Covariance Sketching

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

***Abstract*—Learning covariance matrices from high-dimensional data is an important problem that has received a lot of attention recently. We are particularly interested in the high-dimensional setting, where the number of samples one has access to is fewer than the number of variates. Fortunately, in many applications of interest, the underlying covariance matrix is sparse and hence has limited degrees of freedom. In most existing work however, it is assumed that one can obtain samples of all the variates simultaneously. This could be very expensive or physically infeasible in some applications. As a means of overcoming this limitation, we propose a new procedure that "pools" the covariates into a small number of groups and then samples each pooled group. We show that in certain cases it is possible to recover the covariance matrix from the pooled samples using an efficient convex optimization program, and so we call the procedure "covariance sketching".**

I. Introduction

An important feature of many modern data analysis problems is the presence of a large number of variates relative to the amount of available data. Such high-dimensional settings arise in a range of applications in bioinformatics, climate studies, and economics. A fundamental problem that arises in the high-dimensional regime is the poor behaviour of sample statistics such as empirical covariance matrices [3], [4]. Accordingly, a fruitful and active research agenda over the last few years has been the development of methods for high-dimensional statistical inference and modeling that take into account *structure* in the underlying model. Some examples of structural assumptions on statistical models include models with a few latent factors (leading to low-rank covariance matrices) [10], models specified by banded or sparse covariance matrices [3], [4], and Markov or graphical models [18], [19], [21].

In this paper we consider a scenario in which an unknown high-dimensional covariance matrix Σ is to be estimated. Due to the dimensionality of the variates, it is either infeasible, or prohibitively expensive to acquire samples of each variate. An alternative acquisition mechanism is to pool variates together and acquires samples of the pooled variates. The two questions of interest then are (a) whether it is possible to reliably reconstruct properties of the original high-dimensional covariance matrix Σ from pooled data, and (b) the design of statistically sound methods for pooling data. To understand the recoverability of Σ from this form of pooled data, we study an idealization of the problem that reveals interesting insights into the problem.

To answer the first question we show that, while in general Σ may not be identifiable under such a sensing design, if Σ has a certain form of structured sparsity, it is indeed possible to recover it reliably via convex optimization. The notion of structured sparsity considered in this paper, which we call "distributed sparsity", is natural in many statistical and data analysis settings [3], [9]. It essentially corresponds to the situations where each variate is correlated with only a small fraction of the other variates leading to a dependency graph that is "bounded degree".

The answer to the second question reveals that random pooling achieves near optimal compression properties for distributed sparse covariance matrices. This notion of compression and its analysis has interesting links to expander graphs and their combinatorial properties.

Our work is related to the notion of sketching in computer science; 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 [13] is one such result, and the idea of sketching has been explored in various contexts [1], [15]. Another related line of work is the literature on compressive sensing in the signal processing community [5] that essentially deals with recovering sparse high-dimensional signals from low-dimensional projections. The idea of using random bipartite graphs and their related expansion properties, which play an important role in our analysis, have also been studied in past work [2], [16].

While much related work exists in the aforementioned communities to deal with sparse high-dimensional data, the problem considered here is technically challenging in our setting because the sensing mechanism in our case is constrained to have a tensor product structure. Many standard techniques fail in this setting. Restricted isometry based approaches [6] 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 poor for tensor product sensing operators [8], [14]. Gaussian width based analysis

approaches [7] fail because the kernel of the sensing matrix is not a uniformly random subspace and hence not amenable to an 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.

The rest of this paper is organized as follows. In Section I-A we set the problem up formally. In Section II we setup the necessary preliminaries and notation. In Section III we state the main result in Theorem 1. In Section IV we state a key technical result pertaining to tensor products of random bipartite graphs in Lemma 1 and sketch its proof. In Section V we prove the main result and in Section VI we validate our theory with computational experiments.

A. The Problem Setup and An Idealization

The covariance sketching problem can be stated as follows. Let $\Sigma \in \mathbb{R}^{p\times p}$ be an unknown positive definite matrix and let $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$ be n independent and identically distributed random vectors drawn from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. Now, suppose that one has access to the $m-$dimensional *sketch vectors* Y_i such that

$$Y_i = AX_i, \;\; i = 1, 2, \ldots, n,$$

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 $\{Y_i\}_{i=1}^n$. The sketching matrices we will focus on later will have randomly-generated binary values, so each element of Y_i is a sum (or "pool") of a random subset of the variates.

Notice that the sample covariance matrix computed using the vectors $\{Y_i\}_{i=1}^n$ satisfies the following.

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

where $\hat{\Sigma}^{(n)} := \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ is the (maximum likelihood) estimate of Σ from the samples $X_1 \ldots, X_n$.

In this paper, to gain a better understanding of the covariance sketching problem, we explore a natural idealization that the above calculation suggests. The question we ask is whether and how one can recover $\Sigma \in \mathbb{R}^{p\times p}$ from the "measurement" $\Sigma_Y := A\Sigma A^T \in \mathbb{R}^{m\times m}$. This idealization exposes the most unique and challenging aspects of the covariance sketching problem. First, observe that the following identity is true.

$$\text{vec}\left(\Sigma_Y\right) = \text{vec}\left(A\Sigma A^T\right) = \left(A\otimes A\right)\text{vec}\left(\Sigma\right). \quad (1)$$

Here vec$(\cdot)$ is the "vectorization" operator that transforms a matrix into a long column vector by stacking the columns of the matrix and $A \otimes A$ stands for the tensor (or Kronecker) product of A with itself, i.e., it is the $\mathbb{R}^{m^2\times p^2}$ matrix

$$\begin{bmatrix} a_{11}A & a_{12}A & \cdots & a_{1p}A \\ a_{21}A & a_{22}A & \cdots & a_{2p}A \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}A & a_{m2}A & \cdots & a_{mp}A \end{bmatrix} \quad (2)$$

Upon inspecting equation (1), it is clear that one could think of the (idealized) covariance sketching problem as the problem of inverting an underdetermined linear system. That is, we want to recover a vector vec$(\Sigma) \in \mathbb{R}^{p^2}$ from $(A\otimes A)$vec$(\Sigma) \in \mathbb{R}^{m^2}$, where $m < p$. While this is not possible 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 if vec(Σ) is sparse, i.e., has few non-zeros, then the inversion can be done to find vec(Σ) efficiently. As in the case of compressed sensing, it should then be possible to extend the results to the case where we have only have access to "noisy data", i.e., the samples $Y_1, \ldots, Y_n$.

Note that assuming that vec(Σ) has fewer than p non-zeros does not correspond to realistic instances of the covariance estimation problem, since it must be that each variate depends on at least a few other variates. So, we begin by asking if recovery of Σ is possible when vec(Σ) has $\mathcal{O}(p)$ non-zeros. It turns out that the answer is no, in general. Figure 1 shows two matrices with $\mathcal{O}(p)$ non-zeros. Let us first suppose that the non-zero pattern in Σ looks like that of the matrix on the left side, which we will dub as the "arrow" matrix. Further, suppose that an oracle reveals to us the non-zero pattern of Σ beforehand. It is not hard to see that it is still impossible to recover Σ from $A\Sigma A^T$. For instance, if $v \in \ker(A)$, then the matrix $\tilde{\Sigma}$, with v added to the first column of Σ is such that $A\Sigma A^T = A\tilde{\Sigma}A^T$ and $\tilde{\Sigma}$ is also an arrow matrix and hence indistinguishable from Σ. In what follows, we will show that if the sparsity pattern of Σ is more *distributed*, as in the right side of figure 1, then one can recover Σ and do so efficiently. In fact, our analysis will also reveal what that the size of the sketch m needs to be able to perform this task and we will see that this is very close to being optimal.

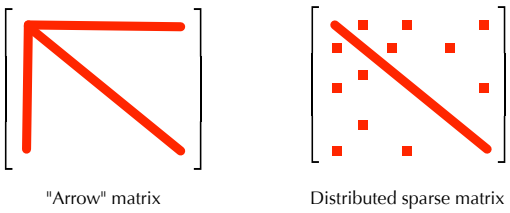

Fig. 1. Two matrices with $\mathcal{O}(p)$ non-zeros. The "arrow" matrix is impossible to recover by covariance sketching while the distributed sparse matrix is.

We finally remark that while Σ is assumed to be distributed sparse, the observed covariance $\Sigma_Y = A\Sigma A^T$ is unstructured (and dense in general). Indeed a pooling mechanism such as the one studied in this paper destroys the independence/correlation structure among the variates, and a central technical challenge is recovering this structure nevertheless.

II. Preliminaries and Notation

For any $p \in \mathbb{N}$, we define $[p] := \{1, 2, \ldots, p\}$. A **graph** $G = ([p], E)$ is defined in the usual sense as an ordered pair with the *vertex set* $[p]$ and the *edge set* E which is a set of 2-element subsets of $[p]$, i.e., $E \subset \binom{[p]}{2}$. We 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 $G = (V, E)$ 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 $i \in [p]$ and $j \in [m]$. A bipartite graph G is said to be $\delta-$**left regular** if for all $i \in [p]$, $\deg(i) = \delta$. 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*.

We let $\mathcal{S}^p$ denote the set of all symmetric matrices in $\mathbb{R}^{p\times p}$. We will be particularly interested in the following subset of $\mathcal{S}^p$.

Definition 2 ($d-$distributed sparse matrices)**.** *We say that a matrix $\Sigma \in \mathcal{S}^p$ is $d-$* **distributed sparse** *if the following hold true.*

1) *For $i = 1, 2, \ldots, p$, $\Sigma_{ii} \neq 0$*
2) *Each row/column of the matrix has no more than d non-zeros.*

We will denote the set of all $d-$distributed sparse matrices in $\mathcal{S}^p$ by $\mathcal{S}^p_d$.

Examples:

- Any diagonal matrix is $d-$distributed sparse for $d \geq 1$.
- The adjacency matrix of a bounded degree graph with maximum degree $d-1$ is $d-$distributed sparse

Let $\Omega \subset [p] \times [p]$ be the support set of a $d-$distributed sparse Σ, i.e.,

$$\Omega := \{(i, j) \in [p] \times [p]\}.$$

Then, Ω has the property that for all $k \in [p]$, (1) $(k, k) \in \Omega$ and, (2) the cardinalities of the sets $\Omega_k := \{(i, j) \in \Omega : i = k\}$ and $\Omega^k := \{(i, j) \in \Omega : j = k\}$ are upper bounded by d. We say that such a set is *$d-$distributed*.

Given a matrix $\Sigma \in \mathbb{R}^{p\times p}$, as mentioned earlier, we write $\text{vec}(\Sigma)$ to denote the $\mathbb{R}^{p^2}$ vector obtained by stacking the columns of Σ. Suppose $\sigma = \text{vec}(\Sigma)$. It will be useful to not forget that σ was actually derived from the matrix Σ and hence we will employ slight abuses of notation as follows:

1) We will say σ is $d-$distributed sparse when we actually mean that the matrix Σ is.
2) We will write (i, j) to denote the index of σ corresponding to Σ_{ij}, i.e.,

$$\sigma_{ij} := \sigma_{(i-1)p+j} = \Sigma_{ij}.$$

Definition 3. *(Tensor graph) Given a bipartite graph $G = ([p], [m], E)$, we define its* **tensor graph** *$G^{\otimes}$ to be the bipartite graph $([p] \times [p], [m] \times [m], E^{\otimes})$ where $E^{\otimes}$ is such that $\{(i, i'), (j, j')\} \in E^{\otimes}$ if and only if $\{i, j\}$ and $\{i', j'\}$ are both in E.*

Notice that if the adjacency matrix of $G = ([p], [m], E)$ is $A' \in \mathbb{R}^{p\times m}$, then the adjacency matrix of $G^{\otimes}$ is $(A \otimes A)' \in \mathbb{R}^{p^2\times m^2}$.

Throughout this paper, unless explicitly stated, the ℓ_1-norm $\|\cdot\|_1$ applied to a matrix X stands for the absolute sum of all the elements of the matrix, i.e.,

$$\|X\|_1 := \|\text{vec}(X)\|_1$$

III. Main Result

In section I-A, we argued why it makes sense to assume that Σ is sparse and in particular, sparse in a distributed manner. Our goal, therefore, is to recover $\Sigma \in \mathcal{S}_d^p$ from the observation matrix $\Sigma_Y := A\Sigma A^T$. A natural, albeit highly impractical, approach to solve this problem is the following: search over all matrices $X \in \mathbb{R}^{p\times p}$ such that AXA^T agrees with Σ_Y and find the sparsest one; the hope being that this would in fact be Σ. Of course, there is no guarantee that this approach might work and worse still, such a search procedure is known to be NP-Hard.

Therefore, we resort to what has now become conventional wisdom again and consider instead the optimization program (P_1), which is a convex relaxation of the naive approach proposed above.

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_1 \\ \text{subject to} \quad & AXA^T = \Sigma_Y. \end{aligned} \tag{P_1}$$

The rest of the paper is devoted to showing that the solution X^* of (P_1) does equal Σ with very high probability. In particular, we prove the following result.

Theorem 1. *Suppose $\Sigma \in \mathcal{S}_d^p$ and $A \in \{0,1\}^{m\times p}$ is chosen as the adjacency matrix of a random bipartite graph with*

$$m = \mathcal{O}(\sqrt{dp}\log^3 p).$$

Then the optimal solution X^* of (P_1) is such that $X^* = \Sigma$ with probability exceeding $1 - p^{-c}$, for some $c > 0$.

We will make the phrase "random bipartite graph" more explicit in section IV, but let us pause here and consider some interesting implications of the above statement:

1) (P_1) does not impose any structural restrictions on X. In other words, even though Σ is assumed to be distributed sparse, this (highly non-convex) constraint need not be factored in to the optimization problem for it to work.
2) Recall that what we measure can be thought of as the $\mathbb{R}^{m^2}$ vector $(A \otimes A)\mathrm{vec}(\Sigma)$. Since $\Sigma \in \mathcal{S}_d^p$, $\mathrm{vec}(\Sigma)$ has $\mathcal{O}(dp)$ non-zeros. Now, even if an oracle were to reveal the exact locations of the non-zeros in $\mathrm{vec}(\Sigma)$, we would require at least $\mathcal{O}(dp)$ measurements to be able to perform the necessary inversion to recover Σ. 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 this sense.
3) While the main result stated in this paper is in the context of covariance estimation, we remark that (P_1) and the proof rely neither on the symmetry of covariance matrices (i.e. that $\Sigma_{ij} = \Sigma_{ji}$), nor on their positive-definiteness ($\Sigma \succ 0$). Indeed our results extend more generally to guarantee that distributed sparse signals of sparsity level $\mathcal{O}(p)$ in p^2 dimensions may be recovered from random tensor product sensing operators of the form $A\otimes A$ (or even $A \otimes B$) where $A \in \mathbb{R}^{\mathcal{O}(\sqrt{p})\times p}$.

Note that if Σ_Y was not observed exactly, but rather available only via an empirical estimate $\hat{\Sigma}_Y$, a natural relaxation to (P_1) would be

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_1 \\ \text{subject to} \quad & \|AXA^T - \hat{\Sigma}_Y\| \le \kappa. \end{aligned} \tag{P_2}$$

While we will not study the properties of (P_2) here, we remark here that if A is the identity matrix, then the optimal solution of this problem corresponds to a thresholded version of $\hat{\Sigma}_Y$ (the threshold parameter being determined by κ) so that one recovers the estimator studied by Bickel and Levina [3].

IV. Uniformly Random Bipartite Graphs and Weak Distributed Expansion

As alluded to earlier, we will choose the sensing matrix A to be the adjacency matrix of a random graph. 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 $i \in [p]$ there exist δ vertices $j_1, j_2, \ldots, j_\delta$ chosen uniformly at random (without replacement) from $[m]$ such that $\{\{i,j_k\}\}_{k=1}^{\delta} \subset E$.*

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

In past work [2], [16], the authors show that a random graph generated as above has the *vertex expansion* property, i.e., for all sets S such that $|S| \le k$, 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. Unfortunately, it turns out that $G^{\otimes}$ does not have this property.

However, we prove that if $A \in \mathbb{R}^{m\times p}$ is picked as in Definition 4, then the tensor graph corresponding to $A \otimes A$, $G^{\otimes} = ([p]\times[p],[m]\times[m],E^{\otimes})$, 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 (P_1) recovers Σ with high probability. We follow up the statement of the lemma with a proof sketch which omits rigorous concentration arguments which can be reconstructed from our exposition.

Lemma 1. *Suppose that $G = ([p], [m], E)$ is a uniformly random $\delta-$left regular bipartite graph with $\delta = \mathcal{O}(\log^3 p)$ and $m = \mathcal{O}(\sqrt{dp}\log^3 p)$. Let Ω be a fixed d-distributed subset of $[p] \times [p]$. Then there exists an $\epsilon \in \left(0, \frac{1}{4}\right)$ such that $G^{\otimes}$ 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* $(i,i') \in \Omega$, $|N(i,i') \cap N(\Omega \setminus (i,i'))| \leq \epsilon\delta^2$.

Proof: Part 1. Since Ω is $d-$distributed, the "diagonal" set $\mathcal{D} := \{(1,1), \ldots, (p,p)\}$ is a subset of Ω. Now, notice that for $j \neq j' \in [m], i \in [p]$,

$$\mathbb{P}\left[(j,j') \in N((i,i))\right] = \frac{\delta(\delta-1)}{m(m-1)} \tag{3}$$

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 have

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

Where the last step follows if p is large enough and m is chosen as prescribed. To complete the proof, it suffices to show that the random quantity $|N(\Omega)|$ does not deviate too much from its mean. While we will not prove this rigorously here, we will sketch the argument. As a first step towards this, we define the random variables $\chi_{jj'} := \mathbf{1}_{\{(j,j') \in N(\mathcal{D})\}}$ and note that

$$|N(\mathcal{D})| = \sum_{jj' \in [m]\times[m]} \chi_{jj'}.$$

Now, notice that each term in the above sum is dependent on no more than $2m-1$ terms. Therefore, using techniques similar to [17], [20], one can bound the deviation of this sum from its expected value.

Part 2. To prove the second part, we note that the probability that a random edge emanating from the vertex $(i,i') \in [p]^2$ 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, if m and δ are chosen as prescribed, the desired expression holds true, in expectation. To show that this quantity concentrates, we again 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, one needs to bound $|N(\Omega)|$ from below and since Ω is $d-$distributed sparse, part 1 of this lemma allows us to do precisely this. Using this, one can show that the result holds with probability exceeding $1 - p^{-c}$ if m and δ are chosen as prescribed.

The proof of part 3 follows along the same lines as above after carefully accounting for the possibility that $i = i'$. ∎

V. Proof of Theorem 1

In this section, we will prove the main theorem. To reduce clutter in our presentation, we will 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 $[p]^2$.

We will consider an arbitrary ordering $\prec$ 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 $\prec$ and so on. Therefore, 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 $\prec$ and the set $E_2^{\otimes} := E^{\otimes} - E_1^{\otimes}$. (We note that a similar proof technique was adopted in Berinde et al. [2]).

Proposition 1 (L_1-RIP)**.** *Suppose $X \in \mathbb{R}^{p\times p}$ is $d-$distributed sparse and is A is the adjacency matrix of a random bipartite $\delta-$left regular graph. Then there exists an $\epsilon > 0$ such that*

$$(1-2\epsilon)\delta^2 \|X\|_1 \leq \|AXA'\|_1 \leq \delta^2 \|X\|_1, \quad (4)$$

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 $X \in \mathbb{R}^{p\times p}$,*

$$\begin{aligned}&\left\|AXA^T\right\|_1 \\ &\geq \delta^2 \|X\|_1 - 2 \sum_{jj'\in[m]^2} \sum_{ii'\in[p]^2} \mathbf{1}_{\{ii',jj'\}\in E_2^{\otimes}} |X_{ii'}|\end{aligned} \quad (5)$$

Proof: In what follows, we will denote the indicator function $\mathbf{1}_{\{ii',jj'\}\in A}$ by $\mathbf{1}_A^{\{ii'jj'\}}$. We begin by observing that

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

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

Now, to complete the proof of Proposition 1, we need to bound the sum in the LHS of (5). Notice that

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

where $r_{ii'}$ is the number of collisions of edges emanating from ii' with all the previous edges as per the ordering $\prec$. 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,

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

This concludes the proof. ∎

Proposition 2 (Nullspace Property)**.** *Suppose $A \in \{0,1\}^{m\times p}$ is the adjacency matrix of a random bipartite $\delta-$left regular graph with $\delta = \mathcal{O}(\log^3 p)$ and $m = \mathcal{O}(\sqrt{dp}\log^3 p)$. Let $X \in \mathbb{S}^p$ be such that $AXA' = 0$ and suppose that $\Omega \subset [p]\times[p]$ is a $d-$distributed set, then with probability exceeding $1-p^{-c}$.*

$$\|X_\Omega\|_1 \leq \frac{\epsilon}{1-2\epsilon} \|X\|_1. \quad (6)$$

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

Proof: Let X be any symmetric matrix such that $AXA^T = 0$. Let $x = \text{vec(X)}$ and note that $\text{vec}\big(AXA^T\big) = (A\otimes A)\,x = 0$. Let Ω be a d-distributed set. 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 abuse notation and use x_Ω to denote $\text{vec}(X_\Omega)$. Note that the following chain of inequalities is true:

$$\begin{aligned}0 &= \left\|(A\otimes A)_{N(\Omega)} x\right\|_1 \\ &= \left\|(A\otimes A)_{N(\Omega)}(x_\Omega + x_{\Omega^c})\right\|_1 \\ &\geq \left\|(A\otimes A)_{N(\Omega)} x_\Omega\right\|_1 - \left\|(A\otimes A)_{N(\Omega)} x_{\Omega^c}\right\|_1 \\ &= \|(A\otimes A) x_\Omega\|_1 - \left\|(A\otimes A)_{N(\Omega)} x_{\Omega^c}\right\|_1 \\ &\overset{(a)}{\geq} (1-2\epsilon)\delta^2 \|X_\Omega\|_1 - \left\|(A\otimes A)_{N(\Omega)} x_{\Omega^c}\right\|_1,\end{aligned}$$

where (a) follows from Proposition 1. Resuming the

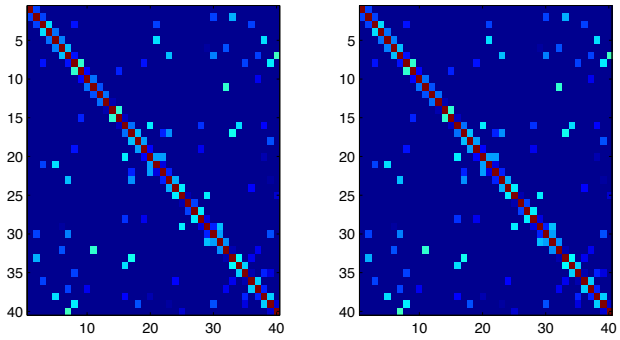

Fig. 2. The matrix on the left is a 40×40 sparse covariance matrix and the matrix on the right is a perfect reconstruction with with $m = 21$.

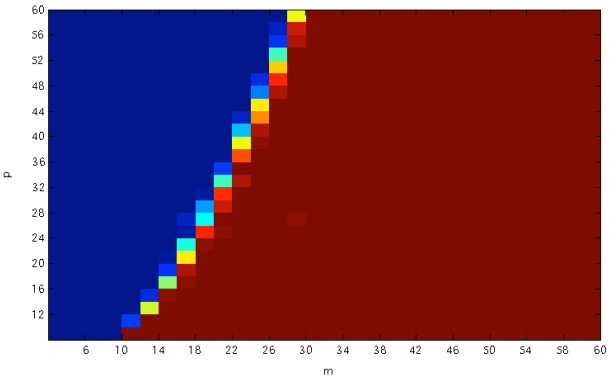

Fig. 3. Phase transition plot. The (i, j)-th pixel shows (an approximation) to the probability of success of the optimization problem (P_1) in recovering a distributed sparse $\Sigma \in \mathbb{R}^{i\times i}$ with sketch-size j.

chain of inequalities, we have:

$$\begin{aligned}
0 &\geq (1-2\epsilon)\delta^2 \left\|X_\Omega\right\|_1 - \sum_{ii'\in\Omega^c} \left\|(A\otimes A)_{N(\Omega)} x_{\{ii'\}}\right\|_1 \\
&\geq (1-2\epsilon)\delta^2 \left\|X_\Omega\right\|_1 - \sum_{\substack{ii':(ii',jj')\in E^\otimes,\\ jj'\in N(\Omega)}} |X_{ii'}| \\
&\geq (1-2\epsilon)\delta^2 \left\|X_\Omega\right\|_1 - \sum_{ii'\in\Omega^c} |E^\otimes(ii' : N(\Omega))|\ |X_{ii'}| \\
&\overset{(b)}{\geq} (1-2\epsilon)\delta^2 \left\|X_\Omega\right\|_1 - \sum_{ii'\in\Omega^c} \epsilon\delta^2\ |X_{ii'}| \\
&\geq (1-2\epsilon)\delta^2 \left\|X_\Omega\right\|_1 - \epsilon\delta^2 \left\|X\right\|_1,
\end{aligned}$$

where (b) follows from the second part of Lemma 1. Rearranging, we get the required result. ∎

Proof of Main Theorem: Let Ω be the support of Σ. Notice that Ω is $d-$distributed. Suppose that there exists an $\tilde{X}$ such that $A\tilde{X}A^T = \Sigma_Y$. Observe that $A\left(\tilde{X} - \Sigma\right)A^T = 0$. Now, consider

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

where (a) follows from proposition 2, and the last line follows from the fact that $\epsilon < \frac{1}{4}$, again from proposition 2. Therefore, the unique solution of (P_1) is Σ with probability exceeding $1 - p^{-c}$, for some $c > 0$. ∎

VI. Experiments

We demonstrate the validity of our theory with some preliminary experiments in this section. Figure 2 shows a 40×40 distributed sparse matrix on the left side. The matrix on the right is a perfect reconstruction using only $m = 21$ samples (i.e., each sketch is only $\sim 50\%$ of the length of the sample vectors). Figure 3 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 $p = i$ and $m = j$. The value at the coordinate (i, j) was generated as follows. A random 4-distributed sparse $\Sigma \in \mathbb{R}^{i\times i}$ 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 [12], [11]. The solution X^* was compared to Σ in the $\|\cdot\|_\infty$ norm (upto numerical precision errors). This was repeated 38 times and the average number of successes was reported in the (i, j)-th spot. In the figure, the deep blue region denotes success during each trial and the deep red region denotes failure in every single trial and there is a sharp *phase transition* between successes and failures. And 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 log factors).

To conclude, 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 set of all X such that $\left\|AXA^T - \hat{\Sigma}_Y^{(n)}\right\|_2 \leq \kappa$ instead. The parameter κ was picked by cross-validation.

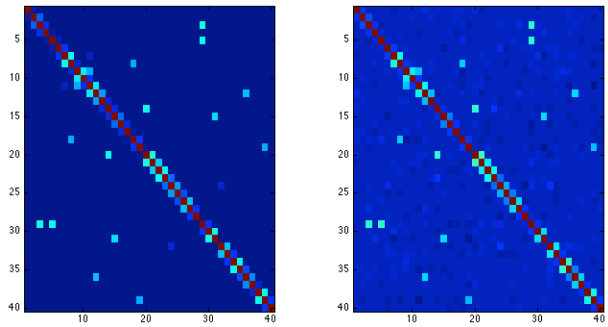

Fig. 4. The matrix on the left is a 40×40 distributed sparse matrix. The matrix on the right was reconstructed using $n = 2100$ samples and with sketches of size $m = 21$.

Figure 4 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.

REFERENCES

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

[2] 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.

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

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

[5] 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.

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

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

[8] 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.

[9] M. Costanzo et al. The genetic landscape of a cell. *Science (New York, N.Y.)*, 327(5964):425–431, January 2010.

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

[11] 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.

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

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

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

[15] 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.

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

[17] 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.

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

[19] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the LASSO. *Annals of Statistics*, 34:1436–1462, 2006.

[20] 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.

[21] 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.