Finding a Large Submatrix of a Gaussian Random Matrix

David Gamarnik* Quan Li†

Abstract

We consider the problem of finding a $k \times k$ submatrix of an $n \times n$ matrix with i.i.d. standard Gaussian entries, which has a large average entry. It was shown in [BDN12] using non-constructive methods that the largest average value of a $k \times k$ submatrix is $2(1 + o(1))\sqrt{\log n/k}$ with high probability (w.h.p.) when $k = O(\log n/ \log\log n)$. In the same paper an evidence was provided that a natural greedy algorithm called Largest Average Submatrix ($\mathcal{LAS}$) for a constant k should produce a matrix with average entry at most $(1+o(1))\sqrt{2\log n/k}$, namely approximately $\sqrt{2}$ smaller, though no formal proof of this fact was provided.

In this paper we show that the matrix produced by the $\mathcal{LAS}$ algorithm is indeed $(1+o(1))\sqrt{2\log n/k}$ w.h.p. when k is constant and n grows. Then by drawing an analogy with the problem of finding cliques in random graphs, we propose a simple greedy algorithm which produces a $k \times k$ matrix with asymptotically the same average value $(1+o(1))\sqrt{2\log n/k}$ w.h.p., for $k = o(\log n)$. Since the greedy algorithm is the best known algorithm for finding cliques in random graphs, it is tempting to believe that beating the factor $\sqrt{2}$ performance gap suffered by both algorithms might be very challenging. Surprisingly, we construct a very simple algorithm which produces a $k \times k$ matrix with average value $(1 + o_k(1) + o(1))(4/3)\sqrt{2\log n/k}$ for $k = o((\log n)^{1.5})$, that is, with asymptotic factor $4/3$ when k grows.

To get an insight into the algorithmic hardness of this problem, and motivated by methods originating in the theory of spin glasses, we conduct the so-called expected overlap analysis of matrices with average value asymptotically $(1 + o(1))\alpha\sqrt{2\log n/k}$ for a fixed value $\alpha \in [1, \sqrt{2}]$. The overlap corresponds to the number of common rows and common columns for pairs of matrices achieving this value (see the paper for details). We discover numerically an intriguing phase transition at $\alpha^* \triangleq 5\sqrt{2}/(3\sqrt{3}) \approx 1.3608.. \in [4/3, \sqrt{2}]$: when $\alpha < \alpha^*$ the space of overlaps is a continuous subset of $[0, 1]^2$, whereas $\alpha = \alpha^*$ marks the onset of discontinuity, and as a result the model exhibits the *Overlap Gap Property (OGP)* when $\alpha > \alpha^*$, appropriately defined. We conjecture that the *OGP* observed for $\alpha > \alpha^*$ also marks the onset of the algorithmic hardness - no polynomial time algorithm exists for finding matrices with average value at least $(1 + o(1))\alpha\sqrt{2\log n/k}$, when $\alpha > \alpha^*$ and k is a mildly growing function of n.

1 Introduction

We consider the algorithmic problem of finding a submatrix of a given random matrix such that the average value of the submatrix is appropriately large. Specifically, consider an $n \times n$ matrix $\mathbf{C}^n$ with i.i.d. standard Gaussian entries. Given $k \leq n$, the goal is to find algorithmically a $k \times k$ submatrix $\mathbf{A}$ of $\mathbf{C}^n$ (not necessarily principal) with average entry as large as possible. The problem has motivations in several areas, including biomedicine, genomics and social networks [SWPN09],[MO04],[For10]. The search of such matrices is called "bi-clustering" [MO04]. The problem of finding asymptotically the largest average entry of $k \times k$ submatrices of $\mathbf{C}^n$ was recently studied by Bhamidi et.al. [BDN12] (see also [SN13] for a related study) and questions arising in this paper constitute the motivation for our work. It was shown in [BDN12] using non-constructive methods that the largest achievable average entry of a $k \times k$ submatrix of $\mathbf{C}^n$ is asymptotically with high probability (w.h.p.) $(1 + o(1))2\sqrt{\log n/k}$

*MIT; e-mail: gamarnik@mit.edu.Research supported by the NSF grants CMMI-1335155.

†MIT; e-mail: quanli@mit.edu

when n grows and $k = O(\log n/\log\log n)$ (a more refined distributional result is obtained). Here $o(1)$ denotes a function converging to zero as $n \to \infty$. Furthermore, the authors consider the asymptotic value and the number of so-called locally maximum matrices. A $k \times k$ matrix $\mathbf{A}$ is locally maximal if every $k \times k$ matrix of $\mathbf{C}^n$ with the same set of rows as $\mathbf{A}$ has a smaller average value than that of $\mathbf{A}$, and every $k \times k$ matrix of $\mathbf{C}^n$ with the same set of columns as $\mathbf{A}$ has a smaller average value than that of $\mathbf{A}$. Such local maxima are natural objects arising as terminal matrices produced by a simple iterative procedure called Large Average Submatrix ($\mathcal{LAS}$), designed for finding a matrix with a large average entry. $\mathcal{LAS}$ proceeds by starting with an arbitrary $k \times k$ submatrix $\mathbf{A}_0$ and finding a matrix $\mathbf{A}_1$ sharing the same set of rows with $\mathbf{A}_0$ which has the largest average value. The procedure is then repeated for $\mathbf{A}_1$ by searching through columns of $\mathbf{A}_1$ and identifying the best matrix $\mathbf{A}_2$. The iterations proceed while possible and at the end some locally maximum matrix $\mathbf{A}_{\mathcal{LAS}}$ is produced as the output. The authors show that when k is constant, the majority of locally maximum matrices of $\mathbf{C}^n$ have the asymptotic value $(1+o(1))\sqrt{2\log n/k}$ w.h.p. as n grows, thus factor $\sqrt{2}$ smaller than the global optimum. Motivated by this finding, the authors suggest that the outcome of the $\mathcal{LAS}$ algorithm should be also factor $\sqrt{2}$ smaller than the global optimum, however one cannot deduce this from the result of [BDN12] since it is not ruled out that $\mathcal{LAS}$ is clever enough to find a "rare" locally maximum matrix with a significantly larger average value than $\sqrt{2\log n/k}$.

The main result of this paper is the confirmation of this conjecture for the case of constant k: the $\mathcal{LAS}$ algorithm produces a matrix with asymptotic average value $(1+o(1))\sqrt{2\log n/k}$ w.h.p. We further establish that the number of iterations of the $\mathcal{LAS}$ algorithm is stochastically bounded as n grows. The proof of this result is fairly involved and proceeds by a careful conditioning argument. In particular, we show that for fixed r, conditioned on the event that $\mathcal{LAS}$ succeeded in iterating at least r steps, the probability distribution of the "new best matrix" which will be used in constructing the matrix for the next iteration is very close to the largest matrix in the $k \times n$ strip of $\mathbf{C}^n$, and which is known to have asymptotic average value of $\sqrt{2\log n/k}$ due to result in [BDN12]. Then we show that the matrix produced in step r and the best matrix in the $k \times n$ strip among the unseen entries are asymptotically independent. Using this we show that given that $\mathcal{LAS}$ proceeded with r steps the likelihood it proceeds with the next $r + 2k + 4$ steps is at most some value $\psi < 1$ which is bounded away from 1 as n grows. As a result the number of steps of $\mathcal{LAS}$ is upper bounded by a geometrically decaying function and thus is stochastically bounded as n grows. We use this as a key result in computing the average value produced by $\mathcal{LAS}$, again relying on the asymptotic independence and the average value of the $k \times n$ strip dominant submatrix.

As it was observed already in [BDN12], the factor $\sqrt{2}$ gap between the global optimum and the performance of $\mathcal{LAS}$ is reminiscent of a similar gap arising in studying of largest cliques of random graphs. Arguably, one of the oldest algorithmic open problems in the field of random graph is the problem of finding a largest clique (a fully connected subgraph) of a random Erdös-Rényi graph $\mathbb{G}(n,p)$, when p is at least $n^{-1+\delta}$ for some positive constant δ. It is known that the value is asymptotically $2\log n/(-\log p)$ and a simple greedy procedure produces a clique with size $\log n/(-\log p)$, namely factor 2 smaller than the global optimum. A similar result holds for the bi-partite Erdös-Rényi graph: the largest clique is asymptotically $2\log n/(-\log p)$ and the greedy algorithm produces a (bi-partite) clique of size asymptotically $\log n/(-\log p)$. Karp in his 1976 paper [Kar76] challenged to find a better algorithm leading to a clique with size say $(1+\epsilon)\log n/(-\log p)$ and this problem remains open. The factor $\sqrt{2}$ appearing in our context is then arguably an analogue of the factor 2 arising in the context of the clique problem in $\mathbb{G}(n,p)$. In order to further investigate the possible connection between the two problems, we propose the following simple algorithm for finding a submatrix of $\mathbf{C}^n$ with a large average entry. Fix a positive threshold θ and consider the random $0,1$ matrix $\mathbf{C}^n_\theta$ obtained by thresholding each Gaussian entry of $\mathbf{C}^n$ at θ. Clearly $\mathbf{C}^n_\theta$ is an adjacency matrix of a bi-partite Erdös-Rényi graph $\mathbb{G}(n,p_\theta)$, where $p_\theta = \mathbb{P}(Z > \theta)$ and Z is a standard Gaussian random variable. Observe that any

$k \times k$ clique of $\mathbb{G}(n, p_\theta)$ corresponds to a $k \times k$ submatrix of $\mathbf{C}^n$ with *each* entry at least θ. Thus any polynomial time algorithm for finding a clique in $\mathbb{G}(n, p_\theta)$ which results in a $k \times k$ clique w.h.p. immediately gives a matrix with average value at least θ w.h.p. Consider the greedy algorithm and adjust θ so that the size of the clique is at least k on each side. Reverse engineering θ from such k, one can find that $\theta \approx \sqrt{2 \log n/k}$ with $p \approx \exp(-\theta^2/2) = n^{-\frac{1}{k}}$ (see the next section for a simple derivation of this fact). Namely, both $\mathcal{LAS}$ and the greedy algorithm have the same asymptotic power! (Note, however, that this analysis extends beyond the $k = O(1)$ unlike our analysis of the $\mathcal{LAS}$ algorithm).

In light of these connections with studying cliques in random graphs and the apparent failure to bridge the factor 2 gaps for cliques, one might suspect that $\sqrt{2}$ is equally challenging to beat for the maximum submatrix problem. Perhaps surprisingly, we establish that this is not the case and construct a very simple algorithm, both in terms of analysis and implementation, which construct a submatrix with average value asymptotically $(1 + o_k(1) + o(1))(4/3)\sqrt{2 \log n/k}$ for $k = o((\log n)^{1.5})$. Here $o_k(1)$ denotes a function decaying to zero as k increases. That is, the asymptotic factor $4/3$ is valid for growing k. The algorithm proceeds by starting with one entry and iteratively building a sequence of $r \times r$ and $r \times (r + 1)$ matrices for $r = 1, \ldots, k$ in a simple greedy fashion. We call this algorithm Incremental Greedy Procedure ($\mathcal{IGP}$), referring to the incremental increase of the matrix size. No immediate simple modifications of $\mathcal{IGP}$ led to the improvement of the $4/3$ factor, unfortunately.

The discussion above raises the following question: where is the true algorithmic hardness threshold value for the maximum submatrix problem if such exists? Short of proving some formal hardness of this problem, which seems out of reach for the currently known techniques both for this problem and the clique problem for $\mathbb{G}(n, p)$, we propose an approach which indirectly suggests the hardness regime for this problem, and this is our last contribution. Specifically, our last contribution is the conjecture for this value based on the *Overlap Gap Property* (OGP) which originates in the theory of spin glasses and which we adopt here in the context of our problem in the following way. We fix $\alpha \in (1, \sqrt{2})$ and let $\mathcal{L}(\alpha)$ denote the set of matrices with average value asymptotically $\alpha\sqrt{2 \log n/k}$. Thus α conveniently parametrizes the range between the achievable value on the one hand, namely $\alpha = 1$ for $\mathcal{LAS}$ and greedy algorithms, $\alpha = 4/3$ for the $\mathcal{IGP}$, and $\alpha = \sqrt{2}$ for the global optimum on the other hand. For every pair of matrices $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{L}(\alpha)$ with row sets I_1, I_2 and column sets J_1, J_2 respectively, let $x(\mathbf{A}_1, \mathbf{A}_2) = |I_1 \cap I_2|/k, y(\mathbf{A}_1, \mathbf{A}_2) = |J_1 \cap J_2|/k$. Namely x and y are the normalized counts of the common rows and common columns for the two matrices. For every $(x, y) \in [0, 1]^2$ we consider the expected number of pairs $\mathbf{A}_1, \mathbf{A}_2$ such that $x(\mathbf{A}_1, \mathbf{A}_2) \approx x, y(\mathbf{A}_1, \mathbf{A}_2) \approx y$, in some appropriate sense to be made precise. We compute this expectation asymptotically. We define $R(x, y) = 0$ if such an expectation converges to zero as $n \to \infty$ and $= 1$ otherwise. Thus the set $\mathcal{R}(\alpha) \triangleq \{(x, y) : R(x, y) = 1\}$ describes the set of achievable in expecation overlaps of pairs of matrices with average value $\alpha\sqrt{2 \log n/k}$. At $\alpha^* \triangleq 5\sqrt{2}/(3\sqrt{3}) \approx 1.3608..$ we observe an interesting phase transition – the set $\mathcal{R}(\alpha)$ is connected for $\alpha < \alpha^*$, and is disconnected for $\alpha > \alpha^*$ (see Figures 6). Namely, for $\alpha > \alpha^*$ the model exhibits the OGP. Namely, the overlaps of two matrices belong to one of the two disconnected regions.

Motivated by this observation, we conjecture that the problem of finding a matrix with the corresponding value $\alpha > \alpha^*$ is not-polynomially solvable when k grows. In fact, by considering multi-overlaps instead of pairwise overlaps, (which we intend to research in future), we conjecture that this hardness threshold might be even lower than α^*. The link between OGP and algorithmic hardness has been suggested and partially established in the context of sparse random constraint satisfaction problems, such as random K-SAT problem, coloring of sparse Erdös-Rényi problem and the problem of finding a largest independent set of a sparse Erdös-Rényi graph problem [ACORT11],[ACO08],[COE11], [GS14a],[RV14],[GS14b],[Mon15]. A more recent link between OGP and tractable algorithms was recently established in the context of sparse high dimensional linear regression problem [GZ17]. Many of these problems exhibit an apparent gap between the best existential values and the best values found by known algorithms, very similar in spirit to the gaps $2, \sqrt{2}$ etc. discussed above in our context. For

example, the largest independent set of a random d-regular graph normalized by the number of nodes is known to be asymptotically $2\log d/d$ as d increases, while the best algorithm can produce sets of size only $\log d/d$ again as d increases. As shown in [COE11],[GS14a] and [RV14] the threshold $\log d/d$ marks the onset of a certain version of OGP. Furthermore, [COE11],[GS14a] show that OGP is the bottleneck for a certain class of algorithms, namely local algorithms (appropriately defined). Roughly speaking, an algorithm is local if for each node of a graph, the decisions resulting in a construction of a solution are conducted by taking account only small neighborhood of a node. Then by a careful coupling construction, it can be shown that solutions resulting from local algorithm exhibit a certain continuity property, which is inconsistent with the OGP. Thus such algorithms cannot overcome the OGP barrier. For the case of sparse linear regression problem in [GZ17], it is shown that local improvement type algorithms also exhibit a continuity property, also inconsistent with the OGP. It is further shown that the onset of the OGP occurs asymptotically around the threshold below which the celebrated compressive sensing methods such as LASSO and Dantzig Selector fail.

A key step observed in [RV14] is that the threshold for multioverlap version of the OGP, namely considering m-tuples of solutions as opposed to pairs of solutions as we do in this paper, lowers the OGP phase transition point. The multioverlap version of OGP was also a key step in [GS14b] in the context of random Not-All-Equal-K-SAT (NAE-K-SAT) problem which also exhibits a marked gap between the regime where the existence of a feasible solution is known and the regime where such a solution can be found by known polynomial time algorithms. We conjecture that the OGP threshold established in this paper can also be lowered by considering overlap of $m \geq 3$ matrices.

The OGP for largest submatrix problem thus adds to the growing class of optimization problems with random input which exhibit a significant gap between the value achieved by an optimal solution and values achieved by currently known tractable algorithmic methods, and where the gap is evidenced by the phase transition associated with the OGP.

The remainder of the paper is structured as follows. In the next section we formally state our four main results: the result regarding the performance of $\mathcal{LAS}$, the result regarding the performance of the greedy algorithm via reduction to random bi-partite graphs, the result regarding the performance of $\mathcal{IGP}$, and finally the result regarding the OGP. The same section provides a short proof for the result regarding the greedy algorithm. Section 3 is devoted to the proof of the result regarding the performance of $\mathcal{IGP}$. Section 4 is devoted to the proof of our result regarding OGP, and Section 5 (which is the most technically involved part of the paper) is devoted to the proof of the result regarding the performance of the $\mathcal{LAS}$ algorithm. We conclude in Section 6 with some open questions.

We close this section with some notational convention. We use standard notations $o(\cdot)$, $O(\cdot)$ and $\Theta(\cdot)$ with respect to $n \to \infty$. $o_k(1)$ denotes a function $f(k)$ satisfying $\lim_{k\to\infty} f(k) = 0$. Given a positive integer n, $[n]$ stands for the set of integers $1, \ldots, n$. Given a matrix A, A^T denotes its transpose. $\Rightarrow$ denotes weak convergence. $\stackrel{d}{=}$ denotes equality in distribution. A complement of event $\mathcal{A}$ is denoted by $\mathcal{A}^c$. For two events $\mathcal{A}$ and $\mathcal{B}$ we write $\mathcal{A} \cap \mathcal{B}$ and $\mathcal{A} \cup \mathcal{B}$ for the intersection (conjunction) and the union (disjunction) of the two events, respectively. When conditioning on the event $\mathcal{A} \cap \mathcal{B}$ we will often write $\mathbb{P}(\cdot|\mathcal{A}, \mathcal{B})$ in place of $\mathbb{P}(\cdot|\mathcal{A} \cap \mathcal{B})$. For non-negative integers $b_1, b_2, \cdots, b_l$ such that $\sum_{i=1}^{l} b_i = n$, the multinomial coefficient is

$$\binom{n}{b_1, b_2, \ldots, b_l} = \frac{n!}{b_1! b_2! \ldots b_l!}.$$

Let $\Phi(u)$ be the cumulative distribution function of the standard normal random variable. When u is large, the function $1 - \Phi(u)$ can be approximated by

$$\frac{1 - 2u^{-2}}{u\sqrt{2\pi}} \exp(-u^2/2) \leq 1 - \Phi(u) \leq \frac{1}{u\sqrt{2\pi}} \exp(-u^2/2). \tag{1}$$

2 Main Results

In this section we formally describe the algorithms we analyze in this paper and state our main results. Given an $n \times n$ matrix A and subsets $I \subset [n], J \subset [n]$ we denote by $A_{I,J}$ the submatrix of A indexed by rows I and columns J. When I consist of a single row i, we use $A_{i,J}$ in place of a more proper notation $A_{\{i\},J}$. Given any $m_1 \times m_2$ matrix B, let $\mathrm{Ave}(B) \triangleq \frac{1}{m_1 m_2} \sum_{i,j} B_{i,j}$ denote the average value of the entries of B.

Let $\mathbf{C} = (\mathbf{C}_{ij}, i, j \geq 1)$ denote an infinite two dimensional array of independent standard normal random variables. Denote by $\mathbf{C}^{n \times m}$ the $n \times m$ upper left corner of $\mathbf{C}$. If $n = m$, we use $\mathbf{C}^n$ instead.

The Large Average Submatrix algorithm is defined as follows.

Large Average Submatrix algorithm ($\mathcal{LAS}$)

Input: *An $n \times n$ matrix A and a fixed integer $k \geq 1$.*
Initialize: *Select k rows I and k columns J arbitrarily.*
Loop: *(Iterate until no improvement is achieved)*
Find the set $\hat{J} \subset [n], |\hat{J}| = k$ such that $\mathrm{Ave}(A_{I,\hat{J}}) \geq \mathrm{Ave}(A_{I,J'})$ for all $J' \subset [n], |J'| = k$. Break ties arbitrarily.
If $\hat{J} = J$, STOP. Otherwise, set $J = \hat{J}$.
Find the set $\hat{I} \subset [n], |\hat{I}| = k$ such that $\mathrm{Ave}(A_{\hat{I},J}) \geq \mathrm{Ave}(A_{I',J})$ for all $I' \subset [n], |I'| = k$. Break ties arbitrarily.
If $\hat{I} = I$, STOP. Otherwise, Set $I = \hat{I}$.
Output: *$A_{I,J}$.*

Since the entries of $\mathbf{C}^n$ are continuous independent random variables the ties in the $\mathcal{LAS}$ algorithm occur with zero probability. Each step of the $\mathcal{LAS}$ algorithm is easy to perform, since given a fixed set of rows I, finding the corresponding set of columns $\hat{J}$ which leads to the matrix with maximum average entry is easy: simply find k columns corresponding to k largest entry sums. Also the algorithm will stop after finitely many iterations since in each step the matrix sum (and the average) increases and the number of submatrices is finite. In fact a major part of our analysis is to bound the number of steps of $\mathcal{LAS}$. Our convention is that in step zero, the $\mathcal{LAS}$ algorithm sets $I_0 = I = \{1, \ldots, k\}$ and $J_0 = J = \{1, \ldots, k\}$. We denote by $T_{\mathcal{LAS}}$ the number of iterations of the $\mathcal{LAS}$ algorithm applied to the $n \times n$ matrix $\mathbf{C}^n$ with i.i.d. standard normal entries. For concreteness, searching for $\hat{I}$ and $\hat{J}$ are counted as two separate iterations. We denote by $\mathbf{C}^n_r$ the matrix produced by $\mathcal{LAS}$ in step (iteration) r, assuming $T_{\mathcal{LAS}} \geq r$. Thus our goal is to obtain asymptotic values of $\mathrm{Ave}\left(\mathbf{C}_{T_{\mathcal{LAS}}}\right)$, as well as the number of iterations $T_{\mathcal{LAS}}$.

Our first main result concerns the performance of $\mathcal{LAS}$ and stated as follows. Let ω_n denote any positive function satisfying $\omega_n = o(\sqrt{\log n})$ and $\log \log n = O(\omega_n)$.

Theorem 2.1. *Suppose a positive integer k is fixed. For every $\epsilon > 0$ there is a positive integer N which depends on k and ϵ only, such that for all $n \geq N$, $\mathbb{P}(T_{\mathcal{LAS}} \geq N) \leq \epsilon$. Furthermore,*

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \mathrm{Ave}(\mathbf{C}^n_{T_{\mathcal{LAS}}}) - \sqrt{\frac{2 \log n}{k}} \right| \leq \omega_n \right) = 1. \tag{2}$$

Theorem 2.1 states that the average of the $k \times k$ submatrix produced by $\mathcal{LAS}$ converges to the value $(1 + o(1))\sqrt{2 \log n / k}$, and furthermore, the number of iterations is stochastically bounded in n. In fact we will show the existence of a constant $0 < \psi < 1$ which depends on k and ϵ only such that

$\mathbb{P}(T_{\mathcal{LAS}} > t) \leq \psi^t, t \geq 1$ for all large enough n. Namely, $T_{\mathcal{LAS}}$ is bounded by a geometric random variable for all large enough n.

Next we turn to the performance of the greedy algorithm applied to the random graph produced from $\mathbf{C}^n$ by first thresholding it at a certain level θ. Given $\mathbf{C}^n$ let $\mathbb{G}(n, n, p(\theta))$ denote the corresponding $n \times n$ bi-partite graph where the edge $(i, j), i, j \in [n]$ is present if $\mathbf{C}^n_{i,j} > \theta$ and is absent otherwise. The edge probability is then $p(\theta) = \mathbb{P}(Z > \theta)$ where Z is a standard normal random variable. A pair of subsets $I \subset [n], J \subset [n]$ is a clique in $\mathbb{G}(n, n, p(\theta))$ if the edge (i, j) exists for every $i \in I, j \in J$. In this case we write $i \sim j$.

Consider the following simple algorithm for generating a clique in $\mathbb{G}(n, n, p(\theta))$, which we call a greedy algorithm for simplicity. Pick node $i_1 = 1$ on the left part of the graph and let $J_1 = \{j : 1 \sim j\}$. Pick any node $j_1 \in J_1$ and let $I_1 = \{i \in [n] : i \sim j_1\}$. Clearly $i_1 \in I_1$. Pick any node $i_2 \in I_1$ different from i_1 and let $J_2 = \{j \in J_1 : i_2 \sim j\}$. Clearly $j_1 \in J_2$. Pick any $j_2 \in J_2$ different from j_1 and let $I_2 = \{i \in I_1 : i \sim j_2\}$, and so on. Repeat this process for as many steps m as possible ending it on the right-hand side of the graph, so that the number of chosen nodes on the left and the right is the same. The end result I_m, J_m is clearly a clique. It is also immediate that $|I_m| = |J_m| = m$. The corresponding submatrix $\mathbf{C}^n_{I_m,J_m}$ of $\mathbf{C}^n$ indexed by rows I_m and columns J_m has every entry at least θ and therefore $\mathrm{Ave}(\mathbf{C}^n_{I_m,J_m}) \geq \theta$. If we can guarantee that θ is small enough so that m is at least k, we obtain a simple algorithm for producing a $k \times k$ matrix with average entry at least θ. From the theory of random graph it is known (and easy to establish) that w.h.p. our greedy algorithm produces a clique of size $\log n / \log(1/p)$ provided that p is at least $n^{-1+\epsilon}$ for some $\epsilon > 0$. Since we need to produce a $k \times k$ clique we obtain a requirement $\log n / \log(1/p) \geq k$ (provided of course the lower bound $n^{-1+\epsilon}$ holds, which we will verify retroactively), leading to

$$p = \mathbb{P}(Z > \theta) \geq n^{-\frac{1}{k}},$$

and in particular $k \geq 2$ is enough to satisfy the $n^{-1+\epsilon}$ lower bound requirement. Now suppose $k = o(\log n)$ implying $n^{-\frac{1}{k}} = o(1)$. Solving for θ_n defined by

$$\mathbb{P}(Z > \theta_n) = n^{-\frac{1}{k}},$$

and using the fact

$$\lim_{t \to \infty} t^{-2} \log(Z > t) = -\frac{1}{2},$$

which is an easy consequence of (1), we conclude that

$$\theta_n = (1 + o(1))\sqrt{\frac{2 \log n}{k}},$$

leading the same average value as the $\mathcal{LAS}$ algorithm! The two algorithms have asymptotically the same performance (though the greedy algorithm guarantees a *minimum* value of $(1 + o(1))\sqrt{\frac{2 \log n}{k}}$ as opposed to just the (same) average value). We summarize our finding as follows.

Theorem 2.2. *Setting $\theta_n = (1 + o(1))\sqrt{\frac{2 \log n}{k}}$, the greedy algorithm w.h.p. produces a $k \times k$ sub-matrix with minimum value θ_n for $k = o(\log n)$.*

Next we turn to an improved algorithm for finding a $k \times k$ submatrix with large average entry, which we call Incremental Greedy Procedure ($\mathcal{IGP}$) and which achieves the $(1 + o_k(1))(4/3)\sqrt{2 \log n / k}$ asymptotics. We first provide a heuristic idea behind the algorithm which ignores certain dependencies,

and then provide the appropriate fix for dealing with the dependency issue. The algorithm is described informally as follows. Fix an arbitrary $i_1 \in [n]$ and in the corresponding row $\mathbf{C}^n_{i_1,[n]}$ find the largest element $\mathbf{C}^n_{i_1,j_1}$. This term is asymptotically $\sqrt{2\log n}$ as the largest of n i.i.d. standard normal random variables (see (21) in Section 5). Then find the largest element $\mathbf{C}^n_{i_2,j_1}$ in the column $\mathbf{C}_{[n],j_1}$ other than $\mathbf{C}_{i_1,j_1}$, which asymptotically is also $\sqrt{2\log n}$. Next in the $2\times n$ matrix $\mathbf{C}^n_{\{i_1,i_2\},[n]}$ find a column $j_2 \neq j_1$ such that the sum of the two elements of the column $\mathbf{C}^n_{\{i_1,i_2\},j_2}$ is larger than the sum for all other columns $\mathbf{C}^n_{\{i_1,i_2\},j}$ for all $j \neq j_1$. Ignoring the dependencies, this sum is asymptotically $\sqrt{2}\sqrt{2\log n}$, though the dependence is present here since the original row $\mathbf{C}_{i_1,[n]}$ is a part of this computation. We have created a 2×2 matrix $\left(\mathbf{C}^n_{i,j},\ i=i_1,i_2;\ j=j_1,j_2\right)$. Then we find a row $i_3 \neq i_1,i_2$ such that the sum of the two elements of the row $\mathbf{C}_{i_3,\{j_1,j_2\}}$ is larger than any other such sum of $\mathbf{C}_{i_3,\{j_1,j_2\}}$ for $i \neq i_1,i_2$. Again, ignoring the dependencies, this average is asymptotically $\sqrt{2}\sqrt{2\log n}$. We continue in this fashion, greedily and incrementally expanding the matrix to a larger sizes, creating in alternation $r\times r$ and $(r+1)\times r$ matrices and stop when $r=k$ and we arrive at the $k\times k$ matrix. In each step, ignoring the dependencies, the sum of the elements of the added row and added column is $\sqrt{r}\sqrt{2\log n}$ when the number of elements in the row and in the column is r, again ignoring the dependency. Thus we expect the total asymptotic size of the final matrix to be

$$2\sum_{1\le r\le k-1}\sqrt{r}\sqrt{2\log n}+\sqrt{k}\sqrt{2\log n}.$$

Approximating $2\sum_{1\le r\le k-1}\sqrt{r}+\sqrt{k}$ by $2\int_1^k\sqrt{x}dx \approx 4k^{3/2}/3$ for growing k and then dividing the expression above by k^2, we obtain the required asymptotics. The flaw in the argument above comes from ignoring the dependencies: when $r\times 1$ row is chosen among the best such rows outside of the already created $r\times r$ matrix, the distribution of this row is dependent on the distribution of this matrix. A simple fix comes from partitioning the entire $n\times n$ matrix into $k\times k$ equal size groups, and only searching for the best $r\times 1$ row within the respective group. The sum of the elements of the r-th added row is then $\sqrt{r}\sqrt{2\log(n/k)}$ which is asymptotically the same as $\sqrt{r}\sqrt{2\log n}$, provided k is small enough. The independence of entries between the groups is then used to estimate rigorously the performance of the algorithm.

We now formalize the approach and state our main result. The proof or the performance of the algorithm is in Section 3. Given $n\in\mathbb{Z}^+$ and $k\in[n]$, divide the set $[n]$ into $k+1$ disjoint subsets, where the first k subsets are

$$P^n_i=\{(i-1)\lfloor n/k\rfloor+1,(i-1)\lfloor n/k\rfloor+2,\dots,i\lfloor n/k\rfloor\},\ \text{for } i=1,2,\dots,k.$$

When n is a multiple of k, the last subset is by convention an empty set. A detailed description of $\mathcal{IGP}$ algorithm is as follows.

$\mathcal{IGP}$ algorithm.

Input: *An $n\times n$ matrix A and a fixed integer $k\ge 1$.*
Initialize: *Select $i_1\in P^n_1$ arbitrarily and set $I=\{i_1\}$, and let $J=\emptyset$.*
Loop: *Proceed until $|I|=|J|=k$*
Find the column $j\in P^n_{|I|}$ such that $\mathrm{Ave}(A_{I,j})\ge\mathrm{Ave}(A_{I,j'})$ for all $j'\in P^n_{|I|}$. Set $J=J\cup\{j\}$.
Find the $i\in P^n_{|I|+1}$ such that $\mathrm{Ave}(A_{i,J})\ge\mathrm{Ave}(A_{i',J})$ for all $i'\in P^n_{|I|+1}$. Set $I=I\cup\{i\}$.
Output: *$A_{I,J}$.*

As shown in Figure 1, the $\mathcal{IGP}$ algorithm at step $2r$ adds a row of r entries (represented by symbol '$\triangle$') with the largest entry sum to the previous $r\times r$ submstrix $\mathbf{C}^{n,2r-1}_{\mathcal{IGP}}$. Similarly, as shown in Figure

2, the $\mathcal{IGP}$ algorithm at step $2r+1$ adds a column of $r+1$ entries (represented by symbol '$\triangle$') with largest entry sum to the previous $(r+1)\times r$ submstrix $\mathbf{C}^{n,2r}_{\mathcal{IGP}}$.

$$\begin{bmatrix} C_{11} & & \cdots & & C_{1n} \\ C_{21} & & r & & C_{2n} \\ & & \begin{bmatrix} * & \cdots & * \\ \vdots & \mathbf{C}^{n,2r}_{\mathcal{IGP}} & \vdots \\ * & \cdots & * \\ \triangle & \cdots & \triangle \end{bmatrix} & & \\ \vdots & r+1 & & & \vdots \\ & & & & \\ C_{n1} & & \cdots & & C_{nn} \end{bmatrix}$$

Figure 1: Step $2r$ of $\mathcal{IGP}$ algorithm

$$\begin{bmatrix} C_{11} & & \cdots & & C_{1n} \\ C_{21} & & r+1 & & C_{2n} \\ & & \begin{bmatrix} * & \cdots & * & \triangle \\ \vdots & \mathbf{C}^{n,2r+1}_{\mathcal{IGP}} & \vdots & \vdots \\ * & \cdots & * & \triangle \end{bmatrix} & & \\ \vdots & r+1 & & & \vdots \\ & & & & \\ C_{n1} & & \cdots & & C_{nn} \end{bmatrix}$$

Figure 2: Step $2r+1$ of $\mathcal{IGP}$ algorithm

Just as for the $\mathcal{LAS}$ algorithm, each step of $\mathcal{IGP}$ algorithm is easy to perform: simply find one column (row) corresponding to the largest entry sum. The algorithm stops after $2k$ steps. We denote by $\mathbf{C}^n_{\mathcal{IGP}}$ the $k\times k$ submatrix produced by $\mathcal{IGP}$ applied to $\mathbf{C}^n$. Our goal is to obtain the asymptotic value of $\mathrm{Ave}(\mathbf{C}^n_{\mathcal{IGP}})$.

Our main result regarding the performance of the $\mathcal{IGP}$ algorithm is as follows.

Theorem 2.3. *Let $f(n)$ be any positive function such that $f(n)=o((\log n)^{1.5})$. Then*

$$\lim_{n\to\infty}\min_{1\le k\le f(n)}\mathbb{P}\left(\left|\mathrm{Ave}(\mathbf{C}^n_{\mathcal{IGP}})-\frac{4}{3}\sqrt{\frac{2\log n}{k}}\right|\le 3\max\left(\frac{1}{k}\sqrt{\frac{\log n}{k}},\frac{\log\log n}{\sqrt{\log n}}\right)\right)=1. \tag{3}$$

The bound on the right hand side is of the order magnitude $O(\sqrt{\log n})$ when k is constant and is $o(\sqrt{\log n/k})$ when k is a growing function of n and $k=o((\log n)^{1.5})$. The asymptotics $(1+o_k(1)+o(1))\frac{4}{3}\sqrt{\frac{2\log n}{k}}$ corresponds to the latter case.

Next we turn to the discussion of the Overlap Gap Property (OGP). Fix $\alpha\in(1,\sqrt{2})$, real values $0<y_1,y_2<1$ and $\delta\in(0,\alpha)$. Let $\mathcal{O}_k(\alpha,y_1,y_2,\delta)$ denote the set of pairs of $k\times k$ submatrices $\mathbf{C}^n_{I_1,J_1},\mathbf{C}^n_{I_2,J_2}$ with average value in the interval $[(\alpha-\delta)\sqrt{2\log n/k},(\alpha+\delta)\sqrt{2\log n/k}]$ and which satisfy $|I_1\cap I_2|/k\in(y_1-\delta,y_1+\delta),|J_1\cap J_2|/k\in(y_2-\delta,y_2+\delta)$. Namely, $\mathcal{O}_k(\alpha,y_1,y_2,\delta)$ is the set of pairs of $k\times k$

matrices with average value approximately $\alpha\sqrt{2\log n/k}$ and which share approximately $y_1 k$ rows and $y_2 k$ columns. Let

$$f(\alpha, y_1, y_2) \triangleq 4 - y_1 - y_2 - \frac{2}{1+y_1 y_2}\alpha^2. \tag{4}$$

The next result says that the expected cardinality of the set $\mathcal{O}_k(\alpha, y_1, y_2, \delta)$ is approximately $n^{kf(\alpha,y_1,y_2)}$ when $f(\alpha, y_1, y_2)$ is positive, and, on the other hand, $\mathcal{O}_k(\alpha, y_1, y_2, \delta)$ is empty with high probability when $f(\alpha, y_1, y_2)$ is negative.

Theorem 2.4. *Fix $\alpha \in (1, \sqrt{2})$. For every $\epsilon > 0$, $c > 0$ and $y_1, y_2 \in (0,1)$, there exists $\delta_0 \in (0, \alpha)$ and $n_0 > 0$ such that for all $n \geq n_0$, $k \leq c \log n$ and $\delta \in (0, \delta_0)$*

$$\left| \frac{\log \mathbb{E}\left[\left|\mathcal{O}_k(\alpha, y_1, y_2, \delta)\right|\right]}{k \log n} - f(\alpha, y_1, y_2) \right| < \epsilon. \tag{5}$$

As a result, when $f(\alpha, y_1, y_2) < 0$, for every $\epsilon > 0$ and $c > 0$, there exists $\delta \in (0, \alpha)$ and $n_0 > 0$ such that for all $n \geq n_0$ and $k \leq c \log n$

$$\mathbb{P}\left(\mathcal{O}_k(\alpha, y_1, y_2, \delta) \neq \emptyset\right) < \epsilon. \tag{6}$$

We see that the region $\mathcal{R}(\alpha) \triangleq \{(y_1, y_2) : f(\alpha, y_1, y_2) \geq 0\}$ identifies the region of achievable in expectation overlaps for matrices with average values approximately $\alpha\sqrt{2\log n/k}$.

Viewing the region $\mathcal{R}(\alpha)$ as a function of α, we establish two phase transition points: the first at $\alpha_1^* = \sqrt{3/2} = 1.2247..$, and the second at $\alpha_2^* = 5\sqrt{2}/(3\sqrt{3}) = 1.3608..$. The derivation of these values is delayed till Section 4. Computing $\mathcal{R}(\alpha)$ numerically we see that it exhibits three qualitatively different behaviors for $\alpha \in (0, \alpha_1^*)$, (α_1^*, α_2^*) and $(\alpha_2^*, \sqrt{2})$, respectively, as shown in Figures 3, 4 and 6. These figures are heat maps of f, where the darker color corresponds to the higher value of f, and the lighter color corresponds to the lower value of f.

(a) When $\alpha \in (1, \sqrt{3}/\sqrt{2})$, $\mathcal{R}(\alpha)$ coincides with the entire region $[0,1]^2$, see Figure 3. In this figure, in particular, the part closer to the origin, we see that the bulk of the overlaps corresponds to the pairs (y_1, y_2) which are very close to the origin. In other words, the picture suggests that most matrices with average value approximately $\alpha\sqrt{2\log n/k}$ tend to be far from each other.

(b) When $\alpha \in (\sqrt{3}/\sqrt{2}, 5\sqrt{2}/(3\sqrt{3}))$, we see that $\mathcal{R}(\alpha)$ is a connected subset of $[0,1]^2$, (Figure 4), but a non-achievable overlap region emerges (colored white on the figure) for pairs of matrices with this average value. At a critical value $\alpha = 5\sqrt{2}/(3\sqrt{3})$ the set is connected through a single point $(1/3, 1/3)$, see Figure 5.

(c) When $\alpha \in (5\sqrt{2}/(3\sqrt{3}), \sqrt{2})$, $\mathcal{R}(\alpha)$ is a disconnected subset of $[0,1]^2$ and the OGP emerges, see Figure 6 for the case $\alpha = 1.364$. In this case, every pair of matrices has either approximately at least $0.4k$ common columns or at most $0.28k$ common columns.

We conjecture that the regime (c) described on Figure 6 corresponds to the hard on average case for which we predict that no polynomial time algorithm exists for non-constant k. Since the OGP was analyzed based on overlaps of two matrices and the overlap of three matrices is likely to push the critical value of OGP even lower, we further conjecture that the hardness regime begins at a value strictly lower than our current estimate $5\sqrt{2}/(3\sqrt{3})$. An interesting open question is to conduct an overlap analysis of m-tuples of matrices and identify the critical value for the onset of disconnectedness.

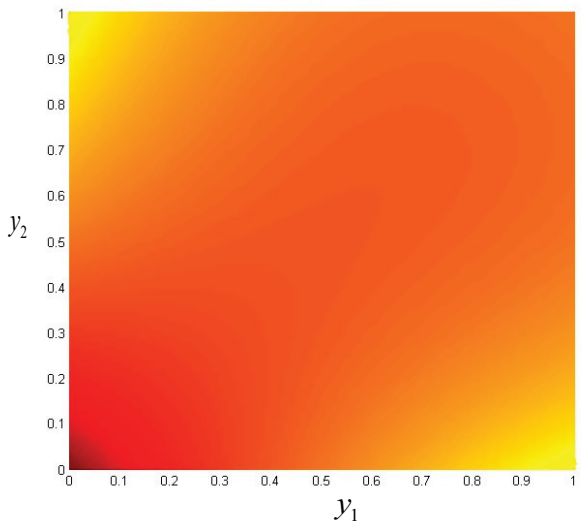

Figure 3: $\mathcal{R}(\alpha)$ for $\alpha \in (0, \sqrt{3}/\sqrt{2})$

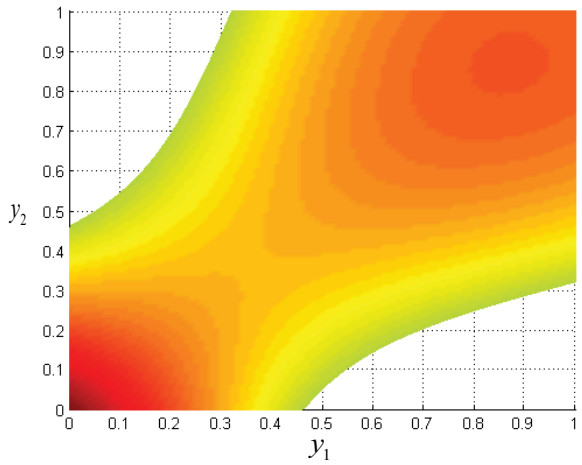

Figure 4: $\mathcal{R}(\alpha)$ for $\alpha \in (\sqrt{3}/\sqrt{2}, 5\sqrt{2}/(3\sqrt{3}))$

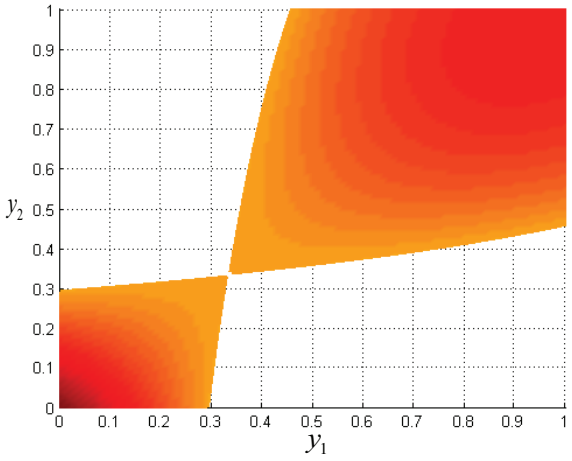

Figure 5: $\mathcal{R}(5\sqrt{2}/(3\sqrt{3}))$

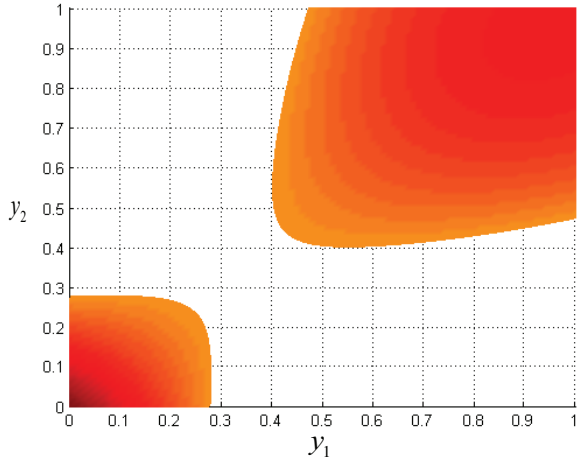

Figure 6: $\mathcal{R}(\alpha)$ for $\alpha \in (5\sqrt{2}/(3\sqrt{3}), \sqrt{2})$

3 Analysis of the $\mathcal{IGP}$ algorithm

This section is devoted to the proof of Theorem 2.3. Denote by I_r^n the set of rows produced by $\mathcal{IGP}$ algorithm in steps $2r$, $r = 0, 1, \ldots, k-1$ and by J_r^n the set of columns produced by $\mathcal{IGP}$ algorithm in steps $2r-1$, $r = 1, \ldots, k$. Their cardinalities satisfy $|I_r^n| = r+1$ for $r = 0, 1, \ldots, k-1$ and $|J_r^n| = r$ for $r = 1, \ldots, k$. In particular, $\mathcal{IGP}$ algorithm chooses $I_0^n = \{i_1\}$ arbitrarily from P_1^n and J_1^n is obtained by finding the column in $\mathbf{C}_{i_1, P_1^n}$ corresponding to the largest entry. Let M_i^n, $i = 1, 2, \cdots, 2k-1$ be the entry sum of the row or column $\mathcal{IGP}$ algorithm adds to the submatrix in the i-th step, namely

$$M_{2r-1}^n \triangleq \max_{j \in P^n_{|I^n_{r-1}|}} \sum_{i \in I^n_{r-1}} C_{i,j} \text{ for } r = 1, 2, \ldots, k,$$
$$M_{2r}^n \triangleq \max_{i \in P^n_{|J^n_r|+1}} \sum_{j \in J^n_r} C_{i,j} \text{ for } r = 1, 2, \ldots, k-1. \tag{7}$$

Introduce

$$b_n := \sqrt{2\log n} - \frac{\log(4\pi \log n)}{2\sqrt{2\log n}}. \tag{8}$$

In order to quantify M_i^n, $i = 1, 2, \ldots, 2k-1$, we now introduce a probabilistic bound on the maximum of n independent standard normal random variables.

Lemma 3.1. *Let Z_i, $i = 1, 2, \ldots, n$ be n independent i.i.d. standard normal random variables. There exists a positive integer N and a constant $c > 0$ such that for all $n > N$*

$$\mathbb{P}\left(\left|\sqrt{2\log n}\left(\max_{1\le i\le n} Z_i - b_n\right)\right| \le \frac{3}{2}\log\log n\right) \ge 1 - c\frac{1}{(\log n)^{1.5}}. \tag{9}$$

Lemma 3.1 is a cruder version of the well-known fact described later in Section 5 as fact (21). For convenience, in what follows, we use n/k in place of $\lfloor n/k \rfloor$. We first establish Theorem 2.3 from the lemma above, the proof of which we delay to the Appendix A.

Proof of Theorem 2.3. Denote by E_{2r-1}^n, $r = 1, 2, \ldots, k$ the event that

$$\left|\sqrt{2\log(n/k)}\left(\frac{M_{2r-1}^n}{\sqrt{|I_{r-1}^n|}} - b_{n/k}\right)\right| \le \frac{3}{2}\log\log(n/k), \tag{10}$$

and by E_{2r}^n, $r = 1, 2, \ldots, k-1$ the event that

$$\left|\sqrt{2\log(n/k)}\left(\frac{M_{2r}^n}{\sqrt{|J_r^n|}} - b_{n/k}\right)\right| \le \frac{3}{2}\log\log(n/k). \tag{11}$$

By Lemma 3.1 and since $k \le f(n) = o((\log n)^{1.5})$, we can choose a positive integer N_1 and $c > 0$ such that for all $n > N_1$

$$\mathbb{P}\left(E_i^n\right) \ge 1 - c\frac{1}{(\log(n/k))^{1.5}}, \qquad \forall\ 1 \le i \le 2k-1. \tag{12}$$

Since M_i^n, $i = 1, 2, \cdots, 2k-1$ corresponds to non-overlapping parts of $\mathbf{C}^n$, they are mutually independent, and so are E_i^n, $i = 1, 2, \cdots, 2k-1$. Choose another positive integer N_2 such that for all $n > N_2$,

$$3\frac{k}{(\log n)^{1.5}} \ge \frac{1}{(\log(n/k))^{1.5}}(2k-1).$$

Let $N \triangleq \max(N_1, N_2)$. Then for all $n > N$ and $k \le f(n) = o((\log n)^{1.5})$ we have

$$\begin{aligned}\mathbb{P}\left(\cap_{i=1}^{2k-1} E_i^n\right) = \prod_{i=1}^{2k-1}\mathbb{P}(E_i^n) &\ge \left(1 - c\frac{1}{(\log(n/k))^{1.5}}\right)^{2k-1} \\ &\ge 1 - c\frac{1}{(\log(n/k))^{1.5}}(2k-1) \ge 1 - 3c\frac{k}{(\log n)^{1.5}}.\end{aligned}$$

As a result, $\cap_{i=1}^{2k-1} E_i^n$ occurs w.h.p.

Under the event $\cap_{i=1}^{2k-1} E_i^n$, we use (10) and (11) to estimate the average value of $\mathbf{C}^n_{\mathcal{IGP}}$:

$$\begin{aligned}\mathrm{Ave}(\mathbf{C}^n_{\mathcal{IGP}}) &\le \frac{1}{k^2}\left(\sum_{r=1}^{k}\left(\sqrt{|I^n_{r-1}|}b_{n/k}+\sqrt{|I^n_{r-1}|}\frac{\frac{3}{2}\log\log(n/k)}{\sqrt{2\log(n/k)}}\right)\right.\\ &\qquad\left.+\sum_{r=1}^{k-1}\left(\sqrt{|J^n_r|}b_{n/k}+\sqrt{|J^n_r|}\frac{\frac{3}{2}\log\log(n/k)}{\sqrt{2\log(n/k)}}\right)\right)\\ &= \frac{\sum_{r=1}^{k}\sqrt{|I^n_{r-1}|}b_{n/k}+\sum_{r=1}^{k-1}\sqrt{|J^n_r|}b_{n/k}}{k^2}+\frac{\sum_{r=1}^{k}\sqrt{|I^n_{r-1}|}+\sum_{r=1}^{k-1}\sqrt{|J^n_r|}}{k^2}\frac{\frac{3}{2}\log\log(n/k)}{\sqrt{2\log(n/k)}}.\end{aligned}$$

Recall $|I^n_{r-1}| = r$ and $|J^n_r| = r$. We can choose a positive integer N_3 such that for all $n > N_3$ and $k \le f(n) = o((\log n)^{1.5})$, $2\log(n/k) \ge \log n$ holds. Also using $b_{n/k} < \sqrt{2\log n}$ and $\log\log(n/k) < \log\log n$, the right hand side of the last equation for all $n > N_3$

$$\begin{aligned}&\le \frac{\sum_{r=1}^{k}\sqrt{2\log n}\sqrt{r}+\sum_{r=1}^{k-1}\sqrt{2\log n}\sqrt{r}}{k^2}+\frac{\frac{3}{2}\log\log n}{\sqrt{\log n}}\\ &= 2\sqrt{\frac{2\log n}{k}}\sum_{r=1}^{k-1}\sqrt{\frac{r}{k}}\frac{1}{k}+\frac{\sqrt{2\log n}}{k^{3/2}}+\frac{\frac{3}{2}\log\log n}{\sqrt{\log n}}\\ &\le 2\sqrt{\frac{2\log n}{k}}\int_0^1\sqrt{x}dx+3\max\left(\frac{1}{k}\sqrt{\frac{\log n}{k}},\frac{\log\log n}{\sqrt{\log n}}\right)\\ &= \frac{4}{3}\sqrt{\frac{2\log n}{k}}+3\max\left(\frac{1}{k}\sqrt{\frac{\log n}{k}},\frac{\log\log n}{\sqrt{\log n}}\right).\end{aligned}$$

Similarly we establish that

$$\mathrm{Ave}(\mathbf{C}^n_{\mathcal{IGP}}) \ge \frac{4}{3}\sqrt{\frac{2\log n}{k}}-3\max\left(\frac{1}{k}\sqrt{\frac{\log n}{k}},\frac{\log\log n}{\sqrt{\log n}}\right).$$

Then (3) follows and the proof is completed. □

4 The Ovelap Gap Property

We delay the derivation of the critical values for the two phase transition points $\alpha_1^* = \sqrt{3}/\sqrt{2}$ and $\alpha_2^* = 5\sqrt{2}/(3\sqrt{3})$ to Appendix B. Now we complete the proof of Theorem 2.4.

Proof of Theorem 2.4. The rest of the section is devoted to establishing part (5) of Theorem 2.4. The second result (6) follows from the Markov inequality.

Fix positive integers k_1, k_2, k and n such that $0 < k_1 \le k \le n$ and $0 < k_2 \le k \le n$. Let X, Y_1 and Y_2 be three mutually independent normal random variables: $X \stackrel{d}{=} \mathcal{N}(0, k_1k_2)$ and $Y_1 \stackrel{d}{=} Y_2 \stackrel{d}{=} \mathcal{N}(0, k^2 - k_1k_2)$. Recall the definition of the multinomial coefficient at the end of introduction. Then

$$\begin{aligned}&\mathbb{E}(|\mathcal{O}_k(\alpha, y_1, y_2, \delta)|)\\ &= \sum_{\substack{k_1\in((y_1-\delta)k,(y_1+\delta)k)\\ k_2\in((y_2-\delta)k,(y_2+\delta)k)}}\binom{n}{k-k_1, k_1, k-k_1, n-2k+k_1}\binom{n}{k-k_2, k_2, k-k_2, n-2k+k_2}\times\\ &\times\mathbb{P}\left(X+Y_1, X+Y_2 \in \left[(\alpha-\delta)\sqrt{\frac{2\log n}{k}}, (\alpha+\delta)k^2\sqrt{\frac{2\log n}{k}}\right]\right). \qquad (13)\end{aligned}$$

For the rest of the proof, we will first estimate the last term in (13), then estimate the first two combinatorial terms in (13), and finally compute $\mathbb{E}(|\mathcal{O}_k(\alpha, y_1, y_2, \delta)|)$ by combining the two estimation results.

We estimate the last term in (13) for the cases $\min(k_1, k_2) < k$ and $k_1 = k_2 = k$, separately. Consider the case $\min(k_1, k_2) < k$. We let $\tau \triangleq (\alpha - \delta)\sqrt{2k_1k_2/(k^2 + k_1k_2)}$ and write

$$\mathbb{P}\left(X + Y_1, X + Y_2 \in \left[(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}}, (\alpha + \delta)k^2\sqrt{\frac{2\log n}{k}}\right]\right) = I_1 + I_2$$

where

$$I_1 = \int_{-\infty}^{\tau k^2\sqrt{\frac{2\log n}{k}}} \mathbb{P}\left((\alpha + \delta)k^2\sqrt{\frac{2\log n}{k}} - x \geq Y_1 \geq (\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} - x\right)^2 \frac{1}{\sqrt{2\pi k_1k_2}} \exp\left(-\frac{x^2}{2k_1k_2}\right) dx,$$

$$I_2 = \int_{\tau k^2\sqrt{\frac{2\log n}{k}}}^{\infty} \mathbb{P}\left((\alpha + \delta)k^2\sqrt{\frac{2\log n}{k}} - x \geq Y_1 \geq (\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} - x\right)^2 \frac{1}{\sqrt{2\pi k_1k_2}} \exp\left(-\frac{x^2}{2k_1k_2}\right) dx.$$

Now we estimate I_1. Let

$$u(x) = \frac{(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} - x}{\sqrt{k^2 - k_1k_2}}.$$

We claim that for $x \leq \tau k^2\sqrt{\frac{2\log n}{k}}$, $u(x)$ diverges to infinity as $n \to \infty$. Namely, $\lim_{n\to\infty} \min u(x) = \infty$ where the minimum is over $x \leq \tau k^2\sqrt{\frac{2\log n}{k}}$. We have

$$\begin{aligned} u(x) &\geq \frac{(\alpha - \delta - \tau)k^2\sqrt{\frac{2\log n}{k}}}{\sqrt{k^2 - k_1k_2}} \\ &= \frac{1 - \sqrt{2k_1k_2/(k^2 + k_1k_2)}}{\sqrt{k^2 - k_1k_2}}(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} \\ &= \frac{1 - \sqrt{1 - (k^2 - k_1k_2)/(k^2 + k_1k_2)}}{\sqrt{k^2 - k_1k_2}}(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}}. \end{aligned}$$

Using the fact $\sqrt{1 - a} \leq 1 - a/2$ for $a = (k^2 - k_1k_2)/(k^2 + k_1k_2) \in [0, 1]$, we have that the expression above is at least

$$\begin{aligned} \frac{\sqrt{k^2 - k_1k_2}}{2(k^2 + k_1k_2)}(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} &\geq \frac{\sqrt{k^2 - k(k-1)}}{4k^2}(\alpha - \delta)k^2\sqrt{\frac{2\log n}{k}} \\ &= \frac{\alpha - \delta}{4}\sqrt{2\log n}, \end{aligned}$$

and the claim is verified. Then using (1) to approximate the first term in the integrand of I_1, we further divide I_1 into two parts as follows

$$\frac{1}{k\log n}\log I_1 = o(1) + \frac{1}{k\log n}\log(I_{11} + I_{12})$$

where

$$I_{11} = \int_{-k^2(\log n)^{2/3}}^{\tau k^2\sqrt{\frac{2\log n}{k}}} \frac{1}{2\pi u(x)^2}\frac{1}{\sqrt{2\pi k_1k_2}} \exp\left(-\frac{((\alpha - \delta)k^2\sqrt{2\log n/k} - x)^2}{2(k^2 - k_1k_2)} \times 2 - \frac{x^2}{2k_1k_2}\right) dx,$$

$$I_{12} = \int_{-\infty}^{-k^2(\log n)^{2/3}} \frac{1}{2\pi u(x)^2}\frac{1}{\sqrt{2\pi k_1k_2}} \exp\left(-\frac{((\alpha - \delta)k^2\sqrt{2\log n/k} - x)^2}{2(k^2 - k_1k_2)} \times 2 - \frac{x^2}{2k_1k_2}\right) dx.$$

Since for any $x \in [-k^2(\log n)^{2/3}, \tau k^2\sqrt{\frac{2\log n}{k}}]$

$$\frac{1}{k\log n}\log(u(x)^2) = o(1),$$

we have

$$\begin{aligned}
&\frac{1}{k\log n}\log I_{11} \\
&= o(1) + \frac{1}{k\log n}\log\int_{-k^2(\log n)^{2/3}}^{\tau k^2\sqrt{\frac{2\log n}{k}}} \frac{1}{\sqrt{2\pi k_1 k_2}}\exp\left(-\frac{((\alpha-\delta)k^2\sqrt{2\log n/k}-x)^2}{2(k^2-k_1k_2)}\times 2 - \frac{x^2}{2k_1k_2}\right)dx
\end{aligned}$$

Then we rewrite the integrand in the equation above in terms of a density function of a normal random variable

$$\begin{aligned}
\frac{1}{k\log n}\log I_{11} &= o(1) - 2(\alpha-\delta)^2\frac{k^2}{k^2+k_1k_2} \\
&\quad + \frac{1}{k\log n}\log\int_{-k^2(\log n)^{2/3}}^{\tau k^2\sqrt{\frac{2\log n}{k}}} \frac{1}{\sqrt{2\pi\frac{k_1k_2(k^2-k_1k_2)}{k^2+k_1k_2}}}\exp\left(-\frac{\left(x-\frac{2k_1k_2k^2(\alpha-\delta)\sqrt{2\log n/k}}{k^2+k_1k_2}\right)^2}{2\frac{k_1k_2(k^2-k_1k_2)}{k^2+k_1k_2}}\right)dx. \quad (14)
\end{aligned}$$

It follows from $\tau = (\alpha-\delta)\sqrt{2k_1k_2/(k^2+k_1k_2)}$ and $\sqrt{a} > a$ for $a \in (0,1)$ that

$$\tau k^2\sqrt{\frac{2\log n}{k}} \geq \frac{2k_1k_2k^2(\alpha-\delta)\sqrt{2\log n/k}}{k^2+k_1k_2}. \quad (15)$$

Also we have as $n \to \infty$

$$\frac{-k^2(\log n)^{2/3} - \frac{2k_1k_2k^2(\alpha-\delta)\sqrt{2\log n/k}}{k^2+k_1k_2}}{\sqrt{\frac{k_1k_2(k^2-k_1k_2)}{k^2+k_1k_2}}} \to -\infty. \quad (16)$$

Since the integrand in (14) is a density function of a normal random variable, (15) and (16) implies that the integral in (14) is in $[1/2+o(1), 1]$. The last term in (14) is $o(1)$ and thus

$$\frac{1}{k\log n}\log I_{11} = o(1) - 2(\alpha-\delta)^2\frac{k^2}{k^2+k_1k_2}.$$

Also we have

$$\frac{1}{k\log n}\log I_{12} \leq \frac{1}{k\log n}\log\int_{-\infty}^{-k^2(\log n)^{2/3}}\exp\left(-\frac{x^2}{2k_1k_2}\right)dx.$$

where the right hand side goes to $-\infty$ as $n \to \infty$.

Now we estimate I_2. We have

$$\begin{aligned}
\frac{1}{k\log n}\log I_2 &\leq \frac{1}{k\log n}\log\int_{\tau k^2\sqrt{\frac{2\log n}{k}}}^{\infty}\exp\left(-\frac{x^2}{2k_1k_2}\right)dx \\
&= o(1) - \tau^2\frac{k^2}{k_1k_2} \\
&= o(1) - 2(\alpha-\delta)^2\frac{k^2}{k^2+k_1k_2}.
\end{aligned}$$

Using $\log(\max(a,b)) \le \log(a+b) \le \log(2\max(a,b))$ for $a,b>0$, we conclude

$$\begin{aligned}
&\frac{1}{k\log n}\log\mathbb{P}\left(X+Y_1, X+Y_2 \in \left[(\alpha-\delta)k^2\sqrt{\frac{2\log n}{k}}, (\alpha+\delta)k^2\sqrt{\frac{2\log n}{k}}\right]\right)\\
&= \frac{1}{k\log n}\log(I_1+I_2)\\
&= o(1) + \frac{1}{k\log n}\max(\log I_1, \log I_2)\\
&= o(1) + \frac{1}{k\log n}\max(\log I_{11}, \log I_{12}, \log I_2)\\
&= o(1) - 2(\alpha-\delta)^2\frac{k^2}{k^2+k_1k_2}. \qquad (17)
\end{aligned}$$

Consider the case $k_1 = k_2 = k$. Observing $Y_1 = Y_2 = 0$ and using (1), we obtain

$$\begin{aligned}
&\frac{1}{k\log n}\log\mathbb{P}\left(X+Y_1, X+Y_2 \in \left[(\alpha-\delta)k^2\sqrt{\frac{2\log n}{k}}, (\alpha+\delta)k^2\sqrt{\frac{2\log n}{k}}\right]\right)\\
&= \frac{1}{k\log n}\log\mathbb{P}\left(X \in \left[(\alpha-\delta)k^2\sqrt{\frac{2\log n}{k}}, (\alpha+\delta)k^2\sqrt{\frac{2\log n}{k}}\right]\right) = o(1) - (\alpha-\delta)^2.
\end{aligned}$$

Hence the equation (17) still holds for the case $k_1 = k_2 = k$.

Now we estimate the first two terms in (13). It is for this part that the assumption $k \le c\log n$ will be used. Let $\beta_1 \triangleq k_1/k$ and $\beta_2 \triangleq k_2/k$. Using the Stirling's approximation $a! \approx \sqrt{2\pi a}(a/e)^a$, $(n-b)\log(n-b) = (n-b)\log n - b(1+o(1))$ for $b = O(\log n)$ and $k \le c\log n$, taking log of the first two terms in the right hand side of (13) gives

$$\begin{aligned}
&\log\left(\frac{n!}{(k-k_1)!k_1!(k-k_1)!(n-2k+k_1)!}\frac{n!}{(k-k_2)!k_2!(k-k_2)!(n-2k+k_2)!}\right)\\
=&O(1) + \log\left(\frac{\sqrt{2\pi n}n^n}{2\pi(k-k_1)(k-k_1)^{2(k-k_1)}\sqrt{2\pi k_1}k_1^{k_1}\sqrt{2\pi(n-2k+k_1)}(n-2k+k_1)^{n-2k+k_1}}\times\right.\\
&\qquad\left.\times\frac{\sqrt{2\pi n}n^n}{2\pi(k-k_2)(k-k_2)^{2(k-k_2)}\sqrt{2\pi k_2}k_2^{k_2}\sqrt{2\pi(n-2k+k_2)}(n-2k+k_2)^{n-2k+k_2}}\right)\\
=&O(1) + (\log n + 2n\log n) - (\log(k-k_1) + 2(k-k_1)\log(k-k_1)) - (\frac{1}{2}\log k_1 + k_1\log k_1)\\
&- (\frac{1}{2}\log(n-2k+k_1) + (n-2k+k_1)\log(n-2k+k_1)) - (\log(k-k_2) + 2(k-k_2)\log(k-k_2))\\
&- (\frac{1}{2}\log k_2 + k_2\log k_2) - (\frac{1}{2}\log(n-2k+k_2) + (n-2k+k_2)\log(n-2k+k_2))\\
=&o(1)k\log n + (\log n + 2n\log n) - (\frac{1}{2}\log(n-2k+k_1) + (n-2k+k_1)\log(n-2k+k_1))\\
&- (\frac{1}{2}\log(n-2k+k_2) + (n-2k+k_2)\log(n-2k+k_2))\\
=&(4-\beta_1-\beta_2+o(1))k\log n. \qquad (18)
\end{aligned}$$

Then it follows from (18) and (17) that

$$\frac{1}{k\log n}\log\mathbb{E}(|\mathcal{O}_k(\alpha,y_1,y_2,\delta)|) = \sup_{\substack{\beta_1\in(y_1-\delta,y_1+\delta)\\ \beta_2\in(y_2-\delta,y_2+\delta)}} 4-\beta_1-\beta_2-\frac{2}{1+\beta_1\beta_2}(\alpha-\delta)^2+o(1)$$
$$= \sup_{\substack{\beta_1\in(y_1-\delta,y_1+\delta)\\ \beta_2\in(y_2-\delta,y_2+\delta)}} f(\alpha-\delta,\beta_1,\beta_2)+o(1)$$

where the region of (β_1,β_2) for the sup above comes from range of the sum in (13). Then (5) follows from the continuity of $f(\alpha,y_1,y_2)$. This completes the proof of Theorem 2.4. □

5 Analysis of the $\mathcal{LAS}$ algorithm

5.1 Preliminary results

We denote by I_r^n the set of rows produced by the $\mathcal{LAS}$ algorithm in iterations $2r, r=0,1,\ldots$ and by J_r^n the set of columns produced by $\mathcal{LAS}$ in iterations $2r-1, r=1,2,\ldots$. Without the loss of generality we set $I_0=\{1,\ldots,k\}$. Then J_1 is obtained by searching the k columns with largest sum of entries in the submatrix $\mathbf{C}^{k\times n}$. Furthermore, $\mathbf{C}^n_{2r+1}=\mathbf{C}^n_{I_r^n,J_{r+1}^n}, r\geq 0$, and $\mathbf{C}^n_{2r}=\mathbf{C}^n_{I_r^n,J_r^n}, r\geq 1$.

Next, for every $r\geq 1$, denote by $\tilde{J}_r^n$ the set of k columns with largest sum of entries in the $k\times(n-k)$ matrix $\mathbf{C}_{I_r^n,[n]\setminus J_r^n}$. In particular, in iteration $2r+1$ the algorithm chooses the best k columns J_{r+1}^n (k columns with largest entry sums) from the $2k$ columns, the k of which are the columns of $\mathbf{C}_{I_r^n,J_r^n}$, and the remaining k of which are columns of $\mathbf{C}_{I_r^n,[n]\setminus J_r^n}$. Similarly, for $r\geq 0$ we define $\tilde{I}_r^n$ to be the set of k rows with largest sum of entries in the $(n-k)\times k$ matrix $\mathbf{C}_{[n]\setminus I_r^n,J_{r+1}^n}$.

The following definition was introduced in [BDN12]:

Definition 5.1. *Let I be a set of k rows and J be a set of k columns in $\mathbf{C}^n$. The submatrix $[\mathbf{C}^n_{ij}]_{i\in I,j\in J}$ is defined to be row dominant in $\mathbf{C}^n$ if*

$$\min_{i\in I}\left\{\sum_{j\in J}\mathbf{C}^n_{ij}\right\}\geq\max_{i\in[n]\setminus I}\left\{\sum_{j\in J}\mathbf{C}^n_{ij}\right\}$$

and is column dominant in $\mathbf{C}^n$ if

$$\min_{j\in J}\left\{\sum_{i\in I}\mathbf{C}^n_{ij}\right\}\geq\max_{j\in[n]\setminus J}\left\{\sum_{i\in I}\mathbf{C}^n_{ij}\right\}.$$

A submatrix which is both row dominant and column dominant is called a locally maximum submatrix.

From the definition above, the $k\times k$ submatrix $\mathcal{LAS}$ returns in each iteration is either row dominant or column dominant, and the final submatrix the $\mathcal{LAS}$ converges to is a locally maximum submatrix.

We now recall the Analysis of Variance (ANOVA) Decomposition of a matrix. Given any $k\times k$ matrix B, let $B_{i\cdot}$ be the average of the ith row , $B_{\cdot j}$ be the average of the jth column, and $B_{\cdot\cdot}:=\text{avg}(B)$ be the average of the matrix B. Then the ANOVA decomposition $\text{ANOVA}(B)$ of the matrix B is defined as

$$\text{ANOVA}(B)_{ij}=B_{ij}-B_{i\cdot}-B_{\cdot j}+B_{\cdot\cdot},\ 1\leq i,j\leq k. \tag{19}$$

The matrix B can then be rewritten as

$$B=\text{avg}(B)\mathbf{1}\mathbf{1}'+\text{Row}(B)+\text{Col}(B)+\text{ANOVA}(B) \tag{20}$$

where $\mathrm{Row}(B)$ denotes the matrix with the ith row entries all equal to $B_{i\cdot} - B_{\cdot\cdot}$ for all $1 \le i \le k$, and similarly $\mathrm{Col}(B)$ denotes the matrix with the jth column entries all equal to $B_{\cdot j} - B_{\cdot\cdot}$ for all $1 \le j \le k$. An essential property of ANOVA decomposition is that, if B consists of independent standard Gaussian variables, the random variables and matrices $B_{\cdot\cdot}, \mathrm{Row}(B), \mathrm{Col}(B)$ and $\mathrm{ANOVA}(B)$ are independent. This property is easily verified by establishing that the corresponding covariances are zero.

Recall the definition of b_n in (8). Let L_n be the maximum of n independent standard normal random variables. It is known that [LLR83]

$$\sqrt{2\log n}(L_n - b_n) \Rightarrow -\log G, \tag{21}$$

as $n \to \infty$, where G is an exponential random variable with parameter 1.

Let (S_1, S_2) be a pair of positive random variables with joint density

$$f(s_1, s_2) = C(\log(1 + s_2/s_1))^{k-1} s_1^{k-1} e^{-(s_1+s_2)}, \tag{22}$$

where C is the normalizing constant to make $f(s_1, s_2)$ a density function. Let $\mathbf{U} = (U_1, \ldots, U_k)$ be a random vector with the Dirichlet distribution with parameter 1. Namely $\mathbf{U}$ is uniformly distributed on the simplex $\{(x_1, \cdots, x_k) \mid \sum_{i=1}^k x_i = 1, x_i \ge 0, 1 \le i \le k\}$. Let

$$\mathbf{C}_\infty^{\mathrm{Row}} \triangleq \left(-\log G, \log(1 + S_1/S_2)\,(k\mathbf{U} - 1)\,\mathbf{1}^T, \mathrm{Col}(\mathbf{C}^k), \mathrm{ANOVA}(\mathbf{C}^k)\right),$$

and

$$\mathbf{C}_\infty^{\mathrm{Col}} \triangleq \left(-\log G, \mathrm{Row}(\mathbf{C}^k), \log(1 + S_1/S_2)\mathbf{1}\,(k\mathbf{U} - 1)^T, \mathrm{ANOVA}(\mathbf{C}^k)\right),$$

where $G, (S_1, S_2), \mathbf{U}$ are independent and distributed as above, and as before $\mathbf{C}^k$ is a $k \times k$ matrix of i.i.d. standard normal random variables independent from $G, (S_1, S_2), \mathbf{U}$.

Denote by $\mathcal{RD}_n$ the event that the matrix $\mathbf{C}^k$ (the top $k \times k$ matrix of $\mathbf{C}^n$) is row dominant. Similarly denote by $\mathcal{CD}_n$ the event that the same matrix is column dominant. Let $\mathbf{D}_{\mathrm{row}}^n$ be a random $k \times k$ matrix distributed as $\mathbf{C}^k$ conditioned on the event $\mathcal{RD}_n$. Similarly define $\mathbf{D}_{\mathrm{col}}^n$.

Introduce the following two operators acting on $k \times k$ matrices A:

$$\Psi_n^{\mathrm{Row}}(A) \triangleq \left(\sqrt{2\log n}(\sqrt{k}\,\mathrm{ave}(A) - b_n), \sqrt{2k\log n}\mathrm{Row}(A), \mathrm{Col}(A), \mathrm{ANOVA}(A)\right) \in \mathbb{R} \times (\mathbb{R}^{k\times k})^3, \tag{23}$$

$$\Psi_n^{\mathrm{Col}}(A) \triangleq \left(\sqrt{2\log n}(\sqrt{k}\,\mathrm{ave}(A) - b_n), \mathrm{Row}(A), \sqrt{2k\log n}\mathrm{Col}(A), \mathrm{ANOVA}(A)\right) \in \mathbb{R} \times (\mathbb{R}^{k\times k})^3. \tag{24}$$

As a result, writing $\Psi_n^{\mathrm{Row}}(A) = (\Psi_{n,j}^{\mathrm{Row}}(A),\ 1 \le j \le 4)$ and applying (20), we have

$$A = \left(\frac{\Psi_{n,1}^{\mathrm{Row}}(A)}{\sqrt{2k\log n}} + \frac{b_n}{\sqrt{k}}\right)\mathbf{1}\mathbf{1}' + \frac{\Psi_{n,2}^{\mathrm{Row}}(A)}{\sqrt{2k\log n}} + \Psi_{n,3}^{\mathrm{Row}}(A) + \Psi_{n,4}^{\mathrm{Row}}(A). \tag{25}$$

A similar expression holds for A in terms of $\Psi_n^{\mathrm{Col}}(A)$.

Bhamidi, Dey and Nobel ([BDN12]) established the limiting distribution result for locally maximum submatrix. For row (column) dominant submatrix, the following result can be easily derived following similar proof.

Theorem 5.2. *For every $k > 0$, the following convergence in distribution takes place as $n \to \infty$:*

$$\Psi_n^{\mathrm{Row}}(\mathbf{D}_{\mathrm{row}}^n) \Rightarrow \mathbf{C}_\infty^{\mathrm{Row}}. \tag{26}$$

Similarly,

$$\Psi_n^{\mathrm{Col}}(\mathbf{D}_{\mathrm{col}}^n) \Rightarrow \mathbf{C}_\infty^{\mathrm{Col}}. \tag{27}$$

Applying ANOVA decomposition (20), the result can be interpreted loosely as follows. $\mathbf{D}^n_{\text{row}}$ is approximately

$$\mathbf{D}^n_{\text{row}} \approx \sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}' + \text{Col}(\mathbf{C}^k) + \text{ANOVA}(\mathbf{C}^k) + O\left(\frac{\log\log n}{\sqrt{\log n}}\right).$$

Indeed the first component of convergence (26) means

$$\text{avg}(\mathbf{D}^n_{\text{row}}) \approx \frac{b_n}{\sqrt{k}} - \frac{\log G}{\sqrt{2k\log n}} = \sqrt{\frac{2\log n}{k}} + O\left(\frac{\log\log n}{\sqrt{\log n}}\right),$$

and the second component of the same convergence means

$$\text{Row}(\mathbf{D}^n_{\text{row}}) = O\left(\frac{1}{\sqrt{\log n}}\right).$$

5.2 Conditional distribution of the row-dominant and column-dominant submatrices

Our next goal is to establish a conditional version of the Theorem 5.2. We begin with several preliminary steps.

Lemma 5.3. *Fix a sequence $Z_1, \ldots, Z_n$ of i.i.d. standard normal random variables and r distinct subsets $I_1, \ldots, I_r \subset [n], |I_\ell| = k, 1 \le \ell \le r$. Let $Y_\ell = k^{-\frac{1}{2}} \sum_{i\in I_\ell} Z_i$. Then there exists a lower triangular matrix*

$$L = \begin{pmatrix} L_{1,1} & 0 & 0 & \cdots & 0 \\ L_{2,1} & L_{2,2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_{r,1} & L_{r,2} & L_{r,3} & \cdots & L_{r,r} \end{pmatrix}, \tag{28}$$

such that

(a) $(Y_1, \ldots, Y_r)^T$ equals in distribution to $L(Y_1, W_2, \ldots, W_r)^T$, where $W_2, \ldots, W_r$ are i.i.d. standard normal random variables independent from Y_1.

(b) The values $L_{i,j}$ are determined completely by the cardinalities of the intersections $I_{\ell_1} \cap I_{\ell_2}, 1 \le \ell_1, \ell_2 \le k$.

(c) $L_{i,1} \in \{0, 1/k, \ldots, (k-1)/k, 1\}$ for all i, with $L_{1,1} = 1$, and $L_{i,1} \le (k-1)/k$, for all $i = 2, \ldots, r$,

(d) $\sum_{1\le i\le r} L^2_{\ell,i} = 1$ for each $\ell = 1, \ldots, r$.

Note that $Y_1, \ldots, Y_r$ are correlated standard normal random variables. The lemma effectively provides a representation of these variables as a linear operator acting on a vector of independent standard normal random variables, where since by condition (c) we have $L_{1,1} = 1$, the first component Y_1 is preserved.

Proof. Let Σ be the covariance matrix of $(Y_1, \ldots, Y_r)$ and let $\Sigma = LL^T$ be its Cholesky factorization. We claim that L has the required property. Note that the elements of Σ are completely determined by the cardinalities of intersections $I_\ell \cap I_{\ell'}, 1 \le \ell, \ell' \le r$ and thus (b) holds. Since Σ is the covariance matrix of $(Y_1, \ldots, Y_r)$ we obtain that this vector equals in distribution $L(W_1, \ldots, W_r)^T$, where $W_i, 1 \le i \le r$ are i.i.d. standard normal and thus (a) holds. We can take W_1 to be Y_1 since Y_1 is also a standard

normal. Note that $L_{1,1}$ is the variance of Y_1 hence $L_{1,1} = 1$. The variance of Y_ℓ is $\sum_{1\le i\le r} L_{\ell,i}^2$ which equals 1 since Y_ℓ is also standard normal, namely (d) holds. Finally, note that $L_{i,1}$ is the covariance of Y_1 with $Y_i, i = 2,\ldots,r$, which takes one of the values $0, 1/k, \ldots, (k-1)/k$, since I_ℓ are distinct subset of $[n]$ with cardinality k. This establishes (c). □

Recall that ω_n denotes any strictly increasing positive function satisfying $\omega_n = o(\sqrt{2\log n})$ and $\log\log n = O(\omega_n)$. We now establish the following conditional version of (21).

Lemma 5.4. *Fix a positive integer $r \ge 2$ and $r \times r$ lower triangular matrix L satisfying $|L_{\ell,i}| \le 1$ and $L_{\ell,1} \le (k-1)/k, \ell = 2,\ldots,r$. Let $\mathbf{Z} = (Z_{i,\ell}, 1 \le i \le n, 1 \le \ell \le r)$ be a matrix of i.i.d. standard normal random variables. Given any $\bar{c} = (c_\ell, 1 \le \ell \le r-1) \in \mathbb{R}^{r-1}$, for each $i = 1,\ldots,n$, let $\mathcal{B}_i = \mathcal{B}_i(\bar{c})$ denote the event*

$$\left[L\left(Z_{i,1}, Z_{i,2}, \ldots, Z_{i,r}\right)^T\right]_\ell \le \sqrt{2\log n} + c_{\ell-1}, \qquad \forall\ 2 \le \ell \le r,$$

where $[\cdot]_\ell$ denotes the ℓ-th component of the vector in the argument. Then for every $w \in \mathbb{R}$

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|_\infty \le \omega_n} \left| \mathbb{P}\left(\sqrt{2\log n}\left(\max_{1\le i\le n} Z_{i,1} - b_n\right) \le w \,\Big|\, \bigcap_{1\le i\le n} \mathcal{B}_i \right) - \exp\left(-\exp\left(-w\right)\right) \right| = 0.$$

Namely, the events $\mathcal{B}_i$ have an asymptotically negligible effect on the weak convergence fact (21), namely that

$$\sqrt{2\log n}(\max_{1\le i\le n} Z_{i,1} - b_n) \Rightarrow -\log G.$$

Proof. Note that the events $\mathcal{B}_i$, $1 \le i \le n$ are independent. Thus we rewrite

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{2\log n}\left(\max_{1\le i\le n} Z_{i,1} - b_n\right) \le w \,\Big|\, \bigcap_{1\le i\le n} \mathcal{B}_i \right) = \mathbb{P}\left(\max_{1\le i\le n} Z_{i,1} \le b_n + \frac{w}{\sqrt{2\log n}} \,\Big|\, \bigcap_{1\le i\le n} \mathcal{B}_i \right) \\
&= \mathbb{P}\left(Z_{1,1} \le b_n + w/\sqrt{2\log n} \;\mid\; \mathcal{B}_i \right)^n \\
&= \left(1 - \frac{\mathbb{P}\left(Z_{1,1} > b_n + w/\sqrt{2\log n}, \mathcal{B}_1\right)}{\mathbb{P}(\mathcal{B}_1)}\right)^n
\end{aligned} \tag{29}$$

Fix any $\delta_1, \delta_2 \in (0, 1/(2k))$. Let $\tilde{\mathcal{B}}_1 = \tilde{\mathcal{B}}_1(\delta_1, \delta_2)$ be the event that

$$Z_{1,1} \le (1+\delta_2) b_n \text{ and } |Z_{1,l}| \le \frac{\delta_1}{r-1} b_n, \qquad \forall\ 2 \le \ell \le r.$$

We claim that $\tilde{\mathcal{B}}_1 \subset \mathcal{B}_1$ for all large enough n and any $\bar{c}$ satisfying $\|\bar{c}\|_\infty \le \omega_n$. Indeed, using $L_{\ell,1} \le (k-1)/k$ and $|L_{\ell,i}| \le 1, \ell = 2,\ldots,r$, the event $\tilde{\mathcal{B}}_1$ implies

$$L_{\ell,1} Z_{1,1} + \sum_{i=2}^{\ell} L_{\ell,i} Z_{1,i} \le (1-1/k)(1+\delta_2) b_n + \delta_1 b_n, \qquad \forall\ 2 \le \ell \le r.$$

Then using $\delta_j \in (0, \frac{1}{2k})$, $\forall j = 1, 2$, we can choose sufficiently large n such that for any $\bar{c}$ satisfying $\|\bar{c}\|_\infty \le \omega_n$,

$$(1-1/k)(1+\delta_2) b_n + \delta_1 b_n \le \sqrt{2\log n} + c_{\ell-1}, \qquad \forall\ 2 \le \ell \le r,$$

from which the claim follows. Then we have

$$1-\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n},\tilde{\mathcal{B}}_1)\geq 1-\frac{\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n},\mathcal{B}_1)}{\mathbb{P}(\mathcal{B}_1)}\geq 1-\frac{\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n})}{\mathbb{P}(\tilde{\mathcal{B}}_1)}. \tag{30}$$

Using (1), we simplify

$$\begin{aligned}\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n},\tilde{\mathcal{B}}_1)&=\mathbb{P}((1+\delta_2)b_n\geq Z_{1,1}>b_n+w/\sqrt{2\log n})\mathbb{P}\left(|Z_{1,\ell}|\leq\frac{\delta_1}{r-1}b_n\right)^{r-1}\\&=\frac{1}{(b_n+w/\sqrt{2\log n})\sqrt{2\pi}}\exp\left(-\frac{(b_n+w/\sqrt{2\log n})^2}{2}\right)(1+o(1)).\end{aligned}$$

Also using $\lim_{n\to\infty}\mathbb{P}(\tilde{\mathcal{B}}_1)=1$, we simplify

$$\frac{\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n})}{\mathbb{P}(\tilde{\mathcal{B}}_1)}=\frac{1}{(b_n+w/\sqrt{2\log n})\sqrt{2\pi}}\exp\left(-\frac{(b_n+w/\sqrt{2\log n})^2}{2}\right)(1+o(1)).$$

The two equations above give the same asymptotics of the two sides in (30). Hence the term in the middle also has the same asymptotics

$$\begin{aligned}1-\frac{\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n},\mathcal{B}_1)}{\mathbb{P}(\mathcal{B}_1)}&=1-\frac{1}{(b_n+w/\sqrt{2\log n})\sqrt{2\pi}}\exp\left(-\frac{(b_n+w/\sqrt{2\log n})^2}{2}\right)(1+o(1))\\&=1-\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n})(1+o(1))\end{aligned} \tag{31}$$

Substituting (31) into (29), we have for any $\bar{c}$ satisfying $\|\bar{c}\|_\infty\leq\omega_n$

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{P}\left(\sqrt{2\log n}\left(\max_{1\leq i\leq n}Z_{i,1}-b_n\right)\leq w\Big|\bigcap_{1\leq i\leq n}\mathcal{B}_i\right)=\lim_{n\to\infty}(1-\mathbb{P}(Z_{1,1}>b_n+w/\sqrt{2\log n}))^n\\&=\lim_{n\to\infty}\mathbb{P}\left(\sqrt{2\log n}\left(\max_{1\leq i\leq n}Z_{i,1}-b_n\right)\leq w\right).\end{aligned}$$

By the limiting distribution of the maximum of n independent standard Gaussians, namely (21),

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{2\log n}\left(\max_{1\leq i\leq n}Z_{i,1}-b_n\right)\leq w\right)=\exp(-\exp(-w)).$$

Then the result follows. □

We now state and prove the main result of this section - the conditional version of Theorem 5.2. By Portmanteau's theorem, a weak convergence $X_n\Rightarrow X$ is established by showing $\mathbb{E}[f(X_n)]\to\mathbb{E}[f(X)]$ for every bounded continuous function f. We use this version in the theorem below.

Theorem 5.5. *Fix a positive integer r and for each n fix any distinct subsets $I_0,\ldots,I_{r-1}\subset[n],|I_\ell|=k,0\leq\ell\leq r-1$, and distinct subsets $J_1,\ldots,J_r\subset[n],|J_\ell|=k,1\leq\ell\leq r$. Fix any sequence $C_1,\ldots,C_{2r-1}$ of $k\times k$ matrices satisfying $\|C_\ell\|_\infty\leq\omega_n,1\leq\ell\leq 2r-1$. Let*

$$\mathcal{E}_r=\mathcal{E}\left(I_i,1\leq i\leq r;J_j,1\leq j\leq r;C_\ell,1\leq\ell\leq 2r-1\right)$$

be the event that $\mathbf{C}^n_{I_{\ell-1},J_\ell}-\sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}'=C_{2\ell-1}$ for each $1\leq\ell\leq r$, $\mathbf{C}^n_{I_\ell,J_\ell}-\sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}'=C_{2\ell}$ for each $1\leq\ell\leq r-1$, and, furthermore, $\sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}'+C_\ell$ is the ℓ-th matrix returned by the algorithm $\mathcal{LAS}$ for

all $\ell = 1, \ldots, 2r-1$. *Namely,* $\mathbf{C}^n_\ell = \sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}' + C_\ell$. *Fix any set of columns* $J \subset [n], |J| = k$ *such that* $J \setminus (\cup_{1\le\ell\le r-1} J_\ell) \neq \emptyset$, *including possibly* J_r, *and let* $\mathbf{D}^n_{\mathrm{Row}}$ *be the* $k \times k$ *submatrix of* $\mathbf{C}^n_{([n]\setminus I_{r-1}),J}$ *with the largest average value and* $\hat{\mathbf{D}}^n_{\mathrm{Row}}$ *be the* $k \times k$ *submatrix of* $\mathbf{C}^n_{\left([n]\setminus\cup_{0\le\ell\le r-1} I_\ell\right),J}$ *with the largest average value. Then, the following holds.*

(a)

$$\lim_{n\to\infty} \inf \mathbb{P}\left(\hat{\mathbf{D}}^n_{\mathrm{Row}} = \mathbf{D}^n_{\mathrm{Row}} | \mathcal{E}_r\right) = 1, \tag{32}$$

where inf *is over all* I_ℓ, J_ℓ *and* $C_\ell, 1 \le \ell \le 2r-1$ *satisfying* $\|C_\ell\|_\infty \le \omega_n$.

(b) Conditional on $\mathcal{E}_r$, $\Psi^{\mathrm{Row}}_n(\mathbf{D}^n_{\mathrm{Row}})$ *converges to* $\mathbf{C}^{\mathrm{Row}}_\infty$ *uniformly in* $(C_\ell, 1 \le l \le 2r-1)$. *Specifically, for every bounded continuous function* $f : \mathbb{R} \times \left(\mathbb{R}^{k\times k}\right)^3 \to \mathbb{R}$ *(and similarly to (26)) we have*

$$\lim_{n\to\infty} \sup \left| \mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{D}^n_{\mathrm{Row}})\right) | \mathcal{E}_r\right] - \mathbb{E}\left[f\left(\mathbf{C}^{\mathrm{Row}}_\infty\right)\right] \right| = 0, \tag{33}$$

where sup *is over all* I_ℓ, J_ℓ *and* $C_\ell, 1 \le \ell \le 2r-1$ *satisfying* $\|C_\ell\|_\infty \le \omega_n$.

(c)

$$\lim_{n\to\infty} \inf \mathbb{P}\left(\|\mathbf{D}^n_{\mathrm{Row}} - \sqrt{\frac{2\log n}{k}}\|_\infty \le \omega_n | \mathcal{E}_r \right) = 1,$$

where inf *is over all* I_ℓ, J_ℓ *and* $C_\ell, 1 \le \ell \le 2r-1$ *satisfying* $\|C_\ell\|_\infty \le \omega_n$.

Similar results of (a), (b) and (c) hold for $\mathbf{D}^n_{\mathrm{Col}}$, $\hat{\mathbf{D}}^n_{\mathrm{Col}}$ *and* $\Psi^{\mathrm{Col}}_n(\mathbf{D}^n_{\mathrm{Col}})$ *when* $I \subset [n], |I| = k$ *is such that* $I \setminus (\cup_{0\le\ell\le r-1} I_\ell) \neq \emptyset$, $\mathbf{D}^n_{\mathrm{Col}}$ *is the* $k \times k$ *submatrix of* $\mathbf{C}^n_{I,([n]\setminus J_r)}$ *with the largest average value and* $\hat{\mathbf{D}}^n_{\mathrm{Col}}$ *is the* $k \times k$ *submatrix of* $\mathbf{C}^n_{I,([n]\setminus\cup_{1\le\ell\le r} J_r)}$ *with the largest average value.*

Regarding the subset of columns J in the theorem above, primarily the special case $J = J_r$ will be used. Note that indeed $J_r \setminus (\cup_{1\le\ell\le r-1} J_\ell) \neq \emptyset$, by applying part (a) of the theorem to the previous step algorithm which claims the identity $\hat{\mathbf{D}}^n_{\mathrm{Col}} = \mathbf{D}^n_{\mathrm{Col}}$ w.h.p.

Proof. Unlike for $\mathbf{D}^n_{\mathrm{Row}}$, in the construction of $\hat{\mathbf{D}}^n_{\mathrm{Row}}$ we only use rows $\mathbf{C}^n_{i,J}$ which are outside the rows $\cup_{0\le\ell\le r-1} I_\ell$ already used in the previous iterations of the algorithm. The bulk of the proof of the theorem will be to establish that claims (b) and (c) of the theorem hold for this matrix instead. Assuming this is the case, (a) then implies that (b) and (c) hold for $\mathbf{D}^n_{\mathrm{Row}}$ as well, completing the proof of theorem.

First we prove part (a) assuming (c) hold for $\hat{\mathbf{D}}^n_{\mathrm{Row}}$. We fix any set of rows $I \subset [n] \setminus I_{r-1}$ with cardinality k satisfying $I \cap (\cup_{0\le\ell\le r-2} I_\ell) \neq \emptyset$. For every $i \in I \cap (\cup_{0\le\ell\le r-2} I_\ell)$ and $j \in J \cap (\cup_{1\le\ell\le r-1} J^n_\ell)$, $\mathbf{C}^n_{i,j}$ is either included in some $\mathbf{C}^n_\ell$, in which case $|\mathbf{C}^n_{i,j} - \sqrt{2\log n/k}| \le \omega_n$ holds under the event $\mathcal{E}_r$, or $\mathbf{C}^n_{i,j}$ is not included in any $\mathbf{C}^n_\ell$, in which case $\mathbf{C}^n_{i,j}$ is $O(1)$ w.h.p. under $\mathcal{E}_r$. Then in both cases we have

$$\lim_{n\to\infty} \inf \mathbb{P}\left(\mathbf{C}^n_{i,j} - \sqrt{\frac{2\log n}{k}} \le \omega_n | \ \mathcal{E}_r \right) = 1,$$

where inf is over all I_ℓ, J_ℓ and $C_\ell, 1 \le \ell \le 2r-1$ satisfying $\|C_\ell\|_\infty \le \omega_n$. Since $|\left(\cup_{0\le\ell\le r-2} I_\ell\right)| \le (r-1)k$ and r is fixed, by the union bound the same applies to all such elements $\mathbf{C}^n_{i,j}$. By part (c) which was assumed to hold for $\hat{\mathbf{D}}^n_{\mathrm{Row}}$, we have

$$\lim_{n\to\infty} \inf \mathbb{P}\left(\sum_{j\in J} \mathbf{C}^n_{i,j} - k\sqrt{\frac{2\log n}{k}} \le k\omega_n, \quad \forall i \in [n]\setminus(\cup_{0\le\ell\le r-1} I_\ell) \mid \mathcal{E}_r \right) = 1,$$

where inf is over the same set of events as above. On the other hand for every $i \in I \cap (\cup_{0\le\ell\le r-2} I_\ell)$ and $j \in J \setminus (\cup_{1\le\ell\le r-1} J_\ell)$, $\mathbf{C}^n_{i,j}$ is not included in any $\mathbf{C}^n_\ell$, $1 \le \ell \le 2r-1$ and hence is $O(1)$ w.h.p. under the event $\mathcal{E}_r$, which gives

$$\lim_{n\to\infty} \sup \mathbb{P}\left(\mathbf{C}^n_{i,j} \ge (1/2)\sqrt{2\log n/k} \mid \mathcal{E}_r\right) = 0. \tag{34}$$

Since $|\cup_{0\le\ell\le r-2} I_\ell| \le (r-1)k$ and r is fixed, by the union bound the same applies to all such elements $\mathbf{C}^n_{i,j}$. It follows, that w.h.p. the average value of the matrix $\mathbf{C}^n_{I,J}$ for all sets of rows $I \in [n] \setminus I_{r-1}$ satisfying $I \cap (\cup_{0\le l\le r-2} I_l) \ne \emptyset$ is at most $(1-1/(2k^2))\sqrt{2\log n/k} + \omega_n$, since by assumption $J \setminus (\cup_{1\le\ell\le r-1} J_\ell) \ne \emptyset$ and thus there exists at least one entry in $\mathbf{C}^n_{I,J}$ satisfying (34). On the other hand by part (c), the average value of $\hat{\mathbf{D}}^n_{\mathrm{Row}}$ is at least $\sqrt{2\log n/k} - \omega_n$ and thus (32) in (a) follows. The proof for $\hat{\mathbf{D}}^n_{\mathrm{Col}}$ is similar.

Thus we now establish (b) and (c) for $\hat{\mathbf{D}}^n_{\mathrm{Row}}$. In order to simplify the notation, we use $\mathbf{D}^n_{\mathrm{Row}}$ in place of $\hat{\mathbf{D}}^n_{\mathrm{Row}}$. We fix I_ℓ, J_ℓ, C_ℓ and J as described in the assumption of the theorem. Let $I^c = [n] \setminus (\bigcup_{0\le\ell\le r-1} I_\ell)$. For each $i \in I^c$ consider the event denoted by $\mathcal{B}^{\mathrm{Row}}_i$ that for each $\ell = 1, \ldots, r-1$ $\mathbf{C}^n_{I_\ell,J_\ell} = \sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}^T + C_{2\ell}$ and

$$\mathrm{Ave}\left(\mathbf{C}^n_{i,J_\ell}\right) \le \min_{i'\in I_\ell} \mathrm{Ave}(\mathbf{C}^n_{i',J_\ell}). \tag{35}$$

Our key observation is that the distribution of the submatrix $\mathbf{C}^n_{I^c,J}$ conditional on the event $\mathcal{E}_r$ is the same as the distribution of the same submatrix conditional on the event $\bigcap_{i\in I^c} \mathcal{B}^{\mathrm{Row}}_i$. Thus in proving parts (b) and (c) we may replace conditioning on the event $\mathcal{E}_r$ by conditioning on the event $\bigcap_{i\in I^c} \mathcal{B}^{\mathrm{Row}}_i$. A similar observation holds for the column version of the statement which we skip.

Now fix any $i \in I^c$. Let $J_0 = J$, and consider the r-vector

$$\left(Y_\ell \triangleq k^{\frac{1}{2}} \mathrm{Ave}(\mathbf{C}^n_{i,J_\ell}), 0 \le \ell \le r-1\right). \tag{36}$$

Without any conditioning the distribution of this vector is the distribution of standard normal random variables with correlation structure determined by the vector of cardinalities of intersections of the sets J_ℓ, namely vector $\sigma \triangleq (|J_\ell \cap J_{\ell'}|, 0 \le \ell, \ell' \le r-1)$. By Lemma 5.3 there exists a $r \times r$ matrix L which depends on σ only and with properties (a)-(d) described in the lemma, such that the distribution of the vector (36) is the same as the one of LZ, where Z is the r-vector of i.i.d. standard normal random variables. We will establish Theorem 5.5 from the following proposition, which is an analogue of Lemma 5.4. We delay its proof for later.

Proposition 5.6. *Let $\mathbf{Z} = (Z_{i,\ell}, 1 \le i \le n, 1 \le \ell \le r-1)$ be a matrix of i.i.d. standard normal random variables independent from the $n \times k$ matrix $\mathbf{C}^{n\times k}$. Given any $\bar{c} = (c_\ell, 1 \le \ell \le r-1) \in \mathbb{R}^{r-1}$, for each $i = 1, \ldots, n$, let $\mathcal{B}_i$ be the event*

$$\left[L\left(k^{\frac{1}{2}} \mathrm{Ave}(\mathbf{C}^{n\times k}_{i,[k]}), Z_{i,1}, \ldots, Z_{i,r-1}\right)^T\right]_{\ell+1} \le \sqrt{2\log n} + \sqrt{k}c_\ell, \qquad \forall\ 1 \le \ell \le r-1,$$

where $[\cdot]_\ell$ denotes the ℓ-th component of the vector in the argument. For every bounded continuous function $f : \mathbb{R} \times \left(\mathbb{R}^{k\times k}\right)^3 \to \mathbb{R}$

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|_\infty \le \omega_n} \left|\mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k)\right) \mid \mathcal{RD}_n, \cap_{1\le i\le n}\mathcal{B}_i\right] - \mathbb{E}\left[f(\mathbf{C}^{\mathrm{Row}}_\infty)\right]\right| = 0. \tag{37}$$

The proposition essentially says that the events $\mathcal{B}_i$ have an asymptotically negligible effect on the distribution of the largest $k \times k$ submatrix of $\mathbf{C}^{n\times k}$.

First we show how this proposition implies part (b) of Theorem 5.5. The event $\bigcap_{i\in I^c} \mathcal{B}_i^{\mathrm{Row}}$ implies that $\|C_{2\ell}\|_\infty \le \omega_n$, for all ℓ and therefore

$$-\omega_n \le c_\ell \triangleq \min_{i'\in I_\ell} \mathrm{Ave}(\mathbf{C}^n_{i',J_\ell}) - \sqrt{\frac{2\log n}{k}} \le \omega_n, \qquad 1 \le \ell \le r-1.$$

The events $\bigcap_{i\in I^c} \mathcal{B}_i^{\mathrm{Row}}$ and $\bigcap_{1\le i\le n} \mathcal{B}_i$ are then identical modulo the difference of cardinalities $|I^c|$ vs n. Since k is a constant, then $|I^c| = n - O(1)$, and the result is claimed in the limit $n \to \infty$. The assertion (b) holds.

We now establish (c). Recalling the representation (25) and the definition of b_n we have

$$\mathbf{D}^n_{\mathrm{Row}} - \sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}' = \frac{\Psi^{\mathrm{Row}}_{n,1}(\mathbf{D}^n_{\mathrm{Row}})}{\sqrt{2k\log n}}\mathbf{1}\mathbf{1}' + \frac{\Psi^{\mathrm{Row}}_{n,2}(\mathbf{D}^n_{\mathrm{Row}})}{\sqrt{2k\log n}} + \Psi^{\mathrm{Row}}_{n,3}(\mathbf{D}^n_{\mathrm{Row}}) + \Psi^{\mathrm{Row}}_{n,4}(\mathbf{D}^n_{\mathrm{Row}}) + O\left(\frac{\log\log n}{\sqrt{\log n}}\right).$$

The claim then follows immediately from part (b), specifically from the uniform weak convergence $\Psi^{\mathrm{Row}}_n(\mathbf{D}^n_{\mathrm{Row}}) \Rightarrow \mathbf{C}^{\mathrm{Row}}_\infty$. □

Proof of Proposition 5.6. According to Theorem 5.2, for every bounded continuous function f,

$$\lim_{n\to\infty} \mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k)\right) \mid \mathcal{RD}_n\right] = \mathbb{E}\left[f(\mathbf{C}^{\mathrm{Row}}_\infty)\right]. \tag{38}$$

Our goal is to show

$$\lim_{n\to\infty} \sup_{\bar c:\|\bar c\|_\infty\le\omega_n} \left|\mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k)\right) \mid \mathcal{RD}_n, \cap_{1\le i\le n}\mathcal{B}_i\right] - \mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k)\right) \mid \mathcal{RD}_n\right]\right| = 0. \tag{39}$$

(37) follows from (38) and (39). We claim that if the following relation holds for any $W \in \mathbb{R}\times(\mathbb{R}^{k\times k})^3$

$$\lim_{n\to\infty} \sup_{\bar c:\|\bar c\|_\infty\le\omega_n} \left|\frac{\mathbb{P}\left(\bigcap_{1\le i\le n}\mathcal{B}_i | \Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W, \mathcal{RD}_n\right)}{\mathbb{P}\left(\bigcap_{1\le i\le n}\mathcal{B}_i\right)} - 1\right| = 0, \tag{40}$$

then (39) follows. By symmetry

$$\mathbb{P}\left(\mathcal{RD}_n | \bigcap_{1\le i\le n}\mathcal{B}_i\right) = \binom{n}{k}^{-1} = \mathbb{P}\left(\mathcal{RD}_n\right).$$

Using the equation above, we compute

$$\begin{aligned}
\mathbb{E}\left[f\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k)\right) \mid \mathcal{RD}_n, \cap_{1\le i\le n}\mathcal{B}_i\right] &= \int f(W) \frac{d\mathbb{P}\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W, \mathcal{RD}_n, \bigcap_{1\le i\le n}\mathcal{B}_i\right)}{\mathbb{P}\left(\mathcal{RD}_n, \bigcap_{1\le i\le n}\mathcal{B}_i\right)} \\
&= \int f(W) \frac{\mathbb{P}\left(\cap_{1\le i\le n}\mathcal{B}_i \mid \Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W, \mathcal{RD}_n\right)}{\mathbb{P}\left(\bigcap_{1\le i\le n}\mathcal{B}_i\right)} \frac{d\mathbb{P}\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W, \mathcal{RD}_n\right)}{\mathbb{P}\left(\mathcal{RD}_n \mid \bigcap_{1\le i\le n}\mathcal{B}_i\right)} \\
&= \int f(W) \frac{\mathbb{P}\left(\cap_{1\le i\le n}\mathcal{B}_i \mid \Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W, \mathcal{RD}_n\right)}{\mathbb{P}\left(\bigcap_{1\le i\le n}\mathcal{B}_i\right)} d\mathbb{P}\left(\Psi^{\mathrm{Row}}_n(\mathbf{C}^k) = W \mid \mathcal{RD}_n\right).
\end{aligned} \tag{41}$$

Substituting (41) into the left hand side of (39) and then using (40) and the boundedness of f, we obtain (39).

The rest of the proof is to show that (40) holds for any $W \in \mathbb{R} \times (\mathbb{R}^{k\times k})^3$. Fix any $W \triangleq (w_1, W_2, W_3, W_4)$ where $w_1 \in \mathbb{R}$ and $W_2,\ W_3,\ W_4 \in \mathbb{R}^{k\times k}$. Conditional on $\Psi_n^{\mathrm{Row}}(\mathbf{C}^k) = W$, and writing $W_2 = (W_{i,j}^2)$ the average value of the i-th row of $\mathbf{C}^k$ is

$$\mathbf{C}_{i\cdot}^k = \frac{W_{i,1}^2}{\sqrt{2k\log n}} + \frac{w_1}{\sqrt{2k\log n}} + \frac{b_n}{\sqrt{k}} \triangleq w_{i,n}, \qquad i = 1, \ldots, k.$$

where we used (25) and the fact that for any matrix $B \in \mathbb{R}^{k\times k}$, the average value of the i-th row of $\mathrm{Col}(B)$ and $\mathrm{ANOVA}(B)$ is zero, for all $i \in [k]$. Let $c_n(W) = \min_{1\le i\le k} w_{i,n}$. Note that

$$w_{i,n} = \sqrt{\frac{2\log n}{k}} + o(1),\ c_n(W) = \sqrt{\frac{2\log n}{k}} + o(1). \tag{42}$$

The event $\mathcal{RD}_n$ is equivalent to the event

$$\max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W).$$

Now observe that by independence of rows of $\mathbf{Z}$

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{1\le i\le n} \mathcal{B}_i | \Psi_n^{\mathrm{Row}}(\mathbf{C}^k) = W, \max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W)\right) \\
&= \mathbb{P}\left(\bigcap_{1\le i\le k} \mathcal{B}_i | \Psi_n^{\mathrm{Row}}(\mathbf{C}^k) = W\right) \mathbb{P}\left(\bigcap_{k+1\le i\le n} \mathcal{B}_i | \max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W)\right).
\end{aligned} \tag{43}$$

By (21) we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W)\right) &= \lim_{n\to\infty} \mathbb{P}\left(\max_{k+1\le i\le n} \sqrt{2\log n}\left(\sqrt{k}\mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) - b_n\right) \le w_1 + \min_{1\le i\le k} W_{i,1}^2\right) \\
&= \exp\left(-\exp\left(-w_1 - \min_{1\le i\le k} W_{i,1}^2\right)\right).
\end{aligned}$$

Furthermore, by Lemma 5.4 we also have

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|_\infty\le\omega_n} \left|\mathbb{P}\left(\max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W) | \bigcap_{k+1\le i\le n} \mathcal{B}_i\right) - \exp\left(-\exp\left(-w_1 - \min_{1\le i\le k} W_{i,1}^2\right)\right)\right| = 0.$$

Applying Bayes rule, we obtain

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|\le\omega_n} \left|\frac{\mathbb{P}\left(\bigcap_{k+1\le i\le n} \mathcal{B}_i | \max_{k+1\le i\le n} \mathrm{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W)\right)}{\mathbb{P}\left(\bigcap_{k+1\le i\le n} \mathcal{B}_i\right)} - 1\right| = 0. \tag{44}$$

Now we claim that

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|\le\omega_n} \left|\mathbb{P}(\bigcap_{1\le i\le k} \mathcal{B}_i | \Psi_n^{\mathrm{Row}}(\mathbf{C}^k) = W) - 1\right| = 0. \tag{45}$$

Indeed the event $\mathcal{B}_i, i \le k$ conditioned on $\Psi_n^{\text{Row}}(\mathbf{C}^k) = W$ is

$$L_{\ell+1,1} k^{\frac{1}{2}} w_{i,n} + L_{\ell+1,2} Z_{i,1} + \cdots L_{\ell+1,r+1} Z_{i,r} \le \sqrt{2\log n} + c_\ell, \qquad 1 \le \ell \le r.$$

Now recall from Lemma 5.3 that $L_{\ell+1,1} \le 1 - 1/k$. Then applying (42) we conclude

$$L_{\ell+1,1} k^{\frac{1}{2}} w_{i,n} \le (1 - 1/k)\sqrt{2\log n} + o(1).$$

Trivially, we have

$$\lim_{n\to\infty} \mathbb{P}\left(L_{\ell,2} Z_{i,1} + \cdots L_{\ell,r+1} Z_{i,r} \le \frac{1}{2k}\sqrt{2\log n},\ \forall\ 1 \le i \le k,\ 1 \le \ell \le r \right) = 1,$$

simply because $\sqrt{\log n}$ is a growing function and the elements of L are bounded by 1. The claim then follows since $|c_\ell| \le \omega_n = o(\sqrt{2\log n})$. Similar to the reasoning of (45), we also have

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|\le\omega_n} \left| \mathbb{P}\Big(\bigcap_{1\le i\le k} \mathcal{B}_i \Big) - 1 \right| = 0. \tag{46}$$

Then if we multiply the denominator of the first term in (44) by $\mathbb{P}(\bigcap_{1\le i\le k} \mathcal{B}_i)$, we still have

$$\lim_{n\to\infty} \sup_{\bar{c}:\|\bar{c}\|\le\omega_n} \left| \frac{\mathbb{P}\left(\bigcap_{k+1\le i\le n} \mathcal{B}_i | \max_{k+1\le i\le n} \text{Ave}(\mathbf{C}_{i\cdot}^{n\times k}) \le c_n(W) \right)}{\mathbb{P}\left(\bigcap_{1\le i\le n} \mathcal{B}_i \right)} - 1 \right| = 0. \tag{47}$$

Applying (45) and (47) for (43) we obtain (40).

□

5.3 Bounding the number of steps of $\mathcal{LAS}$. Proof of Theorem 2.1

Next we obtain an upper bound on the number of steps taken by the $\mathcal{LAS}$ algorithm as well as a bound on the average value of the matrix $\mathbf{C}_r^n$ obtained by the $\mathcal{LAS}$ algorithm in step r, when r is constant, and use these bounds to conclude the proof of Theorem 2.1. For this purpose, we will rely on a repeated application of Theorem 5.5.

We now introduce some additional notations. Fix r and consider the matrix $\mathbf{C}_{2r}^n = \mathbf{C}_{I_r^n, J_r^n}^n$ obtained in step $2r$ of $\mathcal{LAS}$, assuming $T_{\mathcal{LAS}} \ge 2r$. Recall that $\tilde{I}_{r-1}^n$ is the set of k rows with largest sum of entries in $\mathbf{C}_{[n]\setminus I_{r-1}^n, J_r^n}$. Then the matrix $\mathbf{C}_{2r}^n$ is obtained by combining top rows of $\mathbf{C}_{2r-1}^n = \mathbf{C}_{I_{r-1}^n, J_r^n}^n$ and the top rows of $\mathbf{C}_{\tilde{I}_{r-1}^n, J_r^n}^n$. We denote the part of $\mathbf{C}_{2r}^n = \mathbf{C}_{I_r^n, J_r^n}^n$ coming from $\mathbf{C}_{I_{r-1}^n, J_r^n}^n$ by $\mathbf{C}_{2r,1}^n$ and the part coming from $\mathbf{C}_{\tilde{I}_{r-1}^n, J_r^n}^n$ by $\mathbf{C}_{2r,2}^n$. The rows of $\mathbf{C}_{I_{r-1}^n, J_r^n}^n$ leading to $\mathbf{C}_{2r,1}^n$ are denoted by $I_{r,1}^n \subset I_{r-1}^n$ with $|I_{r,1}^n| \triangleq K_1$ (a random variable), and the rows of $\mathbf{C}_{\tilde{I}_{r-1}^n, J_r^n}^n$ leading to $\mathbf{C}_{2r,2}^n$ are denoted by $I_{r,2}^n \subset \tilde{I}_{r-1}^n$ with $|I_{r,2}^n| \triangleq K_2 = k - K_1$. Thus $I_{r,1}^n \cup I_{r,2}^n = I_r^n$ and $\mathbf{C}_{2r,\ell}^n = \mathbf{C}_{I_{r,\ell}^n, J_r^n}^n, \ell = 1, 2$, as shown in Figure 7 where the symbol '△' represents the entries in $\mathbf{C}_{2r}^n$. Our first step is to show that starting from $r = 2$, for every positive real a the average value of $\mathbf{C}_r^n$ is at least $\sqrt{\frac{2\log n}{k}} + a$ with probability bounded away from zero as n increases. We will only show this result for odd r since by monotonicity we also have $\text{Ave}(\mathbf{C}_{r+1}^n) \ge \text{Ave}(\mathbf{C}_r^n)$.

Proposition 5.7. *There exists a strictly positive function $\psi_1 : \mathbb{R}_+ \to \mathbb{R}_+$ which depends only on k, such that for all $r > 0, a > 0$*

$$\liminf_n \mathbb{P}\left(\text{Ave}(\mathbf{C}_{2r+1}^n) \ge \sqrt{\frac{2\log n}{k}} + a \cup \{T_{\mathcal{LAS}} \le 2r\} | T_{\mathcal{LAS}} \ge 2r - 1 \right) \ge \psi_1(a).$$

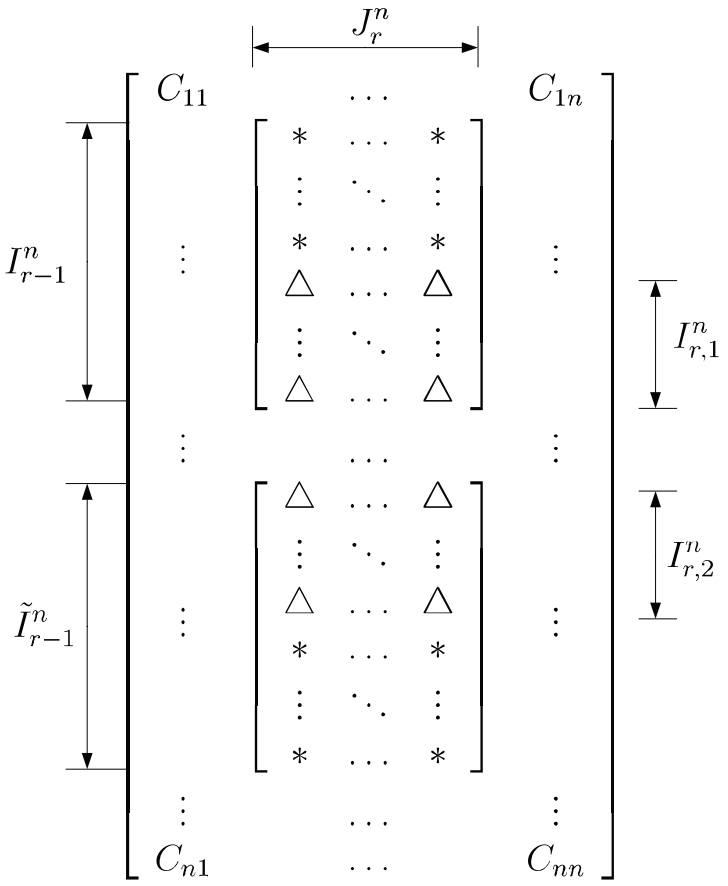

Figure 7: Step $2r$ of $\mathcal{LAS}$ algorithm

Namely, assuming the algorithm proceeds for $2r-1$ steps, with probability at least approximately $\psi_1(a)$ either it stops in step $2r$ or proceeds to step $2r+1$, producing a matrix with average at least $\sqrt{2\log n/k}+a$.

Proof. By Theorem 5.5 the distribution of $\Psi_r^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r})$ conditional on the event $T_{\mathcal{LAS}}\geq 2r-1$ is given by $\mathbf{C}^{\mathrm{Row}}_\infty$ in the limit as $n\to\infty$. In particular, the row averages $\mathrm{Ave}(\mathbf{C}^n_{i,J^n_r}), i\in\tilde{I}^n_{r-1}$ of this matrix are concentrated around $\sqrt{\frac{2\log n}{k}}$ w.h.p. as $n\to\infty$. Motivated by this we write the row averages of $\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}$ as $\sqrt{\frac{2\log n}{k}}+C_1/(\sqrt{2k\log n}),\dots,\sqrt{\frac{2\log n}{k}}+C_k/(\sqrt{2k\log n})$ for the appropriate random values $C_1,\dots,C_k$. Denote the event $\max_j|C_j|\leq\omega_n$ by $\mathcal{L}_{2r}$. Then by Theorem 5.5 we have

$$\lim_{n\to\infty}\mathbb{P}\left(\mathcal{L}^c_{2r}|T_{\mathcal{LAS}}\geq 2r-1\right)=0. \tag{48}$$

If the event $T_{\mathcal{LAS}}\leq 2r-1$ takes place, then $T_{\mathcal{LAS}}\leq 2r$ and therefore the event

$$\left\{\mathrm{Ave}(\mathbf{C}^n_{2r+1})\geq\sqrt{\frac{2\log n}{k}}+a\right\}\cup\{T_{\mathcal{LAS}}\leq 2r\}$$

also takes place. Now consider the event $T_{\mathcal{LAS}}\geq 2r$. On this event the matrices $\mathbf{C}^n_{2r,1}$ and $\mathbf{C}^n_{2r,2}$ are well defined. Recall the notations $I^n_{r,1}$ and $I^n_{r,2}$ for the row indices of $\mathbf{C}^n_{2r,1}$ and $\mathbf{C}^n_{2r,2}$ respectively, and

$0 \le K_1 \le k-1$ and $K_2 = k - K_1$ – their respective cardinalities. Suppose first that

$$\mathrm{Sum}\left(\mathbf{C}^n_{2r,1}\right) > K_1\sqrt{2k\log n} + 2k^2 a, \tag{49}$$

where $\mathrm{Sum}(B)$ denotes the sum of all the entries in a matrix B. Then by the bound $\max_j |C_j| \le \omega_n$ where we recall $\omega_n = o(\sqrt{\log n})$ we have

$$\begin{aligned}\mathrm{Sum}\left(\mathbf{C}^n_{2r}\right) &\ge (K_1+K_2)\sqrt{2k\log n} + 2k^2 a - K_2 k\omega_n/\sqrt{2k\log n} \\ &\ge k^2\sqrt{\frac{2\log n}{k}} + k^2 a,\end{aligned}$$

for large enough n, implying $\mathrm{Ave}\left(\mathbf{C}^n_{2r}\right) \ge \sqrt{\frac{2\log n}{k}} + a$ and therefore either $\mathrm{Ave}\left(\mathbf{C}^n_{2r+1}\right) \ge \sqrt{\frac{2\log n}{k}} + a$, for large enough n or $T_{\mathcal{LAS}} \le 2r$.

Now instead assume the event

$$\mathrm{Sum}\left(\mathbf{C}^n_{2r,1}\right) \le K_1\sqrt{2k\log n} + 2k^2 a, \tag{50}$$

takes place (including the possibility $K_1 = 0$) which we denote by $\mathcal{H}_1$. Then there exists $j_0 \in J^n_r$ such that

$$\mathrm{Sum}\left(\mathbf{C}^n_{I^n_{r,1},j_0}\right) \le K_1\sqrt{\frac{2\log n}{k}} + 2ka.$$

We pick any such column j_0, for example the one which is the smallest index-wise. Consider the event

$$\mathrm{Sum}\left(\mathbf{C}^n_{I^n_{r,2},j_0}\right) \le K_2\sqrt{\frac{2\log n}{k}} - 4k^2 a.$$

which we denote by $\mathcal{H}_2$.

We claim that the probability of the event $\mathcal{H}_2$ conditioned on the events $T_{\mathcal{LAS}} \ge 2r, \mathcal{L}_{2r}$ and $\mathcal{H}_1$ is bounded away from zero as n increases:

$$\liminf_n \mathbb{P}\left(\mathcal{H}_2 | T_{\mathcal{LAS}} \ge 2r, \mathcal{L}_{2r}, \mathcal{H}_1\right) > 0.$$

For this purpose fix any realization of the matrix $\mathbf{C}^n_{2r-1}$ which we write as $\sqrt{\frac{2\log n}{k}} + C$ for an appropriate $k \times k$ matrix C, the realizations $c_1, \ldots, c_k$ of $C_1, \ldots, C_k$, and the realization $j_0 \in J^n_r$, which are all consistent with the events $T_{\mathcal{LAS}} \ge 2r, \mathcal{L}_{2r}, \mathcal{H}_1$. In particular the row averages of $\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}$ are $\sqrt{\frac{2\log n}{k}} + c_1/(\sqrt{2k\log n}), \ldots, \sqrt{\frac{2\log n}{k}} + c_k/(\sqrt{2k\log n})$ and $\max_j |c_j| \le \omega_n$. Note that C and $c_1, \ldots, c_k$ uniquely determine the subsets $I^n_{r,1}$ and $I^n_{r,2}$, and their cardinalities which we denote by I_1, I_2 and k_1, k_2 respectively. Additionally, $c_1, \ldots, c_k$ uniquely determine $\mathrm{Ave}(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r})$:

$$\mathrm{Ave}(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}) = \sqrt{\frac{2\log n}{k}} + \frac{\sum c_j}{k\sqrt{2k\log n}},$$

which we can also write as $\mathrm{Ave}(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}) = \bar{c}/(\sqrt{2k\log n}) + b_n/\sqrt{k}$ where $\bar{c} \triangleq \Psi^{\mathrm{Row}}_{n,1}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right)$. Note that $\max_j |c_j| \le \omega_n = o(\sqrt{\log n})$ also implies $\bar{c} = o(\sqrt{\log n})$. Next we show that

$$\lim_{n\to\infty} \inf_{C,c_1,\ldots,c_k} \mathbb{P}\left(\mathcal{H}_2 | C, c_1 \ldots, c_k\right) \ge \psi_1(a), \tag{51}$$

for some strictly positive function ψ_1 which depends on k only, where $\mathbb{P}(\cdot|C, c_1, \ldots, c_k)$ indicates conditioning on the realizations $C, c_1, \ldots, c_k$ and $\inf_{C,c_1,\ldots,c_k}$ is taken over all choices of $C, c_1, \ldots, c_k$ consistent with the events $T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1$. These realizations imply

$$\Psi_{n,2}^{\text{Row}}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right) = \begin{pmatrix} c_1 - \bar{c} \\ \vdots \\ c_k - \bar{c} \end{pmatrix} \mathbf{1}' + \frac{\log(4\pi \log n)}{2}.$$

where the last term is simply $\sqrt{2\log n}(\sqrt{2\log n} - b_n)$. Thus by representation (25) and by $\bar{c}, c_j = o(\sqrt{\log n})$, we have

$$\begin{aligned}
\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r} &= \frac{\bar{c}}{\sqrt{2k\log n}} + \frac{b_n}{\sqrt{k}}\mathbf{1}\mathbf{1}' + (\sqrt{2k\log n})^{-1}\begin{pmatrix} c_1 - \bar{c} \\ \vdots \\ c_k - \bar{c} \end{pmatrix}\mathbf{1}' + \frac{\log(4\pi\log n)}{2\sqrt{2k\log n}} \\
&\quad + \Psi_{n,3}^{\text{Row}}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right) + \Psi_{n,4}^{\text{Row}}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right) \\
&= \sqrt{\frac{2\log n}{k}}\mathbf{1}\mathbf{1}' + \Psi_{n,3}^{\text{Row}}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right) + \Psi_{n,4}^{\text{Row}}\left(\mathbf{C}^n_{\tilde{I}^n_{r-1},J^n_r}\right) + O\left(\frac{\omega_n}{\sqrt{\log n}}\right),
\end{aligned}$$

(recall that $\log\log n = O(\omega_n)$ and $\omega_n = o(\sqrt{\log n})$). Then by Theorem 5.5 we have

$$\lim_{n\to\infty} \inf_{C,c_1,\ldots,c_k} \mathbb{P}\left(\mathcal{H}_2|C, c_1, \ldots, c_k\right)$$

is the probability that the sum of the entries of $\text{Col}(\mathbf{C}^k) + \text{ANOVA}(\mathbf{C}^k)$ indexed by the subset I_2 and column j_0 is at most $-4k^2a$ which takes some value $\psi(a, |I_2|) > 0$ and depends only on a, k and the cardinality of I_2. Let $\psi_1(a) \triangleq \min_{1\leq|I_2|\leq k}\psi(a, |I_2|)$. Then the claim in (51) follows. We have established

$$\liminf_{n\to\infty} \mathbb{P}\left(\mathcal{H}_2|T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1\right) \geq \psi_1(a).$$

The event $\mathcal{H}_2$ implies that for some column j_0

$$\begin{aligned}
\text{Sum}\left(\mathbf{C}^n_{I^n_r,j_0}\right) &\leq K_1\sqrt{\frac{2\log n}{k}} + 2ka + K_2\sqrt{\frac{2\log n}{k}} - 4k^2a \\
&\leq \sqrt{2k\log n} - 3k^2a.
\end{aligned}$$

By Theorem 5.5 conditional on all of the events $T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1, \mathcal{H}_2$, every column average of $\mathbf{C}_{I^n_r,\tilde{J}^n_r}$ is concentrated around $\sqrt{\frac{2\log n}{k}}$ w.h.p., implying that the column sum is concentrated around $\sqrt{2k\log n}$ w.h.p.. Thus, w.h.p. the j_0-th column will be replaced by one of the column in $\mathbf{C}_{I^n_r,\tilde{J}^n_r}$ (and in particular $T_{\mathcal{LAS}} \geq 2r+1$) and thus during the transition $\mathbf{C}^n_{2r} \to \mathbf{C}^n_{2r+1}$ the sum of the entries increases by $3k^2a - o(1)$, and thus the average value increases by at least $3a - o(1)$ w.h.p. Recall from Theorem 5.2 that w.h.p. $\text{Ave}(\mathbf{C}^n_{2r}) \geq \text{Ave}(\mathbf{C}^n_1) \geq \sqrt{\frac{2\log n}{k}} - a$. Then we obtain $\text{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + 2a - o(1) \geq \sqrt{\frac{2\log n}{k}} + a$ w.h.p. We have obtained

$$\lim_{n\to\infty} \mathbb{P}\left(\text{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + a|T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1, \mathcal{H}_2\right) = 1.$$

By earlier derivation we have

$$\liminf_{n\to\infty} \mathbb{P}\left(\mathcal{H}_2|T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1\right) \geq \psi_1(a),$$

thus implying

$$\liminf_n \mathbb{P}\left(\mathrm{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + a | T_{\mathcal{LAS}} \geq 2r, \mathcal{L}_{2r}, \mathcal{H}_1\right) \geq \psi_1(a).$$

Next recall that $\mathcal{H}_1^c \cap \mathcal{L}_{2r}$ implies either $T_{\mathcal{LAS}} \leq 2r$ or $\mathrm{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + a$ for large enough n, from which we obtain

$$\liminf_n \mathbb{P}\left(\mathrm{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + a \cup \{T_{\mathcal{LAS}} \leq 2r\} | T_{\mathcal{LAS}} \geq 2r-1, \mathcal{L}_{2r}\right) \geq \psi_1(a).$$

Finally, recalling (48) we conclude

$$\liminf_n \mathbb{P}\left(\mathrm{Ave}(\mathbf{C}^n_{2r+1}) \geq \sqrt{\frac{2\log n}{k}} + a \cup \{T_{\mathcal{LAS}} \leq 2r\} | T_{\mathcal{LAS}} \geq 2r-1\right) \geq \psi_1(a).$$

This concludes the proof of Proposition 5.7. □

Now consider the event $T_{\mathcal{LAS}} \geq 2r$, and thus again $\mathbf{C}^n_{2r,1}$ and $\mathbf{C}^n_{2r,2}$ are well-defined. The definitions of $I^n_{r,1}, I^n_{r,2}$ and K_1, K_2 are as above. For any $a > 0$ consider the event for every $j \in J^n_r$ the sum of entries of the column j in $\mathbf{C}^n_{2r,1}$ is at least $K_1\sqrt{\frac{2\log n}{k}} - a$. Denote this event by $\mathcal{F}_{2r}$. Next we show that provided that $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ with probability bounded away from zero as $n \to \infty$, for every fixed r, either the event $\mathcal{F}_{2r+2t}$ takes place for some $t \leq k$ or the algorithm stops earlier. To be more precise

Proposition 5.8. *There exists a strictly positive function $\psi_2 : \mathbb{R}_+ \to \mathbb{R}_+$ which depends on k only such that for every $r > 0$ and $a > 0$*

$$\liminf_{n\to\infty} \mathbb{P}\left(\cup_{0\leq t\leq k}\left(\{T_{\mathcal{LAS}} \leq 2r+2t-1\} \cup \mathcal{F}_{2r+2t}\right) | T_{\mathcal{LAS}} \geq 2r-1, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a\right)$$
$$\geq \psi_2^{2(k+1)}(a).$$

The conditioning on the event $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ will not be used explicitly below. The result just shows that even with this conditioning, the claim still holds, so that this result can be used together with Proposition 5.7.

Proof. On the event $T_{\mathcal{LAS}} \geq 2r-1$, consider the event $\mathcal{G}_{2r}$ defined by

$$\mathcal{G}_{2r} \triangleq \|\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r} - \sqrt{\frac{2\log n}{k}}\|_\infty \leq \frac{a}{4k}. \tag{52}$$

Applying Theorem 5.5, the distribution of $\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}$ conditioned on $T_{\mathcal{LAS}} \geq 2r-1$ and $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ is given asymptotically by $\mathbf{C}^{\mathrm{Row}}_\infty$. Recalling the representation (25) we then have that for a certain strictly positive function ψ_2

$$\liminf_n \mathbb{P}\left(\mathcal{G}_{2r} | T_{\mathcal{LAS}} \geq 2r-1, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a\right) \geq \psi_2(a). \tag{53}$$

If $T_{\mathcal{LAS}} \leq 2r-1$ then the event $\cup_{0\leq t\leq k}\left(\{T_{\mathcal{LAS}} \leq 2r+2t-1\}\cup \mathcal{F}_{2r+2t}\right)$ holds as well. Otherwise assume the event $T_{\mathcal{LAS}} \geq 2r$ takes place and then the matrices $\mathbf{C}^n_{2r,1}$ and $\mathbf{C}^n_{2r,2}$ which constitute $\mathbf{C}^n_{2r} = \mathbf{C}^n_{I^n_r,J^n_r}$ are well-defined. If the event $\mathcal{F}^c_{2r}$ holds then there exists $j_0 \in J^n_r$, such that the sum of entries of the column $\mathbf{C}_{I^n_{r,1},j_0}$ satisfies

$$\mathrm{Sum}\left(\mathbf{C}_{I^n_{r,1},j_0}\right) < |I^n_{r,1}|\sqrt{\frac{2\log n}{k}} - a. \tag{54}$$

The event $\mathcal{G}_{2r}$ implies that the sum of entries of the column $\mathbf{C}^n_{I^n_{r,2},j_0}$ is at most $|I^n_{r,2}|\sqrt{\frac{2\log n}{k}} + a/4$, implying that the sum of entries of the column $\mathbf{C}^n_{I^n_r,j_0}$ is at most

$$|I^n_{r,1}|\sqrt{\frac{2\log n}{k}} - a + |I^n_{r,2}|\sqrt{\frac{2\log n}{k}} + a/4 = \sqrt{2k\log n} - 3a/4. \tag{55}$$

Introduce now the event $\mathcal{G}_{2r+1}$ as

$$\|\mathbf{C}_{I^n_r,\tilde{J}^n_r} - \sqrt{\frac{2\log n}{k}}\|_\infty \leq \frac{a}{4k}. \tag{56}$$

Again applying Theorem 5.5, we have that

$$\liminf_n \mathbb{P}\left(\mathcal{G}_{2r+1}|\mathcal{G}_{2r}, T_{\mathcal{LAS}} \geq 2r, \mathcal{F}^c_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a\right) \geq \psi_2(a), \tag{57}$$

for the same function ψ_2. The event $\mathcal{G}_{2r+1}$ implies that the sum of entries of every column in matrix $\mathbf{C}_{I^n_r,\tilde{J}^n_r}$ is in particular at least $\sqrt{2k\log n} - a/4$. Now recalling (55) this implies that every column $\mathbf{C}^n_{I^n_r,j_0}$ satisfying (54) will be replaced by a new column from $\mathbf{C}_{I^n_r,\tilde{J}^n_r}$ in the transition $\mathbf{C}^n_{2r} \to \mathbf{C}^n_{2r+1}$ (and in particular this transition takes place and $T_{\mathcal{LAS}} \geq 2r+1$). The event $\mathcal{G}_{2r+1}$ then implies that every column $\mathbf{C}^n_{I^n_r,j_0}$ possibly contributing to the event $\mathcal{F}^c_{2r}$ is replaced by a new column in which every entry belongs to the interval $[\sqrt{\frac{2\log n}{k}} - a/(4k), \sqrt{\frac{2\log n}{k}} + a/(4k)]$.

Now if $T_{\mathcal{LAS}} \leq 2r+1$, then also $\cup_{0\leq t\leq k}\left(\{T_{\mathcal{LAS}} \leq 2r+2t-1\}\cup \mathcal{F}_{2r+2t}\right)$. Otherwise, consider $T_{\mathcal{LAS}} \geq 2r+2$. In this case we have a new matrix $\mathbf{C}^n_{2r+2}$ consisting of $\mathbf{C}^n_{2r+2,1}$ and $\mathbf{C}^n_{2r+2,2}$. Note that the event $\mathcal{G}_{2r+1}$ implies that for every subset $I \subset I^n_r$, and for every $j \in \tilde{J}^n_r$, the sum of entries of the sub-column $\mathbf{C}^n_{I,j}$ satisfies

$$\begin{aligned}\mathrm{Sum}\left(\mathbf{C}^n_{I,j}\right) &\geq |I|\left(\sqrt{\frac{2\log n}{k}} - a/(4k)\right)\\ &> |I|\sqrt{\frac{2\log n}{k}} - a.\end{aligned}$$

In particular this holds for $I = I^n_{r+1,1}$ and therefore j does not satisfy the property (54) with $r+1$ replacing r. Thus the columns in $\mathbf{C}^n_{I^n_{r+1,1}}$ satisfying (54) with $r+1$ replacing r can only be the columns which *were not* replaced in the transition $\mathbf{C}^n_{2r} \to \mathbf{C}^n_{2r+1}$. Therefore if the event $\mathcal{F}^c_{2r+2}$ takes place, the columns contributing to this event are one of the original columns of $\mathbf{C}^n_{2r}$.

To finish the proof we use a similar construction inductively and use the fact that the total number of original columns is at most k and thus after $2(k+1)$ iterations all of such columns will be replaced with columns for which (54) cannot occur. Thus assuming the events $\mathcal{G}_{2r},\ldots,\mathcal{G}_{2r+2t-1}$ are defined for some $t \geq 1$, on the event $T_{\mathcal{LAS}} \geq 2r+2t-1$ we let

$$\mathcal{G}_{2r+2t} \triangleq \|\mathbf{C}^n_{\tilde{I}^n_{r+t-1},J^n_{r+t}} - \sqrt{\frac{2\log n}{k}}\|_\infty \leq \frac{a}{4k},$$

and on the event $T_{\mathcal{LAS}} \geq 2r + 2t$

$$\mathcal{G}_{2r+2t+1} \triangleq \|\mathbf{C}^n_{I^n_{r+t},\tilde{J}^n_{r+t}} - \sqrt{\frac{2\log n}{k}}\|_\infty \leq \frac{a}{4k}.$$

Applying Theorem 5.5 we have for $t \geq 0$

$$\liminf_n \mathbb{P}\left(\mathcal{G}_{2r+2t}|\cdot\right) \geq \psi_2(a), \tag{58}$$

where $\cdot$ stands for conditioning on $T_{\mathcal{LAS}} \geq 2r + 2t - 1, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ as well as

$$\left(\mathcal{G}_{2r} \cap \cdots \cap \mathcal{G}_{2r+2t-1}\right) \cap \left(\mathcal{F}^c_{2r} \cap \cdots \cap \mathcal{F}^c_{2r+2t}\right)$$

(here for the case $t = 0$ the event above is assume to be the entire probability space and corresponds to the case considered above). Similarly, for $t \geq 0$

$$\liminf_n \mathbb{P}\left(\mathcal{G}_{2r+2t+1}|\cdot\right) \geq \psi_2(a), \tag{59}$$

where $\cdot$ stands for conditioning on $T_{\mathcal{LAS}} \geq 2r + 2t, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ as well as

$$\left(\mathcal{G}_{2r} \cap \cdots \cap \mathcal{G}_{2r+2t}\right) \cap \left(\mathcal{F}^c_{2r} \cap \cdots \cap \mathcal{F}^c_{2r+2t}\right).$$

By the observation above, since the total number of original columns of $\mathbf{C}^n_{2r-1}$ is k, we have

$$\left(\mathcal{G}_{2r} \cap \cdots \cap \mathcal{G}_{2r+2(k+1)}\right) \cap \left(\mathcal{F}^c_{2r} \cap \cdots \cap \mathcal{F}^c_{2r+2(k+1)}\right) = \emptyset.$$

Iterating the relations (58),(59), we conclude that conditional on the events $T_{\mathcal{LAS}} \geq 2r-1, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ with probability at least $\psi_2^{2(k+1)}(a)$ the event $\cup_{0\leq t\leq k}\left(\{T_{\mathcal{LAS}} \leq 2r + 2t - 1\} \cup \mathcal{F}_{2r+2t}\right)$ takes place. This concludes the proof of the proposition. □

Our next step in proving Theorem 2.1 is to show that if the events $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ and $\mathcal{F}_{2r}$ take place (and in particular $T_{\mathcal{LAS}} \geq 2r$) then with probability bounded away from zero as $n \to \infty$ the algorithm actually stops in step $2r$: $T_{\mathcal{LAS}} \leq 2r$.

On the event $T_{\mathcal{LAS}} \geq 2r - 1$, the matrix $\mathbf{C}^n_{\tilde{I}^n_r,J^n_r}$ is defined. As earlier, we write the row averages of $\mathbf{C}^n_{\tilde{I}^n_r,J^n_r}$ as

$$\sqrt{\frac{2\log n}{k}} + C^n_1/(\sqrt{2k\log n}), \ldots, \sqrt{\frac{2\log n}{k}} + C^n_k/(\sqrt{2k\log n}),$$

for the appropriate values $C^n_1, \ldots, C^n_k$. Denote the event $\max_j |C^n_j| \leq \omega_n$ by $\mathcal{L}_{2r}$. Then by Theorem 5.5

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{L}^c_{2r}|T_{\mathcal{LAS}} \geq 2r - 1, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a\right) = 0. \tag{60}$$

This observation will be used for our next result:

Proposition 5.9. *There exists a strictly positive function $\psi_3 : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $r > 0$ and $a > 0$*

$$\liminf_n \mathbb{P}\left(T_{\mathcal{LAS}} \leq 2r|T_{\mathcal{LAS}} \geq 2r, \mathcal{F}_{2r}, \mathcal{L}_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a\right) \geq \psi_3(a).$$

Proof. Consider any $k \times k$ matrix C, which is a realization of the matrix $\mathbf{C}^n_{2r-1} - \sqrt{\frac{2\log n}{k}}$ satisfying $\mathrm{Ave}(C) \geq a$, namely consistent with the event $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$. Note that the event $\mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ implies that at least one of the row averages of $\mathbf{C}^n_{2r-1}$ is also at least $\sqrt{\frac{2\log n}{k}} + a$. This event and the event $\mathcal{L}_{2r}$ then imply that for large enough n, at least one row of $\mathbf{C}^n_{2r-1}$ will survive till the next iteration $T_{\mathcal{LAS}} = 2r$, provided that this iteration takes place, taking into account the realizations of $C^n_1, \ldots, C^n_k$ corresponding to the row averages of $\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}$.

Now we assume that all of the events $T_{\mathcal{LAS}} \geq 2r, \mathcal{F}_{2r}, \mathcal{L}_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ indeed take place. Consider any constant $1 \leq k_1 < k$ and the subset $I \subset I^n_r$ with cardinality k_1 which corresponds to the k_1 largest rows of C with respect to row averages of C (and therefore of $\mathbf{C}^n_{2r-1}$ as well). Let $A_1, \ldots, A_k$ be the column sums of the $k_1 \times k$ submatrix of C indexed by the rows I. Assume $A_1, \ldots, A_k \geq -a$. Consider the event that $I = I^n_{2r,1}$ corresponds precisely to the rows of $\mathbf{C}^n_{2r-1}$ which survive in the next iteration. Then the column sums of $\mathbf{C}^n_{2r,1}$ are $k_1\sqrt{\frac{2\log n}{k}} + A_j, 1 \leq j \leq k$ consistently with the event $\mathcal{F}_{2r}$. Note that the lower bound $\mathrm{Ave}(C) \geq a$ and the fact that the k_1 row selected are the largest $k_1 \geq 1$ rows in C^n implies

$$\sum_{1 \leq j \leq k} A^n_j \geq k_1 a \geq a. \tag{61}$$

In order for the event above to take place it should be the case that indeed precisely $k_2 = k - k_1 < k$ rows of $\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}$ will be used in creating $\mathbf{C}^n_{2r}$ with the corresponding subset $I^n_{2r,2}, |I^n_{2r,2}| = k_2$. We denote this event by $\mathcal{K}_{k_2}$. Note that whether this event takes place is completely determined by the realization C corresponding to the matrix $\mathbf{C}^n_{2r-1}$, in particular the realization of the row averages of this matrix, and the realizations $C_1, \ldots, C_k$ of $C^n_1, \ldots, C^n_k$ corresponding to the row averages of $\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}$. Furthermore, the realizations $C, C_1, \ldots, C_k$ determine the values $A_1, \ldots, A_k$.

We write the k column sums of $\mathbf{C}^n_{2r,2}$ as $k_2\sqrt{\frac{2\log n}{k}} + U^n_j,\ 1 \leq j \leq k$. Then the column sums of $\mathbf{C}^n_{2r}$ are $\sqrt{2k\log n} + U^n_j + A^n_j, 1 \leq j \leq k$. We claim that for a certain strictly positive function ψ_3 which depends on k only these column sums are all at least $\sqrt{2k\log n} + a/(2k)$:

$$\liminf_n \inf \mathbb{P}\left(\sqrt{2k\log n} + U^n_j + A^n_j \geq \sqrt{2k\log n} + a/(2k),\ j = 1, \ldots, k \mid C^n, C^n_1, \ldots, C^n_k\right) \geq \psi_3(a),$$

where inf is over all sequences $C, C_1, \ldots, C_k$ consistent with the events $T_{\mathcal{LAS}} \geq 2r, \mathcal{F}_{2r}, \mathcal{L}_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$. We first show how this claim implies the claim of the proposition. The claim implies that conditional on the realizations of C, $C_1, \ldots, C_k$ these column sums are at least $\sqrt{2k\log n} + a/(2k)$ with probability $\psi_3(a) - o(1)$. By Theorem 5.5 conditional on $\mathbf{C}^n_{2r}$, the column sums of $\mathbf{C}^n_{I^n_r, \tilde{J}^n_r}$ are concentrated around $\sqrt{2k\log n}$ w.h.p. Thus with high probability all columns of $\mathbf{C}^n_{2r}$ dominate the columns of $\mathbf{C}^n_{I^n_r, \tilde{J}^n_r}$ by at least an additive factor $a/(2k) - o(1)$ and therefore algorithm stops at $T_{\mathcal{LAS}} = 2r$. Integrating over $k_2 = 0, \ldots, k-1$ and realizations $C, C_1, \ldots, C_k$ consistent with the events $T_{\mathcal{LAS}} \geq 2r, \mathcal{F}_{2r}, \mathcal{L}_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$ we obtain the result.

Thus it remains to establish the claim. We have

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{2k\log n} + U^n_j + A^n_j \geq \sqrt{2k\log n} + a/(2k),\ j = 1, \ldots, k \mid C, C_1, \ldots, C_k\right) \\
&\quad = \mathbb{P}\left(U^n_j + A^n_j \geq a/(2k),\ j = 1, \ldots, k \mid C, C_1, \ldots, C_k\right).
\end{aligned}$$

Let $\hat{A}_j^n = \min(A_j^n, 2ka)$. Then

$$\begin{aligned}&\mathbb{P}\left(U_j^n + A_j^n \geq a/(2k),\ j = 1, \ldots, k \mid C, C_1, \ldots, C_k\right)\\ &\quad \geq \mathbb{P}\left(U_j^n + \hat{A}_j^n \geq a/(2k),\ j = 1, \ldots, k \mid C, C_1, \ldots, C_k\right).\end{aligned}$$

The event $\mathcal{L}_{2r}$ implies that $\Psi_{n,1}^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}) = o(\sqrt{\log n})$ and thus $\Psi_{n,1}^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r})/\sqrt{2\log n} = o(1)$. By a similar reason $\Psi_{n,2}^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r})/\sqrt{2\log n} = o(1)$ thus implying from (25) that

$$\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r} = \sqrt{\frac{2\log n}{k}} + \Psi_{n,3}^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}) + \Psi_{n,4}^{\mathrm{Row}}(\mathbf{C}^n_{\tilde{I}^n_{r-1}, J^n_r}) + o(1)$$

Then by Theorem 5.5 we have that

$$\begin{aligned}\lim_{n\to\infty} \sup_{C, C_1, \ldots, C_k} \Big| &\mathbb{P}\left(U_j^n + \hat{A}_j^n \geq a/(2k),\ j = 1, \ldots, k \mid C, C_1, \ldots, C_k\right)\\ &- \mathbb{P}\left(U_j + \hat{A}_j^n \geq a/(2k),\ j = 1, \ldots, k | \hat{A}_1^n, \ldots, \hat{A}_k^n\right)\Big| = 0,\end{aligned}$$

where U_j is the j-th column sum of the matrix of the $k_2 \times k$ submatrix of $\mathrm{Col}(\mathbf{C}^k) + \mathrm{ANOVA}(\mathbf{C}^k)$ indexed by $I^n_{r,2}$ and $\sup_{C, C_1, \ldots, C_k}$ is over the realizations $C, C_1, \ldots, C_k$ consistent with $T_{\mathcal{LAS}} \geq 2r, \mathcal{F}_{2r}, \mathcal{L}_{2r}, \mathrm{Ave}(\mathbf{C}^n_{2r-1}) \geq \sqrt{\frac{2\log n}{k}} + a$. Thus it suffice to show that

$$\inf_{\hat{A}_1^n, \ldots, \hat{A}_k^n} \mathbb{P}\left(U_j + \hat{A}_j^n \geq a/(2k),\ j = 1, \ldots, k | \hat{A}_1^n, \ldots, \hat{A}_k^n\right) \geq \psi_3(a),$$

for some strictly positive function ψ_3 which depends on k only, where the infimum is over $\hat{A}_1^n, \ldots, \hat{A}_k^n$ satisfying $-a \leq \hat{A}_j^n \leq 2ka$ and (61). The joint distribution of $U_j, 1 \leq j \leq k$ is the one of $\left(\sqrt{k_2}(Z_j - \bar{Z}), 1 \leq j \leq k\right)$ where $Z_1, \ldots, Z_k$ are i.i.d. standard normal and $\bar{Z} = k^{-1}\sum_{1\leq j\leq k} Z_j$. Thus our goal is to show that

$$\inf_{\hat{A}_1^n, \ldots, \hat{A}_k^n} \mathbb{P}\left(\sqrt{k_2}(Z_j - \bar{Z}) + \hat{A}_j^n \geq a/(2k), 1 \leq j \leq k | \hat{A}_1^n, \ldots, \hat{A}_k^n\right) \geq \psi_3(a),$$

for some ψ_3. The distribution of the normal $\left(\sqrt{k_2}(Z_j - \bar{Z}), j = 1, \ldots, k\right)$ vector has a full support on the set $\{x = (x_1, \ldots, x_k) : \sum_j x_j = 0\}$.

Consider the set of such vectors $x \in \mathbb{R}^k$ satisfying $\sum_j x_j = 0$ and $x_j + \hat{A}_j^n \geq a/(2k)$. Denote this set by $X(\hat{A}_1^n, \ldots, \hat{A}_k^n)$. By (61) we have $\sum_j (a/(2k) - A_j^n) \leq -a/2$. We claim that in fact

$$\sum_j (a/(2k) - \hat{A}_j^n) \leq -a/2 < 0, \tag{62}$$

and thus the set $X(\hat{A}_1^n, \ldots, \hat{A}_k^n)$ is non-empty. Indeed, if $A_j^n \leq 2ka$, for all j then $\hat{A}_j^n = A_j^n$ and assertion holds from (61). Otherwise, if $A_{j_0}^n > 2ka$ for some j_0, then since $A_j^n \geq -a$ and therefore $\hat{A}_j^n \geq -a$, we have

$$\sum_j (a/(2k) - \hat{A}_j^n) \leq a/2 - 2ka + (k-1)a < -ka < -a/2 < 0.$$

In fact since $a > 0$, the set $X(\hat{A}_1^n, \ldots, \hat{A}_k^n)$ has a non-empty interior and thus a positive measure with respect to the induced Lebesgue measure of the subset $\{x = (x_1, \ldots, x_k) : \sum_j x_j = 0\} \subset \mathbb{R}^k$. As a result the probability

$$\mathbb{P}\left((\sqrt{k_2}(Z_j - \bar{Z}), 1 \leq j \leq k) \in X(\hat{A}_1^n, \ldots, \hat{A}_k^n) | \hat{A}_1^n, \ldots, \hat{A}_k^n\right)$$

is strictly positive. This probability is a continuous function of $\hat{A}_1^n, \ldots, \hat{A}_k^n$ which belong to the bounded interval $[-a, 2ka]$. By compactness argument we then obtain

$$\inf \mathbb{P}\left((\sqrt{k_2}(Z_j - \bar{Z}), 1 \le j \le k) \in X(\hat{A}_1^n, \ldots, \hat{A}_k^n) | A_1^n, \ldots, A_k^n\right) > 0,$$

where the infimum is over $-a \le \hat{A}_1^n, \ldots, \hat{A}_k^n \le 2ka$ satisfying (62). Denoting the infimum by $\psi_3(a)$ we obtain the result. □

We now synthesize Propositions 5.7,5.8 and 5.9 to obtain the following corollary.

Corollary 5.10. *There exists a strictly positive function ψ_4 which depends on k only such that for every $r > k+2$ and $a > 0$*

$$\liminf_n \mathbb{P}\left(T_{\mathcal{LAS}} \le 2r | T_{\mathcal{LAS}} \ge 2r - 2k - 3\right) \ge \psi_4(a).$$

Proof. By Proposition 5.7, we have

$$\liminf_n \mathbb{P}\left(\mathrm{Ave}(\mathbf{C}_{2r-2k-1}) \ge \sqrt{\frac{2\log n}{k}} + a \cup \{T_{\mathcal{LAS}} \le 2r - 2k - 2\} | T_{\mathcal{LAS}} \ge 2r - 2k - 3\right) \ge \psi_1(a).$$

Combining with Proposition 5.8, we obtain that there exists $t, 0 \le t \le k$ such that

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}&\left(\{T_{\mathcal{LAS}} \le 2r - 2t - 1\} \cup \left(\mathcal{F}_{2r-2t} \cap \mathrm{Ave}(\mathbf{C}_{2r-2t-1}) \ge \sqrt{\frac{2\log n}{k}} + a\right) \mid T_{\mathcal{LAS}} \ge 2r - 2k - 3\right) \\ &\ge (k+1)^{-1}\psi_1(a)\psi_2^{2(k+1)}(a).\end{aligned}$$

By observation (60) we also obtain

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{P}&\left(\{T_{\mathcal{LAS}} \le 2r - 2t - 1\} \cup \left(\mathcal{F}_{2r-2t} \cap \mathrm{Ave}(\mathbf{C}_{2r-2t-1}) \ge \sqrt{\frac{2\log n}{k}} + a \cap \mathcal{L}_{2r-2t}\right) | T_{\mathcal{LAS}} \ge 2r - 2k - 3\right) \\ &\ge (k+1)^{-1}\psi_1(a)\psi_2^{2(k+1)}(a).\end{aligned}$$

Finally, applying Lemma 5.9 we obtain

$$\liminf_{n\to\infty} \mathbb{P}\left(\{T_{\mathcal{LAS}} \le 2r - 2t\} | T_{\mathcal{LAS}} \ge 2r - 2k - 3\right) \ge (k+1)^{-1}\psi_1(a)\psi_2^{2(k+1)}(a)\psi_3(a),$$

implying by monotonicity the same result for $T_{\mathcal{LAS}} \le 2r$. Letting $\psi_4(a) \triangleq (k+1)^{-1}\psi_1(a)\psi_2^{2(k+1)}(a)\psi_3(a)$, we obtain the result. □

We are now ready to complete the proof of Theorem 2.1.

Proof of Theorem 2.1. Given $\epsilon > 0$ we fix arbitrary $a > 0$ and find $r = r(\epsilon, a)$ large enough so that $(1 - \psi_4(a))^r < \epsilon$. Applying Corollary 5.10 we obtain for $N = r(2k+4)$

$$\begin{aligned}\mathbb{P}\left(T_{\mathcal{LAS}} \ge N\right) &= \prod_{1 \le t \le r} \mathbb{P}\left(T_{\mathcal{LAS}} \ge t(2k+4) | T_{\mathcal{LAS}} \ge (t-1)(2k+4)\right) \\ &\le (1 - \psi_4(a))^r \\ &\le \epsilon,\end{aligned}$$

which gives the first part of Theorem 2.1. We now show (2). Fix $\epsilon > 0$. We have

$$\begin{aligned}
\mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_{T_{\mathcal{LAS}}}) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n\right) &\leq \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_{T_{\mathcal{LAS}}}) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n, T^n_{\mathcal{LAS}} \leq N\right) + \mathbb{P}\left(T^n_{\mathcal{LAS}} > N\right) \\
&\leq \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_{T_{\mathcal{LAS}}}) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n, T^n_{\mathcal{LAS}} \leq N\right) + \epsilon \\
&= \sum_{1\leq r\leq N} \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_r) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n, T^n_{\mathcal{LAS}} = r\right) + \epsilon \\
&\leq \sum_{1\leq r\leq N} \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_r) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n, T^n_{\mathcal{LAS}} \geq r\right) + \epsilon.
\end{aligned}$$

By part (c) of Theorem 5.5, we have for every r

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_r) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n, T^n_{\mathcal{LAS}} \geq r\right) = 0.$$

We conclude that for every ϵ

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\operatorname{Ave}(\mathbf{C}^n_{T_{\mathcal{LAS}}}) - \sqrt{\frac{2\log n}{k}}\right| > \omega_n\right) \leq \epsilon.$$

Since the left hand-side does not depend on ϵ, we obtain (2). This concludes the proof of Theorem 2.1. □

6 Conclusions and Open Questions

We close the paper with several open questions for further research. In light of the new algorithm $\mathcal{IGP}$ which improves upon the $\mathcal{LAS}$ algorithm by a multiplicative factor 4/3, a natural direction is to obtain a better performing polynomial time algorithm. It would be especially interesting if such an algorithm can improve upon the $5\sqrt{2}/3\sqrt{3}$ threshold since it would then indicate that the OGP is not an obstacle for polynomial time algorithms. Improving the $5\sqrt{2}/3\sqrt{3}$ threshold perhaps by considering multi-overlaps of matrices with fixed asymptotic average value is another important challenge. Based on such improvements obtainable for independent sets in sparse random random graphs [RV14] and for random satisfiability (random NAE-K-SAT) problem [GS14b], it is very plausible that such an improvement is achievable.

Studying the maximum submatrix problem for non-Gaussian distribution is another interesting directions, especially for distributions with tail behavior different from the one of the normal distribution, namely for not sub-Gaussian distributions. Heavy tail distributions are of particular interest for this problem.

Finally, a very interesting version of the maximum submatrix problem is the sparse Principal Component Analysis (PCA) problem for sample covariance data. Suppose, $X_i, 1 \leq i \leq n$ are p-dimensional uncorrelated random variables (say Gaussian), and let Σ be the corresponding sample covariance matrix. When the dimension p is comparable with n the distribution of Σ exhibits a non-trivial behavior. For example the limiting distribution of the spectrum is described by the Marcenko-Pastur law as opposed to the "true" underlying covariance matrix which is just the identity matrix. The sparse PCA problem is the maximization problem $\max \beta^T \Sigma \beta$ where the maximization is over p-dimensional vectors β with

$\|\beta\|_2 = 1$ and $\|\beta\|_0 = k$, where $\|a\|_0$ is the number of non-zero components of the vector a (sparsity). What is the limiting distribution of the objective value and what is the algorithmic complexity structure of this problem? What is the solutions space geometry of this problem and in particular, does it exhibit the OGP? The sparse PCA problem has received an attention recently in the hypothesis testing version [BR13a],[BR13b], where it was shown for certain parameter regime, detecting the sparse PCA signal is hard provided the so-called Hidden Clique problem in the theory of random graphs is hard [AKS98]. Here we propose to study the problem from the estimation point of view - computing the distribution of the k-dominating principal components and studying the algorithmic hardness of this problem.

Finally, a bigger challenge is to either establish that the problems exhibiting the OGP are indeed algorithmically hard and do not admit a polynomial time algorithms, or constructing an example where this is not the case. In light of the repeated failure to improve upon the important special case of this problem - largest clique in the Erdös-Rényi graph $\mathbb{G}(n,p)$, this challenge might be out of reach for the existing methods of analysis.

References

[ACO08] Dimitris Achlioptas and Amin Coja-Oghlan, *Algorithmic barriers from phase transitions*, Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on, IEEE, 2008, pp. 793–802.

[ACORT11] D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi, *On the solution space geometry of random formulas*, Random Structures and Algorithms **38** (2011), 251–268.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Random Structures and Algorithms **13** (1998), no. 3-4, 457–466.

[BDN12] Shankar Bhamidi, Partha S Dey, and Andrew B Nobel, *Energy landscape for large average submatrix detection problems in gaussian random matrices*, arXiv preprint arXiv:1211.2284 (2012).

[BR13a] Quentin Berthet and Philippe Rigollet, *Complexity theoretic lower bounds for sparse principal component detection*, Conference on Learning Theory, 2013, pp. 1046–1066.

[BR13b] ———, *Optimal detection of sparse principal components in high dimension*, The Annals of Statistics **41** (2013), no. 4, 1780–1815.

[COE11] A. Coja-Oghlan and C. Efthymiou, *On independent sets in random graphs*, Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2011, pp. 136–144.

[For10] Santo Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), no. 3, 75–174.

[GS14a] David Gamarnik and Madhu Sudan, *Limits of local algorithms over sparse random graphs*, Proceedings of the 5th conference on Innovations in theoretical computer science., ACM, 2014, pp. 369–376.

[GS14b] ———, *Performance of the survey propagation-guided decimation algorithm for the random NAE-K-SAT problem*, arXiv preprint arXiv:1402.0052 (2014).

[GZ17] David Gamarnik and Ilias Zadik, *High-dimensional regression with binary coefficients. Estimating squared error and a phase transition*, arXiv preprint arXiv:1701.04455 (2017).

[Kar76] Richard M Karp, *The probabilistic analysis of some combinatorial search algorithms*, Algorithms and complexity: New directions and recent results **1** (1976), 1–19.

[LLR83] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes*, Springer Series in Statistics, Springer-Verlag, New York, 1983.

[MO04] Sara C Madeira and Arlindo L Oliveira, *Biclustering algorithms for biological data analysis: a survey*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **1** (2004), no. 1, 24–45.

[Mon15] Andrea Montanari, *Finding one community in a sparse graph*, Journal of Statistical Physics **161** (2015), no. 2, 273–299.

[RV14] Mustazee Rahman and Balint Virag, *Local algorithms for independent sets are half-optimal*, arXiv preprint arXiv:1402.0485 (2014).

[SN13] Xing Sun and Andrew B Nobel, *On the maximal size of large-average and anova-fit submatrices in a gaussian random matrix*, Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability **19** (2013), no. 1, 275.

[SWPN09] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel, *Finding large average submatrices in high dimensional data*, The Annals of Applied Statistics (2009), 985–1012.

Appendix A Proof of Lemma 3.1

We have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{2\log n}\left(\max_{1\le i\le n} Z_i - b_n\right)\right| \le \frac{3}{2}\log\log n\right)\\
&= \mathbb{P}\left(\max_{1\le i\le n} Z_i \le \frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right) - \mathbb{P}\left(\max_{1\le i\le n} Z_i < -\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right)\\
&= \mathbb{P}\left(Z_1 \le \frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right)^n - \mathbb{P}\left(Z_1 < -\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right)^n. \qquad (63)
\end{aligned}$$

Next, we use (1) to approximate

$$\begin{aligned}
&\mathbb{P}\left(Z_1 \le \frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right)\\
&= 1 - (1+o(1))\frac{1}{(\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n)\sqrt{2\pi}}\exp\left(-\frac{(\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n)^2}{2}\right)\\
&= 1 - \Theta\left(\frac{1}{n(\log n)^{3/2}}\right) \qquad (64)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(Z_1 < -\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n\right) \\
&= 1 - (1+o(1))\frac{1}{(-\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n)\sqrt{2\pi}} \exp\left(-\frac{(-\frac{3}{2}\log\log n/\sqrt{2\log n} + b_n)^2}{2}\right) \\
&= 1 - \Theta\left(\frac{(\log n)^{3/2}}{n}\right). \qquad (65)
\end{aligned}$$

Now we substitute (64) and (65) into (63)

$$\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{2\log n}\left(\max_{1\le i\le n} Z_i - b_n\right)\right| \le \log\log n\right) \\
&= \left(1 - \Theta\left(\frac{1}{n(\log n)^{3/2}}\right)\right)^n - \left(1 - \Theta\left(\frac{(\log n)^{3/2}}{n}\right)\right)^n \\
&= \left(1 - \Theta\left(\frac{1}{n(\log n)^{3/2}}\right)\right)^n - \exp(-\Theta((\log n)^{3/2})).
\end{aligned}$$

Then the result follows from choosing a positive integer N and a constant $c > 0$ such that for all $n > N$ the following inequality holds

$$\left(1 - \Theta\left(\frac{1}{n(\log n)^{3/2}}\right)\right)^n - \exp(-\Theta((\log n)^{3/2})) \ge 1 - c\frac{1}{(\log n)^{1.5}}.$$

Appendix B Derivation of two phase transition points $\alpha_1^* = \sqrt{3}/\sqrt{2}$ and $\alpha_2^* = 5\sqrt{2}/(3\sqrt{3})$

We start with α_1^* which we define as a critical point such that for any $\alpha > \alpha_1^*$ and $\alpha \in (0, \sqrt{2})$, $\mathcal{R}(\alpha)$ does not cover the whole region $[0,1]^2$, i.e. $[0,1]^2 \setminus \mathcal{R}(\alpha) \neq \emptyset$. We formulate this as follows

$$\alpha_1^* \triangleq \max\{\alpha \in (0,\sqrt{2}) : \min_{y_1,y_2\in[0,1]^2} f(\alpha, y_1, y_2) \ge 0\}. \qquad (66)$$

Since $f(\alpha, y_1, y_2)$ is differentiable with respect to y_1 and y_2, the minimum of $f(\alpha, y_1, y_2)$ for a fixed α appear either at the boundaries or the stationary points. Using the symmetry of y_1 and y_2, we only need to consider the following boundaries

$$\{(y_1, y_2) : y_1 = 0, y_2 \in [0,1]\} \cup \{(y_1, y_2) : y_1 = 1, y_2 \in [0,1]\}.$$

By inspection, $\min_{y_1=0,y_2\in[0,1]} f(\alpha, y_1, y_2) = 3 - 2\alpha^2$ and

$$\min_{y_1=1,y_2\in[0,1]} f(\alpha, y_1, y_2) = \min_{y_2\in[0,1]} \left\{3 - y_2 - \frac{2}{1+y_2}\alpha^2\right\}.$$

Since the objective function above is a concave function with respect to y_2, its minimum is obtained at $y_2 = 0$ or 1, which is $3 - 2\alpha^2$ or $2 - \alpha^2$. Hence the minimum of $f(\alpha, y_1, y_2)$ at the boundaries above is either $3 - 2\alpha^2$ or $2 - \alpha^2$. Both of them being nonnegative requires

$$3 - 2\alpha^2 \ge 0 \text{ and } 2 - \alpha^2 \ge 0 \text{ and } \alpha \in (0,\sqrt{2}) \Rightarrow \alpha \in (0, \sqrt{3}/\sqrt{2}].$$

Next we consider the stationary points of $f(\alpha, y_1, y_2)$ for a fixed α. The stationary points are determined by solving

$$\frac{\partial f(\alpha, y_1, y_2)}{\partial y_1} = 0, \text{ which implies } -1 + \frac{2\alpha^2 y_2}{(1+y_1y_2)^2} = 0,$$
$$\frac{\partial f(\alpha, y_1, y_2)}{\partial y_2} = 0, \text{ which implies } -1 + \frac{2\alpha^2 y_1}{(1+y_1y_2)^2} = 0.$$

Observe from above $y_1 = y_2$. Then we can simplify the equations above by

$$y_1^4 + 2y_1^2 - 2\alpha^2 y_1 + 1 = 0. \tag{67}$$

Using 'Mathematica', we find that the four solutions for the quartic equation above for $\alpha^2 = 3/2$ are complex numbers all with nonzero imaginary parts. Since the equation above does not have real solutions, the optimization problem (66) has maximum at $\alpha = \sqrt{3}/\sqrt{2}$. On the other hand, for any $\alpha > \sqrt{3}/\sqrt{2}$, $f(\alpha, 1, 0) = 3 - 2\alpha^2$ is always negative. Hence, we have $\alpha_1^* = \sqrt{3}/\sqrt{2}$.

We also claim that for any $\alpha \in (0, \sqrt{3}/\sqrt{2})$, $\mathcal{R}(\alpha) = [0,1]^2$. It suffices to show that for any $y \in [0,1]$,

$$y^4 + 2y^2 - 2\alpha^2 y + 1 > 0.$$

Suppose for $\alpha \in (0, \sqrt{3}/\sqrt{2})$ there is a $\hat{y} \in [0,1]$ such that $\hat{y}^4 + 2\hat{y}^2 - 2\alpha^2\hat{y} + 1 \le 0$. Then by $\alpha^2 < 3/2$ and $\hat{y} \neq 0$ we have

$$\hat{y}^4 + 2\hat{y}^2 - 3\hat{y} + 1 < 0.$$

Since $y^4 + 2y^2 - 3y + 1$ is positive at $y = 0$ and negative at $\hat{y}$, the continuity of $y^4 + 2y^2 - 3y + 1$ implies that there is a $y_1 \in [0,1]$ such that (67) holds for $\alpha^2 = 3/2$, which is a contradiction. The claim follows.

Next we introduce α_2^*. Increasing α beyond α_1^*, we are interested in the first point α_2^* at which the function $f(\alpha_2^*, y_1, y_2)$ has at least one real stationary point and the value of $f(\alpha_2^*, y_1, y_2)$ at this point is zero. Observe that at the stationary points $y_1 = y_2$ and y_1 satisfies (67). Then α_2^* is determined by solving

$$\begin{aligned} &y_1^4 + 2y_1^2 - 2\alpha^2 y_1 + 1 = 0, \\ &4 - 2y_1 - \frac{2}{1+y_1^2}\alpha^2 = 0, \\ &y_1 \in [0,1], \quad \alpha \in (\sqrt{3}/\sqrt{2}, \sqrt{2}). \end{aligned}$$

Using 'mathematica' to solve the equations above, we obtain only one real solution $y_1 = 1/3, \alpha = 5\sqrt{2}/(3\sqrt{3})$. Then we have $\alpha_2^* = 5\sqrt{2}/(3\sqrt{3})$ and $f(\alpha_2^*, 1/3, 1/3) = 0$. We verify that $f(\alpha_2^*, 1/3, y_2) < 0$ for $y_2 \in [0,1] \setminus \{1/3\}$ and $f(\alpha_2^*, y_1, 1/3) < 0$ for $y_1 \in [0,1] \setminus \{1/3\}$. By plotting $f(\alpha_2^*, y_1, y_2)$ in Figure 5, we see that the set $\mathcal{R}(\alpha_2^*)$ is connected through a single point $(1/3, 1/3)$.