

Simple Bounds for Recovering Low-complexity Models

Emmanuel Candès* and Benjamin Recht†

June 2011; revised February, 2012

Abstract

This note presents a unified analysis of the recovery of simple objects from random linear measurements. When the linear functionals are Gaussian, we show that an s -sparse vector in \mathbb{R}^n can be efficiently recovered from $2s \log n$ measurements with high probability and a rank r , $n \times n$ matrix can be efficiently recovered from $r(6n - 5r)$ measurements with high probability. For sparse vectors, this is within an additive factor of the best known nonasymptotic bounds. For low-rank matrices, this matches the best known bounds. We present a parallel analysis for block-sparse vectors obtaining similarly tight bounds. In the case of sparse and block-sparse signals, we additionally demonstrate that our bounds are only slightly weakened when the measurement map is a random sign matrix. Our results are based on analyzing a particular dual point which certifies optimality conditions of the respective convex programming problem. Our calculations rely only on standard large deviation inequalities and our analysis is self-contained.

Keywords. ℓ_1 -norm minimization nuclear-norm minimization block-sparsity duality random matrices.

1 Introduction

The past decade has witnessed a revolution in convex optimization algorithms for recovering structured models from highly incomplete information. Work in compressed sensing has shown that when a vector is *sparse*, then it can be reconstructed from a number of nonadaptive linear measurements proportional to a logarithmic factor times the signal's sparsity level [4, 8]. Building on this work, many have recently demonstrated that if an array of user data has low-rank, then the matrix can be re-assembled from a sampling of information proportional to the number of parameters required to specify a low-rank factorization. See [2, 3, 19] for some early references on this topic.

Sometimes, one would like to know precisely how many measurements are needed to recover an s -sparse vector (a vector with at most s nonzero entries) by ℓ_1 minimization or a rank- r matrix by nuclear-norm minimization. This of course depends on the kind of measurements one is allowed to take, and can be empirically determined or approximated by means of numerical studies. At the theoretical level, however, very precise answers—e.g., perfect knowledge of numerical constants—for models of general interest may be very hard to obtain. For instance, in [4], the authors demonstrated that about $20s \log n$ randomly selected Fourier coefficients were sufficient to recover an s -sparse signal, but determining the minimum number that would suffice appears to be a very difficult question. Likewise, obtaining precise theoretical knowledge about the number of randomly selected entries required to recover a rank- r matrix by convex programming seems delicate, to say the least.

*Mathematics and Statistics Departments, Stanford University, Stanford, CA 94305. candes@stanford.edu

†Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, 53706. brecht@cs.wisc.edu

For some special and idealized models, however, this is far easier and the purpose of this note is to make this clear.

In this note, we demonstrate that many bounds concerning Gaussian measurements can be derived via elementary, direct methods using Lagrangian duality. By a careful analysis of a particular Lagrange multiplier, we are able to prove that $2s \log n$ measurements are sufficient to recover an s -sparse vector in \mathbb{R}^n and $r(6n - 5r)$ measurements are sufficient to recover a rank r , $n \times n$ matrix with high probability. These almost match the best-known, non-asymptotic bounds for sparse vector reconstruction ($2s \log(n/s) + 5/4s$ measurements [5, 7]), and match the best known bounds for low-rank matrix recovery in the nuclear norm (as reported in [5, 16]).

The work [5], cited above, presents a unified view of the convex programming approach to inverse problems and provides a relatively simple framework to derive exact, robust recovery bounds for a variety of simple models. As we already mentioned, the authors also provide rather tight bounds on sparse vector and low-rank matrix recovery in the *Gaussian measurement ensemble* by using a deep theorem in functional analysis due to Gordon, which concerns the intersection of random subspaces with subsets of the sphere [11]. Gordon's Theorem has also been used to provide sharp estimates of the *phase transitions* for the ℓ_1 and nuclear norm heuristics in [20] and [16] respectively. Our work complements these results, demonstrating that the dual multiplier ansatz proposed in [10] can also yield very tight bounds for many signal recovery problems.

To introduce our results, suppose we are given information about an object $\mathbf{x}_0 \in \mathbb{R}^n$ of the form $\Phi \mathbf{x}_0 \in \mathbb{R}^m$ where Φ is an $m \times n$ matrix. When Φ has entries i.i.d. sampled from a Gaussian distribution with mean 0 and variance $1/m$, we call it a *Gaussian measurement map*. We want bounds on the number of rows m of Φ to ensure that \mathbf{x}_0 is the unique minimizer of the problem

$$\begin{aligned} & \text{minimize} && \|\mathbf{x}\|_{\mathcal{A}} \\ & \text{subject to} && \Phi \mathbf{x} = \Phi \mathbf{x}_0. \end{aligned} \tag{1.1}$$

Here $\|\cdot\|_{\mathcal{A}}$ is a norm with some suitable properties which encourage solutions which conform to some notion of simplicity. Our first result is the following

Theorem 1.1 *Let \mathbf{x}_0 be an arbitrary s -sparse vector and $\|\cdot\|_{\mathcal{A}}$ be the ℓ_1 norm. Let $\beta > 1$.*

- *For Gaussian measurement maps Φ with $m \geq 2\beta s \log n + s$, the recovery is exact with probability at least $1 - 2n^{-f(\beta, s)}$ where*

$$f(\beta, s) = \left[\sqrt{\frac{\beta}{2s} + \beta - 1} - \sqrt{\frac{\beta}{2s}} \right]^2.$$

- *Let $\epsilon \in (0, 1)$. For binary measurement maps Φ with i.i.d. entries taking on values $\pm m^{-1/2}$ with equal probability, there exist numerical constants c_0 and c_1 such that if $n \geq \exp(c_0/\epsilon^2)$ and $m \geq 2\beta(1-\epsilon)^{-2}s \log n + s$, the recovery is exact with probability at least $1 - n^{1-\beta} - n^{-c_1\beta\epsilon^2}$.*

The algebraic expression $f(\beta, s)$ is positive for all $\beta > 1$ and $s > 0$. For all fixed $\beta > 1$, $f(\beta, s)$ is an increasing function of s so that $\min_{s \geq 1} f(\beta, s) = f(\beta, 1)$. Moreover, observe that $\lim_{s \rightarrow \infty} f(\beta, s) = \beta - 1$. For binary measurement maps, our result states that for any $\delta > 0$, $(2 + \delta)s \log n$ entries suffice to recover an s -sparse signal when n is sufficiently large. We also provide a very similar result for *block-sparse* signals, stated in Section 3.2.

Our third result concerns the recovery of a low-rank matrix.

Theorem 1.2 *Let \mathbf{X}_0 be an arbitrary $n_1 \times n_2$ rank- r -matrix and $\|\cdot\|_{\mathcal{A}}$ be the matrix nuclear norm. For a Gaussian measurement map Φ with $m \geq \beta r(3n_1 + 3n_2 - 5r)$ for some $\beta > 1$, the recovery is exact with probability at least $1 - 2e^{(1-\beta)n/8}$, where $n = \max(n_1, n_2)$.*

Our results are 1) nonasymptotic and 2) demonstrate sharp constants for sparse signal and low-rank matrix recovery, perhaps the two most important cases in the general model reconstruction framework. Further, our bounds are proven using elementary concepts from convex analysis and probability theory. In fact, the most elaborate result from probability that we employ concerns the largest singular value of a Gaussian random matrix, and this is only needed to analyze the rank minimization problem.

We show in Section 2 that the same construction and analysis can be applied to prove Theorems 1.1 and 1.2. The method, however, handles a variety of complexity regularizers including the ℓ_1/ℓ_2 -norm as well. When specialized in Section 3, we demonstrate sharp constants for exact model reconstruction in all three of these cases (ℓ_1 , ℓ_1/ℓ_2 and nuclear norms). We conclude the paper with a brief discussion of how to extend these results to other measurement ensembles. Indeed, with very minor modifications, we can achieve almost the same constants for subgaussian measurement ensembles in some settings such as sign matrices as reflected by the second part of Theorem 1.1.

2 Dual Multipliers and Decomposable Regularizers

Definition 2.1 *The dual norm is defined as*

$$\|\mathbf{x}\|_{\mathcal{A}}^* = \sup \{ \langle \mathbf{x}, \mathbf{a} \rangle : \|\mathbf{a}\|_{\mathcal{A}} \leq 1 \}. \quad (2.1)$$

A consequence of the definition is the well-known and useful dual-norm inequality

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_{\mathcal{A}} \|\mathbf{y}\|_{\mathcal{A}}^*. \quad (2.2)$$

The supremum in (2.1) is always achieved and thus the dual norm inequality (2.2) is tight in the sense that for any \mathbf{x} , there is a corresponding \mathbf{y} that achieves equality. Additionally, it is clear from the definition that the subdifferential of $\|\cdot\|_{\mathcal{A}}$ at \mathbf{x} is $\{\mathbf{v} : \langle \mathbf{v}, \mathbf{x} \rangle = \|\mathbf{x}\|_{\mathcal{A}}, \|\mathbf{v}\|_{\mathcal{A}}^* \leq 1\}$.

2.1 Decomposable Norms

We will restrict our attention to norms whose subdifferential has very special structure effectively penalizing “complex” solutions. In a similar spirit to [15], the following definition summarizes the necessary properties of a good complexity regularizer:

Definition 2.2 *A norm $\|\cdot\|_{\mathcal{A}}$ is decomposable at \mathbf{x}_0 if there is a subspace $T \subset \mathbb{R}^n$ and a vector $\mathbf{e} \in T$ such that the subdifferential at \mathbf{x}_0 has the form*

$$\partial\|\mathbf{x}_0\|_{\mathcal{A}} = \{ \mathbf{z} \in \mathbb{R}^n : \mathcal{P}_T(\mathbf{z}) = \mathbf{e} \text{ and } \|\mathcal{P}_{T^\perp}(\mathbf{z})\|_{\mathcal{A}}^* \leq 1 \}$$

and for any $\mathbf{w} \in T^\perp$, we have

$$\|\mathbf{w}\|_{\mathcal{A}} = \sup_{\substack{\mathbf{v} \in T^\perp \\ \|\mathbf{v}\|_{\mathcal{A}}^* \leq 1}} \langle \mathbf{v}, \mathbf{w} \rangle.$$

Above, \mathcal{P}_T (resp. \mathcal{P}_{T^\perp}) is the orthogonal projection onto T (resp. orthogonal complement of T).

When a norm is decomposable at \mathbf{x}_0 , the norm essentially penalizes elements in T^\perp independently from \mathbf{x}_0 . The most common decomposable regularizer is the ℓ_1 norm on \mathbb{R}^n . In this case, if \mathbf{x}_0 is an s -sparse vector, then T denotes the set of coordinates where \mathbf{x}_0 is nonzero and T^\perp the complement of T in $\{1, \dots, n\}$. We denote by $\mathbf{x}_{0,T}$ the restriction of \mathbf{x}_0 to T and by $\text{sgn}(\mathbf{x}_{0,T})$ the vector with ± 1 entries depending upon the signs of those of $\mathbf{x}_{0,T}$. The dual norm to the ℓ_1 norm is the ℓ_∞ norm. The subdifferential of the ℓ_1 norm at \mathbf{x}_0 is given by

$$\partial\|\mathbf{x}_0\|_1 = \{\mathbf{z} \in \mathbb{R}^n : \mathcal{P}_T(\mathbf{z}) = \text{sgn}(\mathbf{x}_{0,T}) \text{ and } \|\mathcal{P}_{T^\perp}(\mathbf{z})\|_\infty \leq 1\} .$$

That is, \mathbf{z} is equal to the sign of \mathbf{x}_0 on T and has entries with magnitudes bounded above by 1 on the orthogonal complement. As we will discuss in Section 3, the ℓ_1/ℓ_2 norm and the matrix nuclear norm are also decomposable. The following Lemma gives conditions under which \mathbf{x}_0 is the unique minimizer of (1.1).

Lemma 2.3 *Suppose that Φ is injective on the subspace T and that there exists a vector \mathbf{y} in the image of Φ^* (the adjoint of Φ) obeying*

1. $\mathcal{P}_T(\mathbf{y}) = \mathbf{e}$, where \mathbf{e} is as in Definition 2.2,
2. $\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^* < 1$.

Then \mathbf{x}_0 is the unique minimizer of (1.1).

Proof The proof is an adaptation from a standard argument. Consider any perturbation $\mathbf{x}_0 + \mathbf{h}$ where $\Phi\mathbf{h} = \mathbf{0}$. Since the norm is decomposable, there exists a $\mathbf{v} \in T^\perp$ such that $\|\mathbf{v}\|_{\mathcal{A}}^* \leq 1$ and $\langle \mathbf{v}, \mathcal{P}_{T^\perp}(\mathbf{h}) \rangle = \|\mathcal{P}_{T^\perp}(\mathbf{h})\|_{\mathcal{A}}$. Moreover, we have that $\mathbf{e} + \mathbf{v}$ is a subgradient of $\|\cdot\|_{\mathcal{A}}$ at \mathbf{x}_0 . Hence,

$$\begin{aligned} \|\mathbf{x}_0 + \mathbf{h}\|_{\mathcal{A}} &\geq \|\mathbf{x}_0\|_{\mathcal{A}} + \langle \mathbf{e} + \mathbf{v}, \mathbf{h} \rangle \\ &= \|\mathbf{x}_0\|_{\mathcal{A}} + \langle \mathbf{e} + \mathbf{v} - \mathbf{y}, \mathbf{h} \rangle \\ &= \|\mathbf{x}_0\|_{\mathcal{A}} + \langle \mathbf{v} - \mathcal{P}_{T^\perp}(\mathbf{y}), \mathcal{P}_{T^\perp}(\mathbf{h}) \rangle \\ &\geq \|\mathbf{x}_0\|_{\mathcal{A}} + (1 - \|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^*) \|\mathcal{P}_{T^\perp}(\mathbf{h})\|_{\mathcal{A}} . \end{aligned}$$

Since $\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^*$ is strictly less than one, this last inequality holds strictly unless $\mathcal{P}_{T^\perp}(\mathbf{h}) = 0$. But if $\mathcal{P}_{T^\perp}(\mathbf{h}) = 0$, then $\mathcal{P}_T(\mathbf{h})$ must also be zero because we have assumed that Φ is injective on T . This means that \mathbf{h} is zero proving that \mathbf{x}_0 is the unique minimizer of (1.1). \blacksquare

2.2 Constructing a Dual Multiplier

To construct a \mathbf{y} satisfying the conditions of Lemma 2.3, we follow the program developed in [10] and followed by many researchers in the compressed sensing literature. Namely, we choose the least squares solution of $\mathcal{P}_T(\Phi^*\mathbf{q}) = \mathbf{e}$, and then prove that $\mathbf{y} := \Phi^*\mathbf{q}$ has dual norm strictly less than 1 on T^\perp .

Let Φ_T and Φ_{T^\perp} denote the restriction of Φ to T and T^\perp respectively. Let d_T denote the dimension of the space T . Observe that if Φ_T is injective, then

$$\mathbf{q} = \Phi_T(\Phi_T^*\Phi_T)^{-1}\mathbf{e}, \tag{2.3}$$

$$\mathcal{P}_{T^\perp}(\mathbf{y}) = \Phi_{T^\perp}^*\mathbf{q}. \tag{2.4}$$

The key fact we use to derive our bounds in this note is that, when Φ is a Gaussian map, \mathbf{q} and $\Phi_{T^\perp}^*$ are independent, no matter what T is. This follows from the isotropy of the Gaussian ensemble. This property is also true in the sparse-signal recovery setting whenever the columns of Φ are independent. Another way to express the same idea is that given the value of \mathbf{q} , one can infer the distribution of $\mathcal{P}_{T^\perp}(\mathbf{y})$ with no knowledge of the values of the matrix Φ_T .

We assume in the remainder of this section that Φ is a Gaussian map. Conditioned on \mathbf{q} , $\mathcal{P}_{T^\perp}(\mathbf{y})$ is distributed as

$$\iota_{T^\perp} \mathbf{g},$$

where ι_{T^\perp} is an isometry from \mathbb{R}^{n-d_T} onto T^\perp and $\mathbf{g} \sim \mathcal{N}(0, \frac{\|\mathbf{q}\|_2^2}{m} \mathbf{I})$ (here and in the sequel, $\|\cdot\|_2$ is the ℓ_2 norm). Also, Φ_T is injective as long as $m \geq d_T$ and to bound the probability that the optimization problem (1.1) recovers \mathbf{x}_0 , we therefore only need to bound

$$\mathbb{P}[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^* \geq 1] \leq \mathbb{P}[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^* \geq 1 \mid \|\mathbf{q}\|_2 \leq \tau] + \mathbb{P}[\|\mathbf{q}\|_2 \geq \tau] \quad (2.5)$$

for some value of τ greater than 0. The first term in the upper bound will be analyzed on a case-by-case basis in Section 3. As we have remarked, once we have conditioned on \mathbf{q} , this term just requires us to analyze the large deviations of Gaussian random variables in the dual norm. What is more surprising is that the second term can be tightly upper bounded in a generic fashion for the Gaussian ensemble, independent of the regularizer under study.

To see this, observe that \mathbf{q} has squared norm

$$\|\mathbf{q}\|_2^2 = \langle \mathbf{e}, (\Phi_T^* \Phi_T)^{-1} \mathbf{e} \rangle.$$

By assumption, $(\Phi_T^* \Phi_T)^{-1}$ is a $d_T \times d_T$ inverse Wishart matrix with m degrees of freedom and covariance $m^{-1} \mathbf{I}_{d_T}$. Since the Gaussian distribution is isotropic, we have that $\|\mathbf{q}\|_2^2$ is distributed as $\|\mathbf{e}\|_2^2 m B_{11}$, where B_{11} is the first entry in the first column of an inverse Wishart matrix with m degrees of freedom and covariance \mathbf{I}_{d_T} .

To estimate the large deviations of $\|\mathbf{q}\|_2$, it thus suffices to understand the large deviations of B_{11} . A classical result in statistics states that B_{11} is distributed as an inverse chi-squared random variable with $m - d_T + 1$ degrees of freedom (see, [14, page 72] for example)¹. We can thus lean on tail bounds for the chi-squared distribution to control the magnitude of B_{11} . For each $t > 0$,

$$\begin{aligned} \mathbb{P} \left[\|\mathbf{q}\|_2 \geq \sqrt{\frac{m}{m - d_T + 1 - t}} \|\mathbf{e}\|_2 \right] &= \mathbb{P}[z \leq m - d_T + 1 - t] \\ &\leq \exp \left(-\frac{t^2}{4(m - d_T + 1)} \right). \end{aligned} \quad (2.6)$$

Here z is a chi-squared random variable with $m - d_T + 1$ degrees of freedom, and the final inequality follows from the standard tail bound for chi-square random variables (see, for example, [13]).

To summarize, we have proven the following

Proposition 2.4 *Let $\|\cdot\|_{\mathcal{A}}$ be a decomposable regularizer at \mathbf{x}_0 and let $t > 0$. Let \mathbf{q} and \mathbf{y} be defined as in (2.3) and (2.4). Then \mathbf{x}_0 is the unique optimal solution of (1.1) with probability at least*

$$1 - \mathbb{P} \left[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\mathcal{A}}^* \geq 1 \mid \|\mathbf{q}\|_2 \leq \sqrt{\frac{m}{m - d_T + 1 - t}} \|\mathbf{e}\|_2 \right] - \exp \left(-\frac{t^2/4}{m - d_T + 1} \right). \quad (2.7)$$

¹The reader not familiar with this result can verify with linear algebra that $1/B_{11}$ is equal to the squared distance between the first column of Φ_T and the linear space spanned by all the others. This squared distance is a chi-squared random variable with $m - d_T + 1$ degrees of freedom.

3 Bounds

Using Proposition 2.4, we can now derive non-asymptotic bounds for exact recovery of sparse vectors, block-sparse vectors, and low-rank matrices in a unified fashion.

3.1 Compressed Sensing in the Gaussian Ensemble

Let \mathbf{x}_0 be an s -sparse vector in \mathbb{R}^n . In this case, T denotes the set of coordinates where \mathbf{x}_0 is nonzero and T^\perp the complement of T in $\{1, \dots, n\}$. As previously discussed, the dual norm to the ℓ_1 norm is the ℓ_∞ norm and the subdifferential of the ℓ_1 norm at \mathbf{x}_0 is given by

$$\partial\|\mathbf{x}_0\|_1 = \{\mathbf{z} \in \mathbb{R}^n : \mathcal{P}_T(\mathbf{z}) = \text{sgn}(\mathbf{x}_{0,T}) \text{ and } \|\mathcal{P}_{T^\perp}(\mathbf{z})\|_\infty \leq 1\} .$$

Here, $\dim(T) = s$, the sparsity of \mathbf{x}_0 , and $\mathbf{e} = \text{sgn}(\mathbf{x}_0)$ so that $\|\mathbf{e}\|_2 = \sqrt{s}$.

For $m \geq s$, set \mathbf{q} and \mathbf{y} as in (2.3) and (2.4). To apply Proposition 2.4, we only need to estimate the probability that $\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_\infty$ exceeds 1 conditioned on the event that $\|\mathbf{q}\|_2$ is bounded. Conditioned on \mathbf{q} , the components of $\mathcal{P}_{T^\perp}(\mathbf{y})$ in T^\perp are i.i.d. $\mathcal{N}(0, \|\mathbf{q}\|_2^2/m)$. Hence, for any $\tau > 0$, the union bound gives

$$\begin{aligned} \mathbb{P}[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_\infty \geq 1 \mid \|\mathbf{q}\|_2 \leq \tau] &\leq (n-s) \mathbb{P}[|z| \geq \sqrt{m}/\tau] \\ &\leq n \exp\left(-\frac{m}{2\tau^2}\right), \end{aligned} \tag{3.1}$$

where $z \sim \mathcal{N}(0, 1)$. We have made use above of the elementary inequality $\mathbb{P}(|z| \geq t) \leq e^{-t^2/2}$ which holds for all $t \geq 0$. For $\beta > 1$, select

$$\tau = \sqrt{\frac{ms}{m-s+1-t}} \quad \text{with} \quad t = 2\beta \log(n) \left(\sqrt{1 + \frac{2s(\beta-1)}{\beta}} - 1 \right) .$$

Here, t is chosen to make the two exponential terms in our probability equal to each other. We can put all of the parameters together and plug (3.1) into (2.7). For $m = 2\beta s \log n + s$, $\beta > 1$, a bit of algebra gives the first part of Theorem 1.1.

3.2 Block-Sparsity in the Gaussian Ensemble

In simultaneous sparse estimation, signals are block-sparse in the sense that \mathbb{R}^n can be decomposed into a decomposition of subspaces

$$\mathbb{R}^n = \bigoplus_{b=1}^M V_b \tag{3.2}$$

with each V_b having dimension B [9, 17]. We assume that signals of interest are only nonzero on a few of the V_b 's and search for a solution which minimizes the norm

$$\|\mathbf{x}\|_{\ell_1/\ell_2} = \sum_{b=1}^M \|\mathbf{x}_b\|_2,$$

where \mathbf{x}_b denotes the projection of \mathbf{x} onto V_b .

Suppose \mathbf{x}_0 is block-sparse with k active blocks. T here denotes the coordinates associated with the groups where \mathbf{x}_0 has nonzero energy. T^\perp is equal to all of the coordinates of the groups where $\mathbf{x}_0 = 0$. The dual norm to the ℓ_1/ℓ_2 norm is the ℓ_∞/ℓ_2 norm

$$\|\mathbf{x}\|_{\ell_\infty/\ell_2} = \max_{1 \leq b \leq M} \|\mathbf{x}_b\|_2.$$

The subdifferential of the ℓ_1/ℓ_2 norm at \mathbf{x}_0 is given by

$$\partial\|\mathbf{x}_0\|_{\ell_1/\ell_2} = \left\{ \mathbf{z} \in \mathbb{R}^n : \mathcal{P}_T(\mathbf{z}) = \sum_{b \cap T \neq \emptyset} \frac{\mathbf{x}_{0,b}}{\|\mathbf{x}_{0,b}\|_2} \text{ and } \|\mathcal{P}_{T^\perp}(\mathbf{z})\|_{\ell_\infty/\ell_2} \leq 1 \right\}.$$

Much like in the ℓ_1 case, T denotes the span of the set of active subspaces and T^\perp is the set of inactive subspaces. In this formulation, $\dim(T) = kB$ and

$$\mathbf{e} = \sum_{b \cap T \neq \emptyset} \frac{\mathbf{x}_{0,b}}{\|\mathbf{x}_{0,b}\|_2}.$$

Note also that $\|\mathbf{e}\|_2 = \sqrt{k}$.

With the parameters we have just defined, we can define \mathbf{q} and \mathbf{y} by (2.3) and (2.4). If we again condition on $\|\mathbf{q}\|_2$, the components of \mathbf{y} on T^\perp are i.i.d. $\mathcal{N}(0, \|\mathbf{q}\|_2^2/m)$. Using the union bound, we have

$$\mathbb{P} \left[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_{\ell_\infty/\ell_2} \geq 1 \mid \|\mathbf{q}\|_2 \leq \tau \right] \leq \sum_{b \in T^\perp} \mathbb{P} \left[\|\mathbf{y}_b\|_2 \geq 1 \mid \|\mathbf{q}\|_2 \leq \tau \right]. \quad (3.3)$$

Conditioned on \mathbf{q} , $\frac{m}{\|\mathbf{q}\|_2^2} \|\mathbf{y}_b\|_2^2$ is identically distributed as a chi-squared random variable with B degrees of freedom. Letting $u = \sqrt{\chi_B}$, the Borell inequality [21, Proposition 5.34] gives

$$\mathbb{P}(u \geq \mathbb{E}u + t) \leq e^{-t^2/2}.$$

Since $\mathbb{E}u \leq \sqrt{B}$, we have $\mathbb{P}(u \geq \sqrt{B} + t) \leq e^{-t^2/2}$. Using this inequality, with

$$\tau = \sqrt{\frac{mk}{m - kB + 1 - t}},$$

we have that the probability of failure is upper bounded by

$$M \exp \left(-\frac{1}{2} \left[\sqrt{\frac{m - kB + 1 - t}{k}} - \sqrt{B} \right]^2 \right) + \exp \left(-\frac{t^2/4}{m - kB + 1} \right). \quad (3.4)$$

Choosing $m \geq (1 + \beta)k(\sqrt{B} + \sqrt{2 \log M})^2 + kB$ and setting $t = (\beta/2)k(\sqrt{B} + \sqrt{2 \log M})^2$, we can then upper bound (3.4) by

$$\begin{aligned} M \exp \left(-\frac{1}{2} \left[\sqrt{1 + \beta/2}(\sqrt{B} + \sqrt{2 \log M}) - \sqrt{B} \right]^2 \right) \\ + \exp \left(-\frac{\beta^2}{16(1 + \beta)} k(\sqrt{B} + \sqrt{2 \log M})^2 \right) \leq M^{-\beta/4} + M^{-\beta^2/(8+8\beta)}. \end{aligned}$$

This proves the following

Theorem 3.1 *Let \mathbf{x}_0 be a block-sparse signal with M blocks of size B and k active blocks under the decomposition (3.2). Let $\|\cdot\|_{\mathcal{A}}$ be the ℓ_1/ℓ_2 norm. For Gaussian measurement maps Φ with*

$$m > (1 + \beta)k(\sqrt{B} + \sqrt{2\log M})^2 + kB$$

the recovery is exact with probability at least $M^{-\beta/4} + M^{-\beta^2/(8+8\beta)}$.

The bound on m obtained by this theorem is identical to that of [18], and is, to our knowledge, the tightest known non-asymptotic bound for block-sparse signals. For example, when the block size B is much greater than $\log M$, the results asserts that roughly $2kB$ measurements are sufficient for the convex programming to be exact. Since there are kB degrees of freedom, one can see that this is quite tight.

Note that the theorem gives a recovery result for sparse vectors by setting $B = 1$, $k = s$, and $M = n$. In this case, Theorem 3.1 gives a slightly looser bound and requires a slightly more complicated argument as compared to Theorem 1.1. However, Theorem 3.1 provides bounds for more general types of signals, and we note that the same analysis would handle other ℓ_1/ℓ_p block regularization schemes defined as $\|\mathbf{x}\|_{\ell_1/\ell_p} = \sum_{b=1}^M \|\mathbf{x}_b\|_p$ with $p \in [2, \infty]$. Indeed, the ℓ_1/ℓ_p norm is decomposable and its dual is the ℓ_∞/ℓ_q norm with $1/p + 1/q = 1$. The only adjustment would consist in bounding $\|\mathbf{y}_b\|_q$; up to a scaling factor, this is a sum of independent standard normals and our analysis goes through. We omit the details.

3.3 Low-Rank Matrix Recovery in the Gaussian Ensemble

To apply our results to recovering low-rank matrices, we need a little bit more notation, but the argument is principally the same. Let \mathbf{X}_0 be an $n_1 \times n_2$ matrix of rank r with singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. Without loss of generality, impose the conventions $n_1 \leq n_2$, $\mathbf{\Sigma}$ is $r \times r$, \mathbf{U} is $n_1 \times r$, \mathbf{V} is $n_2 \times r$.

In the low-rank matrix reconstruction problem, the subspace T is the set of matrices of the form $\mathbf{U}\mathbf{Y}^* + \mathbf{X}\mathbf{V}^*$ where \mathbf{X} and \mathbf{Y} are arbitrary $n_1 \times r$ and $n_2 \times r$ matrices. The span of matrices of the form $\mathbf{U}\mathbf{Y}^*$ has dimension $n_1 r$, the span of $\mathbf{X}\mathbf{V}^*$ has dimension $n_2 r$, and the intersection of these two spans has dimension r^2 . Hence, we have $d_T = \dim(T) = r(n_1 + n_2 - r)$. T^\perp is the subspace of matrices spanned by the family $(\mathbf{x}\mathbf{y}^*)$, where \mathbf{x} (respectively \mathbf{y}) is any vector orthogonal to \mathbf{U} (respectively \mathbf{V}). The spectral norm denoted by $\|\cdot\|$ is dual to the nuclear norm. The subdifferential of the nuclear norm at \mathbf{X}_0 is given by

$$\partial\|\mathbf{X}_0\|_* = \{\mathbf{Z} : \mathcal{P}_T(\mathbf{Z}) = \mathbf{U}\mathbf{V}^* \text{ and } \|\mathcal{P}_{T^\perp}(\mathbf{Z})\| \leq 1\}.$$

Note that the Euclidean norm of $\mathbf{U}\mathbf{V}^*$ is equal to \sqrt{r} .

For matrices, a Gaussian measurement map takes the form of a linear operator whose i th component is given by

$$[\Phi(\mathbf{Z})]_i = \text{Tr}(\Phi_i^* \mathbf{Z}).$$

Above, Φ_i is an $n_1 \times n_2$ random matrix with i.i.d., zero-mean Gaussian entries with variance $1/m$. This is equivalent to defining Φ as an $m \times (n_1 n_2)$ dimensional matrix acting on $\text{vec}(\mathbf{Z})$, the vector composed of the columns of \mathbf{Z} stacked on top of one another. In this case, the dual multiplier is a matrix taking the form

$$\mathbf{Y} = \Phi^* \Phi_T (\Phi_T^* \Phi_T)^{-1} (\mathbf{U}\mathbf{V}^*).$$

Here, Φ_T is the restriction of Φ to the subspace T . Concretely, one could define a basis for T and write out Φ_T as an $m \times d_T$ dimensional matrix. Note that none of the abstract setup from Section 2.2 changes for the matrix recovery problem: \mathbf{Y} exists as soon as $m \geq \dim(T) = r(n_1 + n_2 - r)$ and $\mathcal{P}_T(\mathbf{Y}) = \mathbf{U}\mathbf{V}^*$ as desired. We need only guarantee that $\|\mathcal{P}_{T^\perp}(\mathbf{Y})\| < 1$. We still have that

$$\mathcal{P}_{T^\perp}(\mathbf{Y}) = \sum_{i=1}^m q_i \mathcal{P}_{T^\perp}(\mathbf{A}_i)$$

where $\mathbf{q} = \Phi_T(\Phi_T^* \Phi_T)^{-1}(\mathbf{U}\mathbf{V}^*)$ is given by (2.3) and, importantly, \mathbf{q} and $\mathcal{P}_{T^\perp}(\mathbf{A}_i)$ are independent for all i .

With such a definition, we can again straightforwardly apply (2.7) once we obtain an estimate of $\|\mathcal{P}_{T^\perp}(\mathbf{Y})\|$ conditioned on \mathbf{q} . Observe that $\mathcal{P}_{T^\perp}(\mathbf{Y}) = \mathcal{P}_{U^\perp} \mathbf{Y} \mathcal{P}_{V^\perp}$, where \mathcal{P}_{U^\perp} (respectively \mathcal{P}_{V^\perp}) is a projection matrix onto the orthogonal complement of U (respectively V). It follows that $\mathcal{P}_{T^\perp}(\mathbf{Y})$ is identically distributed to a rotation of an $(n_1 - r) \times (n_2 - r)$ Gaussian random matrix whose entries have mean zero and variance $\|\mathbf{q}\|_2^2/m$. Using the Davidson-Szarek concentration inequality for the extreme singular values for Gaussian random matrices [6], we have

$$\mathbb{P}[\|\mathcal{P}_{T^\perp}(\mathbf{Y})\| > 1 \mid \|\mathbf{q}\|_2 \leq \tau] \leq \exp\left(-\frac{1}{2} \left(\frac{\sqrt{m}}{\tau} - \sqrt{n_1 - r} - \sqrt{n_2 - r}\right)^2\right).$$

We are again in a position ready to prove (2.7). To guarantee matrix recovery, with $\tau = \sqrt{\frac{mr}{m-d_T+1-t}}$, we thus need

$$\sqrt{\frac{m-d_T+1-t}{r}} - \sqrt{n_1-r} - \sqrt{n_2-r} \geq 0.$$

This occurs if

$$m \geq r(n_1 + n_2 - r) + \left(\sqrt{r(n_1 - r)} + \sqrt{r(n_2 - r)}\right)^2 + t - 1$$

But since $(a+b)^2 \leq 2(a^2 + b^2)$, we can upper bound

$$\left(\sqrt{r(n_1 - r)} + \sqrt{r(n_2 - r)}\right)^2 \leq 2r(n_1 + n_2 - 2r).$$

Setting $t = (\sqrt{2r+1} - 1)(\beta - 1)(3n_1 + 3n_2 - 5r)$ in (2.7) then yields Theorem 1.2.

4 Discussion

We note that with minor modifications, the results for sparse and block-sparse signals can be extended to measurement matrices whose entries are i.i.d. subgaussian random variables. In this case, we can no longer use the theory of inverse Wishart matrices, but Φ_{T^\perp} and Φ_T are still independent, and we can bound the norm of \mathbf{q} using bounds on the smallest singular value of rectangular matrices. For example, Theorem 39 in [21] asserts that there exist positive constants θ and γ such that the smallest singular value obeys the deviation inequality

$$\mathbb{P}\left[\sigma_{\min}(\Phi_T) \leq 1 - \theta\sqrt{d_T/m} - t\right] \leq e^{-\gamma mt^2} \quad (4.1)$$

for $t > 0$.

We use this concentration inequality to prove the second part of Theorem 1.1. Since $\|\Phi_T(\Phi_T^*\Phi_T)^{-1}\| = \sigma_{\min}^{-1}(\Phi_T)$, we have that

$$\|\mathbf{q}\|_2 \leq \frac{\sqrt{s}}{1 - \theta\sqrt{s/m} - t} := \rho$$

with probability at least $1 - e^{-\gamma mt^2}$. This is the analog of (2.6). Now whenever $\|\mathbf{q}\|_2 \leq \rho$, Hoeffding's inequality [12] implies that (3.1) still holds. Thus, we are in the same position as before, and obtain

$$\mathbb{P}[\|\mathcal{P}_{T^\perp}(\mathbf{y})\|_\infty \geq 1] \leq 2(n-s) \exp\left(-\frac{m}{2\rho^2}\right) + \exp(-\gamma mt^2).$$

Setting $t = \epsilon/2$ proves the second part of Theorem 1.1.

For block-sparse signals, a similar argument would apply. The only caveat is that we would need the following concentration bound which follows from Lemma 5.2 in [1]: Let \mathbf{M} be an $d_1 \times d_2$ dimensional matrix with i.i.d. entries taking on values ± 1 with equal probability. Let \mathbf{v} be a fixed vector in \mathbb{R}^{d_2} . Then

$$\mathbb{P}[\|\mathbf{M}\mathbf{v}\|_2 \geq 1] \leq \exp\left(-\frac{\|\mathbf{v}\|_2^2 - d_1}{4}\right)$$

provided $\|\mathbf{v}\| \leq \sqrt{d_1}$. Plugging this bound into (3.3) gives an analogous threshold for block-sparse signals in the Bernoulli model:

Theorem 4.1 *Let \mathbf{x}_0 be a block-sparse signal with M blocks and k active blocks under the decomposition (3.2). Let $\|\cdot\|_{\mathcal{A}}$ be the ℓ_1/ℓ_2 norm. Let $\beta > 1$ and $\epsilon \in (0, 1)$. For binary measurement maps Φ with i.i.d. entries taking on values $\pm m^{-1/2}$ with equal probability, there exist numerical constants c_0 and c_1 such that if $M \geq \exp(c_0/\epsilon^2)$ and $m \geq 4k\beta(1-\epsilon)^{-2} \log M + 2kB$, the recovery is exact with probability at least $1 - M^{1-\beta} - M^{-c_1\beta\epsilon^2}$.*

For low-rank matrix recovery, the situation is more delicate. With general subgaussian measurement matrices, we no longer have independence between the action on the subspaces T and T^\perp unless the singular vectors somehow align serendipitously with the coordinate axes. In this case, it unfortunately appears that we need to resort to more complicated arguments and will likely be unable to attain such small constants through the dual multiplier without a conceptually new argument.

Acknowledgements

We thank anonymous referees for suggestions and corrections that have improved the presentation. EC is partially supported by NSF via grants CCF-0963835 and the 2006 Waterman Award; by AFOSR under grant FA9550-09-1-0643; and by ONR under grant N00014-09-1-0258. BR is partially supported by ONR award N00014-11-1-0723 and NSF award CCF-1139953.

References

- [1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and Systems Science*, 66(4):671–687, 2003. Special issue of invited papers from PODS'01.
- [2] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

- [3] E. J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- [4] 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.
- [5] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. Submitted for publication. Preprint available at arxiv.org/1012.0621, 2010.
- [6] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach spaces*, pages 317–366. Elsevier Scientific, 2001.
- [7] D. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.
- [8] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [9] Y. C. Eldar and H. Bolcskei. Block-sparsity: Coherence and efficient recovery. In *ICASSP, The International Conference on Acoustics, Signal and Speech Processing*,, 2009.
- [10] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50:1341–1344, 2004.
- [11] Y. Gordon. On Milman’s inequality and random subspaces which escape through a mesh in \mathbb{R}^n . In *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics 1317, pages 84–106. Springer, 1988.
- [12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [13] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.
- [14] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.
- [15] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 2009.
- [16] S. Oymak and B. Hassibi. New null space results and recovery thresholds for matrix rank minimization. Submitted for publication. Preprint available at arxiv.org/abs/1011.6326, 2010.
- [17] F. Parvaresh and B. Hassibi. Explicit measurements with almost optimal thresholds for compressed sensing. In *ICASSP, The International Conference on Acoustics, Signal and Speech Processing, C*, 2008.
- [18] N. Rao, B. Recht, and R. Nowak. Universal measurement bounds for structured sparse signal recovery. In *Proceedings of AISTATS*, 2102.
- [19] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [20] M. Stojnic. Various thresholds for ℓ_1 -optimization in compressed sensing. Preprint available at arxiv.org/abs/0907.3666, 2009.
- [21] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press. To Appear. Preprint available at <http://www-personal.umich.edu/~romanv/papers/papers.html>.