A SIMPLE ALGORITHM FOR HORN'S PROBLEM AND TWO RESULTS ON DISCREPANCY

By

WILLIAM COLE FRANKS

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Mathematics

Written under the direction of

Michael Saks

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2019

ABSTRACT OF THE DISSERTATION

A simple algorithm for Horn's problem and two results on discrepancy

By WILLIAM COLE FRANKS

Dissertation Director: Michael Saks

In the second chapter we consider the discrepancy of permutation families. A k–permutation family on n vertices is a set-system consisting of the intervals of k permutations of $[n]$. Both 1– and 2–permutation families have discrepancy 1, that is, one can color the vertices red and blue such that the number of reds and blues in each edge differs by at most one. That is, their discrepancy is bounded by one. Beck conjectured that the discrepancy of for 3–permutation families is also $O(1)$, but Newman and Nikolov disproved this conjecture in 2011. We give a simpler proof that Newman and Nikolov's sequence of 3-permutation families has discrepancy at least logarithmic in n. We also exhibit a new, tight lower bound for the related notion of root-mean-squared discrepancy of a 6–permutation family, and show new upper bounds on the root–mean–squared discrepancy of the union of set–systems.

In the third chapter we study the discrepancy of random matrices with m rows and $n \gg m$ independent columns drawn from a bounded lattice random variable, a model motivated by the Komlós conjecture. We prove that, with high probability, the discrepancy is at most twice the ℓ_∞-covering radius of the lattice. As a consequence, the discrepancy of a $m \times n$ random t-sparse matrix is at most 1 with high probability for $n \geq m^3 \log^2 m$, an exponential improvement over Ezra and Lovett (Ezra and Lovett,

Approx+Random, 2015). More generally, we show polynomial bounds on the size of n required for the discrepancy to become at most twice the covering radius of the lattice with high probability.

In the fourth chapter, we obtain a simple algorithm to solve a class of linear algebraic problems. This class includes Horn's problem, the problem of finding Hermitian matrices that sum to zero with prescribed spectra. Other problems in this class arise in algebraic complexity, analysis, communication complexity, and quantum information theory. Our algorithm generalizes the work of (Garg et. al., 2016) and (Gurvits, 2004).

Acknowledgements

I sincerely thank my thesis committee, Michael Saks, Leonid Gurvits, Jeff Kahn, and Shubhangi Saraf, for their help and support.

I was extremely lucky to be advised by Michael Saks. When I first met Mike, I was not fully aware of his reputation for fundamental work on, quote, "so many problems;" this lasted until several people mistakenly assumed that I, too, must be familiar with a positive fraction of these topics. I am glad Mike has helped get me up to speed, but I am equally grateful for his enthusiastic problem-solving style, his limitless curiosity, and his ability to lead through selfless, open-minded encouragement.

I am further indebted to my advisor for providing a number of funding, workshop, and internship opportunities. I was fortunate to be a research intern under Navin Goyal at Microsoft Research India. It was Navin who set me on the path that led to the final chapter of this thesis. I was also lucky to spend some weeks at the Centrum Wiskunde & Informatica, made possible and enjoyable by Michael Walter. The work on which this thesis is based has been partially supported by a United States Department of Defense NDSEG fellowship, a Rutgers School of Arts and Sciences fellowship, and by Simons Foundation award 332622.

My collaborators showed me how research is meant to be done. I have been lucky to work alongside Michael Saks, Sasho Nikolov, Navin Goyal, Aditya Potokuchi, Rafael Oliveira, Avi Wigderson, Ankit Garg, Peter Bürgisser, and Michael Walter. Michael Saks and Sasho Nikolov introduced me to the problem addressed in the second chapter of this thesis, and the work benefitted greatly from discussions with them as well as Shachar Lovett. The third chapter of this thesis is based on the joint work [FS18] with Michael Saks, and has also benefitted from discussions with Aditya Potokuchi. Working with Rafael Oliveira, Avi Wigderson, Ankit Garg, Peter Bürgisser, and Michael Walter

on a the follow-up paper [BFG$^+$18] was very helpful in writing the the fourth and final chapter of this thesis, which is based on [Fra18a] (which also appeared in the conference proceedings [Fra18b]). I would also like to thank Akshay Ramachandran, Christopher Woodward, Shachar Lovett, Shravas Rao, Nikhil Srivastava, Thomas Vidick, and Leonid Gurvits for enlightening discussions.

I would like to thank Rutgers for providing a stimulating learning environment, and my cohort and elders in the mathematics department for their much-needed support. I am eternally grateful to Daniel and Sriranjani for sitting in my corner and helping me keep things in perspective.

Dedication

For Mary Jo, Wade, and Max

Table of Contents

List of Figures

Chapter 1

Introduction

This thesis contains two results in combinatorial matrix theory, and one algorithmic result in linear algebra.

1.1 Balancing stacks of balanceable matrices

The discrepancy of an $m \times n$ matrix M is the degree to which the rows of the matrix can be simultaneously balanced by splitting the columns into two groups; formally,

$$\operatorname{disc}_\infty(M) = \min_{x \in \{+1,-1\}^n} \|Mx\|_\infty.$$

We study how the discrepancy of matrices can grow when placed one atop the other. That is, how the discrepancy of

$$M = \begin{bmatrix} -M_1- \\ -M_2- \\ \vdots \\ -M_k- \end{bmatrix}$$

compares with $\operatorname{disc}_\infty(M_1), \operatorname{disc}_\infty(M_2), \ldots, \operatorname{disc}_\infty(M_k)$. The discrepancy of M can be much larger, but if we instead consider the more robust *hereditary discrepancy*, there is a meaningful relationship. Define $\operatorname{herdisc}_\infty(A)$ to be the maximum discrepancy of any subset of columns of A. Matousek [Mat13] showed that if M is $m \times n$, then

$$\operatorname{herdisc}_\infty(M) = O\left(\sqrt{k}(\log^{3/2} m) \max_i \operatorname{herdisc}_\infty(M_i)\right).$$

Improving Matousek's bound is an interesting open problem. Lower bounds so far are consistent with the following:

Conjecture 1.1 (Matousek [Mat13])**.**

$$\operatorname{herdisc}_\infty(M) = O\left(\sqrt{k}(\log m) \cdot \max_i \operatorname{herdisc}_\infty(M_i)\right) \tag{1.1}$$

To the author's knowledge, it is not known that Eq. (1.1) would be tight. It is known that the factors $\sqrt{k}$ and $\log m$ are individually necessary, but there seems to be no lower bound with a factor more than $f(k)\log m$ where $f(k) = \omega(1)$. We conjecture that $f(k) = \sqrt{k}$.

Conjecture 1.2 (Matousek's conjecture is tight even for constant k)**.** *There is a constant $C > 0$ such that, for every $k \in \mathbb{N}$, there is a family of tuples of matrices $(M_1, \dots, M_k)$ with*

$$\operatorname{herdisc}_\infty(M) \geq C\sqrt{k}(\log m) \cdot \max_i \operatorname{herdisc}_\infty(M_i).$$

One interesting example are *permutation families* in which M_i are each column-permuted copies of following upper-triangular matrix:

$$P = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Because M_i is the incidence matrix of the intervals of some permutation, σ, i.e. the set-system $\{\{\}, \{\sigma(1)\}, \{\sigma(1), \sigma(2)\}, \dots\}$, M is called a k–permutation family. By choosing alternating signs for the columns of P one can immediately see that $\operatorname{disc}_\infty(M_i) = \operatorname{disc}_\infty(P)$ is 1, and in fact $\operatorname{herdisc}_\infty(M_i) = 1$ for $1 \leq i \leq k$. Though less trivial, the hereditary discrepancy of 2-permutation families is also 1. This suggests that the discrepancy of 3–permutation families should be small. However, disproving a conjecture of Beck, Newman and Nikolov [NN11] showed that the discrepancy of 3–permutation families can be $\Omega(\log n)$, providing another[1] example showing that a $\log m$ factor[2] in Matousek's bound is necessary. We provide, using only straightforward calculations, a

[1] The first such example is due to Pálvölgyi [Mat13].

[2] in 3-permutation families, $m = \theta(n)$, so $\log n = \log m + O(1)$.

simpler proof of their lower bound on the same counterexample. The proof technique extends naturally to higher values of k, and we hope to use it to show the following result towards Conjecture 1.2.[3]

Conjecture 1.3 (Michael Saks)**.** *There is a function $f \in \omega(1)$ and a family of k-permutation families M_n on n vertices such that*

$$\operatorname{disc}_\infty(M_n) \geq f(k) \log n$$

for all $n, k \in \mathbb{N}$.

Our analysis also yields a new result for the root–mean–squared discrepancy disc_2 of permutation families, where

$$\operatorname{disc}_2(M) := \frac{1}{\sqrt{m}} \min_{x \in \to \{\pm 1\}^n} \|Mx\|.$$

Theorem 1.4. *There is a sequence of 6–permutation families on n vertices with root–mean–squared discrepancy $\Omega(\sqrt{\log n})$.*

Define the *hereditary root–mean–squared discrepancy*, denoted $\operatorname{herdisc}_2(M)$, to be the largest root–mean–squared discrepancy of any subset of columns of M. The paper [Lar17] exhibits a lower bound for $\operatorname{herdisc}_2$ which approximates the hereditary root–mean–squared discrepancy to within a $\sqrt{\log n}$ factor, but this quantity is constant on families of constantly many permutations. Thus, Theorem 1.4 shows that the $\sqrt{\log n}$ approximation factor between $\operatorname{herdisc}_2(M)$ and the lower bound in [Lar17] is best possible. As a side–benefit, we also observe that the lower bound from [Lar17] behaves well under unions, which implies the following:

Theorem 1.5. $\operatorname{herdisc}_2(M) = O\left(k\sqrt{\log n}\ \max_{i \in [k]}\ \operatorname{herdisc}_2(M_i)\right)$.

It is an interesting open direction to improve k to $\sqrt{k}$ in Theorem 1.5.

1.2 Balancing random wide matrices

[*A paper joint with Michael Saks*] We continue our study of discrepancy, shifting our focus to random matrices. It is not hard to see that, if the columns of M are chosen

[3] See Footnote 2

independently at random from some distribution on $\mathbb{R}^n$, the discrepancy is asymptotically constant as n grows while m is left fixed. Under very minimal assumptions on the column distribution, we study how large n must be as a function of m for this behavior to take place. For example, if M is a *random t-sparse matrix*, i.e. the columns are random vectors with t ones and $m-t$ zeroes, it was known that $n = \Omega(\binom{m}{t}\log\binom{m}{t})$ suffices [EL16]. We improve the bound on n to a polynomial in m.

Theorem 1.6. *Let M be a random t-sparse matrix for $0 < t < cm/2$ for $c = 1$. For $n = \Omega(m^3 \log m)$, then*

$$\operatorname{disc}(M) = 1$$

with probability $1 - 2^{-\Omega(m)} - O(\sqrt{m/n}\log m)$.

We explicitly describe the distribution of $\operatorname{disc}_\infty(M)$ for large n, which asymptotically depends on n only in that it depends on the parity of n, and has support contained in $\{0, 1, 2\}$. The main technical ingredient in our proof is a local central limit theorem for random signed sums of the columns of M which holds with high probability over the choice of M. The main open problem remaining is to improve bounds on the value of n. We conjecture that $n = \theta(m \log m)$ is the true answer.

For distributions on arbitrary lattices $\mathcal{L}$, we show that the discrepancy becomes at most twice the ℓ_∞–covering radius $\rho_\infty(\mathcal{L})$ of the lattice when the number of columns is at least polynomial in the number of rows and a few natural parameters of the distribution. Ideally, this would allow one to conclude that the discrepancy becomes constant when the columns are drawn from a distribution of unit vectors, but this would require the following open conjecture:

Conjecture 1.7. *There is an absolute constant C such that for any lattice $\mathcal{L}$ generated by unit vectors, $\rho_\infty(\mathcal{L}) \le C$.*

1.3 Sums of Hermitian matrices and related problems

Culminating a long line of work, Knutson and Tao [KT00] proved *Horn's conjecture*, which posits that the answer to the following problem is a polyhedral cone with facets given by a certain recursively defined set of linear inequalities.

Problem 1.8 (Horn's problem)**.** *What is the set of spectra of $m \times m$ Hermitian matrices A, B, C satisfying*

$$A + B = C?$$

We consider the problem of *finding* the matrices A, B, and C from Problem 1.8.

Problem 1.9 (Constructive variant of Horn's problem)**.** *Given three nonincreasing sequences $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ of m real numbers, construct (if they exist) $m \times m$ Hermitian matrices A, B, C with spectra $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ satisfying*

$$A + B = C. \tag{1.2}$$

We give a simple, iterative algorithm for solving Problem 1.9 to *arbitrary precision* ε. This solves Problem 1.9 in the sense that the sequence of outputs of the algorithm as $\varepsilon \to 0$ have a subsequence converging to a solution of Problem 1.9. Informally, the algorithm proceeds as follows:[4]

1. Define $C = \operatorname{diag}(\boldsymbol{\gamma})$. Choose random real matrices U_A, U_B, and define

$$A = U_A \operatorname{diag}(\boldsymbol{\alpha}) U_A^\dagger \text{ and } B = U_B \operatorname{diag}(\boldsymbol{\beta}) U_B^\dagger.$$

2. Alternately repeat the following steps until Eq. (4.1) holds to the desired precision:

   (a) Enforce $A + B = C$ by simultaneous left–multiplication of U_A, U_B by the same lower-triangular matrix.

   (b) Make U_A orthogonal by right–multiplication with an upper-triangular matrix. Do the same for U_B.

3. Output A, B, C.

The above algorithm is an example of *alternating minimization*, a primitive in optimization. Specializations of our algorithm also solve the below problems, though the specialization of our algorithm to Problem 1.10 was discovered by Sinkhorn long ago [Sin64], and the original characterization of Problem 1.11 was algorithmic [Hor54].

[4] As written, the algorithm requires $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ to be positive. This is without loss of generality by a simple (and standard) rescaling argument.

Problem 1.10 (Matrix scaling)**.** *Given a nonnegative matrix A and nonnegative row- and column-sum vectors* $\mathbf{r}$ *and* $\mathbf{c}$*, construct (if they exist) nonnegative diagonal matrices X, Y such that the row and column sums of $A' = XAY$ are, respectively,* $\mathbf{r}$ *and* $\mathbf{c}$ *(if possible).*

Problem 1.11 (Schur–Horn)**.** *Given vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^m$, construct (if it exists) a symmetric matrix with spectrum $\boldsymbol{\beta}$ and with $\boldsymbol{\alpha}$ as its main diagonal.*

Problem 1.12 (Quantum channels)**.** *Given a completely positive map T and mixed quantum states ρ_1, ρ_2, construct (if they exist) invertible linear maps g, h such that $T' : X \mapsto g^\dagger T(hXh^\dagger)g$ is a quantum channel sending ρ_1 to ρ_2.*

The analysis of the algorithm proceeds by "lifting" the steps of the algorithm to a larger instance for which the algorithm is known to converge by the work of Gurvits [Gur04]. The lifting map is similar to a trick used by Derksen and Wayman [DW00] to provide an alternate proof of Horn's conjecture using quiver representations.

Problems 1.8 to 1.12 are more than linear algebraic curiosities: they have a deep relationship with representation theory. For group actions on vector spaces, invariant varieties such as closures of group orbits are central objects of study. It is often fruitful to study such varieties through the representation theory of their coordinate rings. For instance, the *geometric complexity theory* approach to lower bounds in algebraic complexity relies on a strategy to show a certain orbit closure is not contained in another by proving that there is an irreducible representation occuring in one coordinate ring but not the other. Mumford showed that asymptotic information about the representation theory of the coordinate rings is encoded in a convex polytope known as the *moment polytope* [NM84]. For example, the set described in Problem 1.11 is succinctly described as a moment polytope.

The complexity of testing moment polytope membership is open, except in very special cases such as Horn's problem. Our central open problem is to improve our algorithms so that they can test membership in the moment polytope in polynomial time.

Another interesting avenue for the future concerns approximation algorithms and

Van der Waerden–type conjectures, which were the original motivations of matrix scaling. Among many other examples, scaling problems arose in proofs of Van der Waerden-like conjectures [Gur08], and in [LSW98] it was observed that polynomial time algorithms for matrix scaling result in deterministic algorithms to approximate the permanent to a singly exponential factor. The permanent can be viewed as a sum of squares of evaluations of invariant polynomials of the action of an Abelian group, but there has been no approximation algorithm or Van der Waerden type theorem for the analogous quantities in the *non-Abelian* case. This reflects the lack (to the author's knowledge) of noncommutative analogues of hyperbolic and log–concave polynomials.

1.4 Common notation

Here are a few conventions followed throughout the chapters:

- Unless otherwise specified, $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{C}^n$ or $\mathbb{R}^n$. The corresponding norm is written $\| \cdot \|$.
- On complex or real matrices, $\| \cdot \|$ denotes the Frobenius norm, i.e. $\|M\|^2 = \operatorname{tr} M^\dagger M$, where $M^\dagger$ denotes the conjugate transpose of M. $\|M\|_2$ denotes the spectral norm of M.
- The symbol $\succeq$ denotes the Loewner ordering on matrices, i.e. $A \succeq B$ if $A - B$ is positive-semidefinite. $A \succ B$ if $A - B$ is positive-definite.
- For $m \in \mathbb{N}$, we denote by I_m the $m \times m$ identity matrix.
- Bold letters such as $\boldsymbol{x}$ indicate vectors, or more generally tuples of objects. Random vectors are denoted by capital letters X. The i^{th} entry of the tuple $\boldsymbol{x}$ will be denoted x_i.
- If S is a set in a universe U, $\mathbf{1}_S$ denotes the characteristic vector of S in $\mathbb{R}^U$.
- For $n \in \mathbb{N}$, we let $[n]$ denote the set $\{1, 2, \ldots, n\}$.

Chapter 2

Discrepancy of permutation families

This chapter is based on the work [Fra18c] of the author.

2.1 Introduction

The discrepancy of a set-system is the extent to which the sets in a set-system can be simultaneously split into two equal parts, or two–colored in a balanced way. Let $\mathcal{A}$ be a collection (possibly with multiplicity) of subsets of a finite set Ω. The discrepancy of a two–coloring $\chi : \Omega \to \{\pm 1\}$ of the set-system $(\Omega, \mathcal{A})$ is the maximum imbalance in color over all sets S in $\mathcal{A}$. The discrepancy of $(\Omega, \mathcal{A})$ is the minimum discrepancy of any two–coloring of Ω. Formally,

$$\operatorname{disc}_\infty(\Omega, \mathcal{A}) := \min_{\chi:\Omega\to\{+1,-1\}} \operatorname{disc}_\infty(\chi, \mathcal{A}), \tag{2.1}$$

where $\operatorname{disc}_\infty(\chi, \mathcal{A}) = \max_{S\in\mathcal{A}} |\chi(S)|$ and $\chi(S) = \sum_{x\in S} \chi(x)$.

A central goal of the study of discrepancy is to bound the discrepancy of set-systems with restrictions or additional structure. Here we will be concerned with set-systems constructed from permutations. A permutation $\sigma : \Omega \to \Omega$ from a set Ω with a total ordering $\leq$ to itself determines the set-system $(\Omega, \mathcal{A}_\sigma)$ where

$$\mathcal{A}_\sigma = \{\{i : \sigma(i) \leq \sigma(j)\} : j \in \Omega\} \cup \{\emptyset\}.$$

For example, if $[3]$ inherits the usual ordering on natural numbers and $e : [3] \to [3]$ is the identity permutation, then $\mathcal{A}_e = \{\emptyset, \{1\}, \{1, 2\}, \{1, 2, 3\}\}$. Equivalently, $\mathcal{A}_\sigma$ is a maximal chain in the poset $2^{[n]}$ ordered by inclusion. If $P = \{\sigma_1, \ldots, \sigma_k\}$ is a set of permutations of Ω, let $\mathcal{A}^P = \mathcal{A}_{\sigma_1} + \cdots + \mathcal{A}_{\sigma_k}$ where $+$ denotes multiset sum (union with multiplicity). Then we say $(\Omega, \mathcal{A}^P)$ is a *k-permutation family*.

By Dilworth's theorem, the maximal discrepancy of a k-permutation family is the same as the maximal discrepancy of a set-system of *width* k, that is, a set-system that contains no antichain of cardinality more than k.

It is easy to see that a 1–permutation family has discrepancy at most 1, and the same is true for 2–permutation families [Spe87]. Beck conjectured that the discrepancy of a 3–permutation family is $O(1)$. More generally, Spencer, Srinivasan and Tetali conjectured that the discrepancy of a k-permutation family is $O(\sqrt{k})$ [JS]. Both conjectures were recently disproven by Newman and Nikolov [NN11]. They showed the following:

Theorem 2.1 ([NN11])**.** *There is a sequence of* 3*–permutation families on* n *vertices with discrepancy* $\Omega(\log n)$*.*

The same authors, together with Neiman, showed in [NNN12] the above lower bound implies a natural class of rounding schemes for the Gilmore–Gomory linear programming relaxation of bin–packing, such as the scheme used in the Kamarkar-Karp algorithm, incur logarithmic error.

Spencer, Srinivasan and Tetali proved an upper bound that matches the lower bound of Newman and Nikolov for $k = 3$.

Theorem 2.2 ([JS])**.** *The discrepancy of a* k*–permutation family on* n *vertices is* $O(\sqrt{k}\log n)$*.*

They showed that the upper bound is tight for for $k \geq n$. However, it is open whether this upper bound is tight for $3 < k = o(n)$. In fact, no one has proved lower bounds with logarithmic dependency on n that grow a function of k.

Conjecture 1.3 (Michael Saks)**.** *There is a function* $f \in \omega(1)$ *and a family of* k*-permutation families* M_n *on* n *vertices such that*

$$\operatorname{disc}_\infty(M_n) \geq f(k)\log n$$

for all $n, k \in \mathbb{N}$*.*

In this chapter, we present a new analysis of the counterexample due to Newman and Nikolov. We replace their case analysis by a simple argument using norms of

matrices, albeit achieving a worse constant ($(4\sqrt{6})^{-1}\log_3 n$ vs their $3^{-1}\log_3 n$). Our analysis generalizes well to larger permutation families, and can hopefully be extended to handle Conjecture 1.3. Our analysis also yields a new result for the root–mean–squared discrepancy, defined as

$$\operatorname{disc}_2(\Omega, \mathcal{A}) = \min_{\chi:[n]\to\{\pm 1\}} \operatorname{disc}_2(\mathcal{A}, \chi)$$

where $\operatorname{disc}_2(\mathcal{A}, \chi) = \sqrt{\frac{1}{|\mathcal{A}|}\sum_{S\in\mathcal{A}} |\chi(S)|^2}$. Define the *hereditary root–mean–squared discrepancy* by

$$\operatorname{herdisc}_2(\Omega, \mathcal{A}) = \max_{\Gamma\subset\Omega} \operatorname{disc}_2(\Gamma, \mathcal{A}|_\Gamma).$$

Theorem 1.4. *There is a sequence of* 6*–permutation families on* n *vertices with root–mean–squared discrepancy* $\Omega(\sqrt{\log n})$.

For $k = 6$, Theorem 1.4 matches the upper bound of $\sqrt{k \log n}$ for the root–mean–squared discrepancy implied by the proof of Theorem 2.2 in [JS]. Further, the lower bound implied by [Lar17] for the hereditary root–mean–squared discrepancy is constant for families of constantly many permutations. This fact was communicated to the author by Aleksandar Nikolov; we provide a proof in Section 2.3 for completeness. The lower bound in [Lar17] is smaller than $\operatorname{herdisc}_2(\Omega, \mathcal{A})$ by a factor of at most $\sqrt{\log n}$, so Theorem 1.4 shows that the $\sqrt{\log n}$ gap between $\operatorname{herdisc}_2(\Omega, \mathcal{A})$ and the lower bound in [Lar17] is best possible.

Remark 2.3 (Odd discrepancy)**.** One can relax to *odd integer* assignments and our lower bounds still hold: define the *odd discrepancy* of $\mathcal{A}$ by the smaller quantity

$$\operatorname{oddisc}_\infty(\Omega, \mathcal{A}) = \min_{\{\chi:\Omega\to(2\mathbb{Z}-1)\}} \operatorname{disc}_\infty(\chi, \mathcal{A}),$$

and define $\operatorname{oddisc}_2(\Omega, \mathcal{A})$ analogously. The stronger versions of Theorem 1.4 and Theorem 2.1 with, respectively, $\operatorname{disc}_\infty$ replaced by $\operatorname{oddisc}_\infty$ and disc_2 replaced by oddisc_2 hold.

Acknowledgements

The author would like to thank Michael Saks and Aleksandar Nikolov for many interesting discussions. The author also thanks Aleksandar Nikolov and Shachar Lovett for

suggesting the application of this argument to the root–mean–squared discrepancy, and especially to Aleksandar Nikolov for communicating Observation 2.22 and suggesting the connection with [Lar17], [NTZ13], and [Mat13].

2.2 The set-system of Newman and Nikolov

Our proof of Theorem 2.1 uses the same set-system as Newman and Nikolov. For completeness, we define and slightly generalize the system here. The vertices of the system will be r–ary strings, or elements of $[r]^d$. For Newman and Nikolov's set-system, $r = 3$. We first set our notation for referring to strings.

Definition 2.4 (String notation)**.**

- Bold letters, e.g. $\boldsymbol{a}$, denote strings in $[r]^d$ for some $d \geq 0$. Here $[r]^0$ denotes the set containing only the empty string $\boldsymbol{\varepsilon}$.
- If $\boldsymbol{a} = a_1 \dots a_d \in [r]^d$ is a string, for $0 \leq k \leq d$ let $\boldsymbol{a}[k]$ denote the string $a_1 \dots a_k$, with $\boldsymbol{a}[0] := \boldsymbol{\varepsilon}$.
- If $\boldsymbol{a}$ is a string, $|\boldsymbol{a}|$ denotes the length of $\boldsymbol{a}$.
- If $\boldsymbol{a}$ and $\boldsymbol{b}$ are strings, their concatentation in $[r]^{|\boldsymbol{a}|+|\boldsymbol{b}|}$ is denoted $\boldsymbol{ab}$.
- If $j \in [r]$, then $\overline{j}$ denotes the all j's string of length d; e.g. $\overline{3} := \underbrace{33..3}_{d}$.
- τ denotes the permutation of $[r]$ given by $\tau(i) = r - i + 1$, the permutation reversing the ordering on $[r]$.

We may now define the set-system of Newman and Nikolov.

Definition 2.5 (The set-system $([r]^d, \mathcal{A}_P)$)**.** Let $<$ be the lexicographical ordering on $[r]^d$. Given a permutation σ of $[r]$, we define a permutation $\boldsymbol{\sigma}$ of $[r]^d$ by acting digitwise by σ. Namely, $\sigma(\boldsymbol{a}) := \sigma(a_1)\sigma(a_2)\dots\sigma(a_d)$. For any subset $P \subset S_r$ of permutations of $[r]$, define the permutation family $\mathcal{A}_P$ by

$$\mathcal{A}_P = \mathcal{A}^{\{\boldsymbol{\sigma}:\sigma\in P\}}.$$

Namely, the edges of $\mathcal{A}_P$ are $\emptyset$ and the sets $\leq_\sigma \boldsymbol{a}$ defined by

$$\leq_\sigma \boldsymbol{a} := \{\boldsymbol{b} \in [r]^{|\boldsymbol{a}|} : \boldsymbol{\sigma}(\boldsymbol{b}) \leq \boldsymbol{\sigma}(\boldsymbol{a})\}$$

as σ ranges over P and $\boldsymbol{a}$ over $[r]^d$. Note that for each $\sigma \in P$ and $\boldsymbol{a} \in [r]^d$, $\boldsymbol{A}_P$ also contains the edges defined not to include $\boldsymbol{a}$:

$$<_\sigma \boldsymbol{a} := \{\boldsymbol{b} \in [r]^{|\boldsymbol{a}|} : \boldsymbol{\sigma}(\boldsymbol{b}) < \boldsymbol{\sigma}(\boldsymbol{a}).$$

Definition 2.6 (The set-system of Newman and Nikolov)**.** The system of Newman and Nikolov is $([3]^d, \mathcal{A}_C)$ with $C = \{e, (1,2,3), (1,3,2)\}$. That is, C is the cyclic permutations of 3.

We first bound the discrepancy of the 6-permutation family $([3]^d, \mathcal{A}_{S_3})$. In fact, we bound the smaller *odd* discrepancy of this family.

Proposition 2.7 (Discrepancy lower bound)**.** *If $r \geq 3$ is odd, then*

$$\operatorname{disc}_\infty([r]^d, \mathcal{A}_{S_r}) \geq \operatorname{oddisc}_\infty([r]^d, \mathcal{A}_{S_r}) \geq \frac{d}{2\sqrt{6}}.$$

Proposition 2.7 is proved in the next section, Section 2.2.1. We bound the discrepancy of $\mathcal{A}_C$ in terms of the discrepancy of $\mathcal{A}_{S_3}$. Theorem 2.1 follows immediately from Observation 2.8 and Proposition 2.7 for $r = 3$.

Observation 2.8 ([NN11])**.**

$$\operatorname{disc}_\infty([3]^d, \mathcal{A}_C) \geq \operatorname{oddisc}_\infty([3]^d, \mathcal{A}_C) \geq \frac{1}{2} \operatorname{oddisc}_\infty([3]^d, \mathcal{A}_{S_3}).$$

Proof. The key observation is that $[3]^d$ is in reverse order under the action of σ and $\tau \circ \sigma$, so for each $\sigma, \boldsymbol{a}$, the edges $\leq_\sigma \boldsymbol{a}$ and $<_{\tau\circ\sigma} \boldsymbol{a}$ partition $[3]^d$.

Let $\chi : [3]^d \to 2\mathbb{Z} - 1$ be an assignment of minimal discrepancy K to $([3]^d, \mathcal{A}_C)$. Let $\sigma \in S_3, \boldsymbol{a} \in [r]^d$ be arbitrary. It suffices to show $|\chi(\leq_\sigma \boldsymbol{a})| \leq 2K$. By the preceding reasoning, $|\chi(\leq_\sigma \boldsymbol{a})| = |\chi([3]^d) - \chi(<_{\tau\circ\sigma} \boldsymbol{a})| \leq 2K$. □

We recall a few facts about the set-system of Newman and Nikolov. Define

$$<_\sigma \boldsymbol{a} = \{\boldsymbol{b} \in [r]^{|\boldsymbol{a}|} : \sigma \cdot \boldsymbol{b} < \sigma \cdot \boldsymbol{a}\}.$$

As observed in [NN11], the quantity $\chi(<_\sigma \boldsymbol{a})$ behaves additively under concatentation of strings. That is, if $\boldsymbol{a} = \boldsymbol{bc}$, then there is a natural way to define colorings $\chi^{\boldsymbol{\varepsilon}}$ and $\chi^{\boldsymbol{b}}$ of $[r]^{|\boldsymbol{b}|}$ and $[r]^{|\boldsymbol{c}|}$, respectively, such that $\chi(<_\sigma \boldsymbol{a}) = \chi^{\boldsymbol{\varepsilon}}(<_\sigma \boldsymbol{b}) + \chi^{\boldsymbol{b}}(<_\sigma \boldsymbol{c})$. Further, this holds even if χ is a coloring of $[r]^d$ with any odd integers (rather than just ± 1).

Definition 2.9 (Extension of colorings)**.** We extend each assignment $\chi : [r]^d \to 2\mathbb{Z} - 1$ to

$$\chi : [r]^0 \cup [r]^1 \cup \cdots \cup [r]^d \to 2\mathbb{Z} - 1$$

by defining $\chi(\boldsymbol{a}) = \sum_{|\boldsymbol{b}|=d-|\boldsymbol{a}|} \chi(\boldsymbol{ab})$ for $|\boldsymbol{a}| \le d$. Crucially, χ the entries of χ are indeed odd. Observe that

$$\chi(\boldsymbol{a}) = \sum_{i \in [r]} \chi(\boldsymbol{a}i) \tag{2.2}$$

for $|\boldsymbol{a}| < d$. For $|\boldsymbol{a}| \le d$, define

$$\chi^{\boldsymbol{a}} : [r]^0 \cup [r]^1 \cup \cdots \cup [r]^{d-|\boldsymbol{a}|}$$

by $\chi^{\boldsymbol{a}}(\boldsymbol{b}) = \chi(\boldsymbol{ab})$ for $|\boldsymbol{b}| \le d - |\boldsymbol{a}|$. In particular, $\chi^{\boldsymbol{\varepsilon}}$ and χ are equal as functions on $[r]^0 \cup [r]^1 \cup \cdots \cup [r]^d$.

Observation 2.10 (Additivity of discrepancy)**.** *For any assignment $\chi : [r]^d \to 2\mathbb{Z} - 1$ and $\boldsymbol{a} = \boldsymbol{bc}$ with $|\boldsymbol{a}| \le d$, we have*

$$\chi(<_\sigma \boldsymbol{a}) = \chi^{\boldsymbol{\varepsilon}}(<_\sigma \boldsymbol{b}) + \chi^{\boldsymbol{b}}(<_\sigma \boldsymbol{c}) \tag{2.3}$$

Proof. If $|\boldsymbol{b}'| = |\boldsymbol{b}|$ and $|\boldsymbol{c}| = |\boldsymbol{c}'|$, then $\boldsymbol{b}'\boldsymbol{c}'$ is in $\le_\sigma \boldsymbol{bc}$ if and only if $\sigma \cdot \boldsymbol{b}' < \sigma \cdot \boldsymbol{b}$ or $\boldsymbol{b}' = \boldsymbol{b}$ and $\sigma \cdot \boldsymbol{c}' < \sigma \cdot \boldsymbol{c}$. Thus,

$$\chi(\le_\sigma \boldsymbol{bc}) = \sum_{\sigma \cdot \boldsymbol{b}' < \sigma \cdot \boldsymbol{b}} \sum_{|\boldsymbol{c}'|=|\boldsymbol{c}|} \chi(\boldsymbol{b}'\boldsymbol{c}') + \sum_{\sigma \cdot \boldsymbol{c}' < \sigma \cdot \boldsymbol{c}} \chi(\boldsymbol{bc}').$$

The right-hand-side is precisely $\chi^{\boldsymbol{\varepsilon}}(<_\sigma \boldsymbol{b}) + \chi^{\boldsymbol{b}}(<_\sigma \boldsymbol{c})$. □

2.2.1 Proof of the lower bound

In this section we prove Proposition 2.7, the lower bound on $\mathrm{oddisc}_\infty([r]^d, \mathcal{A}_{S_r})$.

Proof of Proposition 2.7. To show the discrepancy $\mathrm{oddisc}_\infty([r]^d, \mathcal{A}_{S_r})$ is at least K, it is enough to show that given an assignment $\chi : [r]^d \to 2\mathbb{Z} - 1$, we can choose $\sigma \in S_r$ and $\boldsymbol{a} \in [r]^d$ so that $|\chi(<_\sigma \boldsymbol{a})|$ is at least K.

We do this in two steps. First, define some vector $M_\chi(\boldsymbol{a})$ depending on χ and the choice of $\boldsymbol{a}$ such that if $\|M_\chi(\boldsymbol{a})\| \geq K$, then there is a σ with $|\chi(<_\sigma \boldsymbol{a})| \geq K$. Next, we choose $\boldsymbol{a}$ to maximize $\|M_\chi(\boldsymbol{a})\|$. The correct object M_χ turns out to be an $r \times r$ matrix valued function of $\boldsymbol{a}$, and we will measure its size using a certain seminorm $\|\cdot\|_{S_r}$. We will define the two such that for any $\boldsymbol{a} \in [r]^d$,

$$\max_\sigma |\chi(<_\sigma \boldsymbol{a})| = \|M_\chi(\boldsymbol{a})\|_{S_r}.$$

Definition 2.11 (The seminorm $\|\cdot\|_{S_r}$)**.** For $M \in \mathrm{Mat}_{r\times r}(\mathbb{R})$ and $\sigma \in S_r$, define $\sigma \cdot M := \sum_{i,j\in[r],\, \sigma(i)>\sigma(j)} M_{i,j}$. Now let

$$\|M\|_{S_r} = \max_{\sigma \in S_r} |\sigma \cdot M|\,.$$

Remark 2.12. This seminorm is well–studied; if M is the $0,1$ adjacency matrix of a directed graph G, then $\|M\|_{S_r}$ is the maximum size of an acyclic subgraph of G. In [GMR08] it is shown that, assuming the unique games conjecture, $\|M\|_{S_r}$ is **NP**–hard to approximate even for M antisymmetric.

We will define M such that, in particular,

$$\chi(<_\sigma \boldsymbol{a}) = \sigma \cdot M_\chi(\boldsymbol{a}). \tag{2.4}$$

Recall how we extended χ in Definition 2.9. If we define M_χ to be an *additive* function on $[r]^0 \cup [r]^1 \cup \cdots \cup [r]^d$, i.e.

$$M_\chi(\boldsymbol{ab}) = M_{\chi^{\boldsymbol{\varepsilon}}}(\boldsymbol{a}) + M_{\chi^{\boldsymbol{a}}}(\boldsymbol{b}), \tag{2.5}$$

then by linearity of $\sigma \cdot M$ in M we only need check that Eq. (2.4) holds for $r = 1$. This motivates our definition of M_χ.

Definition 2.13 (The matrix $M_\chi(\boldsymbol{a})$)**.** Let $\chi : [r]^d \to 2\mathbb{Z} - 1$. For $d = 0$, define $M_\chi(\boldsymbol{\varepsilon}) = 0$. For $a \in [r]$, i.e. $d = 1$, define $M_\chi(a)$ to be the $r \times r$ matrix with only the

a'th row nonzero, and the entries of this row given by $\chi(1), \chi(2) \ldots, \chi(r)$. Equivalently,

$$M_\chi(a)_{i,j} = \delta_{i,a}\chi(j) \text{ for } a \in [r]. \tag{2.6}$$

For $d > 1$, define

$$M_\chi(\boldsymbol{a}) = \sum_{k=1}^{|\boldsymbol{a}|} M_{\chi^{\boldsymbol{a}[k-1]}}(a_k). \tag{2.7}$$

Note that $(\boldsymbol{a}, \chi) \mapsto M_\chi(\boldsymbol{a})$ is unique $r \times r$ matrix–valued function on

$$([r]^0 \cup [r]^1 \cup \cdots \cup [r]^d) \times (2\mathbb{Z} - 1)^{[r]^d}$$

satisfying $M_\chi(\boldsymbol{\varepsilon}) = 0$, Eq. (2.6), and Eq. (2.5).

We now prove that this matrix and seminorm have the promised property.

Claim 2.14. *For all $\chi : [r]^d \to 2\mathbb{Z}-1$, Eq. (2.4) holds. It follows that $\max_\sigma |\chi(<_\sigma \boldsymbol{a})| = \|M_\chi(\boldsymbol{a})\|_{S_r}$.*

Proof of Claim 2.14. By the additivity (Eq. (2.5)) of M_χ, it suffices to prove the claim for $d = 1$. This is a straightforward calculation. By Eq. (2.6), for $a \in [r]$,

$$\sigma \cdot M_\chi(a) = \sum_{i,j \in [r],\, \sigma(i) > \sigma(j)} \delta_{i,a}\chi(j) = \sum_{j \in [r] : \sigma(j) < \sigma(a)} \chi(j).$$

The right-hand-side is exactly $\chi(<_\sigma a)$. □

Now that we have Claim 2.14, it remains to bound $\min_\chi \max_{\boldsymbol{a}} \|M_\chi(\boldsymbol{a})\|_{S_r}$ below. This quantity is at least the value of the following d-round game played between a "minimizer" and a "maximizer."

Definition 2.15 (The seminorm unbalancing game)**.** The states of the seminorm balancing game are $r \times r$ integer matrices M. The matrix M is updated in each round as follows:

1. The minimizer chooses a row vector v in $(2\mathbb{Z}-1)^r$; that is, a list of r odd numbers.
2. The maximizer chooses a number $i \in [r]$ and adds v to the i^{th} row of M.

The value for the maximizer is the value of $\|M\|_{S_r}$ at the end of the game. ◊

The coloring $\chi : [r]^d \to (2\mathbb{Z}-1)$ determines the following strategy for the minimizer: if the maximizer chose rows $\boldsymbol{a} = a_1, \ldots, a_{k-1}$ in rounds $1, \ldots, k-1$, the minimizer chooses the vector $v = \chi(\boldsymbol{a}1), \ldots, \chi(\boldsymbol{a}r)$ in round k, where χ on $[r]^k$ is determined by χ on $[r]^d$ as in Definition 2.13. If the minimizer plays this strategy and the maximizer plays $\boldsymbol{a} \in [r]^d$, the matrix after the k^{th} round will be $M_\chi(\boldsymbol{a}[k])$, because $M_{\chi^{\boldsymbol{a}[k-1]}}(a_k)$ has v in the a_k^{th} row and zeroes elsewhere. If the minimizer is constrained to choose w, v in the $(k-1)^{st}$ and k^{th} rounds, respectively, such that $\sum_{i=1}^r v_i = w_{a_{k-1}}$, then by Eq. (2.3) the strategy of the minimizer is determined by some coloring χ as above. However, the value of the game is $\Omega(d)$ even without this constraint on the minimizer.

To show this, we first bound the seminorm below by a simpler quantity. Recall that $\|M\|$ denotes the Frobenius norm of the matrix M, i.e. is the square root of the sum of squares of its entries.

Lemma 2.16. *For $\sigma \in S_r$ chosen uniformly at random,*

$$\|M\|_{S_r} \geq \sqrt{\mathbb{E}_\sigma(\sigma \cdot M)^2} \geq \frac{1}{2\sqrt{6}}\|M - M^\dagger\|.$$

Proof of Lemma 2.16. The first inequality is immediate. Let J be the all–ones matrix. For the second inequality, we use the identity

$$\mathbb{E}_\sigma(\sigma \cdot M)^2 = \frac{1}{4}(\operatorname{tr} J(M + M^\dagger))^2 + \frac{1}{4}\mathbb{E}_\sigma(\sigma \cdot (M - M^\dagger))^2. \tag{2.8}$$

Eq. (2.8) follows because the expectation of the square of a random variable is its mean squared plus its variance, and $\mathbb{E}_\sigma \sigma \cdot M = \frac{1}{2}\operatorname{tr} J(M + M^\dagger)$. The second term is the variance because $M = \frac{1}{2}(M + M^\dagger) + \frac{1}{2}(M - M^\dagger)$ and for any $\sigma \in S_r$, we have $\sigma \cdot \frac{1}{2}(M + M^\dagger) = \mathbb{E}_\sigma \sigma \cdot M$.

Set $A = M - M^\dagger$. In particular, A is antisymmetric. Write

$$\begin{aligned}
\mathbb{E}_\sigma(\sigma \cdot A)^2 &= \sum_{i,j,k,l} A_{i,j}A_{k,l}\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(k)>\sigma(l)}] \\
&= \frac{1}{4}\sum_{|\{i,j,k,l\}|=4} A_{i,j}A_{k,l} + \frac{1}{3}\sum_{|\{i,j,k\}|=3} 2A_{i,j}A_{i,k} \\
&+ \frac{1}{6}\sum_{|\{i,j,k\}|=3} 2A_{i,j}A_{j,k} + \frac{1}{2}\sum_{|\{i,j\}|=2} A_{i,j}A_{i,j}.
\end{aligned}$$

This expression is obtained by computing $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(k)>\sigma(l)}]$ in each of the cases and using antisymmetry of A.

- If $|\{i,j,k,l\}| = 4$, then $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(k)>\sigma(l)}] = 1/4$.
- If $|\{i,j,k\}| = 3$, then $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(i)>\sigma(k)}] = 1/3$, $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(j)>\sigma(k)}] = 1/6$.
- If $|\{i,j\}| = 2$, then $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(i)>\sigma(j)}] = 1/2$, $\mathbb{E}[\mathbf{1}_{\sigma(i)>\sigma(j)}\mathbf{1}_{\sigma(j)>\sigma(i)}] = 0$.

Because A is antisymmetric, the sum over $|\{i,j,k,l\}| = 4$ is zero. Dropping this term, combining the two terms with $|\{i,j,k\}| = 3$, and observing that $\sum_{|\{i,j\}|=2} A_{i,j}A_{i,j} = \|A\|^2$, we have

$$\mathbb{E}_\sigma(\sigma \cdot A)^2 = \frac{1}{3}\left(\sum_i \left(\sum_{j\neq i} A_{i,j}\right)^2 - \|A\|^2\right) + \frac{1}{2}\|A\|^2 \geq \frac{1}{6}\|A\|^2 \tag{2.9}$$

for any antisymmetric matrix A. Combining Eq. (2.9) and Eq. (2.8) completes the proof. □

By Lemma 2.16, it suffices to exhibit a strategy for the maximizer that enforces $\|M - M^\dagger\| \geq d$ after d rounds. This is rather easy – we may accomplish this by focusing only on two entries of M: the maximizer only tries to control the $1,r$ and $2,r$ entries. If in the k^{th} round, minimizer chooses v with $v_r > 0$, the maximizer sets $a_k = 1$. Else, maximizer sets $a_k = 2$. Crucially, the entries of v are odd numbers; in particular, they are greater than 1 in absolute value. Further, all but the first and second rows of M are zero throughout the game. Thus, in the d^{th} round, $|(M - M^\dagger)_{2,r}| + |(M - M^\dagger)_{1,r}| \geq d$, so $\|M - M^\dagger\| \geq d$. □

Remark 2.17 (Improving the lower bound for higher values of r)**.** To prove Conjecture 1.3, it suffices to show the maximizer can achieve $\|M\|_{S_r} = f(r)d$ where $f(r) = \omega(\log r)$. A promising strategy is to replace $\|\cdot\|_{S_r}$ by another seminorm $\|\cdot\|_*$ and show that the maximizer can enforce $\|\cdot\|_* \geq f(r)\|Id\|_{S_r\to *}$, where Id is the identity map on $\mathrm{Mat}_{r\times r}(\mathbb{R})$. Obvious candidates such as $\|M - M^\dagger\|$ and $\|M - M^\dagger\|_1$ do not suffice. Here $\|B\|_1$ is the sum of the absolute values of entries of B. For instance, the minimizer can enforce $\|M - M^\dagger\| = O(d)$ or $\|M - M^\dagger\|_1 = O(\sqrt{r}d)$, and even

antisymmetric matrices A can achieve $\|A\|_{S_r} \leq \|A\|$ and $\|A\|_{S_r} \leq \frac{\sqrt{\log r}}{\sqrt{r}}\|A\|_1$. The first inequality is very easy to achieve, and a result of Erdos and Moon shows the second is achieved by random ± 1 antisymmetric matrices [EM65]. By the inapproximability result mentioned in Remark 2.12, it is not likely that any of the easy–to–compute norms $\|\cdot\|_*$ have both $\|Id\|_{S_r \to *}$ and $\|Id\|_{* \to S_r}$ bounded by constants independent of r. A candidate seminorm is the cut–norm of the top–right $1/3r \times 2/3r$ submatrix of M: it is not hard to see that this seminorm is a lower bound for $\|M\|_{S_r}$.

2.3 Root–mean–squared discrepancy of permutation families

This section is concerned with the proof of Theorem 1.4. Before the proof, we discuss the relationship between Theorem 1.4 and the previous lower bounds in [Mat13], [NTZ13], [Lar17], presented below. The original lower bound was for the usual ℓ_∞ discrepancy.

Theorem 2.18 ([LSV86])**.** *Denote by A be the $|\Omega| \times |\mathcal{A}|$ incidence matrix of $(\Omega, \mathcal{A})$, and define*

$$\mathrm{detlb}(\Omega, \mathcal{A}) = \max_k \max_B |\det(B)|^{1/k}.$$

where B runs over all $k \times k$ submatrices of A. Then

$$\mathrm{herdisc}_\infty(\Omega, \mathcal{A}) := \max_{\Gamma \subset \Omega} \mathrm{disc}_\infty(\Omega, \mathcal{A}) \geq \mathrm{detlb}(\Omega, \mathcal{A}).$$

It was proved in [Mat13] that this lower bound behaves well under unions:

Theorem 2.19 ([Mat13])**.**

$$\mathrm{detlb}(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k) = O\left(\sqrt{k} \max_{i \in [k]} \mathrm{detlb}(\Omega, \mathcal{A}_i)\right),$$

where $+$ denotes the multiset sum (union with multiplicity).

Next, consider the analogue of the determinant lower bound for disc_2.

Theorem 2.20 (Theorem 6 of [Lar17]; corollary of Theorem 11 of [NTZ13] up to constants)**.** *Denote by A be the $|\Omega| \times |\mathcal{A}|$ incidence matrix of $(\Omega, \mathcal{A})$, and define*

$$\mathrm{detlb}_2(\Omega, \mathcal{A}) = \max_{\Gamma \subset \Omega} \sqrt{\frac{m|\Gamma|}{8\pi e}} \det(A|_S^\dagger A|_S)^{\frac{1}{2|\Gamma|}}.$$

Then $\mathrm{herdisc}_2(\Omega, \mathcal{A}) \geq \mathrm{detlb}_2(\Omega, \mathcal{A})$.

Theorem 2.21 (Consequence of the proof of Theorem 7 of [Lar17])**.**

$$\text{herdisc}_2(\Omega, \mathcal{A}) = O(\sqrt{\log n}\,\text{detlb}_2(\Omega, \mathcal{A})).$$

The main point of Theorem 2.21 and Theorem 2.20 is that detlb_2 is a $\sqrt{\log n}$ approximation to herdisc_2. Taken together with Theorem 2.19, we obtain the following bound.

Observation 2.22 (Communicated by Aleksandr Nikolov)**.**

$$\text{herdisc}_2(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k) = O\left(\sqrt{k \log n} \max_{i \in [k]} \text{herdisc}_\infty(\Omega, \mathcal{A}_i)\right)$$

Proof. Applying the Cauchy-Binet identity to $\det(A^\dagger A)$ implies

$$\text{detlb}_2(\Omega, \mathcal{A}) = O(\text{detlb}(\Omega, \mathcal{A})).$$

By Theorem 2.21, Theorem 2.19, and Theorem 2.18,

$$\begin{aligned}
\text{herdisc}_2(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k) &= O\left(\sqrt{\log n}\,\text{detlb}_2(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k)\right) \\
&= O\left(\sqrt{\log n}\,\text{detlb}(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k)\right) \\
&= O\left(\sqrt{k \log n} \max_{i \in [k]} \text{detlb}(\Omega, \mathcal{A}_i)\right). \\
&= O\left(\sqrt{k \log n} \max_{i \in [k]} \text{herdisc}_\infty(\Omega, \mathcal{A}_i)\right)
\end{aligned}$$

□

If $(\Omega, \mathcal{A})$ is a 1–permutation family, then $\text{herdisc}_\infty(\Omega, \mathcal{A}) = 1$. Combined with Observation 2.22, we immediately recover the bound from [Spe87].

Corollary 2.23. *If $(\Omega, \mathcal{A})$ is a k–permutation family, then* $\text{herdisc}_2(\Omega, \mathcal{A}) \leq \sqrt{k \log n}$.

Theorem 1.4 implies that, for constant k, Corollary 2.23 and Observation 2.22 are tight. Further, the reasoning for Observation 2.22 shows that for k constant, $\text{detlb}_2(\Omega, \mathcal{A})$ is constant for k–permutation families $(\Omega, \mathcal{A})$. Thus, Theorem 1.4 shows that Theorem 2.21 is best possible in the sense that there can be a $\Omega(\sqrt{\log n})$ gap between $\text{detlb}_2(\Omega, \mathcal{A})$ and $\text{herdisc}_2(\Omega, \mathcal{A})$.

We now proceed with the proof of Theorem 1.4, which follows immediately from the below proposition.

Proposition 2.24 (Root–mean–discrepancy of a 6–permutation family)**.**

$$\mathrm{disc}_2([3]^d, \mathcal{A}_{S_3}) \geq \mathrm{oddisc}_2([3]^d, \mathcal{A}_{S_3}) = \Omega(\sqrt{d}).$$

Fix a coloring $\chi : [3]^d \to 2\mathbb{Z} - 1$. We must show $\mathrm{disc}_2(\mathcal{A}_{S_3}, \chi)^2 = \Omega(d)$. By Lemma 2.16 and Eq. (2.4),

$$\mathrm{disc}_2(\mathcal{A}_{S_3}, \chi)^2 = \mathbb{E}\boldsymbol{a}[|\chi(<_\sigma \boldsymbol{a})|^2] \geq \frac{1}{2\sqrt{6}}\mathbb{E}_{\boldsymbol{a}}\|M_\chi(\boldsymbol{a}) - M_\chi(\boldsymbol{a})^\dagger\|^2. \tag{2.10}$$

Consider again the seminorm unbalancing game of Definition 2.15. Let $(M_i : i \in [d])$ be the sequence of random matrices determined by minimizer playing strategy χ against maximizer choosing the sequence of rows $\boldsymbol{a}$ uniformly at random, so that $\mathbb{E}_{\boldsymbol{a}}\|M_\chi(\boldsymbol{a}) - M_\chi(\boldsymbol{a})^\dagger\|^2 = \mathbb{E}_{\boldsymbol{a}}\|M_d - M_d^\dagger\|^2$. It is enough to show $\mathbb{E}_{\boldsymbol{a}}\|M_d - M_d^\dagger\|^2 = \Omega(d)$. Consider the sequence of random variables $Y_i = (M_i - M_i^\dagger)_{1,2} + (M_i - M_i^\dagger)_{2,3} - (M_i - M_i^\dagger)_{1,3}$. By the Cauchy-Schwarz inequality, $\|M_d - M_d^\dagger\|^2 \geq |Y_d|^2/3$, so it is enough to show that $\mathbb{E}_{\boldsymbol{a}}[Y_d^2] = \Omega(d)$. We will instead show the following, which clearly implies Proposition 2.24.

Claim 2.25. *Let $\chi : [3]^d \to 2\mathbb{Z} - 1$. If $\mathrm{disc}_2(\mathcal{A}_{S_3}, \chi) \leq 0.2(1.9/\sqrt{3})^d$ then $\mathbb{E}_{\boldsymbol{a}}Y_d^2 \geq 10^{-4}d$.*

We now make a few observations to motivate and aid in the proof of Claim 2.25. The sequence Y_i is a martingale with respect to M_i, because $Y_i - Y_{i-1}|M_{i-1}$ is equally likely to be $v_2 - v_3$, $v_3 - v_1$, or $v_1 - v_2$ if the minimizer chooses v in round i. Because Y_i is a martingale,

$$\mathbb{E}_{\boldsymbol{a}}Y_d^2 = \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{a}[i-1]}\left[\frac{(v_2 - v_3)^2 + (v_1 - v_3)^2 + (v_1 - v_2)^2}{3}\middle| \boldsymbol{a}[i-1]\right]. \tag{2.11}$$

There are strategies for the minimizer that make the above quantity small, but they are bad strategies if they come from a coloring χ. If $(v_2 - v_3)^2 + (v_1 - v_3)^2 + (v_1 - v_2)^2$ is small, then v_1, v_2, v_3 are typically equal. However, strategies induced by χ satisfy that $v_1^k + v_2^k + v_3^k = v_{a_{k-1}}^{k-1}$ if the minimizer chose v^{k-1}, v^k in round $k-1, k$, respectively, and the maximizer chose a_{k-1} in round $k-1$. v^k typically having equal entries should lead to the entries of v^k exponentially decreasing with k, which means that for k small they must be very large. This leads to a high discrepancy.

We now make this intuition precise. Firstly, because $|\mathcal{A}_{S_3}| \le 6 \cdot 3^d$, if $\mathrm{disc}_2(\mathcal{A}_{S_3}, \chi)^2 \le 0.25 \cdot 1.9^{2d}/(6 \cdot 3^d)$, then $|\chi(E)| \le 0.5 \cdot 1.9^d$ for *every* $E \in \mathcal{A}_{S_3}$. It follows that $|\chi(\boldsymbol{\varepsilon})| = |\chi(<_e \overline{3}) + \chi(\overline{3})| = |\chi(<_e \overline{3}) + \chi(<_\tau \overline{3}2)| \le 1.9^d$ where $\tau = (1, 3)$. In summary,

Observation 2.26. *Let* $\chi : [3]^d \to 2\mathbb{Z} - 1$. *If* $\mathrm{disc}_2(\mathcal{A}_{S_3}, \chi)^2 \le 1.9^{2d}/(24 \cdot 3^d)$ *then* $|\chi(\boldsymbol{\varepsilon})| \le 1.9^d$.

Next, we show that the assumption $|\chi(\boldsymbol{\varepsilon})| \le 1.9^d$ implies many cancellations, and that this implies Eq. (2.11) is large. For $\boldsymbol{a} \in [r]^0 \cup [r]^1 \cup \cdots \cup [r]^d$, define the *cancellation* of χ at $\boldsymbol{a}$ by

$$C_\chi(\boldsymbol{a}) = \sum_{i \in [3]} |\chi(\boldsymbol{a}i)| - |\chi(\boldsymbol{a})| \,. \tag{2.12}$$

For $i \in [d]$, define the average cancellation $\overline{C_\chi^i} = \mathbb{E}_{\boldsymbol{a} \in [r]^i} C_\chi(\boldsymbol{a})$. The following two propositions, along with Observation 2.26, imply Claim 2.25.

Proposition 2.27. *Let* $\chi : [3]^d \to 2\mathbb{Z} - 1$. $\mathbb{E} Y_d^2 \ge \frac{1}{d} \left(\sum_{i=1}^d \overline{C_\chi^i} \right)^2$.

Proposition 2.28. *Let* $\chi : [3]^d \to 2\mathbb{Z} - 1$. *If* $|\chi(\boldsymbol{\varepsilon})| \le 1.9^d$ *and* $d \ge 400$, *then* $\sum_{i=1}^d \overline{C_\chi^i} \ge 0.01d$.

Proof of Proposition 2.27. In response to $\boldsymbol{a} = a_1 \dots a_{k-1}$, the maximizer plays the vector $v = (\chi(\boldsymbol{a}1), \chi(\boldsymbol{a}2), \chi(\boldsymbol{a}3))$. Then

$$C_\chi(\boldsymbol{a})^2 = \left(|v_1| + |v_2| + |v_3| - |v_1 + v_2 + v_3| \right)^2 \tag{2.13}$$

$$\le \left(|v_1 - v_2| + |v_2 - v_3| + |v_3 - v_1| \right)^2 \tag{2.14}$$

$$\le 3 \left(|v_1 - v_2|^2 + |v_2 - v_3|^2 + |v_3 - v_1|^2 \right) . \tag{2.15}$$

Eq. (2.14) is the inequality $|a| + |b| + |c| - |a + b + c| \le |a - b| + |b - c| + |c - a|$, which can be proved by cases: without loss of generality, $a \le b \le c$, if all are positive, then both sides vanish; else, without loss of generality $a \le 0 \le b$. In this case the left–hand side is $2|a|$, but the right–hand side is $2|a| + 2|c|$. Thus, if the strategy of the minimizer

is induced by χ, using Eq. (2.15) and Eq. (2.11) we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}} Y_d^2 &= \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{a}[i-1]} \left[\frac{(v_2-v_3)^2+(v_1-v_3)^2+(v_1-v_2)^2}{3} \middle| \boldsymbol{a}[i-1] \right] \\ &\geq \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{a}\in[r]^i} \left[C_\chi(\boldsymbol{a})^2 \right] \geq \sum_{i=1}^{d} \overline{C_\chi^i}^2 \geq \frac{1}{d} \left(\sum_{i=1}^{d} \overline{C_\chi^i} \right)^2 . \end{aligned} \tag{2.16}$$

□

Proof of Proposition 2.28. Define the average absolute value $\overline{|\chi_i|} = \mathbb{E}_{\boldsymbol{a}\in[r]^i} |\chi(\boldsymbol{a})|$. Note that $\overline{|\chi_i|} \geq 1$. Thus, there exists $j \in \{1, \ldots, \lceil .99d \rceil\}$ such that $\overline{|\chi_{j-1}|} \leq 2\overline{|\chi_j|}$, else $|\chi(\boldsymbol{\varepsilon})| = \overline{|\chi_0|} \geq 2^{.99d} > 1.9^d$.

Taking the expectation of both sides of the definition Eq. (2.12) of cancellation yields the identity

$$\overline{C_\chi^i} = 3\overline{|\chi_{i+1}|} - \overline{|\chi_i|},$$

so

$$\sum_{i=j}^{d-1} \overline{C_\chi^i} = 3\overline{|\chi_d|} - \overline{|\chi_{j-1}|} + 2\sum_{i=j}^{d-1} \overline{|\chi_i|} \geq 2 \sum_{i=j+1}^{d-1} \overline{|\chi_i|} \geq 2(\lfloor 0.01d \rfloor - 2).$$

The right–hand side is at least $0.01d$ provided d is at least 400. □

2.4 Root–mean–squared discrepancy under unions

The proof of Theorem 2.21 proceeds through an intermediate quantity defined in [Lar17]. Denote by A be the $|\Omega| \times |\mathcal{A}|$ incidence matrix of $(\Omega, \mathcal{A})$, and let λ_l be the l^{th} largest eigenvalue of $A^\dagger A$. Define

$$\begin{aligned} \mathrm{kgl}(\Omega, \mathcal{A}) &= \max_{1 \leq l \leq \min\{|\Omega|, |\mathcal{A}|\}} \frac{l}{e} \sqrt{\frac{\lambda_l}{8\pi |\Omega||\mathcal{A}|}} \\ \text{and } \mathrm{herkgl}(\Omega, \mathcal{A}) &= \max_{\Gamma \subset \Omega} \mathrm{kgl}(\Gamma, \mathcal{A}|_\Gamma). \end{aligned}$$

Theorem 2.29 (Corollary 2 and consequence of the proof of Theorem 7 of [Lar17])**.**

$$\mathrm{herkgl}(\Omega, \mathcal{A}) \leq \mathrm{detlb}_2(\Omega, \mathcal{A}) \leq \mathrm{herdisc}_2(\Omega, \mathcal{A}) = O(\sqrt{\log n}\, \mathrm{herkgl}(\Omega, \mathcal{A})).$$

Like detlb, the quantity herkgl behaves nicely under unions.

Observation 2.30. $\mathrm{herkgl}(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k) \leq k \max_{i\in[k]} \mathrm{herkgl}(\Omega, \mathcal{A}_i)$.

Proof of Observation 2.30. Let $C = \max_{i\in[k]} \text{herkgl}(\Omega, \mathcal{A}_i)$. It is enough to show $\text{kgl}(\Gamma, (\mathcal{A}_1 + \cdots + \mathcal{A}_k)|_\Gamma) \le kC$ for any $\Gamma \subset \Omega$. Let $|\Gamma| = n$, $m_i = |\mathcal{A}_i|$, and $\sum m_i = m$. If A_i is the incidence matrix of $(\Gamma, \mathcal{A}_i|_\Gamma)$ and A that of $(\Gamma, (\mathcal{A}_1 + \cdots + \mathcal{A}_k)|_\Gamma)$, then

$$A^\dagger A = A_i^\dagger A_i + \cdots + A_i^\dagger A_i.$$

Weyl's inequality on the eigenvalues of Hermitian matrices asserts that if H_1 and H_2 are $n \times n$ Hermitian matrices then $\lambda_{i+j-1}(H_1 + H_2) \le \lambda(H_1)_i + \lambda(H_2)_j$ for all $1 \le i, j \le i+j-1 \le n$. Applying this inequality inductively, $\lambda_l(A^\dagger A) \le \sum_{i=1}^k \lambda_{\lceil l/k \rceil}(A_i^\dagger A_i)$. Thus,

$$\begin{aligned}
\text{kgl}(\Gamma, (\mathcal{A}_1 + \cdots + \mathcal{A}_k)|_\Gamma) &= \max_{1\le l\le \min\{n,m\}} \frac{l}{e}\sqrt{\frac{\lambda_l(A^\dagger A)}{8\pi mn}} \\
&\le \max_{1\le l\le \min\{n,mk\}} \frac{l}{e}\sqrt{\frac{\sum_{i=1}^k \lambda_{\lceil l/k \rceil}(A_i^\dagger A_i)}{8\pi mn}} \\
&\le kC.
\end{aligned}$$

where in the last line we used $\sum m_i = m$ and $\lambda_{\lceil l/k \rceil}(A_i^\dagger A_i) \le 8\pi m_i n \left(\frac{Cek}{l}\right)^2$ from our assumption that $\text{kgl}(\Gamma, \mathcal{A}_i|_\Gamma) \le \text{herkgl}(\Omega, \mathcal{A}_i) \le C$. □

The following cousin of Observation 2.22 is a pleasant consequence of Observation 2.30 and Theorem 2.29.

Corollary 2.31 (Theorem 1.5 restated)**.**

$$\text{herdisc}_2(\Omega, \mathcal{A}_1 + \cdots + \mathcal{A}_k) = O\left(k\sqrt{\log n}\, \max_{i\in[k]} \text{herdisc}_2(\Omega, \mathcal{A}_i)\right).$$

Improving k to $\sqrt{k}$ in Observation 2.30, would strengthen Observation 2.22 and generalize Theorem 2.2.

Chapter 3

Discrepancy of random matrices with many columns

This work is based on the joint work [FS18] of the author and Michael Saks.

3.1 Introduction

We continue our study of discrepancy, turning to the discrepancy of random matrices and set-systems. The discrepancy of a matrix $M \in \mathrm{Mat}_{m\times n}(\mathbb{C})$ or $\mathrm{Mat}_{m\times n}(\mathbb{R})$ is

$$\mathrm{disc}(M) = \min_{\boldsymbol{v}\in\{+1,-1\}^n} \|M\boldsymbol{v}\|_\infty. \tag{3.1}$$

If M is the incidence matrix of the set-system $(\Omega, \mathcal{S})$, then Eq. (3.1) and Eq. (2.1) agree. Using a clever linear-algebraic argument, Beck and Fiala showed that the discrepancy of a set-system $(\Omega, \mathcal{S})$ is bounded above by a function of its *maximum degree* $\Delta(\mathcal{S}) := \max_{x\in\Omega} |\{S \in \mathcal{S} : x \in S\}|$. If $\Delta(\mathcal{S})$ is at most t, we say $(\Omega, \mathcal{S})$ is *t-sparse*.

Theorem 3.1 (Beck-Fiala [BF81])**.** *If $(\Omega, \mathcal{S})$ is t-sparse, then* $\mathrm{disc}(\Omega, \mathcal{S}) \le 2t - 1$.

Beck and Fiala conjectured that $\mathrm{disc}(\mathcal{S})$ is actually $O(\sqrt{t})$ for t-sparse set-systems $(\Omega, \mathcal{S})$. The following stronger conjecture is due to Komlós:

Conjecture 3.2 (Komlós Conjecture; see [Spe87])**.** *If every column of M has Euclidean norm at most* 1*, then* $\mathrm{disc}(M)$ *is bounded above by an absolute constant independent of n and m.*

This conjecture is still open. The current record is due to Banaszczyk [Ban98], who showed $\mathrm{disc}(M) = O(\sqrt{\log n})$ if every column of M has norm at most 1. This implies $\mathrm{disc}(\Omega, \mathcal{S}) = O(\sqrt{t \log n})$ if $(\Omega, \mathcal{S})$ is t-sparse.

3.1.1 Discrepancy of random matrices.

Motivated by the Beck-Fiala conjecture, Ezra and Lovett initiated the study of the discrepancy of random t-sparse matrices [EL16]. Here, motivated by the Komlós conjecture, we study the discrepancy of random $m \times n$ matrices with independent, identically distributed columns.

Question 3.3. *Suppose M is an $m \times n$ random matrix with independent, identically distributed columns drawn from a vector random variable that is almost surely of Euclidean norm at most one. Is there a constant C independent of m and n such that for every $\varepsilon > 0$, $\operatorname{disc}(M) \leq C$ with probability $1 - \varepsilon$ for n and m large enough?*

The Komlós conjecture, if true, would imply an affirmative answer to this question. We focus on the regime where $n \gg m$, i.e., the number of columns is much larger than the number of rows.

A few results are known in the regime $n = O(m)$. The theorems in this direction actually control the possibly larger *hereditary discrepancy*. Define the hereditary discrepancy $\operatorname{herdisc}(M)$ by

$$\operatorname{herdisc}(M) = \max_{Y \subset [n]} \operatorname{disc}(M|_Y),$$

where $M|_Y$ denotes the $m \times |Y|$ matrix whose columns are the columns of M indexed by Y. Again, this agrees with the definition of the hereditary discrepancy of a set-system if M is the incidence matrix of the system.

Clearly $\operatorname{disc}(M) \leq \operatorname{herdisc}(M)$. Often the Komlós conjecture is stated with disc replaced by herdisc. While the Komlós conjecture remains open, some progress has been made for *random t-sparse matrices*. To sample a random t-sparse matrix M, choose each column of M uniformly at random from the set of vectors with t ones and $m - t$ zeroes. Ezra and Lovett showed the following:

Theorem 3.4 ([EL16])**.** *If M is a random t-sparse matrix and $n = O(m)$, then* $\operatorname{herdisc}(M) = O(\sqrt{t \log t})$ *with probability* $1 - \exp(-\Omega(t))$.

The above does not imply a positive answer to Question 3.3 due to the factor of $\sqrt{\log t}$, but is better than the worst-case bound $\sqrt{t \log n}$ due to Banaczszyk.

We now turn to the regime $n \gg m$. It is well-known that if $\mathrm{disc}(M|_Y) \leq C$ holds for all $|Y| \leq m$, then $\mathrm{disc}(M) \leq 2C$ [AS04]. However, this observation is not useful for analyzing random matrices in the regime $n \gg m$. Indeed, if n is large enough compared to m, the set of submatrices $M|_Y$ for $|Y| \leq m$ is likely to contain a matrix of the largest possible discrepancy among t-sparse $m \times m$ matrices, so improving discrepancy bounds via this observation is no easier than improving the Beck-Fiala theorem. The discrepancy of random matrices when $n \gg m$ behaves quite differently than the discrepancy when $n = O(m)$. For example, the discrepancy of a random t-sparse matrix with $n = O(m)$ is only known to be $O(\sqrt{t \log t})$, but it becomes $O(1)$ with high probability if n is large enough compared to m.

Theorem 3.5 ([EL16])**.** *Let M be a random t-sparse matrix. If $n = \Omega\left(\binom{m}{t} \log \binom{m}{t}\right)$ then $\mathrm{disc}(M) \leq 2$ with probability $1 - \binom{m}{t}^{-\Omega(1)}$.*

3.1.2 Discrepancy versus covering radius

Before stating our results, we describe a simple relationship between the covering radius of a lattice and a certain variant of discrepancy. We'll need a few definitions.

- For $S \subseteq \mathbb{R}^m$, let $\mathrm{span}_{\mathbb{R}} S$ denote the linear span of of S, and $\mathrm{span}_{\mathbb{Z}} S$ denote the integer span of S.
- A lattice is a discrete subroup of $\mathbb{R}^m$. Note that the set $\mathrm{span}_{\mathbb{Z}} S$ is a subgroup of $\mathbb{R}^m$, but need not be a lattice. If S is linearly independent or lies inside a lattice, $\mathrm{span}_{\mathbb{Z}} S$ *is* a lattice. Say a lattice in $\mathbb{R}^m$ is *nondegenerate* if $\mathrm{span}_{\mathbb{R}} \mathcal{L} = \mathbb{R}^m$.
- For any norm $\|\cdot\|_*$ on $\mathbb{R}^m$, we write $d_*(x, y)$ for the associated distance, and for $S \subseteq \mathbb{R}^m$, $d_*(x, S)$ is defined to be $\inf_{y \in S} d_*(x, y)$.
- The *covering radius* $\rho_*(S)$ of a subset S with respect to the norm $\|\cdot\|_*$ is $\sup_{x \in \mathrm{span}_{\mathbb{R}} S} d_*(x, S)$ (which may be infinite.)
- The discrepancy may be defined in other norms than ℓ_∞. If M is an $m \times n$ matrix and $\|\cdot\|_*$ a norm on $\mathbb{R}^m$, define the $*$-*discrepancy* $\mathrm{disc}_*(M)$ by

$$\mathrm{disc}_*(M) := \min_{y \in \{\pm 1\}} \|My\|_*.$$

In particular, $\mathrm{disc}(M)$ is $\mathrm{disc}_\infty(M)$.

A natural relaxation of $*$-discrepancy is the *odd$_*$ discrepancy*, denoted $\mathrm{oddisc}_*(M)$. Instead of assigning ± 1 to the columns, one could minimize $\|M\mathbf{y}\|_*$ for $\mathbf{y}$ with odd entries. This definition is consistent with the odd discrepancy of a set-system. By writing each odd integer as 1 plus an even number, it is easy to see that the odd$_*$ discrepancy of M is equal to

$$\mathrm{oddisc}_*(M) = d_*(M\mathbf{1}, 2\mathcal{L}) \leq 2\rho_*(\mathcal{L}).$$

where $\mathcal{L}$ is the lattice generated by the columns of M and $\mathbf{1}$ is the all-ones vector. In fact, by standard argument which can be found in [LSV86], the maximum odd$_*$ discrepancy of a matrix whose columns generate $\mathcal{L}$ is sandwiched between $\rho_*(\mathcal{L})$ and $2\rho_*(\mathcal{L})$.

In general, $\mathrm{disc}_*(M)$ can be arbitrarily large compared to $\mathrm{oddisc}_*(M)$, even for $m = 1, n = 2$. If $r \in \mathbb{Z}$ then $M = [2r+1, r]$ has $\rho_*(\mathrm{span}_{\mathbb{Z}}\, M) = 1/2$ but $\mathrm{disc}_*(M) = r+1$. However, the discrepancy of a *random* matrix with many columns drawn from $\mathcal{L}$ behaves more like the odd discrepancy.

Proposition 3.6. *Suppose X is a random variable on $\mathbb{R}^m$ whose support generates a lattice $\mathcal{L}$. Then for any $\varepsilon > 0$, there is an $n_0(\varepsilon)$ so that for $n > n_0(\varepsilon)$, a random $m \times n$ matrix with independent columns generated from X satisfies*

$$\mathrm{disc}_*(M) \leq d_*(M\mathbf{1}, 2\mathcal{L}) \leq 2\rho_*(\mathcal{L})$$

with probability at least $1 - \varepsilon$.

Proof. Let S be the support of S. For every subset T of S, let s_T be the sum of the elements of T. Let C be large enough that for all T, there is an integer combination v_T of elements of S with even coefficients at most C such that $\|v_T - s_T\| \leq d_*(s_T, 2\mathcal{L})$.

Choose $n_0(\varepsilon)$ large enough so that with probability at least $1 - \varepsilon$, if we take $n_0(\varepsilon)$ samples of X, every element of S appears at least $C+1$ times. Let $n \geq n(\varepsilon)$ and let M be a random matrix obtained by selecting n columns according to X. With probability at least $1-\varepsilon$ every vector in S appears at least C times. We claim that if this happens,

$\mathrm{disc}_*(M) \leq d_*(M\mathbf{1}, 2\mathcal{L})$. This is because if T is the subset of S that appeared an odd number of times in M, $d_*(M\mathbf{1}, 2\mathcal{L}) = d_*(s_T, 2\mathcal{L})$, but because each element of S appears at least $C+1$ times, we may choose $\mathbf{y} \in \{\pm 1\}^n$ so that $M\mathbf{y} = s_T - v_T$ for $\|v_T - s_T\| \leq d_*(s_T, 2\mathcal{L})$. □

3.1.3 Our results

The above simple result says nothing about the number of columns required for M to satisfy the desired inequality with high probability. The focus of this chapter is on obtaining quantitative upper bounds on the function $n_0(\varepsilon)$. We will consider the case when $\mathrm{span}_{\mathbb{Z}} \mathrm{supp}(X)$ is a lattice $\mathcal{L}$. The bounds we obtain will be expressed in terms of m and several quantities associated to the lattice $\mathcal{L}$, the random variable X and the norm $\|\cdot\|_*$. Without loss of generality, we assume X is symmetric, i.e. $\Pr[X = x] = \Pr[X = -x]$ for all x. For a real number $L > 0$ we write $B(L)$ for the set of points in $\mathbb{R}^m$ of (Euclidean) length at most L.

- The $\|\cdot\|_*$ covering radius $\rho_*(\mathcal{L})$.
- The *distortion* R_* of the norm $\|\cdot\|_*$, which is defined to be maximum Euclidean length of a vector x such that $\|x\|_* = 1$. For example, $R_\infty = \sqrt{m}$.
- The determinant $\det \mathcal{L}$ of the lattice $\mathcal{L}$, which is the determinant of any matrix whose columns form a basis of $\mathcal{L}$.
- The determinant $\det \Sigma$, where $\Sigma = \mathbb{E}[XX^\dagger]$ is the $m \times m$ covariance matrix of X.
- The smallest eigenvalue σ of Σ.
- The maximum Euclidean length $L = L(Z)$ of a vector in the support of $Z = \Sigma^{-1/2} X$.
- A parameter $s(X)$ called the *spanningness*. The definition of this crucial parameter is technical and is given in Section 3.1.4; roughly speaking, it is large if X is not heavily concentrated near some proper sublattice of $\mathcal{L}$.

We now state our main quantitative theorem about discrepancy of random matrices.

Theorem 3.7 (Main discrepancy theorem)**.** *Suppose X is a random variable on a nondegenerate lattice $\mathcal{L}$. Let $\Sigma := \mathbb{E}XX^\dagger$ have least eigenvalue σ. Suppose* $\operatorname{supp} X \subset \Sigma^{1/2}B(L)$ *and that $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$. If $n \geq N$ then*

$$\operatorname{disc}_*(M) \leq d_*(M\mathbf{1}, 2\mathcal{L}) \leq 2\rho_*(\mathcal{L})$$

with probability at least

$$1 - O\left(L\sqrt{\frac{\log n}{n}}\right).$$

Here N, given by Eq. (3.22) *in Section 3.2, is a polynomial in the quantities m, $s(\Sigma^{-1/2}X)^{-1}$, L, R_*, $\rho_*(\mathcal{L})$, and* $\log(\det \mathcal{L}/\det \Sigma)$.

Remark 3.8 (degenerate lattices)**.** Our assumption that $\mathcal{L}$ is nondegenerate is without loss of generality; if $\mathcal{L}$ is degenerate, we may simply restrict to $\operatorname{span}_{\mathbb{R}} \mathcal{L}$ and apply Theorem 3.7. Further, the assumptions that $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$ and $\mathcal{L}$ is nondegenerate imply $\sigma > 0$.

Remark 3.9 (weaker moment assumptions)**.** Our original motivation, the Kómlos conjecture, led us to study the case when the random variable X is bounded. This assumption is not critical. We can prove a similiar result under the weaker assumption that $(\mathbb{E}\|X\|^\eta)^{1/\eta} = L < \infty$ for some $\eta > 2$. The proofs do not differ significantly, so we give a brief sketch in Section 3.4.4. ◊

Obtaining bounds on the spanningness is the most difficult aspect of applying Theorem 3.7. We'll do this for random t-sparse matrices, for which we extend Theorem 3.5 to the regime $n = \Omega(m^3 \log^2 m)$. For comparison, Theorem 3.5 only applies for $n \gg \binom{m}{t}$, which is superpolynomial in m if $\min(t, m-t) = \omega(1)$.

Theorem 3.10 (discrepancy for random t-sparse matrices)**.** *Let M be a random t-sparse matrix. If $n = \Omega(m^3 \log^2 m)$ then*

$$\operatorname{disc}(M) \leq 2$$

with probability at least $1 - O\left(\sqrt{\frac{m \log n}{n}}\right)$.

Remark 3.11. We refine this theorem later in Theorem 3.24 of Section 3.3 to prove that the discrepancy is, in fact, usually 1. ◊

Using analogous techniques to the proof of Theorem 3.7, we also prove a similar result for a non-lattice distribution, namely the matrices with random unit vector columns.

Theorem 3.12 (random unit vector discrepancy)**.** *Let M be a matrix with i.i.d random unit vector columns. If $n = \Omega(m^3 \log^2 m)$, then*

$$\operatorname{disc} M = O(e^{-\sqrt{\frac{n}{m^3}}})$$

with probability at least $1 - O\left(L\sqrt{\frac{\log n}{n}}\right)$.

One might hope to conclude a positive answer to Question 3.3 in the regime $n \gg m$ from Theorem 3.7. This seems to require the following weakening of the Komlós conjecture:

Conjecture 3.13. *There is an absolute constant C such that for any lattice $\mathcal{L}$ generated by unit vectors, $\rho_\infty(\mathcal{L}) \leq C$.*

3.1.4 Proof overview

In what follows we focus on the case when X is isotropic, because we may reduce to this case by applying a linear transformation. The discrepancy result for the isotropic case, Theorem 3.20, is stated in Section 3.2, and Theorem 3.7 is an easy corollary. We now explain how the parameters in Theorem 3.7 arise.

The theorem is proved via *local central limit theorems* for sums of vector random variables. Suppose M is a fixed $m \times n$ matrix with bounded columns and consider the distribution over $M\mathbf{v}$ where $\mathbf{v}$ is chosen uniformly at random from $(\pm 1)^n$. Multidimensional versions of the central limit theorem imply that this distibution is approximately normal. We will be interested in local central limit theorems, which provide precise estimates on the probability that $M\mathbf{v}$ falls in a particular region. By applying an appropriate local limit theorem to a region around the origin, we hope to show that the probability of being close to the origin is strictly positive, which implies that there is a ± 1 assignment of small discrepancy.

We do not know suitable local limit theorems that work for all matrices M. We will consider random matrices of the form $M = M^X(n)$, where X is a random variable taking values in some lattice $\mathcal{L} \subset \mathbb{R}^m$, and $M^X(n)$ has n columns selected independently according to X. We will show that, for suitably large n (depending on the distribution X), such a random matrix will, with high probability, satisfy a local limit theorem. The relative error in the local limit theorem will decay with n, and our bounds will provide quantitative information on this decay rate. In order to understand our bounds, it helps to understand what properties of X cause the error to decay slowly with n.

We'll seek local limit theorems that compare $\Pr_{\mathbf{y}}[M\mathbf{y} = \boldsymbol{w}]$ to something proportional to $e^{-\frac{1}{2}\boldsymbol{w}^\dagger(MM^\dagger)^{-1}\boldsymbol{w}}$. One cannot expect such precise control if the lattice is very fine. If the spacing tends to zero, we approach the situation in which X is not on a lattice, in which case the probability of expressing any particular element could always be zero! In fact, in the nonlattice situation the covering radius can be zero but the discrepancy can typically be nonzero. For this reason our bounds will depend on $\log(\det \mathcal{L})$ and on L.

We also need $\rho_*(\mathcal{L})$ and the distortion R_* to be small to ensure $e^{-\frac{1}{2}\boldsymbol{w}^\dagger(MM^\dagger)^{-1}\boldsymbol{w}}$ is not too small for some vector $\boldsymbol{w}$ that we want to show is hit by $M\boldsymbol{y}$ with positive probability over $\boldsymbol{y} \in \{\pm 1\}^n$.

Finally, we need that X does not have most of its mass on or near a smaller sublattice $\mathcal{L}'$. This is the role of spanningness, which is analogous to the spectral gap for Markov chains. Since we assume X is symmetric, choosing the columns M and then choosing $\boldsymbol{y}$ at random is the same as adding n identically distributed copies of X. Intuitively, this means that if M is likely to have $M\boldsymbol{y}$ distributed according to a lattice Gaussian, then the sum of n copies of X should also tend to the lattice Gaussian on $\mathcal{L}$. If the support of X is contained in a smaller lattice $\mathcal{L}'$, then clearly X cannot obey such a local central limit theorem, because sums of copies of X are also contained in $\mathcal{L}'$. In fact, this is essentially the only obstruction up to translations. We may state the above obstruction in terms of the *dual lattice* and the Fourier transform of X.

Definition 3.14 (Dual lattice)**.** If $\mathcal{L}$ is a lattice, the *dual lattice* $\mathcal{L}^*$ of $\mathcal{L}$ is the set

$$\mathcal{L}^* = \{\boldsymbol{z} : \langle \boldsymbol{z}, \boldsymbol{\lambda} \rangle \in \mathbb{Z} \text{ for all } \boldsymbol{\lambda} \in \mathcal{L}\}.$$

The Fourier transform $\widehat{X}$ of X is the function defined on $\boldsymbol{\theta} \in \mathbb{R}^m$ by $\widehat{X}(\boldsymbol{\theta}) = \mathbb{E}[\exp(2\pi i \langle X, \boldsymbol{\theta} \rangle)]$. Note that $|\widehat{X}(\boldsymbol{\theta})|$ is always 1 for $\boldsymbol{\theta} \in \mathcal{L}^*$. In fact, if $|\widehat{X}(\boldsymbol{\theta})| = 1$ also *implies* that $\boldsymbol{\theta} \in \mathcal{L}^*$, then the support of X is contained in no (translation of a) proper sublattice of $\mathcal{L}$! This suggests that, in order to show that a local central limit theorem holds, it is enough to rule out vectors $\boldsymbol{\theta}$ outside the dual lattice with $|\widehat{X}(\boldsymbol{\theta})| = 1$.

In this work, the obstructions are points $\boldsymbol{\theta}$ far from the dual lattice with

$$\mathbb{E}[|\langle \boldsymbol{\theta}, X \rangle \bmod 1|^2]$$

small, where the range of $\bmod 1$ is defined to be $(-1/2, 1/2]$. However, we know that for $\boldsymbol{\theta}$ very close to the dual lattice we have $|\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle \bmod 1|^2 = |\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle|^2$ for all $x \in \operatorname{supp} X$, so $\mathbb{E}[|\langle \boldsymbol{\theta}, X \rangle \bmod 1|^2]$ is exactly $d(\boldsymbol{\theta}, \mathcal{L}^*)^2$. The spanningness measures the value of $\mathbb{E}[|\langle \boldsymbol{\theta}, X \rangle \bmod 1|^2]$ where this relationship breaks down.

Definition 3.15 (Spanningness for isotropic random variables)**.** Suppose that Z is an isotropic random variable defined on the lattice $\mathcal{L}$. Let

$$\tilde{Z}(\boldsymbol{\theta}) := \sqrt{\mathbb{E}[|\langle \boldsymbol{\theta}, Z \rangle \bmod 1|^2]},$$

where $y \bmod 1$ is taken in $(-1/2, 1/2]$, and say $\boldsymbol{\theta}$ is *pseudodual* if $\tilde{Z}(\boldsymbol{\theta}) \leq d(\boldsymbol{\theta}, \mathcal{L}^*)/2$. Define the *spanningness* $s(Z)$ of Z by

$$s(Z) := \inf_{\mathcal{L}^* \not\ni \boldsymbol{\theta} \text{ pseudodual}} \tilde{Z}(\boldsymbol{\theta}).$$

It is a priori possible that $s(Z) = \infty$.

Spanningness is, intuitively, a measure of how far Z is from being contained in a proper sublattice of $\mathcal{L}$. Indeed, $s(Z) = 0$ if and only if this the case. Bounding the spanningness is the most difficult part of applying our main theorem. Our spanningness bounds for t-sparse random matrices use techniques from the recent work of Kuperberg, Lovett and Peled [KLP12], in which the authors proved local limit theorems for $M\boldsymbol{y}$ for non-random, highly structured M. Our discrepancy bounds also apply to the lattice random variables considered in [KLP12] with the spanningness bounds computed in that paper; this will be made precise in Lemma 3.33 of Section 3.3.1.

Related work

We submitted a draft of this work in April 2018, and during our revision process Hoberg and Rothvoss posted a paper on arXiv using very similar techniques on a closely related problem [HR18]. They study random $m \times n$ matrices M with independent entries that are 1 with probability p, and show that for $\operatorname{disc} M = 1$ with high probability in n provided $n = \Omega(m^2 \log m)$. The results are closely related but incomparable: our results are more general, but when applied to their setting we obtain a weaker bound of $n \geq \Omega(m^3 \log^2 m)$. Costello [Cos09] obtained very precise results in every norm when X is Gaussian, which imply the discrepancy is constant with high probability for $n = O(m \log m)$.

Organization of the chapter

- In Section 3.2 we build the technical machinery to carry out the strategy from the previous section. We state our local limit theorem and show how to use it to bound discrepancy.
- In Section 3.3 we recall some techniques for bounding spanningness, the main parameter that controls our local limit theorem, and use these bounds to prove Theorem 3.10 on the discrepancy of random t-sparse matrices.
- Section 3.4 contains the proofs of our local limit theorems.
- In Section 3.5 we use similar techniques to bound the discrepancy of matrices with random unit columns.

Notation

If not otherwise specified, M is a random $m \times n$ matrix with columns drawn independently from a distribution X on a lattice $\mathcal{L}$ that is supported only in a ball $B(L)$, and the integer span of the support of X (denoted $\operatorname{supp} X$) is $\mathcal{L}$. Σ denotes $\mathbb{E} X X^\dagger$. D will denote the Voronoi cell of the dual lattice $\mathcal{L}^*$ of $\mathcal{L}$. $\|\cdot\|_*$ denotes an arbitrary norm. Throughout the chapter there are several constants $c_1, c_2, \ldots$. These are assumed to be

absolute constants, and we will assume they are large enough (or small enough) when needed.

3.2 Likely local limit theorem and discrepancy

Here we show that with high probability over the choice of M, the random variable My resembles a Gaussian on the lattice $\mathcal{L}$. We also show how to use the local limit theorem to bound discrepancy.

For ease of reference, we define the rate of growth n must satisfy in order for our local limit theorems to hold.

Definition 3.16. Define $N_0 = N_0(m, s(X), L, \det \mathcal{L})$ by

$$N_0 := c_{14} \max \left\{ m^2 L^2 (\log m + \log L)^2, s(X)^{-4} L^{-2}, L^2 \log^2 \det \mathcal{L} \right\}, \tag{3.2}$$

where c_{14} is a suitably large absolute constant.

A few definitions will be of use in the next theorem.

Definition 3.17 (Lattice Gaussian)**.** For a matrix M, define the lattice Gaussian with covariance $\frac{1}{2}MM^\dagger$ by

$$G_M(\boldsymbol{\lambda}) = \frac{2^{m/2} \det(\mathcal{L})}{\pi^{m/2} \sqrt{\det(MM^\dagger)}} e^{-2\boldsymbol{\lambda}^\dagger (MM^\dagger)^{-1} \boldsymbol{\lambda}}.$$

Theorem 3.18. *Let X be a random variable on a lattice $\mathcal{L}$ such that $\mathbb{E} XX^\dagger = I_m$, $\operatorname{supp} X \subset B(L)$, and $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$. For $n \geq N_0$, with probability at least $1 - c_{13} L \sqrt{\frac{\log n}{n}}$ over the choice of columns of M, for all $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2} M \mathbf{1}$,*

$$\left| \Pr_{y_i \in \{\pm 1/2\}} [M\mathbf{y} = \boldsymbol{\lambda}] - G_M(\boldsymbol{\lambda}) \right| = G_M(0) \cdot \frac{2m^2 L^2}{n}. \tag{3.3}$$

where G_M is as in Definition 3.17.

Equipped with the local limit theorem, we may now bound the discrepancy. We use a special case of a result by Rudelson.

Theorem 3.19 ([Rud99])**.** *Suppose X is an isotropic random vector in $\mathbb{R}^m$ such that $\|X\| \leq L$ almost surely. Let the n columns of the matrix M be drawn i.i.d from X. For*

some absolute constant c_4 independent of m, n

$$\mathbb{E}\left\|\frac{1}{n}MM^\dagger - I_m\right\|_2 \le c_4 L\sqrt{\frac{\log n}{n}}.$$

In particular, there is a constant c_5 such that with probability at least $1 - c_5 L\sqrt{\frac{\log n}{n}}$ we have

$$MM^\dagger \preceq 2nI_m \qquad \text{(concentration)}$$

$$\text{and } MM^\dagger \succeq \frac{1}{2}nI_m \qquad \text{(anticoncentration)}$$

We restate Theorem 3.7 using N_0.

Theorem 3.20 (discrepancy for isotropic random variables)**.** *Suppose X is an isotropic random variable on a nondegenerate lattice $\mathcal{L}$ with $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$ and $\operatorname{supp} X \subset B(L)$. If $n \ge N$ then*

$$\operatorname{disc}_*(M) \le d_*(M\mathbf{1}, 2\mathcal{L}) \le 2\rho_*(\mathcal{L})$$

with probability $1 - c_6 L\sqrt{\frac{\log n}{n}}$, where

$$N_1 = c_{15} \max\left\{R_*^2 \rho_*(\mathcal{L})^2, N_0\left(m, s(X), L, \det \mathcal{L}\right)\right\} \tag{3.4}$$

for N_0 as in Eq. (3.2)*.*

Proof. By the definition of the covering radius of a lattice, there is a point $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$ with $\|\boldsymbol{\lambda}\|_* \le d_*(\frac{1}{2}M\mathbf{1}, \mathcal{L}) \le \rho_*(\mathcal{L})$. It is enough to show that, with high probability over the choice of M, the point $\boldsymbol{\lambda}$ is hit by $M\boldsymbol{y}$ with positive probability over $\boldsymbol{y} \in \{\pm 1/2\}^n$. If so, $2\boldsymbol{y}$ is a coloring of M with discrepancy $2\rho_*(\mathcal{L})$.

Let n be at least $N_0(m, s(X), L, \det \mathcal{L}')$. By Theorem 3.19, the event $MM^\dagger \succeq \frac{1}{2}nI_m$ and the event in Theorem 3.18 simultanously hold with probability at least $1 - c_6 L\sqrt{\frac{\log n}{n}}$ We claim that if the event in Theorem 3.18 occurs, then $\boldsymbol{\lambda}$ is hit by $M\boldsymbol{y}$ with positive probability. Indeed, conditioned on this event, for all $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$ with $e^{-2\boldsymbol{\lambda}^\dagger (MM^\dagger)^{-1}\boldsymbol{\lambda}} > 2m^2L^2/n$ we have

$$\Pr_{\boldsymbol{y} \in \{\pm 1/2\}^n}(M\boldsymbol{y} = \boldsymbol{\lambda}) > 0.$$

Because $n \ge N$, $e^{-1} \ge 2m^2L^2/n$. Thus, it is enough to show $\boldsymbol{\lambda}^\dagger (MM^\dagger)^{-1}\boldsymbol{\lambda} < \frac{1}{2}$. This is true because $MM^\dagger \succeq \frac{1}{2}nI_m$ and so $\boldsymbol{\lambda}^\dagger (MM^\dagger)^{-1}\boldsymbol{\lambda} \le 2\|\boldsymbol{\lambda}\|_*^2 \frac{R_*^2}{n} \le 2\frac{R_*^2}{n}\rho_*(\mathcal{L})^2$. □

Now Theorem 3.7 is an immediate corollary of Theorem 3.20.

Theorem 3.21 (Restatement of Theorem 3.7)**.** *Suppose X is a random variable on a nondegenerate lattice $\mathcal{L}$. Suppose $\Sigma := \mathbb{E}[XX^\dagger]$ has least eigenvalue σ, $\operatorname{supp} X \subset \Sigma^{1/2}B(L)$, and that $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$. If $n \geq N$ then*

$$\operatorname{disc}_*(M) \leq d_*(M\mathbf{1}, 2\mathcal{L}) \leq 2\rho_*(\mathcal{L})$$

with probability at least $1 - c_{13}L\sqrt{\frac{\log n}{n}}$, where

$$N = c_{15} \max\left\{ \frac{R_*^2 \rho_*(\mathcal{L})^2}{\sigma}, N_0\left(m, s(\Sigma^{-1/2}X), L, \frac{\det \mathcal{L}}{\sqrt{\det \Sigma}}\right)\right\} \tag{3.5}$$

for N_0 as in Eq. (3.2)*.*

Proof. Note that $\sigma > 0$, because $\mathcal{L}$ is nondegenerate and

$$\mathcal{L} = \operatorname{span}_Z \operatorname{supp} X \subset \operatorname{span}_{\mathbb{R}} \operatorname{supp} X.$$

Thus, $\sigma = 0$ contradicts $\operatorname{span}_{\mathbb{R}} \operatorname{supp} X \subsetneq \mathbb{R}^m$.

Let $Z := \Sigma^{-1/2}X$ so that $\mathbb{E}[ZZ^\dagger] = I_m$; we'll apply Theorem 3.20 to the random variable Z, the norm $\|\cdot\|_{**}$ given by $\|v\|_{**} := \|\Sigma^{1/2}v\|_*$, and the lattice $\mathcal{L}' = \Sigma^{-1/2}\mathcal{L}$. The distortion R_0 is at most $R_*/\sigma^{1/2}$, the lattice determinant becomes $\det \mathcal{L}' = \det \mathcal{L}/\sqrt{\det \Sigma}$, and $\operatorname{supp} Z \subset B(L)$. The covering radius of $\rho_0(\mathcal{L}')$ is exactly $\rho_*(\mathcal{L})$. Since the choice of N in Eq. (3.22) is N_1 of Theorem 3.20 for $Z, \|\cdot\|_{**}$, and $\mathcal{L}'$, we have from Theorem 3.20 that

$$\operatorname{disc}_*(M) = \operatorname{disc}_0(\Sigma^{-1/2}M) \leq 2\rho_0(\mathcal{L}') = 2\rho_*(\mathcal{L})$$

with probability at least $1 - c_{13}L\sqrt{\frac{\log n}{n}}$. □

3.3 Discrepancy of random t-sparse matrices

Here we will state our spanningness bounds for t-sparse matrices, and before proving them, compute the bounds guaranteed by Theorem 3.7.

Fact 3.22 (random t-sparse vector)**.** *A* random t-sparse vector *is $\mathbf{1}_S$ for S drawn uniformly at random from $\binom{[m]}{t}$. Let X be a random t-sparse vector with $0 < t < m$.*

The lattice $\mathcal{L} \subset \mathbb{Z}^m$ is the lattice of integer vectors with coordinate sum divisible by t; we have $\rho_\infty(\mathcal{L}) = 1$. Observe that $\mathcal{L}^* = \mathbb{Z}^m + \mathbb{Z}\frac{1}{t}\mathbf{1}$, where $\mathbf{1}$ is the all ones vector. Since $e_1, \ldots, e_{m-1}, \frac{1}{t}\mathbf{1}$ is a basis for $\mathcal{L}^*$, $\det \mathcal{L} = 1/\det \mathcal{L}^* = t$.

$$\Sigma_{i,j} = \mathbb{E}[XX^\dagger]_{ij} = \begin{cases} \frac{t}{m} & i = j \\ \\ \frac{t(t-1)}{m(m-1)} & i \neq j \end{cases}$$

The eigenvalues of Σ are $\frac{t^2}{m}$ with multiplicity one, and $\frac{t(m-t)}{m(m-1)}$ with multiplicity $m-1$.

The next lemma is the purpose of the next two subsections.

Lemma 3.23. *There is a constant c_{10} such that the spanningness is at least $c_{10}m^{-1}$; that is,*

$$s(\Sigma^{-1/2}X) \geq c_{10}m^{-1}.$$

Before proving this, we plug the spanningness bound into Theorem 3.21 to bound the discrepancy of t-sparse random matrices.

Proof of Theorem 3.10. If X is a random t-sparse matrix, $\|\Sigma^{-1/2}X\|$ is $\sqrt{m}$ with probability one. This is because $\mathbb{E}\|\Sigma^{-1/2}X\|^2 = m$, but by symmetry $\|\Sigma^{-1/2}x\|$ is the same for every $x \in \operatorname{supp} X$. Hence, we may take $L = \sqrt{m}$. By Fact 3.22, σ is $\frac{t(m-t)}{m(m-1)}$. Now N from Theorem 3.21 is at most

$$c_{15} \max \left\{ \underbrace{m \cdot \frac{m(m-1)}{t(m-t)}}_{\frac{R_\infty^2 \rho_\infty(\mathcal{L})^2}{\sigma}}, \underbrace{m^3 \log^2 m}_{m^2 L^2 (\log M + \log L)^2}, \underbrace{m^3}_{s(X)^{-4} L^{-2}} \underbrace{m \log^2 t}_{L^2 \log^2 \det \mathcal{L}} \right\}, \tag{3.6}$$

which is $O(m^3 \log^2 m)$. □

We can refine this theorem to obtain the limiting distribution for the discrepancy.

Theorem 3.24 (discrepancy of random t-sparse matrices)**.** *Let M be a random t-sparse matrix for $0 < t < m$. Let $Y = \mathrm{B}(m, 1/2)$ be a binomial random variable with m trials and success probability $1/2$. Suppose $n = \Omega(m^3 \log^2 m)$. If n is even, then*

$$\Pr[\operatorname{disc}(M) = 0] = 2^{-m+1} + O\left(\sqrt{(m/n)\log n}\right)$$

$$\Pr[\operatorname{disc}(M) = 1] = (1 - 2^{-m+1}) + O\left(\sqrt{(m/n)\log n}\right)$$

and if n is odd then

$$\Pr[\mathrm{disc}(M) = 0] = 0$$
$$\Pr[\mathrm{disc}(M) = 1] = \Pr[Y \geq t | Y \equiv t \bmod 2] + O\left(\sqrt{(m/n)\log n}\right)$$
$$\Pr[\mathrm{disc}(M) = 2] = \Pr[Y < t | Y \equiv t \bmod 2] + O\left(\sqrt{(m/n)\log n}\right)$$

with probability at least $1 - O\left(\sqrt{\frac{m \log n}{n}}\right)$. *Note that*

$$\Pr[Y \leq s | Y \equiv t \bmod 2] = 2^{-m+1} \sum_{k \equiv t \bmod 2}^{s} \binom{m}{k}.$$

We use a standard lemma, which we prove in Appendix B.1.

Lemma 3.25 (Number of odd rows)**.** *Suppose X_n is a sum of n uniformly random vectors of Hamming weight $0 < t < m$ in $\mathbb{F}_2^m$ and Z_n is a uniformly random element of $\mathbb{F}_2^m$ with Hamming weight having the same parity as nt. If d_{TV} denotes the total variation distance, then*

$$d_{TV}(X_n, Z_n) = O(e^{-(2n/m)+m}).$$

Proof of Theorem 3.24. The proof is identical to that of Theorem 3.10 apart from making use of the upper bound $d_*(M\mathbf{1}, 2\mathcal{L})$ rather than the cruder $2\rho_*(\mathcal{L})$. There are two cases:

Case 1: n is odd. The coordinates of $M\mathbf{1}$ sum to nt. The lattice $2\mathcal{L}$ is the integer vectors with even coordinates whose sum is divisible by $2t$. Thus, in order to move $M\mathbf{1}$ to $2\mathcal{L}$, each odd coordinate must be changed to even and the total sum must be changed by an odd number times t. The number of odd coordinates has the same parity as t, so we may move M to $2\mathcal{L}$ by changing each coordinate by at most 1 if and only if the number of odd coordinates is at least t.

Case 2: n is even. In this case, the total sum of the coordinates must be changed by an *even* number times t. The parity of the number of odd coordinates is even, so the odd coordinates can all be changed to even preserving the sum of all the coordinates. This shows may move M to $2\mathcal{L}$ by changing each coordinate by at most 1, and by at most 0 if all the coordinates of $M\mathbf{1}$ are even.

Thus, in the even case the discrepancy is at most 1 with the same failure probability and 0 with the probability all the row sums are even, and in the odd case the discrepancy is at most 1 provided the number of odd coordinates of $M\mathbf{1}$ is at least t. Observe that the vector of row sums of a $m \times n$ random t-sparse matrix taken modulo 2 is distributed as the sum of n random vectors of Hamming weight t in $\mathbb{F}_2^m$. Lemma 3.25 below shows that the Hamming weight of this vector is at most $O(e^{-2n/m+3m})$ in total variation distance from a binomial $\mathrm{B}(m, 1/2)$ conditioned on having the same parity as nt. Because this is dominated by $\sqrt{(m/n)\log n}$ for $n \geq m^3 \log^2 m$, the theorem is proved. □

We'll now discuss a general method for bounding the spanningness of lattice random variables.

3.3.1 Spanningness of lattice random variables

Suppose X is a finitely supported random variable on $\mathcal{L}$. We wish to bound the spanningness $s(X)$ below. The techniques below nearly identical to those in [KLP12], in which spanningness is bounded for a very general class of random variables.

We may extend spanningness for nonisotropic random variables.

Definition 3.26 (nonisotropic spanningness)**.** A distribution X with finite, nonsingular covariance $\mathbb{E}XX^\dagger = \Sigma$ determines a metric d_X on $\mathbb{R}^m$ given by $d_X(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_X$ where the square norm $\|\boldsymbol{\theta}\|_X^2$ is given by $\boldsymbol{\theta}^\dagger \Sigma \boldsymbol{\theta} = \mathbb{E}[\langle X, \boldsymbol{\theta}\rangle^2]$. Let

$$\tilde{X}(\boldsymbol{\theta}) := \sqrt{\mathbb{E}[|\langle \boldsymbol{\theta}, X\rangle \bmod 1|^2]},$$

where $y \bmod 1$ is taken in $(-1/2, 1/2]$, and say $\boldsymbol{\theta}$ is *pseudodual* if $\tilde{X}(\boldsymbol{\theta}) \leq d_X(\boldsymbol{\theta}, \mathcal{L}^*)/2$. Define the *spanningness* $s(X)$ of X by

$$s(X) := \inf_{\mathcal{L}^* \not\ni \boldsymbol{\theta} \text{ pseudodual}} \tilde{X}(\boldsymbol{\theta}).$$

This definition of spanningness is invariant under invertible linear transformations $X \leftarrow AX$ and $\mathcal{L} \leftarrow A\mathcal{L}$; in particular, $s(X)$ is the same as $s(\Sigma^{-1/2}X)$ for which we have $\|\boldsymbol{\theta}\|_{\Sigma^{-1/2}X} = \|\boldsymbol{\theta}\|$. Hence, this definition extends the spanningness of Definition 3.15.

Strategy for bounding spanningness

Our strategy for bounding spanningness follows [KLP12]; we present it here for completeness and consistency of notation. To bound spanningness, we need to show that if $\boldsymbol{\theta}$ is *pseudodual but not dual*, i.e., $0 < \tilde{X}(\boldsymbol{\theta}) \leq d(x, \mathcal{L}^*)/2$, then $\tilde{X}(\boldsymbol{\theta})$ is large. We do this in the following two steps.

1. Find a δ such that if all $|\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1| \leq \frac{1}{\beta}$ for all $\boldsymbol{x} \in \operatorname{supp} X$, then $\tilde{X}(\boldsymbol{\theta}) \geq d_X(\boldsymbol{\theta}, \mathcal{L}^*)$. Such $\boldsymbol{\theta}$ cannot be pseudodual without being dual.

2. X is α-*spreading*: for all $\boldsymbol{\theta}$,

$$\tilde{X}(\boldsymbol{\theta}) \geq \alpha \sup_{\boldsymbol{x} \in \operatorname{supp} X} |\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1|$$

Together, if $\boldsymbol{\theta}$ is pseudodual, then $\tilde{X}(\boldsymbol{\theta}) \geq \alpha/\beta$.

To achieve the first item, we use bounded integral spanning sets as in [KLP12]. The following definitions and lemmas are nearly identical to arguments in the proof of Lemma 4.6 in [KLP12].

Definition 3.27 (bounded integral spanning set)**.** Say B is an integral spanning set of a subspace H of $\mathbb{R}^m$ if $B \subset \mathbb{Z}^m$ and $\operatorname{span}_{\mathbb{R}} B = H$. Say a subspace $H \subset \mathbb{R}^m$ has a β-*bounded integral spanning set* if H has an integral spanning set B with $\max\{\|\boldsymbol{b}\|_1 : \boldsymbol{b} \in B\} \leq \beta$.

Definition 3.28. Let A_X denote the matrix whose columns are the support of X (in some fixed order). Say X is β-*bounded* if $\ker A_X$ has a β-bounded integral spanning set.

Lemma 3.29. *Suppose X is β-bounded. Then either*

$$\max_{\boldsymbol{x} \in \operatorname{supp}(X)} |\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1| \geq \frac{1}{\beta}$$

or

$$\tilde{X}(\boldsymbol{\theta}) \geq d_X(\boldsymbol{\theta}, \mathcal{L}^*).$$

Proof. To prove Lemma 3.29 we use a claim from [KLP12] , which allows us to deduce that if $\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$ is very close to an integer for all $\boldsymbol{x}$ then we can "round" $\boldsymbol{\theta}$ to an element of the dual lattice to get rid of the fractional parts.

Claim 3.30 (Claim 4.12 of [KLP12])**.** *Suppose X is β bounded, and define $r_{\boldsymbol{x}} := \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1 \in (-1/2, 1/2]$ and $k_{\boldsymbol{x}}$ to be the unique integer such that $\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle = k_{\boldsymbol{x}} + r_{\boldsymbol{x}}$. If*

$$\max_{\boldsymbol{x} \in \operatorname{supp}(X)} |r_{\boldsymbol{x}}| < 1/\beta$$

then there exists $l \in \mathcal{L}^*$ with

$$\langle \boldsymbol{x}, l \rangle = k_{\boldsymbol{x}}$$

for all $\boldsymbol{x} \in \operatorname{supp}(X)$.

Now, suppose $\max_{\boldsymbol{x} \in \operatorname{supp}(X)} |\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1| = \max_{\boldsymbol{x} \in \operatorname{supp}(X)} |r_{\boldsymbol{x}}| < 1/\beta$. By Claim 3.30, exists $l \in \mathcal{L}^*$ with $\langle \boldsymbol{x}, l \rangle = k_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \operatorname{supp}(X)$. By assumption,

$$\tilde{X}(\boldsymbol{\theta}) = \sqrt{\mathbb{E}(\langle X, \boldsymbol{\theta} \rangle \bmod 1)^2} = \sqrt{\mathbb{E} r_X^2} = \sqrt{\mathbb{E} \langle X, \boldsymbol{\theta} - l \rangle^2} \geq d_X(\boldsymbol{\theta}, \mathcal{L}^*),$$

proving Lemma 3.29. □

In order to apply Lemma 3.29, we will need to bound $\tilde{X}(\boldsymbol{\theta})$ below when there is some $\boldsymbol{x}$ with $|\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle|$ fairly large.

Definition 3.31 (Spreadingness)**.** Say X is α-spreading if for all $\boldsymbol{\theta} \in \mathbb{R}^m$,

$$\tilde{X}(\boldsymbol{\theta}) \geq \alpha \cdot \sup_{\boldsymbol{x} \in \operatorname{supp} X} |\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle \bmod 1|.$$

Combining Lemma 3.29 with Definition 3.31 yields the following bound.

Corollary 3.32. *Suppose X is β-bounded and α-spreading. Then $s(X) \geq \frac{\alpha}{\beta}$.*

A lemma of [KLP12] immediately gives a bound on spanningness for random variables that are uniform on their support.

Lemma 3.33 (Lemma 4.4 from [KLP12])**.** *Suppose X is uniform on $\operatorname{supp} X \subset B_\infty(L')$ and the* stabilizer group *of $\operatorname{supp} X$ acts transitively on $\operatorname{supp} X$. That is, for any*

two elements $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{supp} X$ *there is an invertible linear transformation* A *such that* $A \operatorname{supp} X = \operatorname{supp} X$ *and* $A\boldsymbol{x} = \boldsymbol{y}$*. Then* X *is*

$$\Omega\left(\frac{1}{(m \log(L'm))^{3/2}}\right) \text{-spreading.}$$

In particular, if X *is* β*-bounded, then*

$$s(X) = \Omega\left(\frac{1}{\beta(m \log(L'm))^{3/2}}\right).$$

3.3.2 Proof of spanningness bound for t–sparse vectors

Using the techniques from the previous section, we'll prove Lemma 3.23, which states that t-sparse random vectors have spanningness $\Omega(m^{-1})$. In particular, we'll prove that t-sparse random vectors are 4-bounded and $\Omega(m^{-1})$-spreading and apply Corollary 3.32.

Random t-sparse vectors are $\Omega(m^{-1})$-spreading

Lemma 3.33 implies t-sparse vectors are $\Omega\left(\frac{1}{(m \log(m))^{3/2}}\right)$-spreading, but we can do slightly better due to the simplicity of the distribution X.

In order to show that t-sparse vectors are c-spreading, recall that we must show that if a *single* vector $\mathbf{1}_S$ has $|\langle \boldsymbol{\theta}, \mathbf{1}_S \rangle \bmod 1| > \delta$, then $\mathbb{E}[|\langle \boldsymbol{\theta}, X \rangle \bmod 1|^2] \geq c^2\delta^2$. We cannot hope for $c = o(m^{-1})$, because for small enough δ the vector $\boldsymbol{\theta} = \delta(\frac{1}{t}\mathbf{1}_{[t]} - \frac{1}{m-t}\mathbf{1}_{m\setminus[t]})$ has $\langle \boldsymbol{\theta}, \mathbf{1}_{[t]} \rangle \bmod 1 = \delta$ but $\mathbb{E}[|\langle \boldsymbol{\theta}, X \rangle \bmod 1|^2] = \frac{1}{m-1}\delta^2$. Our bound is worse than this, but the term in Eq. (3.6) depending on the spanningness is not the largest anyway, so this does not hurt our bounds.

Lemma 3.34. *There exists an absolute constant* $c_{10} > 0$ *such that random* t*-sparse vectors are* $\frac{c_{10}}{m}$*-spreading.*

For $i \in [m]$, $e_i \in \mathbb{R}^m$ denotes the standard basis vector. For $U \in \binom{[m]}{t}$, let $\boldsymbol{e}^U$ denote the standard basis vector in $\mathbb{R}^{\binom{[m]}{t}}$.

Proof. If $t = 0$ or $t = m$, then t-sparse vectors are trivially 1-spreading. Suppose there is some t-subset of $[m]$, say $[t]$ without loss of generality, satisfying $|\langle \boldsymbol{\theta}, \mathbf{1}_{[t]} \rangle \bmod 1| = \delta > 0$. For convenience, for $S \in \binom{[m]}{t}$, define

$$w(S) := |\langle \boldsymbol{\theta}, \mathbf{1}_S \rangle \bmod 1|.$$

We need to show that $w([t]) = \delta$ implies $\mathbb{E}_S w(S)^2 = \Omega(m^{-2}\delta^2)$. To do this, we will define random integer coefficients $\boldsymbol{\lambda} = \left(\lambda_S : S \in \binom{[m]}{t}\right)$ such that

$$\mathbf{1}_{[t]} = \sum_{S \in \binom{[m]}{t}} \lambda_S \mathbf{1}_S.$$

Because $|a + b \bmod 1| \leq |a \bmod 1| + |b \bmod 1|$ for our definition of mod1, we have the lower bound

$$\delta = w([t]) \leq \mathbb{E}_{\boldsymbol{\lambda}} \sum_{S \in \binom{[m]}{t}} w(S)|\lambda_S| = \sum_{S \in \binom{[m]}{t}} w(S) \cdot \mathbb{E}_{\boldsymbol{\lambda}}|\lambda_S|. \tag{3.7}$$

It is enough to show $\mathbb{E}_{\boldsymbol{\lambda}}|\lambda_S|$ is small for all S in $\binom{m}{t}$, because then $\mathbb{E}[w(S)]$ is large and

$$\mathbb{E}[w(S)^2] \geq \mathbb{E}[w(S)]^2.$$

We now proceed to define $\boldsymbol{\lambda}$. Let σ be a uniformly random permutation of $[n]$ and let $Q = \sigma([t])$. We have

$$\mathbf{1}_{[t]} = \mathbf{1}_Q + \sum_{i \in [t] : i \neq \sigma(i)} \boldsymbol{e}_i - \boldsymbol{e}_{\sigma(i)}, \tag{3.8}$$

where $\boldsymbol{e}_i$ is the i^{th} standard basis vector. Now for each $i \in [t] : i \neq \sigma(i)$ pick R_i at random conditioned on $i \in R_i$ but $\sigma(i) \notin R_i$. Then

$$\boldsymbol{e}_i - \boldsymbol{e}_{\sigma(i)} = \mathbf{1}_{R_i} - \mathbf{1}_{R_i - i + \sigma(i)}. \tag{3.9}$$

To construct $\boldsymbol{\lambda}$, recall that $\boldsymbol{e}^U$ denotes the U^{th} standard basis vector in $\mathbb{R}^{\binom{[m]}{t}}$, and define

$$\boldsymbol{\lambda} = \boldsymbol{e}^Q - \sum_{i \in [t] : i \neq \sigma(i)} \boldsymbol{e}^{R_i} - \boldsymbol{e}^{R_i - i + \sigma(i)}.$$

By Eq. (3.8) and Eq. (3.9), this choice satisfies $\sum \lambda_S \mathbf{1}_S = \mathbf{1}_{[t]}$.

It remains to bound $\mathbb{E}_{\boldsymbol{\lambda}}|\lambda_S|$ for each S. We have

$$\mathbb{E}_{\boldsymbol{\lambda}}|\lambda_S| \leq \Pr[Q = S] \tag{3.10}$$

$$+ \sum_{i=1}^{t} \Pr[\sigma(i) \neq i \text{ and } R_i = S] \tag{3.11}$$

$$+ \sum_{i=1}^{t} \Pr[\sigma(i) \neq i \text{ and } R_i - i + \sigma(i) = S]. \tag{3.12}$$

since Q is a uniformly random t-set, *Eq.* (3.10) $= \binom{m}{t}^{-1}$. Next we have $\Pr[\sigma(i) \neq i \text{ and } R_i = S] = \frac{m-1}{m}\Pr[R_i = S]$. However, R_i is chosen uniformly at random among the t-sets containing i, so

$$\Pr[R_i = S] = \binom{m-1}{t-1}^{-1}\mathbf{1}_{i\in S} = \frac{m}{t}\binom{m}{t}^{-1}\mathbf{1}_{i\in S}.$$

Thus *Eq.* (3.11) $\leq (m-1)\binom{m}{t}^{-1}$. Similarly, $R_i - i + \sigma(i)$ is chosen uniformly at random among sets *not* containing i, so $\Pr[R_i - i + \sigma(i) = S] = \binom{m}{t-1}^{-1}\mathbf{1}_{i\notin S} = \frac{m-t+1}{t}\binom{m}{t}^{-1}\mathbf{1}_{i\notin S}$. Thus *Eq.* (3.11) $\leq (m-1)\binom{m}{t}^{-1}$. Thus, for every S we have $\mathbb{E}_{\boldsymbol{\lambda}}|\lambda_S| \leq 2m\binom{m}{t}^{-1}$. Combining this with Eq. (3.7) we have

$$\mathbb{E}[w(S)^2] \geq \mathbb{E}[w(S)]^2 \geq (2m)^{-2}\delta^2.$$

We may take $c_{10} = 1/2$. □

Random t-sparse vectors are 4-bounded

Recall that A_X is a matrix whose columns consist of the finite set

$$\operatorname{supp} X = \left\{\mathbf{1}_S : S \in \binom{[m]}{t}\right\}.$$

We index the columns of A_X by $\binom{[m]}{t}$.

Lemma 3.35. *X is 4-bounded. That is,* $\ker A_X$ *has a 4-bounded integral spanning set.*

Definition 3.36 (the directed graph G)**.** For $S, S' \in \binom{[m]}{t}$ we write $S' \to_j S$ if $1 \in S'$, $j \notin S'$ and S is obtained by replacing 1 by j in S'. Let G be the directed graph with $V(G) = \binom{[m]}{t}$ and $S'S \in E(G)$ if and only if $S' \to_j S$ for some $j \in S \setminus S'$. Thus every set containing 1 has out-degree $m - t$ and in-degree 0 and every set not containing 1 has in-degree t and out-degree 0.

The following proposition implies Lemma 3.35. Note that if $S' \to_j S$, then $\mathbf{1}_{S'} - \mathbf{1}_S = \boldsymbol{e}_1 - \boldsymbol{e}_j$.

Proposition 3.37.

$$\mathcal{S} = \bigcup_{j=2}^{m}\{\boldsymbol{e}^{S'} - \boldsymbol{e}^{S} + \boldsymbol{e}^{Q} - \boldsymbol{e}^{Q'} : S' \to_j S \text{ and } Q' \to_j Q\}$$

is a spanning set for $\ker A_X$.

Proof of Proposition 3.37. Clearly $\mathcal{S}$ is a subset of $\ker A_X$, because if $S' \to_j S$, then $\mathbf{1}_{S'} - \mathbf{1}_S = \boldsymbol{e}_1 - \boldsymbol{e}_j$, and so $A_X(\boldsymbol{e}^{S'} - \boldsymbol{e}^S) = \mathbf{1}_{S'} - \mathbf{1}_S = \boldsymbol{e}_1 - \boldsymbol{e}_j$. Thus, if $S' \to_j S$ and $Q' \to_j Q$, $A_X(\boldsymbol{e}^{S'} - \boldsymbol{e}^S + \boldsymbol{e}^Q - \boldsymbol{e}^{Q'}) = 0$. If $S' \to_j S$, then $A_X(\boldsymbol{e}^{S'} - \boldsymbol{e}^S) = \mathbf{1}_{S'} - \mathbf{1}_S = \boldsymbol{e}_1 - \boldsymbol{e}_j$. Thus, if $S' \to_j S$ and $Q' \to_j Q$, $A_X(\boldsymbol{e}^{S'} - \boldsymbol{e}^S + \boldsymbol{e}^Q - \boldsymbol{e}^{Q'}) = 0$, so $\boldsymbol{e}^{S'} - \boldsymbol{e}^S + \boldsymbol{e}^Q - \boldsymbol{e}^{Q'} \in \ker A_X$.

Next we try to prove $\mathcal{S}$ spans $\ker A_X$. Note that $\dim \ker A_X = \binom{m}{t} - m$, because the column space of A_X is of dimension m (as we have seen, $\boldsymbol{e}_1 - \boldsymbol{e}_j$ are in the column space of A_X for all $1 < j \leq m$; together with some $\mathbf{1}_S$ for $1 \notin S \in \binom{[m]}{t}$ we have a basis of $\mathbb{R}^m$). Thus, we need to show $\dim \operatorname{span}_{\mathbb{R}} \mathcal{S}$ is at least $\binom{m}{t} - m$.

For each $j \in [m] - 1$, there is some pair $Q_j, Q_j' \in \binom{[m]}{t}$ such that $Q_j' \to_j Q_j$. For $j \in \{2, \ldots, m\}$, pick such a pair and let $f_j := \boldsymbol{e}_{Q_j'} - \boldsymbol{e}_{Q_j}$. As there are only $m-1$ many f_j's, $\dim \operatorname{span}\{f_j : j \in [m] - 1\} \leq m - 1$. By the previous argument, if $S' \to_j S$, then $\boldsymbol{e}^{S'} - \boldsymbol{e}^S - f_j \in \ker A_X$. Because $\bigcup_{j=2}^m \{\boldsymbol{e}^{S'} - \boldsymbol{e}^S - f_j : S' \to_j S\} \subset \mathcal{S}$, it is enough to show that

$$\dim \operatorname{span}_{\mathbb{R}} \bigcup_{j=2}^{m} \{\boldsymbol{e}^{S'} - \boldsymbol{e}^S - f_j : S' \to_j S\} \geq \binom{m}{t} - m.$$

We can do this using the next claim, the proof of which we delay.

Claim 3.38.

$$\dim \operatorname{span}_{\mathbb{R}} \bigcup_{j=2}^{m} \{\boldsymbol{e}^{S'} - \boldsymbol{e}^S : S' \to_j S\} = \binom{m}{t} - 1.$$

Let's see how to use Claim 3.38 to finish the proof:

$$\dim \operatorname{span}_{\mathbb{R}} \bigcup_{j=2}^{m} \{\boldsymbol{e}^{S'} - \boldsymbol{e}^S - f_j : S' \to_j S\} \geq$$

$$\dim \operatorname{span}_{\mathbb{R}} \bigcup_{j=2}^{m} \{\boldsymbol{e}^{S'} - \boldsymbol{e}^S : S' \to_j S\} - \dim \operatorname{span}_{\mathbb{R}} \{f_j : 1 \neq j \in [m]\} \geq$$

$$\binom{m}{t} - 1 - (m-1) = \binom{m}{t} - m.$$

The last inequality is by Claim 3.38. □

Now we finish by proving Claim 3.38.

Proof of Claim 3.38. If a directed graph H on $[l]$ is *weakly connected*, i.e. H is connected when the directed edges are replaced by undirected edges, then $\operatorname{span}\{\boldsymbol{e}_i - \boldsymbol{e}_j : ij \in$

$E(H)\}$ is of dimension $l-1$. To see this, consider a vector $\boldsymbol{v} \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{e}_i - \boldsymbol{e}_j : ij \in E(H)\}^{\perp}$. For any $ij \in E$, we must have that $v_i = v_j$. As H is weakly connected, we must have that $v_i = v_j$ for all $i, j \in [l]$, so $\dim \operatorname{span}_{\mathbb{R}}\{\boldsymbol{e}_i - \boldsymbol{e}_j : ij \in E(H)\}^{\perp} \leq 1$. Clearly $\mathbf{1} \in \operatorname{span}_{\mathbb{R}}\{\boldsymbol{e}_i - \boldsymbol{e}_j : ij \in E(H)\}^{\perp}$, so $\dim \operatorname{span}_{\mathbb{R}}\{\boldsymbol{e}_i - \boldsymbol{e}_j : ij \in E(H)\}^{\perp} = 1$.

In order to finish the proof of the claim, we need only show that our digraph G is weakly connected. This is trivially true if $t = 0$, so we assume $t \geq 1$. Ignoring direction of edges, the operations we are allowed to use to get between vertices of G (sets in $\binom{[m]}{t}$, that is) are the addition of 1 and removal of some other element or the removal of 1 and addition of some other element. Thus, each set containing 1 is reachable from some set not containing 1. If S does not contain one and also does not contain some $i \neq 1$, we can first remove any j from S and add 1, then remove 1 and add i. This means $S - j + i$ is reachable from S. If there is no such i, then $S = \{2, \ldots, m\}$. This implies the sets not containing 1 are reachable from one another, so G is weakly connected. □

3.4 Proofs of local limit theorems

3.4.1 Preliminaries

We use a few facts for the proof of Theorem 3.18. Throughout this section we assume X is in isotropic position, i.e. $\mathbb{E}[XX^{\dagger}] = I_m$. This means $D_X = D$ and $B_X(\varepsilon) = B(\varepsilon)$.

Fourier analysis

Definition 3.39 (Fourier transform)**.** If Y is a random variable on $\mathbb{R}^m$, $\widehat{Y} : \mathbb{R}^m \to \mathbb{C}$ denotes the Fourier transform

$$\widehat{Y}(\boldsymbol{\theta}) = \mathbb{E}[e^{2\pi i \langle Y, \boldsymbol{\theta} \rangle}].$$

◊

We will use the Fourier inversion formula, and our choice of domain will be the Voronoi cell in the dual lattice.

Definition 3.40 (Voronoi cell)**.** Define the Voronoi cell D of the origin in $\mathcal{L}^*$ to be the points as close to the origin as anything else in $\mathcal{L}^*$, or

$$D := \{\boldsymbol{r} \in \mathbb{R}^m : \|\boldsymbol{r}\| \leq \inf_{\boldsymbol{t} \in \mathcal{L}^* \setminus \{0\}} \|\boldsymbol{r} - \boldsymbol{t}\|\}.$$

Note that $\operatorname{vol} D = \det \mathcal{L}^* = 1/\det \mathcal{L}$, where $\det \mathcal{L}$ is the volume of any domain whose translates under $\mathcal{L}$ partition $\mathbb{R}^m$.

Fact 3.41 (Fourier inversion for lattices, [KLP12])**.** *For any random variable Y taking values on a lattice $\mathcal{L}$ (or even a lattice coset $\boldsymbol{v} + \mathcal{L}$),*

$$\Pr(Y = \boldsymbol{\lambda}) = \det(\mathcal{L}) \int_D \widehat{Y}(\boldsymbol{\theta}) e^{-2\pi i \langle \boldsymbol{\lambda}, \boldsymbol{\theta} \rangle} d\boldsymbol{\theta}.$$

for all $\boldsymbol{\lambda} \in \mathcal{L}$ (resp. $\boldsymbol{\lambda} \in \boldsymbol{v} + \mathcal{L}$). Here D is the Voronoi cell as in Definition 3.40, but we could take D to be any fundamental domain of $\mathcal{L}$.

3.4.2 Dividing into three terms

This section contains the plan for the proof of Theorem 3.18. The proof compares the Fourier transform of the random variable $M\boldsymbol{y}$ to that of a Gaussian; the integral to compute the difference of the Fourier transforms will be split up into three terms, which we will bound separately.

Let M be a matrix whose columns $\boldsymbol{x}_i$ are fixed vectors in $\mathcal{L}$, and let Y_M denote the random variable $M\boldsymbol{y}$ for $\boldsymbol{y}$ chosen uniformly at random from $\{\pm 1/2\}^n$. This choice is made so that the random variable Y_M takes values in the lattice coset $\mathcal{L} - \frac{1}{2}M\mathbf{1}$. Let Σ_M be the covariance matrix of Y_M, which is given by

$$\Sigma_M = \frac{1}{4} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\dagger = \frac{1}{4} M M^\dagger.$$

Let Y be a centered Gaussian with covariance matrix Σ_M. That is, Y has the density

$$G_M(\boldsymbol{\lambda}) = \frac{1}{(2\pi)^{m/2} \sqrt{\det \Sigma_M}} e^{-\frac{1}{2} \boldsymbol{\lambda}^\dagger \Sigma_M^{-1} \boldsymbol{\lambda}}.$$

Observe that Eq. (3.21) in Theorem 3.18 is equivalent to

$$|\Pr(Y_M = \boldsymbol{\lambda}) - \det(\mathcal{L}) G_M(\boldsymbol{\lambda})| \leq \frac{1}{(2\pi)^{m/2} \sqrt{\det \Sigma_M}} \cdot 2m^2 L^2 n^{-1}$$

for $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$. To accomplish this, we will show that $\widehat{Y_M}$ and $\widehat{Y}$ are very close. By Fourier inversion, for all $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$,

$$|\Pr(Y_M = \boldsymbol{\lambda}) - \det(\mathcal{L})G_M(\boldsymbol{\lambda})| = \det(\mathcal{L})\left|\int_D \widehat{Y_M}(\boldsymbol{\theta})e^{-2\pi i\langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle}d\boldsymbol{\theta} - \int_{\mathbb{R}^m}\widehat{Y}(\boldsymbol{\theta})e^{-2\pi i\langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle}d\boldsymbol{\theta}\right|;$$

recall the Voronoi cell D from Definition 3.40. Let $B(\varepsilon) \subset \mathbb{R}^m$ denote the Euclidean ball of radius ε about the origin. If $B(\varepsilon) \subset D$, then for all $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$,

$$|\Pr(Y_M = \boldsymbol{\lambda}) - \det(\mathcal{L})G_M(\boldsymbol{\lambda})| \leq$$
$$= \det(\mathcal{L})\left(\underbrace{\int_{B(\varepsilon)}|\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{I_1} + + \underbrace{\int_{\mathbb{R}^m\setminus B(\varepsilon)}|\widehat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{I_2} + \underbrace{\int_{D\setminus B(\varepsilon)}|\widehat{Y_M}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{I_3}\right). \tag{3.13}$$

We now show that this is decomposition holds for reasonably large ε, i.e. $B(\varepsilon) \subset D$.

Lemma 3.42. *Suppose $\varepsilon \leq \frac{1}{2L}$. Then Then $B(\varepsilon) \subset D$; in particular, Eq. (3.13) holds.*

Proof. Suppose $\boldsymbol{\theta} \in B(\varepsilon)$; we need to show that any nonzero element of the dual lattice has distance from $\boldsymbol{\theta}$ at least ε. It is enough to show that any such dual lattice element has norm at least 2ε. Suppose $0 \neq \alpha \in \mathcal{L}^*$. As $\mathrm{supp}(X)$ spans $\mathbb{R}^m$, for some $x \in \mathrm{supp}(X)$, we have $0 \neq \langle\boldsymbol{\alpha}, \boldsymbol{x}\rangle \in \mathbb{Z}$, so $\|\boldsymbol{x}\|\|\boldsymbol{\alpha}\| \geq |\langle\boldsymbol{\alpha}, \boldsymbol{x}\rangle| \geq 1$; in particular $\|\boldsymbol{\alpha}\| \geq \frac{1}{L} \geq 2\varepsilon$. □

Proof plan

We bound I_1 using the Taylor expansion of $\widehat{Y_M}$ to see that, near the origin, $\widehat{Y_M}$ is very close to the unnormalized Gaussian $\widehat{Y}$. We bound I_2 using standard tail bounds for the Gaussian. The bounds for the first two terms hold for *any* matrix M satisfying Eq. (concentration) and Eq. (anticoncentration) and for the correct choice of ε. Finally, we bound I_3 in *expectation* over the choice of M. This is the only bound depending on the spanningness.

The term I_1: near the origin

Here we show how to compare $\widehat{Y_M}$ to $\widehat{Y}$ near the origin in order to bound I_1 from Eq. (3.13). The Fourier transform of the Gaussian Y is

$$\widehat{Y}(\boldsymbol{\theta}) = \exp(-2\pi^2 \boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta}).$$

There is a very simple formula for $\widehat{Y_M}$, the Fourier transform of Y_M.

Proposition 3.43. *If M has columns $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$, then*

$$\widehat{Y_M}(\boldsymbol{\theta}) = \prod_{j=1}^{n} \cos(\pi \langle \boldsymbol{x}_j, \boldsymbol{\theta} \rangle). \tag{3.14}$$

Proof.

$$\widehat{Y_M}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y} \in_R \{\pm 1/2\}^n}[e^{2\pi i \langle \sum_{j=1}^n y_j \boldsymbol{x}_j, \boldsymbol{\theta} \rangle}] = \prod_{j=1}^{n} \mathbb{E}_{y_j}[e^{2\pi i \langle y_j \boldsymbol{x}_j, \boldsymbol{\theta} \rangle}] = \prod_{j=1}^{n} \cos(\pi \langle \boldsymbol{x}_j, \boldsymbol{\theta} \rangle).$$

□

We can bound the first term by showing that near the origin, $\widehat{Y_M}$ is very close to a Gaussian. Recall that by Proposition 3.43,

$$\widehat{Y_M}(\boldsymbol{\theta}) = \prod_{j=1}^{n} \cos(\pi \langle \boldsymbol{x}_j, \boldsymbol{\theta} \rangle).$$

For $\boldsymbol{\theta}$ near the origin, $\langle v_j, \boldsymbol{\theta} \rangle$ will be very small. We will use the Taylor expansion of cosine near zero.

Proposition 3.44. *For $x \in (-1/2, 1/2)$, $\cos(\pi x) = \exp(\frac{\pi^2 x^2}{2} + O(x^4))$.*

Proof. Let $\cos(\pi x) = 1 - y$ where $y \in [0, 1)$. Then $\log(\cos(\pi x)) = \log(1 - y) = 1 - y + O(y^2)$. Since $\cos(\pi x) = 1 - \frac{\pi^2 x^2}{2} + O(x^4)$, we have that $y = \frac{\pi^2 x^2}{2} + O(x^4)$. Thus $\log(\cos(\pi x)) = 1 - \frac{\pi^2 x^2}{2} + O(x^4)$. The proposition follows. □

We may now apply Proposition 3.44 for $\|\boldsymbol{\theta}\|$ small enough.

Lemma 3.45. *Suppose M satisfies Eq.* (concentration) *and $\|\boldsymbol{\theta}\| < \frac{1}{2L}$. Then there exists a constant $c_{11} > 0$ such that*

$$\widehat{Y_M}(\boldsymbol{\theta}) \leq \exp\left(-2\pi^2 \boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta} + E\right)$$

for $|E| \leq c_{11} n L^2 \|\boldsymbol{\theta}\|^4$.

Proof. Because for all $i \in [n]$ we have $|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| \leq \|\boldsymbol{x}_i\|\|\boldsymbol{\theta}\| < 1/2$, Proposition 3.44 applies for all $i \in [n]$ and immediately yields that there is a constant c such that

$$\widehat{Y_M}(\boldsymbol{\theta}) = \exp\left(-2\pi^2 \boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta} + E\right).$$

for $|E| \leq c\sum_{j=1}^n \langle x_j, \boldsymbol{\theta}\rangle^4$. Next we bound the quartic part of E by

$$\begin{aligned}
\sum_{j=1}^n \langle \boldsymbol{x}_j, \boldsymbol{\theta}\rangle^4 &\leq \max_{j\in[n]} \|\boldsymbol{x}_j\|^2 \|\boldsymbol{\theta}\|^2 \sum_{j=1}^n \langle \boldsymbol{x}_j, \boldsymbol{\theta}\rangle^2 \\
&\leq L^2\|\boldsymbol{\theta}\|^2 \boldsymbol{\theta}^\dagger \left(\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j^\dagger\right) \boldsymbol{\theta} \\
&\leq 2nL^2\|\boldsymbol{\theta}\|^4,
\end{aligned}$$

and take $c_{11} = 2c$. □

Lemma 3.46 (First term)**.** *Suppose M satisfies Eq.* (anticoncentration) *and Eq.* (concentration)*. Further suppose that $L^2 n\varepsilon^4 < 1$, and that $\varepsilon < \frac{1}{2L}$. There exists c_{12} with*

$$I_1 \leq c_{12} \frac{m^2 L^2 n^{-1}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.$$

Proof. By concentration and Lemma 3.45,

$$I_1 = \int_{B(\varepsilon)} |\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta})| d\boldsymbol{\theta} \leq \int_{B(\varepsilon)} \widehat{Y}(\boldsymbol{\theta}) \left| e^{c_{11} L^2 n \|\boldsymbol{\theta}\|^4} - 1 \right| d\boldsymbol{\theta}.$$

Let the constant c be such that $|e^{c_{11}x} - 1| \leq c|x|$ for $x \in [-1, 1]$. Thus

$$I_1 \leq cL^2 n \int_{B(\varepsilon)} \widehat{Y}(\boldsymbol{\theta}) \|\boldsymbol{\theta}\|^4 d\boldsymbol{\theta}.$$

By Eq. (anticoncentration),

$$I_1 \leq cL^2 n^{-1} \int_{B(\varepsilon)} \widehat{Y}(\boldsymbol{\theta}) \left(\boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta}\right)^2 d\boldsymbol{\theta}. \tag{3.15}$$

Note that $(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}\widehat{Y}$ is equal to the density of $W = \frac{1}{2\pi}\Sigma_M^{-1/2} G$, where G

is a Gaussian vector with identity covariance matrix. $\Sigma_M^{-1/2}$ exists because Eq. (anticoncentration) holds. Further, $W^\dagger \Sigma_M W = \frac{1}{4\pi^2}\|G\|^2$. Therefore

$$\begin{aligned}\int_{\mathbb{R}^m} \widehat{Y}(\boldsymbol{\theta}) \left(\boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta}\right)^2 d\boldsymbol{\theta} &= \frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} \mathbb{E}_W \left[\left(W^\dagger \Sigma_M W\right)^2\right] \\ &= \frac{1}{16\pi^4(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} \mathbb{E}_G \left[\|G\|^4\right] \\ &= \frac{1}{16\pi^4(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} (2m + m^2) \\ &\leq \frac{3m^2}{16\pi^4(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.\end{aligned}$$

Plugging this into (3.15) and setting $c_{12} = \frac{3}{\pi^4}c$ completes the proof.

□

The term I_2: Bounding Gaussian mass far from the origin

Here we bound the term I_2 of Eq. (3.13), which is not too difficult.

Lemma 3.47 (Third term)**.** *Suppose M satisfies Eq.* (anticoncentration) *holds and that $\varepsilon^2 \geq \frac{16m}{\pi^2 n}$. Then*

$$I_2 \leq \frac{e^{-\frac{\pi^2}{8}\varepsilon^2 n}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.$$

Proof. If M satisfies Eq. (anticoncentration), then $B(\varepsilon) \supset \frac{1}{2}\{\boldsymbol{\theta} : \boldsymbol{\theta}^\dagger \Sigma_M \boldsymbol{\theta} \geq n\varepsilon\} := B_M(\varepsilon/2)$. If we integrate over $B_M(\varepsilon/2)$ and change variables, it remains only to calculate how much mass of a standard normal distribution is outside a ball of radius larger than the average norm. By Lemma 4.14 of [KLP12], a standard Gaussian tail bound, if $\varepsilon^2 \geq \frac{16m}{\pi^2 n}$ then

$$\int_{\mathbb{R}^m \setminus B_M(\varepsilon/2)} |\widehat{Y}(\boldsymbol{\theta})| d\boldsymbol{\theta} \leq \frac{e^{-\frac{\pi^2}{8}\varepsilon^2 n}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.$$

□

The term I_3: Bounding the Fourier transform far from the origin

It remains only to bound the term I_3 of Eq. (3.13) which is given by

$$I_3 = \int_{D \setminus B(\varepsilon)} |\widehat{Y_M}(\boldsymbol{\theta})| d\boldsymbol{\theta}.$$

This is the only part in which spanningness plays a role. If ε is at most the spanningness (see Definition 3.15), we can show I_3 is very small with high probability by bounding it in expectation over the choice of M. The proof is a simple application of Fubini's theorem.

Lemma 3.48. *If $\mathbb{E}XX^\dagger = I_m$ and $\varepsilon \le s(X)$, then*

$$\mathbb{E}[I_3] \le \det(\mathcal{L}^*)e^{-2\varepsilon^2 n}$$

Proof. By Fubini's theorem,

$$\begin{aligned}\mathbb{E}_M[I_3] &= \int_{D\setminus B(\varepsilon)} \mathbb{E}|\widehat{Y_M}(\boldsymbol{\theta})|d\boldsymbol{\theta} \\ &\le \det(\mathcal{L}^*)\sup\{\mathbb{E}[|\widehat{Y_M}(\boldsymbol{\theta})|] : \boldsymbol{\theta} \in D \setminus B(\varepsilon)\}. \end{aligned} \tag{3.16}$$

By Proposition 3.43 and the independence of the columns of n,

$$\mathbb{E}_M[|\widehat{Y_M}(\boldsymbol{\theta})|] = (\mathbb{E}|\cos(\pi\langle X, \boldsymbol{\theta}\rangle)|)^n .$$

Thus,

$$\sup\{\mathbb{E}[|\widehat{Y_M}(\boldsymbol{\theta})|] : \boldsymbol{\theta} \in D \setminus B(\varepsilon)\} \le (\sup\{\mathbb{E}[|\cos(\pi\langle X, \boldsymbol{\theta}\rangle)|] : \boldsymbol{\theta} \in D \setminus B(\varepsilon)\})^n . \tag{3.17}$$

$|\cos(\pi x)|$ is periodic with period $1/2$, so it is enough to consider $\langle X, \boldsymbol{\theta}\rangle \bmod 1$, where $x \bmod 1$ is taken to be in $[-1/2, 1/2)$. Note that for $|x| \le 1/2$, $|\cos(\pi x)| = \cos(\pi(x)) \le 1 - 4x^2$, so

$$\mathbb{E}[|\cos(\pi\langle X, \boldsymbol{\theta}\rangle)|] \le 1 - 4\mathbb{E}[(\langle X, \boldsymbol{\theta}\rangle \bmod 1)^2] = 1 - 4\tilde{X}(\boldsymbol{\theta})^2$$

By the definition of spanningness and the assumption in the hypothesis that $\varepsilon \le s(X)$, we know that every vector with $\tilde{X}(\boldsymbol{\theta}) \le d(\boldsymbol{\theta}, \mathcal{L}^*)/2 = \|\boldsymbol{\theta}\|/2$ is either in $\mathcal{L}^*$ or has $\tilde{X}(\boldsymbol{\theta}) \ge \varepsilon$. Thus, for all $\boldsymbol{\theta} \in D$, $\tilde{X}(\boldsymbol{\theta}) \ge \max\{\|\boldsymbol{\theta}\|/2, \varepsilon\}$, which is at least $\varepsilon/2$ for $\boldsymbol{\theta} \in D \setminus B(\varepsilon)$. Combining this with Eq. (3.17) and using $1 - x \le e^{-x}$ implies

$$\sup\{\mathbb{E}[|\widehat{Y_M}(\boldsymbol{\theta})|] : \boldsymbol{\theta} \in D \setminus B(\varepsilon)\} \le e^{-2\varepsilon^2 n}.$$

Plugging this into Eq. (3.16) completes the proof. □

3.4.3 Combining the terms

Finally, we can combine each of the bounds to prove Theorem 3.18.

Proof of Theorem 3.18. Recall the strategy: we have some conditions (the hypotheses of Lemma 3.42) under which we can write the difference between the two probabilities of interest as a sum of three terms, and we have bounds for each of the terms (Lemma 3.46, Lemma 3.48, and Lemma 3.47) respectively. Our expression depends on ε, and so we must choose ε satisfying the hypotheses of those lemmas. These are as follows:

(i) To apply 3.42 we need $\varepsilon \leq \frac{1}{2L}$,

(ii) for Lemma 3.46 we need $L^2 n\varepsilon^4 \leq 1$,

(iii) to apply Lemma 3.47, we need $\varepsilon^2 \geq \frac{16m}{\pi^2 n}$, and

(iv) for Lemma 3.48 we need $\varepsilon \leq s(X)$.

It is not hard to check that setting

$$\varepsilon = L^{-1/2} n^{-1/4}$$

will satisfy the four constraints provided

1. $n \geq 16L^2$,
2. $n \geq (16mL)^2/\pi^4$, and
3. $n \geq s(X)^{-4} L^{-2}$.

However, (1) follows from (2) because $L \geq \sqrt{m}$ (this follows from $\mathbb{E}XX^\dagger = I_m$, which implies $\mathbb{E}[\|X\|^2] = m$), so

$$n \geq (16mL)^2/\pi^4 \text{ and } n \geq s(X)^{-4} L^{-2}$$

suffice. By 3.42 we have

$$|\Pr(Y_M = \boldsymbol{\lambda}) - \det(\mathcal{L}) G_M(\boldsymbol{\lambda})| \leq$$

$$= \det(\mathcal{L}) \left(\underbrace{\int_{B(\varepsilon)} |\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{I_1} + + \underbrace{\int_{\mathbb{R}^m \setminus B(\varepsilon)} |\widehat{Y}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{I_2} + \underbrace{\int_{D \setminus B(\varepsilon)} |\widehat{Y_M}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{I_3} \right).$$

By Lemma 3.48 and Markov's inequality, I_3 is at most $e^{-\varepsilon^2 n}$ with probability at least $1 - e^{-\varepsilon^2 n} \det(\mathcal{L}^*)$. By Theorem 3.19, Eq. (anticoncentration) and Eq. (concentration) hold for M with probability at least $1 - c_5 L\sqrt{(\log n)/n}$. If n is at least a large enough constant times $L^2 \log^2 \det \mathcal{L}$, $e^{-\varepsilon^2 n} \det(\mathcal{L}^*)$ is at most $L\sqrt{(\log n)/n}$. Thus, all three events hold with probability at least $1 - c_{13} L\sqrt{(\log n)/n}$ over the choice of M. Condition on these three events, and plug in the bounds given by Lemma 3.46 and Lemma 3.47 for I_1 and I_2 and the bound $e^{-\varepsilon^2 n} = e^{-\sqrt{n}/L}$ for I_3 to obtain the following:

$$
\begin{aligned}
&|\Pr(Y_M = \boldsymbol{\lambda}) - \det(\mathcal{L}) G_M(\boldsymbol{\lambda})| \\
&\leq \det(\mathcal{L}) \left(\frac{m^2 L^2 n^{-1}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} + \frac{e^{-\frac{\pi^2}{8}\sqrt{n}/L}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} + e^{-\sqrt{n}/L} \right). \\
&\leq \frac{\det(\mathcal{L})}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} \left(m^2 L^2 n^{-1} + e^{-\frac{\pi^2}{8}\sqrt{n}/L} + (2\pi)^{m/2}\sqrt{\det(\Sigma_M)} e^{-\sqrt{n}/L} \right) \\
&\leq \frac{\det(\mathcal{L})}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} \left(m^2 L^2 n^{-1} + 2e^{\frac{m}{2}\log(4\pi n) - \sqrt{n}/L} \right), \qquad (3.18)
\end{aligned}
$$

where the last inequality is by Eq. (concentration). If c_{14} is large enough, the quantity in parentheses in Eq. (3.18) is at most $2m^2 L^2/n$ provided

$$
n \geq N_0 = c_{14} \max\left\{ m^2 L^2 (\log m + \log L)^2, s(X)^{-4} L^{-2}, L^2 \log^2 \det \mathcal{L} \right\}, \qquad (3.19)
$$

and the combined failure probability of the required events is at most $c_{13} L\sqrt{\frac{\log n}{n}}$. □

3.4.4 Weaker moment assumptions

We now sketch how to extend the proof of Theorem 3.18 to the case $(\mathbb{E}\|X\|^\eta)^{1/\eta} = L < \infty$ for some $\eta > 2$, weakening the assumption that $\operatorname{supp} X \subset B(L)$.

Theorem 3.49 (Lattice local limit theorem for > 2 moments)**.** *Let X be a random variable on a lattice $\mathcal{L}$ such that $\mathbb{E}XX^\dagger = I_m$, $(\mathbb{E}\|X\|^\eta)^{1/\eta} = L < \infty$ for some $\eta > 2$, and $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$. Let G_M be as in Theorem 3.18. There exists*

$$
N_2 = \operatorname{poly}(m, s(X), L, \frac{1}{\eta - 2}, \log(\det \mathcal{L}))^{1 + \frac{1}{\eta - 2}} \qquad (3.20)
$$

such that for $n \geq N_2$, with probability at least $1 - 3n^{-\frac{\eta-2}{2+\eta}}$ over the choice of columns of M, the following two properties of M hold:

1. *$MM^\dagger \succeq \frac{1}{2}nI_m$; that is, $MM^\dagger - \frac{1}{2}nI_m$ is positive-semidefinite.*

2. *For all $\boldsymbol{\lambda} \in \mathcal{L} - \frac{1}{2}M\mathbf{1}$,*

$$\left| \Pr_{y_i \in \{\pm 1/2\}} [M\mathbf{y} = \boldsymbol{\lambda}] - G_M(\boldsymbol{\lambda}) \right| \leq G_M(0) \cdot 2m^2 L^2 n^{-\frac{\eta-2}{2+\eta}}. \tag{3.21}$$

Before proving the theorem, note that it allows us to extend our discrepancy result to this regime. The proof of the next corollary from Theorem 3.49 is identical to the proof of Theorem 3.21 from Theorem 3.18.

Corollary 3.50 (Discrepancy for > 2 moments)**.** *Suppose X is a random variable on a nondegenerate lattice $\mathcal{L}$. Suppose $\Sigma := \mathbb{E}[XX^\dagger]$ has least eigenvalue σ, $(\mathbb{E}\|Z\|^\eta)^{1/\eta} = L < \infty$ for some $\eta > 2$ where $Z := \Sigma^{-1/2}X$, and that $\mathcal{L} = \operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$. If $n \geq N_3$ then*

$$\operatorname{disc}_*(M) \leq 2\rho_*(\mathcal{L})$$

with probability at least $1 - 3n^{-\frac{\eta-2}{2+\eta}}$, where

$$N_3 = c_{15} \max\left\{ \frac{R_*^2 \rho_*(\mathcal{L})^2}{\sigma}, N_2\left(m, s(\Sigma^{-1/2}X), L, \frac{\det \mathcal{L}}{\sqrt{\det \Sigma}}\right) \right\} \tag{3.22}$$

for N_0 as in Eq. (3.2)*.*

Proof sketch of Theorem 3.49. We review each step of the proof of Theorem 3.18 and show how it needs to be modified to accomodate the weaker assumptions. Recall that, to prove Theorem 3.18, we had some conditions (the hypotheses of Lemma 3.42) under which we can write the difference between the two probabilities of interest as a sum of three terms, and we have bounds for each of the terms (Lemma 3.46, Lemma 3.48, and Lemma 3.47) respectively. We also need an analogue of Theorem 3.19 which tells us that Eq. (anticoncentration) and Eq. (concentration) hold with high probability. Neither Lemma 3.48 nor Lemma 3.47 use bounds on the moments of $\|X\|$, so they hold as-is. Let's see how the remaining lemmas must be modified:

Matrix concentration: By Theorem 1.1 in [SV$^+$13], Eq. (concentration) and Eq. (anticoncentration) hold with probability at least $1 - n^{-\frac{\eta-2}{2+\eta}}$ provided

$$n \geq \operatorname{poly}(m, L, \frac{1}{\eta - 2})^{1+\frac{1}{\eta-2}}.$$

Lemma 3.42: The bound $\varepsilon \leq \frac{1}{2L}$ becomes $\varepsilon \leq \frac{1}{4}L^{-\frac{\eta}{\eta-2}}$. To prove Lemma 3.42 it was enough to show $\boldsymbol{\alpha}$ was at least twice the desired bound on ε for $\boldsymbol{\alpha} \neq 0 \in \mathcal{L}^*$. Here we do the same, but to show $\boldsymbol{\alpha}$ is large we consider the random variable $Y \geq 1$ defined by conditioning $|\langle \boldsymbol{\alpha}, X\rangle|$ on $\langle \boldsymbol{\alpha}, X\rangle \neq 0$. By assumption, X is isotropic. Let $p = \Pr[\langle \boldsymbol{\alpha}, X\rangle \neq 0]$, so that $\|\boldsymbol{\alpha}\|^2 = p\mathbb{E}[Y^2]$ and $L\|\alpha\| \geq (\mathbb{E}|\langle \boldsymbol{\alpha}, X\rangle|^\eta)^{\frac{1}{\eta}} = p^{\frac{1}{\eta}}(\mathbb{E}[Y^\eta])^{\frac{1}{\eta}} \geq p^{\frac{1}{\eta}}(\mathbb{E}[Y^2])^{\frac{1}{2}}$ by Hölder's inequality. Cancelling p from the two inequalities and using $Y \geq 1$ yields the desired bound.

Lemma 3.46: The analogue of this lemma will require $L^2 n^{1+\frac{4}{2+\eta}}\varepsilon^4 < 1$ and $\varepsilon < \frac{1}{4L}n^{-\frac{2}{2+\eta}}$, and will hold with probability at least $1 - n^{-\frac{\eta}{4+\eta}}$ over the choice of columns of M. The numerator of the right-hand side becomes $m^2L^2n^{-\frac{\eta-2}{2+\eta}}$. Lemma 3.46 followed from Lemma 3.45. Here the analogue of Lemma 3.45 holds with $|E| \leq c_{11}n^{1+\frac{4}{2+\eta}}L^2\|\boldsymbol{\theta}\|^4$ if $\|z_i\| \leq Ln^{\frac{2}{2+\eta}}$ for all $i \in [n]$, which holds with probability $1 - n^{-\frac{\eta-2}{2+\eta}}$ by Markov's inequality. The rest of the proof proceeds the same.

The new constraints on ε will be satisfied if we take

$$\varepsilon = n^{-\frac{4+\eta}{12+3\eta}},$$

and $n \geq \max\left\{(4L)^{\frac{12+6\eta}{\eta-2}}, \frac{16}{\pi^2}m^{\frac{6+3\eta}{2\eta-4}}\right\}$. The remainder of the proof proceeds as for Theorem 3.18. □

3.5 Random unit columns

Let X be a uniformly random element of the sphere $\mathbb{S}^{m-1}$. Again, let M be an $m \times n$ matrix with columns drawn independently from X. Note that X is not a lattice random variable. This time $\Sigma = \frac{1}{m}I_m$, and $\|\Sigma^{-1/2}X\|$ is always at most m.

We will again prove a local limit theorem, only this time we will not precisely control the probability of hitting a point but rather the expectation of a particular function. The function is similar to the indicator of the cube, but it will be modified a bit to make it easier to handle. Let B be the function, which we will determine later. Recall that, once M is chosen, Y_M is the random variable obtained by summing the columns of

M with i.i.d $\pm 1/2$ coefficients. Σ_M is $MM^\dagger/4$, and Y is the Gaussian with covariance matrix Σ_M. We will try to show that, with high probability over the choice of M, $\mathbb{E}B(Y_M) \sim \mathbb{E}B(Y)$. If B is supported only in $[-K, K]^m$, to show that $\operatorname{disc} M < K$ it suffices to show that

$$|\mathbb{E}B(Y_M) - \mathbb{E}B(Y)| < \mathbb{E}B(Y).$$

3.5.1 Nonlattice likely local limit

We now investigate a different extreme case in which $\operatorname{span}_{\mathbb{Z}} \operatorname{supp} X$ is dense in $\mathbb{R}^m$. In this case the "dual lattice" is $\{0\}$, so we define pseudodual vectors to be those vectors with $\tilde{X}(\boldsymbol{\theta}) \le \|\boldsymbol{\theta}\|_X/2$, and the spanningness to be the least value of $\tilde{X}(\boldsymbol{\theta})$ at a nonzero pseudodual vector.

Theorem 3.51. *Suppose $\mathbb{E}XX^\dagger = I_m$, $\operatorname{supp} X \subset B(L)$, and that $s(X)$ is positive. Let $B : \mathbb{R}^m \to \mathbb{R}$ be a nonnegative function with $\|B\|_1 \le 1$ and $\|\widehat{B}\|_1 \le \infty$. If*

$$n \ge N_1 = c_{17} \max\left\{m^2 L^2 (\log M + \log L)^2, s(X)^{-4} L^{-2}, L^2 \log^2 \|B\|_1\right\},$$

then with probability at least $c_{13} L \sqrt{\frac{\log n}{n}}$ over the choice of M we have

$$|\mathbb{E}[B(Y_M)] - \mathbb{E}[B(Y)]| \le 2m^2 L^2 n^{-1}$$

and $MM^\dagger \succeq \frac{1}{2} n I_m$.

Proof. By Plancherel's theorem,

$$\mathbb{E}[B(Y_M)] - \mathbb{E}[B(Y)] = \int_{\mathbb{R}^m} \widehat{B}(\boldsymbol{\theta})(\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta}))d\boldsymbol{\theta}.$$

Again, we can split this into three terms:

$$\left|\int_{\mathbb{R}^m} \widehat{B}(\boldsymbol{\theta})(\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta}))d\boldsymbol{\theta}\right| \le \underbrace{\int_{B(\varepsilon)} |\widehat{B}(\boldsymbol{\theta})||\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{J_1} + \underbrace{\int_{R^m \setminus B(\varepsilon)} |\widehat{B}(\boldsymbol{\theta})\widehat{Y_M}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{J_2} + \underbrace{\int_{R^m \setminus B(\varepsilon)} |\widehat{B}(\boldsymbol{\theta})\widehat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta}}_{J_3}. \tag{3.23}$$

The proofs of the next two lemmas are identical to that of Lemma 3.46 and Lemma 3.47, respectively, except one uses the assumption $\|B\|_1 \le 1$, which implies $\|\widehat{B}\|_\infty \le 1$, to remove $\widehat{B}$ from the integrand.

Lemma 3.52 (First term)**.** *Suppose Eq.* (anticoncentration) *and Eq.* (concentration) *hold. Further suppose that* $L^2 n\varepsilon^4 < 1$, $\varepsilon < \frac{1}{2L}$, *and that* $\|B\|_1 \leq 1$. *There exists* c_{16} *with*

$$J_1 \leq c_{16} \frac{m^2 L^2 n^{-1}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.$$

Lemma 3.53 (Third term)**.** *Suppose Eq.* (anticoncentration) *holds,* $\varepsilon^2 \geq \frac{16m}{\pi^2 n}$, *and* $\|B\|_1 \leq 1$. *Then*

$$J_3 \leq \frac{e^{-\frac{\pi^2}{8}\varepsilon^2 n}}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}}.$$

The proof of the next lemma is the same as that of Lemma 3.47, except in the derivation of Eq. (3.16) instead of integrating over D one must integrate over the whole of $\mathbb{R}^m \setminus B(\varepsilon)$ against $\widehat{B}$, hence $\det L^*$ is replaced by $\|\widehat{B}\|_1$.

Lemma 3.54. *If* X *is in isotropic position and* $\varepsilon \leq s(X)$, *then*

$$\mathbb{E}[J_2] \leq \|\widehat{B}\|_1 e^{-2\varepsilon^2 n}$$

We now proceed to combine the termwise bounds. As before, we may choose $\varepsilon = n^{1/2}L^{-1/2}$ provided

$$n \geq (16mL)^2/\pi^4 \text{ and } n \geq s(X)^{-4}L^{-2},$$

and with probability at least $1 - \|\widehat{B}\|_1 e^{-\varepsilon^2 n} - c_5 L\sqrt{\frac{\log n}{n}}$, we have J_2 at most $e^{-\varepsilon^2 n}$ and Eq. (concentration), Eq. (anticoncentration) hold. Condition on these events. As in the proof of *Theorem* 3.18, we have

$$\left| \int_{\mathbb{R}^m} \widehat{B}(\boldsymbol{\theta})(\widehat{Y_M}(\boldsymbol{\theta}) - \widehat{Y}(\boldsymbol{\theta}))d\boldsymbol{\theta} \right| \leq \frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} \left(m^2 L^2 n^{-1} + 2e^{\frac{m}{2}\log(4\pi n) - \sqrt{n}/L} \right). \tag{3.24}$$

If c_{17} is large enough, the quantity in parentheses in Eq. (3.18) is at most $2m^2L^2/n$ and the combined failure probability of all the required events is at most $c_{13}L\sqrt{\frac{\log n}{n}}$ provided

$$n \geq N_1 = c_{17} \max\left\{ m^2 L^2 (\log M + \log L)^2, s(X)^{-4}L^{-2}, L^2 \log^2 \|B\|_1 \right\}. \tag{3.25}$$

□

3.5.2 Discrepancy for random unit columns

Lemma 3.55. *Let X be a random unit vector. Then*

$$s(X) \geq c_{18}.$$

for some fixed constant c_{18}.

Before we prove the lemma, let's show how to use it and Theorem 3.51 to prove discrepancy bounds.

Proof of Theorem 3.12. Let X be a random unit vector. We need to choose our function B.

Definition 3.56. For $K > 0$, let $B = \frac{1}{(2K)^{2m}} \mathbf{1}_{[-K,K]^m} * \mathbf{1}_{[-K,K]^m}$. Alternately, one can think of B as the density of a sum of two random vectors from the cube $[-K, K]^m$. $\lozenge$

It's not hard to show $\|B\|_1 = 1$ using that B is a density and that, by Plancherel's theorem, $\|\widehat{B}\|_1 = \frac{1}{(2K)^m}$. Next, we apply Theorem 3.51 to $Z = \sqrt{m}X$; in order to apply the theorem we need

$$n \geq N_2 := c_{19} \max\left\{m^3 \log^2 m, m^{-1}, m^3 \log^2(1/K)\right\}.$$

Thus, we may take

$$n \geq c_2 m^3 \log^2 m \text{ and } K = c_3 e^{-\sqrt{\frac{n}{m^3}}}.$$

We also need to obtain a lower bound on $\mathbb{E}[B(Y)]$ in order to use the bound on $|\mathbb{E}[B(Y_M)] - \mathbb{E}[B(Y)]|$ to deduce that $\mathbb{E}[B(Y_M)] > 0$, or equivalently that $\operatorname{disc} M \leq 2K$. The quantity $\mathbb{E}B(Y)$ is at least the least density of Y on the support of B. The support of B is contained within a $2K\sqrt{m}$ Euclidean ball. Using the property $MM^\dagger \geq \frac{1}{2} n I_m$, the density of Y takes value at least

$$\frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} e^{-2\sigma_{min}(MM^\dagger)^{-1} 4K^2 m} \geq \frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} e^{-16K^2 m/n}.$$

Since the error term in Theorem 3.51 is at most $\frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_M)}} 2m^2 L^2 n^{-1}$, to deduce $\operatorname{disc} M \leq K$ it is enough to show $e^{-16K^2 m/n} > 2m^3 n^{-1}$; indeed this is true with $K = c_3 e^{-\sqrt{\frac{n}{m^3}}}$ and $n \geq m^3 \log^2 m$ for suitably large c_3. $\square$

Spanningness for random unit vectors

We now lower bound the spanningness for random unit vectors. We use the fact that for large m the distribution of a random unit vector behaves much like a Gaussian upon projection to a one-dimensional subspace.

Proof of Lemma 3.55. Let $\|\boldsymbol{\theta}\|_X = \frac{1}{\sqrt{m}}\delta > 0$. We split into two cases. In the first, we show that if $\delta = O(\sqrt{m})$, then $\boldsymbol{\theta}$ is not pseudodual. In the second, we show that if $\delta = \Omega(\sqrt{m})$ then $\tilde{X}(\boldsymbol{\theta})$ is at least a fixed constant. This establishes that $s(X)$ is at least some constant.

By rotational symmetry we may assume $\boldsymbol{\theta}$ is δe_1, where e_1 is the first standard basis vector, so

$$\tilde{X}(\boldsymbol{\theta})^2 = \mathbb{E}[(\langle X, \boldsymbol{\theta}\rangle \bmod 1)^2] = \mathbb{E}[(\delta X_1 \bmod 1)^2].$$

We now try to show $\boldsymbol{\theta}$ is not a pseudodual vector if $\delta = O(\sqrt{m})$. Recall that X is a random unit vector; it is easier to consider the density of X_1. The probability density function of δX_1 for $x < \delta$ is proportional to $(1-(x/\delta)^2)^{\frac{m-3}{2}} =: f_\delta(x)$. Integrating this density gives us

$$\begin{aligned}\int_{-\delta}^{\delta}\left(1-\left(\frac{x}{\delta}\right)^2\right)^{\frac{m-3}{2}}dx &= \delta\int_0^{\delta}(1-x)^{\frac{m-3}{2}}x^{-\frac{1}{2}}dx\\ &= \frac{\delta\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\\ &= \frac{\delta\sqrt{\pi}}{\sqrt{m}}\left(1+o(1)\right).\end{aligned}$$

Therefore, we obtain

$$\mathbb{E}[(\delta X_1 \bmod 1)^2] = \frac{\Gamma\left(\frac{m}{2}\right)}{\delta\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\int_{-\delta}^{\delta}(x \bmod 1)^2\left(1-\left(\frac{x}{\delta}\right)^2\right)^{\frac{m-3}{2}}dx.$$

If we simply give up on all the x for which $|x| > 1/2$, we obtain the following lower

bound for the above quantity:

$$\begin{aligned}
&\frac{\Gamma\left(\frac{m}{2}\right)}{\delta\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\left[\int_{-\delta}^{\delta}x^2\left(1-\left(\frac{x}{\delta}\right)^2\right)^{\frac{m-3}{2}}dx-2\int_{1/2}^{\delta}x^2\left(1-\left(\frac{x}{\delta}\right)^2\right)^{\frac{m-3}{2}}dx\right]\\
&=\frac{\delta^2}{m}-\frac{2\Gamma\left(\frac{m}{2}\right)}{\delta\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}\int_{1/2}^{\delta}x^2\left(1-\left(\frac{x}{\delta}\right)^2\right)^{\frac{m-3}{2}}dx\\
&\geq\frac{\delta^2}{m}-(1+o(1))\frac{2\sqrt{m-3}}{\delta\sqrt{\pi}}\int_{1/2}^{\infty}x^2e^{-\frac{(m-3)x^2}{2\delta^2}}dx.\\
&=\frac{\delta^2}{m}-(1+o(1))\frac{2^{3/2}\delta^2}{m}\left(\frac{1}{\sqrt{2\pi}}\int_{\frac{\sqrt{m-3}}{2\delta}}^{\infty}u^2e^{-\frac{u^2}{2}}du\right)
\end{aligned}$$

The integral in parentheses is simply the contribution to the variance of the tail of a standard gaussian, and can be made an arbitrarily small constant by making $\delta/\sqrt{m}$ small. Thus, for δ at most $\delta \leq c_{20}\sqrt{m}$, the last line above expression is at least $.6^2\frac{\delta^2}{m} = .6^2\|\boldsymbol{\theta}\|_X^2$, so $\boldsymbol{\theta}$ is not pseudodual.

Next we must handle δ larger than $c_{20}\sqrt{m}$; we will show that in this case $\tilde{X}(\boldsymbol{\theta})$ is at least some fixed constant. We use the fact that f_δ is unimodal, so for any $k \neq 0$, $\int_{k-1/2}^{k+1/2}(x \bmod 1)^2 f_\delta(x)dx$ is at least the mass of $f_\delta(x)$ between $k-1/2$ and k times the integral of $(x \bmod 1)^2$ on this region (that is, $1/24$). This product is then at least which is at least one 48th of the mass of $f_\delta(x)$ between $k-1/2$ and $k+1/2$. Taken together, we see that

$$\mathbb{E}[(\delta X_1 \bmod 1)^2] \geq \frac{1}{48}\Pr[|\delta X_1| \geq 1/2]. \tag{3.26}$$

We will lower bound the left-hand side by a small constant for $\delta = \Omega(\sqrt{m})$. We can do so by bounding the ratio of $\int_{-1/2}^{1/2} f_\delta(x)$ to $\int_{1/2}^{\infty} f_\delta(x)$. To this end we will translate and scale the function

$$g_\delta(x) = \begin{cases} f_\delta(x) & x \geq 1/2 \\ 0 & x < 1/2 \end{cases}$$

to dominate $f_\delta(x)$ for $x \in [0, 1/2]$. Let us find the smallest scaling $a > 0$ such that $ag_\delta(x+1/2) \geq f_\delta(x)$ for $x \in [0, 1/2]$; equivalently, $af_\delta(x+1/2) \geq f_\delta(x)$ for $x \in [0, 1/2]$. If we find such an a, we'll have $\int_0^{1/2} f_\delta(x) \leq a\int_0^{\infty} g_\delta(x)dx$, or $\Pr[x \in [-1/2, 1/2]] \leq a(1 - \Pr[x \in [-1/2, 1/2]])$. We need

$$a(1-((x+1/2)/\delta)^2)^{\frac{m-3}{2}} \geq (1-(x/\delta)^2)^{\frac{m-3}{2}},$$

or

$$\begin{aligned} a &= \max_{x\in[0,1/2]} \left(\frac{1-(x/\delta)^2}{1-((x+1/2)/\delta)^2}\right)^{\frac{m-3}{2}} \\ &= \max_{x\in[0,1/2]} \left(\frac{\delta^2-x^2}{\delta^2-(x+1/2)^2}\right)^{\frac{m-3}{2}} \\ &\le \left(\frac{\delta^2}{\delta^2-1}\right)^{\frac{m-3}{2}} = \left(1-\frac{1}{\delta^2}\right)^{-\frac{m-3}{2}} \\ &\le e^{\frac{(m-3)}{2\delta^2}}. \end{aligned}$$

As discussed, we now have $\Pr[x \in [-1/2, 1/2]] \le a(1 - \Pr[x \in [-1/2, 1/2]])$. Equivalently, $\Pr[x \in [-1/2, 1/2]] \le a/(1+a)$. Therefore

$$\Pr[|x| > 1/2] \ge 1/(1+a) \ge .5e^{-\frac{m-3}{2\delta^2}}.$$

If $\delta \ge c_{20}\sqrt{m}$, this and Eq. (3.26) imply $\mathbb{E}[(\delta X_1 \bmod 1)^2] = \mathbb{E}[(\langle \boldsymbol{\theta}, X\rangle \bmod 1)^2$ is at least some constant. Thus, if $\delta \ge c_{20}$ then $\tilde{X}(\boldsymbol{\theta})$ is at least some constant. □

3.6 Open problems

There is still a gap in understanding for t-sparse vectors.

Question 3.57. *Let M be an $m \times n$ random t-sparse matrix. What is the least N such that for all $n \ge N$, the discrepancy of M is at most one with probability at least $1/2$? We know that for t not too large or small, $m \le N \le m^3 \log^2 m$. The lower bound is an easy exercise.*

Next, it would be nice to understand Question 3.3 for more column distributions in other regimes such as $n = O(m)$. In particular, it would be interesting to understand a distribution where combinatorial considerations probably won't work. For example,

Question 3.58. *Suppose M is a random t-sparse matrix plus some Gaussian noise of of variance $\sqrt{t/m}$ in each entry. Is $\operatorname{disc} M = o(\sqrt{t \log m})$ with high probability? How much Gaussian noise can existing proof techniques handle?*

The quality of the nonasymptotic bounds in this chapter depend on the spanningness of the distribution X, which depends on how far X is from lying in a proper sublattice

of $\mathcal{L}$. If X actually *does* lie in a proper sublattice of $\mathcal{L}' \subset \mathcal{L}$, we may apply our theorems with $\mathcal{L}'$ instead. This suggests the following:

Question 3.59. *Is there an N depending on only the parameters in Eq.* (3.22) other than spanningness *such that for all $n \geq N$,*

$$\operatorname{disc} M \leq \max_{\mathcal{L}' \subset \mathcal{L}} \rho_\infty(\mathcal{L}')$$

with probability at least $1/2$*?*

Next, the techniques in this chapter are suited to show that essentially any point in a certain coset of the lattice generated by the columns may be expressed as the signed discrepancies of a coloring. This is why we obtain twice the ℓ_∞-covering radius for our bounds. In order to bound the discrepancy, we must know $\rho_\infty(\mathcal{L})$. However, the following question (Conjecture 3.13 from the introduction) is still open, which prevents us from concluding that discrepancy is $O(1)$ for an arbitrary bounded distribution:

We could also study a random version of the above question:

Question 3.60. *Let $v_1, \ldots, v_m$ be drawn i.i.d from some distribution X on $\mathbb{R}^m$, and let $\mathcal{L}$ be their integer span. Is $\rho_\infty(\mathcal{L}) = O(1)$ with high probability in m?*

Here we also bring attention to an open–ended question asked in [KLP12, HR18]. Interestingly, though we use probabilistic tools to deduce the existence of low-discrepancy assignments, the proof does not yield any obvious efficient randomized algorithm to find them.

Question 3.61. *If an object can be proved to exist by a suitable local central limit theorem, is there an efficient randomized algorithm to find it?*

Chapter 4

A simple algorithm for Horn's problem and its cousins

This chapter is based on the work [Fra18a], which also appeared in the conference proceedings [Fra18b].

4.1 Introduction

A primary motivation for this chapter is a constructive variant of the following problem.

Problem 1.8 (Horn's problem)**.** *Which triples* $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ *of nonincreasing sequences of* m *real numbers are the respective spectra of* $m \times m$ *Hermitian matrices* A, B, C *satisfying*

$$A + B = C?$$

In a non-constructive sense, Horn's problem is solved: the desired Hermitian matrices exist for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ if and only if $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is in the *Horn polytope*, a certain polyhedral cone defined by a recursively defined set of linear inequalities with binary integer coefficients. This result, conjectured by Horn, was proved in [KT00] which culminated a long line of work by many authors; for the history we refer the reader to the survey [Ful00]. Horn's conjecture is remarkable because, a priori, one does not expect such $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ to form a polyhedral cone. The solution [KT00] of Horn's problem implies a polynomial time algorithm for the decision problem [MNS12], but the algorithm does not help construct the matrices.

We consider the problem of *constructing* Hermitian matrices A, B, C with respective spectra $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ satisfying $A+B+C = 0$. Exact solutions may involve irrational numbers which cannot be represented in finite space even if $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are rational, so we consider an approximate version.

Problem 4.1 (Constructive Horn's problem)**.** *Given a triple* $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ *of nonincreasing sequences of* m *real numbers,* either

1. *Correctly determine that* $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ *is not in the Horn polytope,* or
2. *Construct a sequence of triples of Hermitian matrices* $(A_k, B_k, C_k)_{k=1}^{\infty}$ *with respective spectra* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ *such that* $\lim_{k\to\infty} A_k + B_k - C_k = 0$.

For a given triple $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, exactly one of Tasks (1) and (2) is possible. That the impossibility of Task (1) implies the possibility of Task (2) is immediate from the definition of the Horn polytope. On the other hand, if Task (1) is possible, by compactness of the set of matrices of bounded spectral norm, the sequence (A_k, B_k, C_k) has a convergent subsequence whose limit (A, B, C) satisfies $A+B = C$ and has spectra $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, certifying that $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is in the Horn polytope.

We present a simple iterative algorithm (Algorithm 1) to solve Problem 4.1. Algorithm 1 solves a slightly different problem, but one to which Problem 4.1 can easily be reduced by a simple rescaling argument. [1]

Theorem 4.2 (Correctness of Algorithm 1)**.** *Let* $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3)$ *be a triple of nonincreasing sequences of* m positive *real numbers. Either*

1. *There do not exist Hermitian* $H_1, H_2,$ *and* H_3 *with respective spectra* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ *satisfying* $H_1 + H_2 + H_3 = I_m$*, or*
2. *For every* $\varepsilon > 0$*, with probability one, Algorithm 1 outputs real, symmetric matrices* $H_1, H_2,$ *and* H_3 *with respective spectra* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ *satisfying*

$$\|H_1 + H_2 + H_3 - I_m\| \leq \varepsilon. \tag{4.1}$$

Remark 4.3 (Rescaling argument)**.** To use Algorithm 1 to solve Problem 4.1 for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, let $\varepsilon_k \to 0$ and let $(H_1^k, H_2^k, H_3^k)_{k=1}^{\infty}$ be the sequence of outputs for $\boldsymbol{\lambda}_1 = \frac{1}{a}(\boldsymbol{\alpha} + b\mathbf{1}), \boldsymbol{\lambda}_2 = \frac{1}{a}(\boldsymbol{\beta} + b\mathbf{1}), \boldsymbol{\lambda}_3 = \frac{1}{a}(-\boldsymbol{\gamma} + (a - 2b)\mathbf{1})$ for $b > \max\{-\alpha_m, -\beta_m\}$ and

[1] This algorithm is slightly more convenient to analyze than the informal algorithm in Section 1.3, but both converge.

$a > 2b + \gamma_1$. Alternatively, one can use the sequence of (H_1, H_2, H_3) obtained in Step (b) of the algorithm with $\varepsilon = 0$. The sequence

$$(A^k, B^k, C^k)_{k=1}^{\infty} = \left(aH_1^k - bI_m, aH_2^k - bI_m, -aH_3^k + (2b-a)I_m\right)_{k=1}^{\infty}$$

is the solution to Problem 4.1. ◊

Input: Nonincreasing sequences $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ of m positive real numbers and $\varepsilon \geq 0$.

Output: A triple (H_1, H_2, H_3) of real, symmetric matrices with respective spectra $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ and

$$\|H_1 + H_2 + H_3 - I_m\| \leq \varepsilon.$$

Algorithm:

1. Choose each entry of $U_1, U_2, U_3 \in \mathrm{Mat}_m(\mathbb{R})$ independently and uniformly at random from $[-1, 1]$. For $i \in \{1, 2, 3\}$, **define** $H_i = U_i \operatorname{diag}(\boldsymbol{\lambda}_i) U_i^{\dagger}$ throughout.
2. **while** $\|H_1 + H_2 + H_3 - I_m\| > \varepsilon$ **do:**
   (a) Choose $B \in \mathrm{Mat}_m \mathbb{R}$ lower triangular such that
   $$B\,(H_1 + H_2 + H_3)\,B^{\dagger} = I_m,$$
   and set $U_i \leftarrow BU_i$ for $i \in \{1, 2, 3\}$.
   (b) For $i \in \{1, 2, 3\}$, choose $B_i \in \mathrm{Mat}_m(\mathbb{R})$ upper triangular such that $U_i B_i$ is orthogonal and set $U_i \leftarrow U_i B_i$.
3. **output** H_1, H_2, H_3.

Algorithm 1: Algorithm for Problem 4.1.

Algorithm 1 can be seen as a generalization of Sinkhorn's algorithm [Sin64] which, given a nonnegative matrix A, constructs a sequence (X^k, Y^k) of pairs of nonnegative diagonal matrices so that $X^k A Y^k$ is $1/k$–doubly stochastic (if possible). Sinkhorn's algorithm and its relatives fall into a framework within optimization known as *alternating minimization*. To minimize a function with several inputs, one alternates between minimizing the function on each input with the others fixed. We will see that Problem 4.1 can be cast as a minimization problem. The diagonalizing orthogonal matrices U_1, U_2, U_3 such that $H_i = U_i \operatorname{diag}(\boldsymbol{\lambda}_i) U_i^{\dagger}$ satisfies Eq. (4.1) are approximate solutions to the minimization problem. The steps (a) and (b) from Algorithm 1 arise as alternating

minimization steps; step (a) enforces *orthogonality* of the U_i, while step (b) enforces Eq. (4.1). One can find the upper triangular matrices using the Cholesky decomposition. The surprising aspect of the algorithm is that performing (a) does erase the progress made by (b), and vice versa.

Remark 4.4 (Implementability)**.** As written, Algorithm 1 is not implementable by a computer because it requires uniform samples from $[-1, 1]$, and it requires *exact* Cholesky decomposition (which may require irrational numbers). In reality, we must do these steps to some finite precision. We address these issues in Section 4.5. $\Diamond$

To analyze Algorithm 1, we prove convergence of a similar class of algorithms solving a common generalization of Sinkhorn's problem and Horn's problem. Our algorithm is very similar to the first algorithm, due to Gurvits [Gur04], for the "scaling" of completely positive maps.

Related work

A variant of Algorithm 1 was suggested in [GP15] in the more general setting of completely positive maps, which will be explained in the next section. Instead of the Cholesky decomposition, [GP15] uses square roots, and does not include a randomization step. In [Fri16] it was shown the algorithm in [GP15] converges for certain special cases that do not include Problem 4.1.

The follow–up work [BFG$^+$18] shows that Algorithm 1 extends naturally to finding tensors with prescribed marginals. In a forthcoming work with the same authors, we exhibit generalizations of the results of this chapter and [BFG$^+$18] to general moment polytopes.

Acknowledgements

The author would like to thank Michael Saks for many insightful discussions, and Rafael Oliveira for interesting observations and pointers to relevant literature. The author would like to further thank Ankit Garg, Rafael Oliveira, Avi Widgerson and Michael Walter for discussions concerning a forthcoming joint work that greatly simplified the

presentation of the reduction to the doubly stochastic case, and Christopher Woodward for helpful conversations.

Layout of the chapter

- In Section 4.2, we describe quiver representations, the general setting in which our algorithms work.
- In Section 4.3, we describe a reduction to simpler instances which will be used in the analysis of our algorithm. Section 4.3 alone implies that Algorithm 1 terminates in finite time on rational $\lambda_1, \lambda_2, \lambda_3$.
- In Section 4.4, we show our algorithms work even for irrational inputs. In the process we generalize Gurvits' theorem on scalability of completely positive maps.
- In Section 4.5 we analyze the running time of our algorithms, including an implementable version of Algorithm 1.

4.2 Completely positive maps and quiver representations

We will reduce the constructive version of Horn's problem to a problem known in computer science as *operator scaling* [GGOW16b, Gur04]. The objects to be considered are *completely positive maps*, linear maps between spaces of matrices that preserve positive-semidefiniteness in a strong sense. Completely positive maps describe the physically possible operations on a quantum system [NC02], but for our purposes a completely positive map $T : \mathrm{Mat}_{n\times n}(\mathbb{C}) \to \mathrm{Mat}_{m\times m}(\mathbb{C})$ is a map of the form

$$T : X \mapsto \sum_{i=1}^{r} A_i X A_i^\dagger,$$

where $A_i : \mathbb{C}^n \to \mathbb{C}^m, i \in [r]$ are linear maps called *Kraus operators* of T. Note that T preserves positive-semidefiniteness. Considered as a map between normed spaces, T has an adjoint T^* known as the *dual* of T, given by

$$T^* : X \mapsto \sum_{i=1}^{r} A_i^\dagger X A_i.$$

Example 4.5 (Nonnegative matrices as completely positive maps)**.** Suppose A is a nonnegative $m \times n$ matrix. For $i \in [m], j \in [n]$, define e_{ij} to be the $m \times n$ matrix with a one in the ij entry and zeros elsewhere. Let $T_{\sqrt{A}} : \mathrm{Mat}_{n\times n}\,\mathbb{C} \to \mathrm{Mat}_{m\times m}\,\mathbb{C}$ be the completely positive map with Kraus operators $E_{ij} = \sqrt{A_{ij}}e_{ij}$, $i \in [m], j \in [n]$. Then for $\boldsymbol{x} \in \mathbb{C}^n$,

$$T_{\sqrt{A}}(\mathrm{diag}(\boldsymbol{x})) = \mathrm{diag}(A\boldsymbol{x}).$$

If A is a Markov chain transition matrix, i.e. A is stochastic, then $T_{\boldsymbol{A}}$ is a *quantum channel*, a completely positive map that preserves traces. ◊

In analogy with Markov chains and nonnegative matrices, say T is *doubly stochastic* if both T and its dual send the identity to the identity, i.e. $T(I) = I$ and $T^*(I) = I$. Re-weighting the rows and columns of a nonnegative matrix by positive numbers is a natural "scaling" operation on nonnegative matrices. Analogously, say a completely positive map T' is a *scaling* of another completely positive map T by invertible maps $g \in \mathrm{GL}_m(\mathbb{C}), h \in \mathrm{GL}_n(\mathbb{C})$ if $T'(X) = g^\dagger T(hXh^\dagger)g$. That is, T' consists of T post- and pre-composed with changes of basis on the inner products on $\mathbb{C}^m, \mathbb{C}^n$ by g and h. Motivations in polynomial identity testing [Gur04, GGOW16b] and analysis [BCCT08] lead one to ask when T has a doubly stochastic scaling.

As in Example 4.5, often the domain and range of T have a direct sum structure that must be preserved by the scalings in order to capture the original problem. For instance, the classic matrix scaling problem considered by Sinkhorn [Sin64] is equivalent to whether $T_{\sqrt{A}}$ has a doubly stochastic scaling by *diagonal* matrices g, h. To formulate these restrictions on the scalings, we express our completely positive maps through *quiver representations*.

4.2.1 Quiver representations

Quiver representations are an especially convenient framework to express the problems solved by algorithms like Algorithm 1; we do not address deep issues usually considered in the study of quiver representations, nor do we use heavy tools from that field. [2]

[2] apart from a powerful degree bound of [DM17].

Definition 4.6 (Quiver representations)**.** A *quiver* Q is a directed multigraph (Ω, E) with vertex set Ω and edge set E. [3]

- If $e \in E$ is from x to y, then $t(e)$ denotes the *tail* x and $h(e)$ denotes the *head* y.
- A *representation* $\boldsymbol{V}$ of Q is an assignment $x \mapsto V_x$ of a finite dimensional complex inner product space to each vertex and an assigment $e \mapsto (A_e : V_{t(e)} \to V_{h(e)})$ of a linear map to each edge $e \in E$.
- The dimension vector $\boldsymbol{d} \in \mathbb{Z}^\Omega$ is the list of dimensions $\boldsymbol{d} : x \mapsto \dim V_x$. Throughout, if $\boldsymbol{x}$ is a vector we allow $\sum \boldsymbol{x}$ to denote the sum of its entries, and we define $\dim \boldsymbol{V} = \sum \boldsymbol{d}$.
- $\mathrm{Rep}(Q, \boldsymbol{d})$ denotes the space of quiver representations with dimension vector $\boldsymbol{d}$.

Our quiver representations will always have $V_x = \mathbb{C}^{d(x)}$ for $x \in \Omega$, meaning we have chosen some orthonormal bases for the vector spaces.

Example 4.7 (The Kronecker quiver and completely positive maps)**.** Representations of the *Kronecker quiver* with r arrows, as shown below, correspond to completely positive maps with r Kraus operators.

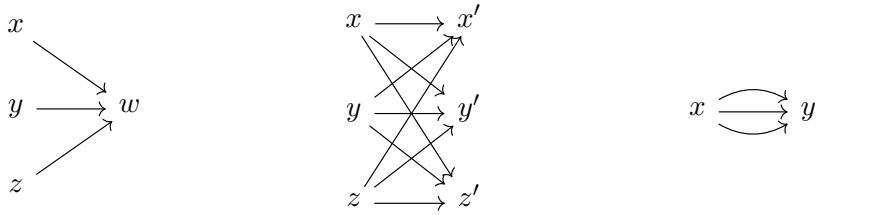
$$\boldsymbol{V} = \left(\mathbb{C}^n \underset{A_r}{\overset{A_1}{\overset{A_2}{\vdots}}} \mathbb{C}^m \right)$$

Here $A_i \in \mathrm{Mat}_{m \times n}(\mathbb{C})$. ◊

Our motivating examples come from only three families of quivers, seen in Fig. 4.1. The quiver in Fig. 4.1b is the *complete bipartite quiver* with three sources and three sinks, and the quiver in Fig. 4.1c is the Kronecker quiver with three edges. In particular, all the quivers of interest here have a bipartite structure.

Definition 4.8 (Bipartite quivers)**.** A quiver is *bipartite* if $\Omega = L \amalg R$, and every edge is from L to R.

[3] here we differ from the usual tradition of denoting the vertex set Q_0 and the edge set Q_1.



(a) Horn's problem (b) Matrix scaling (c) Operator scaling

Figure 4.1: Examples of relevant quivers

- A quiver representation is bipartite if its underlying quiver is bipartite.
- Whenever there is an assignment $\boldsymbol{x}$ of some objects to the vertices of Ω, $\boldsymbol{x}_L$ will denote the projection to only the vertices in L and $\boldsymbol{x}_R$ the projection to R.

For example, $\boldsymbol{d}_L \in \mathbb{Z}^L$ is the assignment $x \mapsto \dim V_x$, and so $\dim \boldsymbol{V}_L = \sum \boldsymbol{d}_L$.

Example 4.9. A bipartite quiver Q and a representation of Q.

$$Q = \begin{pmatrix} x & \longrightarrow & y \\ & \searrow & \\ z & \longrightarrow & w \end{pmatrix}, \boldsymbol{V} = \begin{pmatrix} \mathbb{C}^2 & \xrightarrow{[7 \quad 5]} & \mathbb{C}^1 \\ & \searrow^{[8i \quad 2]} & \\ \mathbb{C}^1 & \xrightarrow{[-5]} & \mathbb{C}^1 \end{pmatrix}$$

We write the dimension vector $\boldsymbol{d}$ of $\boldsymbol{V}$ as $(2, 1; 1, 1)$. ◊

We next discuss how to use a quiver $\boldsymbol{V}$ to define a map $T_{\boldsymbol{V}} : \bigoplus_{x \in L} \mathrm{Mat}_{d(x)}(\mathbb{C}) \to \bigoplus_{y \in R} \mathrm{Mat}_{d(y)}(\mathbb{C})$ between direct sums of matrix spaces, analogously to the completely positive maps between matrix spaces. We view

$$\bigoplus_{x \in L} \mathrm{Mat}_{d(x)}(\mathbb{C}) \text{ and } \bigoplus_{y \in R} \mathrm{Mat}_{d(y)}(\mathbb{C})$$

as included in $\mathrm{Mat}_{\dim \boldsymbol{V}_L}(\mathbb{C}), \mathrm{Mat}_{\dim \boldsymbol{V}_R}(\mathbb{C})$, respectively, as block diagonal matrices in the natural way. We want the maps $T_{\boldsymbol{V}}$ to be completely positive and to preserve the block structure; namely, if $\boldsymbol{X} \in \mathrm{Mat}_{\dim \boldsymbol{V}_L}(\mathbb{C})$ is block diagonal, i.e. in the image of $\bigoplus_{x \in L} \mathrm{Mat}_{d(x)}(\mathbb{C})$ under inclusion, then $T_{\boldsymbol{V}}(\boldsymbol{X})$ should also be. This leads one to consider Kraus operators in $\mathrm{Mat}_{\dim \boldsymbol{V}_L \times \dim \boldsymbol{V}_R}(\mathbb{C})$ that are nonzero on exactly one "block". We can describe such maps more directly as follows.

Definition 4.10 (Completely positive maps for bipartite quivers)**.** Using the completely positive map associated to the representation of a Kronecker quiver as in Example 4.7, to each bipartite quiver representation $\boldsymbol{V}$ we may assign a completely positive map

$$T_{\boldsymbol{V}} : \bigoplus_{x \in L} \mathrm{Mat}_{d(x)}(\mathbb{C}) \to \bigoplus_{y \in R} \mathrm{Mat}_{d(y)}(\mathbb{C}).$$

To do this, for each pair of vertices $x \in L$ and $y \in R$, consider the representation of the Kronecker quiver on $\{x, y\}$ by restricting $\boldsymbol{V}$ to only the edges incident to x and y. Let $T_{x,y} : \mathrm{Mat}_{d(x)}(\mathbb{C}) \to \mathrm{Mat}_{d(y)}(\mathbb{C})$ be the corresponding completely positive map. We then define

$$T_{\boldsymbol{V}} : \boldsymbol{X} \mapsto \bigoplus_{y \in R} \sum_{x \in L} T_{x,y}(X_x).$$

In words, to determine the y component $:= T_y(\boldsymbol{X})$ of $T_{\boldsymbol{V}}(\boldsymbol{X})$, sum $A_e X_{t(e)} A_e^\dagger$ over all edges e with head y. ◊

We now relate this to the discussion of block matrices before the definition. Though [Gur04, GGOW16b] only discuss the Kronecker quiver, it is well-known that, for quivers without oriented cycles, the problems we will consider can be reduced to those on Kronecker quivers by reductions of [DZ01]. For bipartite quivers, this reduction amounts to contracting L and R to single vertices x and y assigned the subspaces $\oplus_{x \in L} V_x, \oplus_{y \in R} V_y$, respectively, and replacing every edge e by an edge e' from x to y which is assigned the block matrix $A'_e \in \mathrm{Mat}_{\dim \boldsymbol{V}_L, \dim \boldsymbol{V}_R}(\mathbb{C})$ with A_e in the $h(e), t(e)$ block and zeroes elsewhere. For example,

$$\boldsymbol{V} = \left(\begin{array}{ccc} \mathbb{C}^2 & \xrightarrow{\begin{bmatrix} 7 & 5 \end{bmatrix}} & \mathbb{C}^1 \\ & \searrow^{\begin{bmatrix} 8i & 2 \end{bmatrix}} & \\ & & \mathbb{C}^1 \end{array} \right) \xmapsto{\text{reduction of [DZ01]}} \boldsymbol{V}' = \left(\begin{array}{ccc} & \begin{bmatrix} 7 & 5 \\ 0 & 0 \end{bmatrix} & \\ \mathbb{C}^2 & \rightleftarrows & \mathbb{C}^2 \\ & \begin{bmatrix} 0 & 0 \\ 8i & 2 \end{bmatrix} & \end{array} \right)$$

Definition 4.11 (Duals)**.** The quiver $\boldsymbol{V}^*$ is obtained by reversing the directions of all the arrows in V and replacing A_e by $A_e^\dagger$. The dual $T^*_{\boldsymbol{V}}$ of $T_{\boldsymbol{V}}$ is defined to be $T_{\boldsymbol{V}^*}$. ◊

Remark 4.12 (Inner products and positivity)**.** Viewing $\bigoplus_{x\in L} \mathrm{Mat}_{d(x)}(\mathbb{C})$ and $\bigoplus_{y\in R} \mathrm{Mat}_{d(y)}(\mathbb{C})$ as included as block-diagonal matrices in $\mathrm{Mat}_{\dim \boldsymbol{V}_L}(\mathbb{C}), \mathrm{Mat}_{\dim \boldsymbol{V}_R}(\mathbb{C})$, respectively, determines traces, inner products $\langle A, B\rangle := \operatorname{tr} A^\dagger B$, and positive-semidefiniteness on the two spaces. We write the induced norms as $\|\cdot\|$, and the Loewner orderings by $\succeq$. The map $T_{\boldsymbol{V}^*}$ is then the adjoint of $T_{\boldsymbol{V}}$; this follows from the formula

$$\operatorname{tr} \boldsymbol{Y} T_{\boldsymbol{V}}(\boldsymbol{X}) = \operatorname{tr} T^*_{\boldsymbol{V}}(\boldsymbol{Y})\boldsymbol{X}, \tag{4.2}$$

an easy consequence of the cyclic identity for trace. $T_{\boldsymbol{V}}$ maps positive-semidefinite elements to positive-semidefinite elements. ◊

Extending from completely positive maps, say a bipartite quiver representation $\boldsymbol{V}$ is *doubly stochastic* if $T_{\boldsymbol{V}}(\boldsymbol{I}_L) = \boldsymbol{I}_R$ and $T^*_{\boldsymbol{V}}(\boldsymbol{I}_R) = \boldsymbol{I}_L$, where $\boldsymbol{I}$ denotes the element $\bigoplus_{x\in\Omega} I_{V_x}$ To capture Horn's problem, we define a more general notion of balancedness.

Definition 4.13 (Quiver data)**.** Say $(\boldsymbol{V}, \boldsymbol{p})$ is a *datum* of the quiver Q if $\boldsymbol{V}$ is a representation of Q and $x \mapsto \boldsymbol{p}(x) \in \mathbb{R}^{d(x)}_{\geq 0} \setminus \{0\}$ is an assignment of nonzero, nonincreasing vectors with nonnegative entries to the vertices of Q.

- Say $(\boldsymbol{V}, \boldsymbol{p})$ is *positive* if all the entries of $\boldsymbol{p}$ are positive.
- P_x denotes the operator $\operatorname{diag}(\boldsymbol{p}(x))$ for $x \in \Omega$, and

$$\boldsymbol{P} := \bigoplus_{x\in\Omega} P_x \in \bigoplus_{x\in\Omega} P(V_x) := \boldsymbol{P}(\boldsymbol{V}).$$

- If Q is bipartite, say $(\boldsymbol{V}, \boldsymbol{p})$ is *ε-balanced* if

$$\|T_{\boldsymbol{V}}(\boldsymbol{P}_L) - \boldsymbol{I}_R\| \leq \varepsilon \text{ and } \|T^*_{\boldsymbol{V}}(\boldsymbol{P}_R) - \boldsymbol{I}_L\| \leq \varepsilon,$$

  and *balanced* if it is 0-balanced.

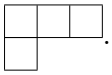
$(\boldsymbol{V}, \boldsymbol{1})$ denotes the datum where each vertex is assigned a vector of all ones. Clearly $(\boldsymbol{V}, \boldsymbol{1})$ is balanced if and only if $\boldsymbol{V}$ is doubly stochastic.

Remark 4.14 (Depicting quiver data)**.** Because we assume $V_x = \mathbb{C}^{d(x)}$, we do not depict the vector space assigned to each vertex, but rather depict quiver data as below

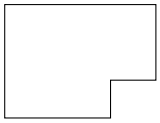
(with the maps suppressed if they are not relevant)

$$(\boldsymbol{V}, (\boldsymbol{p}_x, \boldsymbol{p}_y)) = \left(\begin{array}{c} \boldsymbol{p}_x \rightrightarrows \boldsymbol{p}_y \end{array} \right). \tag{4.3}$$

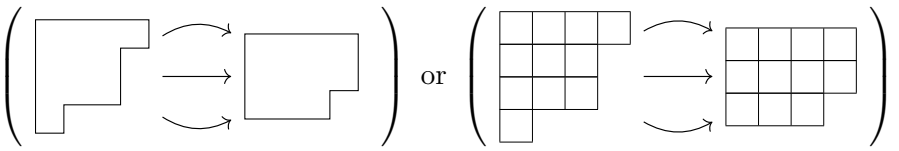
The nonincreasing sequences assigned by integral $\boldsymbol{p}$ are naturally viewed as partitions. Recall that a partition $\boldsymbol{\lambda}$ of a nonnegative integer l with k parts is a weakly decreasing sequence $(\lambda_1, \ldots, \lambda_k)$ of nonnegative integers summing to l. A partition $\boldsymbol{\lambda}$ is often depicted by a Young diagram, a left-justified collection of boxes with λ_i boxes in the i^{th} row from the top. For example, if $\boldsymbol{\lambda} = (3, 1)$, then

$\boldsymbol{\lambda} =$  .

We also use such diagrams to represent non-integral vectors $\boldsymbol{q}$, e.g if $\boldsymbol{q} = (2, 2, \sqrt{2})$ then

$\boldsymbol{q} =$ 

Thus, for $\boldsymbol{p}_x = (4, 3, 3, 1), \boldsymbol{p}_y = (4, 4, 3)$ we may pictorially represent $(\boldsymbol{V}, (\boldsymbol{p}_x, \boldsymbol{p}_y))$ of Eq. (4.3) as



Example 4.15 (Horn's problem)**.** Consider the quiver

$$Q = \left(\begin{array}{ccc} 1 & & \\ & \searrow & \\ 2 & \longrightarrow & 0 \\ & \nearrow & \\ 3 & & \end{array} \right).$$

Let $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ be nonincreasing sequences of m nonnegative numbers. There exist $m \times m$ Hermitian matrices A, B, C with respective spectra $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ satisfying $A + B + C = I_m$ if and only if there is a balanced datum $(\boldsymbol{V}, \boldsymbol{p})$ of Q with dimension vector $\boldsymbol{d} = (m, m, m; m)$ and $\boldsymbol{p} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{1}_m)$. Indeed, if the below datum

$$(\boldsymbol{V}, \boldsymbol{p}) = \left(\begin{array}{ccc} \boldsymbol{\alpha} & & \\ & \overset{U}{\searrow} & \\ \boldsymbol{\beta} & \xrightarrow{V} & \mathbf{1}_m \\ & \underset{W}{\nearrow} & \\ \boldsymbol{\gamma} & & \end{array} \right).$$

Is balanced then $U^\dagger U, V^\dagger V, W^\dagger W$ are the identity, so U, V, W are unitary, and

$$U \operatorname{diag}(\boldsymbol{\alpha}) U^\dagger + V \operatorname{diag}(\boldsymbol{\beta}) V^\dagger + W \operatorname{diag}(\boldsymbol{\gamma}) W^\dagger = I_m,$$

so we may take $A = U \operatorname{diag}(\boldsymbol{\alpha}) U^\dagger$, etc. Likewise, if $A = U \operatorname{diag}(\boldsymbol{\alpha}) U^\dagger$ etc, for U, V, W unitary, then assigning U, V, W to the edges of Q make $(\boldsymbol{V}, \boldsymbol{p})$ balanced.

4.2.2 Scaling of quiver representations

We next consider a right action of products of linear groups on quivers, generalizing the action of $\mathrm{GL}_m(\mathbb{C}) \times \mathrm{GL}_n(\mathbb{C})$ on completely positive maps.

Definition 4.16 (Scalings of quiver representations)**.** If $\boldsymbol{V}$ is a representation of the quiver Q, then $\mathrm{GL}(\boldsymbol{V})$ denotes $\prod_{x \in \Omega} \mathrm{GL}(V_x)$, the elements of which are assignments $g : x \mapsto g_x$ of invertible matrices to the elements of Ω.

- Let $S_g : \mathrm{Rep}(Q, \boldsymbol{d}) \to \mathrm{Rep}(Q, \boldsymbol{d})$ denote the linear map sending $\boldsymbol{V}$ to the quiver $S_g \boldsymbol{V}$ assigning $g_{h(e)}^\dagger A_e g_{t(e)}$ to edge $e \in E$.
- If G is a subgroup of $\mathrm{GL}(\boldsymbol{V})$, $\boldsymbol{W}$ is called a *scaling* of $\boldsymbol{V}$ by G if there exists $g \in G$ with $S_g \boldsymbol{V} = \boldsymbol{W}$, i.e. $\boldsymbol{W}$ is in the orbit of $\boldsymbol{V}$ under G.

If $\boldsymbol{V}$ is bipartite, let $\mathrm{GL}(\boldsymbol{V})_L, \mathrm{GL}(\boldsymbol{V})_R$ denote $\prod_{x \in L} \mathrm{GL}(V_x), \prod_{x \in R} \mathrm{GL}(V_x)$, respectively, so that $\mathrm{GL}(\boldsymbol{V}) = \mathrm{GL}(\boldsymbol{V})_L \times \mathrm{GL}(\boldsymbol{V})_R$. ◊

We next show how the scaling of quiver representations recovers the scaling of matrices considered in [Sin64].

Example 4.17 (Matrix scaling)**.** Consider the complete bipartite quiver $K_{[n] \to [n]}$ with vertices $[n] \amalg [n]$ with one edge xy with $t(xy) = x, h(xy) = y$ for each pair $(x, y) \in [n] \times [n]$. If $\boldsymbol{V}$ is a representation of this quiver with $\boldsymbol{d} = \mathbf{1}$, i.e. $V_x = \mathbb{C}$ for all $x \in [n] \amalg [n]$, then $(A_{xy} : (x, y) \in [n] \times [n])$ is an array of complex numbers naturally viewed as a complex matrix, and $T_{\boldsymbol{V}} : \mathbb{C}^n \to \mathbb{C}^n$ is given by

$$T_y(v) = \sum_{x \in [n]} |A_{xy}|^2 v_x.$$

Thus, we view $T_{\boldsymbol{V}}$ as the nonnegative matrix $A = (|A_{xy}|^2 : (x, y) \in [n] \times [n])^\dagger$. Then $T^*_{\boldsymbol{V}}$ is the nonnegative matrix $A^\dagger$. Now $\mathrm{GL}(\boldsymbol{V})_L = \mathrm{GL}(\boldsymbol{V})_R = \mathbb{C}^n_\times$ and for $g, h \in \mathrm{GL}(\boldsymbol{V})_L \times \mathrm{GL}(\boldsymbol{V})_R$, the completely positive map associated to $\boldsymbol{V}_{(g,h)}$ is the nonnegative matrix $D_h A D_g$ where $D_h = \mathrm{diag}(|h_x|^2 : x \in [n])$ and $D_g = \mathrm{diag}(|g_x|^2 : x \in [n])$.

Problem 4.18 (Balanced scalings of quivers)**.** *Given a bipartite quiver datum* $(\boldsymbol{V}, \boldsymbol{p})$*, find (if possible) an* ε*–balanced scaling* $\boldsymbol{W}$ *of* $\boldsymbol{V}$ *for every* $\varepsilon > 0$*.*

The $\boldsymbol{p}$ for which this problem has a solution form a convex polytope known as the *moment polytope* of the orbit of $\boldsymbol{V}$ for the action of $U(\boldsymbol{V})$ (the unitary subgroup of $\mathrm{GL}(\boldsymbol{V})$), and the corresponding *moment map* is $\boldsymbol{V} \mapsto (T_{\boldsymbol{V}}(\boldsymbol{I}_L), T^*_{\boldsymbol{V}}(\boldsymbol{I}_R))$ [NM84]. We do not use technology for moment maps here, as it is largely non-algorithmic.

Remark 4.19 (Trace condition)**.** Eq. (4.2) implies that

$$\mathrm{tr}\, \boldsymbol{P}_L = \sum \boldsymbol{p}_L = \sum \boldsymbol{p}_R = \mathrm{tr}\, \boldsymbol{P}_R \tag{4.4}$$

is necessary for a positive solution to Problem 4.18. ◊

Example 4.20 (Scaling for Horn's problem)**.** If the following datum has a

$$(\boldsymbol{V}, \boldsymbol{p}) = \left(\begin{array}{ccc} \boldsymbol{\alpha} & \overset{I_m}{\searrow} & \\ \boldsymbol{\beta} & \xrightarrow{I_m} & \mathbf{1}_m \\ \boldsymbol{\gamma} & \underset{I_m}{\nearrow} & \end{array} \right).$$

has a balanced scaling by $(g_1, g_2, g_3; g)$, then by the calculation in Example 4.22 the matrices $g_1^\dagger g, g_2^\dagger g, g_3^\dagger g$ the diagonalizing unitaries needed for Horn's problem.

We provide an algorithm (Algorithm 2) to solve Problem 4.18 to within arbitrary precision, if possible. Algorithm 1 is instance of Algorithm 2 applied to the quiver representation from Example 4.20. The informal algorithm in Section 1.3 is roughly Algorithm 2 applied to the quiver datum

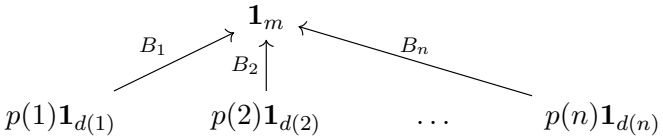
$$(\boldsymbol{V}, \boldsymbol{p}) = \left(\begin{array}{ccc} \boldsymbol{\alpha} & \overset{C^{-1/2}}{\searrow} & \\ & & \boldsymbol{\gamma} \\ \boldsymbol{\beta} & \underset{C^{-1/2}}{\nearrow} & \end{array} \right),$$

where $C = \mathrm{diag}(\gamma)$ and with the change of variables $g_x \leftarrow C^{1/2} g_x C^{-1/2}$ applied to the scalings in step (a).

Definition 4.21 (Scalability)**.** Say $(\boldsymbol{V}, \boldsymbol{p})$ is *approximately scalable* if there exists an ε–balanced scaling of $(\boldsymbol{V}, \boldsymbol{p})$ for every $\varepsilon > 0$. ◊

Note that $(\boldsymbol{V}, \mathbf{1})$–scalability is the same as scalability to doubly stochastic.

Example 4.22 (Brascamp Lieb data, Forster's problem)**.** Let $\boldsymbol{p} \in \mathbb{R}^n_{\geq 0}$. The quiver datum



is called a *Brascamp Lieb datum*, and is approximately scalable if and only if there is a positive constant C such that

$$\int_{\mathbb{R}^m} \prod_{i=1}^n f_i \left(B_i^\dagger x \right)^{p_i} dx \leq C \prod_{i=1}^n \left(\int_{\mathbb{R}^{d(i)}} f_j(x) dx \right)^{p_j}$$

for all nonnegative functions $f_j : \mathbb{R}^{d(j)} \to \mathbb{R}_{\geq 0}, j \in [n]$ [Bar98, BCCT08]. This class of inequalities is called Brascamp-Lieb inequalities. Algorithms for finding C were found in [GGOW16a]. Forster [For02] used that a generic datum with $\boldsymbol{d} = (\mathbf{1}_n; m)$ and $\boldsymbol{p} = \mathbf{1}_n$ is approximately scalable, first proved in [GS02], to prove communication complexity lower bounds. ◊

Theorem 4.23 (Correctness of Algorithm 2)**.** *Let $(\boldsymbol{V}, \boldsymbol{p})$ be a positive bipartite quiver datum. Either*

1. *$(\boldsymbol{V}, \boldsymbol{p})$ is not approximately scalable, or*
2. *For every $\varepsilon > 0$, with probability one, Algorithm 2 outputs an ε–balanced scaling of $(\boldsymbol{V}, \boldsymbol{p})$.*

Algorithm 2 applied to the quiver from Example 4.22 is exactly Algorithm 1.

Remark 4.24 (Implementability II)**.** Algorithm 2 suffers from the same implementability issues as Algorithm 1 highlighted in Remark 4.4. Both are addressed in Section 4.5.

Input: A bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{p})$ and $\varepsilon > 0$.

Output: An ε–balanced scaling $\boldsymbol{W}$ of $\boldsymbol{V}$.
Algorithm:

1. Choose $g \in \mathrm{GL}(\boldsymbol{V})$ by picking each entry of g_x independently at random in $[-1, 1]$ for $x \in \Omega$. Set $\boldsymbol{W} = S_g \boldsymbol{V}$.
2. **while** $\|T_{\boldsymbol{W}}(\boldsymbol{P}_L) - \boldsymbol{I}_R\| > \varepsilon$ **do:**
   (a) Choose $g \in \mathrm{GL}(\boldsymbol{V})_R$ upper triangular such that
   $$g^\dagger T_{\boldsymbol{W}}(\boldsymbol{P}_L) g = \boldsymbol{I}_R,$$
   and set $\boldsymbol{W} \leftarrow S_{(\boldsymbol{I}_L, g)} \boldsymbol{W}$.
   (b) Choose $h \in \mathrm{GL}(\boldsymbol{V})_L$ upper triangular such that
   $$h^\dagger T^*_{\boldsymbol{W}}(\boldsymbol{P}_R) h = \boldsymbol{I}_L,$$
   and set $\boldsymbol{W} \leftarrow S_{(h, \boldsymbol{I}_R)} \boldsymbol{W}$.
3. **output** $\boldsymbol{W}$.

Algorithm 2: Algorithm for Problem 4.18.

4.3 Reduction to Gurvits' problem

The analysis of the algorithm hinges on two steps: firstly, finding arbitrarily precise scalings of $(\boldsymbol{V}, \boldsymbol{p})$ is equivalent to a approximate scaling of $(S_g \boldsymbol{V}, \boldsymbol{p})$ by a smaller group for generic g, hence the randomization step in Algorithm 2. The smaller group is a group of block–upper triangular matrices in a certain basis.

The second step is a further reduction from Problem 4.26 to an instance of Problem 4.18 where $\boldsymbol{p} = \boldsymbol{1}$, which was already addressed by Gurvits. The simple alternating algorithm in Theorem 4.27 applied to the quiver obtained from the reduction "lifts" to the upper triangular scaling steps of Algorithm 2. This reduction only works when $\boldsymbol{p}$ is integral, so by the end of this section the reader will be convinced that Algorithm 2 works for rational $\boldsymbol{p}$. We proceed to set some notation to define the new scaling problem.

Upper–triangular matrices in $\mathrm{GL}_n(\mathbb{C})$ are exactly those that fix *the standard flag*

$$F_\bullet = \{\langle e_1\rangle, \langle e_1, e_2\rangle, \ldots, \langle e_1, \ldots, e_n\rangle\}.$$

More generally, a partial flag $F_\circ$ on a vector space V is a set of nontrivial subspaces that form a chain in the inclusion ordering, and a complete flag $F_\bullet$ is a partial flag of cardinality n. If $F_\circ$ is a partial flag on V, let $\mathrm{GL}(F_\circ)$ denote the subgroup of $\mathrm{GL}(V)$ fixing $F_\circ$, sometimes called *a parabolic subgroup* of $\mathrm{GL}(V)$. Concretely, $\mathrm{GL}(F_\circ) \subset \mathrm{GL}_n(\mathbb{C})$ consists of invertible, block–upper triangular matrix with blocks demarcated by $\{i : E + i \in F_\circ\}$.

Definition 4.25 (Flags and sequences)**.** If $\boldsymbol{p}$ is a decreasing subsequence of length n, let $F_\circ^{\boldsymbol{p}}$ denote the partial flag $\{F_i : p(i) \neq p(i+1), i \in [n]\}$ where $p(n+1) := 0$. $\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$ denotes the group $\prod_{x\in\Omega} \mathrm{GL}(F_\circ^{\boldsymbol{p}_x}) \subset \mathrm{GL}(V)$. That is, each g_x is an invertible, block–upper triangular matrix with blocks demarcated by $\{i : p_i(x) \neq p_{i+1}(x)\}$. $\lozenge$

We are ready to state the problem which generically captures scalability of $(\boldsymbol{V}, \boldsymbol{p})$.

Problem 4.26 (Approximate parabolic scaling)**.** *Given a bipartite quiver datum* $(\boldsymbol{V}, \boldsymbol{p})$ *find (if possible) for every* $\varepsilon > 0$ *an* ε*–balanced scaling of* $(\boldsymbol{V}, \boldsymbol{p})$ *by* $\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$.

The main purpose of using block–upper triangular matrices rather than upper–triangular ones is to avoid using unnecessary randomness, and so the problem is the same as approximate scaling for $\boldsymbol{p} = \mathbf{1}$. If a solution for Problem 4.26 exists for $(\boldsymbol{V}, \boldsymbol{p})$, say $(\boldsymbol{V}, \boldsymbol{p})$ is *approximately parabolic scalable.*

Theorem 4.27 ([Gur04])**.** *Let* $(\boldsymbol{V}, \mathbf{1})$ *be a datum of a bipartite quiver* Q.

1. $(\boldsymbol{V}, \mathbf{1})$ *is approximately scalable.*
2. $\boldsymbol{V}$ *is rank–nondecreasing, i.e. for* $\boldsymbol{X} \succeq 0$, $\operatorname{rank} T_{\boldsymbol{V}}(\boldsymbol{X}) \geq \operatorname{rank} \boldsymbol{X}$.
3. $\dim \boldsymbol{V}_L = \dim \boldsymbol{V}_R$ *and* $0 < \operatorname{cap}(\boldsymbol{V}) := \inf_{\boldsymbol{X}\succ 0} \frac{\det T_{\boldsymbol{V}}(\boldsymbol{X})}{\det \boldsymbol{X}}$.
4. *For all* $\varepsilon > 0$*, the below algorithm outputs an* ε*-scaling* $\boldsymbol{W}$ *of* $\boldsymbol{V}$.
   1. *Set* $\boldsymbol{W} \leftarrow \boldsymbol{V}$.
   2. ***while*** $\|T_{\boldsymbol{V}}(\boldsymbol{I}_L) - \boldsymbol{I}_R\| > \varepsilon$ ***do:***

(a) Choose any $g \in \mathrm{GL}(\boldsymbol{V})_R$ *such that* $g^\dagger T_{\boldsymbol{V}}(\boldsymbol{I}_L)g = \boldsymbol{I}_R$*, and set* $\boldsymbol{W} \leftarrow S_{(I_L,g)}\boldsymbol{W}$*.*

(b) Choose any $h \in \mathrm{GL}(\boldsymbol{V})_L$ *such that* $h^\dagger T^*_{\boldsymbol{V}}(\boldsymbol{I}_R)h = \boldsymbol{I}_R$*, and set* $\boldsymbol{W} \leftarrow S_{(h,I_R)}\boldsymbol{W}$*.*

3. ***output*** S*.*

Gurvits did not originally state this theorem in the language of quivers, but rather in the language of completely positive maps. However, it is very easy to prove the slightly more general Theorem 4.27 from Gurvits' original formulation using the standard reduction of [DZ01] from a bipartite quiver to a Kronecker quiver as in Example 4.7.

The combination of the reduction to parabolic scaling and the reduction from parabolic scaling to Gurvits' problem will show that Algorithm 2 converges on approximately scalable data $(\boldsymbol{V}, \boldsymbol{p})$. In the process of analyzing the running time, however, we will provide (with hindsight) a more self–contained proof mirroring the proof of Theorem 4.27.

4.3.1 Reduction from parabolic scaling to Gurvits' problem

In this section we prove Theorem 4.28 below. The reduction for Brascamp–Lieb data in [GGOW16a] is a special case of our reduction, which also bears some similarity to the reduction from representation theoretic version of Horn's problem to quiver semi–invariants in [DW00].

Say a bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{p})$ is *integral* is $\boldsymbol{p}$ is an integer vector.

Theorem 4.28. *There exists a map* Red *assigning to each integral, bipartite quiver datum* $(\boldsymbol{V}, \boldsymbol{p})$ *a bipartite quiver* $\boldsymbol{V}' = \mathrm{Red}(\boldsymbol{V}, \boldsymbol{p})$ *and to each element* $g \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$ *an element* $\mathrm{Red}\, g \in \mathrm{GL}(\boldsymbol{V}')$ *such that*

1. $(\boldsymbol{V}', \mathbf{1})$ *is approximately scalable if and only if* $(\boldsymbol{V}, \boldsymbol{p})$ *is approximately parabolic scalable.*

2. *If the group elements in Algorithm 3 on input* $(\boldsymbol{V}, \boldsymbol{p})$ *are* $g_1, g_2, \ldots$ *then* $\mathrm{Red}\, g_1, \mathrm{Red}\, g_2, \ldots$ *are valid choices in Algorithm 3 on input* $(\boldsymbol{V}', \mathbf{1})$*.*

In in particular, Algorithm 3 on $(\boldsymbol{V}', \mathbf{1})$ is identical to the algorithm in Theorem 4.27

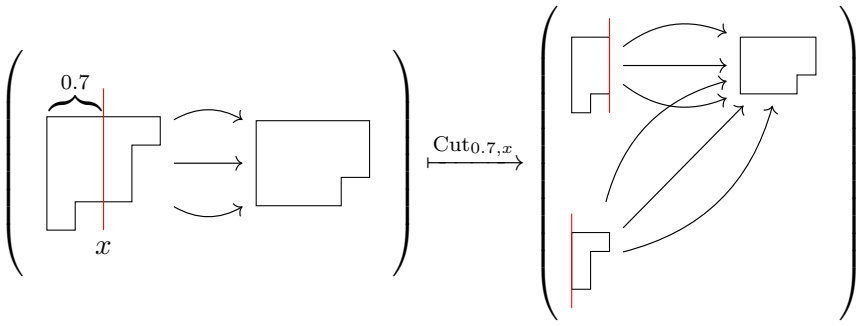
Input: A bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{p})$ and $\varepsilon > 0$.

Output: An ε–balanced parabolic scaling $\boldsymbol{W}$ of $\boldsymbol{V}$.
Algorithm:

1. Set $\boldsymbol{W} \leftarrow \boldsymbol{V}$.
2. **while** $(\boldsymbol{W}, \boldsymbol{p})$ is not ε–balanced, **do:**
   (a) Choose $h \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})_R$ such that
   $$h^\dagger T_{\boldsymbol{W}}(\boldsymbol{P}_L) h = \boldsymbol{I}_R,$$
   set $g = (\boldsymbol{I}_L, h)$, and set $\boldsymbol{W} \leftarrow S_g \boldsymbol{W}$.
   (b) Choose $h \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})_L$ such that
   $$h^\dagger T^*_{\boldsymbol{W}}(\boldsymbol{P}_R) h = \boldsymbol{I}_L,$$
   set $g = (h, \boldsymbol{I}_R)$, and set $\boldsymbol{W} \leftarrow S_g \boldsymbol{W}$.
3. **output** $\boldsymbol{W}$.

Algorithm 3: Algorithm for Problem 4.26.

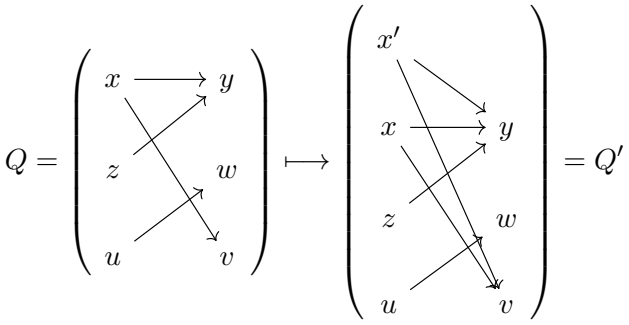
on $\boldsymbol{V}'$. The reduction is easy to describe. It will be a sequence of "cuts", denoted $\mathrm{Cut}_{p,x}$, which operate on one vertex of a datum at a time as shown below.



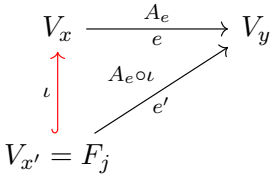
Definition 4.29 (Cuts)**.** Given a bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{p})$ of Q, $x \in L$ and $0 \leq p < p_1(x)$, let $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ be a datum $(\boldsymbol{V}', \boldsymbol{p}')$ for the quiver Q' described as follows.

1. Q' contains all edges and vertices of Q, but Q' has one new vertex x' that is a duplicate of x. That is, for each edge e with $t(e) = x$ there is a new edge e' with

$t(e') = x'$ and $t(e') = z$. E.g.



2. $\boldsymbol{V}'$ is the representation of Q' assigning the same values to the edges and vertices of Q, and assigning to x' the element F_j of the standard flag for $j = |\{i : p_i(x) > p\}|$, and to each edge e' the map $A_e \circ \iota$ where $\iota : F \hookrightarrow V_x$ is the inclusion map.



3. $\boldsymbol{p}'$ assigns all vertices of Q' but x and x' their original values, and to x it assigns the sequence $\boldsymbol{p}'(x) = (\min\{p, p_i(x) : i \in [d(x)])$ and to x' the sequence $\boldsymbol{p}'(x') = (p_i(x) - p : i \in [d(x')])$. For example,

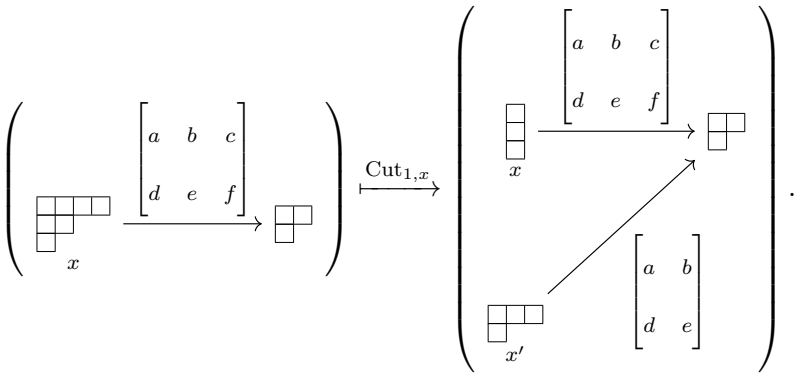
$$\underset{\boldsymbol{p}_x}{(3,2,1)} \overset{p=1.5}{\longmapsto} \underset{\boldsymbol{p}'_x}{(1.5,1.5,1)}, \underset{\boldsymbol{p}'(x')}{(1.5,.5)}$$

4. For $x \in \Omega'$, if $\boldsymbol{p}_x = 0$ then remove x from Q'_0 and all its incoming and outgoing edges. Note that this step will only change Q' if $p = 0$.

For fixed $\boldsymbol{p}$, we abuse notation by allowing $\mathrm{Cut}_{p,x} : \mathrm{Rep}(Q, \boldsymbol{d}) \to \mathrm{Rep}(Q', \boldsymbol{d}')$ to denote the injective, linear map that sends $\boldsymbol{V} \to \boldsymbol{V}'$. ◊

Example 4.30. Below $\boldsymbol{p}(x) = (4, 2, 1)$, which contains two values bigger than one.

Hence $V_{x'} = F_2 = \operatorname{span}\{e_1, e_2\}$, so $\iota = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^\dagger$ and



◊

Remark 4.31 (Observations about cuts)**.**

- By design, one recovers $\boldsymbol{p}(x)$ by adding $\boldsymbol{p}'(x)$ to $\boldsymbol{p}''(x)$ padded with the appropriate number of zeroes. In particular, $\sum \boldsymbol{p} = \sum \boldsymbol{p}'$ and $P_x = P'_x + \iota P'_{x'} \iota^\dagger$.
- So far we have only seen how to perform "left cuts" on vertices of $\boldsymbol{L}$ to perform a cut on a vertex in R, simply do the cut on the *dual* $(\boldsymbol{V}, \boldsymbol{p})^* := (\boldsymbol{V}^*, \boldsymbol{p}_R \oplus \boldsymbol{p}_L)$ and then take the dual of the result.
- Some cuts do not change the quiver, i.e. $\operatorname{Cut}_{0,x}$ when $\boldsymbol{p}_{d(x)} > 0$. Call these trivial cuts.
- Cuts commute. If $x \neq y$, then $\operatorname{Cut}_{p,x} \operatorname{Cut}_{p',y} = \operatorname{Cut}_{p',y} \operatorname{Cut}_{p,x}$. Similarly, for $p < p'$, $\operatorname{Cut}_{p,x} \operatorname{Cut}_{p',x} = \operatorname{Cut}_{p'-p,x'} \operatorname{Cut}_{p,x}$.

Before describing more properties of cuts, we describe the reduction. If the entries of $\boldsymbol{p}$ are positive integers, we may "cut" a datum to the point that the datum is of the form $(\boldsymbol{W}, \mathbf{1})$. If we only cut by integers, we may only perform nontrivial cuts for so long, and we may only do it in one way.

Definition 4.32 (Reduction as a sequence of cuts)**.** Let $(\boldsymbol{V}, \boldsymbol{p})$ be a bipartite quiver datum. $\operatorname{Red}(V, \boldsymbol{p})$ is the unique quiver representation obtained by any maximal sequence of nontrivial cuts of the form $\operatorname{Cut}_{i,x}$ where i is a positive integer. By Remark 4.31, $\dim \operatorname{Red} V = \sum \boldsymbol{p}$. ◊

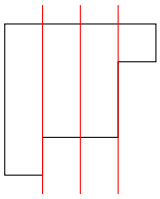


Figure 4.2: Maximal sequence of cuts of $(4, 3, 3, 1)$

Using the first observation in Remark 4.31, we can show cuts preserve balancedness.

Lemma 4.33. *If $x \in L$ and $(\boldsymbol{V}', \boldsymbol{p}') = \mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$, then*

$$T_{\boldsymbol{V}'}(\boldsymbol{P}'_{L'}) = T_{\boldsymbol{V}}(\boldsymbol{P}_L) \tag{4.5}$$

Furthermore, $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ is left (resp. right)–balanced if and only if $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ is left (resp. right)–balanced, and if $p > 0$ then $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ is ε–balanced then $(\boldsymbol{V}, \boldsymbol{p})$ is ε–balanced.

Proof. Let $(\boldsymbol{V}', \boldsymbol{p}') = \mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$. Eq. (4.5) implies cuts preserve left–balancednes. To prove the equation, enough to show that, for a given neighbor y of x, $T_{x,y}(P_x) = T_{x,y}(P'_x) + T_{x',y}(P'_{x'})$. This follows because, by Remark 4.31, $p_x = P'_x + \iota P'_{x'} \iota^\dagger$, and $T_{x',y}(X) = T_{x',y}(\iota X \iota^\dagger)$. To check that right–balancedness is preserved, observe that $T'^*(\boldsymbol{P}'_R) = T'^*(\boldsymbol{P}_R)$ and we need only check x, x'. We have $T'^*(\boldsymbol{P}_R)_x = T^*(\boldsymbol{P}_R)_x$ and $T'^*(\boldsymbol{P}_R)_{x'} = \iota^\dagger T^*(\boldsymbol{P}_R)_x \iota$, which are both the identity if and only if $T^*(\boldsymbol{P}_R)_x$ is. □

Further, if $(\boldsymbol{W}, \boldsymbol{p}) = \mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$, then $(\boldsymbol{V}, \boldsymbol{p})$ is approximately (resp. exactly) parabolically scalable if and only if $(\boldsymbol{W}, \boldsymbol{p})$ is approximately (resp. exactly) scalable by a certain subgroup of $\mathrm{GL}(\boldsymbol{W}, \boldsymbol{p})$.

Definition 4.34 (Cutting the group)**.** Let $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p}) = (\boldsymbol{V}', \boldsymbol{p}')$. By abuse of notation, define the map $\mathrm{Cut}_{p,x} : \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \to \mathrm{GL}(\boldsymbol{V}', \boldsymbol{p}')$ by $(\mathrm{Cut}_{p,x}\, g)_z = g_z$ if $z \neq x'$, and

$$(\mathrm{Cut}_{p,x}\, g)_{x'} = \iota^\dagger g_{x'} \iota.$$

$\mathrm{Cut}_{p,x}\, \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \subset \mathrm{GL}(\boldsymbol{V}', \boldsymbol{p}')$ denotes the image of this map. The last claim follows from Eq. (4.5) and the fact that $\|T'^*(\boldsymbol{P}_R) - \boldsymbol{I}_{L'}\|^2$ is $\|T^*(\boldsymbol{P}_R) - \boldsymbol{I}_L\|^2 + \|T'^*_{x'}(\boldsymbol{P}_R) - I_{x'}\|^2$.

Lemma 4.35 (Cuts preserve scalability)**.** *If $p > 0$, $\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ is approximately (resp. exactly) parabolically scalable if and only if $(\boldsymbol{V}, \boldsymbol{p})$ is approximately (resp. exactly) parabolically scalable by $\mathrm{Cut}_{p,x} \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$.*

Proof. Let $S_g : \mathrm{Rep}\, Q \to \mathrm{Rep}\, Q$ denote scaling by g. For the exact scalability part of the proof, we'd like to show that the group orbit $\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \boldsymbol{V}$ contains an element which is balanced when paired with $\boldsymbol{p}$ if and only if the group orbit $\mathrm{Cut}_{p,x} \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \mathrm{Cut}_{p,x} \boldsymbol{V}$ does. By Lemma 4.33 and the injectivity of $\mathrm{Cut}_{p,x}$, this would follow from $\mathrm{Cut}_{p,x}(\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \boldsymbol{V}) = \mathrm{Cut}_{p,x} \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \mathrm{Cut}_{p,x} \boldsymbol{V}$, for which it is enough to show that the below diagram commutes.

$$\begin{array}{ccc} \mathrm{Rep}(Q, \boldsymbol{d}) & \xrightarrow{\mathrm{Cut}_{p,x}} & \mathrm{Rep}(Q', \boldsymbol{d}') \\ \Big\downarrow{\scriptstyle S_g} & & \Big\downarrow{\scriptstyle S_{\mathrm{Cut}_{p,x}\, g}} \\ \mathrm{Rep}(Q, \boldsymbol{d}) & \xrightarrow{\mathrm{Cut}_{p,x}} & \mathrm{Rep}(Q', \boldsymbol{d}') \end{array} \tag{4.6}$$

For this we need only look at each pair x, e with $x \in E$ and $t(e) = x$. Let $g_{x'}$ denote $(\mathrm{Cut}_{p,x}\, g)_{x'}$. We must show that $g_{h(e)} A_{e'} g_{x'} = g_{h(e)}^{\dagger} A_e g_x \iota$, but this follows because $g_{h(e)} A_{e'} g_{x'} = g_{h(e)}^{\dagger} A_e \iota \iota^{\dagger} g_x \iota$, and $\iota \iota^{\dagger} g_x \iota = g_x \iota$ because $\iota \iota^{\dagger}$ is the orthogonal projection to a an element of $F_\circ(x)$, which is by definition fixed by g_x.

In the approximate case, consider the map $f_{\boldsymbol{Q}} : \boldsymbol{W} \mapsto (T_{\boldsymbol{W}}(\boldsymbol{Q}_L), T^*_{\boldsymbol{Q}} \boldsymbol{Q}_R))$. $f_{\boldsymbol{Q}}$ is continuous and by positivity of $\boldsymbol{q}$ the preimage under $f^{-1}B$ is compact for a closed ball of B of radius $1/2$ about the identity in $\boldsymbol{P}(\boldsymbol{V}^*)$. Thus f is closed on $f^{-1}B$, and so it is enough to show $\mathrm{Cut}_{p,x}(\overline{\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \boldsymbol{V}}) = \overline{\mathrm{Cut}_{p,x} \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) \cdot \mathrm{Cut}_{p,x} \boldsymbol{V}}$, but this also follows from Eq. (4.6) because $\mathrm{Cut}_{p,x}$ is injective and linear. □

Proof of Theorem 4.28. By Lemma 4.35, if $(V, \boldsymbol{p})$ is approximately parabolically scalable then $(\mathrm{Red}(\boldsymbol{V}, \boldsymbol{p}), \boldsymbol{1})$ is approximately scalable by the subgroup

$$\mathrm{Red}\, \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}) = \mathrm{Cut}_{i_1,x_1} \circ \cdots \circ \mathrm{Cut}_{i_t,x_t} \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p}).$$

On the other hand, if $(\mathrm{Red}(\boldsymbol{V}, \boldsymbol{p}), \boldsymbol{1})$ is approximately scalable, the sequence of group elements balancing $(\mathrm{Red}(\boldsymbol{V}, \boldsymbol{p}), \boldsymbol{1})$ might not be elements of $\mathrm{Red}\, \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$ in general. If they were, we could pull them back to elements of $\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$ by Eq. (4.6). However, by

Theorem 4.27, the sequence of elements can be taken to be steps of Gurvits' algorithm, which *can* be taken in $\operatorname{Red}\mathrm{GL}(\boldsymbol{V},\boldsymbol{p})$. We prove this for Cut in Lemma 4.36 below, which implies it for Red by induction. Theorem 4.28 then follows from Theorem 4.27, and the fact that Red is an injective linear map. □

Lemma 4.36 (Algorithm 3 and cuts)**.** *If $g^1, g^2, \ldots$ are the sequence of elements in Algorithm 3 on input $(\boldsymbol{V},\boldsymbol{p})$ with $\varepsilon = 0$, then*

$$\operatorname{Cut}_{p,x} g^1, \operatorname{Cut}_{p,x} g^2, \ldots$$

are a valid choice of elements for Algorithm 3 on input $\operatorname{Cut}_{p,x}(\boldsymbol{V},\boldsymbol{p})$ with $\varepsilon = 0$.

Proof. Let $(\boldsymbol{V}',\boldsymbol{p}') = \operatorname{Cut}_{p,x}(\boldsymbol{V},\boldsymbol{p})$, and let $\boldsymbol{W}_i, \boldsymbol{W}'_i$ be the sequence of quiver representations in the algorithm run on $(\boldsymbol{V},\boldsymbol{p}), (\boldsymbol{V}',\boldsymbol{p}')$, respectively. By Eq. (4.6), we have that $\boldsymbol{W}'_i = \operatorname{Cut}_{p,x} \boldsymbol{W}_i$ throughout. Without loss of generality we assume $x \in R$, which implies that for i even we have $\operatorname{Cut}_{p,x} g^i = \operatorname{Cut}_{p,x}(h^i, I_R) = (h^i, I_{R'})$. As in the proof of Lemma 4.33, we have $T^*_{\boldsymbol{W}'_{i-1}}(\boldsymbol{P}'_L) = T^*_{\boldsymbol{W}_{i-1}}(\boldsymbol{P}_L)$, which implies $(h^i, I_{R'})$ is still valid. For odd i, we may assume $i = 1$. Let $T = T_{\boldsymbol{W}_1}$ and $T' = T_{\boldsymbol{W}'_1}$, and suppose $g_{x'} = \operatorname{Cut}_{p,x}(I_L, h^1)_{x'}$. We need only show that $g_{x'} T'(\boldsymbol{P}_L) g_{x'} = I_{V_{x'}}$. As observed in the proof of Lemma 4.33 and by the definition of $g_{x'}$, we have $T'(\boldsymbol{P}_L) = \iota^\dagger g_x \iota \iota^\dagger T(\boldsymbol{P}_L) \iota \iota^\dagger g_x \iota$. As in the proof Lemma 4.35, $\iota\iota^\dagger g_x \iota = g_x \iota$. Hence, $T'(\boldsymbol{P}_L) = \iota^\dagger \iota = I_{V_{x'}}$. □

For $\boldsymbol{p}$ integral, any cut of $(\boldsymbol{V},\boldsymbol{p})$ can be further cut to obtain $\operatorname{Red}(\boldsymbol{V},\boldsymbol{p})$, which proves the next corollary.

Corollary 4.37. *For $\boldsymbol{p}$ integral, $\operatorname{Cut}_{p,x}(\boldsymbol{V},\boldsymbol{p})$ is approximately parabolically scalable if and only if $(\boldsymbol{V},\boldsymbol{p})$ is.*

4.3.2 Randomized reduction to parabolic problem

Here we show that Problem 4.18 for $(\boldsymbol{V},\boldsymbol{p})$ is equivalent to Problem 4.26 for $(V_g,\boldsymbol{p})$ generic. If $(\boldsymbol{V},\boldsymbol{p})$ is scalable, it is unsurprising that there *exists* g such that $(S_g\boldsymbol{V},\boldsymbol{p})$ is parabolic scalable. If the g such that $(S_g\boldsymbol{V},\boldsymbol{p})$ are parabolic scalable is a Zariski–open set, then the claim follows. This true for integral $\boldsymbol{p}$ because Red is a regular map and

the $\boldsymbol{V}$ such that $(\boldsymbol{V}, \mathbf{1})$ is not scalable form an affine variety known as the *null–cone*. The null–cone for quiver representations is rather well–understood.

Theorem 4.38 (Derksen, Makam [DM17])**.** *The null–cone for bipartite quiver representations is an affine variety generated by polynomials of degree at most $N(N-1)$, where $N = \dim \boldsymbol{V}_L$.*

The map $(g, h^\dagger) \mapsto \operatorname{Red} S_{g,h}\boldsymbol{V}$ is actually bilinear in (g, h), which yields the following.

Corollary 4.39. *Let $\boldsymbol{p}$ be integral with $\sum \boldsymbol{p}_R = N$. The set of $(g, h^\dagger) \in \mathrm{GL}(\boldsymbol{V})_L \times \mathrm{GL}(\boldsymbol{V})_R = \mathrm{GL}(\boldsymbol{V})$ such that $(S_{(g,h)}\boldsymbol{V}, \boldsymbol{p})$ is* not *parabolic scalable is an affine variety in $\mathrm{GL}(\boldsymbol{V})$ generated by polynomials of degree at most $N(N-1)$.*

Some slight care is required, but we have essentially proved our randomized reduction. In order to apply the following to rational $\boldsymbol{p}$, simply scale $\boldsymbol{p}$ by a large enough integer. This does not change scalability.

Theorem 4.40. *Suppose $\boldsymbol{p}$ is integral with $\sum \boldsymbol{p} = N$. If $(\boldsymbol{V}, \boldsymbol{p})$ is approximately scalable, then $(S_g\boldsymbol{V}, \boldsymbol{p})$ is approximately parabolically scalable for almost every g. In particular, if the entries of g are chosen at random in $[\alpha N(N-1)]$ then $(S_g\boldsymbol{V}, \boldsymbol{p})$ is scalable with probability at least $1 - 1/\alpha$.*

Proof. If $(\boldsymbol{V}, \boldsymbol{p})$ is approximately scalable, then for every $\varepsilon > 0$ there exists g such that $(S_g\boldsymbol{V}, \boldsymbol{p})$ is ε–balanced. This implies $(\operatorname{Red}(S_g\boldsymbol{V}, \boldsymbol{p}), \mathbf{1})$ is $\sqrt{N}\varepsilon$ balanced. By [Gur04], which we will show reprove later as a consequence of Lemma 4.56, if $\varepsilon\sqrt{N} \leq \frac{1}{\sqrt{N}}$, then $\operatorname{Red} S_g\boldsymbol{V}$ is rank–nondecreasing. Thus, it is scalable by Theorem 4.27. By Theorem 4.28, $S_g\boldsymbol{V}$ is parabolic scalable, so by Corollary 4.39 the g such that $S_g\boldsymbol{V}$ are scalable form a Zariski open set. The last point is an application of the Schwarz-Zippel lemma. □

4.4 Analysis of the algorithm

So far, we have only proved Theorem 4.23 for rational $\boldsymbol{p}$. Furthermore, the running time depends on a common denominator for the $\boldsymbol{p}$, though the morally the running

time should depend on geometrical properties of $\boldsymbol{p}$. In this section we show that this the case. In the process, we will prove an analogue of Theorem 4.27 with generalized notions of the *capacity* $\operatorname{cap}(\boldsymbol{V})$ and rank–nondecreasingness for $\boldsymbol{p} \neq 1$.

4.4.1 Generalization of Gurvits' theorem

For this part, it will be slightly more convenient to use a different notion of balancedness.

Definition 4.41 (Outer scalings)**.** Say $(\boldsymbol{V}, \boldsymbol{p})$ is *ε-outer balanced* if

$$\|T_{\boldsymbol{V}}(\boldsymbol{I}_L) - \boldsymbol{P}_R\| < \varepsilon \text{ and } \|T^*_{\boldsymbol{V}}(\boldsymbol{I}_R) - \boldsymbol{P}_L\| < \varepsilon,$$

Say $(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer scalable if it has an ε–outer balanced scaling for every $\varepsilon > 0$. Say $(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer parabolically scalable if ε–outer balanced scaling by $\operatorname{GL}(\boldsymbol{V}, \boldsymbol{p})$ for every $\varepsilon > 0$. ◊

Remark 4.42 (Outer scalings vs scalings)**.** If $(\boldsymbol{V}, \boldsymbol{p})$ is approximately scalable (resp. parabolically scalable), then $(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer scalable (resp. parabolically scalable). If $(\boldsymbol{V}, \boldsymbol{p})$ is positive, then the converse is also true. This is true because if $(\boldsymbol{V}, \boldsymbol{p})$ is balanced then as $g \to \boldsymbol{P}^{1/2}$, $(S_g \boldsymbol{V}, \boldsymbol{p})$ converges to outer balanced. If $\boldsymbol{p}$ is positive we can set $g = \boldsymbol{P}^{-1/2}$ for the reverse implication.

Theorem 4.43. *Let $(\boldsymbol{V}, \boldsymbol{p})$ be a bipartite quiver datum.*

1. *$(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer parabolically scalable.*
2. *$(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing.*
3. *$\sum \boldsymbol{p}_L = \sum \boldsymbol{p}_R$ and $0 < \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$.*

Moreover, if any of the above hold and $(\boldsymbol{V}, \boldsymbol{p})$ is positive, Algorithm 3 terminates on $(\boldsymbol{V}, \boldsymbol{p})$ for all $\varepsilon > 0$.

It will be immediate from our definition of that the rank–nondecreasingness of $(\boldsymbol{V}, \boldsymbol{p})$ is preserved under parabolic scalings and is equivalent to $\boldsymbol{p}$ lying in a polytope $\mathcal{P}(\boldsymbol{V})$ whose vertices lie in a finite set that depends only on $\boldsymbol{d}$. That the parabolically scalable data takes this general form is due to [Fra02]. The capacity $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$ is a nonnegative

function that is log-concave in $\boldsymbol{p}$ and is supported on P. $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$ acts as a potential function in the analysis of Algorithm 3.

In geometric invariant theory terms, the equivalence of the first three items is a statement about the semistability of certain quivers, so is probably known to practitioners even if it is not explicitly stated in the literature. We use Theorem 4.43 to show our full characterization of scalability:

Theorem 4.44. *Let $(\boldsymbol{V}, \boldsymbol{p})$ be a bipartite quiver datum.*

1. *$(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer scalable.*
2. *$(S_g\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing for g in a Zariski-open subset of $\operatorname{GL}(\boldsymbol{V})$.*
3. *$\operatorname{cap}(S_g\boldsymbol{V}, \boldsymbol{p}) > 0$ for g in a Zariski-open subset of $\operatorname{GL}(\boldsymbol{V})$.*

Moreover, if any of the above hold and $(\boldsymbol{V}, \boldsymbol{p})$ is positive, Algorithm 2 terminates on $(S_g\boldsymbol{V}, \boldsymbol{p})$ for all $\varepsilon > 0$ for g in a Zariski-open subset of $\operatorname{GL}(\boldsymbol{V})$.

To prove this, we need one unsurprising technical result which will be proved in the next section.

Lemma 4.45 (Nearly balanced implies rank–nondecreasing)**.** *There exists $\varepsilon(\boldsymbol{p}) > 0$ such that every ε–outer balanced datum $(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing.*

Proof of Theorem 4.44. We need only prove that 1 implies 2. The argument is due to [Fra02]. The only care required is in avoiding the integrality assumption in Corollary 4.39. To circumvent this, we use that $\boldsymbol{p}$ such that $(S_g\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing forms a polytope $\mathcal{P}(S_g\boldsymbol{V})$ whose vertices fall in a finite subset of $\mathbb{Q}^{\dim \boldsymbol{V}}$ depending only on $\boldsymbol{d}$. If g is approximately scalable, then by Lemma 4.45, there exists g such that $(S_g\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing, i.e. $\mathcal{P}(S_g\boldsymbol{V}) := \mathcal{P}$ contains $\boldsymbol{p}$. Now we consider the finitely many *rational* vertices $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_t$ of $\mathcal{P}$. Because there exists g such that $(V_g, \boldsymbol{p}_1), \ldots, (V_g, \boldsymbol{p}_t)$ are rank–nondecreasing, by Theorem 4.40 they are all in $\mathcal{P}(S_g\boldsymbol{V})$ for generic g. By convexity, $(S_g\boldsymbol{V}, \boldsymbol{p}) \in \mathcal{P}(S_g\boldsymbol{V})$ for generic g. □

The proof above yields more. Clearly $(\boldsymbol{V}, \boldsymbol{p})$ is scalable if it is in the union U of $\mathcal{P}(S_g\boldsymbol{V})$ for all $g \in \operatorname{GL}(\boldsymbol{V})$. However, the vertices of $\mathcal{P}(S_g\boldsymbol{V})$ lie in some fixed finite

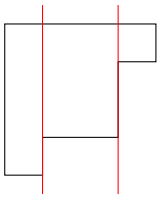


Figure 4.3: Reducing $(4, 3, 3, 1)$ to uniform

set S independent of g, and if there exists g such that a point is in $\mathcal{P}(S_g \boldsymbol{V})$ then that point is in $\mathcal{P}(S_g \boldsymbol{V})$ for almost every g. In particular, there exsts g such that $\mathcal{P}(S_g \boldsymbol{V})$ contains all points of S in U. Thus we have the next corollary.

Corollary 4.46 (The scalable data form a convex polytope)**.** *The set $\mathcal{K}(\boldsymbol{V})$ of $\boldsymbol{p}$ such that $(\boldsymbol{V}, \boldsymbol{p})$ is approximately scalable forms a convex polytope with rational vertices.*

Before we begin the definitions and proofs, we define another useful reduction.

Definition 4.47 (Reduction to uniform quiver datum)**.** Say a quiver datum $(V, \boldsymbol{p})$ is *uniform* if for every x we have $\boldsymbol{p}_x = p(x)\mathbf{1}$ for $p(x) \in \mathbb{R}$. That is, all $\boldsymbol{p}_x$ are proportional to the all–ones vector. If $(V, \boldsymbol{p})$ is a bipartite quiver datum, let

$$\operatorname{Unif}(V, \boldsymbol{p})$$

be the uniform bipartite quiver datum obtained by applying a minimal sequence of cuts to $(\boldsymbol{V}, \boldsymbol{p})$. That is, by applying $\operatorname{Cut}_{p,x}$ for each distinct value p appearing in $\boldsymbol{p}_x$ for $x \in \Omega$. We also define $\operatorname{Unif} g$ for $g \in \operatorname{GL}(\boldsymbol{V})$ as $\operatorname{Red} g, \operatorname{Cut}_{p,x} g$.

Remark 4.48 (Algorithm on $\operatorname{Unif}(\boldsymbol{V}, \boldsymbol{p})$)**.** By Lemma 4.36, if elements g_i are valid choices of group elements in Algorithm 3 on $(\boldsymbol{V}, \boldsymbol{p})$ then $\operatorname{Unif} g_i$ are valid choices for Algorithm 3 on $\operatorname{Unif}(\boldsymbol{V}, \boldsymbol{p})$. ◊

4.4.2 Rank–nondecreasingness

A *Hall blocker* in a bipartite graph on $[n] \amalg [n]$ is an independent set of cardinality greater than n; the existence of a perfect matching is equlivalent to the nonexistence of a hall blocker. A collection of subspaces $W_x \subset V_x$ will play the role of a Hall blocker in

obstructing scalings. We assign a certain "mass" $\boldsymbol{p}(W_x)$ to each subspace in $\boldsymbol{W}$, and the generalized notion of nonadjacency between x and y is that $A_e W_{t(e)} \subset W_{h(e)}^{\perp}$ for every edge e from y to x. We call such a collection $\boldsymbol{W}$ of subspaces a $\boldsymbol{V}$*–independent set*; if the total mass is too large, $\boldsymbol{W}$ obstructs scalability.

Definition 4.49 (Rank–nondecreasingness)**.** If $\boldsymbol{p} \in \mathbb{R}^n$ is a nonincreasing sequence, and W a subspace of $\mathbb{C}^n$, define

$$\boldsymbol{p}(W) = \sum_{i=1}^{n} (p_i - p_{i+1}) \dim(F_i \cap W).$$

If $(\boldsymbol{V}, \boldsymbol{p})$ is a quiver datum and $\boldsymbol{W} : x \mapsto W_x \subset V_x$ is an assignment of subspaces to vertices, define

$$\boldsymbol{p}(\boldsymbol{W}) = \sum_{x \in \Omega} \boldsymbol{p}_x(W_x)$$

Say a bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{P})$ of a bipartite quiver Q is *rank–nondecreasing*[4] if

$$\sum \boldsymbol{p}_L = \sum \boldsymbol{p}_R \tag{4.7}$$

and for every $\boldsymbol{V}$–independent pair $\boldsymbol{W}$,

$$\boldsymbol{p}(\boldsymbol{W}) \leq \sum \boldsymbol{p}_L. \tag{4.8}$$

Remark 4.50 (Facts about rank–nondecreasingness)**.**

- Most importantly, rank–nondecreasingness is preserved under the action of $\mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$. This follows because for $g \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$, g_x fixes F_i if $\boldsymbol{p}_x(i) \neq \boldsymbol{p}_x(i+1)$, and $g^{-1}\boldsymbol{W} : x \mapsto g_x^{-1} W_x$ is still a $\boldsymbol{V}$–independent set.

- Rank–nondecreasingness of $(\boldsymbol{V}, \mathbf{1})$ is equivalent to the rank–nondecreasingness of $\boldsymbol{V}$ in Theorem 4.27.

- Rank–nondecreasingness is manifestly symmetrical; $(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing if and only if $(\boldsymbol{V}, \boldsymbol{p})^*$ is rank–nondecreasing.

[4] In the language of quiver representations, rank–nondecreasingness is the same as semistability of some augmented quiver for $\boldsymbol{p}$ integral.

- We can rewrite $\sum_{i=1}^{n}(p(i)-p(i+1))\dim(W\cap F_i)$ as

$$\sum_{i=1}^{n} p_i(\dim(W\cap F_i)-\dim(W\cap F_{i-1}))$$

for $F_0 := \{0\}$; this shows that rank–nondecreasingness is characterized by a system of inequalities with coefficients in $\{-1,0,1\}$.

Example 4.51 (Perfect matchings)**.** If Q is a complete bipartite graph (directed from left to right) and $\boldsymbol{d} = \mathbf{1}$ (as in Example 4.17), then $(\boldsymbol{V},\mathbf{1})$ is rank–nondecreasing if and only if the matrix $[A_{xy}]$ supports a perfect matching. In this case, rank–nondecreasingness amounts to Hall's condition.

Example 4.52 (Prescribed row and column sums)**.** If $(\boldsymbol{V},(\boldsymbol{r},\boldsymbol{c}))$ where $\boldsymbol{V}$ is as above, rank–nondecreasingness is equivalent to the condition of Rothblum and Schneider [RS89] for $[A_{xy}]$ to have a scaling by diagonals to with row and column sums $\boldsymbol{r},\boldsymbol{c}$, respectively. That is, $\sum \boldsymbol{c} = \sum \boldsymbol{r}$ and every zero submatrix $I\times J$ satisfies $\sum_{x\in I} r(x)+\sum_{y\in J} c(y)\le \sum_{x\in L}^{n} c(y)$.

Example 4.53 (Block matrices with prescribed row and column "sums")**.** If $\boldsymbol{V}$ is on the same quiver as above but with $\boldsymbol{d}\neq\mathbf{1}$, then the condition of rank–nondecreasingness of $(\boldsymbol{V},(\boldsymbol{r};\boldsymbol{c}))$ generalizes that of Dvir et. al. [DGOS16] for uniform data, i.e. when $\boldsymbol{r}_x = r(x)\mathbf{1}, \boldsymbol{c}_x = c(x)\mathbf{1}$ are proportional to all–ones vectors, which states that $\sum \boldsymbol{r} = \sum \boldsymbol{c}$ and

$$\sum_{x\in L} r(x)\dim W_x + \sum_{x\in R} c(x)\dim W_x \le \sum r$$

for any W with $A_{xy}W_y \subset W_x^{\perp}$ for all $y\in L, x\in R$. ◊

The next lemma is proved in Appendix A.1.

Lemma 4.54 (Cuts preserve rank–nondecreasingness)**.** $\mathrm{Cut}_{x,p}(\boldsymbol{V},\boldsymbol{p})$ *is* rank-nondecreasing *if and only if* $(\boldsymbol{V},\boldsymbol{p})$ *is. In particular, cuts preserve* $\boldsymbol{p}_L(\boldsymbol{V}),\boldsymbol{p}_R(\boldsymbol{V})$ *and the maximum value of* $\boldsymbol{p}(\boldsymbol{W})$ *over all* $\boldsymbol{V}$*–independent sets* $\boldsymbol{W}$*.*

Corollary 4.55 (The reduction preserves rank–nondecreasingness)**.** *For* $(\boldsymbol{V},\boldsymbol{p})$ *integral,* $(\mathrm{Red}(\boldsymbol{V},\boldsymbol{p}),\mathbf{1})$ *is rank–nondecreasing if and only if* $(\boldsymbol{V},\boldsymbol{p})$ *is.*

We next show that balanced data are rank–nondecreasing. Applying the following proposition to $\left(S_{\boldsymbol{p}^{1/2}}\boldsymbol{V}, \boldsymbol{p}\right)$ implies Lemma 4.45. The proof is in Appendix A.1.

Lemma 4.56 (Approximate balancedness implies rank–nondecreasingness)**.** *Suppose* $(\boldsymbol{V}, \boldsymbol{p})$ *is an* ε*–outer balanced bipartite quiver datum. Then* $|\sum \boldsymbol{p}_L - \sum \boldsymbol{p}_R| \leq 2\varepsilon\sqrt{\dim \boldsymbol{V}}$ *and*

$$\boldsymbol{p}(\boldsymbol{W}) \leq \sum \boldsymbol{p}_L + 2\varepsilon \dim \boldsymbol{V} \tag{4.9}$$

for all $\boldsymbol{V}$*–independent sets* $\boldsymbol{W}$*.*

4.4.3 Capacity

Here we describe the analogue of the *capacity*, the function $\operatorname{cap}(V)$ from Theorem 4.27. Our modified capacity, $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$, is *log-concave* in $\boldsymbol{p}$, and we will show that it is supported on the set of $\boldsymbol{p}$ such that $(\boldsymbol{V}, \boldsymbol{p})$ is rank–increasing.

Definition 4.57 (Relative determinant)**.** Let $\iota_j : F_j \to \mathbb{C}^k$ be the inclusion map, or as an $n \times j$ matrix,

$$\iota_j = \begin{bmatrix} I_j \\ 0_{n-j} \end{bmatrix}. \tag{4.10}$$

If $\boldsymbol{p}$ is a nonincreasing sequence in $\mathbb{R}^n_{\geq 0}$, define $\det(X, \boldsymbol{p}) : \operatorname{Mat}_n(\mathbb{C}) \to \mathbb{C}$ by

$$\det(X, \boldsymbol{p}) = \prod_{j=1}^{n} |\det \iota_j^\dagger X \iota_j|^{p_j - p_{j+1}}, \tag{4.11}$$

where $0^0 := 1$. For an assignment $(\boldsymbol{X}, \boldsymbol{p}) : x \to X_x \in \operatorname{Mat}_{d(x)} \mathbb{C}, \boldsymbol{p}_x \in \mathbb{R}^n_{\geq 0}$ nonincreasing and x in a finite set S, define

$$\det(\boldsymbol{X}, \boldsymbol{p}) = \prod_{x \in S} \det(X_x, \boldsymbol{p}_x).$$

We use a multiplicativity property of this determinant, which we prove in Appendix A.2.

Lemma 4.58 (Multiplicativity of determinant)**.** *If* $g \in \operatorname{GL}(F_\circ^{\boldsymbol{p}})$*, then for all* $X \in \operatorname{Mat}_n(\mathbb{C})$*,*

$$\det(g^\dagger X g, \boldsymbol{p}) = \det(g^\dagger g, \boldsymbol{p}) \det(X, \boldsymbol{p}). \tag{4.12}$$

We are now ready to define the capacity:

Definition 4.59 (Capacity)**.** If $(\boldsymbol{V}, \boldsymbol{p})$ is a bipartite quiver datum, define

$$\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) = e^{-H(\boldsymbol{p}_L)} \inf_{\boldsymbol{X} \succ 0} \det(T_{\boldsymbol{V}}(\boldsymbol{X}), \boldsymbol{p}_R) \det(\boldsymbol{X}^{-1}, \boldsymbol{p}_L) \tag{4.13}$$

where $H(\boldsymbol{q}) = -\sum q_i \log q_i$ is the Shannon entropy. ◊

Firstly, note that the capacity is log–concave.

Lemma 4.60 (log–concavity of capacity)**.** $e^{H(\boldsymbol{p}_L)} \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$ *is log–concave in* $\boldsymbol{p}$*.*

Proof. We must show that $\inf_{X \succ 0} \log(\det(T_{\boldsymbol{V}}(\boldsymbol{X}), \boldsymbol{p}_R) \det(\boldsymbol{X}^{-1}, \boldsymbol{p}_L))$ is concave in $\boldsymbol{p}$. However, the quantity being infimized is linear in $\Delta\boldsymbol{p} : (x, i) \mapsto \boldsymbol{p}_x(i) = \boldsymbol{p}_x(i+1) \in \mathbb{R}_{\geq 0}^{\dim \boldsymbol{V}}$, which is linear in $\boldsymbol{p}$. In particular, it takes the form $g : \Delta\boldsymbol{p} \mapsto \mathbb{R} \cup \{-\infty\}$ where

$$g(\Delta\boldsymbol{p}) = \inf_{y \in S} \Delta\boldsymbol{p} \cdot \boldsymbol{f}(y)$$

for some set S and $\boldsymbol{f}$ a vector valued function on S with components in $\mathbb{R} \cup \{-\infty\}$. The convention $0 \log 0 = 0$, which is consistent with our convention in defining $\det(X, \boldsymbol{a})$, and the nonnegativity of $\Delta\boldsymbol{p}$, ensures that for $\lambda \in (0, 1)$ we have

$$\begin{aligned} \inf_{y \in S} (\lambda \Delta\boldsymbol{p}_1 + (1-\lambda)\Delta\boldsymbol{p}_2) \cdot \boldsymbol{f}(y) &= \inf_{y \in S} \lambda(\Delta\boldsymbol{p}_1 \cdot \boldsymbol{f}(y)) + (1-\lambda)(\Delta\boldsymbol{p}_2 \cdot \boldsymbol{f}(y)) \\ &\geq \lambda \inf_{y \in S} \Delta\boldsymbol{p}_1 \cdot \boldsymbol{f}(y) + (1-\lambda) \inf_{y \in S} \Delta\boldsymbol{p}_2 \cdot \boldsymbol{f}(y). \end{aligned}$$

□

This determinant behaves well under cuts. It is not hard to see that for $(\boldsymbol{V}', \boldsymbol{p}') = \operatorname{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})$ we have

$$\det(X_x, \boldsymbol{p}_x) = \det(X_x, \boldsymbol{p}_x) \det(\iota^\dagger X_x \iota, \boldsymbol{p}'_{x'}). \tag{4.14}$$

We prove the next lemma in Appendix A.2.

Lemma 4.61 (Cuts preserve capacity)**.** *If* $(\boldsymbol{V}, \boldsymbol{p})$ *is a bipartite quiver datum, then*

$$\operatorname{cap}(\operatorname{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})) = \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}). \tag{4.15}$$

Corollary 4.62. *If* $\boldsymbol{p}$ *is integral, then* $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) = \operatorname{cap}(\operatorname{Red}(\boldsymbol{V}, \boldsymbol{p}))$.

4.4.4 Analysis of the algorithm

Here we show that if $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$, Algorithm 3 converges. Let $\boldsymbol{W}_i$ be the quiver representation in the i^{th} iteration. We show that if $(\boldsymbol{W}_i, \boldsymbol{p})$ is ε-far from balanced, then $\mathrm{cap}(\boldsymbol{W}_{i+1}, \boldsymbol{p}) > \mathrm{cap}(\boldsymbol{W}_i, \boldsymbol{p}) + f(\varepsilon)$, and throughout $\mathrm{cap}(\boldsymbol{W}_i, \boldsymbol{p}) \leq 1$. It is easier to analyze the algorithm on $\mathrm{Unif}(\boldsymbol{V}, \boldsymbol{p})$, which is equivalent by Remark 4.48. We first state the lemmas and prove that the algorithm converges, and then proceed to prove the lemmas. We will need a notion of distance from doubly stochastic.

Definition 4.63 (Distance to doubly stochastic)**.** If $(\boldsymbol{V}, \boldsymbol{p})$ is a uniform bipartite quiver datum, define

$$\mathrm{ds}(\boldsymbol{V}, \boldsymbol{p})^2 = \sum_{x \in L} p(x) \|T_x(\boldsymbol{P}_R) - I_{d(y)}\|^2 + \sum_{y \in R} p(y) \|T_y^*(\boldsymbol{P}_L) - I_{d(y)}\|^2.$$

In particular, if $p_{\min}, p_{\max}$ are the the maximum and minimum entries of $\boldsymbol{p}$, then

$$\begin{aligned} & p_{\min}(\|T(\boldsymbol{P}_L) - \boldsymbol{I}_R\|^2 + \|T^*(\boldsymbol{P}_R) - \boldsymbol{I}_L\|^2) \\ \leq \mathrm{ds}(\boldsymbol{V}, \boldsymbol{p})^2 \leq & \; p_{\max}(\|T(\boldsymbol{P}_L) - \boldsymbol{I}_R\|^2 + \|T^*(\boldsymbol{P}_R) - \boldsymbol{I}_L\|^2) \qquad (4.16) \\ & \qquad (4.17) \end{aligned}$$

so if $(\boldsymbol{V}, \boldsymbol{p})$ is positive then it is ε–balanced if $\mathrm{ds}(\boldsymbol{V}, \boldsymbol{p})^2 \leq \varepsilon^2 p_{\min}$ and only if $\mathrm{ds}(\boldsymbol{V}, \boldsymbol{p})^2 \leq 2\varepsilon^2 p_{\max}$. ♢

Proving that the algorithm makes progress each step is the only ingredient that requires care. We delay the proof until after the proof of Theorem 4.68.

Lemma 4.64 (Substantial progress)**.** *Suppose $(\boldsymbol{V}, \boldsymbol{p})$ is a uniform datum with $\boldsymbol{p}_L(\boldsymbol{V}) = \boldsymbol{p}_R(\boldsymbol{V})$. Suppose $(\boldsymbol{V}, \boldsymbol{p})$ is left–balanced, $\mathrm{ds}(\boldsymbol{V}, \boldsymbol{p})^2 \geq \varepsilon$, and $g = (h, \boldsymbol{I}_R) \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$ satisfies $h^\dagger T^*_{\boldsymbol{V}}(\boldsymbol{P}_R) h = \boldsymbol{I}_L$. Then*

$$\mathrm{cap}(S_g \boldsymbol{W}, \boldsymbol{p}) \geq e^{.06 \min\{\epsilon, p_{\min}\}} \mathrm{cap}(\boldsymbol{W}, \boldsymbol{p}).$$

The same holds if $(\boldsymbol{V}, \boldsymbol{p})$ right–balanced and $g = (\boldsymbol{I}_L, h)$ where $h^\dagger T_{\boldsymbol{V}}(\boldsymbol{P}_L) h = \boldsymbol{I}_R$.

Lemma 4.65 (Evolution of capacity)**.** *If $g \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})$, then*

$$\mathrm{cap}(S_g \boldsymbol{V}, \boldsymbol{p}) = \det(g^\dagger g, \boldsymbol{p}) \, \mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}).$$

Proof. Let $g = (h_L, h_R)$ for $h_L \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})_L$ and $h_R \in \mathrm{GL}(\boldsymbol{V}, \boldsymbol{p})_R$. By the change of variables $\boldsymbol{Y} = h_L \boldsymbol{X} h_L^\dagger$, we have

$$\begin{aligned}\inf_{\boldsymbol{X} \succ 0} \det(T_{S_g \boldsymbol{V}}(\boldsymbol{X}), \boldsymbol{p}_R) \det(\boldsymbol{X}^{-1}, \boldsymbol{p}_L) &= \inf_{\boldsymbol{X} \succ 0} \det(h_R^\dagger T_{\boldsymbol{V}}(h_L \boldsymbol{X} h_L^\dagger) h_R, \boldsymbol{p}_R) \det(\boldsymbol{X}^{-1}, \boldsymbol{p}_L) \\ &= \inf_{\boldsymbol{Y} \succ 0} \det(h_R^\dagger T_{\boldsymbol{V}}(\boldsymbol{Y}) h_R, \boldsymbol{p}_R) \det(h_L^\dagger \boldsymbol{Y}^{-1} h_L, \boldsymbol{p}_L)\end{aligned}$$

Applying 4.12 to both factors completes the proof. □

The next lemma is proved in Appendix A.2.

Lemma 4.66 (Nonsingular operators map to nonsingular operators)**.** *Suppose* $(\boldsymbol{V}, \boldsymbol{p})$ *is a uniform positive bipartite quiver datum with* $\sum \boldsymbol{p}_L = \sum \boldsymbol{p}_R$ *and* $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$*. Then both* $T_{\boldsymbol{V}}$ *and* $T_{\boldsymbol{V}}^*$ *map positive definite matrices to positive definite matrices.*

Lemma 4.67 (capacity upper bound)**.** *Suppose* $(\boldsymbol{V}, \boldsymbol{p})$ *is left– or right– balanced and* $\boldsymbol{p}_L(\boldsymbol{V}) = \boldsymbol{p}_R(\boldsymbol{V}) = 1$*. Then* $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) \leq 1$*.*

Proof. By Lemma 4.61, it is enough to prove the claim assuming $(\boldsymbol{V}, \boldsymbol{p})$ is uniform. Let $T = T_{\boldsymbol{V}}$. By taking $\boldsymbol{X} = \boldsymbol{P}_L$, we have

$$\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) \leq e^{-H(\boldsymbol{p}_L)} \det(T(\boldsymbol{P}_L), \boldsymbol{p}_R) \det(\boldsymbol{P}_L^{-1}, \boldsymbol{p}_L),$$

which is equal to 1 if $(\boldsymbol{V}, \boldsymbol{p})$ is left–balanced. If $(\boldsymbol{V}, \boldsymbol{p})$ is instead right–balanced, then $1 = \operatorname{tr} \boldsymbol{P}_L = \operatorname{tr} \boldsymbol{P}_L T^*(\boldsymbol{P}_R) = \operatorname{tr} T(\boldsymbol{P}_L) \boldsymbol{P}_R$. This shows that $\sum_{y \in R} \sum_{i=1}^{d(y)} p(y) \lambda_y(i) = 1$ if $\boldsymbol{\lambda}_y$ denotes the spectrum of $T_x(\boldsymbol{P}_L)$. Then

$$\begin{aligned}\log \mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) &\leq -H(\boldsymbol{p}_L) + \sum_{x \in R} p(y) \log\det T_y(\boldsymbol{P}_{\boldsymbol{L}}) - \sum_{x \in L} p(x) \log\det P_x \\ &= \sum_{y \in R} p(y) \log\det \sum_{i=1}^{d(y)} \log \lambda_y(i).\end{aligned}$$

The last line is at most one by Jensen's inequality. □

Theorem 4.68 (Running time of Algorithm 3)**.** *Suppose* $(\boldsymbol{V}, \boldsymbol{p})$ *is a positive bipartite quiver datum with* $\boldsymbol{p}_L(\boldsymbol{V}) = \boldsymbol{p}_R(\boldsymbol{V})$*. If* $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$*, then Algorithm 3 terminates in*

$$O\left(\frac{-\log \mathrm{cap}(\boldsymbol{V}_1, \boldsymbol{p})}{p_{\min} \varepsilon^2}\right)$$

steps, where $\boldsymbol{V}_1$ *is the datum obtained after one normalization step.*

Proof. First, let $(\boldsymbol{V}', \boldsymbol{p}') = \mathrm{Unif}(\boldsymbol{V}, \boldsymbol{p})$. By Remark 4.48, the steps of the algorithm on $(\boldsymbol{V}, \boldsymbol{p})$ determine valid steps on $(\boldsymbol{V}', \boldsymbol{p}')$. By Lemma 4.33, $(\boldsymbol{V}', \boldsymbol{p}')$ is ε–balanced, then $(\boldsymbol{V}, \boldsymbol{p})$ is ε–balanced. Unif preserves $p_{\min}$. By Eq. (4.16), it is enough to run the algorithm on $(\boldsymbol{V}', \boldsymbol{p}')$ until $(\boldsymbol{W}, \boldsymbol{p}')$ satisfies $\mathrm{ds}(\boldsymbol{W}, \boldsymbol{p}') \leq \varepsilon'^2 := p_{\min}\varepsilon^2$.

By Lemma 4.61, $\mathrm{cap}(\boldsymbol{V}', \boldsymbol{p}') = \mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$, so Lemma 4.65 implies $\mathrm{cap}(\boldsymbol{W}, \boldsymbol{p}') > 0$ throughout the algorithm. The existence of the Cholesky decomposition and Lemma 4.66 imply each step is possible; e.g. in the (a) steps g will be a Cholesky factor of $T_{\boldsymbol{W}}(\boldsymbol{P}'_L)^{-1}$. Lemma 4.64 implies that in every step of the algorithm after the first,

$$\mathrm{cap}(S_g \boldsymbol{W}, \boldsymbol{p}') \geq e^{.06 \min\{p_{\min}, \epsilon'^2\}} \mathrm{cap}(\boldsymbol{W}, \boldsymbol{p}').$$

provided $\mathrm{ds}(\boldsymbol{W}, \boldsymbol{p}') \geq \epsilon'$. By Lemma 4.67, $\mathrm{cap}(\boldsymbol{W}, \boldsymbol{p}') \leq 1$ throughout. Taking logarithms and using $\min\{p_n, \varepsilon'^2\} = p_{\min}\varepsilon^2$ completes the proof. □

Proof of Lemma 4.64. Let $T = T_{\boldsymbol{W}}$. First note that $g = (h, \boldsymbol{I}_R)$ such that $h^\dagger T^*(\boldsymbol{P}_R)h = \boldsymbol{I}_L$. Because the datum is uniform, $\det(h^\dagger h, \boldsymbol{p}_L) = \det(T^*(\boldsymbol{P}_R)^{-1}, \boldsymbol{p}_L)$. By Lemma 4.65 it is enough to show $\det(T^*(\boldsymbol{P}_R)^{-1}, \boldsymbol{p}_L) \geq e^{.06 \min\{p_{\min}, \varepsilon\}}$. The analogous inequality for the (a) steps follows by symmetry.

Because the datum is uniform, we may equivalently show

$$\sum_{x \in L} p(x) \log\det T^*_x(\boldsymbol{P}_R) \leq -.06 \min\{p_{\min}, \varepsilon\}.$$

As in [GGOW16b, Gur04], the proof is a stability version of Jensen's inequality. The statement is invariant under rescaling of $\boldsymbol{p}$, so we assume $\sum p(x)d(x) = 1$. If $\boldsymbol{\lambda}_x$ is the spectrum of $T^*_x(\boldsymbol{P}_R)$, let X be the random variable obtained by choosing $\lambda_x(i)$ with probability $p(x)$. The left–hand side of the above inequality is $\mathbb{E}[\log X]$. If any $\lambda_x(i)$ are zero, $\mathbb{E}[\log X] = -\infty$ because $p(x) > 0$ for all x, so we assume this is not the case. Because $\mathbb{E}[X] = \mathrm{tr}\, T^*(\boldsymbol{P}_R) = \mathrm{tr}\, \boldsymbol{P}_R = 1$, Jensen's inequality shows $\mathbb{E}[\log X] \leq 0$, however, we need $\mathbb{E}[\log X] \leq -.3 \min\{p_{\min}\}$. We will use that this random variable has nonzero variance. By assumption, we have

$$\sum_{x \in L} p(x) \sum_{i=1}^{d(x)} (\lambda_x(i) - 1)^2 = \mathbb{V}[X] \geq \varepsilon.$$

Because the logarithm punishes outliers only very weakly, no relationship between the variance of X and expectation of $\log X$ is true without some assumption on X. To handle this, we apply a simple technical fact.

Fact 4.69. *If X is a discrete random variable with $\mathbb{E}[X] = 1$ that takes every value in its support with probability at least $p_{\min}$, then*

$$\mathbb{E}[\log X] \leq -.3 \min\{\mathbb{V}[X], p_{\min}\}.$$

We prove this fact in Appendix A. □

4.4.5 Proof of the generalization of Gurvits' theorem

Before the proof, we need a technical lemma to handle when $(\boldsymbol{V}, \boldsymbol{p})$ is not positive. To deal with this, we just work with the datum obtained by applying $\mathrm{Cut}_{0,x}$ for every vertex x in sequence.

Lemma 4.70. *Let $(\boldsymbol{V}', \boldsymbol{p}') = \mathrm{Cut}_{0,x}(\boldsymbol{V}, \boldsymbol{p})$. If $(\boldsymbol{V}', \boldsymbol{p}')$ is approximately outer parabolically scalable, then $(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer parabolically scalable.*

Proof. Let $T = T_{\boldsymbol{V}}, T' = T'_{\boldsymbol{V}'}$. Suppose $x \in R$. Suppose $(S_g \boldsymbol{V}', \boldsymbol{p}')$ is an ε–outer balanced scaling of $(\boldsymbol{V}', \boldsymbol{p}')$. Let h_x approach $\iota g_{x'} \iota^\dagger$ and $h_y = g_y$ for $y \neq x$. Then $h_x T_x(\boldsymbol{I}_L) h_x^\dagger$ approaches

$$\begin{aligned} &\iota g_{x'} \iota^\dagger T_x(\boldsymbol{I}_L) \iota^\dagger g'_{x'} \iota^\dagger \\ &= \iota g_{x'} T'_{x'}(\boldsymbol{I}_L) g_{x'} \iota^\dagger, \end{aligned}$$

which is at most ε from P_x. The proof is similar for $x \in L$. □

Proof of Theorem 4.43. Define $\mathcal{C}(\boldsymbol{V}), \mathcal{S}(\boldsymbol{V}), \mathcal{P}(\boldsymbol{V})$ to be the sets of $\boldsymbol{p}$ with $\sum p = 1$ such that, respectively, $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$, $(\boldsymbol{V}, \boldsymbol{p})$ is approximately outer parabolically scalable, and $(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing. So far, we have shown that

$$\mathcal{C}(\boldsymbol{V}) \subset \mathcal{S}(\boldsymbol{V}) \subset \mathcal{P}(\boldsymbol{V}).$$

To obtain the leftmost inclusion, suppose $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$. Let $(\boldsymbol{V}', \boldsymbol{p}')$ be the datum obtained by applying $\mathrm{Cut}_{0,x}$ to all vertices; by Lemma 4.61, we still have $\mathrm{cap}(\boldsymbol{V}', \boldsymbol{p}') > 0$

and $(\boldsymbol{V}', \boldsymbol{p}')$ is positive. Now Theorem 4.68 and Remark 4.42 show $\boldsymbol{p}' \in \mathcal{S}(\boldsymbol{V}')$, and Lemma 4.70 shows that $\boldsymbol{p} \in \mathcal{S}(\boldsymbol{V})$. The rightmost inclusion is Lemma 4.45.

The reduction characterizes exactly the intersection of each of these sets with $\mathbb{Q}^{\boldsymbol{V}}$ (we may always scale so that rational $\boldsymbol{p}$ become integral without changing scalability, rank-nondecreasingness, or nonvanishing of capacity). By Theorem 4.27, Corollary 4.55, and Corollary 4.62,

$$\mathcal{C}(\boldsymbol{V}) \cap \mathbb{Q}^{\dim \boldsymbol{V}} = \mathcal{P}(\boldsymbol{V}) \cap \mathbb{Q}^{\dim \boldsymbol{V}}.$$

Since $\mathcal{P}(\boldsymbol{V})$ is a convex polytope with rational vertices, $\mathcal{P}(\boldsymbol{V}) \cap \mathbb{Q}^{\dim V}$ contains the vertices of $\mathcal{P}(\boldsymbol{V})$! However, $\mathcal{C}(\boldsymbol{V})$ is convex by Lemma 4.60, so it contains $\mathcal{P}(\boldsymbol{V})$. The three sets must be the same; this completes the proof. □

4.5 Running time

In order to use Theorem 4.68 to bound the running time of Algorithm 3, we must compute a lower bound $\operatorname{cap}(\boldsymbol{V})$. For this we will need to use a nontrivial lower bound on $\operatorname{cap}(\boldsymbol{V})$ from [GGOW16b].

Theorem 4.71 (Garg et. al. [GGOW16b])**.** *Suppose $\boldsymbol{V}$ is a rank–nondecreasing bipartite quiver representation with r edges with Gaussian integer entries, and let $N = \dim \boldsymbol{V}_R$. Then*

$$\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) \geq e^{-N \log(rN^4)}.$$

This is implicit in the proof of Theorem 2.21 in [GGOW16b], which gives the bound $\exp(-O(N^2 \log(RN^4))$, the degree bound used there was since improved to the bound in Theorem 4.38.

To obtain a good enough bound, one can simply apply the above bound to the $\operatorname{Red}(\boldsymbol{V}, \boldsymbol{p})$. Slightly more surprising is that our lower bound for a capacity which does not depend on $\boldsymbol{p}$ at all provided $\sum \boldsymbol{p}_L = 1$! This follows from the log convexity of $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$. Recall from Remark 4.50 that $\mathcal{L}(\boldsymbol{V})$ is determined by inequalities and equalities with coefficients in $\{-1, 0, +1\}$. The next observation follows by standard estimates.

Observation 4.72. *The there exists an integer $M \leq (\dim \boldsymbol{V})^{(\dim \boldsymbol{V})/2}$ such that entries of the vertices of $\mathcal{P}(\boldsymbol{V})$ are of the form k/M for $k \in \mathbb{Z}$.*

Theorem 4.73. *Suppose $\boldsymbol{V}$ has at most r edges whose entries are Gaussian integer entries. If $(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing, $\sum \boldsymbol{p}_L = 1$, and $\dim \boldsymbol{V} = n$, then*

$$\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) \geq \exp(-7n \log n - \log r).$$

Proof. Firstly, we compute our bound for rational $\boldsymbol{p}$ of bit–complexity B. Choose an integer $M \leq 2^B$ such that $\mathbf{q} = M\boldsymbol{p}$ has integer entries. An easy calculation shows $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) = \frac{1}{M} \operatorname{cap}(\boldsymbol{V}, M\boldsymbol{p})^{1/M}$, and by Corollary 4.62 $\operatorname{cap}(\boldsymbol{V}, M\boldsymbol{p}) = \operatorname{cap}(\boldsymbol{V}')$ where $\boldsymbol{V}' = \operatorname{Red}(\boldsymbol{W}', M\boldsymbol{p})$. Further, $\boldsymbol{V}'$ has $\dim \boldsymbol{V}'_L = M$ and has at most $M^2 p_{\max}^2 r \leq M^2 r$ edges whose matrices have Gaussian integer entries. By Corollary 4.55, $\boldsymbol{V}'$ is rank–nondecreasing. By Theorem 4.71, $\operatorname{cap}(\boldsymbol{V}') > e^{M \log(M^6 r)}$. Combining these observations, $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) \geq \frac{1}{M} e^{\log(M^6 r)} \geq e^{-7 \log M - \log r}$.

We now remove the dependence on M. By the log-convexity of $e^{H(\boldsymbol{p})} \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$ in $\boldsymbol{p}$, for $\boldsymbol{V}$ fixed $e^{H(\boldsymbol{p})} \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$ takes a minimum at some vertex of the polytope $\mathcal{P}(\boldsymbol{V})$ of $\boldsymbol{q}$ such that $(\boldsymbol{V}, \boldsymbol{q})$ is rank–nondecreasing. By Observation 4.72, we may assume $M \leq n^{n/2}$. □

We also need to check that the first step of the algorithm does not change the capacity very much, because our running time bound in Theorem 4.68 depends on the capacity after one step.

Lemma 4.74. *Suppose $\boldsymbol{V}$ has at most r edges whose entries are Gaussian integer entries of absolute value at most M. Further suppose that $(\boldsymbol{V}, \boldsymbol{p})$ is rank–nondecreasing, $\sum \boldsymbol{p}_L = 1$, $\dim \boldsymbol{V} = n$, and let $\boldsymbol{V}_1$ be the representation obtained from the first step of Algorithm 3 applied to $(\boldsymbol{V}, \boldsymbol{p})$. Then*

$$\operatorname{cap}(\boldsymbol{V}_1, P, Q) \geq e^{-9n \log n - 2 \log M - 2 \log r}.$$

Proof. By Lemma 4.65, $\operatorname{cap}(\boldsymbol{V}_1, \boldsymbol{p}) = \operatorname{cap}(S_g \boldsymbol{V}, \boldsymbol{p}) = \det(g^\dagger g, \boldsymbol{p}) \operatorname{cap}(\boldsymbol{V}, \boldsymbol{p})$. Let $T = T_{\boldsymbol{V}}$. Then $g = (\boldsymbol{I}_L, h)$ where $hT(\boldsymbol{P}_L)h^\dagger = \boldsymbol{I}_R$, so Eq. (4.12) shows $\det(h^\dagger h, \boldsymbol{p}_R) = \det(T(\boldsymbol{P}_L)^{-1}, \boldsymbol{p}_R)$. One can immediately check that $T(\boldsymbol{P}_L) \preceq n^2 r M^2 \boldsymbol{I}_L$, so

$$\det(T(\boldsymbol{P}_L)^{-1}, \boldsymbol{p}_R) \geq (n^2 r M^2)^{-1}$$

Multiplying the above by the bound from Theorem 4.73 implies Lemma 4.74. □

Combining with Theorem 4.68, we obtain the following.

Corollary 4.75 (Running time of Algorithm 3)**.** *Suppose* $\boldsymbol{V}$ *has at most* r *edges whose entries are Gaussian integer entries of absolute value at most* M*. Further suppose that* $(\boldsymbol{V}, \boldsymbol{p})$ *is rank–nondecreasing,* $\sum \boldsymbol{p}_L = 1$*, and define* $\dim \boldsymbol{V} = n$*. Then Algorithm 3 terminates in*

$$O\left(\frac{n \log n + \log M + \log r}{p_{\min} \varepsilon^2}\right)$$

steps on input $(\boldsymbol{V}, \boldsymbol{p}), \varepsilon$*.*

4.5.1 Running time of the general linear scaling algorithm

In this section we bound the time required for Algorithm 4, which is very similar to Algorithm 2. The only differences are the use of less–restrictive parabolic scalings instead of triangular ones, and randomization using the integers instead of reals.

Input: A bipartite quiver datum $(\boldsymbol{V}, \boldsymbol{p})$ and $\varepsilon > 0$.

Output: An ε–balanced scaling $\boldsymbol{W}$ of $\boldsymbol{V}$.
Algorithm:

1. Let $n = \dim \boldsymbol{V}$. Choose the entries of $g \in \bigoplus_{x \in \Omega} \mathrm{Mat}_{d(x)}(\mathbb{C})$ independently and uniformly at random in $[6n^{n+1}]$.
2. Let $\boldsymbol{W}$ be the output of Algorithm 3 on $(S_g \boldsymbol{V}, \boldsymbol{p}), \varepsilon$; if any step was not possible, output **ERROR**.
3. **output** $\boldsymbol{W}$.

Algorithm 4: Algorithm for Problem 4.26.

Theorem 4.76 (Algorithm 4 running time)**.** *Let* $(\boldsymbol{V}, \boldsymbol{p})$ *be a bipartite quiver datum with at most* r *edges each with Gaussian integer entries of absolute value at most* M*,* $\sum \boldsymbol{p}_L = 1$*. Let* $n = \dim \boldsymbol{V}$*. If* $(\boldsymbol{V}, \boldsymbol{p})$ *is approximately scalable, then Algorithm 4 outputs* ***ERROR*** *with probability at most* $1/3$ *and requires*

$$O\left(\frac{n \log n + \log M + \log r}{p_{\min} \varepsilon^2}\right)$$

steps of Algorithm 3.

Proof. $(\boldsymbol{V}, \boldsymbol{p})$ is approximately scalable, by Corollary 4.46 the point $\boldsymbol{p}$ is in $\mathcal{K}(\boldsymbol{V})$. In particular, $\boldsymbol{p}$ is contained in the convex hull some set S of $n+1$ vertices of $\boldsymbol{p}$. Because every vertex of $\mathcal{K}(\boldsymbol{V})$ is a vertex of $\mathcal{P}(S_g\boldsymbol{V})$ for some $g \in \mathrm{GL}(\boldsymbol{V})$, Observation 4.72 implies that for each element $\boldsymbol{q} \in S$ there is a number $0 < M \leq n^{n/2}$ such that every entry of $\boldsymbol{q}$ takes the form k/M for $k \in \mathbb{Z}$. Replacing $\boldsymbol{q}$ by $M\boldsymbol{q}$ preserves rank–nondecreasingness and $\sum \boldsymbol{q} = M$, so Theorem 4.40 shows $\boldsymbol{q} \notin \mathcal{P}(S_g\boldsymbol{V})$ with probability at most $\frac{1}{3(n+1)}$. By the union bound, all the elements of S, and hence $\boldsymbol{p}$, are contained in $\mathcal{P}(S_g\boldsymbol{V})$ with probability at least $2/3$. If this occurs, g must be nonsingular because $\mathrm{cap}(S_g\boldsymbol{V}, \boldsymbol{p}) > 0$ is and if g is singular then $T_{\mathrm{Unif}\, S_g\boldsymbol{V}}, T^*_{\mathrm{Unif}\, S_g\boldsymbol{V}}$ do not map nonsingular operators to nonsingular operators, contradicting Lemma 4.66.

Condition on $\boldsymbol{p} \in \mathcal{P}(S_g\boldsymbol{V})$. Then $\boldsymbol{V}' = S_g\boldsymbol{V}$ has at most r edges whose entries have absolute values at most $6Mn^n$, so by Corollary 4.75 Algorithm 3 terminates in at most

$$O\left(\frac{n\log n + \log M + \log r}{p_{\min}\varepsilon^2}\right)$$

steps. □

Remark 4.77 (Numerical issues)**.** We should not expect to be able to compute each step of Algorithm 3 exactly, but rather to polynomially many bits of precision. The paper on which this chapter is based had a rather messy analysis of the rounding; [BFG+18] contains a much more pleasant analysis, which we now sketch.

In each iteration of Algorithm 3, simply compute g to some precision 2^{-t} (ensuring that they are upper triangular). We need to check that progress is still made per step and that the capacity is bounded above the entire time. For this we need to emulate the proof of Lemma 4.64 with some error; it is enough to show that the rounded g' satisfies $\det(g^\dagger g, \boldsymbol{p}) \approx \det(g'^\dagger g', \boldsymbol{p})$. Because the left–hand side is either $\det(T^*(\boldsymbol{P}_R)^{-1}, \boldsymbol{p}_L)$ or $\det(T(\boldsymbol{P}_L)^{-1}, \boldsymbol{p}_R)$, it is enough to show that the least eigenvalues of $T^*(\boldsymbol{P}_R)^{-1}$ (resp $T(\boldsymbol{P}_L)^{-1}$) are singly exponentially bounded throughout the algorithm. If this holds, the capacity is also bounded because $T(\boldsymbol{P}_L) \approx \boldsymbol{I}_L$ or $T^*(\boldsymbol{P}_R) \approx \boldsymbol{I}_L$. Though it requires proof, this can be done by choosing t polynomial in the number of steps of the algorithm and the bit complexity of the input.

4.5.2 Running time for Horn's problem

We assume $\sum \boldsymbol{\alpha} + \sum \boldsymbol{\beta} + \sum \boldsymbol{\gamma} = 1$, because there are no solutions to $H_1 + H_2 + H_3 - \frac{1}{m} I_m$ otherwise. By rescaling $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and adding multiples of the all–ones vectors, we may further assume

$$\alpha_m, \beta_m, \gamma_m \geq \frac{1}{3m+1}. \tag{4.18}$$

Theorem 4.78. *If there exist Hermitian matrices H_1, H_2, H_3 satisfying*

$$H_1 + H_2 + H_3 - \frac{1}{m} I_m$$

*with respective spectra $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ with total sum 1 satisfying Eq. (4.18), then Algorithm 4.79 outputs **ERROR** with probability at most $1/3$ and terminates in at most $O(\varepsilon^{-2} m^2 \log m)$ steps.*

Input: Nonincreasing sequences $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ of m positive real numbers and $\varepsilon > 0$.

Output: Real, symmetric matrices H_1, H_2, H_3 with respective spectra $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3$ and

$$\|H_1 + H_2 + H_3 - \frac{1}{m} I_m\| \leq \varepsilon.$$

Algorithm:

1. Choose each entry of $U_1, U_2, U_3 \in \mathrm{Mat}_m(\mathbb{R})$ independently and uniformly at random from $[24m^{4m}]$. For $i \in \{1, 2, 3\}$, define $H_i = U_i \operatorname{diag}(\boldsymbol{\lambda}_i) U_i^\dagger$.
2. **while** $\|H_1 + H_2 + H_3 - I_m\| > \varepsilon$ **do:** If any step is impossible, output **ERROR.**
   (a) Choose $B \in \mathrm{Mat}_m \mathbb{R}$ lower triangular such that
   $$B\,(H_1 + H_2 + H_3)\,B^\dagger = \frac{1}{m} I_m,$$
   and set $U_i \leftarrow B U_i$ for $i \in \{1, 2, 3\}$.
   (b) For $i \in \{1, 2, 3\}$, choose $B_i \in \mathrm{Mat}_m(\mathbb{R})$ upper triangular such that $U_i B_i$ is orthogonal and set $U_i \leftarrow U_i B_i$.
3. **output** H_1, H_2, H_3.

Algorithm 5: Implementable algorithm for Problem 4.1.

Proof of Theorem 4.78. Algorithm 4.79 is just Algorithm 4 run on the quiver datum

$$(\boldsymbol{V}, \boldsymbol{p}) = \left(\begin{array}{ccc} \boldsymbol{\lambda}_1 & & \\ & \overset{I_m}{\searrow} & \\ \boldsymbol{\lambda}_2 & \xrightarrow{I_m} & \frac{1}{m}\mathbf{1}_m \\ & \underset{I_m}{\nearrow} & \\ \boldsymbol{\lambda}_3 & & \end{array} \right).$$

By the calculations in Example 4.20 and Example 4.22, the above quiver datum is approximately scalable if and only if the desired matrices H_1, H_2, H_3 exist. If the quiver datum is approximately scalable, then by Theorem 4.76, Algorithm 4 terminates in $O(\varepsilon^{-2} m^2 \log m)$ with probability at least 2/3. □

Remark 4.79 (Convergent sequences)**.** The sequences of triples (H_1, H_2, H_3) after step (b) in have a convergent subsequence, because in step (b) they are set to lie within the compact space of triples with largest eigenvalues at most $\lambda_1, \lambda_2, \lambda_3$, respectively. ♢

Remark 4.80 (Running time for negative inputs to Problem 4.1)**.** Further, inverting the rescaling for Eq. (4.18) will increase $\|H_1 + H_2 + H_3 - \frac{1}{m} I_m\|$ by a factor of at most $O(m)$, so applying the rescaling before running the algorithm and then inverting only decreases the ϵ needed in the algorithm by a factor of $O(m)$, so there is an algorithm for positive data not satisfying Eq. (4.18) that runs in time at most $O(\varepsilon^{-2} m^4 \log m)$. However, to solve Problem 4.1 for data that might be negative, the inversion may increase the error by a factor of $O(m^2 \max\{\alpha_1 - \alpha_m, \beta_1 - \beta_m, \gamma_1 - \gamma_m\})$, so the running time for that problem is

$$O(\varepsilon^{-2} m^6 \max\{\alpha_1 - \alpha_n, \beta_1 - \beta_m, \gamma_1 - \gamma_m\}^2 \log m).$$

Appendix A

A simple algorithm for Horn's problem and its cousins

A.1 Rank–nondecreasingness calculations

Lemma 4.54. $\mathrm{Cut}_{x,p}(\boldsymbol{V},\boldsymbol{p})$ *is rank–nondecreasing if and only if* $(\boldsymbol{V},\boldsymbol{p})$ *is. In particular, cuts preserve* $\boldsymbol{p}_L(\boldsymbol{V}),\boldsymbol{p}_R(\boldsymbol{V})$ *and the maximum value of* $\boldsymbol{p}(\boldsymbol{W})$ *over all* $\boldsymbol{V}$*–independent sets* $\boldsymbol{W}$*.*

Proof. Let $(\boldsymbol{V}',\boldsymbol{p})' = \mathrm{Cut}_{x,p}(\boldsymbol{V},\boldsymbol{p})$. Checking that $\boldsymbol{p}_{L'}(\boldsymbol{V}') = \boldsymbol{p}_{R'}(\boldsymbol{V}')$ if and only if $\boldsymbol{p}_L(\boldsymbol{V}) = \boldsymbol{p}_R(\boldsymbol{V})$ is straightforward. By taking duals if need be, we may assume $x \in L$. Suppose $(\boldsymbol{V},\boldsymbol{p})$ is not rank–nondecreasing and let $\boldsymbol{W}$ be a $\boldsymbol{V}$–inependent set with $\boldsymbol{p}(\boldsymbol{W}) > \boldsymbol{p}_R(\boldsymbol{V})$. Let $\boldsymbol{W}'$ match $\boldsymbol{W}$ for all but $W_{x'}$, which we define to be $W_x \cap V_{x'} = W_x \cap F_{d(x')}$. We claim that $\boldsymbol{W}'$ is a $\boldsymbol{V}'$–independent set with $\boldsymbol{p}'(\boldsymbol{W}') > \boldsymbol{p}_{R'}(\boldsymbol{V}') = \boldsymbol{p}_R(\boldsymbol{V})$. Firstly, $\boldsymbol{W}'$ is a $\boldsymbol{V}$–independent set because for $t(e) = x$, $A_{e'}W_{x'} = A_e\iota(W_x \cap V_x) = A_e(W_x \cap V_x) \subset W_{h(e)}^{\perp}$. It is enough to show that $\boldsymbol{p}_x(W_x) = \boldsymbol{p}'_x(W_x) + \boldsymbol{p}'_{x'}(W'_{x'})$. Observe that $W_{x'} \cap F_i = W_x \cap F_i$ for $i \le d(x')$. Thus, the right–hand side of the desired equality is

$$\begin{aligned}\sum_{i=1}^{d(x)}(p'_i(x) - p'_{i+1}(x))\dim W_x \cap F_i + \sum_{i=1}^{d(x')}(p'_i(x') - p'_{i+1}(x'))\dim W_x \cap F_i \\ = \sum_{i=1}^{d(x)}\dim W_x \cap F_i(\min\{p_i(x),p\} - \min\{p_{i+1}(x),p\} \\ + \min\{p_i(x)-p,0\} - \min\{p_{i+1}(x)-p,0\}) \\ = \sum_{i=1}^{d(x)}(p_i(x) - p_{i+1}(x))\dim W_x \cap F_i,\end{aligned}$$

which is exactly $\boldsymbol{p}(W_x)$. On the other hand, suppose $\boldsymbol{W}'$ is a $\boldsymbol{V}$–independent set with $\boldsymbol{p}'(\boldsymbol{W}') > \boldsymbol{p}_{R'}(\boldsymbol{V}') = \boldsymbol{p}_R(\boldsymbol{V})$. We again take $W_y = W'_y$ for all $y \neq x'$. The replacement

W'_x by $\iota W'_{x'} + W'_x$ preserves $\boldsymbol{V}'$–independence and may not decrease $\boldsymbol{p}'(\boldsymbol{W}')$. Thus we may assume $W'_{x'} = W'_{x'} \cap F_{d(x')}$ and take $W_x = W'_x$, in which case the previous argument shows $\boldsymbol{p}'(\boldsymbol{W}') = \boldsymbol{p}(\boldsymbol{W})$. □

Lemma 4.56. *Suppose $(\boldsymbol{V}, \boldsymbol{p})$ is an ε–outer balanced bipartite quiver datum. Then $|\sum \boldsymbol{p}_L - \sum \boldsymbol{p}_R| \le 2\varepsilon\sqrt{\dim \boldsymbol{V}}$ and*

$$\boldsymbol{p}(\boldsymbol{W}) \le \sum \boldsymbol{p}_L + 2\varepsilon \dim \boldsymbol{V} \tag{A.1}$$

for all $\boldsymbol{V}$–independent sets $\boldsymbol{W}$.

Proof. Let $T = T_{\boldsymbol{V}}$. Firstly,

$$\begin{aligned} |\sum \boldsymbol{p}_L - \sum \boldsymbol{p}_R| &\le |\operatorname{tr} \boldsymbol{P}_L - \operatorname{tr} T(\boldsymbol{I}_L)| + \varepsilon \dim \boldsymbol{V} \\ &\le |\operatorname{tr} \boldsymbol{P}_L - \operatorname{tr} T^*(\boldsymbol{I}_R)| + \varepsilon\sqrt{\dim \boldsymbol{V}} \\ &\le 2\varepsilon\sqrt{\dim \boldsymbol{V}}. \end{aligned}$$

To show Eq. (A.1), we argue assuming $(\boldsymbol{V}, \boldsymbol{p}) = \operatorname{Unif}(\boldsymbol{V}, \boldsymbol{p})$. This will change ε to $\varepsilon' \le \varepsilon\sqrt{\dim V}$, and by Lemma 4.54, it does not change the maximum value of $\boldsymbol{p}(\boldsymbol{W})$ over $\boldsymbol{V}$–independent sets $\boldsymbol{W}$. Suppose $\boldsymbol{W}$ is a $\boldsymbol{V}$–independent set, and let $\pi_L = \oplus_{x \in L} \pi_{\boldsymbol{W}_x}$ and define π_R analogously. By independence, $T(\pi_L)\pi_{\boldsymbol{R}} = 0$ and $T^*(\pi_R)\pi_L = 0$. By our assumption that $\boldsymbol{p}_x = p(x)\mathbf{1}$, $p_R(\boldsymbol{W}) = \operatorname{tr} \boldsymbol{P}_L \pi_L$. Thus,

$$\begin{aligned} \boldsymbol{p}_R(\boldsymbol{W}) &= \operatorname{tr} \boldsymbol{P}_R \pi_R \\ &\le \operatorname{tr} T(\boldsymbol{I}_L)\pi_R + \varepsilon'\sqrt{\dim V} \\ &= \operatorname{tr} T^*(\pi_R) + \varepsilon\sqrt{\dim V} \\ &= \operatorname{tr} T^*(\pi_R)(I - \pi_L) + \varepsilon'\sqrt{\dim V} \\ &\le \operatorname{tr} T^*(I_R)(I - \pi_L) + \varepsilon'\sqrt{\dim V} \\ &\le \operatorname{tr} \boldsymbol{P}_L(I - \pi_L) + 2\varepsilon'\sqrt{\dim V} \\ &= \boldsymbol{p}_L(\boldsymbol{V}) - \boldsymbol{p}_L(\boldsymbol{W}) + 2\varepsilon'\sqrt{\dim V}. \end{aligned}$$

□

A.2 Capacity calculations

Lemma 4.58. *If $g \in \mathrm{GL}(F_\circ^{\boldsymbol{p}})$, then for all $X \in \mathrm{Mat}_n(\mathbb{C})$,*

$$\det(g^\dagger X g, \boldsymbol{p}) = \det(g^\dagger g, \boldsymbol{p}) \det(X, \boldsymbol{p}). \tag{A.2}$$

Proof. We have $\det(g^\dagger X g, \boldsymbol{p}) = \prod_{j=1}^n |\det \iota_j^\dagger g^\dagger X g \iota_j|^{p_j - p_{j+1}}$. Because g fixes F_j if $p_j - p_{j+1} > 0$, the right–hand side is

$$\begin{aligned}\det(g^\dagger X g, \boldsymbol{p}) &= \prod_{j=1}^n |\det \iota_j^\dagger g^\dagger \iota_j \iota_j^\dagger X \iota_j \iota_j^\dagger g \iota_j|^{p_j - p_{j+1}} \\ &= \det(X, \boldsymbol{p}) \prod_{j=1}^n \det |\iota_j^\dagger g^\dagger \iota_j \iota_j^\dagger g \iota_j| \\ &= \det(X, \boldsymbol{p}) \det(g^\dagger g).\end{aligned}$$

□

Lemma 4.61. *If $(\boldsymbol{V}, \boldsymbol{p})$ is a bipartite quiver datum, then*

$$\mathrm{cap}(\mathrm{Cut}_{p,x}(\boldsymbol{V}, \boldsymbol{p})) = \mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}). \tag{A.3}$$

Proof. Let $T = T_{\boldsymbol{V}}$ and $T' = T_{\boldsymbol{V}'}$. First suppose $x \in R$. By Eq. (4.14),

$$\begin{aligned}\det(T'_x(\boldsymbol{X}), \boldsymbol{p}'_{x'}) \det(T'_{x'}(\boldsymbol{X}), \boldsymbol{p}'_x) &= \det(T_x(\boldsymbol{X}), \boldsymbol{p}'_x) \det(\iota^\dagger T_x(\boldsymbol{X}) \iota, \boldsymbol{p}'_x) \\ &= \det(T_x(\boldsymbol{X}), \boldsymbol{p}_x).\end{aligned}$$

Thus, $\det(T'(\boldsymbol{X}), \boldsymbol{p}') = \det(T(\boldsymbol{X}), \boldsymbol{p})$ for all $\boldsymbol{X}$. Now suppose $x \in L$. We first change variables. Using the Cholesky decomposition we write $\boldsymbol{X} = h \boldsymbol{P}_L h^\dagger$ with h upper triangular, so that by Eq. (4.12) we have $\det(\boldsymbol{X}^{-1}, \boldsymbol{p}_L) = \det((h^\dagger h)^{-1}, \boldsymbol{p}_L) \det(\boldsymbol{p}_L^{-1}, \boldsymbol{p}_L) = e^{H(\boldsymbol{p}_L)} \det((h^\dagger h)^{-1}, \boldsymbol{p}_L)$. Thus

$$\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) = \inf_{h \text{ upper triangular}} \det(T(h \boldsymbol{P}_L h^\dagger), \boldsymbol{p}_R) \det((h h^\dagger)^{-1}, \boldsymbol{p}_L). \tag{A.4}$$

Next, observe that it is enough to prove Eq. (4.15) when p is at least the second largest number appearing in $(p_x(1), \ldots, p_x(d(x)+1))$. This is because the cut p can be refined by a such a sequence of cuts, and cuts commute. Suppose this is the case and write $i = d(x')$. We will have $\boldsymbol{p}'_{x'} = (p_1(x) - p)\mathbf{1}_{d(x')} := q 1_{d(x')}$. We need only show that the

infimum does not increase if performed only over h' of the form $\mathrm{Cut}_{p,x} h$, for then we obtain

$$\begin{aligned}
\mathrm{cap}(\boldsymbol{V}', \boldsymbol{p}') &= \inf_{h'=\mathrm{Cut}_{p,x} h} \det(T(h' \boldsymbol{P}'_L h'^{\dagger}), \boldsymbol{p}_R) \det((h'h'^{\dagger})^{-1}, \boldsymbol{p}'_L) \\
&= \inf_{g=(h,\boldsymbol{I}_R) \text{ upper triangular}} \det(T_{S_g \boldsymbol{V}'}(\boldsymbol{P}'_L), \boldsymbol{p}_R) \det((hh^{\dagger})^{-1}, \boldsymbol{p}_L) \\
&= \inf_{g=(h,\boldsymbol{I}_R) \text{ upper triangular}} \det(T_{S_g \boldsymbol{V}}(\boldsymbol{P}_L), \boldsymbol{p}_R) \det((hh^{\dagger})^{-1}, \boldsymbol{p}_L) \\
&= \inf_{h \text{ upper triangular}} \det(T(h \boldsymbol{P}_L h^{\dagger}), \boldsymbol{p}_R) \det((hh^{\dagger})^{-1}, \boldsymbol{p}_L) \\
&= \mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}).
\end{aligned}$$

where for the second inequality we used Eq. (4.14) and Eq. (4.6) and in the third we used Eq. (4.5). To show that we only need only infimize over $h' = \mathrm{Cut}_{p,x} h$, it is enough to show that we can set the upper–left $d(x') \times d(x')$ corner $\iota^{\dagger} h'_x \iota$ of h'_x equal to $h'_{x'}$ while holding $h'_x P'_x h'_x + q \iota h'_{x'} h'_{x'} \iota^{\dagger} = Y$ fixed and not increasing $\det((h'_x h'^{\dagger}_x)^{-1}, \boldsymbol{p}'_x) \det(h'_{x'} h'^{\dagger}_{x'})^{-q}$. To show this, let $h = h'_x$ and $h' = h'_{x'}$, $P' = P'_x$, $n = d(x), n' = d(x')$. Because h is upper triangular, $\det((hh^{\dagger})^{-1}, \boldsymbol{p}') = \prod_{i=1}^{n} |h_{ii}|^{-2p'_i}$. Modifying the upper–left $n' \times n'$ corner of h does not change $\prod_{i>n'}^{n} |h_{ii}|^{-2p'_i}$, and $\prod_{i=1}^{n'} |h_{ii}|^{-2p'_i} = \det(\iota^{\dagger} h \iota \iota^{\dagger} h^{\dagger} \iota)^{-p}$. Set $X = \iota^{\dagger} h \iota \iota^{\dagger} h^{\dagger} \iota$ and $h' h'^{\dagger} = Z$; we need only show that $p \log\det X + q \log\det(Z)$ does not decrease. To hold $h P' h^{\dagger} + q \iota h' h'^{\dagger} = Y$ fixed while modifying only $\iota^{\dagger} h \iota$ it is enough to hold $pX + q h' h'^{\dagger} := pX + qZ = Y$ fixed. The log-concavity of determinant implies $p \log\det X + q \log\det Z \leq (p+q) \log\det \left(p \frac{X}{p+q} + q \frac{Z}{p+q} \right) = (p+q) \log\det \left(\frac{Y}{p+q} \right)$, so replacing X and Z by $\frac{Y}{p+q}$ has the desired properties. The convention $0 \log 0 = 0$ ensures the above argument is valid even if $p = 0$. □

Lemma 4.66. *Suppose $(\boldsymbol{V}, \boldsymbol{p})$ is a uniform positive bipartite quiver datum with $\sum p_L = \sum p_R$ and $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$. Then both $T_{\boldsymbol{V}}, T^*_{\boldsymbol{V}}$ maps positive definite matrices to positive definite matrices.*

Proof of Lemma 4.66. If $T_{\boldsymbol{V}}(\boldsymbol{X})$ is nonsingular, then $\det(T_{\boldsymbol{V}}(\boldsymbol{X}), \boldsymbol{p}_R) = 0$, which contradicts $\mathrm{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$. On the other hand, if $T^*_{\boldsymbol{V}}(\boldsymbol{X})$ is singular for some positive definite $\boldsymbol{X}$, then $T^*_{\boldsymbol{V}}(\boldsymbol{I}_L)$ is singular because $\alpha \boldsymbol{I}_L \preceq \boldsymbol{X}$ for some $\alpha > 0$. Then for some nonzero projection $\boldsymbol{\pi} = \oplus_{x \in L} \pi_x$ we have $\mathrm{tr}\, T^*_{\boldsymbol{V}}(\boldsymbol{I}_R) \boldsymbol{\pi} = 0$ and hence $\mathrm{tr}\, T_{\boldsymbol{V}}(\boldsymbol{\pi}) = 0$.

Thus $T_{\boldsymbol{V}}(\boldsymbol{\pi}) = 0$, so $T_{\boldsymbol{V}}(\boldsymbol{\pi} + (\boldsymbol{I}_L - \boldsymbol{\pi})) = \varepsilon T_{\boldsymbol{V}}(\boldsymbol{I}_L - \boldsymbol{\pi})$. Take $\varepsilon \to 0$, so that

$$\det(T(\pi + \varepsilon(\boldsymbol{I}_L - \boldsymbol{\pi})), \boldsymbol{p}_R) \det(\boldsymbol{\pi} + \varepsilon(I - \boldsymbol{\pi}), \boldsymbol{p}_L) = O\left(\varepsilon^{\sum \boldsymbol{p}_R - \sum_{x \in L} p(x)(d(x) - \operatorname{rank} \pi_x)}\right).$$

By $\sum \boldsymbol{p}_R = \sum \boldsymbol{p}_L$, the above is $o(1)$. This contradicts $\operatorname{cap}(\boldsymbol{V}, \boldsymbol{p}) > 0$. □

A.3 Probabilistic fact

Fact 4.69. *If X is a discrete random variable with $\mathbb{E}[X] = 1$ that takes every value in its support with probability at least $p_{\min}$, then*

$$\mathbb{E}[\log X] \leq -.06 \min\{\mathbb{V}[X], p_{\min}\}.$$

Proof. Using the Lagrange remainder bound for a quadratic approximation about $x = 1$ and concavity, the logarithm is bounded by the piecewise function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} (x-1) - .1(x-1)^2 & 0 \leq x \leq 1.2 \\ .95(x-1) & x > 1.2. \end{cases}$$

Thus,

$$\begin{aligned} \mathbb{E}[\log X] &\leq \mathbb{E}[f(X)] \\ &= \mathbb{E}[(X-1) - .1(X-1)^2 | X \leq 1.2] \Pr[X \leq 1.2] + .95\mathbb{E}[(X-1)|X > 1.2] \Pr[X > 1.2] \\ &= -.1\mathbb{E}[(X-1)^2 | X \leq 1.2] \Pr[X \leq 1.2] + \mathbb{E}[(X-1)] - .05\mathbb{E}[(X-1)|X > 1.2] \Pr[X\ 1.2] \\ &\leq -.1\mathbb{E}[(X-1)^2 | X \leq 1.2] \Pr[X \leq 1.2] - .06 \Pr[X > 1.2] \end{aligned}$$

If $\Pr[X > 1.2] = 0$, then the first term is the variance $\mathbb{V}[X]$. If not, $\Pr[X > 1.2] \geq p_{\min}$. This completes the proof. □

Appendix B

Discrepancy of random matrices with many columns

B.1 Random walk over $\mathbb{F}_2^m$

Lemma 3.25. *Suppose X_n is a sum of n uniformly random vectors of Hamming weight $0 < t < m$ in $\mathbb{F}_2^m$ and Z_n is a uniformly random element of $\mathbb{F}_2^m$ with Hamming weight having the same parity as nt. If d_{TV} denotes the total variation distance, then*

$$d_{TV}(X_n, Z_n) = O(e^{-(2n/m)+m}).$$

Proof. Though we will not use the language of Markov chains, the following calculation consists of showing that a random walk on the group $\mathbb{F}_2^m$ mixes rapidly by showing it has a spectral gap.

Let X be a random element $\mathbb{F}_2^m$ of Hamming weight t. Let f be the probability mass function of X. Let h_n be the probability mass function of Z_n. By the Cauchy-Schwarz inequality, it is enough to show that the probability mass function f_n of the sum of n i.i.d. copies of X satisfies

$$\sum_{\boldsymbol{x}\in\mathbb{F}_2^m} |f_n(\boldsymbol{x}) - h_n(\boldsymbol{x})|^2 = O(e^{-2n/m})$$

For $\boldsymbol{y} \in \mathbb{F}_2^m$, let $\chi_{\boldsymbol{y}} : \mathbb{F}_2^m \to \{\pm 1\}$ be the Walsh function $\chi_{\boldsymbol{y}}(\boldsymbol{x}) = (-1)^{\boldsymbol{y}\cdot\boldsymbol{x}}$. The Fourier transform of a function $g : \mathbb{F}_2^m \to \mathbb{R}$ is the function $\widehat{g} : \mathbb{F}_2^m \to \mathbb{R}$ given by $\widehat{g}(\boldsymbol{y}) = \sum_{\boldsymbol{x}\in\mathbb{F}_2^m} g(\boldsymbol{x})\chi_{\boldsymbol{y}}(\boldsymbol{x})$. The function f_n satisfies $\widehat{f_n} = (\widehat{f}\,)^n$. Note that $\widehat{h}_n(0) = \widehat{f_n}(0) = 1$, $\widehat{h}_n(\mathbf{1}) = \widehat{f_n}(\mathbf{1}) = (-1)^{nt}$, and $\widehat{h} = 0$ elsewhere. By Plancherel's identity,

$$\sum_{\boldsymbol{x}\in\mathbb{F}_2^m} |f_n(\boldsymbol{x}) - h_n(\boldsymbol{y})|^2 = \mathbb{E}_{\boldsymbol{y}\in\mathbb{F}_2^m} |\widehat{f}(\boldsymbol{y})^n - \widehat{h}_n(\boldsymbol{y})|^2 \tag{B.1}$$

$$= \sum_{\boldsymbol{y}\in\mathbb{F}_2^m\setminus\{0,\mathbf{1}\}} 2^{-n} |\widehat{f}(\boldsymbol{y})|^{2n}. \tag{B.2}$$

Now we claim that $|\widehat{f}(\boldsymbol{y})| \leq 1 - \frac{1}{m}$ for $\boldsymbol{y} \notin \{0, \mathbf{1}\}$, which would imply Eq. (B.2) is at most $(1 - 1/m)^{2n} \leq e^{-2n/m}$. Indeed, if the Hamming weight of $\boldsymbol{y}$ is s, then $\widehat{f}(\boldsymbol{y})$ is exactly the expectation of $(-1)^{|S \cap T|}$ where T is a random t-set and S a fixed s-set. By symmetry we may assume $t \leq s$, and since we are only concerned with the absolute value of this quantity, by taking the complement of S we may assume $s \leq m/2$. We may choose the elements of T in order; it is enough to show that the expectation of $(-1)^{|S \cap T|}$ is at most $1 - 1/m$ in absolute value even after conditioning on the choice of the first $t - 1$ elements of T. Indeed, the value of $(-1)^{|S \cap T|}$ is not determined by this choice, so the conditional expectation is a rational number in $(-1, 1)$ with denominator at most m, and hence at most $1 - 1/m$ in absolute value. $\square$

References

[AS04] Noga Alon and Joel H Spencer, *The probabilistic method*, John Wiley & Sons, 2004.

[Ban98] Wojciech Banaszczyk, *Balancing vectors and gaussian measures of n-dimensional convex bodies*, Random Structures and Algorithms **12** (1998), no. 4, 351–360.

[Bar98] Franck Barthe, *On a reverse form of the Brascamp–Lieb inequality*, Inventiones mathematicae **134** (1998), no. 2, 335–361.

[BCCT08] Jonathan Bennett, Anthony Carbery, Michael Christ, and Terence Tao, *The Brascamp–Lieb inequalities: finiteness, structure and extremals*, Geometric and Functional Analysis **17** (2008), no. 5, 1343–1415.

[BF81] József Beck and Tibor Fiala, *"integer-making" theorems*, Discrete Applied Mathematics **3** (1981), no. 1, 1–8.

[BFG+18] Peter Bürgisser, Cole Franks, Ankit Garg, Rafael Oliveira, Michael Walter, and Avi Wigderson, *Efficient algorithms for tensor scaling, quantum marginals and moment polytopes*, arXiv preprint arXiv:1804.04739 (2018).

[Cos09] Kevin P Costello, *Balancing Gaussian vectors*, Israel Journal of Mathematics **172** (2009), no. 1, 145–156.

[DGOS16] Zeev Dvir, Ankit Garg, Rafael Oliveira, and József Solymosi, *Rank bounds for design matrices with block entries and geometric applications*, arXiv preprint arXiv:1610.08923 (2016).

[DM17] Harm Derksen and Visu Makam, *Polynomial degree bounds for matrix semi-invariants*, Advances in Mathematics **310** (2017), 44–63.

[DW00] Harm Derksen and Jerzy Weyman, *Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients*, Journal of the American Mathematical Society **13** (2000), no. 3, 467–479.

[DZ01] Mátyás Domokos and Alexander N Zubkov, *Semi-invariants of quivers as determinants*, Transformation groups **6** (2001), no. 1, 9–24.

[EL16] Esther Ezra and Shachar Lovett, *On the Beck-Fiala conjecture for random set systems*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (2016).

[EM65] P Erdős and JW Moon, *On sets of consistent arcs in a tournament*, Canad. Math. Bull **8** (1965), 269–271.

[For02] Jürgen Forster, *A linear lower bound on the unbounded error probabilistic communication complexity*, Journal of Computer and System Sciences **65** (2002), no. 4, 612–625.

[Fra02] Matthias Franz, *Moment polytopes of projective G-varieties and tensor products of symmetric group representations*, J. Lie Theory **12** (2002), no. 2, 539–549.

[Fra18a] Cole Franks, *Operator scaling with specified marginals*, arXiv preprint arXiv:1801.01412 (2018).

[Fra18b] ———, *Operator scaling with specified marginals*, Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2018, pp. 190–203.

[Fra18c] ———, *A simplified disproof of Beck's three permutations conjecture and an application to root-mean-squared discrepancy*, arXiv preprint arXiv:1811.01102 (2018).

[Fri16] Shmuel Friedland, *On Schrodinger's bridge problem*, arXiv preprint arXiv:1608.05862 (2016).

[FS18] Cole Franks and Michael Saks, *On the discrepancy of random matrices with many columns*, arXiv preprint arXiv:1807.04318 (2018).

[Ful00] William Fulton, *Eigenvalues, invariant factors, highest weights, and schubert calculus*, Bulletin of the American Mathematical Society **37** (2000), no. 3, 209–249.

[GGOW16a] Ankit Garg, Leonid Gurvits, Rafael Oliveira, and Avi Wigderson, *Algorithmic aspects of Brascamp–Lieb inequalities*, arXiv preprint arXiv:1607.06711 (2016).

[GGOW16b] ———, *A deterministic polynomial time algorithm for non-commutative rational identity testing*, Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on, IEEE, 2016, pp. 109–117.

[GMR08] Venkatesan Guruswami, Rajsekar Manokaran, and Prasad Raghavendra, *Beating the random ordering is hard: Inapproximability of maximum acyclic subgraph*, Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on, IEEE, 2008, pp. 573–582.

[GP15] Tryphon T Georgiou and Michele Pavon, *Positive contraction mappings for classical and quantum Schrödinger systems*, Journal of Mathematical Physics **56** (2015), no. 3, 033301.

[GS02] Leonid Gurvits and Alex Samorodnitsky, *A deterministic algorithm for approximating the mixed discriminant and mixed volume, and a combinatorial corollary*, Discrete & Computational Geometry **27** (2002), no. 4, 531–550.

[Gur04] Leonid Gurvits, *Classical complexity and quantum entanglement*, Journal of Computer and System Sciences **69** (2004), no. 3, 448–484.

[Gur08] ———, *Van der Waerden/Schrijver-Valiant like conjectures and stable (aka hyperbolic) homogeneous polynomials: one theorem for all*, the electronic journal of combinatorics **15** (2008), no. 1, 66.

[Hor54] Alfred Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, American Journal of Mathematics **76** (1954), no. 3, 620–630.

[HR18] Rebecca Hoberg and Thomas Rothvoss, *A Fourier-analytic approach for the discrepancy of random set systems*, arXiv preprint arXiv:1806.04484 (2018).

[JS] P. Tetali J.H. Spencer, A. Srinivasan, *The discrepancy of permutation families*.

[KLP12] Greg Kuperberg, Shachar Lovett, and Ron Peled, *Probabilistic existence of rigid combinatorial structures*, Proceedings of the forty-fourth annual ACM symposium on Theory of computing, ACM, 2012, pp. 1091–1106.

[KT00] Allen Knutson and Terence Tao, *Honeycombs and sums of Hermitian matrices*.

[Lar17] Kasper Green Larsen, *Constructive discrepancy minimization with hereditary l2 guarantees*, arXiv preprint arXiv:1711.02860 (2017).

[LSV86] László Lovász, Joel Spencer, and Katalin Vesztergombi, *Discrepancy of set-systems and matrices*, European Journal of Combinatorics **7** (1986), no. 2, 151–160.

[LSW98] Nathan Linial, Alex Samorodnitsky, and Avi Wigderson, *A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents*, Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM, 1998, pp. 644–652.

[Mat13] Jiří Matoušek, *The determinant bound for discrepancy is almost tight*, Proceedings of the American Mathematical Society **141** (2013), no. 2, 451–460.

[MNS12] Ketan D Mulmuley, Hariharan Narayanan, and Milind Sohoni, *Geometric complexity theory iii: on deciding nonvanishing of a littlewood–richardson coefficient*, Journal of Algebraic Combinatorics **36** (2012), no. 1, 103–110.

[NC02] Michael A Nielsen and Isaac Chuang, *Quantum computation and quantum information*, 2002.

[NM84] Linda Ness and David Mumford, *A stratification of the null cone via the moment map*, American Journal of Mathematics **106** (1984), no. 6, 1281–1329.

[NN11] Alantha Newman and Aleksandar Nikolov, *A counterexample to Beck's conjecture on the discrepancy of three permutations*, arXiv preprint arXiv:1104.2922 (2011).

[NNN12] Alantha Newman, Ofer Neiman, and Aleksandar Nikolov, *Beck's three permutations conjecture: A counterexample and some consequences*, Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, IEEE, 2012, pp. 253–262.

[NTZ13] Aleksandar Nikolov, Kunal Talwar, and Li Zhang, *The geometry of differential privacy: the sparse and approximate cases*, Proceedings of the forty-fifth annual ACM symposium on Theory of computing, ACM, 2013, pp. 351–360.

[RS89] Uriel G Rothblum and Hans Schneider, *Scalings of matrices which have prespecified row sums and column sums via optimization*, Linear Algebra and its Applications **114** (1989), 737–764.

[Rud99] Mark Rudelson, *Random vectors in the isotropic position*, Journal of Functional Analysis **164** (1999), no. 1, 60–72.

[Sin64] Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879.

[Spe87] Joel H Spencer, *Ten lectures on the probabilistic method*, vol. 52, Society for Industrial and Applied Mathematics Philadelphia, PA, 1987.

[SV+13] Nikhil Srivastava, Roman Vershynin, et al., *Covariance estimation for distributions with 2 plus epsilon moments*, The Annals of Probability **41** (2013), no. 5, 3081–3111.