Analysis of High-Dimensional Data

Dolev Dublon
Moria Grohar
Shirel Zecharia
Roy Harel
Hadas Evers

March 8, 2024

Abstract

This Survey explores the fascinating world of high-dimensional data analysis, focusing on the properties of random matrices. These matrices play a crucial role in various fields such as computer science, statistics, and machine learning, especially concerning their singularity and the analysis of their smallest singular values.

1 Introduction

In the realm of high-dimensional data analysis, random matrices have captivated the attention of mathematicians and scientists due to their ubiquitous presence in various domains, including computer science, statistics, and machine learning. Understanding their behavior is paramount, particularly regarding their singularity (non-invertibility) and the analysis of their smallest singular values. This article delves into ten research papers that shed light on these aspects.

2 Mathematical Context Main

[7] The article "Random Symmetric Matrices Are Almost Surely Non-Singular" by K. Costello, T. Tao, and V. Vu presents a significant result in the field of random matrices. The main finding of the article is the proof that a random symmetric matrix Q_n with independent and identically distributed (with the same distribution) Bernoulli variables as its upper diagonal entries is almost surely non-singular, with a probability of $1 - O(n^{-1/8+\delta})$ for any $\delta > 0$. This result extends previous results for random matrices to more general models of random matrices. The article presents the history of the non-singularity problem in random matrices, namely whether it is true that a random matrix A_n with independent Bernoulli variables is almost surely non-singular. This question was positively answered by Komlós in 1967, and later he generalized the result to more general models of random matrices. In a recent paper, Tao and Vu found a different proof for random matrices that provides a precise estimate for the absolute value of the determinant of the matrix A_n. Building upon these previous proofs, the authors develop a quadratic version of Littlewood-Offord type results concerning the concentration of

random variables to prove the non-singularity of Q_n- a random symmetric matrix. This method allows researchers to overcome the challenge of the row and column transpose, which was a hurdle in previous proofs for random matrices due to the dependence between the row vectors of the matrix Q_n. The article raises open questions for future research in the field of random matrices:

Determinant Estimation: The article raises the question of estimating the determinant of random matrices. The estimation provided in the article is: $|det\ Q_n| = n^{(1/2-o(1))n}$.

Singularity Probability: Another open question raised in the article relates to estimating the probability that a random matrix is singular. The authors estimate that the probability of Q_n being a singular matrix is $(1/2 + o(1))^n$.

The quadratic variant of the Littlewood-Offord : Let Q be a quadratic random variable defined as: $Q = \sum_{1 \le i,j \le n} c_{ij} z_i z_j$ where z_i are random variables, $1, \ldots, n = U_1 \cup U_2$ is a non-trivial partition, and S is a non-empty subset of U_1. For each $i \in S$, let d_i be the number of indices $j \in U_2$ such that $|c_{ij}| \ge 1$. If $d_i \ge 1$ for each $i \in S$, and I is an interval of length 1, then: $P(Q \in I) = O(|S|^{-1/2} + |S|^{-1} \sum_{i \in S} d_i{}^{-1/2})^{1/4}$

3 roy article1 summary

[21]

4 roy article2 summary

[6]

5 shirel Singularity of random symmetric matrices revisited summery

[5] Singularity of Random Symmetric Matrices: This paper studies the probability, denoted as $P(det(M_n) = 0)$, of a singular random $n \times n$ matrix M_n drawn uniformly from matrices with entries of -1 and 1. It's a long-standing problem with the conjecture that $P(det(M_n) = 0)$ goes to zero exponentially fast with n, written as $P(det(M_n) = 0) = (1 + o(1))n^2 2^{-n+1}$. Prior work established bounds on this probability but couldn't overcome a natural barrier of $exp(-c\sqrt{n}logn)$ for some constant c, where the randomness in the matrix isn't "reused."
This paper breaks the barrier by introducing a "rough" inverse Littlewood-Offord theorem, proving that $P(det(M_n) = 0) \le exp(-cpnlogn)$ for some constant c and sufficiently large n. The authors build upon previous work that divides vectors v into structured and unstructured and analyzes their contribution to $P(det(M_n) = 0)$. Their key improvement lies in a simpler and stronger "rough" inverse Littlewood-Offord theorem, which defines concepts like
$N_\mu(w) := x \in \mathbb{Z}_p : P(X\mu(w) = x) > 2^{-1} P(X_\mu(w) = 0)$ to analyze the "neighborhood" of a vector w under a random walk $X_\mu(v) := \varepsilon_1 v_1 + ... + \varepsilon_n v_n$, where ε_i are independent and take values -1, 0, or 1 with equal probability $\mu/2$.

6 moria Singularity of random symmetric matrices revisited summery

[14]

7 moria Singularity of discrete random matrices summery

[12]

8 shirel On the smoothed analysis of of the smallest singular value with discrete noise

[13]

9 dolev article1 summery of graph

ANTICONCENTRATION IN RAMSEY GRAPHS AND A PROOF OF THE ERDŐS-MCKAY CONJECTURE

The paper 'ANTICONCENTRATION IN RAMSEY GRAPHS AND A PROOF OF THE ERDŐS-MCKAY CONJECTURE' [19] by Matthew Kwan, Ashwin Sah, Lisa Saurmann, and Methaab Sawhney presents an examination of edge-statistics within C-Ramsey graphs, contributing insights that advance our understanding toward the resolution of the Littlewood-Offord problem. By analyzing the distribution of edges in random vertex subsets of these graphs, we uncover new patterns that echo the fundamental tenets of the Littlewood-Offord theory, marking a notable progression.

Lets begin with a few basics. An induced subgraph of a graph is called homogeneous if it is a clique or independent set (i.e., all possible edges are present, or none are). One of the most fundamental results in Ramsey theory, proved in 1935 by Erdős and Szekeres [9] states that every n-vertex graph contains a homogeneous subgraph with at least $\frac{1}{2}\log_2 n$ vertices.On the other hand, Erdős [8] famously used the probabilistic method to prove that, for all $n \geq 3$, there is an n-vertex graph with no homogeneous subgraph on $2\log_2 n$ vertices. an upper bound and lower bound.

A C-Ramsey graph is an n-vertex graph without cliques or independent sets larger than $C\log_2 n$, where C is a constant.

9.1 Theorem1

Fix $C > 0$ and $\eta > 0$. Let G be a C-Ramsey graph on n vertices, where n is sufficiently large relative to C and η. Then, for any integer x with $0 \leq x \leq (1-\eta)e(G)$, there exists a subset $U \subseteq V(G)$ inducing exactly x edges. The theorem declares that for any large enough C-Ramsey graph and integer within a specified range, there exists a vertex subset inducing exactly that many edges. [1].

9.2 Edge statistics and low-degree polynomials

For a graph G with vertex set $V(G) = \{1, \ldots, n\}$ and edge set $E(G)$, subset of vertices $U \subseteq V(G)$ with a vector $\xi \in \{0,1\}^n$, where $\xi_i = 1$ if vertex i is in U and $\xi_i = 0$ otherwise. The number of edges $e(G[U])$ in the induced subgraph $G[U]$ can then be represented as the evaluation of a quadratic polynomial: $f(\xi) = \sum_{\{i,j\}\in E(G)} \xi_i\xi_j$ where the sum runs over all edges in $E(G)$.

the statement that G has an induced subgraph with exactly x edges is precisely equivalent to the statement that there is a binary vector $\bar{\xi} \in \{0,1\}^n$ with $f(\bar{\xi}) = x$

Recent studies focus on random variables e(G[U]) for graphs G and random vertex sets U, inspired by conjectures from Alon, Hefetz, Krivelevich, and Tyomkyn. [2].

9.3 Theorem2

Fix $C, \lambda > 0$, let G be a C-Ramsey graph on n vertices, and let $\lambda \leq p \leq 1-\lambda$. Then if U is a random subset of $V(G)$ obtained by independently including each vertex with probability p, we have $\sup_{x\in\mathbb{Z}} \Pr[e(G[U]) = x] \leq K_{C\lambda} n^{-3/2}$ *for some $K_{C\lambda} > 0$ depending only on C and λ. Furthermore, for every fixed $A > 0$, we have*

$\inf_{x\in\mathbb{Z},|x-p^2e(G)|\leq An^{3/2}} \Pr[e(G[U]) = x] \geq \kappa_{CA\lambda} n^{-3/2}$ *for some $\kappa_{CA\lambda} > 0$ depending only on C, A, and λ, if n is sufficiently large in terms of C, λ, and A.* In the contexts

of this Theorem it's noted that the distribution of the number of edges, $e(G[U])$, in a subset U of a graph G, adheres to a central limit theorem [[3], [4], [10], [11]], suggesting $\Pr[e(G[U]) \leq x] = \Phi\left(\frac{x-\mu}{\sigma}\right) + o\left(\frac{1}{\sigma}\right)$, with Φ representing the Gaussian CDF, and μ, σ the mean and standard deviation. However, due to the degree sequence's additive structure, a local central limit theorem, which would adjust to $\Pr[e(G[U]) = x] = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma} + o\left(\frac{1}{\sigma}\right)$ using the Gaussian density function, does not generally hold.

9.4 Small-ball probability for quadratic Gaussian chaos

The study of low-degree polynomials of independent random variables, frequently referred to as chaoses,possesses noteworthy contributions within this domain. for example in the paper of Kim-Vu polynomial concentration [16].... This area of research offers insights into the behavior and characteristics of such polynomials and described as the fundamental tools in probabilistic combinatorics, high-dimensional statistics, the analysis of boolean functions and mathematical modeling. Much of this study has focused on low-degree polynomials of Gaussian random variables, which enjoy certain symmetry properties that make them easier to study. While this direction may not seem obviously relevant to the previous theorem, part of the proof related to applying the celebrated Gaussian invariance principle of Mossel O'Donnell, and Oleszkiewicsz [20] to compare our random variables of interest with certain "Gaussian analogs". Hence, an essential phase in the demonstration of Theorem entails examining the small-ball probability associated with quadratic polynomials of Gaussian random variables. The Carbery-Wright theorem represents the foundational principle within this field of study Which says that for $0 < \epsilon < 1$ and any real quadratic polynomial $f = f(Z_1, \ldots, Z_n)$ of independent standard Gaussian random variable $Z_1, \ldots, Z_n$ $\mathcal{N}(0,1)$ we have $\sup_{x \in \mathbb{R}} \Pr[|f - x| \leq \varepsilon] = O\left(\sqrt{\frac{\varepsilon}{\sigma(f)}}\right)$ or any quadratic polynomial of independent standard Gaussian variables and for any small epsilon, the supremum of the probability that the polynomial's deviation from any real value is within epsilon scales optimally with epsilon over the standard deviation of the polynomial Let $Z = (Z_1, \ldots, Z_n)$ be a vector of independent standard Gaussian random variables, and consider a real quadratic polynomial $f(Z) = Z^T F Z + f^T Z + f_0$, where F is a nonzero symmetric matrix in $\mathbb{R}^{n \times n}$, f is a vector in $\mathbb{R}^n$, and f_0 is a real number. Suppose that for some positive constant η, for any symmetric matrix G of rank at most 2, the smallest eigenvalue of $F - G$ in the Frobenius norm is at least η. Then, for any $\epsilon > 0$ and any real number x, there exists a constant C_η, depending only on η, such that $\Pr\left(|f(Z) - x| < \epsilon\right) \leq \frac{C_\eta \epsilon}{\sigma(f(Z))}$, where $\sigma(f(Z))$ denotes the standard deviation of $f(Z)$.

quadratic forms with robust rank 2, such as $Z_1^2 - Z_2^2$, may not conform due to logarithmic scaling in their probability measures as ϵ approaches zero. Furthermore, the theorem is indicative of an inverse theorem, stating that atypical small-ball behavior corresponds to the form's closeness to a low-rank quadratic form, reflecting the insights from the Littlewood-Offord problem. Kane's structure theorem [15] further elaborates that quadratic polynomials of Gaussian variables can be 'decomposed' into parts with standard small-ball behavior, facilitating a simpler analysis of their overall behavior.

Regarding Ramsey graphs adjacency matrices, which are observed to have robustly high rank. This property is related for applying the previous theorem, in particular, a deeper examination into the rank partitioning into submatrices (Lemma 10.1 in the paper). The relationship between graph rank and the absence of large homogeneous sets suggests intriguing implications for communication complexity, specifically in the context

of the log-rank conjecture. Additionally, the research extends to demonstrate that binary matrices, which approximate a low-rank real matrix, closely mirror a low-rank binary matrix (Proposition 10.2 in the paper), presenting potential for broader applications beyond Ramsey graph analysis.

Limitations in precisely determining edge counts $e(G[U])$ in Ramsey graphs via Fourier-analytic estimates, introduces an averaged version of the **switching method.** This method quantifies transitions between outcomes of events, focusing on small perturbations within a random set U. By examining moments of these transition counts and applying the Cauchy-Schwarz inequality, the approach refines probability estimates, potentially extending its application beyond Ramsey graphs.

10 shirel on graph article

[17]

11 Haddas last article

[18]

12 Results

// in the mean time we don't fill this

References

[1] Noga Alon, Michael Krivelevich, and Benny Sudakov. "Induced subgraphs of prescribed size". In: *Journal of Graph Theory* 43.4 (2003), pp. 239–251.

[2] Noga Alon et al. "Edge-statistics on large graphs". In: *Combinatorics, Probability and Computing* 29.2 (2020), pp. 163–189.

[3] Ross Berkowitz. "A local limit theorem for cliques in G (n, p)". In: *arXiv preprint arXiv:1811.03527* (2018).

[4] Ross Berkowitz. "A quantitative local limit theorem for triangles in random graphs". In: *arXiv preprint arXiv:1610.01281* (2016).

[5] Marcelo Campos et al. "Singularity of random symmetric matrices revisited". In: *Proceedings of the American Mathematical Society* 150.7 (2022), pp. 3147–3159.

[6] Kevin P Costello. "Bilinear and quadratic variants on the Littlewood-Offord problem". In: *Israel Journal of Mathematics* 194 (2013), pp. 359–394.

[7] Kevin P Costello, Terence Tao, and Van Vu. "Random symmetric matrices are almost surely nonsingular". In: (2006).

[8] Paul Erdös. "Some remarks on the theory of graphs". In: (1947).

[9] Paul Erdös and George Szekeres. "A combinatorial problem in geometry". In: *Compositio mathematica* 2 (1935), pp. 463–470.

[10] Justin Gilmer and Swastik Kopparty. "A local central limit theorem for triangles in a random graph". In: *Random Structures & Algorithms* 48.4 (2016), pp. 732–750.

[11] Boris Vladimirovich Gnedenko. "On a local limit theorem of the theory of probability". In: *Uspekhi Matematicheskikh Nauk* 3.3 (1948), pp. 187–194.

[12] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. "On the smallest singular value of symmetric random matrices". In: *Combinatorics, Probability and Computing* 31.4 (2022), pp. 662–683.

[13] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. "On the smoothed analysis of the smallest singular value with discrete noise". In: *Bulletin of the London Mathematical Society* 54.2 (2022), pp. 369–388.

[14] Vishesh Jain, Ashwin Sah, and Mehtaab Sawhney. "Singularity of discrete random matrices". In: *Geometric and Functional Analysis* 31 (2021), pp. 1160–1218.

[15] Daniel Kane. "A structure theorem for poorly anticoncentrated polynomials of Gaussians and applications to the study of polynomial threshold functions". In: (2017).

[16] Jeong Han Kim and Van H Vu. *Concentration of multivariate polynomials and its applications*. Vol. 20. 3. Budapest: Akademiai Kiado, 1981-, 2000, pp. 417–434.

[17] Matthew Kwan and Lisa Sauermann. "An algebraic inverse theorem for the quadratic Littlewood-Offord problem, and an application to Ramsey graphs". In: *arXiv preprint arXiv:1909.02089* (2019).

[18] Matthew Kwan and Lisa Sauermann. "Resolution of the quadratic Littlewood–Offord problem". In: *arXiv preprint arXiv:2312.13826* (2023).

[19] Matthew Kwan et al. "Anticoncentration in Ramsey graphs and a proof of the Erdős–McKay conjecture". In: *Forum of Mathematics, Pi*. Vol. 11. 2023, e21.

[20] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. *Noise stability of functions with low influences: invariance and optimality*. 2005, pp. 21–30.

[21] Mark Rudelson and Roman Vershynin. "The Littlewood–Offord problem and invertibility of random matrices". In: *Advances in Mathematics* 218.2 (2008), pp. 600–633.