WEAK RECOVERY CONDITIONS FROM GRAPH PARTITIONING BOUNDS AND ORDER STATISTICS

ALEXANDRE D'ASPREMONT AND NOUREDDINE EL KAROUI

ABSTRACT. We study a weaker formulation of the nullspace property which guarantees recovery of sparse signals from linear measurements by ℓ_1 minimization. We require this condition to hold only with high probability, given a distribution on the nullspace of the coding matrix A. Under some assumptions on the distribution of the reconstruction error, we show that testing these weak conditions means bounding the optimal value of two classical graph partitioning problems: the k-Dense-Subgraph and MaxCut problems. Both problems admit efficient, relatively tight relaxations and we use a randomization argument to produce new approximation bounds for k-Dense-Subgraph. We test the performance of our results on several families of coding matrices.

1. INTRODUCTION

Given a coding matrix $A \in \mathbf{R}^{q \times n}$ and a signal $e \in \mathbf{R}^n$, we focus on conditions under which the solution x_0 to the following minimum cardinality problem

$$
\begin{aligned}
x_0 = \quad & \text{min.} \quad && \mathbf{Card}(x) \\
& \text{subject to} \quad && Ax = Ae,
\end{aligned}
\tag{ℓ_0-recov.}
$$

which is a combinatorial problem in $x \in \mathbf{R}^n$, can be recovered by solving

$$
\begin{aligned}
x^{\text{lp}} = \quad & \text{min.} \quad && \|x\|_1 \\
& \text{subject to} \quad && Ax = Ae,
\end{aligned}
\tag{ℓ_1-recov.}
$$

which is a convex program in $x \in \mathbf{R}^n$. Problem (ℓ_0-recov.) arises in various fields ranging from signal processing to statistics. Suppose for example that we make a few linear measurements of a high dimensional signal, which admits a sparse representation in a well chosen basis (e.g. Fourier, wavelet). Under certain conditions, solving (ℓ_1-recov.) will allow us to reconstruct the signal exactly (Donoho, 2004; Donoho and Tanner, 2005; Donoho, 2006). In a coding application, suppose we transmit a message which is corrupted by a few errors, solving (ℓ_1-recov.) will then allow us to reconstruct the message exactly (Candès and Tao, 2005, 2006). Finally, problem (ℓ_1-recov.) is directly connected to variable selection and penalized regression problems (e.g. LASSO Tibshirani (1996)) arising in statistics (Zhao and Yu, 2006; Meinshausen and Yu, 2008; Meinshausen et al., 2007; Candes and Tao, 2007; Bickel et al., 2007; Candès and Plan, 2009). Of course, in all these fields, problems (ℓ_0-recov.) and (ℓ_1-recov.) are overly simplified. In practice for example, the observations could be noisy, approximate solutions might be sufficient and we might have strict computational limits on the decoding side. While important, these extensions are outside the scope of this work.

Based on results by Vershik and Sporyshev (1992) and Affentranger and Schneider (1992), Donoho and Tanner (2005) showed that when the solution x_0 of (ℓ_0-recov.) is sparse with $\mathbf{Card}(x_0) = k$ and the coefficients of A are i.i.d. Gaussian, then w.h.p. the solution of the (convex) problem in (ℓ_1-recov.) will always match that of the combinatorial problem in (ℓ_0-recov.) provided k is below an explicitly computable *strong recovery* threshold k_S. They also show that if k is below another (larger) *weak recovery* threshold k_W, then these solutions match with an exponentially small probability of failure.

Generic conditions for strong recovery based on sparse extremal eigenvalues, or *restricted isometry properties* (RIP), were also derived in Candès and Tao (2005) and Candès and Tao (2006), who proved that certain

---

random matrix classes satisfied these conditions near optimal values of k with an exponentially small probability of failure. Simpler, weaker conditions which can be traced back to Donoho and Huo (2001), Zhang (2005) or Cohen et al. (2009) for example, are based on properties of the nullspace of A. When the signal cardinality $\mathbf{Card}(e) \leq k$ and the *Nullspace Property* (NSP) holds, i.e. when there is a constant $\alpha_k < 1/2$ such that

$$\|x\|_{k,1} \leq \alpha_k \|x\|_1 \qquad \text{(det-NSP)}$$

for all vectors $x \in \mathbf{R}^n$ with $Ax = 0$, then solving the convex problem (ℓ_1-recov.) will recover the global solution to the combinatorial problem (ℓ_0-recov.). Condition (det-NSP) can be understood as an incoherence measure, i.e. it means that not all of the mass in x can be concentrated among only k coefficients, in other words:

Good coding matrices have incoherent nullspace vectors.

In particular, this condition means that the nullspace of A cannot contain sparse vectors. Furthermore, the constant α_k can be used to explicitly bound the reconstruction error when solving the ℓ_1-recovery problem in (ℓ_1-recov.). This is illustrated in Proposition 2.1 below, directly adapted from Cohen et al. (2009, Th. 4.3).

One fundamental issue with the sparse recovery conditions described above is that, except for explicit thresholds available for certain types of random matrices (with high probability), testing these conditions on generic (deterministic) matrices is potentially *harder* than solving the combinatorial ℓ_0-norm minimization problem in (ℓ_0-recov.) for example as it implies either solving a combinatorial problem to compute α_k in (det-NSP), or computing sparse eigenvalues. Recent results in Candès and Plan (2009) show that the traditional (and tractable) incoherence conditions ensure recovery of sparse signals with high probability, given a uniform distribution on the signal. These incoherence conditions lack universality however, in the sense that contrary to the combinatorial conditions mentioned above, they cannot be used to guarantee recovery of *all* signals of near-optimal size k. Convex relaxation relaxation bounds were used in d'Aspremont et al. (2008) (on sparse eigenvalues), Juditsky and Nemirovski (2011) or d'Aspremont and El Ghaoui (2011) (on NSP) to test sparse recovery conditions similar to (det-NSP) on arbitrary matrices. Unfortunately, the performance (tightness) of these relaxations is still very insufficient: for matrices satisfying the sparse recovery conditions in Candès and Tao (2005) up to signal cardinality k^*, these three relaxations can only certify that the conditions hold up to cardinality $\sqrt{k^*}$ and are also likely to provide poor bounds on reconstruction error.

In what follows, we seek to enforce a weaker version of condition (det-NSP). We will bound the incoherence measure α_k in (det-NSP) with high probability over a random sample of vectors in the nullspace of A. Another way to look at this approach is to remember that, if x^{lp} solves the ℓ_1-decoding problem in (ℓ_1-recov.), the vector $x^{\mathrm{lp}} - e$ is always in the nullspace of A and Proposition 2.1 below shows that enforcing condition (det-NSP) on the reconstruction error $x^{\mathrm{lp}} - e$ allows us to bound the magnitude of this error.

Here, because we cannot efficiently test condition (det-NSP) over all vectors in the nullspace of A, we will instead require condition (det-NSP) to hold only with high probability on the nullspace of A, given a distribution on this subspace. Let us assume for simplicity that $\mathbf{Rank}(A) = q$, and let $F \in \mathbf{R}^{n \times m}$ with $m = n - q$ be a basis for the nullspace of A (not necessarily orthogonal or normalized). We will require that the NSP condition (det-NSP) discussed above, which reads

$$\|Fy\|_{k,1} \leq \alpha_k \|Fy\|_1 \qquad \text{(proba-NSP)}$$

be satisfied with high probability, given a distribution on y. We will start by assuming that y is Gaussian. In this case, we will see that both sides of condition (proba-NSP) can be explicitly controlled by the solution of classic graph partitioning problems. These combinatorial problems admit tight, efficiently computable approximations which will allow us to bound the probability that (proba-NSP) holds. We will then extend these results to more general distributions on the nullspace and show that the same quantities which controlled concentration in the Gaussian case, also control fluctuations in the more general model.

Of course, assuming the true distribution on the signal e is either sparse or follows a power law, our simple model on the nullspace of A error could have zero measure with respect to the true (structured) distribution of $x^{\mathrm{lp}} - e$, especially since x^{lp} is dependent on A. In fact, at first sight, we are implicitly positing a model on the reconstruction error, then ultimately use the model to bound this same reconstruction error,

an apparent circular reference. Our main objective however is not to directly bound the error but rather to isolate efficiently computable quantities which will be good proxies for this error, sacrificing some statistical accuracy in favor of computational efficiency. Moreover, our main result is to efficiently approximate the Lipschitz constants of the two norms in (proba-NSP), constants which are likely to play a critical role *whatever the model* on the reconstruction error. Overall, the phase transition for signal recovery is usually very sharp (or "bang-bang") meaning that either all signals are recovered perfectly or none of them are. This means that choosing a realistic statistical model is probably not that crucial.

Current results in compressed sensing provide universal recovery guarantees using intractable conditions (which can only be tested with high probability on *random matrices*). Our objective here is to do the opposite and isolate *tractable* measures of performance that can be computed on *arbitrary matrices*, even if this means losing some confidence in our signal recovery guarantees. Numerical experiments detailed at the end of this work, using simple models for e, seem to suggest that our assumptions on $x^{\text{lp}} - e$ are not completely unreasonable (cf. Figure 1). Furthermore, the fact that the true signal e is inherently structured means that, in principle, these statistical fidelity questions would arise with *any* model on e.

Our contribution here is twofold. First, assuming a Gaussian model or bounded independent model on the nullspace of the matrix A in (ℓ_0-recov.), we show that testing if the NSP condition (proba-NSP) holds with high probability amounts to bounding the value of two classic graph partitioning problems: MaxCut and k-Dense-Subgraph. Second, we show new approximation results for semidefinite relaxations of the k-Dense-Subgraph problem when the graph weight matrix is positive semidefinite but has coefficients of *arbitrary sign*. This result has applications outside of the compressed sensing context discussed in this paper, and is directly related to *correlation clustering* for example. Solving a k-Dense-Subgraph problem on a (positive semidefinite) correlation matrix (modeling similarities between variables) isolates a highly correlated k-cluster of variables. Here, we use these approximation results to show that our weak recovery conditions can be certified in polynomial time for arbitrary matrices even when the target cardinality k is near the true recovery threshold k^* (i.e. a log term away). This allows us to break the $\sqrt{k^*}$ barrier that plagued all tractable conditions for recovery developed so far (d'Aspremont et al., 2008; Juditsky and Nemirovski, 2011; d'Aspremont and El Ghaoui, 2011). The current state of the art in checking recovery conditions is that we have conditions that we can trust but cannot fully test. We focus here on conditions we can test, but cannot fully trust.

The paper is organized as follows. Conditions for sparse recovery with high probability, given a model on the nullspace of the sampling matrix are derived in Section 2. The performance of these conditions on random matrices and links with the restricted isometry property are discussed in Section 3. Section 4 derives semidefinite relaxations and approximation results for the graph partitioning problems used in testing these weak recovery conditions. Section 5 briefly discusses the complexity of solving these relaxations. Section 6 shows that the relaxations detailed in Section 4 allow us to certify weak recovery for near optimal values of the signal cardinality k, thus breaking the $\sqrt{n}$ barrier that plagues tractable (uniform, deterministic) recovery conditions. Finally, we present some numerical experiments in Section 7.

Notation. For $x \in \mathbf{R}^n$, we write $\|x\|_{k,1}$ the sum of the magnitudes of the k largest coefficients of x. When $X \in \mathbf{R}^{m \times n}$, X_i is the ith row of X, $\|X\|_2$ the spectral norm and $\|X\|_F$ the Frobenius (Euclidean) norm of X. For matrices $A, B \in \mathbf{R}^{m \times n}$, we write $A \otimes B$ their Kronecker product and $A \circ B$ their Schur (componentwise) product. We write $\mathbf{NumRank}(X)$ the numerical rank of the matrix X, with $\mathbf{NumRank}(X) = \|X\|_F^2/\|X\|_2^2$, and $\mathbf{NumCard}(x)$ is the numerical cardinality of a vector x, with $\mathbf{NumCard}(x) = \|x\|_1^2/\|x\|_2^2$. Finally, we write $x \preceq_c y$ when $\mathbf{E}[f(x)] \leq \mathbf{E}[f(y)]$ for any *convex* function $f : \mathbf{R}^n \to \mathbf{R}$ (this is usually called convex majorization).

2. WEAK RECOVERY CONDITIONS

To highlight the central role of the NSP condition in ℓ_1 decoding, we begin by adapting a result from Cohen et al. (2009, Th. 4.3) which uses the constant α_k to bound the reconstruction error when decoding the observations Ae by solving problem (ℓ_1-recov.). Recall that x^{lp} is the solution to the linear program in (ℓ_1-recov.) and e the true signal.

Proposition 2.1. *Suppose that $\|x^{\mathrm{lp}}-e\|_{k,1} \leq \alpha_k\|x^{\mathrm{lp}}-e\|_1$ for some $\alpha_k < 1/2$, where $e \in \mathbf{R}^n$ and $x^{\mathrm{lp}} \in \mathbf{R}^n$ solves problem (ℓ_1-recov.), then $A(x^{\mathrm{lp}} - e) = 0$ and*

$$\|x^{\mathrm{lp}} - e\|_1 \leq \frac{2}{(1 - 2\alpha_k)} \ \min_{\{y \in \mathbf{R}^n : \, \mathbf{Card}(y) \leq k\}} \|y - e\|_1, \tag{1}$$

where the right-hand side is proportional to the best ℓ_1 reconstruction error on e using a signal with cardinality k.

Proof. We adapt the proof of Cohen et al. (2009, Th. 4.3). Because x^{lp} solves (ℓ_1-recov.), we have $\|x^{\mathrm{lp}}\|_1 \leq \|e\|_1$ since e is feasible. Denoting by T the indices of the k largest coefficients in absolute value of e and by $\eta = x^{\mathrm{lp}} - e$ the reconstruction error, we write

$$\|x_T^{\mathrm{lp}}\|_1 + \|x_{T^c}^{\mathrm{lp}}\|_1 \leq \|e_T\|_1 + \|e_{T^c}\|_1$$

and triangular inequalities yield

$$\|e_T\|_1 - \|\eta_T\|_1 + \|\eta_{T^c}\|_1 - \|e_{T^c}\|_1 \leq \|e_T\|_1 + \|e_{T^c}\|_1$$

hence

$$\|\eta_{T^c}\|_1 \leq \|\eta_T\|_1 + 2\|e_{T^c}\|_1.$$

Note that by definition of T, we have $\|e_{T^c}\|_1 = \min_{\{y \in \mathbf{R}^n : \, \mathbf{Card}(y) \leq k\}} \|y - e\|_1$. From our assumption on η and by definition of $\|\cdot\|_{k,1}$, $|T| = k$ means

$$\|\eta_T\|_1 \leq \|\eta\|_{k,1} \leq \alpha_k\|\eta\|_1 = \alpha_k(\|\eta_T\|_1 + \|\eta_{T^c}\|_1)$$

hence

$$\|\eta_T\|_1 \leq \frac{\alpha_k}{1 - \alpha_k}\|\eta_{T^c}\|_1$$

which then yields

$$\|\eta_{T^c}\|_1 \leq \frac{(2 - 2\alpha_k)}{(1 - 2\alpha_k)} \ \min_{\{y \in \mathbf{R}^n : \, \mathbf{Card}(y) \leq k\}} \|y - e\|_1.$$

Using the fact that

$$\|\eta\|_1 = \|\eta_T\|_1 + \|\eta_{T^c}\|_1 \leq \left(1 + \frac{\alpha_k}{1 - \alpha_k}\right)\|\eta_{T^c}\|_1$$

we get $\|\eta\|_1 \leq \|\eta_{T^c}\|_1/(1 - \alpha_k)$, which produces the desired result. ∎

This last result shows that whenever the reconstruction error satisfies (det-NSP) with constant $\alpha_k < 1$, then the magnitude of this error is at most $2/(1 - 2\alpha_k)$ times the best possible reconstruction error achievable using a signal of size k.

2.1. Invariance properties.

Remark that condition (det-NSP), which guarantees recovery of all signals of cardinality less than k can be written

$$\|Fy\|_{k,1} \leq \alpha_k\|Fy\|_1, \quad \text{for all } y \in \mathbf{R}^m, \tag{2}$$

for some $\alpha_k < 1/2$, where $F \in \mathbf{R}^{n \times m}$ is a basis of the nullspace of A. This condition is clearly independent of the choice of basis, hence if F satisfies (2) then so does FQ where Q is any orthogonal matrix.

2.2. Gaussian model.

In what follows, we will use concentration inequalities to bound both sides of the probabilistic Nullspace Property inequality (proba-NSP), namely check that

$$\|Fy\|_{k,1} \leq \alpha_k\|Fy\|_1$$

holds with high probability when y is Gaussian with $y \sim \mathcal{N}(0, \mathbf{I}_m)$, where F is a basis of the nullspace of A. Of course, this means that we implicitly assume that the reconstruction error $x^{\mathrm{lp}} - e$ follows a Gaussian model. Outside of tractability benefits, there is no fundamental reason to pick a Gaussian distribution on the nullspace of A here, except that its rotational invariance means the basis matrix F only has to be defined up to a rotation. This is consistent with the fact that recovery performance, as characterized by the nullspace property (det-NSP), is only a function of the nullspace of A and not of its matrix representation.

Concentration inequalities on Lipschitz functions of Gaussian variables then translate (proba-NSP) into explicit conditions on the matrix F. We begin by the following lemma controlling the left-hand side of this inequality.

Lemma 2.2. *Suppose $F \in \mathbf{R}^{n \times m}$ and $y \sim \mathcal{N}(0, \mathbf{I}_m)$, then*

$$\mathbf{Prob}\left[\|Fy\|_{k,1} \geq \mathbf{E}[\|Fy\|_{k,1}] + x\right] \leq e^{-\frac{x^2}{2\sigma_k^2(F)}}$$

where

$$\sigma_k^2(F) \triangleq \max_{\{u \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} u^T \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes FF^T u. \tag{3}$$

and

$$\mathbf{E}[\|Fy\|_{k,1}] \leq \sigma_k(F)\sqrt{2\log(2^k \binom{n}{k})} \leq \sigma_k(F)\sqrt{2k\left(1 + \log\left(\frac{2n}{k}\right)\right)}.$$

Proof. We can write the left-hand side of inequality (proba-NSP) as

$$\|Fy\|_{k,1} = \max_{\{u=(u_+,u_-)\in\{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} (u_+ - u_-)^T Fy$$

which means that $\|Fy\|_{k,1}$ is the maximum of Gaussian variables. Concentration results detailed in (Massart, 2007, Th. 3.12) for example show that

$$\mathbf{Prob}\left[\|Fy\|_{k,1} \geq \mathbf{E}[\|Fy\|_{k,1}] + x\right] \leq e^{-\frac{x^2}{2\sigma_k^2(F)}}$$

where $\sigma_k(F)$ is defined as

$$\sigma_k^2(F) = \max_{\{u=(u_+,u_-)\in\{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \mathbf{E}\left[\left((u_+ - u_-)^T Fy\right)^2\right].$$

We have

$$\begin{aligned}
\mathbf{E}\left[\left((u_+ - u_-)^T Fy\right)^2\right] &= \left\|(u_+ - u_-)^T F\right\|_2^2 \\
&= (u_+ - u_-)^T FF^T(u_+ - u_-) \\
&= \begin{pmatrix} u_+ \\ u_- \end{pmatrix}^T \begin{pmatrix} FF^T & -FF^T \\ -FF^T & FF^T \end{pmatrix} \begin{pmatrix} u_+ \\ u_- \end{pmatrix},
\end{aligned}$$

and we recover (3) after setting $u = (u_+, u_-)$. Note that we also have

$$\|Fy\|_{k,1} = \max_{\{v \in \mathcal{V}_k\}} v^T Fy,$$

where $\mathcal{V}_k$ is the set of vectors of size n with exactly k entries equal to $+1$ or -1, and $n - k$ zeroes. Each $v^T Fy$ is Gaussian with zero mean and variance $v^T FF^T v$, so $\|Fy\|_{k,1}$ is the maximum of $2^k \binom{n}{k}$ Gaussian random variables. Using (Massart, 2007, Lem. 2.3) we can therefore bound the expectation as follows

$$\mathbf{E}[\|Fy\|_{k,1}] \leq \sigma_k(F)\sqrt{2\log(2^k \binom{n}{k})}$$

and

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \left(\frac{ne}{k}\right)^k$$

yields the desired result. ∎

Note that the bound in $\exp(-x^2/2\sigma_k^2(F))$ can be replaced by $2(1-N(x/\sigma_k(F)))$ (see e.g (Massart, 2007, Thm 3.8), where $N(x)$ is the Gaussian CDF, which is smaller for larger values of x. Expression (3) means $\sigma_k^2(F)$ is the optimum value of a k-Dense-Subgraph problem. Several efficient approximation algorithms have been derived for this graph partitioning problem and will be discussed in Section 4. We now apply similar concentration results to control the fluctuations of the right hand side of inequality (proba-NSP).

Lemma 2.3. *Suppose $F \in \mathbf{R}^{n \times m}$ and $y \sim \mathcal{N}(0, \mathbf{I}_m)$, then*

$$\mathbf{Prob}\left[\|Fy\|_1 \leq \mathbf{E}[\|Fy\|_1] - x\right] \leq e^{-\frac{x^2}{2L^2(F)}}$$

where

$$\mathbf{E}[\|Fy\|_1] = \sqrt{\tfrac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2$$

and $L^2(F) = \max_{v \in \{-1,1\}^n} v^T F F^T v \ (= \sigma_n^2(F))$ is bounded by the following MaxCut relaxation

$$
\begin{aligned}
\tfrac{2}{\pi} L_{\mathrm{mxct}}^2(F) \leq L^2(F) \leq L_{\mathrm{mxct}}^2(F) \triangleq \quad & \textit{max.} \quad \mathbf{Tr}(XFF^T) \\
& \textit{s.t.} \quad \mathbf{diag}(X) = \mathbf{1}, X \succeq 0,
\end{aligned}
\tag{4}
$$

with, in particular, $L_{\mathrm{mxct}}(F) \leq \sqrt{n}\|F\|_2$.

Proof. We can write

$$\|Fy\|_1 = \max_{v \in \{-1,1\}^n} v^T Fy$$

and (Massart, 2007, Th. 3.12) shows that

$$\mathbf{Prob}\left[\|Fy\|_1 \leq \mathbf{E}[\|Fy\|_1] - x\right] \leq e^{-\frac{x^2}{2L^2(F)}}.$$

The fact that $\mathbf{E}[|g|] = \sqrt{2/\pi}V$ whenever $g \sim \mathcal{N}(0, V^2)$ produces the expectation, and the Lipschitz constant $L^2(F)$ in this inequality is given by the largest variance

$$L^2(F) = \max_{v \in \{-1,1\}^n} v^T FF^T v,$$

hence is the solution of a graph partitioning problem similar to MaxCut. Relaxation results in (Goemans and Williamson, 1995) (in the case where the matrix is nonnegative) and (Nesterov, 1998a) show that this combinatorial problem can be bounded by solving

$$
\begin{aligned}
\textit{maximize} \quad & \mathbf{Tr}(XFF^T) \\
\textit{subject to} \quad & \mathbf{diag}(X) = \mathbf{1}, X \succeq 0,
\end{aligned}
$$

which is a semidefinite relaxation in $X \in \mathbf{S}_n$ of the maximum variance problem (tight up to a factor $\pi/2$). Its dual is written

$$
\begin{aligned}
\textit{minimize} \quad & \mathbf{1}^T w \\
\textit{subject to} \quad & FF^T \preceq \mathbf{diag}(w),
\end{aligned}
$$

which is another semidefinite program in the variable $w \in \mathbf{R}^n$. By weak duality, any feasible point of this last problem gives an upper bound on $L_{\mathrm{mxct}}(F)$. In particular, the point $w = \lambda_{\max}(FF^T)\mathbf{1}$ is dual feasible and yields $L_{\mathrm{mxct}}(F) \leq \sqrt{n}\|F\|_2$. ∎

The bound detailed in Lemma 2.3 is directly related to the $MatrixNorm$ problem discussed in Nemirovski (2001) and Nemirovski (2005) or the spin glass models of statistical mechanics. In particular, our approximation bound on $L(F)$ can be directly deduced from the bound on the induced matrix norm $\|\cdot\|_{2,1}$ derived in Nemirovski (2005, Prop. 1.4). Note also that the mean $\mathbf{E}[\|Fy\|_1] = \sqrt{2/\pi} \sum_{i=1}^{n} \|F_i\|_2$ is typically much larger than the factor $L(F)$ controlling concentration. In fact, we can write $\sum_{i=1}^{n} \|F_i\|_2 = \|F\|_F \mathbf{NumCard}(\{\|F_i\|_2\})^{1/2} = \|F\|_2 \mathbf{NumRank}(F)^{1/2} \mathbf{NumCard}(\{\|F_i\|_2\})^{1/2}$. Combining the last two lemmas, we show the following proposition, which is our main recovery condition.

Proposition 2.4. *If $F \in \mathbf{R}^{n \times m}$ satisfies*

$$\left(\sqrt{2k\left(1 + \log\frac{2n}{k}\right)} + \beta\right) \sigma_k(F) \leq \left(\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F)\right) \alpha_k \tag{5}$$

for some $\beta > 0$, where $\sigma_k(F)$ was defined in (3) and $L(F)$ in (4), then the sparse recovery condition (proba-NSP) will be satisfied with probability $1 - 2e^{-\beta^2/2}$ when $y \sim \mathcal{N}(0, \mathbf{I}_m)$.

Proof. We combine the bounds of Lemmas 2.2 and 2.3, requiring them to hold with probability $1 - e^{-\beta^2/2}$. ∎

We finish this section by showing that the function $\sigma_k(F)$ defined in (3) is strictly increasing with k whenever the diagonal of FF^T is positive, which will prove useful in the results that follow.

Lemma 2.5. *Let $F \in \mathbf{R}^{n \times m}$, the function $\sigma_k(F)$ is strictly increasing in $k \in [1, n]$ whenever the diagonal of FF^T is positive, with*

$$\sigma_1(F) = \max_{i=1,\ldots,n} (FF^T)_{ii} \quad and \quad \sigma_n(F) = L(F)$$

where $L(F)$ is defined in Lemma 2.3.

Proof. We can write

$$\begin{aligned}
\sigma_k^2(F) &= \max_{\{(u_+,u_-)\in\{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \left\| (u_+ - u_-)^T F \right\|_2^2 \\
&= \max_{\{v\in\{0,1\}^n, \mathbf{1}^T v \leq k, u \in \{-1,1\}^n\}} u^T (vv^T \circ FF^T) u.
\end{aligned}$$

Let us call $v(k), u(k)$ the optimal solutions of the maximization problem with optimal value $\sigma_k^2(F)$, and let $J = \{i \in [1, n] : v(k)_i \neq 0\}$ be the support of $v(k)$. If we pick $i \in [1, n]$, outside of J, we have

$$\begin{aligned}
\sigma_{k+1}^2(F) &\geq u(k)^T (v(k)v(k)^T \circ FF^T) u(k) + (FF^T)_{ii} + \max_{u_i \in \{-1,1\}} 2u_i \left(\sum_{j \in J} u(k)_j (FF^T)_{ij} \right) \\
&= \sigma_k^2(F) + (FF^T)_{ii} + 2 \left| \sum_{j \in J} u(k)_j (FF^T)_{ij} \right|
\end{aligned}$$

Hence the difference between $\sigma_{k+1}^2(F)$ and $\sigma_k^2(F)$ is at least $\max_{j \in J} (FF^T)_{jj}$. This means that $\sigma_k(F)$ is increasing and bounded by

$$\max_{u \in \{-1,1\}^n} u^T FF^T u,$$

which is the maximization problem defining $L^2(F)$ in Lemma 2.3. ∎

2.3. **Independent, bounded model.** The previous section showed that enforcing condition (proba-NSP) with high probability for Gaussian vectors y meant controlling the ratio between the Lipschitz constant $\sigma_k(F)$ of the norm $\|Fy\|_{1,k}$ and the norm $\sum_{i=1}^n \|F_i\|_2$. In what follows, we will show that the same quantities control the concentration of $\|Fy\|_{1,k}$ and $\|Fy\|_1$ when the coefficients of y are independent and bounded. Once again, because F is defined up to a rotation here, these results are easily extended to the case where $y = Qu$ with $Q^T Q = \mathbf{I}$ and the variables u are independent and bounded. We can write a weak recovery condition for this bounded model, similar to condition (5).

Proposition 2.6. *Let $F \in \mathbf{R}^{n \times m}$ and suppose*

$$\mathbf{E}[\|Fy\|_{1,k}] + \beta\sigma_k(F) \leq (\mathbf{E}[\|Fy\|_1] - \beta L(F))\alpha_k \tag{6}$$

for some $\beta > 0$, where $\sigma_k(F)$ was defined in (3) and $L(F)$ in (4), then the sparse recovery condition (proba-NSP)

$$\|Fy\|_{k,1} \leq \alpha_k \|Fy\|_1$$

will be satisfied with probability $1 - 2ce^{-\beta^2/c\Delta^2}$, where $c > 0$ is an absolute constant, when the coefficients of $y \in \mathbf{R}^m$ are independent and bounded, with $\|y\|_\infty \leq \Delta$.

Proof. As pointwise suprema of affine functions in y, the functions $\|Fy\|_{1,k}$ and $\|Fy\|_1$ are convex and Lipschitz with constants bounded by $\sigma_k(F)$ and $L(F)$ respectively (see the proofs of Lemmas 2.2 and 2.3). If the coefficients of $y \in \mathbf{R}^m$ are independent and bounded, with $\|y\|_\infty \leq \Delta$, Talagrand's inequality (Ledoux, 2005, Corr. 4.10) then shows that

$$\mathbf{Prob}\left[|\, \mathbf{E}[\|Fy\|_{1,k}] - \|Fy\|_{1,k}| \geq t \right] \leq Ce^{-\frac{t^2}{c\sigma_k^2(F)\Delta^2}}$$

and

$$\mathbf{Prob}\left[|\,\mathbf{E}[\|Fy\|_1] - \|Fy\|_1|\geq t\right] \leq Ce^{-\frac{t^2}{cL^2(F)\Delta^2}}$$

where c is an absolute constant, hence the desired result. ∎

The parallel with the Gaussian case can be made even more explicit using the following simple majorization result.

Lemma 2.7. *Let $V \subset \mathbf{R}^n$ be a finite set. Suppose the variables $\{y_i\}_{i=1,\dots,n}$ are independent with support in $[-1, 1]$, then*

$$\mathbf{E}[\sup_{v\in V} v^T y] \leq \sigma\sqrt{\pi \log |V|}$$

where $\sigma = \max_{v\in V} \|v\|_2$.

Proof. If the variables y_i are independent, supported in $[-1, 1]$, then $y \preceq_c g$ where $g \sim \mathcal{N}(0, \frac{\pi}{2}\mathbf{I}_n)$ is a Gaussian vector (Ben-Tal et al., 2009, Prop. 10.3.2). The supremum $\sup_{v\in V} v^T y$ is a pointwise maximum of affine functions of y, hence is convex in y, so $y \preceq_c g$ implies $\mathbf{E}[\sup_{v\in V} v^T y] \leq \mathbf{E}[\sup_{v\in V} v^T g]$. Finally, (Massart, 2007, Th. 3.12) shows that $\mathbf{E}[\sup_{v\in V} v^T g] \leq \sigma\sqrt{\pi \log |V|}$. ∎

If we take V in Lemma 2.7 to be the set of vectors of size n with exactly k entries equal to $+1$ or -1, and $n - k$ zeroes, this result shows that, when the coefficients of y are supported on $[-1, 1]$ and independent, then $\mathbf{E}[\|Fy\|_{1,k}]$ is bounded by $\frac{\pi}{2}\mathbf{E}[\|Fg\|_{1,k}]$ with g Gaussian. Alternatively, both expectations in (6) can be evaluated efficiently. In fact Hoeffding's inequality shows that if we need to estimate these quantities with precision ϵ and confidence $1 - \beta$, we need at least N samples of either $\|Fy\|_{1,k}$ or $\|Fy\|_1$, with

$$N = \frac{D^2 \log(2/\beta)}{2\epsilon^2}$$

where $D = \max_{\|y\|_\infty \leq \Delta} \|Fy\|_1$ is an upper bound on both norms whenever $\|y\|_\infty \leq \Delta$.

3. Weak recovery and restricted isometry

In this section, we show that some random matrices satisfy our weak recovery condition (5) for near optimal values of the cardinality k (i.e. in scenarios where the number m of linear samples required to recover a signal is a small multiple of the number k of nonzero components in that signal). We show in particular that in some cases, matrices satisfying the restricted isometry property defined in Candès and Tao (2005) also satisfy condition (5). Here however, the restricted isometry is tested on the nullspace basis F instead of the coding matrix A, so the compressed sensing interpretation is lost, but this connection allows us to *recycle all known results on restricted isometry thresholds for random matrices* and easily derive weak recovery thresholds from condition (5). In the next section, we will see that the most important difference between RIP and the weak condition detailed here is that (5) can be tested efficiently while RIP is intractable. Here, we simply check that our weak condition (5) is indeed satisfied by good coding matrices.

We first show that the k-Dense-Subgraph problem computing $\sigma_k(F)$ in (3) is inherently simpler than the sparse eigenvalue problem used in testing the restricted isometry property. We then show that matrices F such that F^T satisfies the restricted isometry property defined in Candès and Tao (2005) at a near-optimal cardinality k, also satisfy our weak recovery condition (5) for similar values of k. This allows us to recycle all known results on the RIP for random matrices and show in particular that Gaussian matrices satisfy condition (5) at near optimal values of k.

Roughly speaking, our main objective here is to show that for good coding matrices $\sigma_k(F)$ grows as $\sqrt{k}$ while $L(F)$ is of order $\sqrt{n}$ and $\sum_{i=1}^{n} \|F_i\|_2$ is of order n (up to a normalizing factor in condition (5)). For completeness, we have also included a direct proof of these facts in the appendix, using standard concentration arguments instead of RIP.

3.1. **Sparse eigenvalues and k-Dense-Subgraph.** We will see in the next section that approximating the k-Dense-Subgraph problem is significantly easier than testing RIP or the nullspace property. There is in fact a direct connection between the sparse eigenvalue and k-Dense-Subgraph problems. The k-Dense-Subgraph problem used in bounding $\sigma_k(F)$ is written

$$\sigma_k^2(F) = \max_{\substack{u \in \{0,1\}^{2n} \\ \mathbf{1}^T u \leq k}} u^T M u \quad \text{where} \quad M = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes F F^T$$

in the variable $u \in \{0,1\}^{2n}$. On the other hand, the problem of computing a sparse maximum eigenvalue to check the restricted isometry property can be written

$$\lambda_{\max}^k(FF^T) = \max_{\substack{u \in \{0,1\}^n \\ \mathbf{1}^T u \leq k}} \max_{\|x\|=1} u^T(FF^T \circ xx^T)u$$

in the variables $x \in \mathbf{R}^n$, $u \in \{0,1\}^n$. We observe that computing sparse eigenvalues (hence test RIP) means solving a k-Dense-Subgraph problem over the result of an inner eigenvalue problem in x, while bounding $\sigma_k(F)$ only requires solving a k-Dense-Subgraph problem over a fixed matrix M, hence is significantly easier.

3.2. **Weak NSP and random matrices.** Following Candès and Tao (2005), we will say that a matrix $A \in \mathbf{R}^{m \times n}$ satisfies the *restricted isometry property* (RIP) at cardinality $k > 0$ if there is a constant $\delta_k > 0$ such that

$$\|x\|_2^2(1 - \delta_k) \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

for all sparse vectors $x \in \mathbf{R}^n$ such that $\mathbf{Card}(x) \leq k$. We now show that the RIP allows us to closely control the values of $\sigma_k(F)$ and $L(F)$. This will allow us to directly recycle all known results on RIP for random matrices and apply them to the weak recovery condition considered here. We start by a technical lemma bounding the values of $\sigma_k(F)$ and $\|F_i\|_2$ for RIP matrices.

Lemma 3.1. *Suppose the matrix $F^T \in \mathbf{R}^{m \times n}$ satisfies the restricted isometry property with constant $\delta_k > 0$ at cardinality k, then*

$$\sigma_k(F) \leq \sqrt{k(1 + \delta_k)} \quad \text{and} \quad \|F_i\|_2 \geq \sqrt{1 - \delta_1}. \tag{7}$$

and $(k/n)^2 L^2(F) \leq \sigma_k^2(F)$.

Proof. We get

$$
\begin{aligned}
\sigma_k^2(F) &= \max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \left\| (u_+ - u_-)^T F \right\|_2^2 \\
&= \max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} (u_+ - u_-)^T F F^T (u_+ - u_-) \\
&\leq (1 + \delta_k) \max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \|u_+ - u_-\|_2^2 \\
&\leq (1 + \delta_k)k
\end{aligned}
$$

because F^T satisfies the RIP and $\mathbf{Card}(u_+ - u_-) \leq k$. Plugging Euclidean basis vectors in the RIP also means $(1 - \delta_1) \leq \|F_i\|_2^2$ for $i = 1, \dots, n$. Lemma 2.5 showed that $L(F) = \sigma_n(F)$ and combining this with the lower bound in (Srivastav and Wolf, 1998, Lem.1) on the performance of the greedy algorithm in §4.2.1 shows that $(k/n)^2 L^2(F) \leq \sigma_k^2(F)$. ∎

We now use this last lemma to show the main result of this section, which proves that if a matrix F satisfies RIP then F^T will satisfy the weak recovery condition (5) in the optimal regime where k is proportional to n. In other words, this result shows that our weak recovery condition is satisfied by optimal matrices, hence is indeed weaker than existing recovery conditions.

Proposition 3.2. *Suppose $F^T \in \mathbf{R}^{m \times n}$ satisfies the restricted isometry property with constant δ_k with $0 < \delta_k < c < 1$ at cardinality k, where c is an absolute constant. Suppose that $k \leq n$, $k \to \infty$ as $n \to \infty$ and $\limsup_{n \to \infty} k/n = \kappa$. Then F satisfies condition (5) for n large enough with $\alpha_k < 1/2$, provided that $f_c(\kappa) < 1/2$, where f_c is defined as*

$$f_c(x) = x \frac{\sqrt{\pi(1+c)}}{\sqrt{1-c}} \sqrt{\left(1 + \log \frac{2}{x}\right)},$$

and $f_c(0) = 0$.

Proof. When F^T satisfies the RIP, Lemma 3.1 above shows

$$\sigma_k(F) \leq \sqrt{k(1+\delta_k)}$$

and, using $L(F) \leq (n/k)\sigma_k(F)$ (see Lemma 3.1), we then get $L(F) \leq nk^{-1/2}\sqrt{(1+\delta_k)}$. Therefore,

$$\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F) \geq n\sqrt{\frac{2(1-\delta_1)}{\pi}} - \beta n \sqrt{(1+\delta_k)/k}$$

for any $\beta > 0$. We also note that $\delta_1 \leq \delta_k < c$ so that

$$\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F) \geq n\sqrt{\frac{2(1-\delta_k)}{\pi}} - \beta n \sqrt{(1+\delta_k)/k}$$

$$> n \left(\sqrt{\frac{2(1-c)}{\pi}} - \beta\sqrt{(1+c)/k} \right).$$

Using the fact that $\sigma_k(F) \leq \sqrt{k(1+c)}$, it is clear that if

$$\alpha_k \left[\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F) \right] \geq \sqrt{k(1+c)} \left[\sqrt{2k\left(1 + \log \frac{2n}{k}\right)} + \beta \right],$$

then Equation (5) holds. Therefore, if

$$\alpha_k n \left(\sqrt{\frac{2(1-c)}{\pi}} - \beta\sqrt{(1+c)/k} \right) \geq \sqrt{k(1+c)} \left[\sqrt{2k\left(1 + \log \frac{2n}{k}\right)} + \beta \right],$$

or equivalently, assuming $k > \pi\beta^2(1+c)/2(1-c)$, if

$$\alpha_k \geq \frac{k}{n} \frac{\sqrt{1+c}}{\sqrt{\frac{2(1-c)}{\pi}} - \beta\sqrt{\frac{1+c}{k}}} \left[\sqrt{2\left(1 + \log \frac{2n}{k}\right)} + \frac{\beta}{\sqrt{k}} \right] \triangleq \Gamma(k, n, c, \beta),$$

then Equation (5) holds. It is therefore clear that if $\Gamma(k, n, c, \beta) < 1/2$, we can find $\alpha_k < 1/2$ such that Equation (5) holds. Notice that as $k \to \infty$, we have

$$\Gamma(k, n, c, \beta) \sim \frac{k}{n} \frac{\sqrt{\pi(1+c)}}{\sqrt{1-c}} \sqrt{1 + \log \frac{2n}{k}} = f_c(k/n).$$

Elementary analysis shows that f_c is a continuous increasing function on $[0, 1]$.

Recall now, that by assumption, $k \to \infty$ as $n \to \infty$ and $\limsup_{n \to \infty} \frac{k}{n} = \kappa \geq 0$ with κ such that $f_c(\kappa) < 1/2$. We therefore conclude - by considering $\limsup_{n \to \infty} \Gamma(k, n, c, \beta)$ - that when n is large enough, Equation (5) holds with $\alpha_k < 1/2$ under our assumptions. ∎

This last result shows that F satisfies the weak recovery condition in (5) at cardinalities near k when F^T satisfies the RIP at cardinality k, in the optimal regime where k is proportional to n.

4. Bounds on $L(F)$ and $\sigma_k(F)$ using graph partitioning relaxations

In Section 2.2, we showed that if the matrix $F \in \mathbf{R}^{n \times m}$ satisfied the weak recovery condition (5), which read

$$\left(\sqrt{2k \log\left(1 + \frac{2n}{k}\right)} + \beta\right) \sigma_k(F) \leq \left(\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F)\right) \alpha_k,$$

for some $\beta > 0$, then the recovery condition in (proba-NSP) would be satisfied with probability $1 - 2e^{-\beta^2/2}$ when y is Gaussian. Testing this weak recovery condition essentially hinged on bounding the Lipschitz constants $\sigma_k(F)$ and $L(F)$. In Section 2.3 we showed that the same quantities allowed us to check the weak recovery condition in a more general model where y is bounded. As we will see below, efficient approximation results on these graph partitioning problems produce relatively tight bounds on both $\sigma_k(F)$ and $L(F)$. In particular, these bounds are tight enough to allow condition (proba-NSP) to be tested in polynomial time at near-optimal values of the cardinality k.

4.1. **Bounding $L(F)$: MaxCut.**

We have observed in Lemma 2.3 that the constant $L(F)$ on the right hand side of condition (5) is defined as

$$L^2(F) = \max_{v \in \{-1,1\}^n} v^T F F^T v. \tag{8}$$

This is an instance of a graph partitioning problem similar to MaxCut. Goemans and Williamson (1995) (when the matrix is nonnegative) and Nesterov (1998a) show that the following relaxation

$$L^2(F) \leq L^2_{\mathrm{mxct}}(F) = \begin{array}{ll} \max. & \mathbf{Tr}(XFF^T) \\ \text{s.t.} & \mathbf{diag}(X) = \mathbf{1}, X \succeq 0, \end{array} \tag{9}$$

which is a (convex) semidefinite program in the variable $X \in \mathbf{S}_n$, is tight up to a factor $\pi/2$. This means that $\sqrt{2/\pi} L_{\mathrm{mxct}}(F) \leq L(F) \leq L_{\mathrm{mxct}}(F)$. The dual of this last program is written

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T w \\ \text{subject to} & FF^T \preceq \mathbf{diag}(w), \end{array}$$

which is another semidefinite program in the variable $w \in \mathbf{R}^n$. By weak duality, any feasible point of this last problem gives an upper bound on $L(F)$.

4.2. **Bounding $\sigma_k(F)$: k-Dense-Subgraph.**

On the left hand side of (5), the constant $\sigma_k^2(F)$ is computed as

$$\begin{array}{ll} \sigma_k^2(F) = & \max. \quad u^T M u \\ & \text{s.t.} \quad \mathbf{1}^T u \leq k \\ & \qquad u \in \{0, 1\}^{2n}, \end{array} \tag{10}$$

in the binary variable u, where $M \in \mathbf{S}_{2n}$ is positive semidefinite, with

$$M = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes FF^T, \tag{11}$$

here. This is a graph partitioning problem known as k-Dense-Subgraph, which seeks to find a subgraph S of the graph of M, with at most k nodes and maximum edge weight $\sum_{(i,j) \in S} M_{ij}$, see Kortsarz and Peleg (1993); Arora et al. (1995); Feige et al. (2001); Feige and Langberg (2001); Han et al. (2002a); Billionnet and Roupin (2006) among others for details. Note that in our application here, M is typically dense and its coefficients can take negative values while most of the references cited above consider graphs with non-negative (often sparse) weight matrices. The k-DenseSubgraph problem can also be seen as an instance of the Quadratic Knapsack problem (see Lin (1998); Pisinger (2007) for a general overview). We will see that elementary greedy or random sampling algorithms already produce satisfactory approximations. However, their crudeness means that they are outperformed in practice by linear programming or semidefinite relaxation bounds, and we begin by outlining a few of these relaxations below.

4.2.1. *A Greedy Algorithm.* We now recall the greedy elimination procedure described by e.g. Srivastav and Wolf (1998), which extracts a k-subgraph out of a larger graph containing the optimal solution. Suppose we are given a weight matrix $M \in \mathbf{S}_n$, and assume we know an index set $I \in [1, n]$ such that the weight $w(I) = \sum_{i,j \in I} M_{ij}$ of the subgraph with vertices in I is an upper bound on $\sigma_k^2(F)$ of the k-Dense-Subgraph problem in (10). If $|I| \leq k$, then I is optimal, otherwise we can greedily prune $|I| - k$ vertices from the graph and Srivastav and Wolf (1998, Lem.1) show that the pruned subgraph must have weight at least

$$\frac{k(k-1)}{|I|(|I|-1)} w(I).$$

When the weight matrix M is nonnegative, the full graph weight $w([1, n])$ produces an obvious upper bound on $w(I^*)$. The situation is slightly more complex when M has negative coefficients, as in the particular instance considered here in (3). In Proposition 6.1, we show how to produce an upper bound $w(I)$ by solving the MaxCut relaxation (9).

4.2.2. *Semidefinite Relaxation.* Many different relaxations have been developed for the k-Dense-Subgraph and Quadratic Knapsack problem and we highlight some of them in what follows. Semidefinite relaxations were derived in Helmberg et al. (2000) to bound $\sigma_k^2(F)$. In particular, the SQK2 relaxation in Helmberg et al. (2000) yields

$$
\begin{aligned}
\sigma_k^2(F) \leq \quad \text{max.} \quad & \mathbf{Tr}(MX) \\
\text{s.t.} \quad & \mathbf{1}^T X \mathbf{1} \leq k^2 \\
& X - \mathbf{diag}^2(X) \succeq 0,
\end{aligned}
\tag{12}
$$

which is a semidefinite program in the variable $X \in \mathbf{S}_n$. Note that the constraint $X - \mathbf{diag}^2(X)$ is a Schur complement, hence is convex in X. Adaptively adding further constraints as in Helmberg et al. (2000) can further tighten this relaxation. In particular, adding constraints of the type

$$\sum_{j=1}^n X_{ij} \leq k X_{ii} \quad \text{or} \quad \sum_{j=1}^n (X_{jj} - X_{ij}) \leq (1 - X_{ii}) \tag{13}$$

for some $i = 1, \ldots, n$, sometimes significantly improves tightness. Another simple relaxation formulated in Helmberg et al. (2000) bounds (8) when $k \geq 2$ by solving

$$
\begin{aligned}
\sigma_k^2(F) \leq \quad \text{max.} \quad & \mathbf{Tr}(MX) \\
\text{s.t.} \quad & \mathbf{Tr}((\mathbf{11}^T - \mathbf{I})X) \leq k(k-1) \\
& X - \mathbf{diag}^2(X) \succeq 0,
\end{aligned}
\tag{14}
$$

in the variable $X \in \mathbf{S}_n$. This last relaxation is tighter than (12) but not as tight as its refinements using the additional constraints in (13). Another relaxation detailed in Feige and Langberg (2001) first writes (10) as a binary optimization problem over $\{-1, 1\}^n$, then bounds it by solving

$$
\begin{aligned}
\text{maximize} \quad & \mathbf{Tr}(M(\mathbf{11}^T + y\mathbf{1}^T + \mathbf{1}y^T + Y)) \\
\text{subject to} \quad & Y\mathbf{1} = y(2k - n) \\
& \mathbf{diag}(Y) = \mathbf{1}, Y \succeq 0,
\end{aligned}
\tag{15}
$$

which is a semidefinite program in the variable $Y \in \mathbf{S}_n$. We refer the reader to Helmberg et al. (2000) for details on the tightness and complexity of these various semidefinite relaxations.

Fortunately, even though the k-Dense-Subgraph problem is NP-Hard, simple randomized or greedy algorithms reach good approximation ratios (Arora et al. (1995) even produced a PTAS in the dense nonnegative case). While many tightness results have been derived on the semidefinite relaxations detailed above (see e.g. Han et al. (2002b)), most of them producing approximation ratios of k/n or better, existing results do not apply when *the coefficients of M have arbitrary signs*. Here, we show a similar approximation ratio when the graph weight matrix M is allowed to have some negative coefficients but is positive semidefinite.

Proposition 4.1. *Suppose $M \in \mathbf{S}_n$ is positive semidefinite. Define*

$$\mathcal{D}_k(M) = \max_{\substack{u \in \{0,1\}^n \\ \mathbf{1}^T u \leq k}} u^T M u,$$

the relaxation

$$SDP_k(M) = \begin{array}{ll} max. & \mathbf{Tr}\, MX \\ s.t. & 0 \leq X_{ij} \leq 1 \\ & \mathbf{Tr}\, X = k,\, X \succeq 0, \end{array} \tag{16}$$

satisfies, for n large enough and $k \geq n^{1/3}$,

$$\frac{k}{n}\mu(n,k)\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}SDP_k(M)\right) \leq \mathcal{D}_k(M) \leq SDP_k(M),$$

where

$$\mu(n,k) = \left(1 - \frac{2}{k^{1/3}}\right)\left(\frac{1}{1 - \frac{2\pi n^2}{k^2}e^{-\frac{n^{1/9}}{3}}}\right) \xrightarrow[n\to\infty]{} 1$$

and $G_{ij} = \sqrt{X_{ii}X_{jj}}$, $i,j = 1,\ldots,n$, so in particular $\mathbf{Tr}\, MG \geq 0$.

Proof. We use a hybrid randomization procedure, mixing the sparse sampling strategy in Feige and Seltser (1997) with the correlation argument in Nesterov (1998a). Let X be an optimal solution to problem (16), w.l.o.g. we can assume $|X_{ii}| > 0$, and we define the corresponding (positive semidefinite) correlation matrix $C_{ij} = X_{ij}/\sqrt{X_{ii}X_{jj}}$, $i,j = 1,\ldots n$ and sample vectors $z \sim \mathcal{N}(0,C)$. For each sample z, we define

$$y_i = \begin{cases} 1 & \text{if } z_i \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

As in Feige and Seltser (1997), we also sample independent variables $u \in \mathbf{R}^n$ such that

$$u_i = \begin{cases} 1 & \text{with probability } q_i = k\sqrt{X_{ii}}/S, \\ 0 & \text{otherwise.} \end{cases}$$

where $S = \sum_{i=1}^n \sqrt{X_{ii}}$. Note that $0 \leq q_i \leq 1$ because $0 \leq X_{ii} \leq 1$ and $\sum_i X_{ii} = k$. For each sample, we then define $w \in \{0,1\}^n$, with $w_i = u_i y_i$, $i = 1,\ldots,n$, so when $i \neq j$

$$\begin{aligned} \mathbf{E}[w_i w_j] &= \mathbf{Prob}[z_i \geq 0, z_j \geq 0, u_i = u_j = 1] \\ &= \mathbf{Prob}[z_i \geq 0, z_j \geq 0]\,\mathbf{Prob}[u_i = 1]\,\mathbf{Prob}[u_j = 1] \\ &= \left(\frac{1}{4} + \frac{1}{2\pi}\arcsin(C_{ij})\right)\frac{k^2\sqrt{X_{ii}X_{jj}}}{S^2} \end{aligned}$$

and $\mathbf{E}[w_i^2] \geq \mathbf{Prob}[z_i \geq 0]\,\mathbf{Prob}[u_i = 1]^2$. If we define $G \in \mathbf{S}_n$ with $G_{ij} = \sqrt{X_{ii}X_{jj}}$, we conclude that

$$\mathbf{E}[ww^T] \succeq \tfrac{k^2}{S^2}\left[\tfrac{1}{4}G + \tfrac{1}{2\pi}\arcsin(C) \circ G\right].$$

Because $X, M \succeq 0$ with $\mathbf{Tr}\, X = k$, we have $S \leq \sqrt{kn}$, and we thus obtain

$$\begin{aligned} \mathbf{E}[w^T M w] &\geq \frac{k^2}{S^2}\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}\mathbf{Tr}(M(\arcsin(C) \circ G))\right) \\ &\geq \frac{k}{n}\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}SDP_k(M)\right) \end{aligned}$$

because $\arcsin(C) \succeq C$ (Nesterov, 1998b, Corr. 3.2), $\mathbf{Tr}(M(\arcsin(C) \circ G)) = \mathbf{Tr}(\arcsin(C)(M \circ G))$, $C \circ G = X$ and $M, C, G \succeq 0$ so $M \circ G \succeq 0$. Now, let us call $b = \mathbf{Prob}[w^T M w \leq \mathbf{E}[w^T M w]/\beta]$ for some $\beta \geq 1$. By construction, because $w^T M w \leq SDP_n(M)$ whenever $w \in \{0,1\}^n$ and

$$w^T M w \leq \frac{\mathbf{E}[w^T M w]}{\beta}\mathbf{1}_{\{w^T M w \leq \mathbf{E}[w^T M w]/\beta\}} + SDP_n(M)\,\mathbf{1}_{\{w^T M w > \mathbf{E}[w^T M w]/\beta\}}$$

we have

$$\mathbf{E}[w^T M w] \leq b\,\mathbf{E}[w^T M w]/\beta + (1-b)SDP_n(M)$$

so

$$b \leq 1 - \frac{\beta - 1}{\beta SDP_n(M)/\mathbf{E}[w^T M w] - 1}.$$

Now, let us call $Y \in \mathbf{S}_n$ a solution to $SDP_n(M)$; then kY/n is a feasible point of (16), so $SDP_n(M) = \mathbf{Tr}\, MY \leq \frac{n}{k} \mathbf{Tr}\, MX$ and the previous paragraph shows

$$\frac{SDP_n(M)}{\mathbf{E}[w^T M w]} = \frac{\mathbf{Tr}\, MY}{\mathbf{E}[w^T M w]} \leq \frac{2\pi n\, \mathbf{Tr}\, MY}{k\, \mathbf{Tr}\, MX} \leq \frac{2\pi n^2}{k^2}\;.$$

Therefore, for n large enough, setting

$$\beta \geq \frac{1}{1 - \frac{2\pi n^2}{k^2} e^{-k^{1/3}/3}},$$

ensures

$$\beta > \frac{1 - e^{-k^{1/3}/3}}{1 - \frac{SDP_n(M)}{\mathbf{E}[w^T M w]} e^{-k^{1/3}/3}}\;.$$

When the denominator is positive, the previous inequality implies that

$$\frac{\beta - 1}{\beta SDP_n(M)/\mathbf{E}[w^T M w] - 1} > e^{-k^{1/3}/3}\;.$$

Hence, choosing again n large enough to make the denominator positive, we finally have

$$1 - b \geq \frac{\beta - 1}{\beta SDP_n(M)/\mathbf{E}[w^T M w] - 1} > e^{-k^{1/3}/3},$$

Now, using Chernoff's inequality as in (Feige and Seltser, 1997, Lem. 4.1) produces

$$\mathbf{Prob}\left[\mathbf{Card}(u) - \mathbf{1}^T q \geq t \mathbf{1}^T q\right] \leq e^{-\frac{t^2 \mathbf{1}^T q}{3}},$$

where $q_i = \mathbf{Prob}[u_i = 1]$. We note that here $\mathbf{1}^T q = k$ and as in (Feige and Seltser, 1997, Th. 4.1), when $k \geq n^{1/3}$

$$\mathbf{Prob}\left[\mathbf{Card}(u) \geq k\left(1 + k^{-1/3}\right)\right] \leq e^{-k^{1/3}/3}.$$

This last result, together with the bound on b derived above, shows that

$$\mathbf{Prob}[w^T M w \geq \mathbf{E}[w^T M w]/\beta] = 1 - b > e^{-k^{1/3}/3} \geq \mathbf{Prob}\left[\mathbf{Card}(w) \geq k\left(1 + k^{-1/3}\right)\right]\;.$$

Therefore, by sampling enough points w, we can generate a vector $w_0 \in \{0,1\}^n$ such that

$$w_0^T M w_0 \geq \frac{k}{\beta n}\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}SDP_k(M)\right) \quad \text{and} \quad \mathbf{Card}(w_0) \leq k\left(1 + k^{-1/3}\right)$$

If we remove no more than $k^{2/3}$ variables from w_0 using the backward greedy algorithm described in Srivastav and Wolf (1998, Lem.1) we loose at most a factor

$$\frac{k(k-1)}{(k + k^{2/3})(k + k^{2/3} - 1)} = 1 - \frac{2}{k^{1/3}} + o\left(\frac{1}{k^{1/3}}\right)$$

and, from w_0, we obtain a point w_k such that

$$w_k^T A w_k \geq \frac{k}{\beta n}\left(1 - \frac{2}{k^{1/3}}\right)\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}SDP_k(M)\right) \quad \text{and} \quad \mathbf{Card}(w_k) \leq k,$$

when n is large enough, which yields the desired result. ∎

Note that, in the previous result, the condition $k \geq n^{1/3}$ can be replaced by any constraint of the type $k \geq n^\alpha$ where $0 < \alpha < 1$ with $n^{1/9}$ replaced by $n^{\alpha/3}$.

5. COMPLEXITY

Bounding $L(F)$ and $\sigma_k(F)$ using semidefinite relaxations means solving two maximum eigenvalue minimization problems. Problem (9), used for bounding $L(F)$, can be rewritten

$$\min_{w \in \mathbf{R}^n} \; n\lambda_{\max}(FF^T - \mathbf{diag}(w)) - \mathbf{1}^T w \tag{17}$$

while problem (12) bounding $\sigma_k(F)$ can be written

$$\min_{\{w, z \in \mathbf{R}^n, \, y \in \mathbf{R}^n\}} (k+1)\lambda_{\max}\left(\bar{F} + w\bar{H} + z\bar{G} + \sum_{i=1}^n y_i \bar{E}_i\right) - wk(k-1) - z \tag{18}$$

where

$$\bar{F} = \begin{pmatrix} FF^T & 0 \\ 0 & 0 \end{pmatrix}, \; \bar{H} = \begin{pmatrix} \mathbf{11}^T - \mathbf{I} & 0 \\ 0 & 0 \end{pmatrix}, \; \bar{G} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \bar{E}_i = \begin{pmatrix} e_i e_i^T & -e_i/2 \\ -e_i^T/2 & 0 \end{pmatrix}$$

where $e_i \in \mathbf{R}^n$ is the ith Euclidean basis vector. Given a priori bounds on the norm of the solutions, Nesterov (2007) showed that solving problems (18) and (17) up to a target precision ϵ using first-order methods has total complexity growing as

$$O\left(\frac{n^3 \sqrt{\log n}}{\epsilon}\right) \quad \text{and} \quad O\left(\frac{n^{3.5} \sqrt{\log n}}{\epsilon}\right)$$

for problems (17) and (18) respectively.

6. TIGHTNESS

We use the convex relaxation result of Proposition 4.1 to show that if a matrix F satisfies the weak recovery condition (5) up to cardinality k^*, the semidefinite relaxation in (16) will allow us to certify that F satisfies (5) at cardinalities very near k^*.

Proposition 6.1. *Suppose the matrix $F \in \mathbf{R}^{n \times m}$ satisfies the weak recovery condition (5) up to cardinality $k^* = \gamma(n)n$ for some $\gamma(n) \in (0, 1)$, $\beta > 0$ and $\alpha_{k^*} \in [0, 1]$, i.e.*

$$\left(\sqrt{2k^* \log \frac{2n}{k^*}} + \beta\right)\sigma_{k^*}(F) \le \left(\sqrt{\frac{2}{\pi}} \sum_{i=1}^n \|F_i\|_2 - \beta L(F)\right)\alpha_{k^*},$$

and let $SDP_k(\cdot)$ be defined as in (16), we have

$$\left(\sqrt{2k \log \frac{2n}{k}} + \beta\right)(SDP_k(M))^{1/2} \le \left(\sqrt{\frac{2}{\pi}} \sum_{i=1}^n \|F_i\|_2 - \beta L(F)\right)\alpha_{k^*}, \tag{19}$$

for n sufficiently large, when $k \le \gamma(n)(\log n)^{-1}k^*$, with M defined as in (11).

Proof. Applying the result of Proposition 4.1 at cardinality k^* shows

$$(SDP_{k^*}(M))^{1/2} \le \sigma_{k^*}(F)\sqrt{\frac{2\pi n}{k^*}}\left(1 + \frac{o(1)}{k^{*1/3}}\right)^{1/2}.$$

Using $SDP_k(M) \le SDP_{k^*}(M)$, with $k \le \gamma(n)(\log n)^{-1}k^*$ showing

$$\frac{\left(\sqrt{2k \log \frac{2n}{k}} + \beta\right)}{\left(\sqrt{2k^* \log \frac{2n}{k^*}} + \beta\right)}\sqrt{\frac{2\pi n}{k^*}}\left(1 + \frac{o(1)}{k^{*1/3}}\right)^{1/2} = o(1)$$

when $n \to \infty$, yields the desired result. ∎

7. Numerical Results

We start by studying the distribution of the residual error $x^{\mathrm{lp}} - e$ when e is a random sparse signal. We sample a thousand vectors $e \in \mathbf{R}^{100}$ with 15 nonzero i.i.d. uniform coefficients. Our (fixed) design matrix $A \in \mathbf{R}^{m \times n}$ is Gaussian or Bernoulli with $m = 30$. We produce a vector of observations Ae and solve the ℓ_1 reconstruction problem in (ℓ_1-recov.) and record the value of $x^{\mathrm{lp}} - e$ projected along a fixed (randomly chosen) direction v. The histogram of these values is plotted in Figure 1.

Projected Error (Gaussian)   Projected Error (Bernoulli)
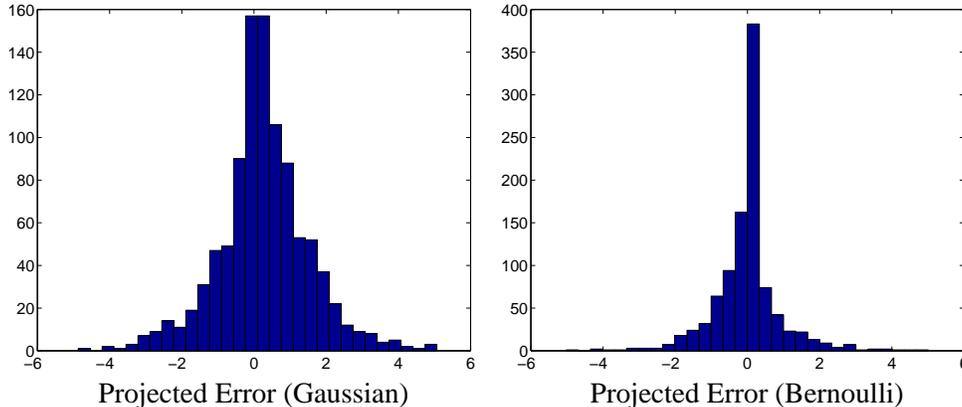
FIGURE 1. Projected reconstruction error $v^T(x^{\mathrm{lp}} - e)$, along a fixed (randomly chosen) direction v, using a single Gaussian (left) or Bernoulli (right) design matrix with $p = 100$, $m = 30$ and a thousand samples of a random sparse signal $e \in \mathbf{R}^{100}$ with 15 i.i.d. uniform coefficients.

On a random Gaussian matrix with $n = 40$ and $m/n = 1/2$, we recall in Table 1 the recovery threshold k/m certified by the semidefinite relaxation (SDP) detailed in (d'Aspremont and El Ghaoui, 2011) and the linear programming (LP) relaxation in (Juditsky and Nemirovski, 2011), strong and weak recovery thresholds from the asymptotic results in Donoho and Tanner (2008).

SDP	LP	Strong D&T	Weak D&T
0.1	0.1	0.1	0.5

TABLE 1. Perfect recovery threshold k/m computed using the semidefinite relaxation (SDP) detailed in (d'Aspremont and El Ghaoui, 2011), the linear programming (LP) relaxation in (Juditsky and Nemirovski, 2011) on a sample Gaussian matrix. We also recall the asymptotic strong and weak recovery thresholds from Donoho and Tanner (2008).

We then sample Gaussian and Bernoulli matrices of increasing dimensions $n \times n/2$ and plot the mean values of the relaxation bounds on $L(F)$ (blue circles), $\sigma_k(F)$ (brown diamonds) together with $\sum_{i=1}^{p} \|F_i\|_2$ (black squares). These quantities are plotted in loglog scale in Figure 2. As expected, the norm grows as n while both $\sigma_k(F)$ and $L(F)$ grow as $\sqrt{n}$. In Figure 3 we plot the empirical (brown squares) versus predicted (blue circles) probability of recovering signals e, where $F \in \mathbf{R}^{n \times m}$ is a Gaussian with $n = 300$ and $m = n/2$, for various values of the relative cardinality k/m. The empirical probability was obtained by solving (ℓ_1-recov.) over one hundred random sparse signal $e \in \mathbf{R}^{100}$ with 15 i.i.d. uniform coefficients. The predicted probability is obtained by computing β from condition (5) after bounding $L(F)$ and $\sigma_k(F)$ using the convex relaxations detailed in Section 4.
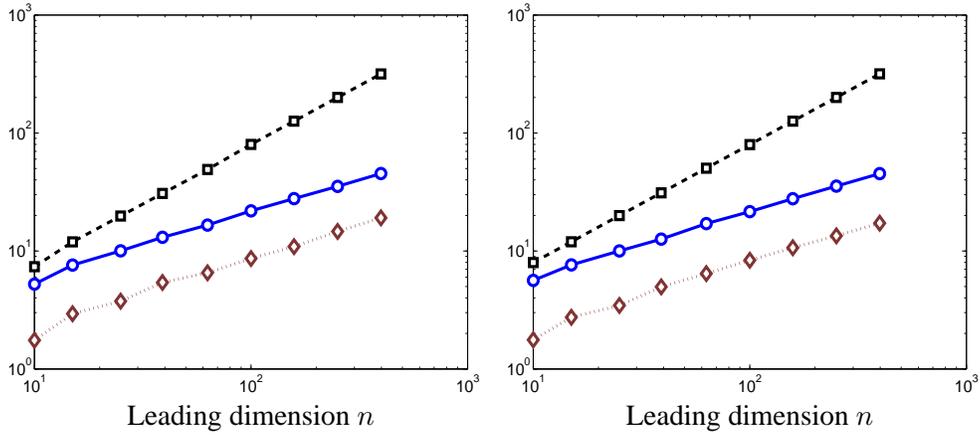
FIGURE 2. *Left:* Loglog plot of mean values of $L(F)$ (blue circles), $\sigma_k(F)$ (brown diamonds) and $\sum_{i=1}^{n} \|F_i\|_2$ (black squares) for Gaussian (left) or Bernoulli (right) matrices of increasing dimensions n, with $m = n/2$.
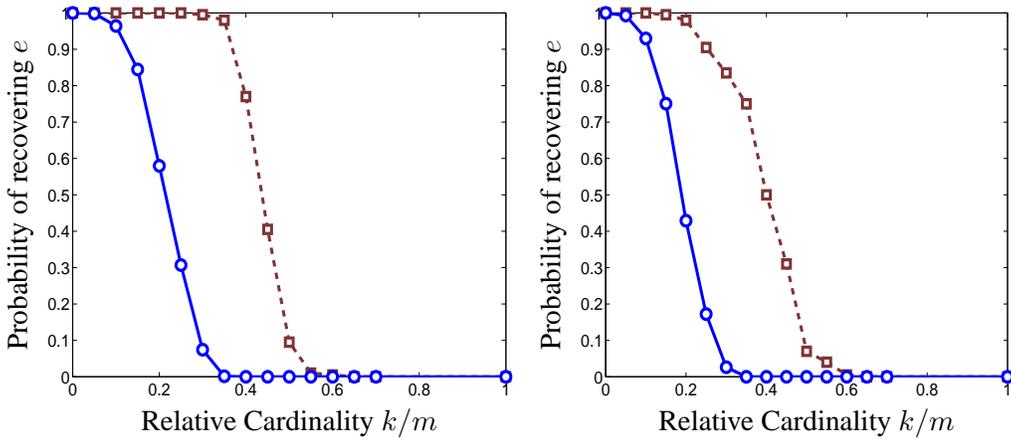

FIGURE 3. Empirical (brown squares) versus predicted (blue circles) probability of recovering the true signal e, where $F \in \mathbf{R}^{n \times m}$ is Gaussian (left) or Bernoulli (right) with $m = n/2$, for various values of the relative cardinality k/m.

8. APPENDIX

Gaussian matrices are known to satisfy the recovery condition (det-NSP) with high probability for near-optimal values of k hence obviously satisfy (proba-NSP). Here we directly verify that these matrices satisfy condition (5) w.h.p. without using RIP. Concentration inequalities have been used in Baraniuk et al. (2008) to derive a simple proof that some classes of random matrices satisfy RIP, we use similar techniques on the weak recovery property (5) here.

We start by bounding the fluctuations of the right hand side of inequality (5) when $F \in \mathbf{R}^{n \times m}$ is a Gaussian random matrix with $F_{ij} \sim \mathcal{N}(0, 1/m)$.

Lemma 8.1. *Let $F \in \mathbf{R}^{n \times m}$ with i.i.d $F_{ij} \sim \mathcal{N}(0, 1/m)$,*

$$\sum_{i=1}^{n} \mathbf{E}\left[\|F_i\|_2\right] = n(1 + O(m^{-1}))$$

as $m \to \infty$.

Proof. In this setting, each $\sqrt{m}\|F_i\|_2$ is χ distributed with m degrees of freedom, so

$$\mathbf{E}[\|F_i\|_2] = \sqrt{\frac{2}{m}} \frac{\Gamma((m+1)/2)}{\Gamma(m/2)}, \quad i = 1, \ldots, n.$$

Using Stirling's formula (Abramowitz and Stegun, 1970, §6.1.37), we get

$$\frac{\Gamma((m+1)/2)}{\Gamma(m/2)} = \frac{\exp\left(\frac{m+1}{2}\log\left(\frac{m+1}{2}\right) - \frac{m}{2}\log\left(\frac{m}{2}\right)\right)}{\sqrt{e(1+1/m)}} + O(m^{-1/2})$$

$$= \sqrt{\frac{m+1}{2}} + O(m^{-1/2})$$

as $m \to \infty$, which is the desired result. ∎

We now use concentration inequalities to bound $\sum_{i=1}^{n} \|F_i\|_2$ in condition (5) with high probability when $F_{ij} \sim \mathcal{N}(0, 1/m)$.

Lemma 8.2. *Let $F \in \mathbf{R}^{n \times m}$ with i.i.d $F_{ij} \sim \mathcal{N}(0, 1/m)$,*

$$\mathbf{Prob}\left[\sum_{i=1}^{n} \|F_i\|_2 \leq \sum_{i=1}^{n} \mathbf{E}\left[\|F_i\|_2\right] - x\right] \leq e^{-\frac{mx^2}{2n}}$$

Proof. For any $U, V \in \mathbf{R}^{m \times n}$, we have

$$\sum_{i=1}^{n} \|U_i\|_2 - \|V_i\|_2 \leq \sum_{i=1}^{n} \|U_i - V_i\|_2$$

$$\leq \sqrt{n}\|U - V\|_F$$

so $\sum_{i=1}^{n} \|F_i\|_2$ is a $\sqrt{n/m}$-Lipschitz function (w.r.t. the Euclidean norm) of nm i.i.d. Gaussian variables $F_{ij}/\sqrt{m} \sim \mathcal{N}(0, 1)$ and (Massart, 2007, Th. 3.4) yields the desired result. ∎

We now turn to the left-hand side of inequality (5) and produce inequalities on $\sigma_k(F)$, using again the fact that it is a Lipschitz function of F.

Lemma 8.3. *Let $F \in \mathbf{R}^{n \times m}$ with i.i.d. $F_{ij} \sim \mathcal{N}(0, 1/m)$,*

$$\mathbf{Prob}\left[\sigma_k(F) \geq \mathbf{E}[\sigma_k(F)] + x\right] \leq e^{-\frac{mx^2}{2k}}$$

Proof. We first note that the max is 1-Lispchitz with respect to the ℓ_∞ norm on $\mathbf{R}^n$. Indeed, if $a, b \in \mathbf{R}^n$

$$|\max_i a_i - \max_j b_j| \leq \max_i |a_i - b_i|,$$

because

$$a_i - \max_j b_j \leq a_i - b_i \leq |a_i - b_i| \leq \max_i |a_i - b_i|, \quad i = 1, \ldots, n.$$

Hence, $\max_i a_i - \max_j b_j \leq \max_i |a_i - b_i|$. The two sequences play symmetric roles so we also have $|\max_j b_j - \max_i a_i| \leq \max_k |a_k - b_k|$. Now our aim is to show that $F \to \sigma_k(F)$ is a Lipschitz function of F with respect to the Euclidian norm. The argument we just gave shows that if F and G are two matrices,

$$|\sigma_k(F) - \sigma_k(G)| \leq \max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \left| \left\|(u_+ - u_-)^T F\right\|_2 - \left\|(u_+ - u_-)^T G\right\|_2 \right|,$$

because $\sigma_k(F)$ and $\sigma_k(G)$ are maxima of finite sequences. We now have

$$
\begin{aligned}
\left| \left\| (u_+ - u_-)^T F \right\|_2 - \left\| (u_+ - u_-)^T G \right\|_2 \right| &\leq \left\| (u_+ - u_-)^T (F - G) \right\|_2 \\
&\leq \left\| (F - G) \right\|_2 \left\| (u_+ - u_-)^T \right\|_2 \\
&\leq \left\| F - G \right\|_F \left\| (u_+ - u_-)^T \right\|_2
\end{aligned}
$$

which shows that

$$
\sigma_k(F) = \max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \left\| (u_+ - u_-)^T F \right\|_2
$$

is a Lipschitz function of the entries of F (with respect to Euclidian norm). Now when the entries of F are i.i.d $\mathcal{N}(0, 1/m)$, $\sigma_k(F)$ is a Lipschitz function of standard Gaussian variables with Lipschitz constant

$$
\max_{\{(u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \frac{\|u_+ - u_-\|_2}{\sqrt{m}} \leq \sqrt{\frac{k}{m}},
$$

and (Massart, 2007, Th. 3.12) yields the desired result. ∎

Next, to bound $\mathbf{E}[\sigma_k(F)]$, we first show a bound on the supremum of an arbitrary number of χ distributed random variables.

Lemma 8.4. *Let $\{y_i\}_{i \in T}$ be χ distributed variables with m degrees of freedom, then*

$$
\mathbf{E}[\sup_{i \in T} y_i] \leq \sqrt{2 \log |T|} + \frac{\sqrt{2}\,\Gamma((m+1)/2)}{\Gamma(m/2)} \leq \sqrt{2 \log |T|} + \sqrt{m} .
$$

We note that the proof we present applies non only to χ distributed random variables but more generally to Lipschitz functions of i.i.d normal random variables.

Proof. Since y_i's have the same mean, we have

$$
\sup_{i \in T} y_i = \mathbf{E}[y_i] + \sup_{i \in T}(y_i - \mathbf{E}[y_i]) .
$$

Here we know that $\mathbf{E}[y_i] = \frac{\sqrt{2}\,\Gamma((m+1)/2)}{\Gamma(m/2)}$ and we also know using Jensen's inequality that $\mathbf{E}[y_i] \leq \sqrt{\mathbf{E}[y_i^2]} = \sqrt{m}$.

The fact that a standard multivariate normal satisfies a log-Sobolev inequality (with constant 1 in the setup of Ledoux (2005, Chap. 5)) implies through the Herbst argument that any 1-Lipschitz function F (with respect to Euclidian norm) of i.i.d Gaussian random variables satisfies (see Ledoux (2005, Eq.5.8))

$$
\log \Psi(z) \triangleq \log \mathbf{E}[\exp\{z(F(X) - \mathbf{E}[F(X)])\}] \leq \frac{z^2}{2} .
$$

The previous inequality naturally applies to y_i's since a χ_m random variable is just the norm of a m-dimensional vector with i.i.d entries (and the norm is 1-Lipschitz by the triangle inequality).

Using a classic approach in probability, namely a "soft-max" inequality, the concavity of the log, the definition of $\Psi(z)$ and the fact that the variables y_i are identically distributed, we now have, if $\tilde{y}_i = y_i - \mathbf{E}[y_i]$,

$$
\begin{aligned}
\mathbf{E}[\sup_{i \in T} \tilde{y}_i] &\leq \frac{1}{z} \mathbf{E}\left[\log \left(\sum_{i \in T} e^{z\tilde{y}_i} \right) \right] \\
&\leq \frac{1}{z} \log \left(\sum_{i \in T} \mathbf{E}\left[e^{z\tilde{y}_i} \right] \right) \\
&\leq \frac{\log |T| + \log \Psi(z)}{z} \leq \frac{\log |T| + z^2/2}{z}
\end{aligned}
$$

for any $z > 0$. Optimizing over z, we get that

$$
\mathbf{E}[\sup_{i \in T} \tilde{y}_i] \leq \sqrt{2 \log |T|} .
$$

This gives the desired result. ∎

Let us now assume that the basis $F \in \mathbf{R}^{n \times m}$ is a Gaussian random matrix (hence A is implicitly defined here as a matrix annihilating F on the left) with $F_{ij} \sim \mathcal{N}(0, 1/m)$. As detailed below and throughout this appendix, standard concentration arguments allow us to directly show that F satisfies condition (5), without resorting to the restricted isometry property. We assume that m scales proportionally to n, with $m = \mu n$ as n goes to infinity. We also assume that k scales as $\kappa m u_m$ with $u_m \to 0$ when m and n go to infinity.

Proposition 8.5. *Suppose $m = \mu n$ and $k = \kappa m u_m$ for some $\mu, \kappa \in (0, 1)$, with $u_m \to 0$ as $m \to \infty$. Let $F \in \mathbf{R}^{n \times m}$ be an i.i.d. Gaussian random matrix with $F_{ij} \sim \mathcal{N}(0, 1/m)$ and $\beta > 0$, then F satisfies condition (5) with high probability as n goes to infinity.*

Proof. We first study the left hand side of (5), which reads

$$\left(\sqrt{2k \left(1 + \log \frac{2n}{k} \right)} + \beta \right) \sigma_k(F) \leq \left(\sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \|F_i\|_2 - \beta L(F) \right) \alpha_k,$$

when n goes to infinity. Because $\sqrt{m/k} \left\| (u_+ - u_-)^T F \right\|_2$ is χ distributed with m d.f. whenever $u = (u_+, u_-) \in \{0,1\}^{2n}$ with $\mathbf{1}^T u = k$ and $u_+^T u_- = 0$, Lemma 8.4 shows that for n large enough

$$
\begin{aligned}
\mathbf{E}[\sigma_k(F)] &= \mathbf{E} \left[\max_{\{u = (u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u \leq k\}} \left\| (u_+ - u_-)^T F \right\|_2 \right] \\
&= \mathbf{E} \left[\max_{\{u = (u_+, u_-) \in \{0,1\}^{2n}, \mathbf{1}^T u = k, u_+^T u_- = 0\}} \left\| (u_+ - u_-)^T F \right\|_2 \right] \\
&\leq \sqrt{\frac{k}{m}} \left[\sqrt{2k \left(1 + \log \left(\frac{2n}{k} \right) \right)} + \sqrt{m} \right].
\end{aligned}
$$

Here we have used the fact that the cardinality of the set T over which we are taking a supremum is such that $\log |T| \leq k \left(1 + \log \left(\frac{2n}{k} \right) \right)$, as shown in the proof of Lemma 2.2. We note that for a constant $c > 0$, we have $k \log \left(\frac{2n}{k} \right) \leq c m u_m \log(1/u_m) \ll m$. Therefore, if c denotes a constant that may change from display to display (but does not depend on n or m), we have

$$\mathbf{E}[\sigma_k(F)] \leq c\sqrt{k} \,,$$

and

$$\left(\sqrt{2k \log \frac{2n}{k}} \right) \frac{\mathbf{E}[\sigma_k(F)]}{m} \leq c \frac{k}{m} \sqrt{-\log(u_m)} \leq c \sqrt{-u_m^2 \log(u_m)} \xrightarrow[n \to \infty]{} 0$$

where $c > 0$ does not depend on n. For some arbitrarily small $\nu > 0$, setting $x = n^\nu \sqrt{2k/m}$ in Lemma 8.3, yields

$$\mathbf{Prob} \left[\sigma_k(F) \geq \mathbf{E}[\sigma_k(F)] + n^\nu \sqrt{2k/m} \right] \leq e^{-n^{2\nu}}.$$

We now focus on the right hand side of (5). Lemma 8.1 shows that

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} \mathbf{E}\left[\|F_i\|_2\right]}{m} = \lim_{n \to \infty} \frac{n}{m} = \frac{1}{\mu}.$$

because $\sqrt{m}\|F_i\|_2$ is χ distributed with m degrees of freedom. Setting $x^2 = n^{\nu+1}/m$ in Lemma 8.2 then yields

$$\mathbf{Prob} \left[\frac{\sum_{i=1}^{n} \|F_i\|_2}{m} \leq \frac{\sum_{i=1}^{n} \mathbf{E}\left[\|F_i\|_2\right]}{m} - \frac{n^{\nu+1/2}\sqrt{2}}{m^{3/2}} \right] \leq e^{-n^{2\nu}}$$

which, together with the inequality on the left hand side derived above, means that for n large enough, the matrix F satisfies condition (5) with probability at least $1 - 2e^{-n^{2\nu}}$. Finally, with $L(F)^2 \leq n\|FF^T\|_2$, the fact that $\|F\|_2$ is 1-Lipschitz (with respect to Euclidian norm as a function of the (Gaussian) entries of F) combined with the bound on $\mathbf{E}[\|F\|_2]$ detailed in Davidson and Szarek (2001, Prop. 2.14) shows that

$$\mathbf{Prob} \left[\|F\|_2 \leq c + \sqrt{2} n^\nu \right] \leq e^{-n^{2\nu}},$$

for some absolute constant $c > 0$. This means that $L(F)/m \to 0$ when n goes to infinity and the second term of the right-hand side of (5) is then negligible compared to the first. ∎

This last result shows that the sufficient condition in (5) is weak enough on Gaussian matrices to hold w.h.p. near optimal values of the cardinality where the number of samples m is almost proportional to the number of nonzero coefficients in the signal.

REFERENCES

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1970.

F. Affentranger and R. Schneider. Random projections of regular simplices. *Discrete and Computational Geometry*, 7 (1):219–226, 1992.

S. Arora, D. Karger, and M. Karpinski. Polynomial time approximation schemes for dense instances of graph problems. In *Proc. of 28th STOC*, pages 193–210, 1995.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

A. Ben-Tal, L. El Ghaoui, and A.S. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Preprint Submitted to the Annals of Statistics*, 2007.

A. Billionnet and F. Roupin. A deterministic approximation algorithm for the densest k-subgraph problem. *International Journal of Operational Research*, 2006.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12): 4203–4215, 2005.

E.J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37:2145–2177, 2009.

E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the AMS*, 22 (1):211–231, 2009.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

Alexandre d'Aspremont and Laurent El Ghaoui. Testing the nullspace property using semidefinite programming. *Mathematical Programming*, 127:123–144, 2011.

K.R. Davidson and S.J. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1:317–366, 2001.

D. L. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. *Stanford dept. of statistics working paper*, 2004.

D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. of the National Academy of Sciences*, 102(27):9446–9451, 2005.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

D.L. Donoho and J. Tanner. Counting the Faces of Randomly-Projected Hypercubes and Orthants, with Applications. *Arxiv preprint arXiv:0807.3590*, 2008.

U. Feige and M. Langberg. Approximation algorithms for maximization problems arising in graph partitioning. *Journal of Algorithms*, 41(2):174–211, 2001.

U. Feige and M. Seltser. On the densest k-subgraph problem. Technical report, Department of Applied Mathematics and Computer Science, The Weizmann Institute, 1997.

U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

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

Q. Han, Y. Ye, and J. Zhang. An improved rounding method and semidefinite programming relaxation for graph partition. *Mathematical Programming*, 92(3):509–535, 2002a.

Q. Han, Y. Ye, and J. Zhang. An improved rounding method and semidefinite programming relaxation for graph partition. *Mathematical Programming*, 92(3):509–535, 2002b. ISSN 0025-5610.

C. Helmberg, F. Rendl, and R. Weismantel. A semidefinite programming approach to the quadratic knapsack problem. *Journal of Combinatorial Optimization*, 4(2):197–215, 2000.

A. Juditsky and A.S. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Mathematical Programming Series B*, 127(57-88), 2011.

G. Kortsarz and D. Peleg. On choosing a dense subgraph. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 692–701, 1993.

M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2005.

E.Y.H. Lin. A Biblographical Survey on Some Wellknown Non-Standard Knapsack Problems. *Information Systems and Operational Research*, 36(4):274–317, 1998.

P. Massart. Concentration inequalities and model selection. *Ecole d'Eté de Probabilités de Saint-Flour XXXIII*, 2007.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

N. Meinshausen, G. Rocha, and B. Yu. A tale of three cousins: Lasso, l2boosting, and danzig. *Annals of Statistics*, 35 (6):2373–2384, 2007.

A.S. Nemirovski. The matrix cube problem: Approximations and applications. *INFORMS*, 2001.

A.S. Nemirovski. *Computation of matrix norms with applications to Robust Optimization*. PhD thesis, Technion, 2005.

Y. Nesterov. *Global quadratic optimization via conic relaxation*. Number 9860. CORE Discussion Paper, 1998a.

Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization methods and software*, 9(1): 141–160, 1998b.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110 (2):245–259, 2007.

D. Pisinger. The quadratic knapsack problem—a survey. *Discrete Applied Mathematics*, 155(5):623–648, 2007.

A. Srivastav and K. Wolf. Finding dense subgraphs with semidefinite programming. *Lecture Notes in Computer Science*, 1444:181–192, 1998.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal statistical society, series B*, 58(1):267–288, 1996.

AM Vershik and PV Sporyshev. Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem. *Selecta Math. Soviet*, 11(2):181–201, 1992.

Y. Zhang. A simple proof for recoverability of ℓ_1-minimization: Go over or under. *Rice University CAAM Technical Report TR05-09*, 2005.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

CMAP, Ecole Polytechnique, UMR CNRS 7641.
E-mail address: `alexandre.daspremont@m4x.org`

Statistics, U.C. Berkeley. Berkeley, CA 94720.
E-mail address: `nkaroui@stat.berkeley.edu`