf{D}$ is the discrete cosine transform (DCT). Some limited numerical simulations suggest that both the Walsh–Hadamard and the DCT matrices behave qualitatively similar to the random matrix $\boldsymbol{Q}$ in the examples considered in this work. This observation is in line with the universality of phase transitions that appear in $\ell_1$ minimization for many classes of measurement matrices [30]. However, more coherent situations may exhibit different behavior, and thus they fall outside the purview of this work.

### 1.2.3 Formulating a convex demixing method

We are ready to introduce a computational framework for demixing structured signals. This approach unifies several related procedures that appear in the literature. See, for example, [47], [17], [83].

Assume we observe the superposition of two structured signals:

$$z_0 = x_0 + Qy_0,$$

where $Q$ is a known orthogonal matrix and the pair $(x_0, y_0)$ is unknown. We include the matrix $Q$ in the formalism because it allows us to model incoherence using the random basis model described in Section 1.2.2 above. Our goal is to demix the pair $(x_0, y_0)$ from the observation $z_0$.

Let $f$ and $g$ be convex complexity measures—such as atomic gauges—associated with the structures we expect to find in $x_0$ and $y_0$. Suppose we have access to the additional side information $\alpha = g(y_0)$. We combine these ingredients to reach the following convex demixing method:

$$
\begin{aligned}
&\text{minimize} \quad f(x) \\
&\text{subject to} \quad g(y) \le \alpha \text{ and } x + Qy = z_0,
\end{aligned}
\tag{1.2}
$$

where the decision variables are $x, y \in \mathbb{R}^d$. The display (1.2) describes a convex program because $f$ and $g$ are convex functions. We say that the convex demixing method (1.2) *succeeds at demixing* $(x_0, y_0)$, or simply *succeeds*, if $(x_0, y_0)$ is the unique optimal point of (1.2); otherwise, it *fails*. In this work, we develop conditions that describe when the convex demixing method (1.2) succeeds and when it fails.

### 1.2.4 The Lagrangian counterpart

In practice, the value $\alpha = g(y_0)$ may not be known. In this case, we may replace the convex demixing method (1.2) with its Lagrangian relative

$$
\begin{aligned}
&\text{minimize} \quad f(x) + \lambda \cdot g(y) \\
&\text{subject to} \quad x + Qy = z_0,
\end{aligned}
\tag{1.3}
$$

where $\lambda > 0$ is a regularization parameter that must be specified. The constrained problem (1.2) is slightly more powerful than (1.3), so its performance dominates the Lagrangian formulation (1.3). It is well known that (1.2) and (1.3) are essentially equivalent when the regularization parameter $\lambda$ is chosen correctly and a mild regularity condition holds; see Appendix A for details. Thus, we can interpret our results as delineating the *best possible performance* of the Lagrange problem (1.3).

This type of best-case analysis has precedents in the literature on sparse approximation, e.g., [82], yet the identification of optimal Lagrange parameters for demixing remains a significant open problem. Several works prove that specific demixing procedures succeed for specific choices of Lagrange parameters under incoherence assumptions [26], [84], [17], [12], [88], but these conservative guarantees fail to identify phase transitions. Since this document was submitted, some additional theoretical guidance choosing Lagrange parameters has appeared [76], [74], [37], [61]. Nevertheless, a comprehensive theory describing the optimal choices of Lagrange parameters for (1.3) does not currently exist.

## 1.3 One hammer, many nails

The convex demixing method (1.2) includes many interesting special cases. Our analysis provides detailed information about when (1.2) is able to separate two structured, incoherent signals. We now describe some applications of this machinery.
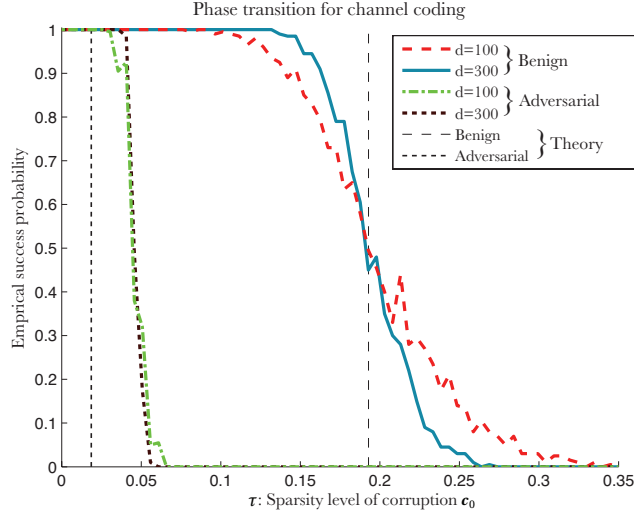
**Figure 3** *Performance of the robust communication protocol.* The heavy curves indicate the empirical probability that problem (1.4) decodes a $d$-bit message $\boldsymbol{m}_0 \in \{\pm 1\}^d$ from the observation $\boldsymbol{z}_0 = \boldsymbol{Q}\boldsymbol{m}_0 + \boldsymbol{c}_0$, where $\boldsymbol{Q}$ is a random orthogonal matrix. The variable $\tau$ on the horizontal axis measures the proportion of nonzero entries in the corruption $\boldsymbol{c}_0$. A benign corruption is independent of $\boldsymbol{Q}$, while an adversarial erasure zeros out the $\lceil \tau d \rceil$ largest components of the transmitted message. The vertical line at $\tau \approx 0.19$ marks the theoretical phase transition for successful decoding under a benign corruption. The vertical line at $\tau \approx 0.018$ provides a uniform guarantee applicable to adversarial corruptions. For $\tau \lesssim 0.018$, with high probability, our protocol will decode a message subject to *any* $\tau$-sparse corruption whatsoever. See Section 1.3.1 and Section 6.2 for more details.

### 1.3.1   A secure communications protocol that is robust to sparse errors

Suppose we wish to securely transmit a binary message across a communications channel. We can obtain strong guarantees of security by modulating the message with a random rotation before transmission [**85**], [**86**]. Our theory shows that decoding the message via demixing also makes this secure scheme perfectly robust to sparse corruptions such as erasures or malicious interference.

Consider the following simple communications protocol. We model the binary message as a sign vector $\boldsymbol{m}_0 \in \{\pm 1\}^d$. Choose a random orthogonal matrix $\boldsymbol{Q} \in \mathsf{O}_d$. The transmitter sends the scrambled message $\boldsymbol{s}_0 = \boldsymbol{Q}\boldsymbol{m}_0$ across the channel, where it is corrupted by an unknown sparse vector $\boldsymbol{c}_0 \in \mathbb{R}^d$. The receiver must determine the original message given only the corrupted signal

$$\boldsymbol{z}_0 = \boldsymbol{s}_0 + \boldsymbol{c}_0 = \boldsymbol{Q}\boldsymbol{m}_0 + \boldsymbol{c}_0$$

and knowledge of the scrambling matrix $\boldsymbol{Q}$.

This signal model is perfectly suited to the demixing recipe of Section 1.2. The discussion in Section 1.2.1 indicates that the $\ell_1$ and $\ell_\infty$ norms are natural complexity measures for the structured signals $\boldsymbol{c}_0$ and $\boldsymbol{m}_0$. Since the message $\boldsymbol{m}_0$ is a sign vector, we also have the side information $\|\boldsymbol{m}_0\|_{\ell_\infty} = 1$. Our receiver then recovers the message with the convex demixing method

$$
\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{c}\|_{\ell_1} \\
\text{subject to} \quad & \|\boldsymbol{m}\|_{\ell_\infty} \le 1 \text{ and } \boldsymbol{c} + \boldsymbol{Q}\boldsymbol{m} = \boldsymbol{z}_0,
\end{aligned}
\tag{1.4}
$$

where the decision variables are $\boldsymbol{c}, \boldsymbol{m} \in \mathbb{R}^d$. This method succeeds if $(\boldsymbol{c}_0, \boldsymbol{m}_0)$ is the unique optimal point of (1.4).

In Section 6.2, we apply the general theory developed in this work to study this communications protocol. Before summarizing the results of this analysis, we fix some notation. Suppose the corruption
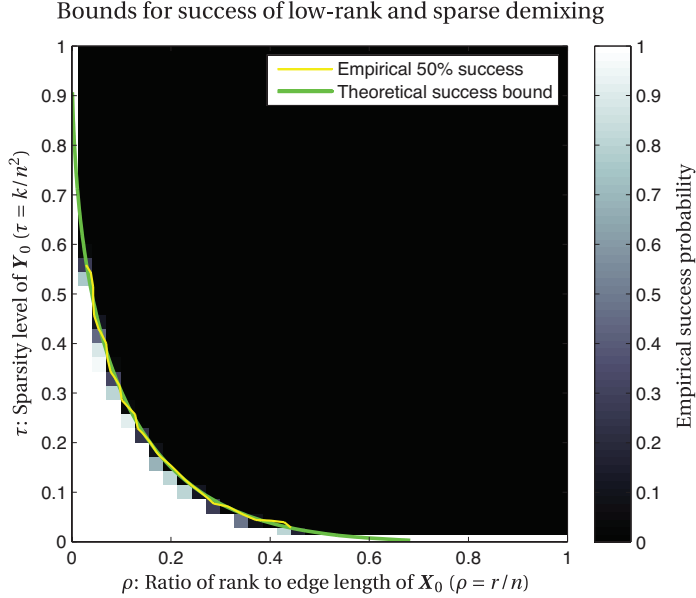
**Figure 4** *Low-rank matrix recovery with sparse corruptions.* The horizontal axis is the normalized rank $\rho = \mathrm{rank}(X_0)/n$, and the vertical axis is the sparsity level $\tau = \mathrm{nnz}(Y_0)/n^2$. The intensity of the background denotes the empirical probability that (1.5) recovers $(X_0, Y_0)$ from the observation $Z_0 = X_0 + \mathcal{Q}(Y_0)$. In the region below the green curve, the convex demixing method (1.5) recovers a low-rank matrix $X_0$ from a randomly rotated sparse corruption $\mathcal{Q}(Y_0)$ with overwhelming probability in high dimensions. See Section 1.3.2 and Section 6.3 for more details.

$c_0 \in \mathbb{R}^d$ is $\tau$-sparse; that is, $\mathrm{nnz}(c_0) = \lceil \tau d \rceil$ for some $\tau \in [0, 1]$. We further distinguish between two types of corruption. A *benign* corruption $c_0$ is independent of the scrambling matrix $Q$. In contrast, an *adversarial* corruption may depend on both $Q$ and $m_0$. Adversarial corruptions also include nonlinear effects that are not necessarily malicious. For example, we can model an erasure at the $i$th time instant by taking $(c_0)_i = -(Q m_0)_i$.

Figure 3 presents the results of a numerical experiment on this communications protocol; the complete experimental procedure is detailed in Section 6.2.3. Briefly, we consider messages of length $d = 100$ and $d = 300$, and we let the sparsity $\tau$ range over the interval $[0, 0.35]$. We test the benign case by adding a $\tau$-sparse corruption that is independent from $Q$. We also consider a particular adversarial corruption in which we set the $\lceil \tau d \rceil$ largest-magnitude entries in the transmitted message $s_0$ to zero. The curves indicate the empirical probability that the protocol succeeds as a function of $\tau$.

In the benign case, our theory shows that there exists a phase transition in the success probability of the convex demixing method (1.4) at sparsity level $\tau \approx 0.19$. The empirical 50% failure threshold for benign corruptions closely matches this prediction. In the adversarial case, our results guarantee that with high probability, our protocol will tolerate all corruptions that affect no more than 1.8% of the components in the received message $z_0$. This bound is conservative for the type of adversarial corruption in the numerical experiment; this is not surprising because we may not have constructed the worst possible corruption.

### 1.3.2 Low-rank matrix recovery with generic sparse corruptions

Consider now the *matrix* observation $Z_0 = X_0 + \mathcal{Q}(Y_0) \in \mathbb{R}^{n \times n}$, where $X_0$ has low rank, $Y_0$ is sparse, and $\mathcal{Q}$ is a random rotation on $\mathbb{R}^{n \times n}$. This type of signal provides a highly stylized model for applications such as latent variable selection [**14**], [**17**] and robust principal component analysis [**12**]. In these settings, $X_0$ has low rank because the underlying data is drawn from a low-dimensional linear model, while $\mathcal{Q}(Y_0)$ represents a

corruption. (Note, however, that this stylized model is not equivalent to pre- and post-multiplying $Y_0$ by independent random rotations of $\mathbb{R}^n$.) We aim to discover the matrix $X_0$ given the corrupted observation $Z_0$ and the rotation $\mathcal{Q}$.

We follow the now-familiar pattern of Section 1.2. The Schatten 1-norm $\|\cdot\|_{S_1}$ serves as a natural complexity measure for the low-rank structure of $X_0$, and the matrix $\ell_1$ norm $\|\cdot\|_{\ell_1}$ is appropriate for the sparse structure of $Y_0$. We further assume the side information $\alpha = \|Y_0\|_{\ell_1}$. We then solve

$$
\begin{aligned}
&\text{minimize} \quad \|X\|_{S_1} \\
&\text{subject to} \quad \|Y\|_{\ell_1} \leq \alpha \text{ and } X + \mathcal{Q}(Y) = Z_0.
\end{aligned}
\tag{1.5}
$$

This convex demixing method succeeds if $(X_0, Y_0)$ is the unique solution to (1.5).

Figure 4 displays the results of a numerical experiment on this approach to rank–sparsity demixing. We take the matrix side length $n = 35$ and draw a random rotation $\mathcal{Q}$ for $\mathbb{R}^{n \times n}$. For parameters $0 \leq \rho, \tau \leq 1$, we generate matrices $X_0$ and $Y_0$ such that $\text{rank}(X_0) = [\rho n]$ and $\text{nnz}(Y_0) = [\tau n^2]$. The background shading indicates the empirical probability that (1.5) succeeds given the observation $Z_0 = X_0 + \mathcal{Q}(Y_0)$. We mark the empirical 50% success probability with a yellow curve. See Section 6.3.1 for the experimental details.

The results of this work show that, with high probability, program (1.5) succeeds so long as the pair $(\rho, \tau)$ lies below the green curve on Figure 4. When the rank parameter $\rho$ is small, our theoretical bound closely tracks the phase transition visible in the numerical experiment, although the bound appears loose when $\rho$ is larger. Section 6.3 provides further details.

### 1.3.3 Matrix demixing mix-and-match

Our results are not restricted to the convex demixing methods (1.1), (1.4) or (1.5). Let us mention a few other situations we can analyze using the theory developed in this work. With a tractable convex program, it is possible to demix

- An orthogonal matrix from a matrix that is sparse in a random orthogonal basis,
- A randomly oriented sign matrix from a sufficiently low-rank matrix, and
- A randomly rotated low-rank matrix from an orthogonal matrix.

See Section 6.4 for the details.

## 1.4 Theoretical insights

Our approach reveals a number of theoretical insights.

**Design of convex demixing methods for incoherent structures.** In the incoherent regime we analyze, the parameters that determine when the convex demixing method (1.2) succeeds reflect the structures in the constituent signals and the associated complexity measures. These summary parameters are independent of the relationship between the two incoherent structures. We discuss this fact and its consequences for the design of demixing procedures in Section 4.2.1.

**Connection with linear inverse problems.** The parameters that determine success of the convex demixing method (1.2) are closely related to number of random linear measurements required to identify a structured signal. In Section 5.2, we leverage this relationship to compute these parameters from Gaussian width bounds developed in [15].

**Phase transitions.** Our theory indicates that there is often a phase transition in the behavior of the demixing method (1.2). See Section 4.2.2 for a discussion of this point.

## 1.5 Outline

This work begins with demixing in the deterministic setting. Section 2 describes the geometry of the convex demixing method (1.2) and provides a geometric characterization of successful demixing.

Section 3 presents a random model for incoherence along with some techniques from spherical integral geometry that allow us to analyze this model. In Section 4, these ideas yield theory that predicts success and failure regimes for the convex demixing method (1.2). Section 5 develops methods for computing the parameters necessary to apply the theorems of Section 4.

In Section 6, we analyze the application problems described in Sections 1.1 and 1.3. Section 7 concludes with a discussion of this work's place in the literature and future directions.

## 1.6 Notation and conventions

All variables are real valued. We write $\lfloor t \rfloor$, $\lceil t \rceil$, and $[t]$ for the floor, ceiling, and rounded integer values of $t$. The signum function is $\text{sgn}(t) := t/|t|$ for $t \neq 0$ and $\text{sgn}(0) := 0$. Bold lowercase letters represent vectors, and bold capital letters are matrices. The $i$th element of a vector is written $x_i$ or $(\boldsymbol{x})_i$, while the $(i,j)$th element of a matrix is $X_{ij}$ or $(\boldsymbol{X})_{ij}$. We express the transpose of $\boldsymbol{X}$ as $\boldsymbol{X}^*$. The vector signum $\text{sgn}(\boldsymbol{x})$ is defined by applying the signum elementwise.

The symbol $\|\cdot\|_{\ell_p}$ stands for the $\ell_p$ vector norm on $\mathbb{R}^d$, defined by $\|\boldsymbol{x}\|_{\ell_p}^p := \sum_{i=1}^d |x_i|^p$ when $1 \leq p < \infty$ and $\|\boldsymbol{x}\|_{\ell_\infty} := \max_{i=1,\dots,d} |x_i|$ when $p = \infty$. The $\ell_p$ norm of a matrix treats the matrix as a vector and applies the corresponding vector $\ell_p$ norm. The Schatten 1-norm $\|\boldsymbol{X}\|_{S_1}$ is the sum of the singular values of a matrix $\boldsymbol{X}$, while the operator, or spectral, norm $\|\boldsymbol{X}\|_{\text{Op}}$ returns the maximum singular value of $\boldsymbol{X}$.

We reserve the symbols $f$ and $g$ for convex functions. A convex function may take the value $+\infty$, but we assume that all convex functions are *proper*; that is, each convex function takes on at least one finite value and never takes the value $-\infty$.

The Euclidean unit sphere in $\mathbb{R}^d$ is the set $\mathsf{S}^{d-1}$. The orthogonal group—the set of $d \times d$ orthogonal matrices—is denoted $\mathsf{O}_d$. The subset of $\mathsf{O}_d$ with determinant one (the *special* orthogonal group) is $\mathsf{SO}_d$. For brevity, we use the term *basis* to refer to an orthogonal matrix from $\mathsf{O}_d$. In the sequel, the letter $\boldsymbol{Q}$ will always refer to a basis.

The symbol $\mathbb{P}$ denotes the probability of an event. Gaussian vectors and matrices have independent standard normal entries. A *random basis* is a matrix drawn from the Haar measure on $\mathsf{O}_d$.

A special note is in order when our observations are matrices. The space of matrices $\mathbb{R}^{m \times n}$ is equipped with a natural isomorphism to $\mathbb{R}^{mn}$ through the $\text{vec}(\cdot)$ operator, which stacks the columns of a matrix on top of one another to form a tall vector. We define a random basis $\mathscr{Q}$ for $\mathbb{R}^{m \times n}$ by $\mathscr{Q}(\boldsymbol{Y}) = \text{vec}^{-1}(\boldsymbol{Q}\,\text{vec}(\boldsymbol{Y}))$, where $\boldsymbol{Q}$ is a random basis for $\mathbb{R}^{mn}$. It is easily verified that such a $\mathscr{Q}$ is a random basis for $\mathbb{R}^{m \times n}$ equipped with the Euclidean structure induced by the Frobenius norm.

We deal frequently with convex cones, which are positively homogeneous convex sets. For any cone $K$, we define the polar cone

$$K^\circ := \{\boldsymbol{y} : \langle \boldsymbol{y}, \boldsymbol{x} \rangle \leq 0 \text{ for all } \boldsymbol{x} \in K\}.$$

A *polyhedral cone* is the intersection of a finite number of closed halfspaces, each containing the origin. One important polyhedral cone is the nonnegative orthant, defined by

$$\mathbb{R}_+^d := \{\boldsymbol{x} \in \mathbb{R}^d : x_i \geq 0, \ i = 1, \dots, d\}.$$

For a set $A \subset \mathbb{R}^d$, we write $\text{conv}(A)$ for its convex hull and $\overline{A}$ for its closure. Two cones $K_1, K_2 \subset \mathbb{R}^d$ are *congruent*, written $K_1 \cong K_2$, if there is a basis $\boldsymbol{U} \in \mathsf{O}_d$ such that $K_1 = \boldsymbol{U}K_2$.

# 2 The geometry of demixing

This short section lays the geometric foundation for the rest of this work. Section 2.1 describes the local behavior of convex functions in terms of special convex cones. In Section 2.2, this geometric view yields a

concise characterization of successful demixing in terms of the configuration of two cones. The results in this section are deterministic, that is, they hold for any fixed basis $\boldsymbol{Q}$.

## 2.1 Feasible cones

The success of the convex demixing method (1.2) depends on the properties of the complexity measures $f$ and $g$ at the structured vectors $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$. The following definition captures the local behavior of a convex function.

**Definition 2.1** (Feasible cone). The *feasible cone* of a convex function $f$ at a point $\boldsymbol{c}\boldsymbol{x}$ is defined as the cone of directions at which $f$ is locally nondecreasing about $\boldsymbol{x}$:

$$\mathscr{F}(f, \boldsymbol{x}) := \bigcup_{\lambda > 0} \{\boldsymbol{\delta} : f(\boldsymbol{x} + \lambda \boldsymbol{\delta}) \leq f(\boldsymbol{x})\}.$$

**Example 2.2** (The feasible cone of the $\ell_\infty$ norm at sign vectors). Let $\boldsymbol{x} \in \{\pm 1\}^d$ be a sign vector. Then it is easy to check that $\|\boldsymbol{x} + \lambda \boldsymbol{\delta}\|_{\ell_\infty} \leq \|\boldsymbol{x}\|_{\ell_\infty} = 1$ for some $\lambda > 0$ if and only if $\operatorname{sgn}(\delta_i) = -x_i$ for all $i = 1, 2, \ldots, d$. Therefore, the feasible cone of the $\ell_\infty$ norm at $\boldsymbol{x}$ is congruent to the nonnegative orthant:

$$\mathscr{F}(\|\cdot\|_{\ell_\infty}, \boldsymbol{x}) = \left\{ \boldsymbol{\delta} : \operatorname{sgn}(\boldsymbol{\delta}) = -\boldsymbol{x} \right\} \cong \mathbb{R}_+^d.$$

The feasible cone is always a convex cone containing zero, but it is not necessarily closed. If $\mathscr{A}$ is a set of atoms and $\boldsymbol{a} \in \mathscr{A}$, the feasible cone $\mathscr{F}(f_{\mathscr{A}}, \boldsymbol{a})$ of the atomic gauge $f_{\mathscr{A}}$ at the atom $\boldsymbol{a}$ tends to be small because the unit ball of $f_{\mathscr{A}}$ is the smallest convex set containing all of the atoms. See Figure 2 for an illustration. The positive homogeneity of atomic gauges further implies that the feasible cones of atomic gauges do not depend on the scaling of a vector, in the sense that

$$\mathscr{F}(f_{\mathscr{A}}, \boldsymbol{x}) = \mathscr{F}(f_{\mathscr{A}}, \lambda \cdot \boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ and any $\lambda > 0$.

**Remark 2.3.** Definition 2.1 is equivalent to the definition of the "tangent cone" appearing in [**15,** Eq. (8)]. However, that definition differs slightly from the standard definition of a tangent cone; cf., [**66,** Thm. 6.9]. The cone of feasible directions [**62,** p. 33] is the closest relative of the feasible cone that we have identified in the literature, and this is the source of our terminology.

## 2.2 A geometric characterization of optimality

The following lemma provides a geometric characterization for success in the convex demixing method (1.2) in terms of the configuration of two feasible cones. This is the main result of this section.

**Lemma 2.4.** *Program* (1.2) *succeeds at demixing* $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ *if and only if* $\mathscr{F}(f, \boldsymbol{x}_0) \bigcap \left(-\boldsymbol{Q}\mathscr{F}(g, \boldsymbol{y}_0)\right) = \{\boldsymbol{0}\}$.

In words, the demixing method (1.2) succeeds if and only if the two feasible cones are rotated so that they intersect trivially. Intuitively, we expect that many bases $\boldsymbol{Q}$ satisfy this condition when the two feasible cones are small. This observation provides further support for choosing atomic gauges as our complexity measures. We illustrate Lemma 2.4 in Figure 5.

The proof of Lemma 2.4 requires two technical propositions. The first is an alternative characterization of feasible cones.

**Proposition 2.5.** *Let $f$ be a convex function. Then $\boldsymbol{\delta} \in \mathscr{F}(f, \boldsymbol{x})$ if and only if there is a number $\lambda_0 > 0$ such that, for all $\lambda \in [0, \lambda_0]$, we have $f(\boldsymbol{x} + \lambda \boldsymbol{\delta}) \leq f(\boldsymbol{x})$.*
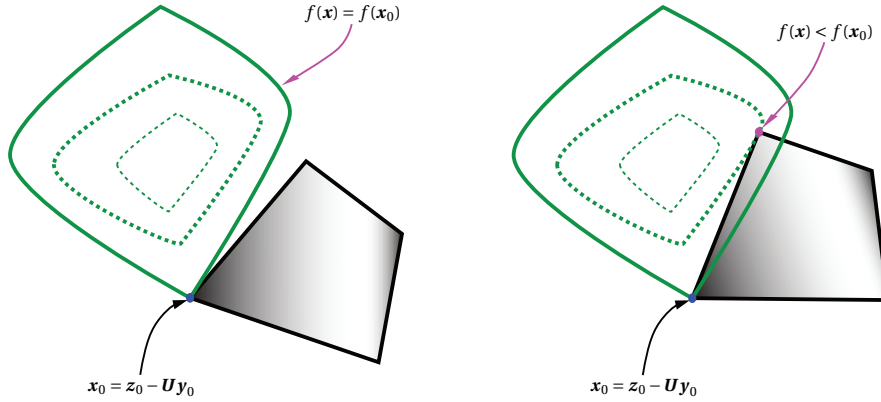
**Figure 5** *Geometry of demixing.* [Left] Demixing succeeds: Every feasible perturbation about $x_0$ *(gray area)* increases the objective function *(green level lines)*. [Right] Demixing fails: Some feasible perturbations decrease the objective value. In each panel, the success or failure of the convex demixing method (1.2) is determined by a configuration of two cones. This fact forms the content of Lemma 2.4.

*Proof.* The "if" part is immediate: given any such $\lambda_0$, the assumption $f(x + \lambda_0 \delta) \leq f(x)$ implies $\delta \in \mathscr{F}(f, x)$ by the definition of feasible cones. The other direction follows from convexity. Indeed, suppose $\delta \in \mathscr{F}(f, x)$, so that there exists a number $\lambda_0 > 0$ for which $f(x + \lambda_0 \delta) \leq f(x)$. By convexity, the map $\lambda \mapsto f(x + \lambda \delta)$ lies below the chord connecting 0 and $\lambda_0$, which is the claim. □

Our second technical proposition is a change of variables formula for the feasible cone under a nondegenerate affine transformation.

**Proposition 2.6.** *Let $g$ be any convex function, and define $h(x) := g\big(A^{-1}(z - x)\big)$ for some invertible matrix $A$. Then $\mathscr{F}(h, x) = -A\mathscr{F}\big(g, A^{-1}(z - x)\big)$.*

The proof of Proposition 2.6 follows directly from the definition of a feasible cone. We omit the details.

*Proof of Lemma 2.4.* Because $Q$ is unitary, we may eliminate the variable $y$ in (1.2) via the equality constraint $y = Q^*(z_0 - x)$. Therefore, $(x_0, y_0)$ is the unique optimum of (1.2) if and only if $x_0$ is the unique optimum of

$$
\begin{aligned}
\text{minimize} \quad & f(x) \\
\text{subject to} \quad & g\big(Q^*(z_0 - x)\big) \leq g\big(Q^*(z_0 - x_0)\big) = g(y_0),
\end{aligned}
\tag{2.1}
$$

with decision variable $x$. The equality in (2.1) follows from the definition $z_0 = x_0 + Q y_0$. The original claim thus reduces to the statement that $x_0$ is the unique optimum of (2.1) if and only if $\mathscr{F}(f, x_0) \cap \big(-Q\mathscr{F}(g, y_0)\big) = \{0\}$. The rest of the proof is devoted to this claim.

($\Leftarrow$) Suppose $\mathscr{F}(f, x_0) \cap \big(-Q\mathscr{F}(g, y_0)\big) = \{0\}$. We show that $x_0$ is the unique optimum of (2.1) by verifying that the strict inequality $f(x) > f(x_0)$ holds for any feasible point $x$ of (2.1) with $x \neq x_0$.

To this end, assume $x$ is feasible for (2.1) and $x \neq x_0$. Feasibility of $x$ is equivalent to

$$
g\big(Q^*(z_0 - x)\big) \leq g\big(Q^*(z_0 - x_0)\big) = g(y_0).
$$

By definition of the feasible cone and the transformation rule of Proposition 2.6, the inequality above implies

$$
x - x_0 \in \mathscr{F}\Big(g\big(Q^*(z_0 - \cdot)\big), x_0\Big) = -Q\mathscr{F}\big(g, Q^*(z_0 - x_0)\big) = -Q\mathscr{F}(g, y_0);
$$

The final equality above follows from $z_0 = x_0 + Q y_0$.

12

Since $x \neq x_0$, the assumption $\mathcal{F}(f, x_0) \cap \left(-Q\mathcal{F}(g, y_0)\right) = \{0\}$ implies $x - x_0 \notin \mathcal{F}(f, x_0)$. By the definition of feasible cones, we must have $f(x) = f\left(x_0 + (x - x_0)\right) > f(x_0)$. We have deduced $f(x) > f(x_0)$ for every feasible $x \neq x_0$, and so conclude that $x_0$ is the unique optimum of (2.1).

($\Rightarrow$) Suppose $x_0$ is the unique optimum of (2.1). Let $\delta$ be some vector in the intersection $\mathcal{F}(f, x_0) \cap \left(-Q\mathcal{F}(g, y_0)\right)$. We must show $\delta = 0$.

As $\delta \in \mathcal{F}(f, x_0)$, Proposition 2.5 implies that $f(x_0) \geq f(x_0 + \lambda\delta)$ for all sufficiently small $\lambda > 0$. Applying Proposition 2.5 to the fact $-Q^*\delta \in \mathcal{F}(g, y_0)$ yields

$$g(y_0) \geq g(y_0 - \lambda Q^*\delta) = g\left(Q^*\left(z_0 - (x_0 + \lambda\delta)\right)\right)$$

for all sufficiently small $\lambda > 0$. We have used the relation $z_0 = x_0 + Qy_0$ again here. In summary, for some small enough $\lambda > 0$, the perturbed point $x_0 + \lambda\delta$ is feasible for (2.1), and its objective value is no larger than $f(x_0)$. In other words, $x_0 + \lambda\delta$ is an optimal point of (2.1). But $x_0$ is the unique optimal point of (2.1) by assumption, so we must have $\delta = 0$. This is the claim. □

# 3   Background from integral geometry

By coupling the random basis model of Section 1.2.2 with the optimality condition of Lemma 2.4, the optimality condition for the demixing method (1.2) boils down to a geometric question: *When does a randomly oriented cone strike a fixed cone?*

This section provides a background in spherical integral geometry, a subfield of integral geometry that studies random configurations of cones and quantities related to these configurations [**69,** Section 6.5]. This theory provides an *exact* expression, called the spherical kinematic formula, for the probability that the convex demixing method (1.2) succeeds under our random model. Unfortunately, the quantities involved in the spherical kinematic formula are typically difficult to compute. To ease this burden, Section 3.3 defines geometric summary parameters that greatly simplify the application of the spherical kinematic formula.

We use subspaces and the nonnegative orthant as running examples to illustrate the concepts from integral geometry. These are not simply toy examples. A subspace plays an important role in the linear inverse problems in Section 5, while the orthant, which is congruent to the feasible cone of the $\ell_\infty$ norm at a sign vector, appears at several points in our examples in Section 6.

## 3.1   Spherical intrinsic volumes

We begin our introduction to integral geometry with fundamental geometric parameters known as spherical intrinsic volumes [**54**]. Spherical intrinsic volumes quantify geometric properties of convex cones such as the fraction of space a cone consumes (a type of volume), the fraction of space taken by the corresponding dual cone, and quantities akin to surface area. The following characterization [**3,** Proposition 4.4.6] is convenient.

**Definition 3.1** (Spherical intrinsic volumes)**.** Let $K \subset \mathbb{R}^d$ be a polyhedral convex cone, and define the Euclidean projection onto $K$ by

$$\Pi_K(x) := \arg\min_{y \in K} \|x - y\|_{\ell_2}.$$

For $i = -1, 0, \ldots, d-1$, we define the $i$th *spherical intrinsic volume of $K$* by

$$v_i(K) := \mathbb{P}\left\{ \begin{array}{l} \Pi_K(\omega) \text{ lies in the relative interior} \\ \text{of an } (i+1)\text{-dimensional face of } K \end{array} \right\},$$

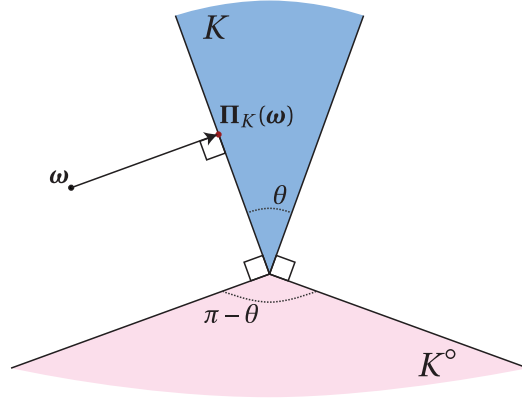where the vector $\omega$ is drawn from the standard Gaussian distribution on $\mathbb{R}^d$.

**Figure 6** *Spherical intrinsic volumes in $\mathbb{R}^2$*. A convex cone $K \subset \mathbb{R}^2$ of solid angle $\theta$ has four faces: one 2-dimensional face *(light blue)*, two 1-dimensional faces *(heavy blue lines)*, and one 0-dimensional face *(blue dot)*. The projection $\Pi_K(\boldsymbol{w})$ of a Gaussian vector $\boldsymbol{\omega}$ onto $K$ lies in the 2-dimensional face when $\boldsymbol{\omega}$ is in the blue region; a 1-dimensional face when $\boldsymbol{\omega}$ is in the white region; and the 0-dimensional face when $\boldsymbol{\omega}$ is in the red region. By Definition 3.1 of the spherical intrinsic volumes, we have $v_1(K) = \theta/(2\pi)$, $v_0(K) = 1/2$, and $v_{-1}(K) = (\pi - \theta)/(2\pi)$.

Although computing spherical intrinsic volumes is often challenging, a direct application of Definition 3.1 can bear fruit in some situations. In Figure 6, we demonstrate how to compute spherical intrinsic volumes for cones in $\mathbb{R}^2$. A subspace of $\mathbb{R}^d$ provides another simple example.

**Proposition 3.2** (Spherical intrinsic volumes of a subspace)**.** *Suppose $L \subset \mathbb{R}^d$ is a linear subspace of dimension $n$. Then $v_i(L) = \delta_{i,n-1}$, where $\delta_{i,j}$ is the Kronecker $\delta$ function.*

*Proof.* A subspace $L$ of dimension $n$ is a polyhedral cone with a single face, with dimension $n$. The projection of any point onto $L$ lies in the relative interior of this face. The claim follows immediately from Definition 3.1. $\qquad\square$

The computation of the spherical intrinsic volumes for the nonnegative orthant requires only some elementary probability theory. We illustrate the following result in Figure 7.

**Proposition 3.3** (Spherical intrinsic volumes of the orthant [**3,** Example 4.4.7])**.** *The spherical intrinsic volumes of the nonnegative orthant $\mathbb{R}^d_+$ are given by the binomial sequence*

$$v_i(\mathbb{R}^d_+) = 2^{-d}\binom{d}{i+1} \tag{3.1}$$

*for $i = -1, 0, \ldots, d-1$.*

*Proof.* The Euclidean projection onto the nonnegative orthant is given by the componentwise threshold operation

$$\left(\Pi_{\mathbb{R}^d_+}(\boldsymbol{\omega})\right)_i = \max\{\omega_i, 0\}.$$

Therefore, the projection $\Pi_{\mathbb{R}^d_+}(\boldsymbol{\omega})$ lies in the relative interior of an $(i+1)$-dimensional face of $\mathbb{R}^d_+$ if and only if $\boldsymbol{\omega}$ has exactly $i+1$ strictly positive values.

When $\boldsymbol{\omega}$ is drawn from a standard Gaussian distribution, the number of positive entries is distributed as a binomial random variable. Hence, $\Pi_{\mathbb{R}^d_+}(\boldsymbol{\omega})$ lies in an $(i+1)$-dimensional face of $\mathbb{R}^d_+$ with probability $2^{-d}\binom{d}{i+1}$. This is precisely the definition of the $i$th spherical intrinsic volume $v_i(\mathbb{R}^d_+)$. $\qquad\square$
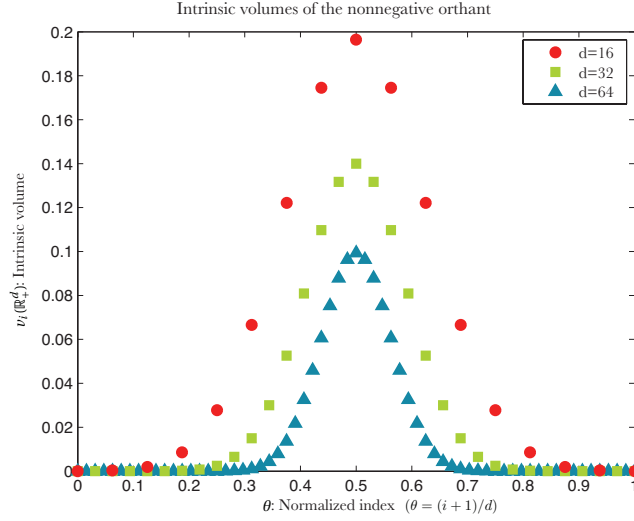
**Figure 7** *Spherical intrinsic volumes of the orthant.* The spherical intrinsic volumes of the nonnegative orthant are given by scaled binomial coefficients (3.1). The scale on the horizontal axis is the normalized index $\theta = (i+1)/d$.

**Remark 3.4** (Extension to nonpolyhedral cones)**.** The definition of spherical intrinsic volumes extends to all closed convex cones by approximation with polyhedral cones, but note that the probabilistic characterization above *does not hold* for nonpolyhedral cones. The technical details involve continuity properties of the spherical intrinsic volumes under the spherical Hausdorff metric. This theory is developed in [**38**], [**39**]. See [**69,** Ch. 6.5] for a self-contained overview of spherical integral geometry developed via polyhedral approximation, or see [**3,** Sec. 4] for a development using tools from differential geometry.

## 3.2 Key facts from integral geometry

We now collect some of the properties of spherical intrinsic volumes that are required for our development. We start with some elementary facts.

**Fact 3.5.** *Let $K \subset \mathbb{R}^d$ be a closed convex cone. Then*

1. *(Positivity) $v_i(K) \geq 0$ for each $i = -1, 0, \dots, d-1$,*

2. *(Unit-sum) $\sum_{i=-1}^{d-1} v_i(K) = 1$, and*

3. *(Basis invariance) For any basis $\boldsymbol{U} \in \mathsf{O}^d$ and index $i = -1, 0, \dots, d-1$, we have $v_i(\boldsymbol{U}K) = v_i(K)$.*

*Proof sketch.* First, assume $K$ is a polyhedral cone. The positivity of spherical intrinsic volumes follows from the positivity of probability. The unit-sum rule follows immediately from the fact that the projection $\Pi_K(\boldsymbol{x})$ lies in the relative interior of a unique face of $K$. Finally, the basis invariance of spherical intrinsic volumes is immediate from the corresponding invariance of the Gaussian distribution. Continuity of the spherical intrinsic volumes under the spherical Hausdorff metric implies that these facts must hold for all closed convex cones by approximation with polyhedral cones. □

We conclude our background discussion with the following remarkable formula for the probability that a randomly oriented cone strikes a fixed cone. In view of Lemma 2.4, this result is fundamental to our understanding of the probability that the convex demixing method (1.2) succeeds under the random basis model.

**Fact 3.6** (Spherical kinematic formula [**69**, p. 261])**.** *Let $K$ and $\tilde{K}$ be closed convex cones in $\mathbb{R}^d$, at least one of which is not a subspace, and let $\boldsymbol{Q}$ be a random basis. Then*

$$\mathbb{P}\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\}\} = \sum_{k=0}^{d-1} \left(1 + (-1)^k\right) \sum_{i=k}^{d-1} v_i(K) \cdot v_{d-1-i+k}(\tilde{K}). \tag{3.2}$$

**Remark 3.7.** Fact 3.6 is usually stated for random rotations drawn according to the Haar measure on the special orthogonal group $\mathsf{SO}_d$, but the unitary invariance given by Fact 3.5.3 readily implies that the spherical kinematic formula holds for $\boldsymbol{Q}$ drawn from the Haar measure on the orthogonal group $\mathsf{O}_d$.

**Remark 3.8.** The spherical Gauss–Bonnet formula (Fact B.2 in Appendix B.1) can be used to eliminate the apparent asymmetry between $K$ and $\tilde{K}$ in (3.2). In particular, the identity

$$\mathbb{P}\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\}\} = \mathbb{P}\{\boldsymbol{Q}K \cap \tilde{K} \neq \{\boldsymbol{0}\}\}$$

holds for any convex cones $K$ and $\tilde{K}$.

## 3.3   High-dimensional decay of spherical intrinsic volumes

The spherical kinematic formula (3.2), coupled with the geometric optimality conditions of Lemma 2.4, provides an exact expression for the probability that the demixing problem (1.2) succeeds. Nevertheless, the formula involves the spherical intrinsic volumes of two cones, and it is challenging to determine these quantities directly from the definition except in simple situations.

To confront this challenge, we seek summary statistics for the intrinsic volumes as the dimension $d \to \infty$. We motivate our approach with the orthant. In Figure 7, we see that the spherical intrinsic volumes of the orthant $v_i(\mathbb{R}^d_+)$ decay rapidly as $d \to \infty$ when the index $i$ falls outside of the region $i \approx d/2$. The intrinsic volumes with indices $i$ far away from $d/2$ will contribute very little to the sum (3.2) appearing in the kinematic formula. This observation simplifies the application of the kinematic formula when one of the cones is congruent to the orthant.

In general, we might hope that for some sequence of cones $K^{(d)} \in \mathbb{R}^d$, the spherical intrinsic volumes $v_i(K^{(d)})$ decay rapidly as $d \to \infty$ for indices $i$ outside of some interval $[\kappa_\star d, \theta_\star d]$. We codify this behavior with *decay thresholds* that indicate which intrinsic volumes are very small as the ambient dimension $d$ grows.

**Definition 3.9** (Decay threshold)**.** Let $\mathscr{D} \subset \mathbb{N}$ be an infinite set of indices, and suppose $\{K^{(d)} : d \in \mathscr{D}\}$ is an ensemble of closed convex cones with $K^{(d)} \subset \mathbb{R}^d$ for each $d \in \mathscr{D}$. We say that $\theta_\star \in [0,1]$ is an *upper decay threshold* for $\{K^{(d)}\}$ if, for every $\theta > \theta_\star$, there exists an $\varepsilon > 0$ such that, for all sufficiently large $d \in \mathscr{D}$, we have

$$v_i(K^{(d)}) \leq \mathrm{e}^{-\varepsilon d} \quad \text{for every } i \geq \lceil \theta d \rceil. \tag{3.3}$$

On the other hand, we say that $\kappa_\star \in [0,1]$ is a *lower decay threshold* for $\{K^{(d)}\}$ if, for every $\kappa < \kappa_\star$, there exists an $\varepsilon > 0$ such that, for all $d \in \mathscr{D}$ sufficiently large, we have

$$v_i(K^{(d)}) \leq \mathrm{e}^{-\varepsilon d} \quad \text{for every } i \leq \lceil \kappa d \rceil.$$

We extend these definitions to non-closed cones by taking the closure. We say $\theta_\star$ is an upper decay threshold for $\{K^{(d)}\}$ if and only if it is an upper decay threshold for $\{\overline{K}^{(d)}\}$. Similarly, $\kappa_\star$ is a lower decay threshold for $\{K^{(d)}\}$ if and only if it is a lower decay threshold for $\{\overline{K}^{(d)}\}$.

When it will not cause confusion, we omit the index set $\mathscr{D}$.

**Remark 3.10** (Nonuniqueness of decay thresholds)**.** The upper decay threshold $\theta_\star$ for an ensemble $\{K^{(d)}\}$ defined above is not unique since any $\theta' > \theta_\star$ is also a decay threshold for $\{K^{(d)}\}$. An analogous comment holds for the lower decay threshold.
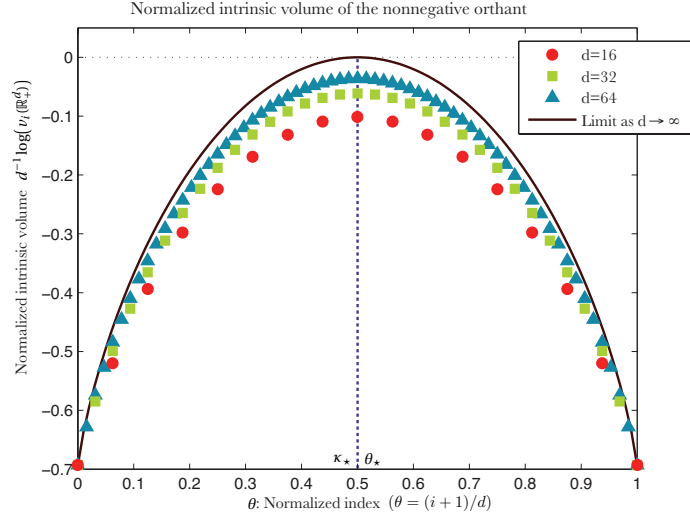
**Figure 8** *Normalized spherical intrinsic volumes of the orthant.* The graph illustrates the computation in Proposition 3.12. The normalized spherical intrinsic volumes $d^{-1}\log(v_i(\mathbb{R}_+^d))$ appear against the normalized index $\theta = (i+1)/d$ for several values of dimension $d$. The solid curve is the uniform limit $H(\theta) - \log(2)$ of these rescaled volumes, where $H(\theta)$ is the bit entropy (3.6). The smallest possible upper decay threshold is the rightmost point where the solid curve crosses the zero level line (dotted), and the largest possible lower decay threshold is the leftmost point where the zero level line crosses the solid curve. The solid curve takes its unique maximum value of zero at $\theta = \frac{1}{2}$ so that $\theta_\star = \frac{1}{2}$ is an upper decay threshold and $\kappa_\star = \frac{1}{2}$ is a lower decay threshold for the ensemble $\{\mathbb{R}^d : d = 1, 2, \ldots\}$ of nonnegative orthants.

### 3.3.1 Examples of decay thresholds

Since the intrinsic volumes are positive and sum to one (Fact 3.5), not every intrinsic volume is exponentially small in the ambient dimension $d$. In particular, the inequality $\kappa_\star \leq \theta_\star$ between lower and upper decay thresholds always holds. In practice, however, we can often find decay thresholds that satisfy the equality $\kappa_\star = \theta_\star$, as the following examples demonstrate.

**Proposition 3.11** (Upper decay threshold for subspaces)**.** *Let $\mathscr{D} \subset \mathbb{N}$ be an infinite set of indices, and let $\{L^{(d)} : d \in \mathscr{D}\}$ be an ensemble of linear subspaces with $L^{(d)} \subset \mathbb{R}^d$ for each $d \in \mathscr{D}$. Suppose there exists a parameter $\sigma \in [0,1]$ such that $\dim(L^{(d)}) = \lceil \sigma d \rceil$. Then $\theta_\star = \sigma$ is an upper decay threshold for the ensemble $\{L^{(d)}\}$, and $\kappa_\star = \sigma$ is a lower decay threshold for the ensemble $\{L^{(d)}\}$.*

*Proof.* Let $n = n^{(d)} = \dim(L^{(d)})$. By Proposition 3.2, we have $v_i(L^{(d)}) = \delta_{i,n-1}$. Then for any $\theta > \sigma$, we have $\lceil \theta d \rceil > \lceil \sigma d \rceil$ for all large enough $d \in \mathscr{D}$, so that

$$v_i(L^{(d)}) = 0 < e^{-d} \tag{3.4}$$

for all $i \geq \lceil \theta d \rceil$ and $d$ sufficiently large. By definition, $\theta_\star = \sigma$ is an upper decay threshold for the ensemble of subspaces $\{L^{(d)} : d \in \mathscr{D}\}$. The demonstration that $\kappa_\star = \sigma$ is a lower decay threshold for $\{L^{(d)} : d \in \mathscr{D}\}$ follows in the same way. $\qquad\square$

**Proposition 3.12** (Upper decay threshold for the nonnegative orthant)**.** *Let $\mathscr{D} \subset \mathbb{N}$ be an infinite index set. The value $\theta_\star = \frac{1}{2}$ is an upper decay threshold for the ensemble $\{\mathbb{R}_+^d : d \in \mathscr{D}\}$ of nonnegative orthants, and $\kappa_\star = \frac{1}{2}$ is a lower decay threshold for $\{\mathbb{R}_+^d : d \in \mathscr{D}\}$.*

*Proof.* It is well known (see, e.g., [**25,** Eq. (3.4)]) that, for any $\theta \in [0, 1]$, we have

$$\frac{1}{d}\log\binom{d}{\lceil \theta d \rceil} \to H(\theta) \text{ uniformly in } \theta \text{ as } d \to \infty, \tag{3.5}$$

17

**Table 1** *Decay of intrinsic volumes.* The decay thresholds developed in this work appear below. The computations for the orthant and subspace appear in Section 3.3.1, while the computations for the feasible cone of the Schatten 1-nrom and spectral norm appear in Section 5.2.1. Other computations appear where indicated.

| Cone ensemble | Level zero | | Level $\psi$ |
|---|---|---|---|
| | Upper threshold $\theta_\star$ | Lower threshold $\kappa_\star$ | $\theta_\star(\psi)$ |
| Subspaces of dimension $\lceil \sigma d \rceil$ | $\sigma$ | $\sigma$ | $\sigma$ |
| Orthants $\mathbb{R}^d_+$ | $\frac{1}{2}$ | $\frac{1}{2}$ | (See Prop. 4.9) |
| Feasible cones of $\ell_1$ norm at $\tau$-sparse vectors | (See Prop. C.3 in Appendix C) | | |
| Feasible cones of the Schatten 1-norm at $n \times n$ rank $\lceil \rho n \rceil$ matrices | $6\rho - 3\rho^2$ (See also Rem. 5.5.) | | |
| Feasible cones of the spectral norm at an orthogonal matrices | $\frac{3}{4}$ | | |

where

$$H(\theta) := -\theta \log(\theta) - (1-\theta)\log(1-\theta) \tag{3.6}$$

is the natural entropy; be aware that the logarithms are base-e rather than the customary base-2 used in information theory. Basic calculus shows that $H(\theta)$ achieves its unique maximum at $\theta_\star = \frac{1}{2}$, where it has maximum value $H\left(\frac{1}{2}\right) = \log(2)$.

Let $\theta > \frac{1}{2}$. By continuity of $H$, there is an $\varepsilon > 0$ such that $H(\tilde{\theta}) < \log(2) - \varepsilon$ for all $\tilde{\theta} \geq \theta > \theta_\star$. Continuity of the exponential and the uniform convergence in (3.5) together imply that, for all sufficiently large $d \in \mathscr{D}$, and any $i \geq \lceil \theta d \rceil$, we have

$$2^{-d}\binom{d}{i} \leq \exp\left(-d\log(2) + d(\log(2) - \varepsilon)\right) = e^{-\varepsilon d}.$$

The left-hand side is equal to the spherical intrinsic volume $v_i(\mathbb{R}^d_+)$ by Proposition 3.3, so we see that $\theta_\star = \frac{1}{2}$ is an upper decay threshold for $\{\mathbb{R}^d_+ : d \in \mathscr{D}\}$.

The proof that $\kappa_\star = \frac{1}{2}$ is a lower decay threshold for $\{\mathbb{R}^d_+ : d \in \mathscr{D}\}$ follows along similar lines. Briefly, for any $\kappa < \frac{1}{2}$, there is an $\varepsilon > 0$ such that $H(\tilde{\kappa}) < \log(2) - \varepsilon$ for every $\tilde{\kappa} \leq \kappa$. For the same reasons as before, when $d$ is large enough, we have

$$2^{-d}\binom{d}{i} \leq \exp\left(-d\log(2) + d(\log(2) - \varepsilon)\right) = e^{-\varepsilon d}$$

for all $i \leq \lceil \kappa d \rceil$. We conclude that $\kappa_\star = \frac{1}{2}$ is a lower decay threshold for the ensemble of nonnegative orthants $\{\mathbb{R}^d_+ : d \in \mathscr{D}\}$. □

Figure 8 illustrates the computation above. We discuss other approaches for finding decay thresholds in Section 5. Table 1 summarizes the decay thresholds determined in this work.

# 4  Success and failure

This section synthesizes the material from Sections 2 and 3 to determine whether the convex demixing method (1.2) succeeds, or fails, with high probability. Section 4.1 introduces the concept of a demixing

ensemble. Our main results arrive in Section 4.2, where we find that the success and failure of the convex demixing method (1.2) are characterized by decay thresholds. Section 4.3 extends our methods to achieve uniform guarantees on the success of method (1.2).

## 4.1 Ensembles of demixing problems

A demixing ensemble is a collection of demixing problems that is indexed by the ambient dimension of the observation. We explain this idea in the context of MCA, and we develop the abstract definition in Section 4.1.2.

### 4.1.1 Example: The MCA demixing ensemble

Recall from Section 1.1 that MCA seeks to demix a superposition of two sparse vectors. Let us fix sparsity levels $\tau_x$ and $\tau_y$ in $[0,1]$. For each pair $(\tau_x, \tau_y)$, we construct an ensemble of demixing problems with one problem per dimension. For each $d \in \mathbb{N}$, let $\boldsymbol{x}_0^{(d)}$ and $\boldsymbol{y}_0^{(d)}$ be vectors in $\mathbb{R}^d$ with

$$\mathrm{nnz}(\boldsymbol{x}_0^{(d)}) = \lceil \tau_x d \rceil \quad \text{and} \quad \mathrm{nnz}(\boldsymbol{y}_0^{(d)}) = \lceil \tau_y d \rceil.$$

In other words, the sparsity of each vector is proportional to the ambient dimension. Draw a random basis $\boldsymbol{Q}^{(d)}$ from $\mathsf{O}_d$. We observe the vector $\boldsymbol{z}_0^{(d)} = \boldsymbol{x}_0^{(d)} + \boldsymbol{Q}^{(d)} \boldsymbol{y}_0^{(d)}$. To set up a convex demixing method, we need to introduce appropriate complexity measures for $\boldsymbol{x}_0^{(d)}$ and $\boldsymbol{y}_0^{(d)}$. For both vectors, the $\ell_1$ norm on $\mathbb{R}^d$ is the natural choice for inducing sparsity. Assume we have access to the side information $\alpha^{(d)} = \|\boldsymbol{y}_0^{(d)}\|_{\ell_1}$.

Together, these data define a demixing ensemble for MCA. We want to study when the MCA problem (1.2) succeeds with high probability for all members of the ensemble with $d$ sufficiently large.

### 4.1.2 Abstract demixing ensembles

It is straightforward to extend this idea to other demixing problems. Let $\mathscr{D} \subset \mathbb{N}$ be an infinite set of indices. A *demixing ensemble* consists of one problem per index. For each $d \in \mathscr{D}$, the data are

- Vectors $\boldsymbol{x}_0^{(d)}, \boldsymbol{y}_0^{(d)} \in \mathbb{R}^d$,
- A random basis $\boldsymbol{Q}^{(d)} \in \mathsf{O}_d$ that is statistically independent of the other ensemble data,
- The observation $\boldsymbol{z}_0^{(d)} = \boldsymbol{x}_0^{(d)} + \boldsymbol{Q}^{(d)} \boldsymbol{y}_0^{(d)} \in \mathbb{R}^d$,
- Complexity measures $f^{(d)}$ and $g^{(d)}$ defined on $\mathbb{R}^d$, and
- The side information $\alpha^{(d)} = g^{(d)}(\boldsymbol{y}_0^{(d)})$.

Given such a demixing ensemble, we seek to determine conditions for which the convex demixing method

$$\begin{aligned} \text{minimize} \quad & f^{(d)}(\boldsymbol{x}) \\ \text{subject to} \quad & g^{(d)}(\boldsymbol{y}) \le \alpha^{(d)} \text{ and } \boldsymbol{x} + \boldsymbol{Q}^{(d)} \boldsymbol{y} = \boldsymbol{z}_0^{(d)}, \end{aligned} \tag{4.1}$$

succeeds with high probability when the dimension $d$ is large. When it does not cause confusion, we omit the superscript $d$.

Our goal in this paper is to describe regions where the demixing program (4.1) succeeds, or fails, with high probability as the dimension $d \to \infty$. In order to avoid cumbersome repetition in our theorems, we make a shorthand definition.

**Definition 4.1** (Overwhelming probability in high dimensions)**.** Given a demixing ensemble as in Section 4.1.2, we say that (4.1) *succeeds with overwhelming probability in high dimensions* if there exists an $\varepsilon > 0$ such that, for every sufficiently large dimension $d \in \mathscr{D}$, program (4.1) succeeds with probability at least $1 - \mathrm{e}^{-\varepsilon d}$ over the randomness in $\boldsymbol{Q}^{(d)}$. Similarly, we say that (4.1) *fails with overwhelming probability in high dimensions* if there exists an $\varepsilon > 0$ such that, for every sufficiently large dimension $d \in \mathscr{D}$, program (4.1) succeeds with probability at most $\mathrm{e}^{-\varepsilon d}$.

Following common practice in asymptotic analysis, the definition above masks the dependence between the probability decay rate $\varepsilon$ and the "sufficiently large" dimension $d_0$. This approach represents a tradeoff. We will find that it provides demixing guarantees in terms of only decay thresholds, but it does not provide explicit probabilistic bounds for finite $d$. Nevertheless, our asymptotic analysis proves quite accurate at predicting the behavior in our experiments. See Remark 4.6 for further discussion.

## 4.2 The main results

We are now in a position to state our main results. The first result shows that the upper decay threshold provides guarantees for the success of demixing under the model of Section 4.1.

**Theorem 4.2** (Success of demixing). *Consider a demixing ensemble as in Section 4.1.2. Suppose the ensembles $\{\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) : d \in \mathscr{D}\}$ and $\{\mathscr{F}(g^{(d)}, \boldsymbol{y}_0^{(d)}) : d \in \mathscr{D}\}$ of feasible cones have upper decay thresholds $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{y}}$. If*

$$\theta_{\boldsymbol{x}} + \theta_{\boldsymbol{y}} < 1,$$

*then program (4.1) succeeds with overwhelming probability in high dimensions.*

The proof of this result appears in Section 4.2.3. This next result shows that the failure of demixing is characterized by the lower decay threshold.

**Theorem 4.3** (Failure of demixing). *Consider a demixing ensemble as in Section 4.1.2. Suppose the ensembles $\{\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) : d \in \mathscr{D}\}$ and $\{\mathscr{F}(g^{(d)}, \boldsymbol{y}_0^{(d)}) : d \in \mathscr{D}\}$ of feasible cones have lower decay thresholds $\kappa_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}}$. If*

$$\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}} > 1,$$

*then program (4.1) fails with overwhelming probability in high dimensions.*

The proof of this second result is similar in spirit to the proof of Theorem 4.2, but it requires additional technical finesse; we defer the details to Appendix B.1.

### 4.2.1 Consequences for the choice of complexity functions

The complementary nature of Theorems 4.2 and 4.3 has striking implications. The success, or failure, of the convex demixing method (4.1) in high dimensions depends only on the sum of the decay thresholds. As a consequence, the upper and lower decay thresholds assess the quality of the complexity measures $f^{(d)}$ and $g^{(d)}$ in high dimensions.

Since the decay thresholds are independent of any interrelationship between the structured vectors $\boldsymbol{x}_0^{(d)}$ and $\boldsymbol{y}_0^{(d)}$ in the superimposed observation $\boldsymbol{z}_0^{(d)}$, the quality of the complexity measure $f^{(d)}$ is independent of the choice $g^{(d)}$. This explains, for instance, the ubiquity of the use of the $\ell_1$ norm for inducing sparsity and the Schatten 1-norm as a complexity measure for rank. Simply put, when a complexity measure is good for one incoherent demixing problem, it is good for another.

### 4.2.2 Sharp phase transitions

Theorems 4.2 and 4.3 have an important consequence for phase transitions in the convex demixing method (4.1). Equality between the lower and upper decay thresholds holds in cases where we have access to exact formulas for the spherical intrinsic volumes. Proposition 3.11 shows that the upper and lower decay thresholds are equal for subspaces whose dimension is proportional to the ambient space, and Proposition 3.12 implies that the upper decay threshold is equal to the lower decay threshold for the ensemble of nonnegative orthants. Moreover, our computations in Appendix C suggest that equality between the upper and lower decay thresholds also holds for the ensemble of feasible cones of the $\ell_1$ norm at vectors with a fixed proportion of nonzero elements.

The equality of the upper and lower decay thresholds explains the close agreement between our theoretical bounds and the empirical experiments. For instance, consider Figure 1. For sparsity levels below the green curve, the sum of the upper decay thresholds is less than one, so Theorem 4.2 implies that demixing succeeds with overwhelming probability in high dimensions. On the other hand, with sparsity levels above the green curve, the sum of the lower decay thresholds exceeds one, so demixing fails with overwhelming probability in high dimensions by Theorem 4.3. (See Section 6.1 for the details of this calculation.)

One may wonder whether the transition between success and failure is sharp in general. In work undertaken after the submission of this article, with collaborators, we have determined that the answer to this question is *yes* for a large class of demixing ensembles. In essence, the result [**5,** Thm. 6.1] indicates that the upper- and lower-decay thresholds for a sufficiently regular ensemble $\{K^{(d)} : d \in \mathcal{D}\}$ are equal. For example, the fact that $d^{-1} W(K^{(d)} \cap \mathsf{S}^{d-1})^2 \to \rho \in (0,1)$ as $d \to \infty$ is sufficient to guarantee that the upper- and lower-decay thresholds $\theta_\star$ and $\kappa_\star$ of the ensemble $\{K^{(d)} : d \in \mathcal{D}\}$ satisfy $\theta_\star = \kappa_\star = \rho$. (The function $W$ is the Gaussian width defined in Section 5.2.1.) We refer the reader to this newer work for details.

### 4.2.3 Proof of Theorem 4.2

The proof of Theorem 4.2 follows readily from a geometric statement concerning the probability that a random cone strikes a fixed cone as the dimension $d$ becomes large.

**Theorem 4.4.** *Suppose $\mathcal{D}$ is an infinite set of indices, and let $\{K^{(d)} \subset \mathbb{R}^d : d \in \mathcal{D}\}$ and $\{\tilde{K}^{(d)} \subset \mathbb{R}^d : d \in \mathcal{D}\}$ be two ensembles of closed convex cones with upper decay thresholds $\theta_\star$ and $\tilde{\theta}_\star$. If $\theta_\star + \tilde{\theta}_\star < 1$, then there exists an $\varepsilon > 0$ such that, for all sufficiently large $d$, we have $\mathbb{P}\{K^{(d)} \cap Q\tilde{K}^{(d)} \neq \{\mathbf{0}\}\} \le \mathrm{e}^{-\varepsilon d}$.*

Before proceeding to the proof, let us explain how Theorem 4.2 follows from Theorem 4.4 and the geometric optimality conditions of Lemma 2.4.

*Proof of Theorem 4.2 from Theorem 4.4.* By the assumptions in Theorem 4.2, the ensembles $\{\overline{\mathscr{F}}(f^{(d)}, \mathbf{x}_0^{(d)}) : d \in \mathcal{D}\}$ and $\{-\overline{\mathscr{F}}(g^{(d)}, \mathbf{y}_0^{(d)}) : d \in \mathcal{D}\}$ of closed cones satisfy the hypothesis of Theorem 4.4. Thus, there is an $\varepsilon > 0$ for which

$$\overline{\mathscr{F}}(f^{(d)}, \mathbf{x}_0^{(d)}) \bigcap \left(-Q\overline{\mathscr{F}}(g^{(d)}, \mathbf{y}_0^{(d)})\right) = \{\mathbf{0}\}$$

except with probability $\mathrm{e}^{-\varepsilon d}$, for all sufficiently large $d$.

Since cones are contained in their closure, the two feasible cones have a trivial intersection at least as frequently as their closures (but see the remark below). Applying our geometric optimality condition, Lemma 2.4, immediately implies that (4.1) succeeds with probability at least $1 - \mathrm{e}^{-\varepsilon d}$ for every sufficiently large $d$. □

**Remark 4.5.** In fact, the probability that randomly oriented convex cones strike is *equal* to the probability that their closures strike. This seemingly innocuous claim appears to have no simple proof from first principles. However, this fact readily follows from the discussion of touching probabilities in [**69,** pp. 258–259].

Figure 9 illustrates the main idea behind the following proof.

*Proof of Theorem 4.4.* The spherical kinematic formula (3.2) only applies when at least one cone is not a subspace, so we split the demonstration into two cases. First, we assume that at least one of the ensembles $\{K^{(d)}\}$ or $\{\tilde{K}^{(d)}\}$ of cones does not contain any subspace. Then, we consider the case where both ensembles of cones contain only subspaces. The argument readily extends to the general case by considering subsequences where one of the two cases above holds. Throughout the proof, we drop the explicit dependence of $K^{(d)}$ and $\tilde{K}^{(d)}$ on the dimension $d$ for notational clarity.
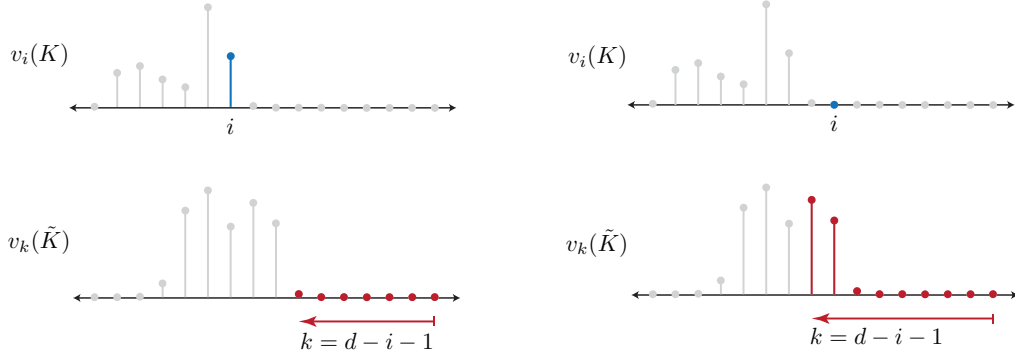
**Figure 9** *Main idea behind the proof of Theorem 4.4.* When the spherical intrinsic volumes are very small for large $i$, the product of the outer (top) and inner (bottom) terms in equation (4.2) is always small. In the left panel, the large value of $v_i(K)$ is offset by the small inner sum $\sum_{k=d-i-1} v_k(\tilde{K})$. On the right, the small value of $v_i(K)$ counteracts the large sum $\sum_{k=d-i-1} v_k(\tilde{K})$. This situation always occurs when the upper decay thresholds satisfy $\theta_\star + \tilde{\theta}_\star < 1$ and the ambient dimension $d$ is large.

We start with the first case: assume that either $K$ or $\tilde{K}$ is not a subspace. Our main tool is the spherical kinematic formula, Fact 3.6. The positivity of the spherical intrinsic volumes and the bound $(1 + (-1)^k) \leq 2$ imply that the probability $P$ of interest satisfies

$$P := \mathbb{P}\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\}\} \leq 2 \sum_{k=0}^{d-1} \sum_{i=k}^{d-1} v_i(K) \cdot v_{d-1-i+k}(\tilde{K})$$

$$= 2 \sum_{i=0}^{d-1} v_i(K) \sum_{k=d-i-1}^{d-1} v_k(\tilde{K}). \tag{4.2}$$

The equality follows by a change in the order of summation and a change of the summation index. Since $\theta_\star + \tilde{\theta}_\star < 1$, there exist parameters $\theta > \theta_\star$ and $\tilde{\theta} > \tilde{\theta}_\star$ for which $\theta + \tilde{\theta} < 1$. We expand the right-hand sum of (4.2) into four terms, say $\Sigma_1, \Sigma_2, \Sigma_3$, and $\Sigma_4$:

$$\frac{1}{2}P \leq \underbrace{\sum_{i=0}^{\lceil \theta d \rceil} v_i(K) \sum_{k=d-i-1}^{\lceil \tilde{\theta} d \rceil} v_k(\tilde{K})}_{=:\Sigma_1} + \underbrace{\sum_{i=\lceil \theta d \rceil + 1}^{d-1} v_i(K) \sum_{k=d-i-1}^{\lceil \tilde{\theta} d \rceil} v_k(\tilde{K})}_{=:\Sigma_2}$$

$$+ \underbrace{\sum_{i=0}^{\lceil \theta d \rceil} v_i(K) \sum_{k=\lceil \tilde{\theta} d \rceil + 1}^{d-1} v_k(\tilde{K})}_{=:\Sigma_3} + \underbrace{\sum_{i=\lceil \theta d \rceil + 1}^{d-1} v_i(K) \sum_{k=\lceil \tilde{\theta} d \rceil + 1}^{d-1} v_k(\tilde{K})}_{=:\Sigma_4}. \tag{4.3}$$

We bound each summand separately. First, the fact $\theta + \tilde{\theta} < 1$ implies that, for all sufficiently large $d$, we have $\lceil \theta d \rceil + \lceil \tilde{\theta} d \rceil < d - 1$. We conclude that $\Sigma_1 = 0$ when $d$ is large enough: The inner sum is empty.

To bound $\Sigma_2$, apply Facts 3.5.1 and 3.5.2 to the inner sum to find

$$\Sigma_2 \leq \sum_{i=\lceil \theta d \rceil + 1}^{d-1} v_i(K) \leq (d-1)\mathrm{e}^{-\varepsilon' d},$$

where the second inequality holds for some $\varepsilon' > 0$ and all sufficiently large $d$ owing to the definition of the upper decay threshold $\theta_\star$. Through analogous reasoning, the definition of $\tilde{\theta}_\star$ gives the exponential bounds

$$\Sigma_3 \leq (d-1)\mathrm{e}^{-\varepsilon'' d} \quad \text{and} \quad \Sigma_4 \leq (d-1)\mathrm{e}^{-\varepsilon''' d},$$

22

for some $\varepsilon'', \varepsilon''' > 0$ and all sufficiently large $d$. Taking $\varepsilon$ sufficiently small (say $\varepsilon = \frac{1}{2} \min\{\varepsilon', \varepsilon'', \varepsilon'''\}$) and $d$ sufficiently large gives the result for the first case.

For the second case, suppose that both $K$ and $\tilde{K}$ are subspaces. Set $n := \dim(K)$ and $\tilde{n} := \dim(\tilde{K})$. Choose parameters $\theta > \theta_\star$ and $\tilde{\theta} > \tilde{\theta}_\star$ such that $\theta + \tilde{\theta} < 1$. The definition of the upper decay threshold and Proposition 3.2 imply that

$$v_i(K) = \delta_{i,n-1} \le e^{-\varepsilon' d}$$

for some $\varepsilon' > 0$, all sufficiently large $d$, and every $i \ge \lceil \theta d \rceil$. In particular, this inequality implies $n - 1 < \lceil \theta d \rceil$ for all $d$ large enough. Similarly, we find $\tilde{n} - 1 < \lceil \tilde{\theta} d \rceil$ for all $d$ sufficiently large.

Since $\theta + \tilde{\theta} < 1$, we have

$$n + \tilde{n} \le \lceil \theta d \rceil + \lceil \tilde{\theta} d \rceil + 2 < d$$

whenever $d$ is large enough. That is, the sum of the dimensions of the subspaces is less than the ambient dimension. Since two randomly oriented subspaces are almost surely in general position, we see that $K \cap Q\tilde{K} = \{\mathbf{0}\}$ with probability one, whenever $d$ is large enough. This completes the second case, so we are done. $\qquad\square$

**Remark 4.6** (Explicit dimensional dependence)**.** The methods above can provide explicit dependence between the sum of the decay thresholds $\theta_x + \theta_y$, the "sufficiently large" dimension and the probability decay rate $\varepsilon$ when detailed information about the intrinsic volumes is available. In the case of the orthant, for example, Stirling's formula with remainder [**59,** Sec. 5.6.1] may provide enough information. For the descent cones of the $\ell_1$ norm, it may be possible to achieve such explicit dependence using the approach of [**32**].

## 4.3 Uniform demixing guarantees

Suppose that $Q$ is drawn at random and fixed. For example, in the MCA problem from Section 1.1, we may observe a number of different images, but the structures that we expect to find (encoded by $Q$) are the same in each image. In this case, we ask whether (1.1) demix *every* sparse pair $(x_0, y_0)$ from the associated observation $z_0 = x_0 + Q y_0$, and thus successfully identify the structures in a large family of images.

More generally, we can study the probability that the generic demixing program (1.2) will demix every structured pair $(x_0, y_0)$ with a random—but fixed—basis $Q$. (These uniform guarantees go by the name of *strong bounds* [**28**], [**25**], [**29**], [**31**].) By the geometric optimality condition of Lemma 2.4, this probability is equal to the probability of the event

$$\mathscr{F}(f, x_0) \bigcap \left(-Q \mathscr{F}(g, y_0)\right) = \{\mathbf{0}\} \text{ for } every \text{ structured pair } (x_0, y_0). \tag{4.4}$$

In this section, we control the probability of (4.4) by coupling the argument leading to Theorem 4.2 with a union bound. This approach does not necessarily limit our methods to a finite number of structured pairs $(x_0, y_0)$, however. In the case of a sparse vectors, for example, the set of feasible cones

$$\left\{ \mathscr{F}(\|\cdot\|_{\ell_1}, x) : x \in \mathbb{R}^d, \text{ nnz}(x) = k \right\} \tag{4.5}$$

consists of $\binom{d}{k} 2^k$ cones because the feasible cone $\mathscr{F}(\|\cdot\|_{\ell_1}, x)$ depends only on the sparsity and sign pattern, and not the magnitude, of the elements of $x$. (See Section 6.1.1 for more details.) Thus, even the simple union bound can offer insight into the behavior of constrained MCA (1.1) for an *infinite* family of images.

Applying a union bound to a finite, but rather large, set of cones such as (4.5) requires stronger probabilistic information than provided by the decay thresholds. Therefore, we define an extension of the upper decay threshold that provides detailed information on the rate of decay of spherical intrinsic volumes.

**Definition 4.7** (Decay at level $\psi$)**.** Let $\mathscr{D} \subset \mathbb{N}$ be an infinite set of indices, let $\{K^{(d)} : d \in \mathscr{D}\}$ be an ensemble of closed convex cones with $K^{(d)} \in \mathbb{R}^d$, and suppose $\psi \ge 0$. We say that $\theta_\star$ is an upper decay threshold *at*

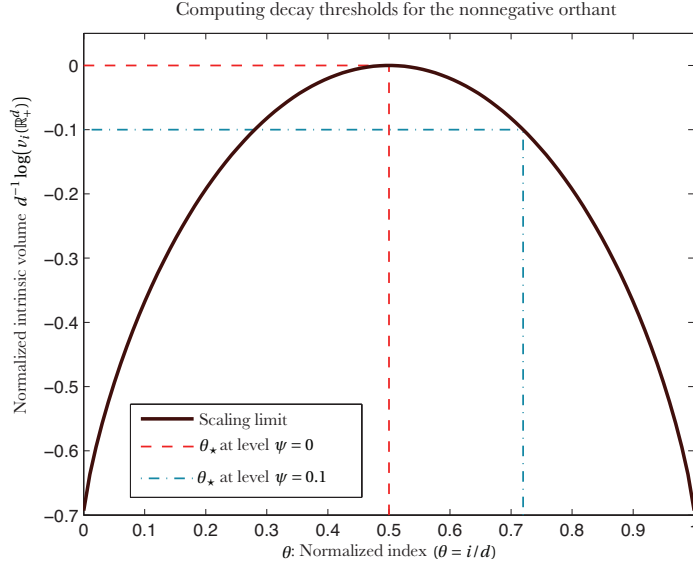Computing decay thresholds for the nonnegative orthant

**Figure 10** *Computation of decay thresholds for the orthant.* The solid curve is the limiting behavior of the spherical intrinsic volumes of the orthant $\mathbb{R}_+^d$ from Figure 8. The rightmost point where the horizontal line at level $-\psi$ crosses this curve is an upper decay threshold at level $\psi$. From the diagram, we see that $\theta_\star = \frac{1}{2}$ is an upper decay threshold at level zero, while $\theta_\star \approx 0.72$ is an upper decay threshold at level $\psi = 0.1$.

*level* $\psi$ for the ensemble $\{K^{(d)}\}$ if, for every $\theta > \theta_\star$, there exists an $\varepsilon > 0$ such that, for all sufficiently large $d$, the inequality

$$v_i(K^{(d)}) \le e^{-d(\psi+\varepsilon)}$$

holds for all $i \ge \lceil \theta d \rceil$. When no level is specified, we take $\psi = 0$ for compatibility with Definition 3.9. As in Definition 3.9, this definition extends to non-closed cones by taking the closure.

Subspaces and orthants again provide useful examples.

**Proposition 4.8** (Decay at level $\psi$ for an ensemble of subspaces). *Let $\{L^{(d)}\}$ be an infinite ensemble of linear subspaces with $L^{(d)} \subset \mathbb{R}^d$ and $\dim(L^{(d)}) = \lceil \sigma d \rceil$, and suppose $\psi \ge 0$. Then $\theta_\star = \sigma$ is an upper decay threshold for $\{L^{(d)}\}$ at level $\psi$.*

*Proof.* The proof is substantially similar to the proof of Proposition 3.11. Let $n = \dim(L^{(d)})$, so that $v_i(L^{(d)}) = \delta_{i,n-1}$ by Proposition 3.2. For any $\theta > \sigma$, we have $\lceil \theta d \rceil > \lceil \sigma d \rceil$ for all large enough $d$. Therefore,

$$v_i(L^{(d)}) = 0 < e^{-(\psi+1)d}$$

for all $i \ge \lceil \theta d \rceil$ and $d$ sufficiently large. By definition, $\theta_\star = \sigma$ is an upper decay threshold at level $\psi$ for $\{L^{(d)}\}$. $\square$

The computation for the orthant is only slightly more involved. The following result is illustrated in Figure 10.

**Proposition 4.9** (Decay at level $\psi$ for $\{\mathbb{R}_+^d\}$). *Suppose $0 \le \psi \le \log(2)$, and let $\mathscr{D} \subset \mathbb{N}$ be an infinite set of indices. Define*

$$\theta_{\mathbb{R}_+^d}(\psi) := \sup\{\theta : H(\theta) \ge \log(2) - \psi\}, \tag{4.6}$$

*where $H(\theta)$ is the entropy defined in (3.6). Then $\theta_{\mathbb{R}_+^d}(\psi)$ is an upper decay threshold at level $\psi$ for the ensemble $\{\mathbb{R}_+^d\}$ of orthants.*

*Proof sketch.* Let $\theta > \theta_{\mathbb{R}_+^d}(\psi)$. As in the proof of Proposition 3.12, there exists an $\varepsilon > 0$, such that, for all $d$ sufficiently large, we have

$$v_i(\mathbb{R}_+^d) \leq \exp\Big(d\big(H(\theta_\psi) - \log(2) + \varepsilon\big)\Big) = \exp\big(-d(\psi + \varepsilon)\big)$$

for every $i \geq \lceil \theta d \rceil$. By definition, $\theta_{\mathbb{R}_+^d}(\psi)$ is an upper decay threshold at level $\psi$ for the ensemble $\{\mathbb{R}_+^d\}$. $\quad\square$

We now extend the decay threshold to cover the family of cones considered in condition (4.4). In our applications, the feasible cones appearing in (4.4) are congruent, so we restrict our attention to this case.

**Definition 4.10** (Decay threshold for an ensemble of sets of cones). For an infinite set of indices $\mathscr{D} \subset \mathbb{N}$, let $\{\mathscr{K}^{(d)} : d \in \mathscr{D}\}$ be an ensemble of *sets* of congruent cones, indexed by the ambient dimension, and let $\{K^{(d)}\}$ be an ensemble of exemplars, that is, $K^{(d)} \in \mathscr{K}^{(d)}$ for every $d \in \mathscr{D}$. We say that $\{\mathscr{K}^{(d)}\}$ has upper decay threshold $\theta$ at level $\psi$ if the sequence of exemplars $\{K^{(d)}\}$ has upper decay threshold $\theta$ at level $\psi$.

By Fact 3.5.3, the decay threshold for an ensemble $\{\mathscr{K}^{(d)}\}$ of sets of congruent cones is independent of the choice of exemplars, so this nomenclature is well defined. We now state an analog to Theorem 4.4 that bounds the probability that a number of cones strike, provided there are not too many cones.

**Theorem 4.11.** *Let $\{\mathscr{K}^{(d)}\}$ and $\{\tilde{\mathscr{K}}^{(d)}\}$ be two ensembles of sets of congruent closed convex cones, indexed by ambient dimension $d$. Suppose the cardinality of $\mathscr{K}^{(d)}$ and $\tilde{\mathscr{K}}^{(d)}$ grows no faster than exponentially: there exist $\psi, \tilde{\psi}$ such that, for every $\eta > 0$ and all $d$ sufficiently large, we have the inequalities $|\mathscr{K}^{(d)}| \leq e^{d(\psi+\eta)}$, and $|\tilde{\mathscr{K}}^{(d)}| \leq e^{d(\tilde{\psi}+\eta)}$. Suppose further $\{\mathscr{K}^{(d)}\}$ and $\{\tilde{\mathscr{K}}^{(d)}\}$ have respective upper decay thresholds $\theta_\star$ and $\tilde{\theta}_\star$, each at level $\psi + \tilde{\psi}$.*

*If $\theta_\star + \tilde{\theta}_\star < 1$, then there exists an $\varepsilon > 0$ such that for every sufficiently large $d$,*

$$\mathbb{P}\left\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\} \text{ for any } K \in \mathscr{K}^{(d)}, \ \tilde{K} \in \tilde{\mathscr{K}}^{(d)}\right\} \leq e^{-\varepsilon d},$$

*where the probability is taken over the random basis $\boldsymbol{Q} \in \mathsf{O}_d$.*

The proof of Theorem 4.11 simply couples a union bound to the proof of Theorem 4.4, so we defer the demonstration to Appendix B.2. Note that the statement of Theorem 4.11 is equivalent to that of Theorem 4.4 when $\mathscr{K}^{(d)}$ and $\tilde{\mathscr{K}}^{(d)}$ are singletons, because we may take $\psi = \tilde{\psi} = 0$ in this case.

Theorem 4.11 can be used to verify that event (4.4) holds with high probability when the dimension $d$ becomes large. In Section 6.1.1, we use this approach to verify that the MCA formulation (1.1) can demix all sufficiently sparse vectors, and in Section 6.2.2, we use Theorem 4.11 to show that the channel coding method (1.4) is robust to adversarial sparse corruptions.

# 5 Computing decay thresholds

Section 4.2 demonstrates that the decay thresholds provide a simple way to analyze demixing under the random basis model. This section describes several methods for computing decay thresholds. We begin by considering direct approaches, where precise formulas for the spherical intrinsic volumes give correspondingly precise thresholds.

The direct method is powerful, but its application is limited to regimes where we have access to formulas for spherical intrinsic volumes. In Section 5.2, we observe that known results on linear inverse problems [15] imply bounds on decay thresholds. This observation allows us to study upper decay thresholds for several structural classes, including low-rank matrices.

## 5.1 Direct approach

There are several situations where the direct approach for calculating decay thresholds is feasible. Propositions 3.11 and 3.12 compute upper and lower decay thresholds for ensembles of subspaces and orthants directly from the definition of spherical intrinsic volumes. In Appendix C, we use the asymptotic polytope angle computations of [25] to compute the decay threshold for ensembles of feasible cones of the $\ell_1$ norm at sparse vectors. The approach follows roughly the same lines as Propositions 3.12 and 4.9, but the argument requires a good deal of background information that is tangential to this work.

## 5.2 Relationship to linear inverse problems

There is a useful link between the number of random linear measurements required to identify a structured signal with a convex complexity measure and the upper decay threshold of the associated feasible cone. Roughly speaking, the upper decay threshold is the ratio between the number of linear measurements required to identify a structured signal and the ambient dimension. This observation provides a powerful method for determining decay thresholds.

Linear inverse problems are closely related to demixing problems. Suppose we observe the linear image $z_0 = Ax_0$, where $A$ is a known matrix and $x_0$ is a structured vector. Given an associated convex complexity measure $f$, Chandrasekaran et al. [15] study the convex optimization program

$$
\begin{aligned}
\text{minimize} \quad & f(x) \\
\text{subject to} \quad & Ax = z_0.
\end{aligned}
\tag{5.1}
$$

These authors consider the question "Given the data $z_0 = Ax_0$, when is $x_0$ the unique optimal point of (5.1)?" The answer to this question is closely related to our demixing problem.

To place the linear inverse problem in our asymptotic framework, we consider an ensemble of problems indexed by the ambient dimension $d$. Fix an undersampling parameter $\sigma \in [0,1]$. For each $d$ in some infinite set $\mathscr{D} \subset \mathbb{N}$ of indices, assume we are given a structured vector $x_0^{(d)} \in \mathbb{R}^d$, a Gaussian measurement matrix $\Omega^{(d)} \in \mathbb{R}^{\lceil \sigma d \rceil \times d}$, the observation $z_0^{(d)} = \Omega^{(d)} x_0^{(d)} \in \mathbb{R}^d$, and a complexity measure $f^{(d)}$ associated with the structure of $x_0^{(d)}$. We attempt to identify $x_0^{(d)}$ by solving the optimization problem

$$
\begin{aligned}
\text{minimize} \quad & f^{(d)}(x) \\
\text{subject to} \quad & \Omega^{(d)} x = z_0^{(d)}
\end{aligned}
\tag{5.2}
$$

with decision variable $x \in \mathbb{R}^d$. This method succeeds when $x_0^{(d)}$ is the unique optimal point of (5.2). The following result shows that the problem of computing the number of random linear measurements needed to identify a structured vector is equivalent to determining an upper decay threshold.

**Lemma 5.1.** *Consider the ensemble described above.*

1. *Suppose the ensemble $\{\mathscr{F}(f^{(d)}, x_0^{(d)})\}$ of feasible cones has an upper decay threshold $\theta_\star < \sigma$. Then (5.2) succeeds with overwhelming probability in high dimensions.*

2. *On the other hand, suppose the linear inverse program (5.2) succeeds with overwhelming probability in high dimensions. Then the ensemble $\{\mathscr{F}(f^{(d)}, x_0^{(d)})\}$ of feasible cones has an upper decay threshold $\theta_\star = \sigma$.*

The proof of Lemma 5.1 appears in Appendix D, but we sketch the main ideas here. The program (5.1) identifies $x_0$ precisely when the null space of $A$ intersects the feasible cone $\mathscr{F}(f, x_0)$ trivially [15, Prop. 2.1]. When $A = \Omega$ is a Gaussian matrix, its null space is a randomly oriented subspace. Combining this fact with the kinematic formula (3.2) lets us compute the decay threshold using the number of observations required to recover a vector with (5.1) under a Gaussian measurement model.

In Appendix C.3, we describe how our computation of the upper decay threshold for the feasible cone of the $\ell_1$ norm at sparse vectors relates to the recovery guarantees for basis pursuit explored in the series of papers [**23**], [**25**], [**29**], [**30**], [**33**]. Our approach also yields a sharp transition between success and failure regimes for basis pursuit. See Appendix C.3.2 for the details.

**Remark 5.2.** A counterpart to Lemma 5.1 that links the failure of linear inverse problems to the lower decay threshold $\kappa_\star$ is readily derivable with the techniques used in this work. This may enable the computation of lower decay thresholds through information-theoretic arguments [**81**].

### 5.2.1 The upper decay threshold from the Gaussian width

Lemma 5.1 provides a powerful tool for computing upper decay thresholds. Define the *Gaussian width* of a cone $K \subset \mathbb{R}^d$ by the expression

$$W(K \cap \mathsf{S}^{d-1}) := \mathbb{E}\left[\sup_{\boldsymbol{x} \in K \cap \mathsf{S}^{d-1}} \langle \boldsymbol{\omega}, \boldsymbol{x} \rangle\right],$$

where the random vector $\boldsymbol{\omega}$ is drawn from the Gaussian distribution on $\mathbb{R}^d$. The following corollary lets us determine upper decay thresholds from Gaussian width bounds. As usual, $\mathscr{D}$ is an infinite subset of the natural numbers.

**Corollary 5.3.** *Consider an ensemble $\{\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) \subset \mathbb{R}^d : d \in \mathscr{D}\}$ of feasible cones. Suppose that*

$$\limsup_{d \to \infty} \frac{1}{d} W\big(\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) \cap \mathsf{S}^{d-1}\big)^2 \leq \theta_\star. \tag{5.3}$$

*Then $\theta_\star$ is an upper decay threshold (at level zero) for $\{\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)})\}$.*

The proof relies the result [**15,** Cor. 3.3(1)], which asserts that the linear inverse problem (5.2) succeeds with high probability when the number of measurements $n^{(d)} \gtrsim W\big(\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) \cap \mathsf{S}^{d-1}\big)^2$. Since $\lceil \theta_\star d \rceil$ exceeds the width for large $d$ by assumption (5.3), taking $n^{(d)} = \lceil \theta_\star d \rceil$ results in high-probability success in (5.2). Corollary 5.3 then follows from the second part of Lemma 5.1. The details appear at the end of Appendix D.

This machinery allows us to compute upper decay thresholds in situations involving matrix observations. For simplicity, we consider only the space $\mathbb{R}^{n \times n}$ of square $n \times n$ matrices, where $n \in \mathbb{N}$. The ambient dimension of this vector space is $d = n^2$, but we will index the observations by the parameter $n$. This poses no difficulty, because it is equivalent to indexing over the set $\mathscr{D} = \{1^2, 2^2, 3^2, \ldots\}$.

**Proposition 5.4.** *Fix $\rho \in [0, 1]$. For each $n \in \mathbb{N}$, let $\boldsymbol{X}_0^{(n)} \in \mathbb{R}^{n \times n}$ be a matrix with $\mathrm{rank}(\boldsymbol{X}_0^{(n)}) = \lceil \rho n \rceil$. Then the corresponding ensemble $\{\mathscr{F}(\|\cdot\|_{S_1}, \boldsymbol{X}_0^{(n)})\}$ of feasible cones has upper decay threshold $\theta_\star = 6\rho - 3\rho^2$.*

*Proof.* Let $r = r^{(n)} = \mathrm{rank}(\boldsymbol{X}_0^{(n)})$. From [**15,** Prop. 3.11], we have

$$W\left(\mathscr{F}(\|\cdot\|_{S_1}, \boldsymbol{X}_0^{(n)}) \cap \mathsf{S}^{n^2-1}\right)^2 \leq 6rn - 3r^2 = \big(6\rho - 3\rho^2\big)n^2. \tag{5.4}$$

Dividing both sides by the ambient dimension $n^2$ and taking limits, we see that the conditions of Corollary 5.3 hold. This gives the result. $\qquad\square$

**Remark 5.5.** An asymptotically sharp upper bound for the Gaussian width in (5.4) is given by the solution to an implicit equation in [**60,** Eq. (94)]. We use this asymptotically precise formula for computing the location of the green curve in Figure 4. See Section 6.3 for more details.

**Proposition 5.6.** *For each $n \in \mathbb{N}$, let $\boldsymbol{X}_0^{(n)}$ be an orthogonal matrix. Then the corresponding ensemble $\{\mathscr{F}(\|\cdot\|_{\mathrm{Op}}, \boldsymbol{X}_0^{(n)})\}$ of feasible cones has upper decay threshold $\theta_\star = \frac{3}{4}$.*
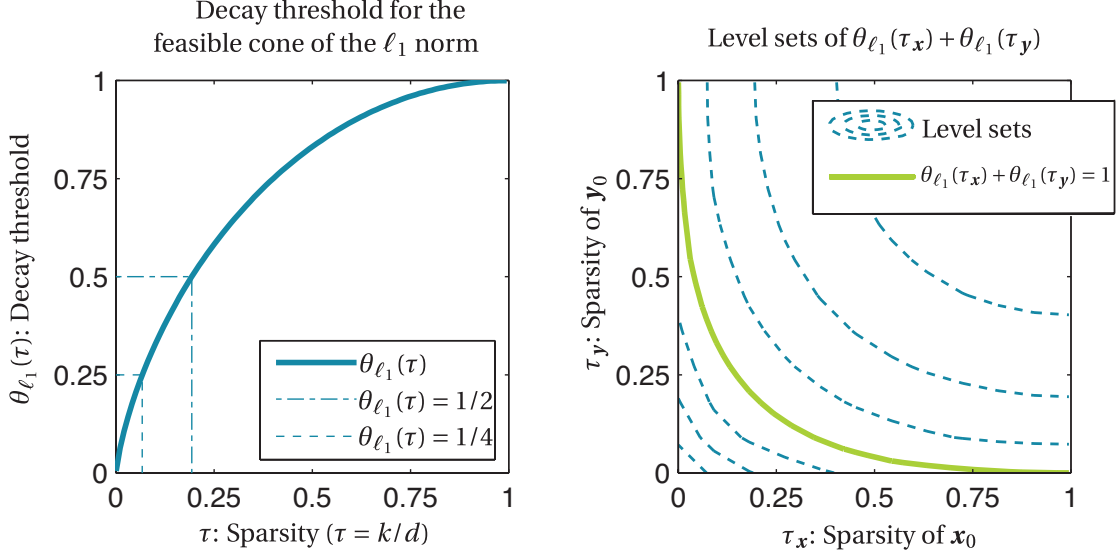
**Figure 11** *Decay threshold for the $\ell_1$ norm.* The left panel shows $\theta_{\ell_1}(\tau)$, the upper decay threshold for the sequence of feasible cones of the $\ell_1$ norm at $\lceil \tau d \rceil$-sparse vectors as a function of the sparsity $\tau$. The narrow lines show that $\theta_{\ell_1}(\tau) < \frac{1}{2}$ for $\tau < 0.19$, while $\theta_{\ell_1}(\tau) < \frac{1}{4}$ for $\tau < 0.06$. The right panel displays level sets of the function $\theta_{\ell_1}(\tau_{\boldsymbol{x}}) + \theta_{\ell_1}(\tau_{\boldsymbol{y}})$. The thick curve marks the level set $\theta_{\ell_1}(\tau_{\boldsymbol{x}}) + \theta_{\ell_1}(\tau_{\boldsymbol{y}}) = 1$; it corresponds to the green curve in Figure 1.

*Proof.* From [**15,** Prop. 3.13], we have

$$W\left(\mathcal{F}(\|\cdot\|_{\mathrm{Op}}, \boldsymbol{X}_0^{(n)}) \cap \mathsf{S}^{n^2-1}\right)^2 \leq \frac{3n^2 - n}{4} = \frac{3}{4}n^2 + O(n).$$

The result follows upon dividing by $n^2$, taking limits, and applying Corollary 5.3. $\qquad\square$

We list all of the bounds computed in our work in Table 1 on page 18, but note that several more bounds are readily derivable from the Gaussian width calculations in [**15,** Sec. 3.4].

# 6 Experiments

We can tackle a variety of scenarios using the theory developed in Section 4 and the decay threshold calculations of Section 5.

## 6.1 Morphological component analysis

We return to the MCA model of Section 1.1. Our goal is to analyze when we can demix two signals that are sparse in incoherent bases. To apply our theoretical results, we consider the demixing ensemble from Section 4.1.1.

Fix sparsity levels $\tau_{\boldsymbol{x}}$ and $\tau_{\boldsymbol{y}}$ in $[0, 1]$. For each dimension $d \in \mathbb{N}$, we construct signals $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ in $\mathbb{R}^d$ that satisfy

$$\mathrm{nnz}(\boldsymbol{x}_0) = \lceil \tau_{\boldsymbol{x}} d \rceil \quad \text{and} \quad \mathrm{nnz}(\boldsymbol{y}_0) = \lceil \tau_{\boldsymbol{y}} d \rceil.$$

Draw a random basis $\boldsymbol{Q}$, and suppose that we observe $\boldsymbol{z}_0 = \boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{y}_0$.

The $\ell_1$ norm is a natural complexity measure for sparse vectors. Given the side information $\alpha = \|\boldsymbol{y}_0\|_{\ell_1}$, we pose the constrained MCA problem

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{x}\|_{\ell_1} \\ \text{subject to} \quad & \|\boldsymbol{y}\|_{\ell_1} \le \alpha \ \text{ and } \ \boldsymbol{x} + \boldsymbol{Q}\boldsymbol{y} = \boldsymbol{z}_0. \end{aligned} \tag{6.1}$$

The theory of Section 4 describes when (6.1) identifies $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ with overwhelming probability in high dimensions in terms of decay thresholds for the ensembles $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0)\}$ and $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{y}_0)\}$ indexed by the dimension $d$.

Observe that, up to rotations, the geometry of the feasible cone $\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{w})$ depends only on the number of nonzero entries in $\boldsymbol{w}$, but not on the positions or magnitudes of the entries. For a fixed sparsity level $\tau \in [0,1]$, consider an ensemble $\{\boldsymbol{w}^{(d)}\}$ with $\boldsymbol{w}^{(d)} \in \mathbb{R}^d$ and $\mathrm{nnz}(\boldsymbol{w}^{(d)}) = \lceil \tau d \rceil$ for each $d \in \mathbb{N}$. In Appendix C, we compute the optimal upper decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{w}^{(d)})\}$. We denote this value by $\theta_{\ell_1}(\tau)$ (see Figure 11, left panel). Therefore, the ensembles $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0)\}$ and $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{y}_0)\}$ have decay thresholds $\theta_{\ell_1}(\tau_{\boldsymbol{x}})$ and $\theta_{\ell_1}(\tau_{\boldsymbol{y}})$.

Theorem 4.2 implies that our demixing method (6.1) for sparse vectors succeeds with overwhelming probability in high dimensions so long as $\theta_{\ell_1}(\tau_{\boldsymbol{x}}) + \theta_{\ell_1}(\tau_{\boldsymbol{y}}) < 1$. The right panel of Figure 11 shows the level sets of the function $\theta_{\ell_1}(\tau_{\boldsymbol{x}}) + \theta_{\ell_1}(\tau_{\boldsymbol{y}})$ for $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$ on the unit square $[0,1]^2$. The green curve is the level set

$$\{(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}}) : \theta_{\ell_1}(\tau_{\boldsymbol{x}}) + \theta_{\ell_1}(\tau_{\boldsymbol{y}}) = 1\}.$$

Theorem 4.2 implies that program (6.1) succeeds with overwhelming probability when the joint sparsity $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$ lies below the green curve.

On the other hand, our computations show that the upper decay threshold $\theta_{\ell_1}(\tau)$ is numerically equal to the lower decay threshold $\kappa_{\ell_1}(\tau)$—see the discussion in Appendix C.2. Theorem 4.3 implies that the sparse demixing method (6.1) fails with overwhelming probability for sparsity levels $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$ in the region above the green curve. In other words, the green curve on the right panel of Figure 11 delineates a sharp transition between success and failure for constrained MCA (6.1). The green curve in Figure 11(b) is the same as the green curve in Figure 1.

### 6.1.1 Strong guarantees

The theory of Section 4.3 allows us to provide a uniform recovery guarantee. For a fixed draw of the random basis, constrained MCA can demix all sufficiently sparse pairs of vectors with overwhelming probability in high dimensions.

Fix the sparsity $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$ and the ambient dimension $d$. Suppose $\boldsymbol{Q} \in \mathsf{O}_d$ is a random basis. By Lemma 2.4, constrained MCA can identify every $(\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})$-sparse pair $(\boldsymbol{x}_0, \boldsymbol{y}_0)$ given the observation $\boldsymbol{z}_0 = \boldsymbol{x}_0 + \boldsymbol{Q}\boldsymbol{y}_0$, provided that the event

$$\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0) \bigcap (-\boldsymbol{Q}\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{y}_0)) = \{\boldsymbol{0}\} \text{ for each } (\tau_{\boldsymbol{x}}, \tau_{\boldsymbol{y}})\text{-sparse pair } (\boldsymbol{x}_0, \boldsymbol{y}_0) \tag{6.2}$$

holds. Theorem 4.11 guarantees that the probability of event (6.2) is large when some associated decay thresholds are small enough; let us describe how to verify the required technical assumptions.

First, the results in Appendix C allow us to compute upper decay threshold at levels $\psi \ge 0$ for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{w}^{(d)})\}$ from Section 6.1 consisting of feasible cones for the $\ell_1$ norm at $\tau$-sparse vectors. We extend our earlier notation by writing this quantity as $\theta_{\ell_1}(\tau, \psi)$. This is the first element required to check the hypotheses of Theorem 4.11.

We also require information on the total number of feasible cones under consideration. Let

$$\mathscr{K}_{\boldsymbol{x}}^{(d)} = \bigcup \{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0)\}, \quad \text{and} \quad \mathscr{K}_{\boldsymbol{y}}^{(d)} = \bigcup \{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{y}_0)\},$$

where the unions take place over all $\lceil \tau_{\boldsymbol{x}} d \rceil$-sparse $\boldsymbol{x}_0 \in \mathbb{R}^d$ and all $\lceil \tau_{\boldsymbol{y}} d \rceil$-sparse $\boldsymbol{y}_0 \in \mathbb{R}^d$. There are exactly $2^k \binom{d}{k}$ different—but congruent—feasible cones of the $\ell_1$ norm at vectors in $\mathbb{R}^d$ with $k$ nonzero entries,

one for each sign/sparsity pattern. This corresponds to the number of $(k-1)$-dimensional faces of the crosspolytope; see, e.g., [**25,** Sec. 3.3]. With $k = \lceil \tau d \rceil$, it follows from the proof of Proposition 3.12 that

$$d^{-1} \log \left( 2^k \binom{d}{k} \right) \to \tau \log(2) + H(\tau) =: E(\tau), \tag{6.3}$$

uniformly as $d \to \infty$. By continuity of the exponential, for every $\eta > 0$ and all sufficiently large $d$, the number of feasible cones is bounded above by

$$|\mathcal{K}_x^{(d)}| = 2^{\lceil \tau_x d \rceil} \binom{d}{\lceil \tau_x d \rceil} \le e^{d(E(\tau_x)+\eta)},$$

for large enough $d$. Similarly, $|\mathcal{K}_y^{(d)}| \le e^{d(E(\tau_y)+\eta)}$ for all sufficiently large $d$.

We have now collected enough information to apply our theory. By Theorem 4.11, the event (6.2) holds with overwhelming probability in high dimensions so long as

$$\theta_{\ell_1}\big(\tau_x, E(\tau_x) + E(\tau_y)\big) + \theta_{\ell_1}\big(\tau_y, E(\tau_x) + E(\tau_y)\big) < 1. \tag{6.4}$$

The blue curve appearing in Figure 1 on page 3 shows the level set

$$\Big\{ (\tau_x, \tau_y) : \theta_{\ell_1}\big(\tau_x, E(\tau_x) + E(\tau_y)\big) + \theta_{\ell_1}\big(\tau_y, E(\tau_x) + E(\tau_y)\big) = 1 \Big\}.$$

When the sparsity level $(\tau_x, \tau_y)$ lies below the blue curve, inequality (6.4) holds, so that the demixing method (6.1) succeeds at demixing *every* pair $(x_0, y_0)$ of $(\tau_x, \tau_y)$-sparse vectors with overwhelming probability in high dimensions.

### 6.1.2   Numerical experiment for constrained MCA

The following numerical experiment illustrates the accuracy of these theoretical results. We fix the dimension $d = 100$. For each pair $(k_x, k_y) \in \{0, 1, \dots, 100\}^2$, we repeat the following procedure 25 times.

1. Draw $x_0, y_0 \in \mathbb{R}^d$ with $k_x$ or $k_y$ nonzero elements, respectively. The locations of the nonzero elements are chosen at random, and these elements are equally likely to be $+1$ or $-1$.
2. Generate a random basis $Q \in O_d$; see Remark 6.1 below.
3. Solve (6.1) for the optimal point $(x_\star, y_\star)$ with the numerical optimization software CVX [**42**], [**43**].
4. Declare success if $\|x_\star - x_0\|_{\ell_\infty} < 10^{-4}$.

The background of Figure 1 shows the results of this experiment as a function of $\tau_x = k_x/d$ and $\tau_y = k_y/d$. The yellow curve marks the empirical 50% success line, and we note that it tracks the green theoretical curve closely. It emerges that $d = 100$ is already large enough to see the high dimensional behavior described by our theoretical results.

**Remark 6.1.** A random basis in $Q \in O_d$ is often defined through a conceptually simple two-step operation. First, draw a square $d \times d$ Gaussian matrix; second, orthogonalize the columns of this matrix via the Gram–Schmidt procedure. Although this definition has the flavor of a numerical algorithm, the conceptual process is not numerically stable. Moreover, standard procedures for stabilizing the orthogonalization do not preserve the Haar measure. For a straightforward, numerically stable approach to generating random bases, see [**55**].

## 6.2   Secure and robust channel coding

Next, we study the secure channel coding scheme of Section 1.3.1. We want to analyze when the receiver can decode a transmitted message that is subject to a sparse corruption. The difficulty depends on the sparsity level $\tau$ of the corruption, where $\tau \in [0, 1]$. Let us introduce an ensemble of demixing problems. In each dimension $d \in \mathbb{N}$, choose a $d$-bit message $\boldsymbol{m}_0 \in \{\pm 1\}^d$ and a corruption $\boldsymbol{c}_0 \in \mathbb{R}^d$ with $\mathrm{nnz}(\boldsymbol{c}_0) = \lceil \tau d \rceil$. Draw a random basis $\boldsymbol{Q}$ known to both the receiver and transmitter. The receiver observes $\boldsymbol{z}_0 = \boldsymbol{Q}\boldsymbol{m}_0 + \boldsymbol{c}_0$, the encoded message plus the sparse interference.

The natural complexity measure for the sparse corruption $\boldsymbol{c}_0$ is the $\ell_1$ norm, while the $\ell_\infty$ norm is the appropriate complexity measure for the sign vector $\boldsymbol{m}_0$. Note that $\|\boldsymbol{m}_0\|_{\ell_\infty} = 1$. To recover the original message, the receiver solves the problem

$$
\begin{aligned}
&\text{minimize} && \|\boldsymbol{c}\|_{\ell_1} \\
&\text{subject to} && \|\boldsymbol{m}\|_{\ell_\infty} \leq 1 \ \text{ and } \ \boldsymbol{c} + \boldsymbol{Q}\boldsymbol{m} = \boldsymbol{z}_0.
\end{aligned}
\tag{6.5}
$$

This approach succeeds when $(\boldsymbol{m}_0, \boldsymbol{c}_0)$ is the unique optimal point of (6.5).

We consider two types of corruptions. *Benign* corruptions are taken in any manner that is independent of $\boldsymbol{Q}$—this would happen, for instance, when the corruption is generated by an adversary with no knowledge of $\boldsymbol{Q}$ or the transmission $\boldsymbol{Q}\boldsymbol{m}_0$. *Adversarial* corruptions are worst-case sparse corruptions; these corruptions may model malicious interference or an erasure channel that sets some of the coordinates of $\boldsymbol{z}_0$ to zero. We first consider with the benign corruption, and then consider the adversarial case in Section 6.2.2.

### 6.2.1   Benign corruptions

Since the corruption is chosen independently of the basis $\boldsymbol{Q}$, the convex demixing method (6.5) succeeds if it can identify the single pair $(\boldsymbol{c}_0, \boldsymbol{m}_0)$. Theorems 4.2 and 4.3 describe where procedure succeeds, or fails, with overwhelming probability in high dimensions, but we must first determine some decay thresholds.

The feasible cone of the $\ell_\infty$ norm at a sign vector is congruent to an orthant (Example 2.2). From Table 1, we see that the sequence $\{\mathbb{R}_+^d : d \in \mathbb{N}\}$ of orthants has an upper decay threshold of $\theta_{\mathbb{R}_+^d} = \frac{1}{2}$ and a matching lower decay threshold $\kappa_{\mathbb{R}_+^d} = \frac{1}{2}$. Appendix C defines an upper decay threshold $\theta_{\ell_1}(\tau)$ for the sequence of feasible cones of the $\ell_1$ norm at $\tau$-sparse vectors.

With these thresholds in hand, Theorem 4.2 guarantees that (6.5) succeeds with overwhelming probability in high dimensions provided that $\theta_{\mathbb{R}_+^d} + \theta_{\ell_1}(\tau) < 1$, or, equivalently, provided that $\theta_{\ell_1}(\tau) < \frac{1}{2}$. On the other hand, Theorem 4.3 shows that (6.5) fails with overwhelming probability in high dimensions if the corresponding upper decay threshold satisfies $\kappa_{\ell_1}(\tau) + \kappa_{\mathbb{R}_+^d} > 1$. This condition holds when $\kappa_{\ell_1}(\tau) > \frac{1}{2}$.

Numerically, we find that $\theta_{\ell_1}(\tau) < \frac{1}{2}$ when the sparsity level $\tau \lesssim 0.193$; see the left panel of Figure 11. The fact that $\kappa_{\ell_1}(\tau) = \theta_{\ell_1}(\tau)$ to numerical precision implies that $\kappa_{\ell_1}(\tau) > \frac{1}{2}$ for $\tau \gtrsim 0.193$. In other words, if fewer than 19% of the entries of $\boldsymbol{c}_0$ are nonzero, the scheme (6.5) succeeds with overwhelming probability in high dimensions; otherwise, it fails with overwhelming probability in high dimensions. This sharp transition corresponds to the location of the dashed line in Figure 3.

### 6.2.2   The adversarial case

In the adversarial case, the corruptions are sparse but may depend on the basis $\boldsymbol{Q}$ and the message $\boldsymbol{m}_0$. To ensure that no corruption with $\lceil \tau d \rceil$ nonzero entries can cause (6.5) to fail, we must verify that (6.5) succeeds at identifying $(\boldsymbol{c}_0, \boldsymbol{m}_0)$ for *every* $\lceil \tau d \rceil$-sparse vector $\boldsymbol{c}_0$. From Lemma 2.4, this is equivalent to the event

$$
\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{c}_0) \cap \left(-\boldsymbol{Q}\mathscr{F}(\|\cdot\|_{\ell_\infty}, \boldsymbol{m}_0)\right) = \{\boldsymbol{0}\} \text{ for every } \lceil \tau d \rceil\text{-sparse vector } \boldsymbol{c}_0.
\tag{6.6}
$$

We can use Theorem 4.11 to verify that event (6.6) holds with overwhelming probability in high dimensions. Let us collect the additional information required to verify the technical assumptions of this theorem.
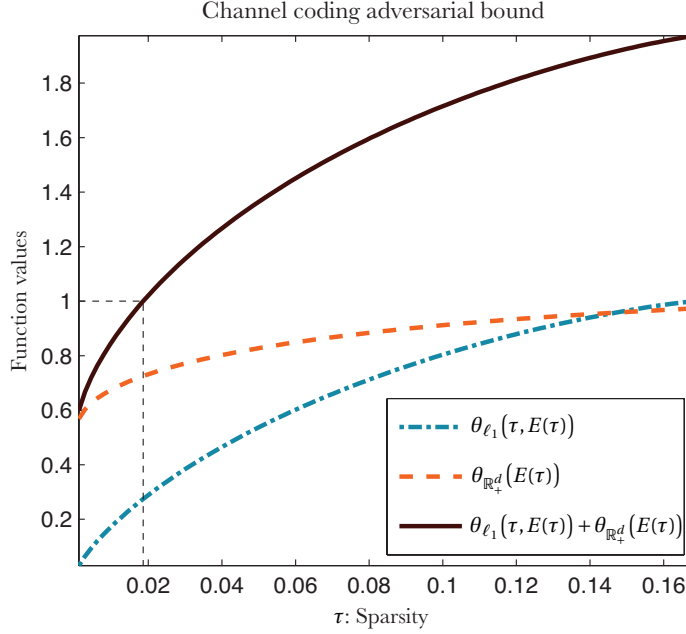
**Figure 12** *Calculating the adversarial guarantees for channel coding.* The lower two curves in the figure correspond to $\theta_{\ell_1}(\tau, E(\tau))$ and $\theta_{\mathbb{R}_+^d}(E(\tau))$ defined by (C.14) and (4.6). The upper curve indicates the sum $\theta_{\ell_1}(\tau, E(\tau)) + \theta_{\mathbb{R}_+^d}(E(\tau))$. For $\tau < 0.018$, the sum lies below one, so Theorem 4.11 implies that event (6.6) holds with overwhelming probability.

For $\boldsymbol{m}_0 \in \{\pm 1\}^d$, define the singleton $\mathcal{K}_{\boldsymbol{m}}^{(d)} = \{\mathcal{F}(\|\cdot\|_{\ell_\infty}, \boldsymbol{m}_0)\}$. The feasible cone $\mathcal{F}(\|\cdot\|_{\ell_\infty}, \boldsymbol{m}_0)$ is congruent to the orthant $\mathbb{R}_+^d$, so the sequence $\mathcal{K}_{\boldsymbol{m}}^{(d)}$ of sets has an upper decay threshold at level $\psi$ of $\theta_{\mathbb{R}_+^d}(\psi)$ defined in Proposition 4.9.

Define the set $\mathcal{K}_{\boldsymbol{c}}^{(d)} = \bigcup \{\mathcal{F}(\|\cdot\|_{\ell_1}, \boldsymbol{c}_0)\}$, where the union occurs over all vectors $\boldsymbol{c}_0$ with $\lceil \tau d \rceil$ nonzero elements. By (6.3), the size of $\mathcal{K}_{\boldsymbol{c}}^{(d)}$ is bounded by

$$\left| \mathcal{K}_{\boldsymbol{c}}^{(d)} \right| = 2^{\lceil \tau d \rceil} \binom{d}{\lceil \tau d \rceil} \leq e^{d(E(\tau) + \eta)}$$

for any $\eta > 0$ and all sufficiently large $d$.

Since $\mathcal{K}_{\boldsymbol{m}}^{(d)}$ is a singleton, Theorem 4.11 implies that event (6.6) holds with overwhelming probability in high dimensions whenever

$$\theta_{\mathbb{R}_+^d}(E(\tau)) + \theta_{\ell_1}(\tau, E(\tau)) < 1. \tag{6.7}$$

Computing $\theta_{\ell_1}(\tau, \psi)$ and $\theta_{\mathbb{R}_+^d}(\psi)$ numerically, we find that for all $\tau \lesssim 0.0186$, inequality (6.7) holds. We conclude that our channel coding scheme is robust to all adversarial corruptions so long as the corruptions have no more than about 1.8% nonzero entries. This computation is illustrated in Figure 12.

### 6.2.3 Numerical experiment

We perform two numerical experiments to complement our theory, one for the benign corruptions and the other for a specific type of malicious erasure.

For dimensions $d = 100$ and $d = 300$ and for each of 70 equally spaced values of $\tau \in [0, 0.35]$, we test the benign corruption case by repeating the following procedure 200 times:

1. Draw a binary vector $\boldsymbol{m}_0 \in \{\pm 1\}^d$ at random.

2. Choose a corruption $c_0$ with $k = \lceil \tau d \rceil$ nonzero elements; the support of $c_0$ is random, and the nonzero elements are taken to be $\pm 1$ with equal probability.

3. Generate a random basis $Q \in O_d$; see Remark 6.1.

4. Solve (6.5) with the observation $z_0 = Qm_0 + c_0$ with the numerical optimization software CVX; call $(m_\star, c_\star)$ the optimal point.

5. Declare success if $\|m_\star - m_0\|_{\ell_\infty} < 10^{-4}$.

The second experiment incorporates a malicious erasure. As in the benign case, the experiment is run for dimensions $d = 100$ and $d = 300$ and for 70 equally spaced values of $\tau$ between zero and one. For each of these parameters, we repeat the following 200 times:

1. Draw a message $m_0 \in \{\pm 1\}^d$ at random, and generate a random basis $Q \in O_d$.

2. Set the observation $z_0 = \text{erase}(Qm_0, \lceil \tau d \rceil)$, where $\text{erase}(x, k)$ sets the $k$ largest-magnitude elements of $x$ to zero.

3. Solve (6.5) with CVX for the optimal point $(m_\star, c_\star)$, and

4. Declare success if $\|m_\star - m_0\|_{\ell_\infty} < 10^{-4}$.

The curves in Figure 3 show the results of these experiments. For benign corruptions, the empirical 50% success rate occurs very near the predicted sparsity value $\tau = 0.193$, and the transition region is more narrow for larger $d$. Thus, the experiment closely match our prediction for the location of the asymptotic phase transition.

The empirical evidence suggests that our adversarial guarantees are conservative for the type of malicious corruption used in the experiment. This is expected, as we have no reason to believe that such erasures correspond to the worst-case corruption. However, the empirical transition between success and failure near $\tau \approx 0.05$ suggests that our adversarial bound lies within a factor of two or three of the best possible guarantee.

## 6.3 Low-rank matrices under sparse corruptions

Consider the problem of separating a low-rank square matrix from a corruption that is sparse in a random basis. The parameters that determine the difficulty are the proportional rank $\rho \in [0, 1]$ and the sparsity level $\tau \in [0, 1]$. Let us introduce a demixing ensemble. For each side length $n \in \mathbb{N}$, choose a low-rank matrix $X_0 \in \mathbb{R}^{n \times n}$ with $\text{rank}(X_0) = \lceil \rho n \rceil$ and a sparse matrix $Y_0 \in \mathbb{R}^{n \times n}$ with $\text{nnz}(Y_0) = \lceil \tau n^2 \rceil$. Draw a random basis $\mathcal{Q}$ for $\mathbb{R}^{n \times n}$, and suppose we observe $Z_0 = X_0 + \mathcal{Q}(Y_0)$.

To promote low-rank, we use the Schatten 1-norm, and to promote sparsity, we use the matrix $\ell_1$ norm. Given the side information $\alpha = \|Y_0\|_{\ell_1}$, we pose the convex demixing method

$$
\begin{aligned}
\text{minimize} \quad & \|X\|_{S_1} \\
\text{subject to} \quad & \|Y\|_{\ell_1} \leq \alpha \text{ and } X + \mathcal{Q}(Y) = Z_0.
\end{aligned}
\tag{6.8}
$$

We study when $(X_0, Y_0)$ is the unique solution to (6.8) with overwhelming probability in high dimensions.

The feasible cone of the $\ell_1$-matrix norm at a matrix $Y_0 \in \mathbb{R}^{n \times n}$ with $k$ nonzero elements is isomorphic to the feasible cone of the $\ell_1$ norm at a sparse vector $y_0 := \text{vec}(Y_0) \in \mathbb{R}^{n^2}$ with $k$ nonzero entries. It follows that the value $\theta_{\ell_1}(\tau)$ from (C.14) is an upper decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, Y_0)\}$ of feasible cones indexed by the ambient dimension $d = n^2$ of the matrix space $\mathbb{R}^{n \times n}$.

By Proposition 5.4, we see that $\theta_{S_1}(\rho) := 6\rho - 3\rho^2$ is an upper decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{S_1}, X_0)\}$ of feasible cones, indexed by the ambient dimension. However, a smaller upper decay threshold $\tilde{\theta}_{S_1}(\rho)$ is available using the results of [60]—see Remark 5.5. The green line in Figure 4 is the level set

$$
\{(\rho, \tau) : \theta_{\ell_1}(\tau) + \tilde{\theta}_{S_1}(\rho) = 1\},
$$

where the $\tilde{\theta}_{S_1}(\rho)$ is given by the asymptotic upper bound on the Gaussian width given implicitly in [**60,** Eq. (94)].[†] For $(\rho, \tau)$ pairs lying below the curve, Theorem 4.2 implies that our demixing method (6.8) succeeds with overwhelming probability in high dimensions.

### 6.3.1 Numerical experiment

Let us summarize the experiment in Figure 4. The matrix side length $n = 35$ is fixed, and for each pair $(\rho, \tau)$ in the set $\left\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}^2$, we repeat the following procedure 25 times:

1. Draw a matrix $X_0 = Q_L \Lambda Q_R \in \mathbb{R}^{n \times n}$ with rank $r = [\rho n]$, where $\Lambda$ is a diagonal matrix that satisfies $\Lambda_{ii} = 1$ for $i = 1, \ldots r$ and $\Lambda_{ii} = 0$ otherwise and $Q_L, Q_R$ are independent random bases in $O_n$.

2. Generate a random matrix $Y_0 \in \mathbb{R}^{n \times n}$ with $[\tau n^2]$ nonzero entries; the nonzero entries in $Y_0$ take the values $+1$ or $-1$ with equal probability.

3. Generate a random basis $\mathcal{Q}$ for $\mathbb{R}^{n \times n}$.

4. Solve (6.8) with the observation $Z_0 = X_0 + \mathcal{Q}(Y_0)$ with CVX, and set $(X_\star, Y_\star)$ to the optimal point.

5. Declare success if $\|X_\star - X_0\|_{\ell_\infty} < 10^{-4}$.

From Figure 4, we see that the theoretical bound closely matches the empirical success curve throughout the entire regime.

## 6.4 Assorted matrix demixing problems

We conclude this section with some other combinations of structured square matrices that we can separate using the convex demixing method (1.2). In each of these applications, we observe a superposition of the form $Z_0 = X_0 + \mathcal{Q}(Y_0) \in \mathbb{R}^{n \times n}$, where $\mathcal{Q}$ is a random basis for the matrix space $\mathbb{R}^{n \times n}$. We consider various structures for $X_0$ and $Y_0$—either low rank, orthogonal, sparse, or sign matrices—and we show that our theory quickly identifies a regime where an appropriate convex demixing method succeeds with overwhelming probability in high dimensions. While we know of no concrete applications for these particular demixing programs, the analysis below illustrates the ease with which our theory extends to new settings.

### 6.4.1 Orthogonal and sparse matrices

Fix a sparsity level $\tau$ in $[0, 1]$. For each side length $n \in \mathbb{N}$, choose an orthogonal matrix $X_0 \in O_n$ and a sparse matrix $Y_0 \in \mathbb{R}^{n \times n}$ with $\mathrm{nnz}(Y_0) = \lceil \tau n^2 \rceil$. We use the operator norm as a complexity measure for $X_0$ and the matrix $\ell_1$ norm as a complexity measure for $Y_0$. Since $\|X_0\|_{\mathrm{Op}} = 1$, we pose the convex demixing method

$$
\begin{aligned}
\text{minimize} \quad & \|Y\|_{\ell_1} \\
\text{subject to} \quad & \|X\|_{\mathrm{Op}} \leq 1 \text{ and } X + \mathcal{Q}(Y) = Z_0.
\end{aligned}
\tag{6.9}
$$

The interchange of the objective and constraint as compared with (1.2) poses no difficulty because the optimality conditions of Lemma 2.4 are symmetric with respect to the objective and constraint.

By Theorem 4.2, the convex demixing method (6.9) succeeds with overwhelming probability in high dimensions so long as

$$\theta_{\ell_1}(\tau) + \theta_{\mathrm{Op}} < 1,$$

where $\theta_{\ell_1}(\tau)$, defined in Appendix C, is an upper decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, Y_0)\}$ and $\theta_{\mathrm{Op}} = \frac{3}{4}$ is an upper decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{\mathrm{Op}}, X_0)\}$ by Proposition 5.6. Therefore, program (6.9) succeeds with overwhelming probability in high dimensions whenever

$$\theta_{\ell_1}(\tau) < \frac{1}{4}.$$

---

[†]Our actual computation uses the simpler, but equivalent, formula given in [**5,** Prop. 4.9].

This occurs for $\tau < 0.06$; see the left panel of Figure 11. We conclude that (6.9) demixes an orthogonal matrix $X_0$ from a matrix sparse in a random basis $\mathcal{Q}(Y_0)$ with high probability when no more than about 6% of the elements of $Y_0$ are nonzero.

### 6.4.2   Low-rank and sign matrices

Fix the proportional rank $\rho$ in $[0, 1]$. We abbreviate the ambient dimension $d = n^2$. For each side length $n \in \mathbb{N}$, choose a low-rank matrix $X_0 \in \mathbb{R}^{n \times n}$ with $\mathrm{rank}(X_0) = \lceil \rho n \rceil$ and a sign matrix $Y_0 \in \{\pm 1\}^{n \times n}$. We use the Schatten 1-norm as a complexity measure for rank and the matrix $\ell_\infty$ norm as a complexity measure for sign matrices. Given that $\|Y_0\|_{\ell_\infty} = 1$, we consider the convex demixing method

$$
\begin{aligned}
\text{minimize} \quad & \|X\|_{S_1} \\
\text{subject to} \quad & \|Y\|_{\ell_\infty} \le 1 \text{ and } X + \mathcal{Q}(Y) = Z_0.
\end{aligned}
\tag{6.10}
$$

We invoke Theorem 4.2 to see that (6.10) succeeds with overwhelming probability in high dimensions whenever $\theta_{S_1}(\rho) + \theta_{\mathbb{R}_+^d} < 1$. Here, $\theta_{S_1}(\rho)$ is an upper decay threshold for the ensemble $\{\mathcal{F}(\|\cdot\|_{S_1}, X_0)\}$ of feasible cones, and $\theta_{\mathbb{R}_+^d}$ is an upper decay threshold for the ensemble of nonnegative orthants. By Proposition 3.12, we have $\theta_{\mathbb{R}_+^d} = \frac{1}{2}$, while Proposition 5.4 gives $\theta_{S_1}(\rho) = 6\rho - 5\rho^2$. Therefore, the convex demixing method (6.10) succeeds with overwhelming probability in high dimensions so long as

$$
6\rho - 5\rho^2 < \frac{1}{2}.
$$

This bound is valid when $\rho \le 0.09$. We conclude that (6.10) can demix a low-rank matrix from a sign matrix in a random basis with overwhelming probability if $\mathrm{rank}(X_0) \le 0.09n$, where $n$ is side length of $X_0$.

### 6.4.3   Low-rank and orthogonal matrices

Let $\rho \in [0, 1]$ be a proportional rank parameter. For each side length $n \in \mathbb{N}$, choose a low-rank matrix $X_0 \in \mathbb{R}^{n \times n}$ with $\mathrm{rank}(X_0) = \lceil \rho n \rceil$ and an orthogonal matrix $Y_0 \in \mathsf{O}_n$. With the usual choice of complexity measures, the convex demixing method is

$$
\begin{aligned}
\text{minimize} \quad & \|X\|_{S_1} \\
\text{subject to} \quad & \|Y\|_{\mathrm{Op}} \le 1 \text{ and } X + \mathcal{Q}(Y) = Z_0.
\end{aligned}
\tag{6.11}
$$

By Theorem 4.2, program (6.11) succeeds with overwhelming probability in high dimensions so long as $\theta_{S_1}(\rho) + \theta_{\mathrm{Op}} < 1$. Propositions 5.4 and 5.6 imply that this occurs whenever

$$
6\rho - 3\rho^2 < \frac{1}{4}.
$$

For instance, it suffices that $\rho \le 0.04$. Therefore, the convex demixing method (6.11) can identify a superposition of a low-rank matrix and an orthogonal matrix with overwhelming probability in high dimensions so long as $\mathrm{rank}(X_0) \le 0.04n$, where $n$ is the side length of $X_0$.

## 7   Prior art and future directions

This work occupies a unique place in the literature on demixing. The analysis is highly general and applies to many problems. At the same time, the results are sharp or nearly sharp. This final section offers a wide-angle view of the field of demixing, from applications to analytical techniques, with a focus on methods based on convex optimization. We conclude by discussing some extensions of our current approach in the hope of encouraging further development in this field.

## 7.1 A short history of convex demixing and incoherence

The use of convex optimization for signal demixing has a long history. Early predecessors to morphological component analysis come from the work of Claerbout & Muir [**19**] and Taylor et al. [**77**], where $\ell_1$ minimization is used to identify sparse spike trains from an observed seismic trace.

Demixing methods based on $\ell_1$ minimization were put on a rigorous footing in the 1980s with the work of Santosa & Symes [**68**] and Donoho & Stark [**27**]. These results, either implicitly or explicitly, rely on incoherence in the form of an uncertainty principle. The work of Donoho & Huo [**26**] formalizes the notion of incoherence. Incoherent models, both random and deterministic, now pervade the sparse demixing literature [**70**], [**71**], [**35**], [**46**], [**75**].

In the last decade, new classes of convex regularizers have been introduced for solving inverse problems in signal processing. In particular, the Schatten 1-norm is used for problems involving low-rank matrices [**36**], [**64**]. Demixing methods that involve the Schatten 1-norm include robust principal component analysis [**12**], [**87**], [**88**], [**52**] and latent variable selection [**14**]. Rigorous theoretical results for these techniques typically involve a spectral incoherence assumption, but no previous work in this area identifies phase transition behavior.

## 7.2 The neighborhood of this work

We take much of our inspiration from the geometric analysis of linear inverse problems in [**15**]. Indeed, the geometric optimality condition (Lemma 2.4) is a direct generalization of a geometric result [**15,** Prop. 2.1] for linear inverse problems. Moreover, the Gaussian width bounds from that work prove useful for computing the decay thresholds in this research.

A related line of work, due to Negahban et al. [**56**], [**57**], is based on the concept of restricted strong convexity. The results in these papers are sharp within constant factors, but they do not yield bounds as precise as ours. Another general approach to demixing appears in of [**46**], where a deterministic incoherence condition leads to recovery guarantees, even in nonconvex settings. The recovery bounds available through this method are not competitive with the guarantees we provide.

Several works also consider demixing two sparse vectors. The works [**84**], [**58**] show that a nearly dense vector could be demixed from a sufficiently sparse vector, but they do not identify phase transition behavior. Recent work [**63**] also offers demixing guarantees for demixing sparse vectors when the sparsity is mildly sublinear in the dimension. Their model is similar to our MCA formulation in Section 1.1, but again the results do not identify the phase transition between success and failure.

### 7.2.1 Random geometry and convex optimization

We now trace the use of methods from integral geometry for understanding randomized convex optimization programs. Vershik & Sporyshev [**80**] use an asymptotic analysis of polytope angles to analyze the average-case behavior of the simplex method for linear programming. The underlying formulas have their roots in the results of Ruben [**67**], although some of the ideas apparently go back to the work Schläfli from the mid-nineteenth century—see Ruben's paper for a discussion.

The analysis of Vershik & Sporyshev fed a line of investigation on the expected face counts of randomly projected polytopes [**2**], [**9**], a topic of theoretical interest in combinatorial geometry. These computations resurfaced in convex optimization in the line of work of Donoho & Tanner [**23**], [**28**], [**25**], [**29**], [**30**], [**31**], [**33**]. These articles characterize the behavior of convex optimization methods for solving several linear inverse problems under a random measurement model. In Appendix C, we leverage the asymptotic polytope angle calculations of Donoho & Tanner to compute decay threshold for the $\ell_1$ norm at sparse vectors.

This asymptotic polytope angle approach also yields stability guarantees for basis pursuit [**89**]. Furthermore, it has been used to establish that iteratively reweighted basis pursuit can provide strictly stronger guarantees than standard basis pursuit [**90**], [**48**], [**49**].

Our approach to random geometry differs from these earlier works because it starts with the modern theory of spherical integral geometry. Previous research was based on an older theory of polytope angles [**44**], [**45**], [**54**]. Spherical integral geometry reached its current state of development in the dissertation [**38**], [**39**]. Chapter 6.5 of [**69**] and the notes therein summarize this research. We also draw on insights from the thesis [**3**].

## 7.3   Conclusions and future directions

The results in this work demonstrate the power of spherical integral geometry in the context of demixing. Our bounds are often tight, and they are broadly applicable. This approach raises many questions worth further attention. We conclude with a list of directions for future work. During the period that our original manuscript was under review, several of these areas have seen significant progress. We augment our original list of open problems with a summary of progress made in the interim.

**Tight results for Lagrangian demixing.**   The Lagrange penalized demixing method (1.3) is important because it requires less knowledge about the unobserved vectors $(x_0, y_0)$ than the corresponding constrained method (1.2). The results in this work give information regarding the potential for, and the limits of, the penalized demixing approach (1.2). Nevertheless, a precise analysis of the penalized problem (1.3) and its dependence on the penalty parameter $\lambda$ would have real practical value.

*While a sharp phase transition characterization for the Lagrange demixing problem* (1.3) *remains open, the recent work* [**37**] *offers theoretical guarantees and explicit choices of Lagrange parameters.*

**Multiple demixing.**   It would be interesting to study demixing problems involving more than two structured vectors.

*A study of this problem appears in* [**83**]. *The present authors provide sharp phase transition characterizations for demixing an arbitrary number of signals in* [**53**].

**Spherical intrinsic volumes for more cones.**   Computation of additional decay thresholds will provide new bounds for convex demixing methods. The sharpest decay thresholds appear to require formulas for spherical intrinsic volumes. For instance, an asymptotic analysis of the spherical intrinsic volumes for feasible cones of the Schatten 1-norm would provide sharp recovery results for low-rank matrix demixing problems. Amelunxen & Bürgisser have made some recent progress in this direction by developing a formula for the spherical intrinsic volumes for the semidefinite cone [**4**].

**Log-concavity of spherical intrinsic volumes.**   Bürgisser & Amelunxen [**10,** Conj. 2.19] conjecture that the sequence of spherical intrinsic volumes is log-concave. This conjecture is closely related to the question of whether the upper and lower decay thresholds match.

*In recent work by the present authors and collaborators, the intrinsic volumes are shown to have a nontrivial log-concave upper bound* [**5,** Sec. 6.1]. *While this result implies that the upper and lower decay thresholds are often equal (Section 4.2.2), the log-concavity conjecture remains open.*

**Extensions to more general probability measures.**   The analysis in this work focuses on a specific random model. It would be interesting to incorporate more general probability measures into our framework. This may be a difficult problem; by the results of Section 5.2, this question is closely related to the observed universality phenomenon in basis pursuit [**30**].

*Bayati et al.* [**6**] *provide a rigorous version the universality property for basis pursuit observed in* [**30**]. *It remains unclear whether their methods adapt to demixing problems considered here.*

## A   Equivalence of the constrained and penalized methods

This appendix provides a geometric account of the equivalence between the constrained (1.2) and penalized (1.3) convex demixing methods. The results in this section let us interpret our conditions for

the success of the constrained demixing method (1.2) as limits on, and opportunities for, the Lagrange demixing method (1.3).

We begin with the following well-known result; it holds without any technical restrictions. We omit the demonstration, which is an easy exercise in proof by contradiction. (See also [**65,** Cor. 28.1.1].)

**Proposition A.1.** *Suppose the Lagrange problem* (1.3) *succeeds for some value* $\lambda > 0$. *Then* (1.2) *succeeds.*

Before stating a partial converse to Proposition A.1, we require a technical definition. We say that a proper convex function $f$ is *typical at* $x$ if $f$ is subdifferentiable at $x$ but does not achieve its minimum at $x$. With this technical condition in place, we have the following complement to Proposition A.1.

**Proposition A.2.** *Suppose* $f$ *is typical at* $x_0$ *and* $g$ *is typical at* $y_0$. *If the constrained method* (1.2) *succeeds, then there exists a parameter* $\lambda > 0$ *such that* $(x_0, y_0)$ *is an optimal point for the Lagrange method* (1.3).

Note that there is a subtlety here: the Lagrange program may have strictly more optimal points than the corresponding constrained problem even for the best choice of $\lambda$, so that we cannot guarantee that $(x_0, y_0)$ is the unique optimum. See [**65,** Sec. 28] for more details.

*Proof of Proposition A.2.* The key idea is the construction of a subgradient that certifies the optimality of the pair $(x_0, y_0)$ for the Lagrange penalized problem (1.3) for an appropriate choice of parameter $\lambda$. As with many results in convex analysis, a separating hyperplane plays an important role.

By Lemma 2.4, the constrained problem (1.2) succeeds if and only if $\mathscr{F}(f, x_0) \cap -Q\mathscr{F}(g, y_0) = \{0\}$. The trivial intersection of the feasible cones implies that there exists a hyperplane that separates these cones. (This fact is a special case of the Hahn–Banach separation theorem for convex cones due to Klee [**50**].) In other words, there exists some vector $u \neq 0$ such that

$$\langle u, x \rangle \leq 0 \text{ for all } x \in \mathscr{F}(f, x_0),$$

and moreover

$$\langle u, y \rangle \geq 0 \text{ for all } y \in -Q\mathscr{F}(g, y_0).$$

In the language of polar cones, the first separation inequality is simply the statement that $u \in \mathscr{F}(f, x_0)^\circ$, while the second inequality is equivalent to $Q^* u \in \mathscr{F}(g, y_0)^\circ$.

We will now show that $u$ generates a subgradient optimality certificate for the point $(x_0, y_0)$ in problem (1.3) for an appropriate choice of parameter $\lambda > 0$. We denote the subdifferential map by $\partial$.

At this point, we invoke our technical assumption. Since $f$ is typical at $x_0$, the polar to the feasible cone is generated by the subdifferential of $f$ at $x_0$ [**65,** Thm 23.7]. In particular, there exists a number $\lambda_f \geq 0$ such that $u \in \lambda_f \partial f(x_0)$. In fact, the stronger inequality $\lambda_f > 0$ holds because $u \neq 0$. For the same reason, there exists a number $\lambda_g > 0$ such that $Q^* u \in \lambda_g \partial g(y_0)$.

Define $h(x) := \lambda_f f(x) + \lambda_g g(Q^*(z_0 - x))$. By standard transformation rules for subdifferentials [**65,** Thms. 23.8, 23.9], we have

$$\partial h(x_0) \supset \lambda_f \partial f(x_0) - \lambda_g Q \partial g(y_0),$$

where $A - B := A + (-B)$ is the Minkowski sum of the sets $A$ and $-B$. Since $u \in \lambda_f \partial f(x_0)$ and $u \in \lambda_g Q \partial g(y_0)$, we see $0 \in \partial h(x_0)$. By the definition of subgradients, $x_0$ is a global minimizer of $h$. Introducing the variable $y = Q^*(z_0 - x)$, it follows that $(x_0, y_0)$ is a global minimizer of

$$\begin{aligned}
\text{minimize} \quad & f(x) + \tfrac{\lambda_g}{\lambda_f} g(y) \\
\text{subject to} \quad & x + Qy = z_0.
\end{aligned}$$

This is Lagrange problem (1.3) with the parameter $\lambda = \lambda_g / \lambda_f > 0$, so we have the result. $\qquad\square$

# B Regions of failure and uniform guarantees

We now present the proofs of the results of Section 4 concerning regions of failure (Theorem 4.3) and strong demixing guarantees (Theorem 4.11) for the convex demixing method. These demonstrations closely follow the pattern laid down by the proof of Theorem 4.2.

## B.1 Regions of failure: The proof of Theorem 4.3

We first state an analog of Theorem 4.4. As usual, $\mathscr{D}$ is an infinite set of indices.

**Theorem B.1.** *Let $\{K^{(d)} \subset \mathbb{R}^d : d \in \mathscr{D}\}$ and $\{\tilde{K}^{(d)} \subset \mathbb{R}^d : d \in \mathscr{D}\}$ be two ensembles of closed convex cones with lower decay thresholds $\kappa_\star$ and $\tilde{\kappa}_\star$. If $\kappa_\star + \tilde{\kappa}_\star > 1$, then there exists an $\varepsilon > 0$ such that $\mathbb{P}\{K^{(d)} \cap \boldsymbol{Q}\tilde{K}^{(d)} \neq \{\boldsymbol{0}\}\} \geq 1 - \mathrm{e}^{-\varepsilon d}$ for all sufficiently large $d$.*

Theorem 4.3 follows from Theorem B.1 in the same way that Theorem 4.2 follows from Theorem 4.4 with one additional technical point regarding closure conditions.

*Proof of Theorem 4.3 from Theorem B.1.* By the assumptions in Theorem 4.3, the ensembles $\{\overline{\mathscr{F}}(f^{(d)}, \boldsymbol{x}_0^{(d)})\}$ and $\{-\overline{\mathscr{F}}(g^{(d)}, \boldsymbol{y}_0^{(d)})\}$ of closed cones satisfy the hypotheses of Theorem B.1. Therefore, there is an $\varepsilon > 0$ such that the closure of the feasible cones have *non*trivial intersection with probability at least $1 - \mathrm{e}^{-\varepsilon d}$, for all large enough $d$.

It follows from Remark 4.5 that the probability of the event $\overline{\mathscr{F}}(f^{(d)}, \boldsymbol{x}_0^{(d)}) \cap -\boldsymbol{Q}\overline{\mathscr{F}}(g^{(d)}, \boldsymbol{y}_0^{(d)}) \neq \{\boldsymbol{0}\}$ is equal to the probability of the event $\mathscr{F}(f^{(d)}, \boldsymbol{x}_0^{(d)}) \cap -\boldsymbol{Q}\mathscr{F}(g^{(d)}, \boldsymbol{y}_0^{(d)}) \neq \{\boldsymbol{0}\}$. Applying the geometric optimality condition of Lemma 2.4 immediately implies that (1.2) fails with probability at least $1 - \mathrm{e}^{-\varepsilon d}$. $\qquad\square$

The proof of Theorem B.1 requires an additional fact concerning spherical intrinsic volumes.

**Fact B.2** (Spherical Gauss–Bonnet formula [69, P. 258])**.** *For any closed convex cone $K \subset \mathbb{R}^d$ that is not a subspace, we have*

$$\sum_{\substack{i=-1 \\ i\ even}}^{d-1} v_i(K) = \sum_{\substack{i=-1 \\ i\ odd}}^{d-1} v_i(K) = \frac{1}{2}.$$

In the proof below, the Gauss–Bonnet formula is crucial for dealing with the parity term $(1 + (-1)^k)$ that arises in the spherical kinematic formula (3.2).

*Proof of Theorem B.1.* Since the Gauss–Bonnet formula only applies to cones that are not subspaces, we split the demonstration into three cases: neither ensemble $\{K^{(d)}\}$ nor $\{\tilde{K}^{(d)}\}$ contains a subspace, one ensemble consists of subspaces, or both ensembles consist of subspaces. We assume without loss that each case holds for every dimension $d \in \mathscr{D}$; the proof extends to the general case by considering subsequences where only a single case applies.

We drop the superscript $d$ for clarity. Assume first that neither $K$ nor $\tilde{K}$ is a subspace. Let $\varpi(k) = (1 + (-1)^k)$ be the parity term in the spherical kinematic formula (3.2). Changing the order of summation in the spherical kinematic formula, we find

$$P := \mathbb{P}\left\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\}\right\} = \sum_{i=0}^{d-1} v_i(K) \sum_{k=d-i-1}^{d-1} \varpi(k-d+1+i) v_k(\tilde{K}).$$

Let $\kappa < \kappa_\star$ and $\tilde{\kappa} < \tilde{\kappa}_\star$ with $\kappa + \tilde{\kappa} > 1$; such scalars exist because $\kappa_\star + \tilde{\kappa}_\star > 1$. By positivity of the spherical intrinsic volumes (Fact 3.5.1), we have

$$P \geq \sum_{i=\lceil \kappa d \rceil + 1}^{d-1} v_i(K) \sum_{k=d-i-1}^{d-1} \varpi(k-d+1+i) v_k(\tilde{K}) \tag{B.1}$$

We will see that the inner sum above is very close to one. Indeed,

$$\sum_{k=d-i-1}^{d-1} \varpi(k-d+1+i)\,v_k(\tilde{K}) = \sum_{k=-1}^{d-1} \varpi(k-d+1+i)\,v_k(\tilde{K}) - \tilde{\xi}_i = 1 - \tilde{\xi}_i, \tag{B.2}$$

where $\tilde{\xi}_i$ is a discrepancy term (see (B.3) below). The second equality follows by the spherical Gauss–Bonnet formula (Fact B.2), and the assumption that $\tilde{K}$ is not a subspace.

We now bound the discrepancy term $\tilde{\xi}_i$ uniformly over $i \geq \lceil \kappa d \rceil + 1$. Since $\kappa + \tilde{\kappa} > 1$, for any $i \geq \lceil \kappa d \rceil + 1$ we have $d - 2 - i \leq \lceil \tilde{\kappa} d \rceil$. By definition of the lower decay threshold, we see that the discrepancy term must be small: for any $i \geq \lceil \kappa d \rceil + 1$,

$$\tilde{\xi}_i = \sum_{k=-1}^{d-2-i} \varpi(k-d+1+i)\,v_k(\tilde{K}) \leq 2 \sum_{k=-1}^{\lceil \tilde{\kappa} d \rceil} v_k(\tilde{K}) \leq 2(d-1)\mathrm{e}^{-\varepsilon' d}, \tag{B.3}$$

for some $\varepsilon' > 0$ and all sufficiently large $d$. Applying (B.2) and (B.3) to (B.1), we find

$$P \geq \sum_{i=\lceil \kappa d \rceil + 1}^{d-1} v_i(K)\big(1 - 2(d-1)\mathrm{e}^{-\varepsilon' d}\big) \geq \left( \sum_{i=\lceil \kappa d \rceil + 1}^{d-1} v_i(K) \right) - 2(d-1)\mathrm{e}^{-\varepsilon' d}, \tag{B.4}$$

where the second inequality follows from Fact 3.5: the spherical intrinsic volumes are positive and sum to one. We now reindex the sum on the right-hand side of (B.4) over $i = -1, 0, \ldots, d-1$ with only exponentially small loss:

$$\sum_{i=\lceil \kappa d \rceil + 1}^{d-1} v_i(K^{(d)}) = \sum_{i=-1}^{d-1} v_i(K) - \xi$$

where the discrepancy $\xi$ satisfies

$$\xi = \sum_{i=-1}^{\lceil \kappa d \rceil} v_k(K) \leq (d-1)\mathrm{e}^{-\varepsilon'' d},$$

for some $\varepsilon'' > 0$ and all sufficiently large $d$ by definition of the lower decay threshold. Applying these observations to (B.4), we deduce that

$$P \geq \sum_{i=-1}^{d-1} v_i(K) - (d-1)\Big(\mathrm{e}^{-\varepsilon' d} + \mathrm{e}^{-\varepsilon'' d}\Big) \geq \sum_{i=-1}^{d-1} v_i(K) - \mathrm{e}^{-\varepsilon d}$$

for some $\varepsilon > 0$ and all sufficiently large $d$. Since $\sum_{i=-1}^{d-1} v_i(K) = 1$ by Fact 3.5.2, this shows the result when both $K$ and $\tilde{K}$ are not subspaces, completing the first case.

For the second case, suppose that only one of the cones is a subspace. Without loss, we may assume $\tilde{K}$ is the subspace by the symmetry of the spherical kinematic formula (see Remark 3.8). Denote the dimension of the subspace $\tilde{K}$ by $\tilde{n} := \dim(\tilde{K})$, and take parameters $\kappa$ and $\tilde{\kappa}$ as above.

By Proposition 3.2, the spherical intrinsic volumes of $\tilde{K}$ are given by $v_i(\tilde{K}) = \delta_{i,\tilde{n}-1}$. Inserting this Kronecker $\delta$ into the spherical kinematic formula (3.2) and simplifying the resulting expression, we find the probability of interest is given by

$$P := \mathbb{P}\big\{K \cap \boldsymbol{Q}\tilde{K} \neq \{\boldsymbol{0}\}\big\} = \sum_{k=d-\tilde{n}}^{d-1} \varpi(k+\tilde{n}-d)\,v_k(K).$$

Reindexing the sum over $k = -1, 0, \ldots, d-1$, we see

$$P = \sum_{k=-1}^{d-1} \varpi(k+\tilde{n}-d)\,v_k(K) - \sum_{i=-1}^{d-\tilde{n}-1} \varpi(k+\tilde{n}-d)\,v_i(K) = 1 - \sum_{i=-1}^{d-\tilde{n}-1} \varpi(k+\tilde{n}-d)\,v_i(K) \tag{B.5}$$

where the second equality holds by the spherical Gauss–Bonnet formula (Fact B.2).

We now show that $\tilde{n}$ is relatively large. The definition of the lower decay threshold implies that there exists an $\varepsilon' > 0$ such that

$$v_i(\tilde{K}) = \delta_{i,\tilde{n}-1} \le e^{-\varepsilon' d} \quad \text{for all } i \le \lceil \tilde{\kappa} d \rceil$$

when $d$ is sufficiently large. This inequality cannot hold if $\tilde{n} - 1 \le \lceil \tilde{\kappa} d \rceil$, so we deduce that $\tilde{n} \ge \tilde{\kappa} d$ for all sufficiently large $d$.

Since $\tilde{n} \ge \tilde{\kappa} d$ and $\kappa + \tilde{\kappa} > 1$, we must have $d - \tilde{n} - 1 < \lceil \kappa d \rceil$ for all sufficiently large $d$. Applying the definition of the lower decay threshold, we find the sum on the right-hand side of (B.5) is exponentially small: there exists an $\varepsilon'' > 0$ such that

$$\sum_{i=-1}^{d-\tilde{n}-1} \varpi(k + \tilde{n} - d) v_i(K) \le 2(d-1)e^{-\varepsilon'' d}$$

for all sufficiently large $d$. The result for the second case follows immediately.

Finally, we consider the case when both of the cones are subspaces. Suppose $K$ has dimension $n$, while $\tilde{K}$ has dimension $\tilde{n}$, and let $\kappa$, $\tilde{\kappa}$ be as above. As in the second case, we find that $n \ge \kappa d$ and $\tilde{n} \ge \tilde{\kappa} d$ when $d$ is sufficiently large. Then the inequality $\kappa_\star + \tilde{\kappa}_\star > 1$ implies that $n + \tilde{n} > d$, that is, the sum of the dimensions of the subspaces is larger than the ambient dimension. A standard fact from linear algebra implies $K \cap Q\tilde{K} \ne \{\mathbf{0}\}$ for any unitary $Q$—in other words, for all $d$ large enough, the probability of nontrivial intersection is one. This completes the third case, and we are done. $\qquad\square$

## B.2   Proof of the strong guarantees of Theorem 4.11

*Proof of Theorem 4.11.* For clarity, we drop the superscript $d$ in this proof. We begin with the union bound: the probability of interest $P$ is bounded above by

$$P := \mathbb{P}\left\{ K \cap Q\tilde{K} \ne \{\mathbf{0}\} \text{ for any } K \in \mathcal{K}, \ \tilde{K} \in \tilde{\mathcal{K}} \right\} \le |\mathcal{K}| \cdot |\tilde{\mathcal{K}}| \cdot \mathbb{P}\left\{ K \cap Q\tilde{K} \ne \{\mathbf{0}\} \right\}. \tag{B.6}$$

From here, the proof closely parallels the proof of Theorem 4.4, so we compress the demonstration. We consider two cases, one where at least one cone is a subspace, the other where neither cone is a subspace; the result extends to the mixed case by considering subsequences.

Suppose first that at least one cone is a subspace. Let $\theta > \theta_\star$ and $\tilde{\theta} > \tilde{\theta}_\star$ with $\theta + \tilde{\theta} < 1$. We bound the probability on the right-hand side of (B.6) by

$$\frac{1}{2}\mathbb{P}\left\{ K \cap Q\tilde{K} \ne \{\mathbf{0}\} \right\} \le \Sigma_1 + \Sigma_2 + \Sigma_3 + \Sigma_4,$$

where the $\Sigma_i$ are given in (4.3). The fact that $\theta + \tilde{\theta} < 1$ implies that $\Sigma_1 = 0$ for sufficiently large $d$, as in the proof of Theorem 4.4. Since $\theta > \theta_\star$, the definition of the upper decay threshold at level $\psi + \tilde{\psi}$ implies

$$\Sigma_2 \le \sum_{i=\lceil \theta d \rceil + 1}^{d-1} v_i(K) \le (d-1)e^{-d(\psi + \tilde{\psi} + \varepsilon')}$$

for some $\varepsilon' > 0$ and all sufficiently large $d$. With analogous reasoning, we find similar exponential bounds for $\Sigma_3$ and $\Sigma_4$:

$$\Sigma_3 \le (d-1)e^{-d(\psi + \tilde{\psi} + \varepsilon'')}, \quad \Sigma_4 \le (d-1)e^{-d(\psi + \tilde{\psi} + \varepsilon''')}$$

again for positive $\varepsilon'', \varepsilon'''$ and all sufficiently large $d$. Summing these inequalities and taking $d$ sufficiently large gives

$$\mathbb{P}\left\{ K \cap Q\tilde{K} \ne \{\mathbf{0}\} \right\} \le e^{-d(\psi + \tilde{\psi} + \hat{\varepsilon})}$$

for some $\hat{\varepsilon} > 0$. The claim then follows from our exponential upper bound on the growth of $|\mathcal{K}|$ and $|\tilde{\mathcal{K}}|$ with $\eta = \hat{\varepsilon}/2$:

$$P \le 2 \cdot |\mathcal{K}| \cdot |\tilde{\mathcal{K}}| \cdot e^{d(\psi + \tilde{\psi} + \hat{\varepsilon}/2) - d(\psi + \tilde{\psi} + \hat{\varepsilon})} = 2e^{-\frac{\hat{\varepsilon}}{2} d}.$$

**Table 2** *Notation translation between this work and* [25]. *Note that in the reference,* $\Psi_{\text{net}}$ *is defined for three arguments, but only depends on two parameters, namely* $v$ *and* $\rho\delta$.

| Parameter | Our notation | Notation of [25] |
|---|---|---|
| Sparsity ratio | $\tau$ | $\rho\delta$ |
| Ratio of measurements to ambient dimension | $\sigma$ | $\delta$ |
| Undersampling ratio | $\tau/\sigma$ | $\rho$ |
| Normalized index | $\theta$ | $v$ |
| Internal exponent | $\Psi_{\text{int}}(\theta,\tau)$ | $\Psi_{\text{int}}(v,\rho\delta)$ |
| External exponent | $\Psi_{\text{ext}}(\theta,\tau)$ | $\Psi_{\text{ext}}(v,\rho\delta)$ |
| Total exponent | $\Psi_{\text{total}}(\theta,\tau)$ | $\Psi_{\text{net}}(v,\rho,\delta) - \Psi_{\text{face}}(\rho\delta)$ |
| Net exponent | $\Psi_{\text{total}}(\theta,\tau) + E(\tau)$ | $\Psi_{\text{net}}(v,\rho,\delta)$ |

Taking $\varepsilon = \hat{\varepsilon}/4$ and $d$ sufficiently large gives the claim in the first case.

Now consider the case where both cones are subspaces, and let $n := \dim(K)$ and $\tilde{n} := \dim(\tilde{K})$. Take parameters $\theta > \theta_\star$, and $\tilde{\theta} > \tilde{\theta}_\star$, such that $\theta + \tilde{\theta} < 1$. As in the proof of Theorem 4.4, the Kronecker $\delta$ expression for the intrinsic volumes of the subspaces $K$ and $\tilde{K}$ given by Proposition 3.2, combined with the definition of the upper decay threshold, reveals that $n \leq \lceil \theta d \rceil$ and $\tilde{n} \leq \lceil \tilde{\theta} d \rceil$ for all sufficiently large $d$. The fact that $\theta + \tilde{\theta} < 1$ implies $n + \tilde{n} < d$ for all sufficiently large $d$. Since randomly oriented subspaces are almost always in general position, the probability that $K \cap QK \neq \{\mathbf{0}\}$ is zero. This is the second case, so we are done. $\qquad\square$

# C   Decay thresholds for feasible cones of the $\ell_1$ norm

This section describes how we compute decay thresholds for the feasible cone of the $\ell_1$ at sparse vectors. The polytope angle calculations appearing in [25] form an important part of this computation. For convenient comparisons, Table 2 provides a map between our notation and that of the reference.

Fix a sparsity parameter $\tau \in [0,1]$, and let $\mathcal{D}$ be an infinite set of indices. For each dimension $d \in \mathcal{D}$, we define a vector $\boldsymbol{x}^{(d)} \in \mathbb{R}^d$ such that $\text{nnz}(\boldsymbol{x}^{(d)}) = \lceil \tau d \rceil$. The following results describes the behavior of the spherical intrinsic volumes of the feasible cone $\mathcal{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)})$ in terms of the sparsity $\tau$ and the normalized index $\theta = i/d$ when $d$ is large.

**Lemma C.1.** *Consider the ensemble above. There exists a function* $\Psi_{\text{total}}$ *such that, for every* $\varepsilon > 0$ *and all sufficiently large* $d \in \mathcal{D}$, *we have*

$$\frac{1}{d} \log\left( v_{\lceil \theta d \rceil}\left( \mathcal{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)}) \right) \right) \leq \Psi_{\text{total}}(\theta, \tau) + \varepsilon \tag{C.1}$$

*for all* $\theta \in [\tau, 1]$, *and*

$$v_{\lceil \theta d \rceil}\left( \mathcal{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)}) \right) = 0 \tag{C.2}$$

*for* $\theta \in [0, \tau)$.

We discuss the definition and computation of the normalized exponent $\Psi_{\text{total}}$ in Section C.1. This function provides decay thresholds for the ensemble $\{ \mathcal{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)}) : d \in \mathcal{D} \}$ in the same way that the limit (3.5) provides decay thresholds for the ensemble of orthants. See Section C.2 for details.

*Proof of Lemma C.1.* We leave the dependence on the dimension implicit for clarity. Define $k := \lceil \tau d \rceil$.

We first show that (C.1) holds. The proof relies on an expression for spherical intrinsic volumes in terms of polytope angles. For a face $F$ of a polytope $P$, we define $\beta(F, P)$ as the internal angle of $P$ at $F$ and $\gamma(F, P)$

as the external angle of $P$ at $F$ (see [**44,** Chapter 14] for the definitions). The following is an important alternative characterization of the spherical intrinsic volumes in terms of these angles.

**Fact C.2** ([**69,** Equation (6.50)]). *Let $K$ be a polyhedral cone, and let $\mathfrak{F}_i(K)$ be the set of all $i$-dimensional faces of $K$. Then*

$$v_i(K) = \sum_{F \in \mathfrak{F}_{i+1}(K)} \beta(\mathbf{0}, F)\gamma(F, K). \tag{C.3}$$

We now specialize Fact C.2 to the case where $K = \mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x})$. Define the sublevel set $S := \{\mathbf{w} : \|\mathbf{w}\|_{\ell_1} \leq \|\mathbf{x}\|_{\ell_1}\}$. Recalling that our assumption $\tau > 0$ implies $\|\mathbf{x}\|_{\ell_1} > 0$, we have

$$\mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x}) = \mathrm{cone}(S - \{\mathbf{x}\}) = \mathrm{cone}\big(C - \{\mathbf{x}/\|\mathbf{x}\|_{\ell_1}\}\big), \tag{C.4}$$

where $C := \{\mathbf{w} : \|\mathbf{w}\|_{\ell_1} \leq 1\}$ is the standard crosspolytope. The fact that $\mathbf{x}$ is $k$-sparse is equivalent to the statement that $\mathbf{x}/\|\mathbf{x}\|_{\ell_1}$ lies in the relative interior of a $(k-1)$-dimensional face of the crosspolytope $C$.

Relationship (C.4) implies that there is a one-to-one correspondence between the $i$-dimensional faces of $\mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x})$ and the $i$-dimensional faces of the crosspolytope $C$ that contain $\mathbf{x}/\|\mathbf{x}\|_{\ell_1}$. Since the internal and external angles only depend on the local structure of a given polytope, we find that for every nonempty face $F$ of $\mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x})$ the internal and external angles satisfy

$$\beta(\mathbf{0}, F) = \beta(\mathbf{x}, \tilde{F}) \text{ and } \gamma(F, K) = \gamma(\tilde{F}, C),$$

where $\tilde{F}$ is the face of the crosspolytope $C$ naturally corresponding to the face $F$ of the feasible cone $\mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x})$.

A number of important relationships due to Böröczky & Henk [**9**] for faces of the crosspolytope $C$ are conveniently collected in [**25,** Section 3.3]. In particular, we will need the following two facts:

1. There are $2^{i-k+2}\binom{d-k}{i-k+2}$ faces of $C$ of dimension $(i+1) \geq (k-1)$ containing a given $(k-1)$-dimensional face of $C$, and

2. The high degree of symmetry of the crosspolytope ensures that the internal and external angles at these faces depend only on the dimensional parameters $k$ and $i$.

Applying the observations above to equation (C.3), we find

$$v_i\big(\mathscr{F}(\|\cdot\|_{\ell_1}, \mathbf{x})\big) = 2^{i-k+2}\binom{d-k}{i-k+2}\beta(T_{k-1}, T_{i+1})\gamma(\tilde{F}_{i+1}, C) \tag{C.5}$$

for $i = \{k-1, \ldots, d-1\}$. Above, $\tilde{F}_{i+1}$ is any $(i+1)$-dimensional face of the crosspolytope $C$, and $T_j$ is the $j$-dimensional regular simplex.

The internal and external angles in (C.5) have explicit expressions due to [**9**] and the work of Ruben [**67**]. Donoho [**25**] conducts an asymptotic investigation of these formulas. To distill the essence of the analysis, Donoho gives continuous functions $\Psi_{\mathrm{int}}(\theta, \tau)$ and $\Psi_{\mathrm{ext}}(\theta)$ such that, for any $\varepsilon > 0$ and all sufficiently large $d$, the inequalities

$$\frac{1}{d}\log\big(\beta(T_{k-1}, T_{i+1})\big) \leq -\Psi_{\mathrm{int}}\big(\tfrac{i}{d}, \tau\big) + \frac{\varepsilon}{3} \tag{C.6}$$

$$\frac{1}{d}\log\big(\gamma(F_{i+1}, C)\big) \leq -\Psi_{\mathrm{ext}}\big(\tfrac{i}{d}\big) + \frac{\varepsilon}{3} \tag{C.7}$$

hold uniformly over $i = \{\lceil \tau d \rceil, \ldots, d-1\}$. Moreover, it follows from Equation (3.5) that for sufficiently large $d$, we have

$$\frac{1}{d}\log\left(2^{i-k+2}\binom{d-k}{i-k+2}\right) \leq \Psi_{\mathrm{cont}}\big(\tfrac{i}{d}, \tau\big) + \frac{\varepsilon}{3},$$
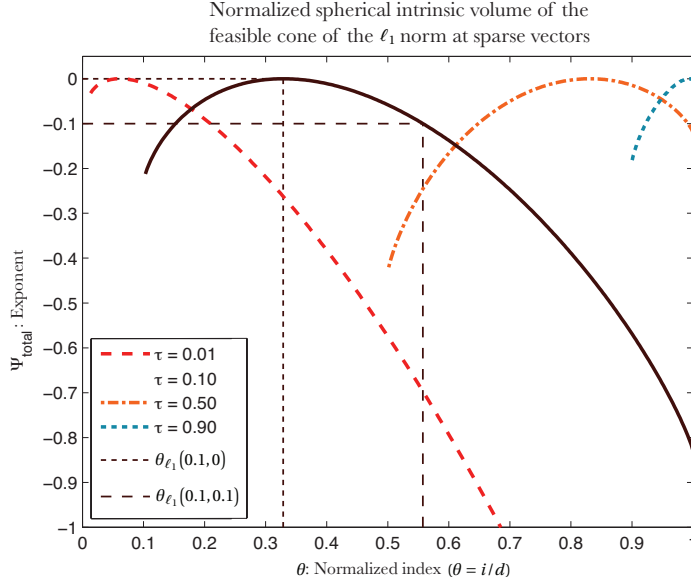
**Figure 13** *Upper bounds for the exponent of the spherical intrinsic volumes.* We plot $\Psi_{\text{total}}(\tau,\cdot)$ for several different values of $\tau$. The best upper decay threshold at level $\psi$ is given by the rightmost $\theta$ for which the curve intersects the horizontal line at $-\psi$. The short dashes show the upper decay threshold $\theta_{\ell_1}(0.1,0)$, while the long dashes show $\theta_{\ell_1}(0.1,0.1)$, defined in Equation (C.14). For each $\tau$, the upper decay threshold at level zero is numerically equal to the lower decay threshold.

where the exponent $\Psi_{\text{cont}}$ for the number of *cont*aining faces is defined by

$$\Psi_{\text{cont}}(\theta,\tau) := (\theta - \tau)\log(2) + (1-\theta)H\left(\frac{\theta-\tau}{1-\tau}\right). \tag{C.8}$$

The function $H(\theta)$ is the entropy defined by (3.6). Equation (C.1) follows by defining

$$\Psi_{\text{total}}(\theta,\tau) := \Psi_{\text{cont}}(\theta,\tau) - \Psi_{\text{int}}(\theta,\tau) - \Psi_{\text{ext}}(\theta) \tag{C.9}$$

and taking logarithms in (C.5). This is the first claim.

We now show that, for any $\theta < \tau$, Equation (C.2) holds for all sufficiently large $d$. Since $\boldsymbol{x}/\|\boldsymbol{x}\|_{\ell_1}$ lies in a $(k-1)$-dimensional face of the crosspolytope $C$, every face of $\mathscr{F}(\|\cdot\|_{\ell_1},\boldsymbol{x})$ has dimension at least $(k-1)$. It follows immediately from Definition 3.1 that

$$v_i\big(\mathscr{F}(\|\cdot\|_{\ell_1},\boldsymbol{x})\big) = 0$$

for all $i < k-1$. Since $k = \lceil \tau d \rceil$, we see that (C.2) holds for all sufficiently large $d$ so long as $\theta < \tau$. This is the second claim. $\qquad\square$

## C.1 Computing the exponents

We now define the functions needed to compute $\Psi_{\text{total}}$ in (C.9). Recall that $\Psi_{\text{cont}}$ is explicitly defined in (C.8). The functions $\Psi_{\text{int}}$ and $\Psi_{\text{ext}}$ are defined in [**25**], but we recapitulate their definitions for completeness. Define implicit parameters $x = x(\theta)$ and $s = s(\theta,\tau)$ as the solutions to the equations

$$\frac{2xG(x)}{G'(x)} = \frac{1-\theta}{\theta}, \tag{C.10}$$

$$M(s) = 1 - \frac{\tau}{\theta}, \tag{C.11}$$

44

where $G(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ is the error function $\text{erf}(x)$ and $M(s)$ is a variant of the Mills' ratio given by

$$M(s) = -s e^{s^2/2} \int_{-\infty}^s e^{-t^2/2} dt = -s\sqrt{\frac{\pi}{2}} \, \text{erfcx}\left(-\frac{s}{\sqrt{2}}\right),$$

where $\text{erfcx}(s) = e^{s^2} \text{erfc}(s)$ is the scaled complementary error function. This second form for $M(s)$ is convenient for numerical computations. It follows from [25] that $x$ and $s$ are well defined. Numerical evaluation of $x(\theta)$ and $s(\theta, \tau)$ is straightforward using, for example, bisection methods.

With these parameters in hand, the exponent for the internal angle is given by[‡]

$$\Psi_{\text{int}}(\theta, \tau) := (\theta - \tau) \log\left(\sqrt{2\pi} \frac{s\theta}{\tau - \theta}\right) - \frac{\tau s^2}{2}, \tag{C.12}$$

where $s = s(\theta, \tau)$ satisfies equation (C.11). The exponent for the external angle is

$$\Psi_{\text{ext}}(\theta) := -(1 - \theta) \log\big(G(x)\big) + \theta x^2, \tag{C.13}$$

where $x = x(\theta)$ is given by (C.10) above.

Figure 13 displays $\Psi_{\text{total}}(\cdot, \tau)$ for a few values of the parameter $\tau$. Empirically, it appears that $\Psi_{\text{total}}(\cdot, \tau)$ is concave for every value of $\tau \in [0, 1]$ and has a unique maximal value of zero.

## C.2 Defining decay thresholds

The exponent $\Psi_{\text{total}}$ provides decay thresholds for the $\ell_1$ norm at proportionally sparse vectors. We define

$$\theta_{\ell_1}(\tau, \psi) := \inf\big\{\theta_\star \in [0, 1] : \Psi_{\text{total}}(\theta, \tau) < -\psi \text{ for all } \theta \in (\theta_\star, 1]\big\}. \tag{C.14}$$

In words, $\theta_{\ell_1}(\tau, \psi)$ is the rightmost point of intersection of the curve $\Psi_{\text{total}}(\cdot, \tau)$ with the horizontal line at the level $-\psi$ (see Figure 13). Further define

$$\kappa_{\ell_1}(\tau) := \sup\big\{\kappa_\star \in [0, 1] : \Psi_{\text{total}}(\theta, \tau) < 0 \text{ for all } \kappa \in [0, \kappa_\star)\big\}. \tag{C.15}$$

This function $\kappa_{\ell_1}(\tau)$ is the leftmost point of intersection of $\Psi_{\text{total}}(\cdot, \tau)$ with the horizontal line at level zero. Equations (C.14) and (C.15) define decay thresholds for the ensemble of feasible cones for the $\ell_1$ norm at proportionally sparse vectors.

**Proposition C.3** (Decay thresholds for the $\ell_1$ norm at proportionally sparse vectors)**.** *Consider the ensemble of Lemma C.1. The function $\theta_{\ell_1}(\tau, \psi)$ is an upper decay threshold at level $\psi$ for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0^{(d)}) : d \in \mathscr{D}\}$, while $\kappa_{\ell_1}(\tau)$ is a lower decay threshold for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0^{(d)}) : d \in \mathscr{D}\}$.*

*Proof.* By definition, $\Psi_{\text{total}}(\theta, \tau) < -\psi$ for every $\theta > \theta_{\ell_1}(\tau, \psi)$. It then follows immediately from Lemma C.1 and Definition 4.7 that $\theta_{\ell_1}(\tau, \psi)$ is an upper decay threshold at level $\psi$ for the ensemble $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)}) : d \in \mathscr{D}\}$. The proof that $\kappa_{\ell_1}(\tau)$ is a lower decay threshold for $\{\mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^{(d)}) : d \in \mathscr{D}\}$ is equally straightforward, so we omit the argument. $\square$

We abbreviate the upper decay threshold at level zero by

$$\theta_{\ell_1}(\tau) := \theta_{\ell_1}(\tau, 0) \tag{C.16}$$

for consistency with Definition 3.9.

Figure 13 illustrates the definition of $\theta_{\ell_1}(\tau, \psi)$. Numerically, it appears that zero is the *unique* maximal value of $\Psi_{\text{total}}(\tau, \cdot)$ for every value $\tau \in (0, 1]$. If this is indeed the case, then we would be able to deduce that $\theta_{\ell_1}(\tau) = \kappa_{\ell_1}(\tau)$ for all values of sparsity $\tau$.

---

[‡]Equation (C.12) requires a significant amount of wholly uninteresting algebraic simplification from the formulas of [25]. The key steps in this simplification follow from the equations on page 638 of the reference. In particular, we write $y$ explicitly in terms of $s$ with their Eq. (6.12), and then write $\xi$ explicitly in terms of $s$ using this expression for $y$—see equation (6.13) in the reference. Noting that the reference defines $\gamma = \frac{\tau}{\theta}$ on page 631 gives (C.12), modulo trivial simplifications.

### C.3 Reconciliation with [25]

We now discuss the relationship between our spherical intrinsic volume approach and the bounds of [25], [28] for basis pursuit. Numerically, it appears that the two approaches provide equivalent success guarantees, but the expressions for the exponents seem to preclude a direct proof of equivalence. We also describe how our approach gives matching upper bounds for region of success of basis pursuit, which shows that our results are the best possible up to numerical accuracy.

#### C.3.1 Reconciliation with the weak threshold

Recall that basis pursuit is the linear inverse problem (5.1) with objective $f(\cdot) = \|\cdot\|_{\ell_1}$. By the first part of Lemma 5.1, basis pursuit with a Gaussian measurement matrix $\boldsymbol{\Omega} \in \mathbb{R}^{\lceil \sigma d \rceil \times d}$ succeeds at recovering a $k = \lceil \tau d \rceil$-sparse vector with overwhelming probability in high dimensions, so long as the pair $(\tau, \sigma)$ satisfies $\theta_{\ell_1}(\tau) < \sigma$.

We now describe the analogous result given in [25, Sec. 7.1]. Define the critical sparsity ratio (compare with the critical proportion [25, Def. 2])

$$\tau_W(\sigma) = \sup\{\hat{\tau} \in [0,\sigma] : \Psi_{\text{total}}(\theta, \hat{\tau}) < 0 \text{ for all } \theta \in [\sigma, 1]\}. \tag{C.17}$$

Then the result [25, Thm. 2] is equivalent to the statement that basis pursuit with a Gaussian matrix $\boldsymbol{\Omega} \in \mathbb{R}^{\lceil \sigma d \rceil \times d}$ succeeds with overwhelming probability in high dimensions whenever the pair $(\tau, \sigma)$ satisfies $\tau < \tau_W(\sigma)$.

These two approaches show strong similarities, and methods provide the same results to numerical precision. Indeed, under the assumption that both $\tau_W(\sigma)$ and $\theta_{\ell_1}(\tau)$ are monotonically increasing functions (this appears to hold empirically), one can show that these approaches are equivalent. Rather than dwell on this fine detail, we present a matching failure region for basis pursuit.

#### C.3.2 Matching upper bound

The following result shows links the lower decay threshold to regions where basis pursuit fails.

**Proposition C.4.** *Suppose $\kappa_{\ell_1}(\tau) > \sigma$. Then basis pursuit with $n = \lceil \sigma d \rceil$ Gaussian measurements fails with overwhelming probability in high dimensions for the $\tau$-sparse ensemble of Lemma C.1.*

Since the function $\Psi_{\text{total}}(\cdot, \tau)$ has a unique maximal value of zero up to our ability to compute the functions involved (see Figure 13), we have the equality $\kappa_{\ell_1}(\tau) = \theta_{\ell_1}(\tau)$ to numerical precision. Coupling Proposition C.4 with our discussion in Section C.3.1 reveals that basis pursuit with a Gaussian measurement matrix exhibits a phase transition between success and failure at $\sigma = \theta_{\ell_1}(\tau)$.

*Proof of Proposition C.4.* Let $\{\boldsymbol{x}_0^{(d)} \in \mathbb{R}^d\}$ be an ensemble of $\tau$-sparse vectors as in Lemma C.1. The null space of an $n \times d$ Gaussian matrix is distributed as $\boldsymbol{Q}L$, where $L$ is a linear subspace of dimension $(d - n)$. It then follows from [15, Prop. 2.1] that basis pursuit with $n = \lceil \sigma d \rceil$ Gaussian measurements succeeds with the same probability that $\boldsymbol{Q}L \cap \mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0) = \{\boldsymbol{0}\}$.

By Proposition 3.11, the value $\kappa_\star = (1 - \sigma)$ is a lower decay threshold for $L$. Our assumption implies $\kappa_{\ell_1}(\tau) + \kappa_\star > 1$, so by Theorem B.1, we see $\boldsymbol{Q}L \cap \mathscr{F}(\|\cdot\|_{\ell_1}, \boldsymbol{x}_0) \neq \{\boldsymbol{0}\}$ with overwhelming probability in high dimensions. We conclude that basis pursuit fails with overwhelming probability in high dimensions. $\square$

## D  Proof of Lemma 5.1 and Corollary 5.3

We begin with the proof of the two claims of Lemma 5.1 concerning the relationship between the upper decay threshold and the linear inverse problem (5.1). The first result is a corollary of [15, Prop. 2.1]. We drop the superscript $d$ for clarity.

*Proof of Lemma 5.1, Part 1.* Let $\boldsymbol{\Omega}$ be the $n \times d$ Gaussian measurement matrix, where $n = \lceil \sigma d \rceil$. The null space of $\boldsymbol{\Omega}$ is distributed as $\boldsymbol{Q}L$, where $\boldsymbol{Q}$ is a random basis, and $L$ is any fixed $(d - n)$-dimensional subspace of $\mathbb{R}^d$. Therefore, the probability that (5.1) succeeds is equal to the probability that $\boldsymbol{Q}L \cap \mathscr{F}(f, \boldsymbol{x}_0) \neq \{\boldsymbol{0}\}$ [**15,** Prop. 2.1]. By Proposition 3.11, the subspace $L$ has an upper decay threshold $\theta_L = 1 - \sigma$, so that the assumption $\theta_\star < \sigma$ implies that $\theta_\star + \theta_L < 1$. The claim follows from Theorem 4.4. $\square$

The second claim of Lemma 5.1 requires additional effort. We require the following technical lemma.

**Lemma D.1.** *Let $\mathscr{D}$ be an infinite set of indices. Let $\{K^{(d)} : d \in \mathscr{D}\}$ be an ensemble of closed convex cones with $K^{(d)} \subset \mathbb{R}^d$ for each $d$, and let $\{L^{(d)} : d \in \mathscr{D}\}$ be an ensemble of linear subspaces of $\mathbb{R}^d$ of dimension $d - \lceil \sigma d \rceil$. If there exists an $\varepsilon > 0$ such that for every sufficiently large $d$,*

$$\mathbb{P}\{K^{(d)} \cap \boldsymbol{Q}L^{(d)} \neq \{\boldsymbol{0}\}\} \leq e^{-\varepsilon d},$$

*then $\{K^{(d)}\}$ has an upper decay threshold $\theta_\star = \sigma$.*

Again, the spherical kinematic formula (3.2) is at the heart of the proof.

*Proof of Lemma D.1.* We split the argument into two cases: first, we consider the case where $K^{(d)}$ is not a subspace, and then consider the case where $K^{(d)}$ is a subspace. The general mixed-cone case follows by applying these arguments to the subsequences consisting of only one type of cone.

We drop the superscript $d$ for clarity. Let $n = \lceil \sigma d \rceil$. We first assume that $K$ is not a subspace. By the spherical kinematic formula, the probability of interest $P$ is given by

$$P := \mathbb{P}\{K \cap \boldsymbol{Q}L \neq \{\boldsymbol{0}\}\} = \sum_{k=0}^{d-1} (1 + (-1)^k) \sum_{i=k}^{d-1} v_i(L) v_{d-1-i+k}(K).$$

By Proposition 3.2, $v_i(L) = \delta_{i, d-n-1}$, so by replacing $k$ with $k - n$, the probability above reduces to

$$P = \sum_{k=n}^{d-1} (1 + (-1)^{k-n}) v_k(K). \tag{D.1}$$

By assumption, we have $P \leq e^{-\varepsilon d}$ for all sufficiently large $d$, so the positivity of spherical intrinsic volumes (Fact 3.5.1) implies

$$v_k(K) \leq e^{-\varepsilon d}, \text{ for any } k \geq n \text{ such that } k \equiv n \mod 2.$$

It requires an additional geometric observation to remove the dependence on parity. Let $\tilde{L}$ be a $(d - n - 1)$-dimensional subspace contained in $L$. By containment, it is immediate that

$$\tilde{P} := \mathbb{P}\{K \cap \boldsymbol{Q}\tilde{L} \neq \{\boldsymbol{0}\}\} \leq \mathbb{P}\{K \cap \boldsymbol{Q}L \neq \{\boldsymbol{0}\}\} \leq e^{-\varepsilon d},$$

where the last inequality is by assumption. But the same manipulations as before show

$$\tilde{P} = \sum_{k=n+1}^{d-1} (1 + (-1)^{k-n-1}) v_k(K)$$

so we have $v_k(K) \leq e^{-\varepsilon d}$ for every $k \geq n + 1$ such that $k \equiv n + 1 \mod 2$. In summary, for every $d$ sufficiently large and any $k \geq n = \lceil \sigma d \rceil$, we have $v_k(K) \leq e^{-\varepsilon d}$. By definition, $\sigma$ is an upper decay threshold for the ensemble of cones $K$. This completes the first case.

Now suppose that $K$ is a subspace and define $m := \dim(K)$. Since $\dim(L) = d - \lceil \sigma d \rceil$, we have

$$P := \mathbb{P}\{K \cap \boldsymbol{Q}L \neq \{\boldsymbol{0}\}\} = \begin{cases} 0, & m < \lceil \sigma d \rceil \\ 1, & \text{otherwise} \end{cases}$$

because randomly oriented subspaces are almost always in general position. The assumption that $P \leq e^{-\varepsilon d}$ for all sufficiently large $d$ requires that $m < \lceil \sigma d \rceil$ for all sufficiently large $d$. By Proposition 3.11, the scalar $\sigma$ is an upper decay threshold for subspace $K$. This is the result for the second case, so we are done. $\square$

47

*Proof of Lemma 5.1, Part 2.* The results of [**15**] imply that the linear inverse problem (5.1) with a Gaussian measurement matrix $\boldsymbol{\Omega}$ succeeds with the same probability that a randomly oriented $(n-d)$-dimensional subspace $\boldsymbol{Q}L$ strikes the feasible cone $\mathscr{F}(f,\boldsymbol{x}_0)$ trivially. The result then follows from Lemma D.1. □

We conclude with the proof of Corollary 5.3, which asserts that $\theta_\star$ is an upper decay threshold for the ensemble $\{\mathscr{F}(f^{(d)},\boldsymbol{x}_0^{(d)}) : d \in \mathscr{D}\}$ whenever the bound (5.3) holds.

*Proof of Corollary 5.3.* To shorten notation, we drop the explicit dependence on $d$, and we define the width $W := W\big(\mathscr{F}(f,\boldsymbol{x}_0) \cap \mathsf{S}^{d-1}\big)$. The result[§] [**15,** Cor. 3.3(1)] states that for any $\varepsilon > 0$,

$$n \ge \big(W + \varepsilon\sqrt{d}\big)^2 + 1 \implies \text{(5.2) fails with probability} \le \mathrm{e}^{-\varepsilon^2 d/2}. \tag{D.2}$$

Fix $\varepsilon \in (0,1)$ and define $\theta_\varepsilon := \theta_\star + 2\varepsilon(\theta_\star^{1/2} + 1)$. We claim that $\theta_\varepsilon$ is an upper decay threshold for the ensemble $\{\mathscr{F}(f,\boldsymbol{x}_0)\}$. To see this, choose the number of measurements $n = \lceil \theta_\varepsilon d \rceil$ in the linear inverse problem (5.2). Then for $d \ge \varepsilon^{-1}$, we have

$$n \ge \theta_\varepsilon d \ge \big[\theta_\star + \varepsilon(2\theta_\star^{1/2} + 1)\big]d + 1$$

by our choice of $n$. We bound the bracketed expression using convexity of the map $\varepsilon \mapsto \big(\theta_\star^{1/2} + \varepsilon\big)^2$:

$$\theta_\star + \varepsilon(2\theta_\star^{1/2} + 1) = (1-\varepsilon)\theta_\star + \varepsilon\big(\theta_\star^{1/2} + 1\big)^2 \ge \big(\theta_\star^{1/2} + \varepsilon\big)^2$$

because $\varepsilon \in (0,1)$ and a convex function lies below the chord connecting its endpoints. Combining the two displayed equations above,

$$n - 1 \ge \big(\theta_\star^{1/2} + \varepsilon\big)^2 d \ge \big(W + \varepsilon\sqrt{d}\big)^2$$

where the second inequality holds for all sufficiently large $d$ by assumption (5.3). The implication (D.2) thus implies that the linear inverse problem (5.2) succeeds with overwhelming probability in high dimensions when $n = \lceil \theta_\varepsilon d \rceil$. From the second part of Lemma 5.1, we conclude that $\theta_\varepsilon$ is an upper decay threshold for $\{\mathscr{F}(f,\boldsymbol{x}_0)\}$, as claimed. The proof is completed by taking $\varepsilon \to 0$ and verifying that a limit of decay thresholds is itself a decay threshold. We omit this straightforward, but technical, argument. □

# References

[1] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio inpainting. *IEEE Trans. Audio, Speech, and Lang. Process.*, 20(3):922–932, Mar. 2012.

[2] F. Affentranger and R. Schneider. Random projections of regular simplices. *Discrete Comput. Geom.*, 7(1):219–226, 1992.

[3] D. Amelunxen. *Geometric analysis of the condition of the convex feasibility problem.* Dissertation, Universität Paderborn, 2011.

[4] D. Amelunxen and P. Bürgisser. Intrinsic volumes of symmetric cones. *preprint*, 2012. arXiv:1205.1863.

[5] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *preprint*, March 2013. Available at arXiv:1303.6672.

[6] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *preprint*, July 2012. arXiv:1207.7321.

[7] J. Bobin, Y. Moudden, and J.-L. Starck. Morphological diversity and source separation. *IEEE Trans. Signal Process.*, 13(7):409–412, 2006.

[8] J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden. Sparsity and morphological diversity in blind source separation. *IEEE Trans. Image Process.*, 16(11):2662–2674, Nov. 2007.

[9] K. Böröczky, Jr. and M. Henk. Random projections of regular polytopes. *Arch. Math. (Basel)*, 73(6):465–473, Dec. 1999.

---

[§]This result is a corollary of a Gaussian process inequality due to Gordon [**40**], [**41**].

[10] P. Bürgisser and D. Amelunxen. Robust smoothed analysis of a condition number for linear programming. *Math. Program.*, Apr. 2010.

[11] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59–73, 2011.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. Assoc. Comput. Mach.*, 58(3):1–37, May 2011.

[13] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[14] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613, Oct. 2010.

[15] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[16] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *SYSID 2009*, Saint-Malo, France, July 2009.

[17] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim*, 21(2):572–596, 2011.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[19] J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

[20] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Process.*, 53(7):2477–2488, 2005.

[21] R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5(2-3):173–187, 1996.

[22] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.5 of 2012-10-01. Online companion to [**59**].

[23] D. L. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report, Stanford University, 2004.

[24] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[25] D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.*, 35(4):617–652, Dec. 2006.

[26] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, Aug. 2001.

[27] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, June 1989.

[28] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452–7, July 2005.

[29] D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 22(1):1–53, 2009.

[30] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*, 367(1906):4273–4293, 2009.

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

[32] D. L. Donoho and J. Tanner. Exponential bounds implying construction of compressed sensing matrices, error-correcting codes, and neighborly polytopes by random sampling. *IEEE Trans. Inform. Theory*, 56(4):2002–2016, 2010.

[33] D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proc. IEEE*, 98(6):913–924, June 2010.

[34] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, 2007.

[35] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Appl. Comput. Harmon. Anal.*, 19(3):340–358, Nov. 2005.

[36] M. Fazel. *Matrix rank minimization with applications*. Dissertation, Stanford University, Stanford, CA, 2002.

[37] R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *preprint*, 2013. arXiv:1305.2524.

[38] S. Glasauer. *Integralgeometrie konvexer Korper im spharischen Raum*. Dissertation, University of Freiburg, 1995.

[39] S. Glasauer. Integral geometry of spherically convex bodies. *Diss. Summaries Math.*, 1:219–226, 1996.

[40] Y. Gordon. Elliptically contoured distributions. *Probab. Theory Related Fields*, 76(4):429–438, 1987.

[41] 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 (GAFA), 1986-87*, page 84. Springer, 1988.

[42] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, London, 2008.

[43] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, October 2010.

[44] B. Grünbaum. *Convex polytopes*, volume XVI of *Pure and Applied Mathematics*. Wiley, London, 1967.

[45] B. Grünbaum. Grassmann angles of convex polytopes. *Acta Math.*, 121(1):293–302, Dec. 1968.

[46] C. Hegde and R. G. Baraniuk. Signal recovery on incoherent manifolds. *IEEE Trans. Inform. Theory*, 58(12):7204–7214, December 2012.

[47] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 964–972. NIPS, 2010.

[48] A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. Improved sparse recovery thresholds with two-step reweighted $\ell_1$ minimization. In *IEEE Int. Symp. Inform. Theory Proc. (ISIT)*, pages 1603–1607, Austin, TX, 2010.

[49] A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi. Analyzing weighted $\ell_1$ minimization for sparse recovery with nonuniform sparse models. *IEEE Trans. Signal Process.*, 59(5):1985–2001, 2011.

[50] V. L. Klee, Jr. Separation properties of convex cones. *Proc. Amer. Math. Soc.*, 6(2):313– 318, 1955.

[51] O. L. Mangasarian and B. Recht. Probability of unique integer solution to a system of linear equations. *European J. Oper. Res.*, 214(1):27–30, Oct. 2011.

[52] M. B. McCoy and J. A. Tropp. Two proposals for robust PCA using semidefinite programming. *Elec. J. Statist.*, 5:1123–1160, 2011.

[53] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. *preprint*, 2013. Submitted to *Math. Program.*. Available at arXiv:1309.7478.

[54] P. McMullen. Non-linear angle-sum relations for polyhedral cones and polytopes. *Math. Proc. Cambridge Philos. Soc.*, 78(02):247, Oct. 1975.

[55] F. Mezzadri. How to generate random matrices from the classical compact groups. *Notices Amer. Math. Soc.*, 54(5):592–604, 2007.

[56] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1348–1356. NIPS, 2009.

[57] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012.

[58] N. H. Nguyen and T. D. Tran. Exact recoverability from dense corrupted observations via $\ell_1$-minimization. *IEEE Trans. Inform. Theory*, 59(4):2017–2035, 2013.

[59] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010. Print companion to [**22**].

[60] S. Oymak and B. Hassibi. New null space results and recovery thresholds for matrix rank minimization. *preprint*, 2010. arXiv:1011.6326.

[61] S. Oymak and B. Hassibi. Asymptotically exact denoising in relation to compressed sensing. *preprint*, May 2013. arXiv:1305.2714.

[62] G. Pataki. The geometry of semidefinite programming. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *The Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 29–65. Kluwer, Boston, 2000.

[63] G. Pope, A. Bracher, and C. Studer. Probabilistic recovery guarantees for sparsely corrupted signals. *IEEE Trans. Inform. Theory*, 59(5):3104–3116, 2013.

[64] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[65] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1997.

[66] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1998.

[67] H. Ruben. On the geometrical moments of skew-regular simplices in hyperspherical space, with some applications in geometry and mathematical statistics. *Acta Math.*, 103:1–23, 1960.

[68] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330, 1986.

[69] R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Probability and its Applications. Springer Verlag, 2008.

[70] J.-L. Starck, D. L. Donoho, and E. J. Candès. Astronomical image representation by the curvelet transform. *Astronom. Astrophys.*, 398(2):785–800, 2003.

[71] J.-L. Starck, M. Elad, and D. L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. Image Process.*, 14(10):1570–1582, Oct. 2005.

[72] M. Stojnic. Strong thresholds for $\ell_1/\ell_2$-optimization in block-sparse compressed sensing. In *ICASSP 2009.*, pages 3025–3028, 2009.

[73] M. Stojnic. Various thresholds for $\ell_1$-optimization in compressed sensing. *preprint*, 2009. arXiv:0907.3666.

[74] M. Stojnic. A framework to characterize performance of lasso algorithms. *preprint*, March 2013. arXiv:1303.7291.

[75] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei. Recovery of sparsely corrupted signals. *IEEE Trans. Inf. Theory*, pages 3115–3130, May 2012.

[76] G. Tang, B. N. Bhaskar, and B. Recht. Near minimax line spectral estimation. *preprint*, 2013. arXiv:1303.4348.

[77] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the l1 norm. *Geophysics*, 44(1):39, 1979.

[78] V. N. Temlyakov. Nonlinear methods of approximation. *Found. Comput. Math.*, 3(1):33–107, 2003.

[79] J. A. Tropp. On the linear independence of spikes and sines. *J. Fourier Anal. Appl*, 14:838–858, 2008.

[80] A. M. Vershik and P. V. Sporyshev. An asymptotic estimate for the average number of steps in the parametric simplex method. *USSR Comput. Maths. Math. Phys.*, 26(3):104–113, 1986.

[81] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[82] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[83] J. Wright, A. Ganesh, K. Min, and Y. Ma. Compressive principal component pursuit. *Information and Inference*, 2(1):32–68, 2013.

[84] J. Wright and Y. Ma. Dense error correction via $\ell^1$-minimization. *IEEE Trans. Inform. Theory*, 56(7):3540–3560, 2010.

[85] A. D. Wyner. An analog scrambling scheme which does not expand bandwidth, part I: Discrete time. *IEEE Trans. Inform. Theory*, 25(3):261–274, May 1979.

[86] A. D. Wyner. An analog scrambling scheme which does not expand bandwidth, part II: Continuous time. *IEEE Trans. Inform. Theory*, 25(4):415–425, 1979.

[87] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2496–2504. NIPS, 2010.

[88] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Inform. Theory*, 58(5):1–24, 2012.

[89] W. Xu and B. Hassibi. Precise stability phase transitions for $\ell_1$ minimization: A unified geometric framework. *IEEE Trans. Inform. Theory*, 57(10):6894–6919, Oct. 2011.

[90] W. Xu, A. Khajehnejad, A. S. Avestimehr, and B. Hassibi. Breaking through the thresholds: an analysis for iterative reweighted $\ell_1$ minimization via the Grassmann angle framework. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5498–5501, Mar. 2010.