Sketching Sparse Matrices

Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Rob Nowak
University of Wisconsin - Madison

March 26, 2013

Abstract

This paper considers the problem of recovering an unknown sparse $p \times p$ matrix X from an $m \times m$ matrix $Y = AXB^T$, where A and B are known $m \times p$ matrices with $m \ll p$.

The main result shows that there exist constructions of the "sketching" matrices A and B so that even if X has $\mathcal{O}(p)$ non-zeros, it can be recovered exactly and efficiently using a convex program as long as these non-zeros are not concentrated in any single row/column of X. Furthermore, it suffices for the size of Y (the sketch dimension) to scale as $m = \mathcal{O}\left(\sqrt{\#\text{ nonzeros in } X} \times \log p\right)$. The results also show that the recovery is robust and stable in the sense that if X is equal to a sparse matrix plus a perturbation, then the convex program we propose produces an approximation with accuracy proportional to the size of the perturbation. Unlike traditional results on sparse recovery, where the sensing matrix produces independent measurements, our sensing operator is highly constrained (it assumes a tensor product structure). Therefore, proving recovery guarantees require non-standard techniques. Indeed our approach relies on a novel result concerning tensor products of bipartite graphs, which may be of independent interest.

This problem is motivated by the following application, among others. Consider a $p \times n$ data matrix D, consisting of n observations of p variables. Assume that the correlation matrix $X := DD^T$ is (approximately) sparse in the sense that each of the p variables is significantly correlated with only a few others. Our results show that these significant correlations can be detected even if we have access to only a sketch of the data $S = AD$ with $A \in R^{m \times p}$.

Keywords. sketching, tensor products, distributed sparsity, ℓ_1 minimization, compressed sensing, covariance sketching, graph sketching, multi-dimensional signal processing.

1 Introduction

An important feature of many modern data analysis problems is the presence of a large number of variables relative to the amount of available resources. Such high dimensionality

occurs in a range of applications in bioinformatics, climate studies, and economics. Accordingly, a fruitful and active research agenda over the last few years has been the development of methods for sampling, estimation, and learning that take into account *structure* in the underlying model and thereby making these problems tractable. A notion of structure that has seen many applications is that of *sparsity*, and methods for sampling and estimating sparse signals have been the subject of intense research in the past few years [11, 10, 20]

In this paper we will study a more nuanced notion of structure which we call *distributed sparsity*. For what follows, it will be convenient to think of the unknown high-dimensional signal of interest as being represented as a matrix X. Roughly, the signal is said to be distributed sparse if every row and every column of X has only a few non-zeros. We will see that it is possible to design efficient and effective acquisition and estimation mechanisms for such signals. Let us begin by considering a few example scenarios where one might encounter distributed sparsity.

- **Covariance Matrices:** Covariance matrices associated to some natural phenomena have the property that each covariate is correlated with only a few other covariates. For instance, it is observed that protein signaling networks are such that there are only a few significant correlations [45] and hence the discovery of the such networks from experimental data naturally leads to the estimation of a covariance matrix (where the covariates are proteins) which is (approximately) distributed sparse. Similarly, the covariance structure corresponding to longitudinal data is distributed sparse [17]. See Section 1.3.1.

- **Multi-dimensional signals:** Multi-dimensional signals such as the natural images that arise in medical imaging [11] are known to be sparse in the gradient domain. When the features in the images are not axis-aligned, not only is the matrix representation of the image gradient sparse, it is also distributed sparse. For a little more on this, see Section 1.3.3

- **Random Sparse Signals and Random Graphs:** Signals where the sparsity pattern is *random* (i.e., each entry is nonzero independently and with a probability q) are also distributed sparse with high probability. The "distributedness" of the sparsity pattern can be measured using the "degree of sparsity" d which is defined to be the maximum number of non-zeros in any row or column. For random sparsity patterns, we have the following:

  Proposition 1. *Consider a random matrix $X \in \mathbb{R}^{p\times p}$ whose entries are independent copies of the Bernoulli(γ) [1] distribution where $p\gamma = \Delta = \Theta(1)$. Then for any $\epsilon > 0$, X has at most d 1's in each row/column with probability at least $1-\epsilon$, where*

$$d = \Delta\left(1 + \frac{2\log(2p/\epsilon)}{\Delta}\right).$$

[1] Recall that if $\chi \sim$ Bernoulli(γ), then $P(\chi = 1) = \gamma$ and $P(\chi = 0) = 1-\gamma$.

(The proof is straightforward, and available in Appendix B.)

In a similar vein, combinatorial graphs have small *degree* in a variety of applications, and their corresponding matrix representation will then be distributed sparse. For instance, Erdos-Renyi random graphs $\mathcal{G}(p, q)$ with $pq = \mathcal{O}(\log p)$ have small degree [9].

1.1 Problem Setup and Main Results

Our goal is to invert an underdetermined linear system of the form

$$Y = AXB^T, \tag{1}$$

where $A = [a_{ij}] \in \mathbb{R}^{m\times p}, B = [b_{ij}] \in \mathbb{R}^{m\times p}$, with $m \ll p$ and $X \in \mathbb{R}^{p\times p}$. Since the matrix $X \in \mathbb{R}^{p\times p}$ is linearly transformed to obtain the smaller dimensional matrix $Y \in \mathbb{R}^{m\times m}$, we will refer to Y as the *sketch* (borrowing terminology from the computer science literature [40]) of X and we will refer to the quantity m as the *sketching dimension*. Since the value of m signifies the amount of compression achieved, it is desirable to have as small a value of m as possible.

Rewriting the above using tensor product notation, with $y = \text{vec}(Y)$ and $x = \text{vec}(X)$, we equivalently have

$$y = (B \otimes A)x, \tag{2}$$

where $\text{vec}(X)$ simply *vectorizes* the matrix X, i.e., produces a long column vector by stacking the columns of the matrix and $B \otimes A$ is the tensor (or Kronecker) product of B and A, given by

$$\begin{bmatrix} b_{11}A & b_{12}A & \cdots & b_{1p}A \\ b_{21}A & b_{22}A & \cdots & b_{2p}A \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1}A & b_{m2}A & \cdots & b_{mp}A \end{bmatrix}. \tag{3}$$

While it is not possible to invert such underdetermined systems of equations in general, the rapidly growing literature on what has come to be known as *compressed sensing* suggests that this can be done under certain assumptions. In particular, taking cues from this literature, one might think that this is possible if x (or equivalently X) has only a few non-zeros.

Let us first consider the case when there are only $k = \Theta(1)$ non-zeros in X, i.e., it is very sparse. Then, it is possible to prove that the optimization program (P_1) recovers X from AXB^T using standard "RIP-based" techniques [11]. We refer the interested reader to the papers by Jokar et al [33] and Duarte et al [21] for more details, but in essence the authors show that if $\delta_r(A)$ and $\delta_r(B)$ are the restricted isometry constants (of order r) [11] for A and B respectively, then the following is true about $B \otimes A$

$$\max\{\delta_r(A), \delta_r(B)\} \leq \delta_r(A \otimes B) = \delta_r(B \otimes A) \leq (1 + \delta_r(A))(1 + \delta_r(B)) - 1.$$

In many interesting problems that arise naturally, as we will see in subsequent sections, a more realistic assumption to make is that X has $\mathcal{O}(p)$ non-zeros and it is this setting we

consider for this paper. Unfortunately, the proof techniques outlined above cannot succeed in such a demanding scenario. As hinted earlier, it will turn out however that one cannot handle arbitrary sparsity patterns and that the non-zero pattern of X needs to be *distributed*, i.e., each row/column of X cannot have more than a few, say d, non-zeros. We will call such matrices $d-$distributed sparse (see Definition 3). We explore this notion of structure in more detail in Section 3.1.

An obvious, albeit highly impractical, approach to recover a (distributed) sparse X from measurements of the form $Y = AXB^T$ is the following: search over all matrices $\tilde{X} \in \mathbb{R}^{p\times p}$ such that $A\tilde{X}B^T$ agrees with $Y = AXB^T$ and find the sparsest one. One might hope that under reasonable assumptions, such a procedure would return X as the solution. However, there is no guarantee that this approach might work and worse still, such a search procedure is known to be computationally infeasible.

We instead consider solving the optimization program (P_1) which is a natural (convex) relaxation of the above approach.

$$\begin{aligned} \underset{\tilde{X}}{\text{minimize}} \quad & \left\|\tilde{X}\right\|_1 \\ \text{subject to} \quad & A\tilde{X}B^T = Y. \end{aligned} \tag{P_1}$$

Here, by $\|\tilde{X}\|_1$ we mean $\sum_{i,j}\left|\tilde{X}_{i,j}\right|$, i.e., the ℓ_1 norm of $\mathrm{vec}(\tilde{X})$.

The main part of the paper is devoted to showing with high probability (P_1) has a unique solution that equals X. In particular, we prove the following result.

Theorem 1. *Suppose that X is $d-$distributed sparse. Also, suppose that $A, B \in \{0,1\}^{m\times p}$ are drawn independently and uniformly from the $\delta-$random bipartite ensemble[2]. Then as long as*

$$m = \mathcal{O}(\sqrt{dp}\log p) \quad \textit{and} \quad \delta = \mathcal{O}(\log p),$$

there exists a $c > 0$ such that the optimal solution X^* of (P_1) equals X with probability exceeding $1 - p^{-c}$. Furthermore, this holds even if B equals A.

In Section 4, we will prove Theorem 1 for the case when $B = A$. It is quite straightforward to modify this proof to the case where A and B are independently drawn, since there is much more independence that can be leveraged.

Let us pause here and consider some implications of this theorem.

1. (P_1) does not impose any structural restrictions on X^*. In other words, even though X is assumed to be distributed sparse, this (highly non-convex) constraint need not be factored in to the optimization problem. This ensures that (P_1) is a Linear Program (see e.g., [6]) and can thus be solved efficiently.

[2]Roughly speaking, the $\delta-$random bipartite ensemble consists of the set of all 0-1 matrices that have almost exactly δ ones per column. We refer the reader to Definition 4 for the precise definition and Section 3.2 for more details.

2. Recall that what we observe can be thought of as the $\mathbb{R}^{m^2}$ vector $(B \otimes A)x$. Since X is $d-$distributed sparse, x has $\mathcal{O}(dp)$ non-zeros. Now, even if an oracle were to reveal the exact locations of these non-zeros, we would require at least $\mathcal{O}(dp)$ measurements to be able to perform the necessary inversion to recover x. In other words, it is absolutely necessary for m^2 to be at least $\mathcal{O}(dp)$. Comparing this to Theorem 1 shows that the simple algorithm we propose is *near optimal* in the sense that it is only a logarithm away from this trivial lower bound. This logarithmic factor also makes an appearance in the measurement bounds in the compressed sensing literature [10].

3. Finally, as mentioned earlier, inversion of under-determined linear systems where the linear operator assumes a tensor product structure has been studied earlier [32, 22]. However, these methods are relevant only in the regime where the sparsity of the signal to be recovered is much smaller than the dimension p. The proof techniques they employ will unfortunately not allow one to handle the more demanding situation the sparsity scales linearly in p and if one attempted an extension of their techniques naively to this situation, one would see that the sketch size m needs to scale like $\mathcal{O}(dp \log^2 p)$ in order to recover a $d-$ distributed sparse matrix X. This is of course uninteresting since it would imply that the size of the sketch is bigger than the size of X.

It is possible that Y was not exactly observed, but rather is only available to us as a corrupted version $\hat{Y}$. For instance, $\hat{Y}$ could be Y corrupted by independent zero mean, additive Gaussian noise or in case of the covariance sketching problem discussed in Section 1.3.1, $\hat{Y}$ could be an empirical estimate of the covariance matrix $Y = AXA^T$. In both these cases, a natural relaxation to (P_1) would be the following optimization program (P_2) (with B set to A in the latter case).

$$\underset{\tilde{X}}{\text{minimize}} \qquad \|A\tilde{X}B^T - \hat{Y}\|_2^2 + \lambda \left\|\tilde{X}\right\|_1 \tag{P_2}$$

Notice that if X was a sparse covariance matrix and if $A = B = I_{p\times p}$, then (P_2) reduces to the soft thresholding estimator of sparse covariance matrices studied in [43].

While our experimental results show that this optimization program (P_2) performs well, we leave its exploration and analysis to future work. We will instead state the following "approximation" result that shows that the solution of (P_1) is close to the optimal $d-$ distributed sparse approximation for any matrix X. The proof is similar to the proof of Theorem 3 in [5] and is provided in Appendix C. Given $p \in \mathbb{N}$, let $[p]$ denote the set $\{1, 2, \ldots, p\}$ and let $\mathfrak{W}_{d,p}$ denote the following collection of subsets of $[p] \times [p]$:

$$\mathfrak{W}_{d,p} := \{\Omega \subset [p] \times [p] : |\Omega \cap \{\{i\} \times [p]\}| \leq d, |\Omega \cap \{[p] \times \{i\}\}| \leq d, \text{ for all } i \in [p]\}.$$

Notice that if a matrix $X \in \mathbb{R}^{p\times p}$ is such that there exists an $\Omega \in \mathfrak{W}_{d,p}$ with the property that $X_{ij} \neq 0$ only if $(i, j) \in \Omega$, then the matrix is $d-$distributed sparse.

Given $\Omega \in \mathfrak{W}_{d,p}$ and a matrix $X \in \mathbb{R}^{p\times p}$, we write X_Ω to denote the projection of X onto the set of all matrices supported on Ω. That is,

$$[X_\Omega]_{i,j} = \begin{cases} X_{i,j} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } (i,j) \in [p] \times [p].$$

Theorem 2. *Suppose that X is an arbitrary $p \times p$ matrix and that the hypotheses of Theorem 1 hold. Let X^* denote the solution to the optimization program (P_1). Then, there exist constants $c > 0$ and $\epsilon \in (0, 1/4)$ such that the following holds with probability exceeding $1 - p^{-c}$.*

$$\|X^* - X\|_1 \leq \frac{2-4\epsilon}{1-4\epsilon} \left(\min_{\Omega \in \mathfrak{W}_{d,p}} \|X - X_\Omega\|_1 \right). \tag{4}$$

The above theorem tells us that even if X is not structured in any way, the solution of the optimization program (P_1) approximates X as well as the best possible $d-$distributed sparse approximation of X (up to a constant factor). This has interesting implications, for instance, to situations where a $d-$distributed sparse X is corrupted by a "noise" matrix N as shown in the following corollary.

Corollary 1. *Suppose $X \in \mathbb{R}^{p\times p}$ is $d-$distributed sparse and suppose that $\hat{X} = X + N$. Then, the solution X^* to the optimization program*

$$\min_{\tilde{X}} \left\| \tilde{X} \right\|_1 \text{ subject to } A\tilde{X}B^T = A\hat{X}B^T$$

satisfies

$$\|X^* - X\|_1 \leq \frac{5-12\epsilon}{1-4\epsilon} \|N\|_1 \tag{5}$$

Proof. Let Ω be the support of X. To prove the result, we will consider the following chain of inequalities.

$$\begin{aligned} \|X^* - X\|_1 &\leq \left\| X^* - \hat{X} \right\|_1 + \left\| \hat{X} - X \right\|_1 \\ &\overset{(a)}{\leq} \frac{2-4\epsilon}{1-4\epsilon} \left\| \hat{X} - \hat{X}_\Omega \right\|_1 + \left\| \hat{X} - X \right\|_1 \\ &\leq \frac{2-4\epsilon}{1-4\epsilon} \left\| \hat{X} - X \right\|_1 + \frac{2-4\epsilon}{1-4\epsilon} \left\| \hat{X}_\Omega - X \right\|_1 + \left\| \hat{X} - X \right\|_1 \\ &\overset{(b)}{\leq} \frac{5-12\epsilon}{1-4\epsilon} \left\| \hat{X} - X \right\|_1 \\ &\overset{(c)}{=} \frac{5-12\epsilon}{1-4\epsilon} \|N\|_1 . \end{aligned}$$

Here (a) follows from Theorem 2 since $\Omega \in \mathfrak{W}_{d,p}$ and (b) follows from the fact that $\left\| \hat{X}_\Omega - X \right\| \leq \left\| \hat{X} - X \right\|$ since X_{Ω^c} is $\mathbf{0}_{p\times p}$. Finally, in (c) we merely plug in the definition of $\hat{X}$. □

1.2 The Rectangular Case and Higher Dimensional Signals

While Theorem 1, as stated, applies only to the case of square matrices X, we can extend our result in a straightforward manner to the rectangular case. Consider a matrix $X \in \mathbb{R}^{p_1 \times p_2}$ where (without loss of generality) $p_1 < p_2$. We assume that the row degree is d_r (i.e. no row of X has more than d_r non-zeros) and that the column degree is d_c. Consider sketching matrices $A \in \mathbb{R}^{m \times p_1}$ and $B \in \mathbb{R}^{m \times p_2}$ and the sketching operation:

$$Y = AXB^T.$$

Then we have the following corollary:

Corollary 2. *Suppose that X is distributed sparse with row degree d_r and colum degree d_c. Also, suppose that $A \in \{0,1\}^{m \times p_1}$, $B \in \{0,1\}^{m \times p_2}$ are drawn independently and uniformly from the $\delta-$random bipartite ensemble. Let us define $p = \max(p_1, p_2)$ and $d = \max(d_r, d_c)$.*

Then if

$$m = \mathcal{O}(\sqrt{dp} \log p) \quad \text{and} \quad \delta = \mathcal{O}(\log p),$$

there exists a $c > 0$ such that the optimal solution X^* of (P_1) equals X with probability exceeding $1 - p^{-c}$.

Proof. Let us define the matrix $\tilde{X} \in \mathbb{R}^{p \times p}$ as

$$\tilde{X} = \begin{bmatrix} X \\ 0 \end{bmatrix},$$

i.e. it is made square by padding additional zero rows. Note that $\tilde{X}$ has degree $d = \max(d_r, d_c)$. Moreover note that the matrix $A \in \mathbb{R}^{m \times p_1}$ can be augmented to $\tilde{A} \in \mathbb{R}^{m \times p}$ via:

$$\tilde{A} = \begin{bmatrix} A & \bar{A} \end{bmatrix}$$

where $\bar{A} \in \mathbb{R}^{m \times (p - p_1)}$ is also drawn from the δ-random bipartite ensemble. Then one has the relation:

$$Y = \tilde{A}\tilde{X}B^T.$$

Thus, the rectangular problem can be reduced to the standard square case considered in Theorem 1, and the result follows.

□

The above result shows that a finer analysis is required for the rectangular case. For instance, if one were to consider a scenario where $p_1 = 1$, then from the compressed sensing literature, we know that the result of Corollary 2 is weak. We believe that determining the right scaling of the sketch dimension(s) in the case when X is rectangular is an interesting avenue for future work.

Finally, we must also state that while the results in this paper only deal with two-dimensional signals, similar techniques can be used to deal with higher dimensional tensors that are distributed sparse. We leave a detailed exploration of this question to future work.

1.3 Applications

It is instructive at this stage to consider a few examples of the framework we set up in this paper. These applications demonstrate that the modeling assumptions we make viz., tensor product sensing and distributed sparsity are important and arise naturally in a wide variety of contexts.

1.3.1 Covariance Estimation from Compressed realizations or Covariance Sketching

One particular application that will be of interest to us is the estimation of covariance matrices from sketches of the sample vectors. We call this *covariance sketching*.

Consider a scenario in which the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of a high-dimensional zero-mean random vector $\xi = (\xi_1, \ldots, \xi_p)^T$ is to be estimated. In many applications of interest, one determines Σ by conducting correlation tests for each pair of covariates ξ_i, ξ_j and computing an estimate of $\mathbf{E}[\xi_i \xi_j]$ for $i, j = 1, \ldots, p$. This requires one to perform correlation tests for $\mathcal{O}(p^2)$ pairs of covariates, a daunting task in the high-dimensional setting. Perhaps most importantly, in many cases of interest, the underlying covariance matrix may have structure, which such an approach may fail to exploit. For instance if Σ is very sparse, it would be vastly more efficient to perform correlation tests corresponding to only the non-zero entries. The chief difficulty of course is that the sparsity pattern is rarely known in advance, and finding this is often the objective of the experiment.

In other settings of interest, one may obtain statistical samples by observing n independent *sample paths* of the statistical process. When ξ is high-dimensional, it may be infeasible or undesirable to sample and store the entire sample paths $\xi^{(1)}, \ldots, \xi^{(n)} \in \mathbb{R}^p$, and it may be desirable to reduce the dimensionality of the acquired samples.

Thus in the high-dimensional setting we propose an alternative acquisition mechanism: pool covariates together to form a collection of new variables $Z_1, \ldots, Z_m$, where $m < p$. For example one may construct:

$$Z_1 = \xi_1 + \xi_2 + \xi_6, \quad Z_2 = \xi_1 + \xi_4 + \xi_8 + \xi_{12}, \quad \ldots$$

and so on; more generally we have measurements of the form $Z = A\xi$ where $A \in \mathbb{R}^{m\times p}$ and typically $m \ll p$. We call the thus newly constructed covariates $Z = (Z_1, \ldots, Z_m)$ a sketch of the random vector ξ.

More formally, the covariance sketching problem can be stated as follows. Let $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)} \in \mathbb{R}^p$ be n independent and identically distributed $p-$variate random vectors and let $\Sigma \in \mathbb{R}^{p\times p}$ be their unknown covariance matrix. Now, suppose that one has access to the $m-$dimensional *sketch vectors* $Z^{(i)}$ such that

$$Z^{(i)} = A\xi^{(i)}, \quad i = 1, 2, \ldots, n,$$

where $A \in \mathbb{R}^{m\times p}, m < p$ is what we call a *sketching matrix*. The goal then is to recover Σ using only $\{Z^{(i)}\}_{i=1}^n$. The sketching matrices we will focus on later will have randomly-generated binary values, so each element of $Z^{(i)}$ will turn out to be a sum (or "pool") of a random subset of the covariates.

Notice that the sample covariance matrix computed using the vectors $\left\{Z^{(i)}\right\}_{i=1}^{n}$ satisfies the following.

$$\begin{aligned}\hat{\Sigma}_Z^{(n)} &:= \frac{1}{n}\sum_{i=1}^{n} Z^{(i)}(Z^{(i)})^T \\ &= A\left(\frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}(\xi^{(i)})^T\right)A^T \\ &= A\hat{\Sigma}^{(n)}A^T,\end{aligned}$$

where $\hat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}(\xi^{(i)})^T$ is the (maximum likelihood) estimate of Σ from the samples $\xi^{(1)}\ldots,\xi^{(n)}$.

To gain a better understanding of the covariance sketching problem, it is natural to first consider the stylized version of the problem suggested by the above calculation. That is, whether it is possible to efficiently recover a matrix $\Sigma \in \mathbb{R}^{p\times p}$ given the ideal covariance matrix of the sketches $\Sigma_Z = A\Sigma A^T \in \mathbb{R}^{m\times m}$. The analysis in the current paper focuses on exactly this problem and thus helps in exposing the most unique and challenging aspects of covariance sketching.

The theory developed in this paper tells us that at the very least, one needs to restrict the underlying random vector ξ to have the property that each ξ_i depends on only a few (say, d) of the other ξ_j's. Notice that this would of course imply that the true covariance matrix Σ will be $d-$distributed sparse. Applying Theorem 1, especially the version in which the matrices A and B are identical, to this stylized situation reveals the following result. If A is chosen from a particular random ensemble and if one gets to observe the covariance matrix $A\Sigma A^T$ of the sketch random vector $Z = A\xi$, then using a very efficient convex optimization program, one can recover Σ exactly.

Now, suppose that ξ and A are as above and that we get to observe n samples $Z^{(i)} = A\xi^{(i)}, i = 1,2,\ldots,n$ of the sketch $Z = A\xi$. Notice that we can can consider $\hat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}(\xi^{(i)})^T$ to be a "noise corrupted version" of Σ since we can write

$$\hat{\Sigma}^{(n)} = \Sigma + (\hat{\Sigma}^{(n)} - \Sigma),$$

where, under reasonable assumptions on the underlying distribution, $\left\|\hat{\Sigma}^{(n)} - \Sigma\right\|_1 \to 0$ almost surely as $n \to \infty$ by the strong law of large of numbers. Therefore, an application of Theorem 2 tells us that solving (P_1) with the observation matrix $\hat{\Sigma}_Z^{(n)}$ gives us an asymptotically consistent procedure to estimate the covariance matrix Σ from sketched realizations.

We anticipate that our results will be interesting in many areas such as quantitative biology where it may be possible to naturally pool together covariates and measure interactions at this pool level. Our work shows that covariance structures that occur naturally are amenable to covariance sketching, so that drastic savings are possible when correlation tests are performed at the pool level, rather than using individual covariates.

The framework we develop in this paper can also be used to accomplish *cross covariance sketching*. That is, suppose that ξ and ζ are two zero mean $p-$variate random vectors and suppose that $\Sigma_{\xi\zeta} \in \mathbb{R}^{p\times p}$ is an unknown matrix such that $[\Sigma_{\xi\zeta}]_{ij} = \mathbb{E}[\xi_i\zeta_j]$. Let $\{\xi^{(i)}\}_{i=1}^n$ and $\{\zeta^{(i)}\}_{i=1}^n$ be $2n$ independent and identically distributed random realizations of ξ and ζ respectively. The goal then, is to estimate $\Sigma_{\xi\zeta}$ from the m dimensional *sketch vectors* $Z^{(i)}$ and $W^{(i)}$ such that

$$Z^{(i)} = A\xi^{(i)}, W^{(i)} = B\zeta^{(i)} \quad i = 1, 2, \ldots, n,$$

where $A, B \in \mathbb{R}^{m\times p}, m < p$.

As above, in the idealized case, Theorem 1 shows that the cross-covariance matrix $\Sigma_{\xi\zeta}$ of ξ and ζ can be exactly recovered from the cross-covariance matrix $\Sigma_{ZW} = A\Sigma_{\xi\zeta}B^T$ of the sketched random vectors W and Z as long as $\Sigma_{\xi\zeta}$ is distributed sparse. In the case we have n samples each of the sketched random vectors, an application of Theorem 2 to this problem tells us that (P_1) is an efficient and asymptotically consistent procedure to estimate a distributed sparse $\Sigma_{\xi\zeta}$ from compressed realization of ξ and ζ.

We note that the idea of pooling information in statistics, especially in the context of meta analysis is a classical one [29]. For instance the classical Cohen's d estimate uses the idea of pooling samples obtained from different distributions to obtain accurate estimates of a common variance. While at a high level the idea of pooling is related, we note that our notion is qualitatively different in that we propose pooling covariates themselves into sketches and obtain samples in this reduced dimensional space.

1.3.2 Graph Sketching

Large graphs play an important role in many prominent problems of current interest; two such examples are graphs associated to communication networks (such as the internet) and social networks. Due to their large sizes it is difficult to store, communicate, and analyze these graphs, and it is desirable to compress these graphs so that these tasks are easier. The problem of compressing or sketching graphs has recently gained attention in the literature [1, 24].

In this section we propose a new and natural notion of compression of a given graph $G = (V, E)$. The resulting "compressed" graph is a weighted graph $\hat{G} = (\hat{V}, \hat{E})$, where $\hat{V}$ has a much smaller cardinality than V. Typically, $\hat{G}$ will be a complete graph, but the edge weights will encode interesting and valuable information about the original graph.

Partition the vertex set $V = V_1 \cup V_2 \cup \ldots \cup V_m$; in the compressed graph $\hat{G}$, each partition V_i is represented by a node. (We note that this need not necessarily be a disjoint partition, and we allow for the possibility for $V_i \cap V_j \neq \emptyset$.) For each pair $V_i, V_j \in \hat{V}$, the associated edge weight is the total number of edges crossing from nodes in V_i to the nodes in V_j in the original graph G. Note that if an index $k \in V_i \cap V_j$, the self edge (k, k) must be included when counting the total number of edges between V_i and V_j. (We point out that the edge $(V_k, V_k) \in \hat{E}$ also carries a non-zero weight; and is precisely equal to the number of edges in G that have both endpoints in V_k. See Fig. 1 for an illustrative example.)

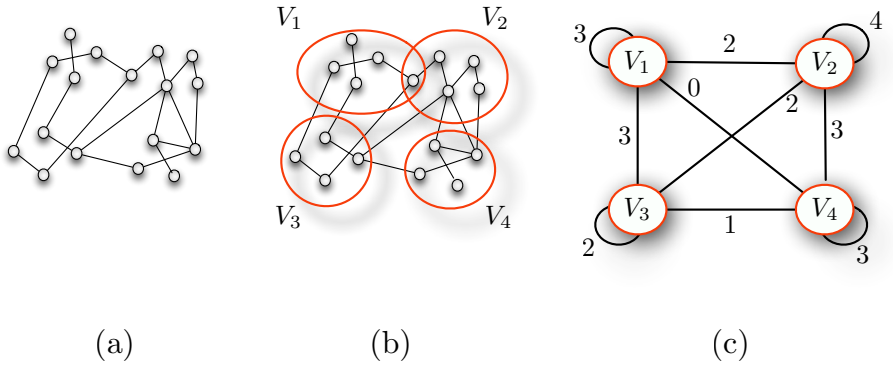
Define A_i to be the (row) indicator vector for the set V_i, i.e.

$$A_{ij} = \begin{cases} 1 & \text{if } j \in V_i \\ 0 & \text{otherwise} \end{cases}$$

If X denotes the adjacency matrix of G, then $Y := AXA^T$ denotes the matrix representation of $\hat{G}$. The sketch Y has two interesting properties:

- The encoding faithfully preserves high-level "cut" information about the original graph. For instance information such as the weight of edges crossing between the partitions V_i and V_j is faithfully encoded. This could be useful for networks where the vertex partitions have a natural interpretation such as geographical regions; questions about the total network capacity between two regions is directly available via this method of encoding a graph. Approximate solutions to related questions such as maximum flow between two regions (partitions) can also be provided by solving the problem on the compressed graph.

- When the graph is bounded degree, the results in this paper show that there exists a suitable random partitioning scheme such that the proposed method of encoding the graph is lossless. Moreover, the original graph G can be unravelled from the smaller sketched graph $\hat{G}$ efficiently using the convex program (P_1).



(a) (b) (c)

Figure 1: An example illustrating graph sketching. (a) A graph G with 17 nodes (b) Partitioning the nodes into four partitions V_1, V_2, V_3, V_4 (c) The sketch of the graph G. The nodes represent the partitions and the edges in the sketch represent the total number of edges of G that cross partitions.

1.3.3 Multidimensional Signal Processing

Multi-dimensional signals arise in a variety of applications, for instance images are naturally represented as two-dimensional signals $f(\cdot, \cdot)$ over some given domain.

Often it is more convenient to view the signal not in the original domain, but rather in a transformed domain. Given some one dimensional family of "mother" functions $\psi_u(t)$

(usually an orthonormal family of functions indexed with respect to the transform variable u), such a family induces the transform for a one dimensional signal $f(t)$ (with domain $\mathcal{T}$) via

$$\hat{f}(u) = \int_{t \in \mathcal{T}} f(t)\psi_u(t).$$

For instance if $\psi_u(t) := \exp(-i2\pi ut)$, this is the Fourier transform, and if $\psi_u(t)$ is chosen to be a wavelet function (where $u = (a, b)$, the translation and scale parameters respectively) this generates the well-known wavelet transform that is now ubiquitous in signal processing.

Using $\psi_u(t)$ to form an orthonormal basis for one-dimensional signals, it is straightforward to extend to a basis for two-dimensional signal by using the functions $\psi_u(t)\psi_v(r)$. Indeed, this defines a two-dimensional transform via

$$\hat{f}(u, v) = \int_{(t,r) \in \mathcal{X} \times \mathcal{X}} f(t, r)\psi_u(t)\psi_v(r).$$

Similar to the one-dimensional case, appropriate choices of ψ yeild standard transforms such as the two-dimensional Fourier transform and the two-dimensional wavelet transform. The advantage of working with an alternate basis as described above is that signals often have particularly simple representations when the basis is appropriately chosen. It is well-known, for instance, that natural images have a sparse representation in the wavelet basis (see Fig. 2). Indeed, many natural images are not only sparse, but they are also distributed sparse, when represented in the wavelet basis. This enables compression by performing "pooling" of wavelet coefficients, as described below.

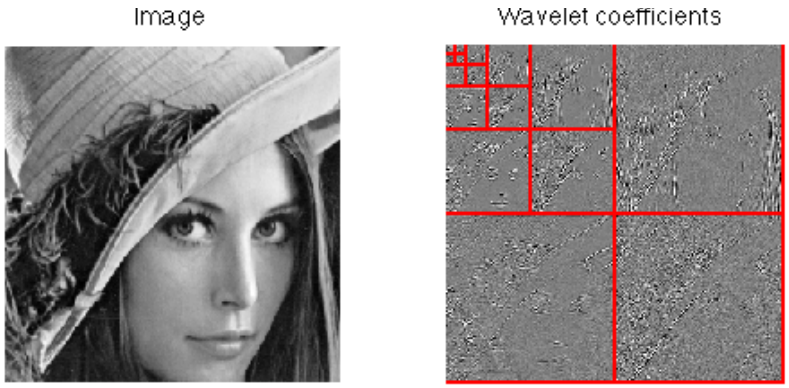


Figure 2: Note that the wavelet representaion of the image is distributed sparse.

In many applications, it is more convenient to work with discrete signals and their transforms (by discretizing the variables (t, r) and the transform domain variables (u, v)). It is natural to represent the discretization of the two-dimensional signal $f(t, r)$ by a matrix $F \in \mathbb{R}^{p \times p}$. The corresponding discretization of $\psi_u(t)$ can be represented as a matrix $\Psi = [\Psi]_{ut}$, and the discretized version of the $\hat{f}(u, v)$, denoted by $\hat{F}$ is given by:

$$\hat{F} = \Psi F \Psi^T.$$

As noted above, in several applications of interest, when the basis Ψ is chosen appropriately, the signal has a succinct representation and the corresponding matrix $\hat{F}$ is sparse. This is true, for instance, when F represents a natural image and $\hat{F}$ is the wavelet transform of F. Due to the sparse representability of the signal in this basis, it is possible to acquire and store the signal in a *compressive* manner. For instance, instead of sensing the signal F using Ψ (which correpsonds to sensing the signal at every value of the transform variable u), one could instead form "pools" of transform variables $S_i = \{u_{i1}, u_{i2} \ldots, u_{ik}\}$ and sense the signal via

$$A\Psi = \begin{bmatrix} \sum_{u \in S_1} \Psi_u \\ \vdots \\ \sum_{u \in S_m} \Psi_u \end{bmatrix},$$

where the matrix A corresponds to the pooling operation. This means of compression corresponds to "mixing" measurements at randomly chosen transform domain values u. (When Ψ is the Fourier transform, this corresponds to randomly chosen frequencies, and when Ψ is the wavelet, this corresponds to mixing randomly chosen translation and scale parameters) . When the signal F is acquired in this manner, we obtain measurements of the form:

$$Y = A\hat{F}A^T,$$

where $\hat{F}$ is suitably sparse. Note that one may choose different random mixtures of measurements for the t and r "spatial" variables, in which case one would obtain measurements of the form:

$$Y = A\hat{F}B^T.$$

The theory developed in this paper shows how one can recover the multi-dimensional signal F from such an undersampled acquisition mechanism. In particular, our results will show that if the pooling of the transform variable is done suitably randomly, then there is an efficient method based on linear programming that can be used to recover the original multi-dimensional signal.

1.4 Related Work and Obstacles to Common Approaches

The problem of recovering sparse signals via ℓ_1 regularization and convex optimization has been studied extensively in the past decade; our work fits broadly into this context. In the signal processing community, the literature on compressed sensing [10, 20] focuses on recovering sparse signals from data. In the statistics community, the LASSO formulation as proposed by Tibshirani, and subsequently analyzed (for variable selection) by Meinshausen and Bühlmann [39], and Wainwright [48] are also closely related. Other examples of structured model selection include estimation of models with a few latent factors (leading to low-rank covariance matrices) [23], models specified by banded or sparse covariance matrices [7, 8], and Markov or graphical models [38, 39, 42]. These ideas have been studied in depth and extended to analyze numerous other model selection problems in statistics and signal processing [4, 14, 12].

Our work is also motivated by the work on sketching in the computer science community; this literature deals with the idea of compressing high-dimensional data vectors via projection to low-dimensions while preserving pertinent geometric properties. The celebrated Johnson-Lindenstrauss Lemma [31] is one such result, and the idea of sketching has been explored in various contexts [2, 34]. The idea of using random bipartite graphs and their related expansion properties, which motivated our approach to the problem, have also been studied in past work [5, 35, 36].

While most of the work on sparse recovery focuses on sensing matrices where each entry is an i.i.d. random variable, there are a few lines of work that explore structured sensing matrices. For instance, there have been studies of matrices with Toeplitz structure [28], or those with random entries with independent rows but with possibly dependent columns [47, 44]. Also related is the work on deterministic dictionaries for compressed sensing [16], although those approaches yield results that are too weak for our setup.

One interesting aspect of our work is that we show that it is possible to use highly constrained sensing matrices (i.e. those with tensor product structure) to recover the signal of interest. Many standard techniques fail in this setting. Restricted isometry based approaches [11] and coherence based approaches [18, 26, 46] fail due to a lack of independence structure in the sensing matrix. Indeed, the restricted isometry constants as well as the coherence constants are known to be weak for tensor product sensing operators [22, 32]. Gaussian width based analysis approaches [13] fail because the kernel of the sensing matrix is not a uniformly random subspace and hence not amenable to a similar application of Gordon's ("escape through the mesh") theorem. We overcome these technical difficulties by working directly with combinatorial properties of the tensor product of a random bipartite graph, and exploiting those to prove the so-called nullspace property [19, 15].

2 Experiments

We demonstrate the validity of our theory with some preliminary experiments in this section. Figure 3 shows a 40×40 distributed sparse matrix on the left side. The matrix on the right is a perfect reconstruction using a sketch dimension of $m = 21$.

Figure 4 is what is known now as the "phase transition diagram". Each coordinate $(i, j) \in \{10, 12, \ldots, 60\} \times \{2, 4, \ldots, 60\}$ in the figure corresponds to an experiment with $p = i$ and $m = j$. The value at the coordinate (i, j) was generated as follows. A random 4-distributed sparse $X \in \mathbb{R}^{i \times i}$ was generated and a random $A \in \mathbb{R}^{j \times i}$ was generated as adjacency matrix of a graph as described in Definition 4. Then the optimization problem (P_1) was solved using the CVX toolbox [30, 25]. The solution X^* was compared to X in the $\|\cdot\|_\infty$ norm (upto numerical precision errors). This was repeated 40 times and the average number of successes was reported in the (i, j)-th spot. In the figure, the white region denotes success during each trial and the black region denotes failure in every single trial and there is a sharp *phase transition* between successes and failures. In fact, the curve that borders this phase transition region roughly looks like the curve $p = \frac{1}{14}m^2$ which is what our theory predicts (upto constants and log factors).

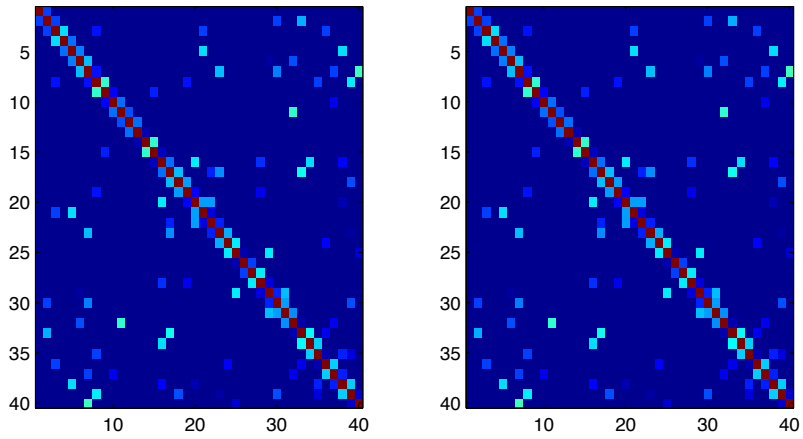


Figure 3: The matrix on the left is a 40×40 sparse matrix and the matrix on the right is a perfect reconstruction with $m = 21$.

We also ran some preliminary tests on trying to reconstruct a covariance matrix from sketches of samples drawn from the original distribution.To factor in the "noise", we replaced the equality constraint in (P_1) with a constraint which restricts the feasible set to be the

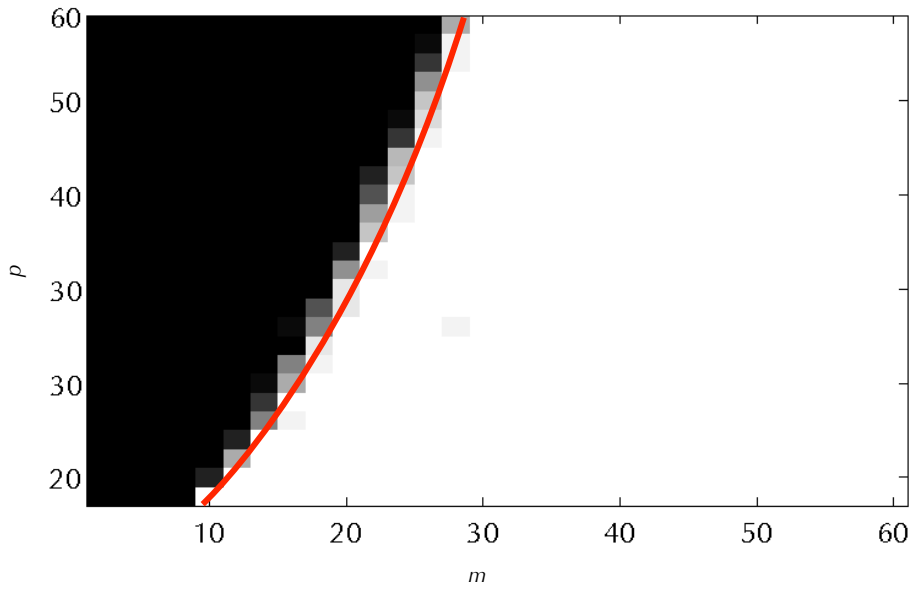


Figure 4: Phase transition plot. The (i, j)-th pixel shows (an approximation) to the probability of success of the optimization problem (P_1) in recovering a distributed sparse $X \in \mathbb{R}^{i \times i}$ with sketch-size j. The (red) solid line shows the boundary of the phase transition regime and is approximately the curve $p = \frac{1}{14}m^2$

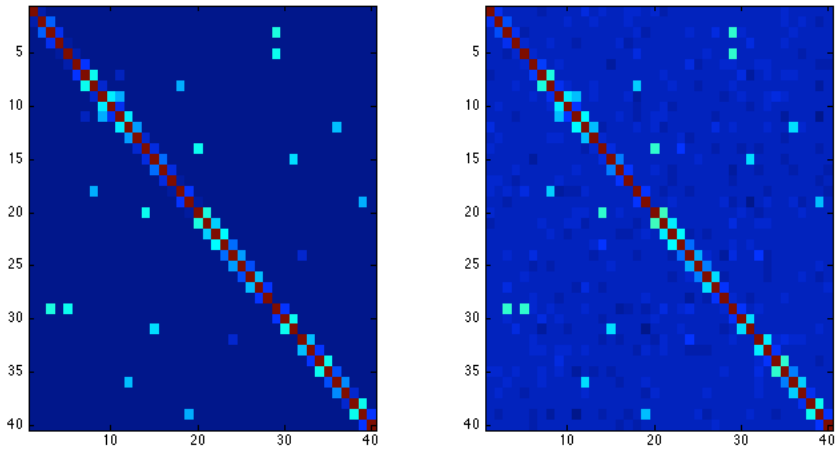


Figure 5: The matrix on the left is a 40×40 distributed sparse matrix. The matrix on the right was reconstructed using $n = 2100$ samples and with sketches of size $m = 21$.

set of all X such that $\left\| AXA^T - \hat{X}_Y^{(n)} \right\|_2 \leq \kappa$ instead. The parameter κ was picked by cross-validation. Figure 5 shows a representative result which is encouraging. The matrix on the left is a 40×40 distributed sparse covariance matrix and the matrix on the right is a reconstruction using $n = 2100$ sketches of size $m = 21$ each.

3 Preliminaries and Notation

We will begin our theoretical discussion by establishing some notation and preliminary concepts that will be used through the rest of the paper.

For any $p \in \mathbb{N}$, we define $[p] := \{1, 2, \ldots, p\}$. A **graph** $G = (V, E)$ is defined in the usual sense as an ordered pair with the *vertex set* V and the *edge set* E which is a set of 2-element subsets of V, i.e., $E \subset \binom{V}{2}$. Henceforth, unless otherwise stated, we will deal with the graph $G = ([p], E)$. We also assume that all the graphs that we consider here include all the self loops, i.e., $\{i, i\} \in E$ for all $i \in [p]$. For any $S \subset [p]$, the set of neighbors $N(S)$ is defined as

$$N(S) = \{j \in [p] : i \in S, \{i, j\} \in E\}.$$

For any vertex $i \in [p]$, the degree $\deg(i)$ is defined as $\deg(i) := |N(i)|$.

Definition 1 (Bounded degree graphs and regular graphs)**.** *A graph $G = ([p], E)$ is said to be a* **bounded degree graph** *with (maximum) degree d if for all $i \in [p]$,*

$$deg(i) \leq d$$

*The graph is said to be $d-$***regular** *if $deg(i) = d$ for all $i \in [p]$.*

We will be interested in another closely related combinatorial object. Given $p, m \in \mathbb{N}$, a **bipartite graph** $G = ([p], [m], E)$ is a graph with the *left set* $[p]$ and *right set* $[m]$ such

that the edge set E only has pairs $\{i, j\}$ where i is the left set and j is in the right set. A bipartite graph $G = ([p], [m], E)$ is said to be $\delta-$**left regular** if for all i in the left set $[p]$, $\deg(i) = \delta$. Given two sets $A \subset [p], B \subset [m]$, we define the set

$$E(A : B) := \{(i, j) \in E : i \in A, j \in B\},$$

which we will find use for in our analysis. This set is sometimes known as the *cut set*. Finally for a set $A \subset [p]$ (resp. $B \subset [m]$), we define $N(A) := \{j \in [m] : i \in A, \{i, j\} \in E\}$ (resp. $N_R(B) := \{i \in [p] : j \in B, \{i, j\} \in E\}$). This distinction between N and N_R is made only to reinforce the meaning of the quantities which is otherwise clear in context.

Definition 2. *(Tensor graphs) Given two bipartite graphs $G_1 = ([p], [m], E_1)$ and $G_2 = ([p], [m], E_2)$, we define their* **tensor graph** *$G_1 \otimes G_2$ to be the bipartite graph $([p] \times [p], [m] \times [m], E_1 \otimes E_2)$ where $E_1 \otimes E_2$ is such that $\{(i, i'), (j, j')\} \in E_1 \otimes E_2$ if and only if $\{i, j\} \in E_1$ and $\{i', j'\} \in E_2$.*

Notice that if the adjacency matrices of G_1, G_2 are given respectively by $A^T, B^T \in \mathbb{R}^{p \times m}$, then the adjacency matrix of $G_1 \otimes G_2$ is $(A \otimes B)^T \in \mathbb{R}^{p^2 \times m^2}$.

As mentioned earlier, we will be particularly interested in the situation where $B = A$. In this case, write the tensor product of a graph $G = ([p], [m], E)$ with itself as $G^{\otimes} = ([p] \times [p], [m] \times [m], E^{\otimes})$. Here $E^{\otimes}$ is such that $\{(i, i'), (j, j')\} \in E^{\otimes}$ if and only if $\{i, j\}$ and $\{i', j'\}$ are in E.

Throughout this paper, we write $\|\cdot\|$ to denote norms of vectors. For instance, $\|x\|_1$ and $\|x\|_2$ respectively stand for the ℓ_1 and ℓ_2 norm of x. Furthermore, for a matrix X, we will often write $\|X\|$ to denote $\|\mathrm{vec}(X)\|$ to avoid clutter. Therefore, the Frobenius norm of a matrix X will appear in this paper as $\|X\|_2$.

3.1 Distributed Sparsity

As promised, we will now argue that distributed sparsity is important. Towards this end, let us turn our attention to Figure 6 which shows two matrices with $\mathcal{O}(p)$ non-zeros. Suppose that the non-zero pattern in X looks like that of the matrix on the left (which we dub as the "arrow" matrix). It is clear that it is impossible to recover this X from AXB^T *even if* we know the non-zero pattern in advance. For instance, if $v \in \ker(A)$, then the matrix $\tilde{X}$, with v added to the first column of X is such that $AXB^T = A\tilde{X}B^T$ and $\tilde{X}$ is also an arrow matrix and hence indistinguishable from X. Similarly, one can "hide" a kernel vector of B in the first row of the arrow matrix. In other words, it is impossible to uniquely recover X from AXB^T.

In what follows, we will show that if the sparsity pattern of X is more *distributed*, as in the right side of Figure 6, then one can recover X and do so efficiently. In fact, our analysis will also reveal what that the size of the sketch Y needs to be able to perform this task and we will see that this is very close to being optimal.
In order to make things concrete, we will now define these notions formally.

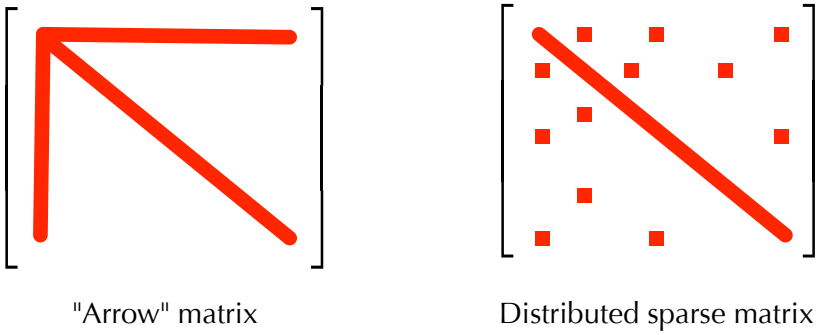


Figure 6: Two matrices with $\mathcal{O}(p)$ non-zeros. The "arrow" matrix is impossible to recover by covariance sketching while the distributed sparse matrix is.

Definition 3 ($d-$distributed sets and $d-$distributed sparse matrices)**.** *We say that a subset* $\Omega \subset [p] \times [p]$ *is* $d-$**distributed** *if the following hold.*

1. *For* $i = 1, 2, \ldots, p$, $(i, i) \in \Omega$.
2. *For all* $k \in [p]$, *the cardinality of the sets* $\Omega_k := \{(i, j) \in \Omega : i = k\}$ *and* $\Omega^k := \{(i, j) \in \Omega : j = k\}$ *is no more than* d.

The set of all $d-$*distributed subsets of* $[p] \times [p]$ *will be denoted as* $\mathfrak{W}_{p,d}$. *We say that a matrix* $X \in \mathbb{R}^{p \times p}$ *is* $d-$**distributed sparse** *if there exists an* $\Omega \in \mathfrak{W}_{d,p}$ *such that* $supp(X) := \{(i, j) \in [p] \times [p] : X_{ij} \neq 0\} \subset \Omega$.

While the theory we develop here is more generally applicable, the first point in the above definition makes the presentation easier. Notice that this forces the number of off-diagonal non-zeros in each row or column of a $d-$distributed sparse matrix X to be at most $d-1$. This is not a serious limitation since a more careful analysis along these lines can only improve the bounds we obtain by at most a constant factor.
Examples:

- Any diagonal matrix is $d-$distributed sparse for $d = 1$. Similarly a tridiagonal matrix is $d-$distributed sparse with $d = 3$.
- The adjacency matrix of a bounded degree graph with maximum degree $d - 1$ is $d-$distributed sparse
- As shown in Proposition 1, *random sparse matrices* are $d-$distributed sparse with $d = \mathcal{O}(\log p)$. While we implicitly assume that d is constant with respect to p in what follows, all arguments work even if d grows logarithmically in p as is the case here.

Given a matrix $X \in \mathbb{R}^{p \times p}$, as mentioned earlier, we write $\text{vec}(X)$ to denote the $\mathbb{R}^{p^2}$ vector obtained by stacking the columns of X. Suppose $x = \text{vec}(X)$. It will be useful in what follows to remember that x was actually derived from the matrix X and hence we will employ slight abuses of notation as follows: We will say x is $d-$distributed sparse when

we actually mean that the matrix X is. Also, we will write (i, j) to denote the index of x corresponding to X_{ij}, i.e.,

$$x_{ij} := x_{(i-1)p+j} = X_{ij}.$$

We finally make two remarks:

- Even if X were distributed sparse, the observed vector $Y = AXB^T$ is usually unstructured (and dense).
- While the results in this paper focus on the regime where the maximum number of non-zeros d per row/column of X is a constant (with respect to p), they can be readily extended to the case when d grows poly-logarithmically in p. Extensions to more general scalings of d is an interesting avenue for future work.

3.2 Random Bipartite Graphs, Weak Distributed Expansion and the Choice of the Sketching Matrices

As alluded to earlier, we will choose the sensing matrices A, B to be the adjacency matrices of certain random bipartite graphs. The precise definition of this notion follows.

Definition 4 (Uniformly Random $\delta-$left regular bipartite graph)**.** *We say that $G = ([p], [m], E)$ is a uniformly random $\delta-$left regular bipartite graph if the edge set E is a random variable with the following property: for each $i \in [p]$ one chooses δ vertices $j_1, j_2, \ldots, j_\delta$ chosen uniformly and independently at random (with replacement) from $[m]$ such that $\{\{i, j_k\}\}_{k=1}^{\delta} \subset E$.*

Remarks:

- Note that since we are sampling with replacement, it follows that the bipartite graph thus constructed may not be simple. If for instance there are two edges from the left node i to the right node i, the corresponding entry $A_{ij} = 2$.
- It is in fact possible to work with a sampling without replacement model in Definition 4 (where the resulting graph is indeed simple) and obtain qualititatively the same results. We work with a "sampling with replacement" model for the ease of exposition.

The probabilistic claims in this paper are made with respect to this probability distribution on the space of all bipartite graphs.

In past work [5, 35], the authors show that a random graph generated as above is, for suitable values of ϵ, δ, a $(k, \delta, \epsilon)-$expander. That is, for all sets $S \subset [p]$ such that $|S| \leq k$, the size of the neighborhood $|N(S)|$ is no less than $(1 - \epsilon)\delta |S|$. If A is the adjacency matrix of such a graph, then it can then be shown that this implies that ℓ_1 minimization would recover a $k-$sparse vector x if one observes the sketch Ax (actually, [5] shows that these two properties are equivalent). Notice that in our context, the vector that we need to recover is $\mathcal{O}(p)$ sparse and therefore, our random graph needs to be a $(\mathcal{O}(p), \delta, \epsilon)-$expander. Unfortunately, this turns out to not be true of $G_1 \otimes G_2$ when G_1 and G_2 are randomly chosen as directed above.

However, we prove that if G_1 and G_2 are picked as in Definition 4, then their tensor graph $G_1 \otimes G_2$, satisfies what can be considered a *weak distributed expansion* property. This roughly says that the neighborhood of a $d-$distributed $\Omega \subset [p] \times [p]$ is large enough. Moreover, we show that this is in fact sufficient to prove that with high probability X can be recovered from AXB^T efficiently. The precise statement of these combinatorial claims follows.

Lemma 1. *Suppose that $G_1 = ([p], [m], E_1)$ and $G_2 = ([p], [m], E_2)$ are two independent uniformly random $\delta-$left regular bipartite graphs with $\delta = \mathcal{O}(\log p)$ and $m = \mathcal{O}(\sqrt{dp} \log p)$. Let $\Omega \in \mathfrak{W}_{d,p}$ be fixed. Then there exists an $\epsilon \in \left(0, \frac{1}{4}\right)$ such that $G_1 \otimes G_2$ has the following properties with probability exceeding $1 - p^{-c}$, for some $c > 0$.*

1. *$|N(\Omega)| \geq p\delta^2(1-\epsilon)$.*
2. *For any $(i, i') \in ([p] \times [p]) \setminus \Omega$ we have $|N(i, i') \cap N(\Omega)| \leq \epsilon\delta^2$.*
3. *For any $(i, i') \in \Omega$, $|N(i, i') \cap N(\Omega \setminus (i, i'))| \leq \epsilon\delta^2$.*

Moreover, all these claims continue to hold when G_2 is the same as G_1.

Remarks:

- Part 1 of Lemma 1 says that if Ω is a $d-$distributed set, then the size of the neghborhood of Ω is large. This can be considered a *weak distributed expansion property*. Notice that while it is reminiscent of the *vertex expansion* property of expander graphs, it is easy to see that it does not hold if Ω is not distributed. Furthermore, we call it "weak" because the lower bound on the size of the neighborhood is $\delta^2 p(1-\epsilon)$ as opposed to $\delta^2 |\Omega| (1-\epsilon) = \delta^2 dp(1-\epsilon)$ as one typically gets for standard expander graphs. It will become clear that this is one of the key combinatorial facts that ensures that the necessary theoretical guarantees hold for covariance sketching.
- Parts 2 and 3 say that the number of collisions between the edges emanating out of a single vertex with the edges emanating out of a distributed set is small. Again, this combinatorial property is crucial for the proof of Theorem 1 to work.

As stated earlier, we are particularly interested in the challenging case when $G_1 = G_2$ (or equivalently, their adjacency matrices A, B are the same). The difficultly, loosely speaking, stems from the fact that since we are not allowed to pick G_1 and G_2 separately, we have much less independence. In Appendix A, we will only prove Lemma 1 in the case when $G_1 = G_2$ since this proof can be modified in a straightforward manner to obtain the proof of the case when G_1 and G_2 are drawn independently.

4 Proofs of Main Results

In this section, we will prove the main theorem. To reduce clutter in our presentation, we will sometimes employ certain notational shortcuts. When the context is clear, the ordered

pair (i, i') will simply be written as ii' and the set $[p] \times [p]$ will be written as $[p]^2$. Sometimes, we will also write $\mathcal{A}$ to mean $A \otimes A$ and if $S \subset [p] \times [p]$, we write $(A \otimes A)_S$ or $\mathcal{A}_S$ to mean the submatrix of $A \otimes A$ obtained by appending the columns $\{A_i \otimes A_j \mid (i, j) \in S\}$.

As stated earlier, we will only provide the proof of Theorem 1 for the case when $A = B$. Some straightforward changes to the proof presented here readily gives one the proof for the case when the matrices A and B are distinct.

4.1 Proof of Theorem 1

We will consider an arbitrary ordering of the set $[p] \times [p]$ and we will order the edges in $E^{\otimes}$ lexicographically based on this ordering, i.e., the first δ^2 edges $e_1, \ldots, e_{\delta^2}$ in $E^{\otimes}$ are those that correspond to the first element as per the ordering on $[p] \times [p]$ and so on. Now, one can imagine that the graph $G^{\otimes}$ is formed by including these edges sequentially as per the ordering on the edges. This allows us to partition the edge set into the set $E_1^{\otimes}$ of edges that do not collide with any of the previous edges as per the ordering and the set $E_2^{\otimes} := E^{\otimes} - E_1^{\otimes}$. (We note that a similar proof technique was adopted in Berinde et al. [5]).

As a first step towards proving the main theorem, we will show that the operator $A \otimes A$ preserves the ℓ_1 norm of a matrix X as long as X is distributed sparse. Berinde et al.,[5] call a similar property RIP-1, taking cues from the restricted isometry property that has become popular in literature [11]. The proposition below can also be considered to be a restricted isometry property but the operator in our case is only constrained to behave like an isometry for distributed sparse vectors.

Proposition 2 (ℓ_1-RIP)**.** *Suppose $X \in \mathbb{R}^{p \times p}$ is $d-$distributed sparse and A is the adjacency matrix of a random bipartite $\delta-$left regular graph. Then there exists an $\epsilon > 0$ such that*

$$(1 - 2\epsilon)\delta^2 \|X\|_1 \leq \left\|AXA^T\right\|_1 \leq \delta^2 \|X\|_1 , \tag{6}$$

with probability exceeding $1 - p^{-c}$ for some $c > 0$.

Proof. The upper bound follows (deterministically) from the fact that the induced (matrix) ℓ_1-norm of $A \otimes A$, i.e., the maximum column sum of $A \otimes A$, is precisely δ^2. To prove the lower bound, we need the following lemma.

Lemma 2. *For any $X \in \mathbb{R}^{p \times p}$,*

$$\left\|AXA^T\right\|_1 \geq \delta^2 \|X\|_1 - 2 \sum_{jj' \in [m]^2} \sum_{ii' \in [p]^2} \mathbf{1}_{\{ii', jj'\} \in E_2^{\otimes}} |X_{ii'}| \tag{7}$$

Proof. In what follows, we will denote the indicator function $\mathbf{1}_{\{ii', jj'\} \in S}$ by $\mathbf{1}_S^{\{ii'jj'\}}$. We begin

by observing that

$$\begin{aligned}
\left\|AXA^T\right\|_1 &= \|\mathcal{A}\,\mathrm{vec}(X)\|_1 \\
&= \sum_{jj'\in[m]^2}\left|\sum_{ii'\in[p]^2}\mathcal{A}_{\{ii'jj'\}}X_{ii'}\right| \\
&= \sum_{jj'\in[m]^2}\left|\sum_{ii'\in[p]^2}\mathbf{1}_{E^{\otimes}}^{\{ii'jj'\}}\,X_{ii'}\right| \\
&= \sum_{jj'\in[m]^2}\left|\sum_{ii'\in[p]^2}\mathbf{1}_{E_1^{\otimes}}^{\{ii'jj'\}}X_{ii'}+\sum_{ii'\in[p]^2}\mathbf{1}_{E_2^{\otimes}}^{\{ii'jj'\}}X_{ii'}\right| \\
&\geq \sum_{jj'\in[m]^2}\left|\sum_{ii'\in[p]^2}\mathbf{1}_{E_1^{\otimes}}^{\{ii'jj'\}}X_{ii'}\right|-\left|\sum_{ii'\in[p]^2}\mathbf{1}_{E_2^{\otimes}}^{\{ii'jj'\}}X_{ii'}\right| \\
&\overset{(a)}{\geq} \sum_{jj'\in[m]^2}\Big(\sum_{ii'\in[p]^2}\mathbf{1}_{E_1^{\otimes}}^{\{ii'jj'\}}\left|X_{ii'}\right|-\sum_{ii'\in[p]^2}\mathbf{1}_{E_2^{\otimes}}^{\{ii'jj'\}}\left|X_{ii'}\right|\Big) \\
&= \sum_{ii'\in[p]^2,jj'\in[m]^2}\mathbf{1}_{E^{\otimes}}^{\{ii'jj'\}}\left|X_{ii'}\right|-2\sum_{ii'\in[p]^2jj'\in[m]^2}\mathbf{1}_{E_2^{\otimes}}^{\{ii'jj'\}}\left|X_{ii'}\right|,
\end{aligned}$$

where (a) follows after observing that the first (double) sum has only one term and applying triangle inequality to the second sum. Since $\sum_{ii'\in[p]^2,jj'\in[m]^2}\mathbf{1}_{E^{\otimes}}^{\{ii'jj'\}}\left|X_{ii'}\right|=\delta^2\left\|X\right\|_1$, this concludes the proof of the lemma. □

Now, to complete the proof of Proposition 2, we need to bound the sum in the LHS of (7). Notice that

$$\sum_{ii'jj':(ii'jj')\in E_2^{\otimes}}\left|X_{ii'}\right|=\sum_{ii'}\left|X_{ii'}\right|r_{ii'}=\sum_{ii'\in\Omega}\left|X_{ii'}\right|r_{ii'}.$$

where $r_{ii'}$ is the number of collisions of edges emanating from ii' with all the previous edges as per the ordering we defined earlier. Since Ω is $d-$distributed, from the third part of Lemma 1, we have that for all $ii'\in\Omega$, $r_{ii'}\leq\epsilon\delta^2$ with probability exceeding $1-p^{-c}$ and therefore,

$$\sum_{ii'\in\Omega}\left|X_{ii'}\right|r_{ii'}\leq\epsilon\delta^2\left\|X\right\|_1.$$

This concludes the proof. □

Next, we will use the fact that $A\otimes A$ behaves as an approximate isometry (in the ℓ_1 norm) to prove what can be considered a nullspace property [19, 15]. This will tell us that the nullspace of $A\otimes A$ is "smooth" with respect to distributed support sets and hence ℓ_1 minimization as proposed in (P_1) will find the right solution.

Proposition 3 (Nullspace Property)**.** *Suppose that $A \in \{0,1\}^{m\times p}$ is the adjacency matrix of a random bipartite $\delta-$left regular graph with $\delta = \mathcal{O}(\log p)$ and $m = \mathcal{O}(\sqrt{dp}\log p)$ and that $\Omega \in \mathfrak{W}_{d,p}$ is fixed. Then, with probability exceeding $1-p^{-c}$, for any $V \in \mathbb{R}^{p\times p}$ such that $AVA^T = 0$, we have*

$$\|V_\Omega\|_1 \leq \frac{\epsilon}{1-3\epsilon}\|V_{\Omega^c}\|_1 . \tag{8}$$

for some $\epsilon \in (0, \frac{1}{4})$ and for some $c > 0$.

Proof. Let V be any symmetric matrix such that $AVA^T = 0$. Let $v = \mathrm{vec(V)}$ and note that $\mathrm{vec}\big(AVA^T\big) = (A\otimes A)\, v = 0$. Let Ω be a d-distributed set. As indicated in Section 3, we define $N(\Omega) \subseteq [m]^2$ to be the set of neighbors of Ω with respect to the graph $G^{\otimes}$. Let $(A\otimes A)^{N(\Omega)}$ denote the submatrix of $A\otimes A$ that contains only those rows corresponding to $N(\Omega)$ (and all columns). We will slightly abuse notation and use v_Ω to denote the vectorization of the projection of V onto the set Ω, i.e., $v_\Omega = \mathrm{vec}(V_\Omega)$, where

$$[V_\Omega]_{i,j} = \begin{cases} V_{i,j} & (i,j)\in\Omega \\ 0 & \text{otherwise} \end{cases}.$$

Now, we can follow the following chain of inequalities:

$$\begin{aligned}
0 &= \left\|(A\otimes A)^{N(\Omega)} v\right\|_1 \\
&= \left\|(A\otimes A)^{N(\Omega)}(v_\Omega + v_{\Omega^c})\right\|_1 \\
&\geq \left\|(A\otimes A)^{N(\Omega)} v_\Omega\right\|_1 - \left\|(A\otimes A)^{N(\Omega)} v_{\Omega^c}\right\|_1 \\
&= \|(A\otimes A) v_\Omega\|_1 - \left\|(A\otimes A)^{N(\Omega)} v_{\Omega^c}\right\|_1 \\
&\geq (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \left\|(A\otimes A)^{N(\Omega)} v_{\Omega^c}\right\|_1 ,
\end{aligned}$$

where the last inequality follows from Proposition 2. Resuming the chain of inequalities, we have:

$$\begin{aligned}
0 &\geq (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \sum_{ii'\in\Omega^c} \left\|(A\otimes A)^{N(\Omega)} v_{\{ii'\}}\right\|_1 \\
&\geq (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \sum_{\substack{ii',jj':(ii',jj')\in E^{\otimes}, \\ jj'\in N(\Omega), ii'\in\Omega^c}} |V_{ii'}| \\
&= (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \sum_{ii'\in\Omega^c} |E^{\otimes}(ii' : N(\Omega))|\, |V_{ii'}| \\
&\overset{(a)}{\geq} (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \sum_{ii'\in\Omega^c} \epsilon\delta^2\ |V_{ii'}| \\
&\geq (1-2\epsilon)\delta^2 \|V_\Omega\|_1 - \epsilon\delta^2 \|V\|_1 ,
\end{aligned}$$

where (a) follows from the second part of Lemma 1. Writing $\|V\|_1 = \|V_\Omega\|_1 + \|V_{\Omega^c}\|_1$ and rearranging, we get the required result. □

Now, we can use this to prove our main theorem.

Proof of Theorem 1. Let Ω be the support of X and notice that Ω is $d-$distributed. Now, suppose that there exists an $\tilde{X} \neq X$ such that $A\tilde{X}A^T = Y$. Observe that $A\left(\tilde{X} - X\right)A^T = 0$. Now, consider

$$\begin{aligned}
\|X\|_1 &\leq \left\|X - \tilde{X}_\Omega\right\|_1 + \left\|\tilde{X}_\Omega\right\|_1 \\
&= \left\|\left(X - \tilde{X}\right)_\Omega\right\|_1 + \left\|\tilde{X}_\Omega\right\|_1 \\
&\overset{(a)}{\leq} \frac{\epsilon}{1-3\epsilon}\left\|\left(X - \tilde{X}\right)_{\Omega^c}\right\| + \left\|\tilde{X}_\Omega\right\|_1 \\
&= \frac{\epsilon}{1-3\epsilon}\left\|-\tilde{X}_{\Omega^c}\right\| + \left\|\tilde{X}_\Omega\right\|_1 \\
&< \left\|\tilde{X}\right\|_1
\end{aligned}$$

where (a) follows from Proposition 3, and the last line follows from the fact that $\epsilon < \frac{1}{4}$, again from Proposition 3. Therefore, the unique solution of (P_1) is X with probability exceeding $1 - p^{-c}$, for some $c > 0$. □

5 Conclusions

In this paper we have introduced the notion of distributed sparsity for matrices. We have shown that when a matrix is X distributed sparse, and A, B are suitable random binary matrices, then it is possible to recover X from under-determined linear measurements of the form $Y = AXB^T$ via ℓ_1 minimization. We have also shown that this recovery procedure is robust in the sense that if X is equal to a distributed sparse matrix plus a perturbation, then our procedure returns an approximation with accuracy proportional to the size of the perturbation. Our results follow from a new lemma about the properties of tensor products of random bipartite graphs. We also describe three interesting applications where our results would be directly applicable.

In future work, we plan to investigate the statistical behavior and sample complexity of estimating a distributed sparse matrix (and its exact support) in the presence of various sources of noise (such as additive Gaussian noise, and Wishart noise). We expect an interesting trade-off between the sketching dimension and the sample complexity.

References

[1] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Graph sketches: sparsification, spanners, and subgraphs. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 5–14. ACM, 2012.

[2] Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David Woodruff. Efficient sketches for earth-mover distance, with applications. In *in FOCS*, 2009.

[3] Dana Angluin and Leslie G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and system Sciences*, 18(2):155–193, 1979.

[4] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Proceedings of the 48th Annual Allerton Conference*, 2010.

[5] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 798 –805, sept. 2008.

[6] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997.

[7] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[8] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[9] B. Bollobás. *Random graphs*, volume 73. Cambridge university press, 2001.

[10] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

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

[12] Emmanuel J Candes, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *arXiv preprint arXiv:0912.3599*, 2009.

[13] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The Convex Geometry of Linear Inverse Problems. *ArXiv e-prints*, December 2010.

[14] Venkat Chandrasekaran, Benjamin Recht, PabloA. Parrilo, and AlanS. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12:805–849, 2012.

[15] Albert Cohen, Wolfgang Dahmen, and Ronald Devore. COMPRESSED SENSING AND BEST k-TERM APPROXIMATION. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

[16] Graham Cormode and S Muthukrishnan. Combinatorial algorithms for compressed sensing. *Structural Information and Communication Complexity*, pages 280–294, 2006.

[17] Peter J Diggle and Arūnas P Verbyla. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, pages 401–415, 1998.

[18] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ?1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[19] David L Donoho and Xiaoming Huo. Uncertaintly Principles and Ideal Atomic Decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[20] David Leigh Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[21] Marco F Duarte and Richard G Baraniuk. Kronecker compressive sensing. *Image Processing, IEEE Transactions on*, 21(2):494–504, 2012.

[22] M.F. Duarte and R.G. Baraniuk. Kronecker product matrices for compressive sensing. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3650 –3653, march 2010.

[23] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):43, 2007.

[24] Anna C Gilbert and Kirill Levchenko. Compressing network graphs. In *Proceedings of the LinkKDD workshop at the 10th ACM Conference on KDD*. Citeseer, 2004.

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

[26] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320–3325, 2003.

[27] András Hajnal and Endre Szemerédi. Proof of a conjecture of erdos. *Combinatorial theory and its applications*, 2:601–623, 1970.

[28] Jarvis Haupt, Waheed U Bajwa, Gil Raz, and Robert Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *Information Theory, IEEE Transactions on*, 56(11):5862–5875, 2010.

[29] Larry V Hedges, Ingram Olkin, Mathematischer Statistiker, Ingram Olkin, and Ingram Olkin. *Statistical methods for meta-analysis*. Academic Press New York, 1985.

[30] CVX Research, Inc. CVX: Matlab software for disciplined convex programming, version 2.0 beta, September 2012.

[31] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1–1, 1984.

[32] S. Jokar. Sparse recovery and kronecker products. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1 –4, march 2010.

[33] Sadegh Jokar and Volker Mehrmann. Sparse solutions to underdetermined kronecker product systems. *Linear Algebra and its Applications*, 431(12):2437–2447, 2009.

[34] Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, STOC '11, pages 745–754, New York, NY, USA, 2011. ACM.

[35] M. Amin Khajehnejad, Alexandros G. Dimakis, Weiyu Xu, and Babak Hassibi. Sparse recovery of positive signals with minimal expansion. *CoRR*, abs/0902.4045, 2009.

[36] MAmin Khajehnejad, Alexandros G Dimakis, Weiyu Xu, and Babak Hassibi. Sparse recovery of nonnegative signals with minimal expansion. *Signal Processing, IEEE Transactions on*, 59(1):196–208, 2011.

[37] H. A. Kierstead and A. V. Kostochka. A short proof of the Hajnal-Szemeredi Theorem on equitable colouring. *Comb. Probab. Comput.*, 17(2):265–270, March 2008.

[38] S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[39] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[40] S Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.

[41] Sriram V. Pemmaraju. Equitable colorings extend Chernoff-Hoeffding bounds. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, SODA '01, pages 924–925, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics.

[42] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 2011.

[43] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[44] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[45] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science Signalling*, 308(5721):523, 2005.

[46] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

[47] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[48] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

Appendix A

Proof of Lemma 1

Lemma 1. *Suppose that $G_1 = ([p], [m], E_1)$ and $G_2 = ([p], [m], E_2)$ are two independent uniformly random $\delta-$left regular bipartite graphs with $\delta = \mathcal{O}(\log p)$ and $m = \mathcal{O}(\sqrt{dp}\log p)$. Let $\Omega \in \mathfrak{W}_{d,p}$ be fixed. Then there exists an $\epsilon \in \left(0, \frac{1}{4}\right)$ such that $G_1 \otimes G_2$ has the following properties with probability exceeding $1 - p^{-c}$, for some $c > 0$.*

1. *$|N(\Omega)| \geq p\delta^2(1-\epsilon)$.*
2. *For any $(i, i') \in ([p] \times [p]) \setminus \Omega$ we have $|N(i, i') \cap N(\Omega)| \leq \epsilon\delta^2$.*
3. *For any $(i, i') \in \Omega$, $|N(i, i') \cap N(\Omega \setminus (i, i'))| \leq \epsilon\delta^2$.*

Moreover, all these claims continue to hold when G_2 is the same as G_1.

Proof. As stated earlier, we will only prove this lemma for the case when $G_1 = G_2$. With a few minor modifications, one can readily get a proof for the easier case when G_1 and G_2 are drawn independently.

Let $\mathcal{E}_1, \mathcal{E}_2$ and, $\mathcal{E}_3$ respectively denote the events that the implications (1), (2) and, (3) are true. Notice that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c\right) &\leq \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c) \\
&= \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c \mid \mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2^c \mid \mathcal{E}_1^c)\mathbb{P}(\mathcal{E}_1^c) \\
&\qquad + \mathbb{P}(\mathcal{E}_3^c \mid \mathcal{E}_1)\mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_3^c \mid \mathcal{E}_1^c)\mathbb{P}(\mathcal{E}_1^c) \\
&\leq 3\mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c \mid \mathcal{E}_1) + \mathbb{P}(\mathcal{E}_3^c \mid \mathcal{E}_1)
\end{aligned}$$

Our strategy will be to upper bound $\mathbb{P}(\mathcal{E}_1^c), \mathbb{P}(\mathcal{E}_2^c \mid \mathcal{E}_1)$ and, $\mathbb{P}(\mathcal{E}_3^c \mid \mathcal{E}_1)$. Suppose the bounds were p_1, p_2 and, p_3 respectively, then it is easy to see that

$$\mathbb{P}\left\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3\right\} \geq 1 - \max\{3p_1, p_2, p_3\}. \tag{9}$$

Part 1. We will first show that $\mathbb{P}(\mathcal{E}_1^c)$ is small. Since Ω is $d-$distributed, the "diagonal" set $\mathcal{D} := \{(1,1), \ldots, (p,p)\}$ is a subset of Ω. Now, notice that for $j \neq j' \in [m], i \in [p]$,

$$\mathbb{P}\left[(j,j') \in N((i,i))\right] = \frac{\delta(\delta-1)}{m(m-1)} \tag{10}$$

This implies that

$$\mathbb{P}\left[(j,j') \notin N(\mathcal{D})\right] = \left(1 - \frac{\delta(\delta-1)}{m(m-1)}\right)^{|\mathcal{D}|}$$

Therefore, we can bound the expected value of $|N(\Omega)|$ as follows.

$$\begin{aligned}
\mathbb{E}\big[\,|N(\Omega)|\,\big] &\geq \mathbb{E}\big[\,|N(\mathcal{D})|\,\big] \\
&= \sum_{jj' \in [m]\times[m]} \mathbb{P}\left[(j,j') \in N(\mathcal{D})\right] \\
&\geq \sum_{\substack{jj' \in [m]\times[m], \\ j \neq j'}} \mathbb{P}\left[(j,j') \in N(\mathcal{D})\right] \\
&= \sum_{\substack{jj' \in [m]\times[m], \\ j \neq j'}} \left(1 - \left(1 - \frac{\delta(\delta-1)}{m(m-1)}\right)^{|\mathcal{D}|}\right) \\
&= m(m-1)\left(1 - \left(1 - \frac{\delta(\delta-1)}{m(m-1)}\right)^{|\mathcal{D}|}\right) \\
&\geq m(m-1)\left(\frac{|\mathcal{D}|\,\delta(\delta-1)}{m(m-1)} - \frac{|\mathcal{D}|^2\,\delta^2(\delta-1)^2}{m^2(m-1)^2}\right) \\
&= |\mathcal{D}|\,\delta^2\left(1 - \left(\frac{1}{\delta} + \frac{(\delta-1)^2\,|\mathcal{D}|}{m(m-1)}\right)\right) \\
&= p\delta^2\left(1 - \epsilon'\right).
\end{aligned}$$

Where in the last step, we set $\epsilon' = \frac{1}{\delta} + \frac{(\delta-1)^2|\mathcal{D}|}{m(m-1)}$.

To complete the proof, we must show that the random quantity $|N(\Omega)|$ cannot be much smaller than $p\delta^2(1-\epsilon')$. As a first step, we define the random variables $\chi_{jj'} := \mathbf{1}_{\{(j,j') \in N(\mathcal{D})\}}$ and notice that the following chain of inequalities hold

$$|N(\Omega)| \geq |N(\mathcal{D})| \geq \sum_{\substack{jj' \in [m]\times[m] \\ j \neq j'}} \chi_{jj'}.$$

Therefore, we have that

$$\mathbb{P}\left[|N(\Omega)| < p\delta^2(1-\epsilon'-\epsilon'')\right] \leq \mathbb{P}\left[\sum_{\substack{jj'\in[m]\times[m]\\ j\neq j'}} \chi_{jj'} < p\delta^2(1-\epsilon'-\epsilon'')\right].$$

Also, since by above, $\mathbb{E}\left[\sum_{j\neq j'}\chi_{jj'}\right] \geq p\delta^2(1-\epsilon')$, we have that

$$\mathbb{P}\left[|N(\Omega)| < p\delta^2(1-\epsilon)\right] \leq \mathbb{P}\left[\sum_{\substack{jj'\in[m]\times[m]\\ j\neq j'}} \chi_{jj'} < \mathbb{E}\left[\sum_{\substack{jj'\in[m]\times[m]\\ j\neq j'}} \chi_{jj'}\right] - p\delta^2\epsilon''\right].$$

Now, notice that the sum $\sum_{j\neq j'}\chi_{jj'}$ has $m(m-1)$ terms and each term in the sum is dependent on no more than $2m-4$ terms. Therefore, one way to bound the required quantity is to extract independent sub-sums from the above sum and bound the deviation of each of those from their means (which is the corresponding sub-sum of the mean). A principled way of doing this is suggested by the celebrated Hajnal-Szemeridi theorem [27, 37]. Consider a graph on the vertex set $[m]\times[m]\setminus\{(1,1),\ldots,(m,m)\}$ where there is an edge between vertices (j,j') and (j_1,j_1') if $j=j_1$ and/or $j'=j_1'$, i.e., exactly when the random variables $\chi_{jj'}$ and $\chi_{j_1j_1'}$ are dependent. Since this graph has degree $\Theta(m)$, Hajnal-Szemeridi theorem tells us that this graph can be equitable colored with $\Theta(m)$ colors. In other words, the above sum can be partitioned into $\Theta(m)$ sub-sums such that each sub-sum has $\Theta(m)$ elements and the random variables in each of them are independent. Along with this and the fact that $m(m-1) > p\delta^2$, we can use the union bound and write

$$\begin{aligned}\mathbb{P}\left[\sum_{\substack{jj'\in[m]\times[m]\\ j\neq j'}} \chi_{jj'} < \mathbb{E}\left[\sum_{\substack{jj'\in[m]\times[m]\\ j\neq j'}} \chi_{jj'}\right] - p\delta^2\epsilon''\right]&\\ \leq \Theta(m)\,\mathbb{P}\left[\frac{1}{|C_1|}\sum_{jj'\in C_1}\chi_{jj'} < \mathbb{E}\left[\frac{1}{|C_1|}\sum_{jj'\in C_1}\chi_{jj'}\right] - \epsilon''\right],&\end{aligned}$$

where C_1 is one of the "colors". Notice that $|C_1| = \Theta(m)$.

Finally, using Chernoff bounds, we have

$$\begin{aligned}\Theta(m)\,\mathbb{P}\left[\frac{1}{|C_1|}\sum_{jj'\in C_1}\chi_{jj'} < \frac{1}{|C_1|}\mathbb{E}\left[\sum_{jj'\in C_1}\chi_{jj'}\right] - \epsilon''\right]&\\ \leq \Theta(m)\ \exp\left\{-2\epsilon''^2\Theta(m)\right\}.&\end{aligned}$$

Finally, since ϵ' can be made as small as possible, setting $\epsilon := \epsilon'+\epsilon''$ yeilds $p_1 < p^{-c_1}$ for some $c_1 > 0$. This technique of generating large deviation bounds when one has limited dependence is not new, see [41].

Part 2: Now, we bound $\mathbb{P}(\mathcal{E}_2^c \mid \mathcal{E}_1)$. Associated to a fixed i one can imagine δ independent random trials that determine the outgoing edges from i. In a similar way there are δ independent random trials associated to the outgoing edges of i'. Let us fix (i, i') and investigate the outgoing edge (in the tensor graph) determined by the first trial of i and the first trial of i'. The probability that this edge emanating from the vertex $(i, i') \in [p]^2 \setminus \{(1,1), \ldots, (p,p)\}$ hits an arbitrary vertex $(j, j') \in [m]^2$ is given by $1/m^2$. The probability that this edge lands in $N(\Omega)$ is, therefore, given by $|N(\Omega)|/m^2$. Since there are δ^2 edges that are incident on (i, i'), the expected size of overlap between $N(i, i')$ and $N(\Omega)$ is upper bounded by

$$\delta^2 \frac{|N(\Omega)|}{m^2}.$$

Again, to show concentration, we employ similar arguments as before and define indicator random variables $\chi_1 \ldots, \chi_{\delta^2}$ each of which corresponds to one of the edges emanating from the vertex (i, i') and then observing that the sum $\sum_{k=1}^{\delta^2} \chi_k$ is precisely equal to the random quantity $|N(i, i') \cap N(\Omega)|$. To conclude that this random quantity concentrates, we first observe that, as above, the δ^2 dependent terms can be divided up into $\Theta(\delta)$ with $\Theta(\delta)$ elements each such that in each set the terms are independent. Therefore, we have

$$\begin{aligned}
\mathbb{P}\left[|N(i,i') \cap N(\Omega)| > \delta^2 \frac{|N(\Omega)|}{m^2}(1+\epsilon')\right] &= \mathbb{P}\left[\sum_{k=1}^{\delta^2} \chi_k > \delta^2 \frac{|N(\Omega)|}{m^2}(1+\epsilon')\right] \\
&\leq \Theta(\delta)\, \mathbb{P}\left[\sum_{k \in C_1} \chi_k > \Theta(\delta) \frac{|N(\Omega)|}{m^2}(1+\epsilon')\right] \\
&\leq \Theta(\delta) \exp\left\{-\delta \frac{|N(\Omega)|}{m^2}\epsilon'\right\},
\end{aligned}$$

where C_1 is one of the colors.

Therefore, since conditioned on $\mathcal{E}_1$, $|N(\Omega)| > \delta^2 p(1-\epsilon)$ if we pick $m = \delta\sqrt{dp}$, and $\delta = \Theta(\log p)$ there is a $c_2' = c_2'(\epsilon') > 2$ such that, $|N(i,i') \cap N(\Omega)| > \delta^2 \frac{|N(\Omega)|}{m^2}(1+\epsilon')$, with probability not exceeding $p^{-c_2'}$. Setting $\epsilon = (1+\epsilon')\,|N(\Omega)|/m^2$, picking m as prescribed, and taking union bound over $(i, i') \in \Omega^c$, we get $p_2 \leq p^{-c_2}$ for some $c_2 > 0$.

Part 3: Next, we bound $\mathbb{P}(\mathcal{E}_3^c \mid \mathcal{E}_1)$. Notice that the proof is very similar to that of part 2 when $i \neq i'$. So, here we will consider the quantity $|N(i,i) \cap N(\Omega \setminus \{(i,i)\})|$.

As explained above to each left node i we associate δ random trials that determine its outgoing edges. Correspondingly, if we fix a left node (i, i) in the tensor graph, and think of its outgoing edges they are determined by the outcome of δ^2 product trials. Let $\mathbf{1}_k(j)$ be the indicator function of the event that in the k^{th} trial of i the outgoing edge is incident on j. The probability that the edge associated to the (k, l) trial associated to (i, i) is incident on (j, j') is the random variable $\mathbf{1}_k(j)\mathbf{1}_l(j')$. Note that $\mathbf{1}_k(j)\mathbf{1}_k(j') = \mathbf{1}_k(j)$ if $j = j'$ and 0 otherwise.

Note that

$$|N(i,i) \cap N(\Omega \setminus (i,i))| = \sum_{k=1}^{\delta} \sum_{l=1}^{\delta} \sum_{(j,j') \in N(\Omega \setminus (i,i))} \mathbf{1}_k(j) \mathbf{1}_l(j') \\ \leq \delta + \sum_{k \neq l} \sum_{(j,j') \in N(\Omega \setminus (i,i))} \mathbf{1}_k(j) \mathbf{1}_l(j')$$

When $k \neq l$, the trials corresponding to $\mathbf{1}_k(j), \mathbf{1}_l(j')$ are independent, and hence $\mathbb{E}\mathbf{1}_k(j)\mathbf{1}_l(j') = \frac{1}{m^2}$. We define $\chi_{k,l} := \sum_{(j,j') \in N(\Omega \setminus (i,i))} \mathbf{1}_k(j)\mathbf{1}_l(j')$ and note that $\mathbb{E}(\chi_{k,l}) = \frac{N(\Omega \setminus (i,i))}{m^2}$. We also note that $\chi_{k,l}$ is binary valued, and

$$|N(i,i) \cap N(\Omega \setminus (i,i))| \leq \delta + \sum_{k \neq l} \chi_{k,l}.$$

Therefore,

$$\begin{aligned} \mathbb{E}(|N(i,i) \cap N(\Omega \setminus (i,i))|) &\leq \delta + (\delta^2 - \delta) \frac{N(\Omega \setminus (i,i))}{m^2} \\ &\leq \delta + (\delta^2 - \delta) \frac{\delta^2 dp}{m^2} \\ &\leq \delta^2 \left(\frac{1}{\delta} + \left(1 - \frac{1}{\delta}\right) \frac{\delta^2 dp}{m^2} \right) \\ &\leq \delta^2 \epsilon. \end{aligned}$$

Next we need to prove that the quantity of interest $\sum_{k \neq l} \chi_{k,l}$ concentrates about its mean. To that end we note that these binary valued variables are such that any particular $\chi_{k,l}$ is dependent on at most $2\delta - 2$ other variables. Using Chernoff concentration bounds in conjunction with the Hajnal-Szemeredi based coloring argument explained in part 1 of this proof, followed by a union bound over $i \in [p]$ we obtain the required probability bounds $p_3 < p^{-c_3}$ for some $c_3 > 0$.

Substituting the bounds for p_1, p_2, p_3 back into (9) concludes the proof. □

Appendix B

Proof of Proposition 1

Proposition 1. *Consider a random matrix $X \in \mathbb{R}^{p \times p}$ such that $X_{ij} \overset{iid}{\sim} \text{Ber}(\gamma)$ where $p\gamma = \Delta = \Theta(1)$, then for any $\epsilon > 0$, X is $d-$distributed sparse with probability at least $1-\epsilon$, where*

$$d = \Delta \left(1 + \frac{2\log(2p/\epsilon)}{\Delta}\right).$$

Proof. Let $X_i, i = 1, \ldots, p$ denote the sparsity of the $i-$th column and let $X_i, i = p+1, \ldots, 2p$, denote the sparsity of the $i-$th row. Notice that the $2p$ random variables $X_1, X_2, \ldots, X_{2p}$ are (dependent) $\text{Bin}(p, \gamma)$ random variables. With the choice of d as indicated in the theorem, we have the following,

$$\begin{aligned}
\mathbb{P}(X_1 > d) &= \mathbb{P}\left(X_1 > \Delta\left(1 + \frac{\log(2p/\epsilon)}{\Delta}\right)\right) \\
&\overset{(a)}{\leq} \exp\left\{-\frac{\beta^2\Delta}{2+\beta}\right\}, \quad \beta = \frac{2\log(2p/\epsilon)}{\Delta} \\
&\overset{(b)}{\leq} \exp\left\{-\frac{\beta\Delta}{2}\right\} \\
&= \frac{\epsilon}{2p}
\end{aligned}$$

where (a) follows from the multiplicative form of the Chernoff Bound [3] and (b) follows as long as $\beta > 2$. The rest of the proof follows from a simple application of the union bound. □

Appendix C

Proof of Theorem 2

Theorem 2. *Suppose that X is a $p \times p$ matrix. Furthermore, suppose that the hypotheses of Theorem 1 hold and let X^* be the solution to the optimization program (P_1). Then, there exists a $c > 0$ and an $\epsilon \in (0, 1/4)$ such that the following holds with probability exceeding $1 - p^{-c}$.*

$$\|X^* - X\|_1 \leq \frac{2 - 4\epsilon}{1 - 4\epsilon}\left(\min_{\Omega \in \mathfrak{W}_{d,p}} \|X - X_\Omega\|_1\right). \tag{11}$$

Proof. Since X^* is the optimum of the optimization program (P_1), we have that $\|X\|_1 \geq \|X^*\|_1$. Let Ω^* be such that $\|X - X_{\Omega^*}\|_1 = \min_{\Omega \in \mathfrak{W}_{d,p}} \|X - X_X\|$. We can proceed as follows

$$\begin{aligned}
\|X\|_1 &\geq \|X^*\|_1 && (12)\\
&= \|(X + X^* - X)_\Omega\|_1 + \|(X + X^* - X)_{\Omega^c}\|_1 && (13)\\
&\geq \|X_\Omega\|_1 - \|(X^* - X)_\Omega\|_1 + \|(X^* - X)_{\Omega^c}\|_1 - \|X_{\Omega^c}\|_1 && (14)\\
&= \|X\|_1 - 2\|X_{\Omega^c}\|_1 + \|X^* - X\|_1 - 2\|(X - X^*)_\Omega\|_1 && (15)\\
&\geq \|X\|_1 - 2\|X_{\Omega^c}\|_1 + \left(1 - \frac{2\epsilon}{1 - 2\epsilon}\right)\|X^* - X\|_1 && (16)
\end{aligned}$$

where in the last step, we have used the fact that since X^* is a feasible point in (P_1), $AX^*B^T = AXB^T$ and therefore, we can apply the result of Proposition 3 to $X^* - X$. This completes the proof. □