How to learn a quantum state

John Wright

CMU-CS-16-108
May 2016

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Ryan O'Donnell, Chair
Anupam Gupta
Venkatesan Guruswami
Aram Harrow, MIT

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Supported by NSF grants CCF-0747250 and CCF-1116594 and also a Simons Fellowship in Theoretical Computer Science. Some of this work was completed while visiting Columbia University.

Keywords: Quantum tomography, property testing, longest increasing subsequences, RSK algorithm, Schur-Weyl duality

To my parents.

Abstract

The subject of this thesis is learning and testing properties of mixed quantum states. A mixed state is described by a density matrix $\rho \in \mathbb{C}^{d\times d}$. In the standard model, one is given access to many identical copies of the mixed state, and the goal is to perform measurements on the copies to infer some information about ρ. In our problem, each copy of ρ plays a role analogous to a sample drawn from a probability distribution, and just as we aim to minimize sample complexity in classical statistics, here we aim to minimize copy complexity. Our results are:

- We give new upper bounds for the number of copies needed to learn the matrix ρ and the best low rank approximation to ρ, matching the lower bounds of [HHJ+16]. This settles the copy complexity of the *quantum tomography* problem (up to constant factors) and gives a first-of-its-kind principal-component-analysis-style guarantee for learning approximately low rank states. In addition, we give new upper bounds for the number of copies needed to learn the entire spectrum of ρ and the largest eigenvalues of ρ. We then show matching lower bounds for these latter problems for a popular spectrum learning algorithm, the *empirical Young diagram algorithm* of [ARS88, KW01].
- We consider testing properties of ρ and its spectrum in the standard property testing model [RS96, BFR+00]. We show matching upper and lower bounds for the number of copies needed to test if ρ is the “maximally mixed state”. This can be viewed as the quantum analogue of Paninski’s sharp bounds for classical uniformity-testing [Pan08]. In addition, we give a new upper bound for testing whether ρ is low rank. Finally, we give almost matching upper and lower bounds for the problem of distinguishing whether ρ is maximally mixed on a subspace of dimension r or of dimension $r + \Delta$.

Our quantum results exploit a new connection to the combinatorial subject of longest increasing subsequences (LISes) of random words and require us to prove new results in this area. These results include:

- We give a new and optimal bound on the expected length of the LIS in a random word. Furthermore, we show optimal bounds for the “shape” of the Young diagram resulting from applying the “RSK algorithm” to a random word.
- We prove a majorization theorem for the RSK algorithm applied to random words. It states, roughly, that random words drawn from more “top-heavy” distributions will tend to produce more “top-heavy” Young diagrams when the RSK algorithm is applied to them.

Acknowledgments

I'm grateful to my advisor (and doppleganger) Ryan O'Donnell for six years of top-notch mentorship. He has always been generous with his time, a great collaborator, and mindful of the difficulties a clueless grad student faces when learning how to research. A couple of years ago we jumped into the abyss of quantum computing together, and looking back it couldn't have gone better. Effortlessly cool, a brilliant mind, and a good taste in blogs too: what more could you ask from an advisor?

As a grad student, I spent a pair of wonderful summers interning outside of Pittsburgh. I'd like to thank Madhur Tulsiani for hosting me when I was a summer intern in Chicago and Rocco Servedio for hosting me when I was a summer casual in New York (♫) (and for reading *Don Quixote* out loud during our meetings). In addition, I'd like to thank my thesis committee, Anupam Gupta, Venkatesan Guruswami, and Aram Harrow, for their time and many great suggestions. Finally, I'd like to thank my coauthors Per Austrin, Boaz Barak, Johan Håstad, Sanxia Huang, Akshay Krishnamurthy, Euiwoong Lee, Rajsekar Manokaran, Ankur Moitra, Ryan O'Donnell, Prasad Raghavendra, Oded Regev, Melanie Schmidt, Rocco Servedio, David Steurer, Xiaorui Sun, Li-Yang Tan, Luca Trevisan, Madhur Tulsiani, Aravindan Vijayaraghavan, Andrew Wan, David Witmer, Chenggang Wu, Yu Zhao, and Yuan Zhou.

Carnegie Mellon University has a fine group of administrative staff who make grad school an all-around more enjoyable experience. Among them, I'd like to single out the all-seeing and all-powerful Deb Cavlovich, who has gone to bat for me more times than I can count, Catherine Copetas, who turned out to be right about *Madama Butterfly*, and Angie Miller, who has been quite helpful lately with the Pittsburgh public transport system.

My six years at CMU have been the best of my life, and for that I have my friends to thank. There are too many to list, so let me instead list some of my favorite memories: the big ones were our spring break in Puerto Rico, the canoeing trip in Quetico, and the roadtrip to Cleveland. I'll always remember all of our ping pong games, lunch conversations, kayaking days, Avalon games, and karaoke nights. A great group of people!

Contents

Chapter 1

Introduction

The subject of this thesis is how to learn a quantum state. A quantum state is described by a $d \times d$ matrix ρ which characterizes the state's behavior under quantum operations. In this thesis, we will give algorithms for learning the matrix ρ—thereby solving the so-called *quantum tomography* problem—and for learning specific properties of ρ, such as its spectrum. Algorithms for these problems are of enormous practical importance for real-world verification of current-day quantum technologies. In addition, these algorithms are of future theoretical importance, as they play key roles in quantum protocols such as entanglement detection [HE02, GT09].

The two scenarios we will typically keep in mind are: (i) you are given a quantum device promised to output a quantum state with a particular matrix ρ, and you'd like to learn the matrix that it *actually* outputs to verify that it works properly; (ii) you have performed a quantum experiment which you have hypothesized will output a quantum state with a particular matrix ρ, and to test your experimental hypothesis you'd like to learn the state. Quantum mechanics provides only one way for classical observers to learn about quantum states: quantum measurements, in which one "observes" the quantum state and receives a random "outcome" depending on ρ. Unfortunately, quantum measurements are (i) destructive, meaning they render the state useless for future measurements (this is the "collapse of the wave function"), and they are (ii) low-information, meaning that they reveal little about the matrix ρ. Either of these in isolation would not be particularly troubling, but in combination they appear to render state learning impossible.

The standard fix is to repeatedly run the device (or experiment) to produce many identical copies of the quantum state and to either use each new copy to perform a different measurement or to perform one giant measurement across all of the copies simultaneously (a so-called *entangled* measurement). Each copy is viewed as being expensive to produce, and so we would like to learn with as few copies as possible. This introduces a new resource measure, the *copy complexity*, which is a quantum analogue of the sample complexity from statistics.

Let us give two examples of state learning in action.

- In [MHS$^+$12], the authors demonstrated long-range quantum teleportation by teleporting the state of a photon (encoded in its polarization) to another photon 143 km away. To verify successful teleportation, they learned the 2×2 matrix of the photon on the receiving end, and checked that this agreed with the matrix of the original state. In total, they used $n = 605$ copies of the state.

- In [HHR+05], the authors constructed a device to generate 8-particle W-states, meaning that the joint state of the 8 particles was given by a particular 256×256 matrix. W-states are potentially useful as resources in fault-tolerant quantum communication protocols. To verify the device worked properly, they generated $n = 656100$ copies of the 8 particles and measured each copy separately, taking 10 hours in total. They then computed the maximum likelihood estimate of the state based on the measurement outcomes and found that it had "fidelity" 0.72 with the desired state.

Quantum state learning dates back to the 1950s [Hua12], and in spite of this, the optimal copy complexity for many basic problems remains poorly understood. For example, as of early 2015, the complexity of quantum state tomography was "shockingly unknown" [Har15]. In this thesis, we settle the copy complexity of quantum tomography, showing that $n = O(d^2/\epsilon^2)$ copies are sufficient to learn ρ up to error ϵ in trace distance (see Definition 1.4.1), matching a lower bound proved by [HHJ+16]. In addition, we give a variety of new algorithms for problems like spectrum learning, principal component analysis, and mixedness testing.

Our second contribution is a new framework for analyzing these quantum state learning algorithms which relates this topic to the combinatorial topic of longest increasing subsequences of random words. Here, given a probability distribution $\alpha = (\alpha_1, \ldots, \alpha_d)$, we let $\boldsymbol{w}$ be an n-letter α-random word, meaning that each $\boldsymbol{w}_i$ is independently distributed according to α. Then the key question is:

What is the expected length of the longest increasing subsequence of an n-letter α-random word?

Our framework shows that sufficiently tight answers to questions like this yield optimal algorithms for learning quantum states. Motivated by this, we answer this question and many other basic questions in this area which were surprisingly unresolved.

The remainder of this chapter expands upon this introduction. It is organized as follows.

- Section 1.1 explains how quantum states are represented mathematically by matrices.
- Section 1.2 gives an introduction to the basics of quantum measurements.
- Section 1.3 surveys probability distribution learning and testing, the classical analogue of the quantum problems we consider in this thesis.
- Section 1.4 states the problems we consider and our main results.
- Section 1.5 explains our methodology and how quantum state learning is connected to longest increasing subsequences.
- Section 1.6 gives the outline for the rest of the thesis.

1.1 Quantum states

To store data in a quantum system, such as an atom, we pick a property of the quantum system and encode our data into the state of this property. We say that it is a *d-level quantum system* if the property exhibits *d perfectly distinguishable states*, meaning that if the quantum system is in one of the states, then there is a measurement that will detect which one with certainty. The state of a d-level quantum system is represented as a vector in $\mathbb{C}^d$. In this work, we will use Dirac's bra-ket notation, in which $|v\rangle \in \mathbb{C}^d$ denotes a column vector and $\langle v| := |v\rangle^\dagger$ denotes a row vector. It follows that $\langle u| \cdot |v\rangle$ is the usual inner product between $|u\rangle$ and $|v\rangle$, which is typically simplified as $\langle u|v\rangle$. The following are some examples of how quantum states are encoded as vectors.

- An electron has a property called *spin*, which has two distinct states: up ($\uparrow$) and down ($\downarrow$). We represent these two states with the column vectors

  $$|{\uparrow}\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad |{\downarrow}\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

  More generally, the system may be in a *superposition* of these two states, in which case its state is given by a vector $|v\rangle = \alpha_\uparrow |{\uparrow}\rangle + \alpha_\downarrow |{\downarrow}\rangle$ satisfying $|\alpha_\uparrow|^2 + |\alpha_\downarrow|^2 = 1$. Equivalently, $|v\rangle$ satisfies $\langle v|v\rangle = 1$.

- A photon has a property called *polarization*, which has two distinct states: horizontal and vertical. These give rise to the basis vectors $|\mathrm{H}\rangle$ and $|\mathrm{V}\rangle$; as in the case of electron spin, allowing for these and any possible superpositions means that the polarization state may be any unit-norm vector $|v\rangle \in \mathbb{C}^2$.

- The spin state of *two* electrons has four distinct possibilities: $\uparrow\uparrow$, $\uparrow\downarrow$, $\downarrow\uparrow$, and $\downarrow\downarrow$, represented by the four vectors

  $$|{\uparrow\uparrow}\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad |{\uparrow\downarrow}\rangle = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad |{\downarrow\uparrow}\rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and} \quad |{\downarrow\downarrow}\rangle = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

  Allowing for superpositions, the state may be given by any unit-norm vector $|v\rangle \in \mathbb{C}^4$. In some cases the spin states of the two electrons act totally independently, meaning that the state of the first electron is given by some $|v_1\rangle \in \mathbb{C}^2$ and the state of the second electron is given by some $|v_2\rangle \in \mathbb{C}^2$, and the state of the whole system is given by the tensor-product vector $|v\rangle = |v_1\rangle \otimes |v_2\rangle$; such states are said to be *unentangled*. However, there are some vectors $|v\rangle \in \mathbb{C}^4$ which cannot be written in this way, and these vectors correspond to *entangled* states.

We also refer to a 2-level system as a *qubit* and a d-level system as a *qudit*. The first two are examples of qubits whereas the third is an example of a $d = 4$ qudit.

Quantum states represented as vectors are called *pure states*. More generally, a quantum system can be in a *mixed state*, in which its state is a random distribution (a statistical mixture) over pure states, as follows.

$$|\boldsymbol{v}\rangle = \begin{cases} |v_1\rangle & \text{with probability } p_1, \\ |v_2\rangle & \text{with probability } p_2, \\ \dots & \\ |v_m\rangle & \text{with probability } p_m. \end{cases} \tag{1.1}$$

(Note that the $|v_i\rangle$'s need not be orthogonal, and so m may be smaller or larger than the dimensionality d.) Mixed states commonly occur in quantum computing. For example, a computer may first flip some coins before deciding which pure state to output, or some noise may be applied to a pure state, perturbing it to a randomly-distributed nearby vector. A third, more exotic example occurs due to quantum entanglement: for example, if the joint spin state of two electrons is given by a vector $|v\rangle \in \mathbb{C}^4$, then the appropriate way of describing the state of the first electron by itself is with a mixed state computed by applying the "partial trace" to $|v\rangle$. (In the case when the state is unentangled, this mixed state will assign all of its probability mass to a single pure state.) Related to this, if one has a multiparticle quantum system and measures some of the particles, then the remaining particles will collapse to a state depending on the random measurement outcome. Hence, they are in a mixed state.

Associated with each mixed state is its *density matrix* ρ, defined as

$$\rho := \sum_{i=1}^{m} p_i \, |v_i\rangle \, \langle v_i| \, . \tag{1.2}$$

Though the mapping $|\boldsymbol{v}\rangle \mapsto \rho$ is lossy (in particular, multiple mixed states may have the same density matrix), the density matrix gives a full characterization of the behavior of the quantum system under all possible quantum operations and measurements. This means that two mixed states with the same density matrix are indistinguishable from one another, and we therefore think of them as being the same. As a result, it is usually convenient to work solely with a quantum system's density matrix, ignoring the issue of which particular ensemble of pure states gave rise to it. This motivates the following definition.

Definition 1.1.1. A *density matrix* $\rho \in \mathbb{C}^{d \times d}$ is any Hermitian positive semidefinite matrix with trace one. Equivalently, if $\alpha_1, \dots, \alpha_d$ are ρ's eigenvalues, then $\alpha_i \geq 0$ for each i, and $\sum_i \alpha_i = 1$. In other words, $(\alpha_1, \dots, \alpha_d)$ is a probability distribution on the set $\{1, \dots, d\}$.

It is easy to check that any matrix of the form (1.2) satisfies this definition. The reverse is true as well: any density matrix corresponds to at least one mixed state. This is because a density matrix has d eigenvalues $\alpha_1, \dots, \alpha_d$ corresponding to d orthonormal eigenvectors $|v_1\rangle, \dots, |v_d\rangle$, and hence represents the mixed state "output $|v_i\rangle$ with probability α_i".

1.2 A quantum primer

The simplest type of measurement is a *basis measurement*. In a basis measurement, we specify in advance d orthonormal vectors $|u_1\rangle, \dots, |u_d\rangle \in \mathbb{C}^d$ corresponding to d distinct outcomes. The result of a quantum measurement is a probabilistic outcome $\boldsymbol{i}$ from the set $[1, \dots, d]$. Here we are using the following notation.

Definition 1.2.1. For a positive integer m, we define $[m] := \{1, \ldots, m\}$.

If our quantum system is described by the pure state $|v\rangle$, then this outcome is distributed as

$$\mathbf{Pr}[\text{observe outcome } i] = |\langle u_i|v\rangle|^2 = \langle u_i|v\rangle \langle v|u_i\rangle .$$

That these d probability values form a valid probability distribution follows from the Pythagorean theorem. More generally, if the quantum system is in a mixed state, as in (1.1), then we can calculate the outcome distribution as

$$\begin{aligned}\mathbf{Pr}[\text{observe outcome } i] &= \sum_{j=1}^{m} p_j \cdot \mathbf{Pr}[\text{observe outcome } i \text{ on } |v_j\rangle] \\ &= \sum_{j=1}^{m} p_j \cdot \langle u_i|v_j\rangle \langle v_j|u_i\rangle = \langle u_i| \rho |u_i\rangle .\end{aligned}$$

Note that this depends only on ρ.

Following the measurement, if the outcome i was observed, then the state of the system *collapses* and becomes the observed pure state $|u_i\rangle$. Thus, any future measurements on the quantum system will yield no additional information about the original state ρ. This is a problem if one is trying to learn any significant amount of information about ρ: a basis measurement returns one of d possibilities, and hence provides at most $\log d$ bits of information about ρ. If, say, one is trying to learn *all* of ρ—the entire $d \times d$ matrix—then one is trying to learn $\Theta(d^2)$ unknowns, and this roughly corresponds to trying to learn $\Theta(d^2)$ bits of information. These two quantities—$\log d$ and d^2—differ by orders of magnitude, meaning that a single measurement is inadequate for the task at hand.

The solution to this problem is to recall the motivating setup from the beginning: we were not just given a single quantum system whose state is represented by the matrix ρ, but a device (or an experiment) capable of generating ρ. If we repeatedly run this device, and can guarantee independence between the different runs, then we will generate many identical copies of ρ, freeing us to perform as many measurements as desired. This motivates the following definition.

Definition 1.2.2. Given a quantum algorithm for learning some property of a quantum state ρ, the *copy complexity* is denoted by n and refers to the number of copies of ρ the algorithm uses.

Copy complexity can be viewed as a quantum mechanical analogue of sample complexity from classical statistics, and in general we aim for algorithms which minimize n. There are numerous reason why copy complexity is an interesting resource to study in this context. For example, the quantum teleportation experiment [MHS+12] only required 605 copies for their application, but generating each copy was a highly error-prone process which took 6.5 hours in total. On the other hand, the experiment of [HHR+05] was able to more easily generate copies of their quantum state, but the large number of copies needed for their application (656100 in total) was itself a bottleneck. In general, the quantum device which outputs ρ may itself be an arbitrary quantum computer for which an individual execution may be expensive in terms of time, space, or money, and so it is desirable to run it as few times as possible.

Finally, let us note that our algorithms will actually use measurements which are more powerful than basis measurements, such as entangled measurements and POVMs. For details, see Section 1.2.1 below.

Example 1.2.3. Suppose our quantum system is described by the density matrix $\rho \in \mathbb{C}^{d\times d}$, and that we would like to learn this density matrix. (As we will mention below, this is referred to as the *quantum tomography* problem.) Suppose further that we knew in advance the eigenvectors $|v_1\rangle, \dots, |v_d\rangle$ of ρ. Then learning ρ reduces to learning the eigenvalues $\alpha_1, \dots, \alpha_d$.

To do this, we claim that it is optimal to perform a basis measurement on each copy of ρ using the basis $|v_1\rangle, \dots, |v_d\rangle$. This is because ρ can be viewed as representing the mixed state

$$|\boldsymbol{v}\rangle = \begin{cases} |v_1\rangle & \text{with probability } \alpha_1, \\ \dots & \\ |v_d\rangle & \text{with probability } \alpha_d, \end{cases}$$

and given $|\boldsymbol{v}\rangle$, one can determine with certainty which of the $|v_i\rangle$'s it is equal to by measuring in the eigenbasis. Having measured in this basis, the outcome is distributed as

$$\mathbf{Pr}[\text{observe outcome } i] = \langle v_i| \rho |v_i\rangle = \alpha_i.$$

Our goal is to learn the probability distribution $\alpha = (\alpha_1, \dots, \alpha_d)$, and each measurement produces an outcome $\boldsymbol{i} \in [d]$ sampled according to α. This is exactly the classical problem of learning an unknown distribution from independent samples, and it is known that $n = \Theta(d/\epsilon^2)$ samples are necessary and sufficient to learn a distribution on d elements. Hence, this same bound holds for the number of measurements and copies of ρ needed to learn α. This example is one instance of the connection between quantumly learning quantum states and classically learning probability distributions, which we explore more in Section 1.3. Note that in general we do *not* know the eigenbasis of ρ, and this is where much of the challenge in quantum state learning arises.

1.2.1 Quantum measurements

In this section, we will introduce the background on quantum computing necessary for this thesis. For a more detailed introduction, see the textbook of Nielsen and Chuang [NC10].

Definition 1.2.4. We will consider the following three types of measurements.

- In a *basis measurement*, one provides an orthonormal basis $|v_1\rangle, \dots, |v_d\rangle \in \mathbb{C}^d$ corresponding to the measurement outcomes. When the measurement is performed on a pure state $|\psi\rangle \in \mathbb{C}^d$, one receives the outcome $i \in [d]$ with probability $|\langle v_i|\psi\rangle|^2$, in which case $|\psi\rangle$ "collapses" to the state $|v_i\rangle$. If instead the measurement is performed on a mixed state $\rho \in \mathbb{C}^{d\times d}$, then outcome i is observed with probability $\langle v_i| \rho |v_i\rangle$, in which case ρ collapses to $|v_i\rangle\langle v_i|$.

- In a *projective measurement*, one provides a set of projection matrices $\Pi_1, \ldots, \Pi_m \in \mathbb{C}^{d\times d}$ satisfying the "completeness condition" $P_1 + \ldots + P_m = I$. If the measurement is performed on a pure state $|\psi\rangle$, then outcome $i \in [m]$ is observed with probability $\langle\psi| \Pi_i |\psi\rangle$, in which case $|\psi\rangle$ collapses to

  $$\frac{\Pi_i |\psi\rangle}{|\Pi_i |\psi\rangle |}.$$

  If the measurement is performed on a mixed state $\rho \in \mathbb{C}^{d\times d}$, then outcome $i \in [m]$ is observed with probability $\mathrm{tr}(\Pi_i \rho)$, in which case ρ collapses to

  $$\frac{\Pi_i \rho \Pi_i}{\mathrm{tr}(\Pi_i \rho)}.$$

- In a *Positive-Operator Valued Measure* (henceforth, always a *POVM*), one provides a set of PSD matrices $E_1, \ldots, E_m$ such that $E_1 + \ldots + E_m = I$. If the measurement is performed on a pure state $|\psi\rangle$, outcome $i \in [m]$ is observed with probability $\mathrm{tr}(E_i |\psi\rangle \langle\psi|)$. If the measurement is performed on a mixed state ρ, then outcome $i \in [m]$ is observed with probability $\mathrm{tr}(E_i \rho)$. The states that $|\psi\rangle$ and ρ collapse to are undefined.

  We also allow for POVMs with an infinite outcome set. In this case, we will specify a set Ω with σ-algebra Σ and a measure $d\omega$ on this set. Each $\omega \in \Omega$ has a corresponding measurement outcome E_ω. The measurement maps a (Borel) subset $B \subseteq \Omega$ to

  $$M(B) := \int_B E_\omega d\omega.$$

  The completeness condition is given by $M(\Omega) = I$, and the probability that an outcome falls inside the subset B is given by either $\mathrm{tr}(M(B) |\psi\rangle \langle\psi|)$ or $\mathrm{tr}(M(B)\rho)$. (In this thesis, we will only consider the case when $\Omega = \mathrm{U}(d)$, the set of unitary matrices, and $d\omega$ is the Haar measure on $\mathrm{U}(d)$.)

We note that (ignoring the fact that POVMs don't define a post-measurement state) each measurement generalizes the previous one: a basis measurement is a projective measurement with projectors $\Pi_i = |v_i\rangle \langle v_i|$, and a projective measurement is a POVM in which $E_i = \Pi_i$. Furthermore, the rules for measuring mixed states follow from the rules for pure states using the interpretation of a mixed state as a probability distribution over pure states.

Definition 1.2.5. If we have n unentangled quantum subsystems described by the states $|v_1\rangle \in \mathbb{C}^{d_1}, \ldots, |v_n\rangle \in \mathbb{C}^{d_n}$, then the joint state of the whole system is described by the tensor product $|v_1\rangle \otimes \cdots \otimes |v_n\rangle$. Similarly, if the n subsystems are described by the mixed states $\rho_1 \in \mathbb{C}^{d_1 \times d_1}, \ldots, \rho_n \in \mathbb{C}^{d_n \times d_n}$, then the joint state is described by the tensor product $\rho_1 \otimes \cdots \otimes \rho_n$. In this thesis, we will commonly consider the case when we are given n identical and unentangled copies of an unknown state $|\psi\rangle \in \mathbb{C}^d$ or $\rho \in \mathbb{C}^{d\times d}$, and so the entire state of the n copies is given by either $|\psi\rangle^{\otimes n}$ or $\rho^{\otimes n}$, respectively.

Definition 1.2.6. We will consider three types of measurements, each of increasing complexity. Given n copies of ρ:

- a *nonadaptive measurement* fixes n measurements (of any type) in advance (by which we mean either the bases, projectors, or POVM elements are fixed), measures each state separately, and then collects the results and tries to infer some property of ρ.

- an *adaptive measurement* measures each copy of ρ one-by-one and is allowed to pick each measurement based on the outcomes of the previous experiments.

- an *entangled measurement* performs any of the three types of measurements on the state $\rho^{\otimes n}$.

It can be shown that entangled measurements generalize adaptive measurements, which in turn generalize nonadaptive measurements. Each generalization increases the complexity of implementing the measurement, and only very simple nonadaptive measurements (e.g. projective measurements consisting of two projectors) can be practically implemented using current-day techniques.

The measurement formalism in quantum mechanics gives one a significant amount of freedom when designing measurements, but this freedom comes at a cost: it is often difficult to determine what the best measurement for a given task is. As we will show in the following proposition, the situation is simplified greatly when ρ is known to be block diagonal. (This generalizes the case considered in Example 1.2.3.)

Proposition 1.2.7. *Suppose ρ is known to be block-diagonal, where the blocks correspond to the known orthogonal projectors $\Pi_1, \ldots, \Pi_m$. Then the following two statements hold.*

1. *Prior to any other measurements, one may without loss of generality first perform a projective measurement on ρ using the Π_i's.*

2. *If, further, within each block, ρ is known to be a multiple of the identity matrix, then this projective measurement is the optimal measurement.*

Proof. Write $|v_1\rangle, \ldots, |v_d\rangle$ and $\alpha_1, \ldots, \alpha_d$ for ρ's eigenvectors and corresponding eigenvalues. Because ρ is block diagonal, each $|v_i\rangle$ falls within a subspace corresponding to one of the projectors Π_i. As we can view ρ as the mixed state "output $|v_i\rangle$ with probability α_i", we may suppose that the system is in the pure state $|v_i\rangle$ for some $i \in [d]$. By the measurement rule for projectors, if $|v_i\rangle$ falls in the subspace corresponding to Π_j, then the projective measurement $\Pi_1, \ldots, \Pi_m$ will always produce the outcome j, and $|v_i\rangle$ will remain unchanged (i.e., it will collapse to itself). Hence, the measurement does not perturb the system, and so we may perform it without loss of generality, proving item 1.

After the projective measurement is made and some outcome j is observed, then ρ collapses to the maximally mixed state on the subspace corresponding to Π_j. At this point, we know that the state ρ collapsed to the maximally mixed state on this subspace, and so nothing can be gained information theoretically from further measurements. This proves item 2. □

1.3 Classical distribution learning and testing

Before moving on to our quantum learning and testing problems, let us first consider the classical special case (from Example 1.2.3) of learning and testing probability distributions. In the standard model of distribution learning, there is an unknown probability distribution α, and the tester is allowed to draw n independent samples from this distribution. We will often state this in terms of random words.

Definition 1.3.1. Let $\mathcal{A}$ be an *alphabet*; i.e., a totally ordered set. Most often we consider $\mathcal{A} = [d]$. A *word* is a finite sequence $(a_1, \dots, a_n)$ of elements from $\mathcal{A}$. We say that $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_n)$ is an *n-letter α-random word* if each letter $\boldsymbol{w}_i$ is independently drawn from the set $\mathcal{A}$ according to the distribution α. We may sometimes also write $\boldsymbol{w} \sim \alpha^{\otimes n}$.

In this thesis, we will reserve $\mathcal{A}$ for alphabets and α for distributions on alphabets. One may of course want to learn distributions on finite sets Ω which are not totally ordered (i.e. are not alphabets). However, the alphabet case is without loss of generality in this setting and will be crucial for our work on longest increasing subsequences.

We will now define some distance measures between probability distributions, and for this it is convenient to consider distributions $\mathcal{D}$ on general sets Ω. The most basic way of measuring the distance between two probability distributions is given by the total variation distance.

Definition 1.3.2. Given a real number $p \geq 1$ and a vector x on a finite set Ω, the ℓ_p norm of x, written as $\|x\|_p$, is defined as

$$\|x\|_p^p := \sum_{\omega \in \Omega} |x_\omega|^p.$$

Given two discrete probability distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on a finite set Ω, the *total variation distance* between them is $d_{\mathrm{TV}}(\alpha, \beta) := \frac{1}{2}\|\mathcal{D}_1 - \mathcal{D}_2\|_1$.

Suppose a random element $\boldsymbol{\omega}$ in Ω was drawn from either $\mathcal{D}_1$ or $\mathcal{D}_2$, and one would like to know which of the two it came from. It is natural to select a subset $S \subseteq \Omega$, guess "$\mathcal{D}_1$" if $\boldsymbol{\omega} \in S$, and guess "$\mathcal{D}_2$" if $\boldsymbol{\omega} \notin S$. The following easy-to-prove statement relates how well the best such strategy works with the total variation distance:

$$d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \max_{S \subseteq [d]} \left\{ \mathop{\mathbf{Pr}}_{\boldsymbol{\omega} \sim \mathcal{D}_1} [\boldsymbol{\omega} \in S] - \mathop{\mathbf{Pr}}_{\boldsymbol{\omega} \sim \mathcal{D}_2} [\boldsymbol{\omega} \in S] \right\}. \tag{1.3}$$

We will also require some nonsymmetric "distances" between probability distributions.

Definition 1.3.3. The *chi-squared distance* is

$$d_{\chi^2}(\mathcal{D}_1, \mathcal{D}_2) := \mathop{\mathbf{E}}_{\boldsymbol{\omega} \sim \mathcal{D}_2} \left[\left(\frac{\mathcal{D}_1(\boldsymbol{\omega})}{\mathcal{D}_2(\boldsymbol{\omega})} - 1 \right)^2 \right].$$

Further, if $\mathrm{supp}(\mathcal{D}_1) \subseteq \mathrm{supp}(\mathcal{D}_2)$, then the *Kullback–Leibler divergence* is

$$d_{\mathrm{KL}}(\mathcal{D}_1, \mathcal{D}_2) := \mathop{\mathbf{E}}_{\boldsymbol{\omega} \sim \mathcal{D}_1} \left[\ln \left(\frac{\mathcal{D}_1(\boldsymbol{\omega})}{\mathcal{D}_2(\boldsymbol{\omega})} \right) \right].$$

To relate these quantities, Cauchy–Schwarz implies that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{1}{2}\sqrt{d_{\chi^2}(\mathcal{D}_1, \mathcal{D}_2)}$, and Pinsker's inequality states that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{1}{\sqrt{2}}\sqrt{d_{\mathrm{KL}}(\mathcal{D}_1, \mathcal{D}_2)}$.

1.3.1 Distribution learning

In distribution learning, we are given an n-letter α-random word $\boldsymbol{w}$, and we would like to learn a feature of α. The most basic problem in this area is to learn the entire distribution α, and a good estimate turns out to be given by the empirical distribution.

Definition 1.3.4. Given an n-letter α-random word $\boldsymbol{w}$, the *empirical distribution* is the probability distribution $\hat{\boldsymbol{\alpha}}$ in which $\hat{\boldsymbol{\alpha}}_i$ is the number of i's in $\boldsymbol{w}$ divided by n.

The most basic fact about the empirical distribution is that by taking $n = \Theta(d/\epsilon^2)$, it is ϵ-close to α in total variation distance with high probability [DL01, pages 10 and 31]. The simplest proof of this, from [Dia14, Slide 6], begins by proving convergence of the empirical distribution in ℓ_2^2 *distance.*

Proposition 1.3.5. *Given $\boldsymbol{w} \sim \alpha^{\otimes n}$, let $\hat{\boldsymbol{\alpha}}$ be the empirical distribution. Then*

$$\mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2 \leq \frac{1}{n}.$$

To relate this to the total variation distance, note that

$$\mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_1 \leq \sqrt{d} \cdot \mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_2 \leq \sqrt{d} \cdot \sqrt{\mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2},$$

where the first step is Cauchy-Schwarz and the second is concavity of the square root. This gives us the following corollary.

Corollary 1.3.6. $\mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_1 \leq \sqrt{d/n}$. *Hence, $\hat{\boldsymbol{\alpha}}$ is ϵ-close to α in total variation distance when $n = O(d/\epsilon^2)$ with high probability.*

(The high probability bound follows from the expectation bound by increasing n and applying Markov's inequality.) This strategy of proving ℓ_1 distance bounds by first switching to ℓ_2^2 distance will prove fruitful later with our quantum state learning results.

Proof of Proposition 1.3.5. Each coordinate of the empirical distribution $\hat{\boldsymbol{\alpha}}_i$ is distributed as $\text{Binomial}(n, \alpha_i)/n$ and hence has mean α_i and variance $\alpha_i(1-\alpha_i)/n$. Then

$$\mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2 = \sum_{i=1}^{d} \mathbf{E}(\hat{\boldsymbol{\alpha}}_i - \alpha_i)^2 = \sum_{i=1}^{d} \mathbf{Var}[\hat{\boldsymbol{\alpha}}_i] = \sum_{i=1}^{d} \frac{\alpha_i(1-\alpha_i)}{n} \leq \frac{1}{n}. \qquad \square$$

A related problem, especially interesting in the case when α has only a few large entries, is to estimate the values of the k largest α_i's. In the $k = 1$ case, for example, this is the problem of estimating α's ℓ_∞ norm. A natural algorithm is to output the k largest entries in the empirical distribution.

Notation 1.3.7. Given $x \in \mathbb{R}^d$, the notation $x_{[i]}$ means the i-th largest value among $x_1, \dots, x_d$.

In other words, given $\boldsymbol{w}$, then we output $(\hat{\boldsymbol{\alpha}}_{[1]}, \dots, \hat{\boldsymbol{\alpha}}_{[k]})$. As this algorithm is agnostic to the order of α, we may assume that α is sorted in decreasing order. In this case, we have the following bound.

Proposition 1.3.8. *Suppose α is a sorted probability distribution. Then for any $k \in [d]$,*

$$\mathbf{E} \sum_{i=1}^{k} |\hat{\boldsymbol{\alpha}}_{[i]} - \alpha_i| \leq \sqrt{\frac{k}{n}}.$$

Thus, we can estimate the k largest α_i's when $n = O(k/\epsilon^2)$ with high probability.

To show this, we first need the following fact.

Fact 1.3.9. *Let $x, y \in \mathbb{R}^d$ be sorted. Then for any permutation $\pi \in \mathfrak{S}(d)$,*

$$\sum_{i=1}^{d} |x_i - y_i| \leq \sum_{i=1}^{d} |x_i - y_{\pi(i)}|.$$

Proof. If π is not the identity permutation, then there are adjacent coordinates $i, i+1$ for which $\pi(i) > \pi(i+1)$. Suppose σ is the permutation formed by transposing these two positions. Then we claim that

$$\sum_{i=1}^{d} |x_i - y_{\sigma(i)}| \leq \sum_{i=1}^{d} |x_i - y_{\pi(i)}|. \tag{1.4}$$

Showing this will prove the fact, as we can repeatedly do this to σ, eventually arriving at the identity permutation.

The only indices where the left-hand and the right-hand sides of Equation (1.4) differ are i and $i+1$. Zeroing in on these reduces to showing the following fact about four real numbers a, b, c, d satisfying $a \geq b$, $c \geq d$:

$$|a - c| + |b - d| \leq |a - d| + |b - c|.$$

This is easily verified using case analysis. □

Proof of Proposition 1.3.8. For $i \in [k]$, write $\boldsymbol{\ell}_i$ for the index of the i-th largest coordinate of $\hat{\boldsymbol{\alpha}}$. Then $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}_i} = \hat{\boldsymbol{\alpha}}_{[i]}$ for all $i \in [k]$. Consider $\boldsymbol{\ell}'_1, \ldots, \boldsymbol{\ell}'_k$, the rearrangement of $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_k$ formed by (i) first setting $\boldsymbol{\ell}'_i = i$ if $i \in \{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_k\}$, for each $i \in [k]$, and then (ii) setting the remaining $(\boldsymbol{\ell}'_i)$'s to be the remaining $\boldsymbol{\ell}_i$'s in any order. By fact 1.3.9,

$$\mathbf{E} \sum_{i=1}^{k} |\hat{\boldsymbol{\alpha}}_{[i]} - \alpha_i| = \mathbf{E} \sum_{i=1}^{k} |\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}_i} - \alpha_i| \leq \mathbf{E} \sum_{i=1}^{k} |\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i} - \alpha_i|. \tag{1.5}$$

Next, consider $\boldsymbol{\ell}^+_1, \ldots, \boldsymbol{\ell}^+_k$, in which $\boldsymbol{\ell}^+_i = \boldsymbol{\ell}'_i$ if $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i} \geq \alpha_i$ and $\boldsymbol{\ell}^+_i = i$ otherwise. Then for each $i \in [k]$, $|\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i} - \alpha_i| \leq |\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}^+_i} - \alpha_{\boldsymbol{\ell}^+_i}|$ because either (i) $\boldsymbol{\ell}'_i = i$ already, in which case the two quantities are the same, or otherwise (ii) $\boldsymbol{\ell}'_i \neq i$. In this case, (i) if $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i} \geq \alpha_i$, then the inequality follows from the fact that $\alpha_{\boldsymbol{\ell}'_i} \leq \alpha_i$, because $\boldsymbol{\ell}'_i \neq i$ and so $\alpha_{\boldsymbol{\ell}'_i}$ is not one of the k

largest α_j's, and (ii) if $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i} < \alpha_i$, then the inequality follows from the fact that $\hat{\boldsymbol{\alpha}}_i \leq \hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}'_i}$, because $\boldsymbol{\ell}'_i \neq i$ and so $\hat{\boldsymbol{\alpha}}_i$ is not one of the k largest $\hat{\boldsymbol{\alpha}}_j$'s. As a result,

$$(1.5) \leq \mathbf{E} \sum_{i=1}^{k} |\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}_i^+} - \alpha_{\boldsymbol{\ell}_i^+}| \leq \mathbf{E} \sqrt{k \cdot \sum_{i=1}^{k} (\hat{\boldsymbol{\alpha}}_{\boldsymbol{\ell}_i^+} - \alpha_{\boldsymbol{\ell}_i^+})^2}$$
$$\leq \mathbf{E} \sqrt{k \cdot \|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2} \leq \sqrt{k \cdot \mathbf{E}\, \|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2} \leq \sqrt{\frac{k}{n}},$$

where the inequalities follow from (in order): (i) the definition of the ℓ_i^+ indices, (ii) Cauchy-Schwarz, (iii) the fact that each index $j \in [d]$ appears at most once among $\{\boldsymbol{\ell}_i^+\}_{i \in [k]}$, (iv) Jensen's inequality, and (v) Proposition 1.3.5. □

There are various other natural properties of α one can learn, and the algorithms and lower bounds for these often involve a substantial amount of cleverness. Famous examples involve estimating the entropy of α up to ϵ-additive error, which can be done with $n = \Theta(\frac{d}{\log(d) \cdot \epsilon})$ samples [VV11a, VV11b], and estimating the support size of α up to ϵn-additive error, for which we know an upper bound of $n = O(\frac{d}{\log(d) \cdot \epsilon^2})$ samples[1] and a nearly matching lower bound of $n = \Omega(\frac{d}{\log d})$ samples for any constant $\epsilon > 0$ [VV11a].

1.3.2 Distribution testing

A related stream of research has dealt with the problem of testing properties of α. This stream operates in the *property testing* model of Rubinfeld and Sudan [RS92, RS96], which was originally introduced in the context of testing algebraic properties of polynomials over finite fields but has since found applications in a wide variety of areas, including testing properties of graphs and of Boolean functions. In the case of testing properties of probability distributions, one is given a sample $\boldsymbol{w} \sim \alpha^{\otimes n}$ with the goal of determining whether α has some property $\mathcal{P}$ or is ϵ-far from $\mathcal{P}$ in total variation distance, meaning that it is ϵ-far from *every* distribution with property $\mathcal{P}$. Formally, property testing is defined as follows.

Definition 1.3.10. In the model of property testing, there is a set of objects $\mathcal{O}$ along with a distance measure dist : $\mathcal{O} \times \mathcal{O} \to \mathbb{R}$. A property $\mathcal{P}$ is a subset of $\mathcal{O}$, and for an object $o \in \mathcal{O}$, we define the distance of o to $\mathcal{P}$ to be[2]

$$\mathrm{dist}(o, \mathcal{P}) \coloneqq \min_{o' \in \mathcal{P}} \{\mathrm{dist}(o, o')\}.$$

If $\mathrm{dist}(o, \mathcal{P}) \geq \epsilon$, then we say that o is ϵ-far from $\mathcal{P}$. A testing algorithm $\mathcal{T}$ tests $\mathcal{P}$ if, given some sort of "access" to $o \in \mathcal{O}$ (e.g., independent samples or queries), $\mathcal{T}$ accepts with high probability (say, probability at least 2/3) if $o \in \mathcal{P}$ and rejects with high probability if o is ϵ-far from $\mathcal{P}$. Generally, the aim is for $\mathcal{T}$ to be efficient according some measure, most typically the number of accesses made to o. (On the other hand, $\mathcal{T}$ is generally allowed unlimited computational power. Nevertheless, as we will see, all of the testers considered in this thesis can be implemented efficiently.)

[1] Here, they also need the additional assumption that any nonzero probability value is at least $1/d$.

[2] Formally, our sets $\mathcal{O}$ will always lie within some $\mathbb{R}^N$ or $\mathbb{C}^N$, and we always require that $\mathcal{P}$ be a *closed* set. Thus the "min" here is well-defined.

Definition 1.3.11. We will instantiate property testing in the following setting:

- **Properties of probability distributions:** $\mathcal{O}$ is the set of probability distributions α on $[d]$, the tester gets i.i.d. draws from α, and $\text{dist} = d_{\text{TV}}$.

In this model, it is possible to test *any* property with $n = O(d/\epsilon^2)$ samples by $\epsilon/2$-estimating α with the empirical distribution $\hat{\boldsymbol{\alpha}}$ (by Corollary 1.3.6) and checking whether $\hat{\boldsymbol{\alpha}}$ is $\epsilon/2$-close to any distribution with property $\mathcal{P}$. As a result, the goal of this area is to find testers for properties which use a *sublinear* (in d) number of samples. That such algorithms could exist is suggested by the following Birthday Paradox-based fact from [GR11, BFR$^+$00] (cf. [Bat01, Theorem 3.24]):

Fact 1.3.12. $\Theta(\sqrt{r})$ *samples are necessary and sufficient to distinguish between the cases when the distribution is uniform on either* r *or* $2r$ *values. (The bound also holds for* r *vs.* r' *when* $r' > 2r$*.)*

Setting $r = \frac{d}{2}$, we see that this fact gives a sublinear algorithm for distinguishing between the uniform distribution and a distribution that is uniform on exactly half of the elements of $\{1, \ldots, d\}$. This fact is also important as it immediately gives a lower bound of $\Omega(\sqrt{d})$ for testing a variety of natural problems, those for which Fact 1.3.12 appears as a special case.

Perhaps the most basic property of probability distributions one can test for is the property of being equal to the uniform distribution.

Definition 1.3.13. We will write Unif_d for the uniform probability distribution $(\frac{1}{d}, \ldots, \frac{1}{d})$.

An $\Omega(\sqrt{d})$ lower bound follows directly from Fact 1.3.12. On the other hand, an $O(\sqrt{d}/\epsilon^4)$ upper bound was shown in the early work of [BFR$^+$00, BFR$^+$13] using techniques of [GR11]. The correct sample complexity was finally pinned down by Paninski in [Pan08], who showed matching upper and lower bounds:

Theorem 1.3.14 ([Pan08])**.** $\Theta(\sqrt{d}/\epsilon^2)$ *samples are necessary and sufficient to test whether* α *is the uniform distribution* Unif_d*.*

This result was recently extended [VV14] to an $O(\sqrt{d}/\epsilon^2)$ upper bound for testing equality to *any* fixed distribution, improving on the previously known [BFF$^+$01] upper bound of $\widetilde{O}(\sqrt{d}/\epsilon^4)$. More precisely, [VV14] upper-bounds the sample complexity of testing equality to a fixed distribution β by $O(f(\beta)/\epsilon^2)$, where $f(\beta)$ is a certain norm which is maximized when β is the uniform distribution. Thus the uniform distribution is the hardest fixed distribution to test equality to.

The property of being the uniform distribution falls within the class of *symmetric* properties of probability distributions.

Definition 1.3.15. We will instantiate property testing in the following setting:

- **Symmetric properties of probability distributions:** As in Definition 1.3.11 above, but $\mathcal{P}$ is any symmetric property of probability distributions, meaning that if $\alpha \in \mathcal{P}$, then $\alpha_\pi = (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(d)}) \in \mathcal{P}$ for any permutation $\pi \in \mathfrak{S}(d)$.

Here we are using the following definition.

Definition 1.3.16. Given an integer d, $\mathfrak{S}(d)$ refers to the symmetric group on d elements.

Note that membership in a symmetric property $\mathcal{P}$ depends only on the multiset $\{\alpha_1, \dots, \alpha_d\}$ and not on the ordering of the α_i's. Other interesting symmetric properties beyond uniformity include having small entropy or small support size. Testing for small support size does not appear to have been precisely addressed in the literature; however the following is easy to derive from known results (in particular, the lower bound follows from the work of [VV11a]):

Theorem 1.3.17. *To test (with ϵ a constant) whether a probability distribution has support size r, $O(r)$ samples are sufficient and $\Omega(r/\log(r))$ samples are necessary.*

Property testing of probability dstributions is a large field beyond the scope of this thesis; see [Can16] for a comprehensive survey.

1.4 Quantum problems and our results

Let us begin by defining some standard distance measures between quantum states.

Definition 1.4.1. If $M \in \mathbb{C}^{d\times d}$ is any Hermitian matrix with eigenvalues $\mu_1, \dots, \mu_d$, the ℓ_1 or *trace norm* of M is

$$\|M\|_1 \coloneqq \operatorname{tr}\left(\sqrt{M^\dagger M}\right) = \sum_{i=1}^{d} |\mu_i|.$$

Similarly, the ℓ_2 or *Frobenius norm* of M, written as $\|M\|_F$, is defined as

$$\|M\|_F^2 \coloneqq \operatorname{tr}(M^\dagger M) = \sum_{i=1}^{d} \mu_i^2.$$

We note that $\|M\|_1 \leq \sqrt{d} \cdot \|M\|_F$ by Cauchy-Schwarz applied to the eigenvalues of M. Given two density matrices ρ and σ, the *trace distance* between them is

$$d_{\mathrm{tr}}(\rho, \sigma) \coloneqq \frac{1}{2}\|\rho - \sigma\|_1.$$

The trace distance is the standard generalization of the total variation distance to mixed states; for example, it satisfies the following generalization of Equation (1.3) [NC10, equation (9.22)]:

$$d_{\mathrm{TV}}(\rho_1, \rho_2) = \max_{\text{projectors } \Pi} \left\{\operatorname{tr}(\Pi\rho_1) - \operatorname{tr}(\Pi\rho_2)\right\}.$$

This statement relates the trace distance to the maximum probability with which two mixed states can be distinguished by a projective measurement. This property makes it the natural choice of distance for property testing of quantum states. We also have the following simple fact:

Fact 1.4.2. *Suppose ρ and σ are diagonal density matrices with diagonal entries $\alpha = (\alpha_1, \dots, \alpha_d)$ and $\beta = (\beta_1, \dots, \beta_d)$, respectively. Then $d_{\mathrm{tr}}(\rho, \sigma) = d_{\mathrm{TV}}(\alpha, \beta)$.*

1.4.1 Quantum state learning

The most basic type of problem we consider is that of computing an estimate $\hat{\rho}$ of the quantum state ρ.

Definition 1.4.3. In *quantum tomography*, one is given n copies of a density matrix $\rho \in \mathbb{C}^{d\times d}$ with sorted spectrum α, and the goal is to output a density matrix $\hat{\rho}$ such that $d_{\mathrm{tr}}(\rho, \hat{\rho}) \leq \epsilon$. In *quantum PCA*, there is an additional parameter $1 \leq k \leq d$, and the goal is to output a rank-k matrix $\hat{\rho}$ which is PSD and has $\mathrm{tr}(\hat{\rho}) \leq 1$ such that

$$\|\rho - \hat{\rho}\|_1 \leq \alpha_{k+1} + \ldots + \alpha_d + \epsilon.$$

We note that $\alpha_{k+1} + \ldots + \alpha_d$ is the error of the best rank-k approximator to ρ.

Related to this is the problem of learning ρ's spectrum.

Definition 1.4.4. In *quantum spectrum estimation*, one is given n copies of a density matrix $\rho \in \mathbb{C}^{d\times d}$ with sorted spectrum α, and the goal is to output a sorted spectrum $\hat{\alpha}$ such that $d_{\mathrm{TV}}(\alpha, \hat{\alpha}) \leq \epsilon$. In *truncated spectrum estimation*, there is an additional parameter $1 \leq k \leq d$, and the goal is that $d_{\mathrm{TV}}^{(k)}(\alpha, \hat{\alpha}) \leq \epsilon$, where $d_{\mathrm{TV}}^{(k)}(\alpha, \beta)$ denotes $\frac{1}{2}\sum_{i=1}^{k} |\alpha_i - \beta_i|$.

Quantum PCA and truncated spectrum estimation correspond to the naturally occuring case when ρ is either pure or low rank but has been subjected to a small amount of noise. This case has been studied previously in, for example, [FGLE12]. Intuitively, tomography is a harder problem than spectrum estimation, as the former requires learning both ρ's eigenvalues and eigenvectors while the latter requires learning only ρ's eigenvalues. Indeed, this relationship can be made quantitative: if $\hat{\rho}$ is an ϵ-approximation to ρ, then $\hat{\rho}$'s spectrum $\hat{\alpha}$ is an ϵ-approximation to α, per the following fact.

Fact 1.4.5. *Suppose the density matrices $\rho_1, \rho_2 \in \mathbb{C}^{d\times d}$ have sorted spectrum $\alpha_1, \alpha_2 \in \mathbb{C}^d$. Then $d_{\mathrm{TV}}(\alpha_1, \alpha_2) \leq d_{\mathrm{tr}}(\rho_1, \rho_2)$.*

Proof. We learned the proof of this fact from Ashley Montanaro [Mon14]. Since $\|\cdot\|_1$ is a unitarily invariant norm, a theorem of Mirsky (see [HJ13, Corollary 7.4.9.3]) states that

$$\|\rho_1 - \rho_2\|_1 \geq \|\rho_1' - \rho_2'\|_1, \tag{1.6}$$

where ρ_1' (respectively, ρ_2') denotes the diagonal density matrix whose entries are the eigenvalues of ρ_1 (respectively, ρ_2) arranged in nonincreasing order. We have $\rho_1' = \mathrm{diag}(\alpha_1)$, and $\rho_2' = \mathrm{diag}(\alpha_2)$. But the left-hand side of (1.6) is $2d_{\mathrm{tr}}(\rho_1, \rho_2)$, and the right-hand side is $2d_{\mathrm{tr}}(\rho_1', \rho_2')$, which in turn equals $2d_{\mathrm{TV}}(\alpha_1, \alpha_2)$ (by Fact 1.4.2). Thus $d_{\mathrm{TV}}(\alpha_1, \alpha_2) \leq d_{\mathrm{tr}}(\rho_1, \rho_2)$, as needed. □

We note that Fact 1.3.9 corresponds to the special case of Fact 1.4.5 when ρ_1 and ρ_2 are diagonal matrices.

As it is the simpler of the two problems, let us begin by discussing our results for spectrum estimation. In this thesis, we consider a particular spectrum estimation algorithm called the *empirical Young diagram (EYD)* algorithm. The EYD algorithm was originally introduced in [ARS88, KW01] and is the most popular and powerful spectrum estimation algorithm in the literature. In fact, is has been suggested that this algorithm can be implemented using current-day experimental techniques [BAH+16]. Given $\rho^{\otimes n}$, it outputs a random estimate $\hat{\boldsymbol{\alpha}}$ of α which can be viewed as a quantum analogue of the empirical distribution. The work of [HM02, CM06] showed that $\hat{\boldsymbol{\alpha}}$ is ϵ-close in total variation distance to α with high probability when $n = O(d^2/\epsilon^2 \cdot \log(d/\epsilon))$, and prior to our work this was the best known bound for spectrum estimation. We improve this bound to $n = O(d^2/\epsilon^2)$. As in the proof of Corollary 1.3.6, we begin by showing that $\hat{\boldsymbol{\alpha}}$ approximates α well in ℓ_2^2 distance.

Theorem 1.4.6. *Given n copies of a mixed state ρ with spectrum α, let $\hat{\boldsymbol{\alpha}}$ be the random output of the EYD algorithm. Then*

$$\mathbf{E}\,\|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2 \leq \frac{d}{n}.$$

Our spectrum estimation bound then follows as an immediate corollary.

Corollary 1.4.7. *Given n copies of a mixed state ρ with spectrum α, let $\hat{\boldsymbol{\alpha}}$ be the random output of the EYD algorithm. Then*

$$\mathbf{E}\,\|\hat{\boldsymbol{\alpha}} - \alpha\|_1 \leq \frac{d}{\sqrt{n}}.$$

As a result, $n = O(d/\epsilon)$ copies suffice to obtain an ϵ-accurate estimate in ℓ_2^2 distance, and $n = O(d^2/\epsilon^2)$ copies suffice to obtain an ϵ-accurate estimate in total variation distance. (These bounds are with high probability; confidence $1-\delta$ may be obtained by increasing the copies by a factor of $\log(1/\delta)$.)

Proof. By Cauchy-Schwarz and then Jensen's inequality,

$$\mathbf{E}\,\|\hat{\boldsymbol{\alpha}} - \alpha\|_1 \leq \sqrt{d}\,\mathbf{E}\,\|\hat{\boldsymbol{\alpha}} - \alpha\|_2 \leq \sqrt{d}\sqrt{\mathbf{E}\,\|\hat{\boldsymbol{\alpha}} - \alpha\|_2^2} \leq \frac{d}{\sqrt{n}}. \qquad \square$$

As we will see later, the behavior of the EYD algorithm depends only on the rank of ρ and not on the ambient dimension d. Hence, if ρ is rank k, then only $O(k^2/\epsilon^2)$ copies of ρ are needed to estimate α in total variation distance. Our next result, which generalizes Corollary 1.4.7, shows that $O(k^2/\epsilon^2)$ copies are sufficient even if ρ is only *approximately* low rank.

Theorem 1.4.8. *Given n copies of a mixed state ρ with spectrum α, let $\hat{\boldsymbol{\alpha}}$ be the random output of the EYD algorithm. Then*

$$\mathbf{E}\,d_{\mathrm{TV}}^{(k)}(\hat{\boldsymbol{\alpha}}, \alpha) \leq \frac{1.92\,k + .5}{\sqrt{n}}.$$

Thus, truncated spectrum estimation can be solved with $n = O(k^2/\epsilon^2)$ copies.

To our knowledge, nothing was previously known about truncated spectrum estimation.

In general, a lower bound of $n = \Omega(d/\epsilon^2)$ copies for spectrum estimation follows from our Theorem 1.4.23 below. In the case of the EYD algorithm, we can improve this lower bound to $\Omega(d^2/\epsilon^2)$, matching the upper bound from Corollary 1.4.7. To do this, we show that the EYD algorithm requires $\Omega(d^2/\epsilon^2)$ copies when trying to estimate the spectrum of a particular mixed state known as the *maximally mixed state*.

Definition 1.4.9. The d-dimensional *maximally mixed state* is the mixed state defined as

$$\frac{I}{d} = \begin{pmatrix} \frac{1}{d} & 0 & \dots & 0 \\ 0 & \frac{1}{d} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{d} \end{pmatrix}.$$

Theorem 1.4.10. *If $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state, then the EYD algorithm* fails *to give an ϵ-accurate estimate in total variation distance with high probability unless $\Omega(d^2/\epsilon^2)$ copies are used.*

Thus, to improve on Corollary 1.4.7, one has to consider a new algorithm.

Next, we extend these results to quantum tomography. For an extremely long time, the best known tomography algorithm was the "textbook" algorithm [NC10, Section 8.4.2], which performed nonadaptive "Pauli" measurements and used $n = O(d^4/\epsilon^2)$ copies [FGLE12, Footnote 2]. The textbook algorithm is particularly easy to implement and is still referred to by practicioners. Only recently in 2014 was this upper bound improved to $n = O(d^3/\epsilon^2)$ by [KRT14], who proposed a new algorithm which performs nonadaptive "random basis measurements". In Section 5.1 below, we give a simpler proof of this bound. As announced by Jeongwan Haah at QIP 2016 [Haa16], he and his coauthors [HHJ$^+$16] have shown a matching lower bound of $\Omega(d^3/\epsilon^2)$ for all algorithms using nonadaptive measurements. Thus, improving on this requires studying algorithms which perform more powerful measurements.

In this work, we consider *two* tomography algorithms: one being the algorithm of Michael Keyl [Key06], and the other being the first of the two *Pretty Good Measurement (PGM)*-inspired tomography algorithms from [HHJ$^+$16]. Keyl's algorithm and the first PGM tomography algorithm share some high level features: both perform highly entangled measurements, both run the EYD algorithm as a subroutine, and it turns out that we can analyze both with substantially overlapping proofs. In doing so, we improve the upper bound of [KRT14], showing that not only can we learn the spectrum using $O(d^2/\epsilon^2)$ copies, we can learn the *entire* state with as many copies. The outline of our tomography results largely follow the outline of our spectrum estimation results. For example, we begin by showing that these algorithms well-approximate ρ in ℓ_2^2 distance.

Theorem 1.4.11. *Given n copies of a mixed state ρ, let $\hat{\boldsymbol{\rho}}$ be the random output of either Keyl's algorithm or the PGM tomography algorithm. Then*

$$\mathbf{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_F^2 \leq \frac{4d-3}{n}.$$

As in Corollary 1.4.7, the ℓ_1 tomography bound follows as an immediate corollary.

Corollary 1.4.12. *Given n copies of a mixed state ρ, let $\hat{\boldsymbol{\rho}}$ be the random output of either Keyl's algorithm or the PGM tomography algorithm. Then*

$$\mathbf{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_1 \leq \sqrt{\frac{4d^2-3d}{n}}. \tag{1.7}$$

As a result, $n = O(d/\epsilon)$ copies suffice to obtain an ϵ-accurate estimate in ℓ_2^2 distance, and $n = O(d^2/\epsilon^2)$ copies suffice to obtain an ϵ-accurate estimate in trace distance. (These bounds are with high probability; confidence $1-\delta$ may be obtained by increasing the copies by a factor of $\log(1/\delta)$.)

As we will see later, both tomography algorithms output a mixed state $\hat{\boldsymbol{\rho}}$ whose rank is at most the rank of ρ. Hence, if ρ is rank k, then the Cauchy-Schwarz step involved in deriving Equation (1.7) incurs a penalty of $\sqrt{2k}$ rather than $\sqrt{d}$, in which case only $n = O(kd/\epsilon^2)$ copies are needed to estimate ρ in total variation distance. If ρ is only *approximately* rank k, then there is a natural way to truncate the output of Keyl's algorithm so that it is always rank k. Our next result, which generalizes Corollary 1.4.12, shows that $O(kd/\epsilon^2)$ copies are sufficient for the truncated version of Keyl's algorithm to approximate ρ in this case.

Theorem 1.4.13. *Given n copies of a mixed state ρ, let $\hat{\boldsymbol{\rho}}$ be the rank k random output of the truncated version of Keyl's algorithm. Then*

$$\mathbf{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_1 \leq \alpha_{k+1} + \cdots + \alpha_d + 6\sqrt{\frac{kd}{n}}.$$

Thus, quantum PCA can be solved with $n = O(kd/\epsilon^2)$ copies.

Prior works have typically considered the case when ρ is exactly rank k, and for this the best previous bound was $n = O(k^2 d/\epsilon^2)$ [KRT14]. To our knowledge, the only prior work to get PCA-type bounds was [FGLE12], which showed how to find a rank-k matrix $\hat{\rho}$ such that

$$d_{\mathrm{tr}}(\rho, \hat{\rho}) \leq C \cdot (\alpha_{k+1} + \ldots + \alpha_d) + \epsilon,$$

for some absolute constant C, when $n = O(\left(\frac{kd}{\epsilon}\right)^2 \log(d))$. We are not aware of any previous work proving a PCA guarantee satisfying Definition 1.4.3.

The focus of the paper [OW16] was analyzing Keyl's algorithm. Independently, the work of [HHJ$^+$16] introduced two PGM-inspired quantum tomography algorithms. Following their paper, we observed that our tomography analysis could also be used to analyze the first of their algorithms [OW15a]. The main result of [HHJ$^+$16] is that if ρ is exactly rank k, then $n = O(kd/\epsilon^2 \cdot \log(d/\epsilon))$ copies are sufficient to approximate ρ in trace distance, and $n = O(kd/\epsilon \cdot \log(d/\epsilon))$ copies are sufficient to approximate ρ in "infidelity" (see [HHJ$^+$16] for this defined). In addition, they show two lower bounds for trace distance tomography: the first is a lower bound of $n = \Omega(d^2/\epsilon^2)$ for the general case, and the second is a lower bound of $n = \Omega(kd/(\epsilon^2 \log(d/\epsilon)))$ if ρ is rank k. Hence, our tomography bound is optimal, and our PCA bound is optimal up to logarithmic factors. We believe that our PCA bound is in fact optimal, though proving the tight lower bound of $n = \Omega(kd/\epsilon^2)$ for tomography of rank k matrices remains an open problem. We can, however, prove this bound without the dependence on ϵ.

Theorem 1.4.14. *Quantum tomography requires $\Omega(kd)$ copies in the case when ρ is rank k.*

1.4.2 Quantum state testing

The second set of problems we consider involve testing properties of a state ρ. Formally:

Definition 1.4.15. We will instantiate property testing in the following setting:

- **Properties of quantum states:** $\mathcal{O}$ is the set of $d \times d$ mixed states ρ, the tester gets unentangled copies of α, and dist $= d_{\mathrm{tr}}$.

An important subclass of properties to test are those that are *unitarily invariant.*

Definition 1.4.16. A property of mixed states $\mathcal{P}$ is *unitarily invariant* if $\rho \in \mathcal{P}$ implies that $U\rho U^\dagger \in \mathcal{P}$ as well, for every unitary U.

Note that whether a mixed state ρ satisfies a unitarily invariant property depends only on the multiset of its eigenvalues $\{\alpha_1, \ldots, \alpha_d\}$; equivalently, it depends only on its sorted spectrum α. Examples of unitarily invariant properties include having rank k and having von Neumann entropy lower than some specified threshold. The property testing model requires us to understand the quantity $d_{\mathrm{tr}}(\rho, \mathcal{P})$, where ρ is an arbitrary mixed state, and for unitarily invariant properties $\mathcal{P}$, which may contain a wide variety of states, this quantity seems difficult to understand. For example, if $\mathcal{P}$ is the set of rank k states, then it is not obvious how to compute $d_{\mathrm{tr}}(\rho, \mathcal{P})$. Ideally, we might hope that because $\mathcal{P}$ is unitarily invariant, then $d_{\mathrm{tr}}(\rho, \mathcal{P})$ is given by the total variation distance between α, ρ's sorted spectrum, and the closest sorted spectrum of any state in $\mathcal{P}$. If this were so, then we would have

$$d_{\mathrm{tr}}(\rho, \mathcal{P}) \stackrel{?}{=} \alpha_{k+1} + \ldots + \alpha_d,$$

but it is not immediately clear that this is indeed the case. Let us anyway define this alternative notion of distance, in which states are not compared by their trace distances but by the total variation distances of their spectra.

Definition 1.4.17. We will now introduce a "permutation-invariant" notion of total variation distance. Suppose α, β are distributions on $[d]$. We define

$$d_{\mathrm{TV}}^{\mathrm{sym}}(\alpha, \beta) \coloneqq d_{\mathrm{TV}}(\alpha^{\downarrow}, \beta^{\downarrow}) = \min_{\pi \in \mathfrak{S}(d)} \{d_{\mathrm{TV}}(\alpha, \beta_\pi)\}.$$

Here, for a vector $x \in \mathbb{R}^d$, $x^{\downarrow}$ denotes the rearrangement of x's coordinates in nonincreasing order, so $x_1^{\downarrow} \geq \cdots \geq x_d^{\downarrow}$. Here, the equivalence of the two expressions is by Fact 1.3.9. By virtue of the permutation-invariance, we may also naturally extend this notation to the case when α and β are simply unordered multisets of nonnegative numbers summing to 1.

If ρ and σ are d-dimensional mixed states with eigenvalues $\{\alpha_1, \ldots, \alpha_d\}$ and $\{\beta_1, \ldots, \beta_d\}$ (thought of as a multiset), respectively, we will use the notation

$$d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \sigma) \coloneqq d_{\mathrm{TV}}^{\mathrm{sym}}(\{\alpha_1, \ldots, \alpha_d\}, \{\beta_1, \ldots, \beta_d\}).$$

Equivalently, if α and β are ρ and σ's sorted spectra, respectively, then

$$d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \sigma) = d_{\mathrm{TV}}^{\mathrm{sym}}(\alpha, \beta) = d_{\mathrm{TV}}(\alpha, \beta).$$

Definition 1.4.18. We will instantiate property testing in the following settings:

- **Unitarily invariant properties of mixed states:** As in Definition 1.4.15, but $\mathcal{P}$ must be unitarily invariant.

- **Quantum spectrum testing:** $\mathcal{O}$ is the set of d-dimensional mixed states, $\mathcal{P}$ must be unitarily invariant, and $\mathrm{dist}(\rho, \sigma) = d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \sigma)$.

The following proposition shows that these two models are in fact equivalent. This equivalence allows us to rephrase the mixed state testing model, in the case of unitarily invariant properties, as a spectrum testing model.

Proposition 1.4.19. *The two models in Definition 1.4.18 are equivalent.*

Proof. We need to show that if $\mathcal{P}$ is a unitarily invariant property of d-dimensional mixed states then $d_{\mathrm{tr}}(\rho, \mathcal{P}) = d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \mathcal{P})$ holds for all mixed states ρ. By performing a unitary transformation, we may assume without loss of generality that ρ is a diagonal matrix with nonincreasing diagonal entries (spectrum).

The easy direction of the proof is showing that $d_{\mathrm{tr}}(\rho, \mathcal{P}) \leq d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \mathcal{P})$. To see this, suppose $\sigma \in \mathcal{P}$ achieves $d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \sigma) = \epsilon$. Let σ' denote the diagonal density matrix whose diagonal entries are the eigenvalues of σ arranged in nonincreasing order. Now σ' is unitarily equivalent to σ, and hence $\sigma' \in \mathcal{P}$ as well. But $d_{\mathrm{tr}}(\rho, \sigma') = \epsilon$ by Fact 1.4.2 and we therefore conclude $d_{\mathrm{tr}}(\rho, \mathcal{P}) \leq \epsilon$, as needed.

The more interesting direction is showing that $d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \mathcal{P}) \leq d_{\mathrm{tr}}(\rho, \mathcal{P})$. However, we have essentially already carried out the proof. Suppose that $\sigma \in \mathcal{P}$ achieves $d_{\mathrm{tr}}(\rho, \sigma) = \epsilon$. Then by Fact 1.4.5, $d_{\mathrm{TV}}(\alpha, \beta) \leq \epsilon$, if α and β are ρ and σ's sorted spectra, respectively. But we are done, as $d_{\mathrm{TV}}^{\mathrm{sym}}(\rho, \sigma) = d_{\mathrm{TV}}(\alpha, \beta)$. □

Having finished the setup, let us now discuss prior work and our results. Analogous to the case of testing properties of probability distributions, in the case of testing properties of mixed states, any property can be tested using $n = O(d^2/\epsilon^2)$ copies of ρ. This is due to our main tomography result (Corollary 1.4.12), which allows us to estimate ρ to $\epsilon/2$-accuracy with this many copies and check whether this estimate is $\epsilon/2$-close to $\mathcal{P}$. As a result, the goal of this area is to find testers for properties which use a *subquadratic* (in d) number of copies. That such algorithms could exist is suggested by the following quantum version of Fact 1.3.12, proven by [CHW07] (see also [DF09], who independently reproved the main combinatorial statement used to prove this theorem).

Theorem 1.4.20. $\Theta(r)$ *copies of a state* ρ *are necessary and sufficient to distinguish between the cases when* ρ*'s spectrum is uniform on either* r *or* $2r$ *values. (The bound also holds for* r *vs.* cr *when* $c > 2$ *is an integer.)*

Property testing of mixed states was first explicitly studied in [MdW13], an excellent survey on property testing in quantum computing. They suggested a variety of testing problems to work on; we include these and others below.

Definition 1.4.21. The following are some examples of testing problems.

- In *identity testing*, there is a known density matrix σ, and the goal is to test whether $\rho = \sigma$.
- *Mixedness testing* is the case of identity testing when σ is the maximally mixed state.
- *Diagonality testing* is the problem of testing whether ρ is diagonal (in a given basis).
- Given two unknown mixed states $\rho_1, \rho_2 \in \mathbb{C}^{d\times d}$, *equality testing* is the problem of testing whether $\rho_1 = \rho_2$.
- The following are some examples of spectrum testing problems:
  - *Rank testing* is the problem of testing whether ρ is rank r (in other words, testing whether α has at most r nonzero elements). As a special case, *purity testing* is the problem of testing whether ρ is pure, i.e. rank one.
  - *Von Neumann entropy testing* is the problem of testing whether ρ's von Neumann entropy is at most β (equivalently, testing whether α has entropy at most β), for some real number β.
  - Mixedness testing can be viewed as the problem of testing whether ρ's spectrum α is $(\frac{1}{d}, \ldots, \frac{1}{d})$.

We can derive two easy lower bounds on these testing problems. First, Theorem 1.4.20 provides a linear lower bound (in d) for all of the spectrum testing problems, as the state whose spectrum is uniform on $d/2$ values is far from the maximally mixed state and has "small" rank and von Neumann entropy, whereas the state whose spectrum is uniform on all d values *is* the maximally mixed state and has maximum rank and von Neumann entropy. Second, recall from Example 1.2.3 that in the case when one knows ρ's eigenvectors, then the optimal measurement returns a statistic isomorphic to a sample $\boldsymbol{w} \sim \alpha^{\otimes n}$. Hence, in this case, testing whether α has a particular property $\mathcal{P}$ is equivalent to classical distribution testing, and so the lower bounds from classical distribution testing apply.

Fact 1.4.22. *Let $\mathcal{P}$ be a symmetric property of probability distributions which requires $f(d, \epsilon)$ samples to test classically. Then testing whether a mixed state's spectrum satisfies $\mathcal{P}$ also requires at least $f(d, \epsilon)$ copies of the mixed state.*

Indeed, testing unitarily invariant properties of mixed states can be viewed as the quantum analogue of testing symmetric properties of probability distributions. In the first, a property is invariant under unitary conjugation, and in the second, the property is invariant under arbitrary permutation. This means that Definition 1.3.15 would be the same if it were defined with $\mathrm{dist} = d_{\mathrm{TV}}^{\mathrm{symm}}$ instead.

Fact 1.4.22 shows that quantum spectrum testing is at least as hard as testing symmetric properties of probability distributions, but there are some interesting nontrivial properties which have the same complexity in both models (up to constant factors). For example, if $\mathcal{P}$ is the property of ρ being a pure state (meaning that α has support size one), then $\Theta(1/\epsilon)$ samples/copies are necessary and sufficient to test $\mathcal{P}$ in both models (see [MdW13] for the $O(1/\epsilon)$ quantum spectrum testing upper bound using the swap test). In general, however, it is known that spectrum testing can require an asymptotically higher complexity (at least in terms of the parameter d).

Our first property testing result is a quantum analogue of Paninski's Theorem 1.3.14.

Theorem 1.4.23. *$\Theta(d/\epsilon^2)$ copies are necessary and sufficient to test whether $\rho \in \mathbb{C}^{d \times d}$ is the maximally mixed state.*

We also remark that given the way we prove Theorem 1.4.23, Childs et al.'s Theorem 1.4.20 can be obtained as a very special case. Our second result gives new bounds for testing whether a state has low rank.

Theorem 1.4.24. *$\Theta(r^2/\epsilon)$ copies are necessary and sufficient to test whether $\rho \in \mathbb{C}^{d \times d}$ has rank r with one-sided error. With two-sided error, a lower bound of $\Omega(r/\epsilon)$ holds.*

Here, by *one-sided error* we mean that the tester must accept rank-r states always but is allowed to err with constant probability on non-rank-r states. On the other hand, *two-sided error* means that the tester is allowed to err in either case. We note that the copy complexity is independent of the ambient dimension d. Compare this to Theorem 1.3.17.

Finally, we extend Childs et al.'s Theorem 1.4.20 to r vs. r' for *any* $r + 1 \leq r' \leq 2r$. A qualitative difference is seen when $r' = r + 1$; namely, nearly quadratically many copies are necessary.

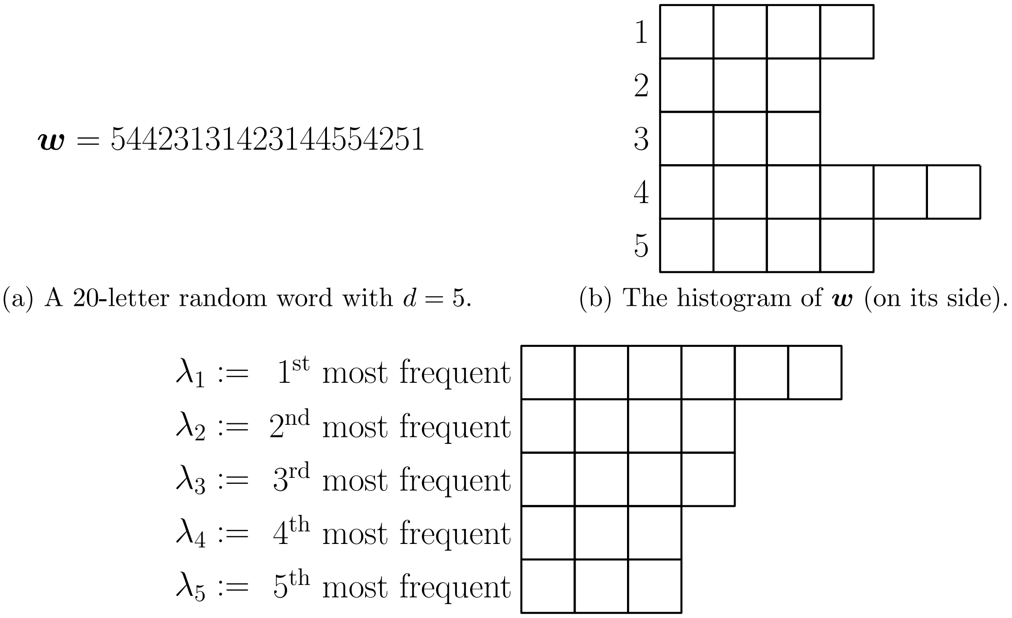
Theorem 1.4.25. *Let $1 \leq \Delta \leq r$. Then $O(r^2/\Delta)$ copies are sufficient to distinguish between the cases when ρ's spectrum is uniform on either r or $r + \Delta$ eigenvalues; further, a nearly matching lower bound of $\widetilde{\Omega}(r^2/\Delta)$ copies holds.*

As above, we note that these bounds are independent of the ambient dimension d. (Note that though one is testing between two options, this theorem does not strictly fall into the property testing model.)

1.5 Our methodology

Let us describe our methodology for learning or testing properties just of ρ's spectrum α. This methodology is also relevant for tomography and PCA, as solving these problems requires first learning some or all of α. The first step of our methodology is standard: we use a symmetry-based argument to show that when learning or testing properties of α, there is an optimal measurement one should use without loss of generality. This reduces our study of spectrum learning and testing to the study of just one measurement, both for our upper and lower bounds. The second step of our methodology is novel: we relate the behavior of this measurement to the combinatorial subject of longest increasing subsequences of random words. This not only gives us many new tools with which to analyze our quantum algorithms, it also allows our analyses to be conceptual rather than technical.

Optimality from symmetry. To illustrate the symmetry-based argument, let us show a similar argument in the context of classically learning or testing symmetric properties of probability distributions. Here, one is given n independent and identically distributed (i.i.d.) samples from the unknown distribution, as in Figure 1.1a.


$\boldsymbol{w} = 54423131423144554251$
1
2
3
4
5
$\lambda_1 :=$ 1st most frequent
$\lambda_2 :=$ 2nd most frequent
$\lambda_3 :=$ 3rd most frequent
$\lambda_4 :=$ 4th most frequent
$\lambda_5 :=$ 5th most frequent

(a) A 20-letter random word with $d = 5$.

(b) The histogram of $\boldsymbol{w}$ (on its side).

(c) The sorted histogram of $\boldsymbol{w}$.

Figure 1.1: The process of creating the sorted histogram of a word.

First, note that because the samples are i.i.d., their order doesn't matter. As a result, we lose nothing if we "forget" the ordering information and retain only the histogram, as in Figure 1.1b. Second, note that because we care only about a symmetric property of the distribution, the "names" of the elements in the distribution don't matter, only the frequencies with which they occur in the sample. As a result, we lose nothing if we also "forget" the names and sort the histogram, retaining only the frequency statistics, as in Figure 1.1c. Formally, we can view this as symmetry under the product group $\mathfrak{S}(n) \times \mathfrak{S}(d)$, and by retaining only the sorted histogram we have "factored out" this symmetry.

The sorted histogram is always a collection of n boxes arranged into d rows whose row lengths $\lambda_1, \ldots, \lambda_d$ are always *nondecreasing*, i.e. $\lambda_1 \geq \ldots \geq \lambda_d$. In the combinatorics and representation theory literature, an object of this type is called a *Young diagram* and is typically named "λ". The d values $(\lambda_1, \ldots, \lambda_d)$ are also referred to as a *partition*, and the fact that they partition n is denoted as $\lambda \vdash n$. The result of the above discussion is that when learning a symmetric property of the probability distribution α, rather than receiving a sample $\boldsymbol{w} \sim \alpha^{\otimes n}$, we can without loss of generality assume that the algorithm received a Young diagram $\boldsymbol{\lambda}$ distributed as follows.

1. Let $\boldsymbol{w}$ be an n-letter α-random word.

2. Output $\boldsymbol{\lambda}$, the sorted histogram of $\boldsymbol{w}$.

Equivalently, $\boldsymbol{\lambda}$ is drawn from the distribution in which for each $\lambda = (\lambda_1, \ldots, \lambda_d)$,

$$\mathbf{Pr}[\boldsymbol{\lambda} = \lambda] = \binom{n}{\lambda} \cdot m_\lambda(\alpha), \tag{1.8}$$

where m_λ is the *monomial symmetric function* (see Section 2.4.1). "Forgetting" the sample and retaining only the sorted histogram is a standard first step in the literature on learning symmetric properties of probability distributions; here, $\boldsymbol{\lambda}$ is referred to as the *fingerprint* of the sample [Bat01, Val08].

A similar situation holds for learning properties of a mixed state ρ's spectrum α. Here the problem exhibits symmetry under the product group $\mathfrak{S}(n) \times \mathrm{U}(d)$: symmetry under $\mathfrak{S}(n)$ because the n copies of ρ are identical, and symmetry under $\mathrm{U}(d)$ because the spectrum is invariant under rotation, i.e. ρ and $U\rho U^\dagger$ have the same spectrum for $U \in \mathrm{U}(d)$. Factoring these out involves a powerful result from representation theory called *Schur-Weyl duality*; the punchline is that there is a measurement called *weak Schur sampling (WSS)* which is the optimal measurement whenever one is trying to learn any property just of the spectrum α. Furthermore, though the measurement itself is difficult to state, its input/output behavior is (relatively) simple: it is an entangled measurement, meaning that it uses up all n copies of ρ simultaneously, and its measurement outcomes correspond to Young diagrams λ with n boxes and d rows. Furthermore, it is possible to explicitly compute the probability that WSS outputs a particular Young diagram λ on a state with spectrum α, and the result is an expression that looks similar to Equation (1.8):

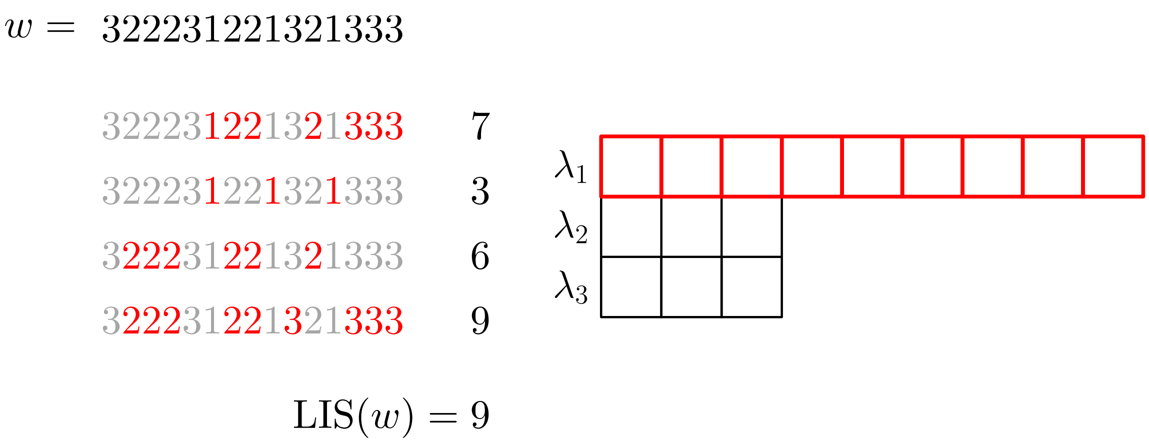
$$\mathbf{Pr}[\boldsymbol{\lambda} = \lambda] = \dim(\mathrm{V}_\lambda^d) \cdot s_\lambda(\alpha),$$

where $\dim(\mathrm{V}_\lambda^d)$ is an absolute constant associated with λ, and s_λ is a symmetric polynomial known as the *Schur polynomial* (see Section 2.4.1). (That s_λ is *symmetric* will prove to be important throughout this thesis, and it is reasonable given that the eigenvalues of ρ have no intrinsic ordering.) Hence, when one performs WSS, one receives a random Young diagram, and the goal is to infer the desired property of α from the diagram. This approach of using WSS to learn properties of the spectrum is widely used in the literature [ARS88, CEM99, KW01, HM02, CM06, CHW07, Mon09], and it is typically analyzed by explicit computations involving formulas for Schur polynomials, the dimension constants $\dim(\mathrm{V}_\lambda^d)$, and the "projectors" (see Section 2.6 below) associated with WSS.

Relating WSS to increasing subsequences. The main conceptual contribution of this thesis is a new interpretation of the output distribution of weak Schur sampling in terms of a certain combinatorial process related to longest increasing subsequences of random words. In particular, if we perform WSS on a mixed state ρ with spectrum α, then the random Young diagram $\boldsymbol{\lambda}$ received has the same distribution as the following process:

1. Let $\boldsymbol{w}$ be an n-letter α-random word.
2. Output $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$.

Here shRSK refers to the *RSK algorithm*, named after its inventors Robinson (a mathematician) [Rob38], Schensted (a physicist) [Sch61], and Knuth (a computer scientist) [Knu70]. It is an elegant combinatorial algorithm which scans the word $\boldsymbol{w}$ from left to right and iteratively constructs the Young diagram $\boldsymbol{\lambda}$ box-by-box by "inserting" the letters of $\boldsymbol{w}$ into $\boldsymbol{\lambda}$,



(a) Various increasing subsequences in w along with their lengths. The longest is at the bottom.

(b) The Young diagram $\lambda = \mathrm{shRSK}(w)$. By Schensted's theorem, the number of boxes in the first row λ_1 is equal to 9 because $\mathrm{LIS}(w) = 9$.

Figure 1.2: An illustration of Schensted's theorem.

"bumping" previously inserted letters in $\boldsymbol{\lambda}$ in the process. See [O'D16] for a video demonstration, [Gog99] for an interactive applet, and Section 3.2 below for a formal definition.[3]

Unfortunately, the iterative nature of the RSK algorithm means that it can be conceptually difficult to analyze. However, there is an alternative interpretation of the RSK algorithm due to Schensted [Sch61] and Greene [Gre74] in terms of the *longest increasing subsequence statistics* of $\boldsymbol{w}$ that is more amenable to analysis. Letting $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$, Schensted showed that the number of boxes in the first row of $\boldsymbol{\lambda}$, i.e. $\boldsymbol{\lambda}_1$, is equal to $\mathrm{LIS}(\boldsymbol{w})$, the length of the longest weakly increasing subsequence of $\boldsymbol{w}$.

Definition 1.5.1. Given a word $w \in \mathcal{A}^n$, a *subsequence* is a set of indices $1 \leq i_1 < \ldots < i_t \leq n$. The subsequence is *weakly increasing* if $w_{i_1} \leq \ldots \leq w_{i_t}$ and *strongly increasing* if $w_{i_1} < \ldots < w_{i_t}$, and we define a subsequence to be weakly or strongly *decreasing* analogously. We will typically call a subsequence *increasing* or *decreasing* if we mean that it is *weakly* increasing or decreasing. Finally, we define $\mathrm{LIS}(w)$ to be the length of the longest increasing subsequence of w and $\mathrm{LDS}(w)$ to be the length of the longest strongly decreasing subsequence.

We illustrate this definition in Figure 1.2a and Schensted's theorem in Figure 1.2b. Note the asymmetry in the definitions of LIS and LDS: the first refers to a *weakly* increasing subsequence whereas the second refers to a *strongly* decreasing subsequence. This distinction is important when w has repeated letters; in the case when w has no repeated letters (e.g. the case when w is a permutation), the distinction is irrelevant.

Following Schensted, Greene provided a significant generalization of Schensted's theorem to the first k rows of $\boldsymbol{\lambda}$, for any $1 \leq k \leq d$. For $k = 2$, he showed that $\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2$, the number of boxes in the first two rows of $\boldsymbol{\lambda}$, is equal to the length of the longest disjoint union of two increasing subsequences in $\boldsymbol{w}$. This is illustrated in Figure 1.3. More generally, he showed that $\lambda_1 + \cdots + \lambda_k$ is equal to the length of the longest disjoint union of k increasing

[3] In fact, the RSK algorithm actually outputs *two* "Young tableaus" P and Q whose common shape is λ. We write this as $(P, Q) = \mathrm{RSK}(w)$ and $\lambda = \mathrm{shRSK}(w)$. See Section 3.2 for fully correct details.

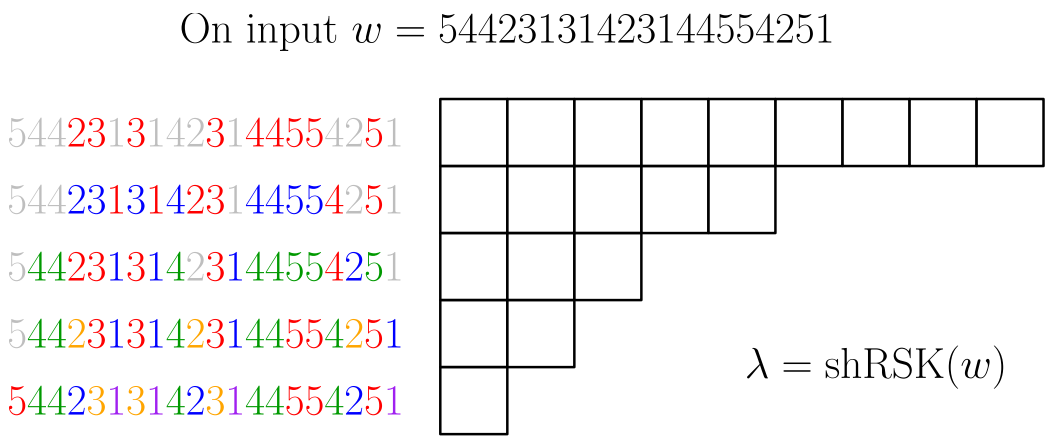


Figure 1.3: An illustration of Greene's theorem. On each row k, we have highlighted the letters in the longest set of k disjoint increasing subsequences. The number of letters highlighted should be equal to $\lambda_1 + \cdots + \lambda_k$, the number of boxes in the first k rows of λ.

subsequences in $\boldsymbol{w}$. As a special case, note that one can always find d disjoint increasing subsequences which cover all of $\boldsymbol{w}$: the all-ones subsequence, the all-twos subsequence, ..., and the all-d's subsequence. Hence, $\lambda_1 + \ldots + \lambda_d$, the number of boxes in the first d rows of $\boldsymbol{\lambda}$, is always equal to n.

At this point, the high level picture is that we have the following two distributions on Young diagrams:

$\boxed{\textbf{D1}}$

1. Run the weak Schur sampling measurement on n copies of ρ.
2. Receive a random measurement outcome $\boldsymbol{\lambda}$.

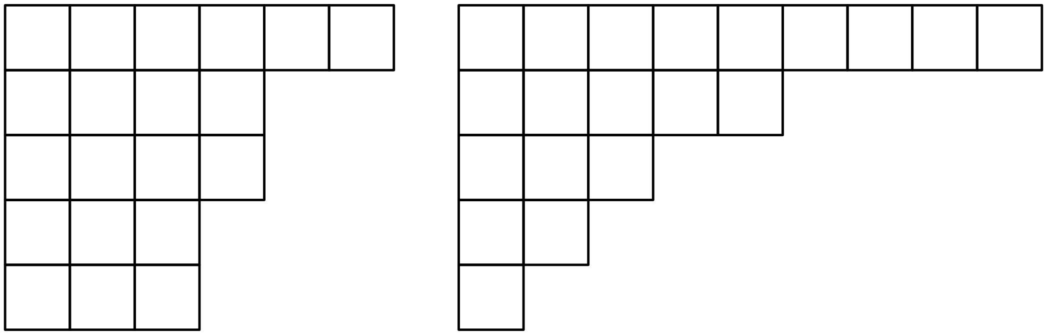
$\boxed{\textbf{D2}}$

1. Let $\boldsymbol{w}$ be an n-letter α-random word.
2. Set $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$.

For our application in quantum state learning, we would like to understand distribution **D1**, but this may be difficult due to its origins in representation theory and quantum computing. However, we have argued that these two distributions are the same, and so we can instead analyze $\boldsymbol{\lambda}$ as if it came from **D2**! But topics such as longest increasing subsequences of random words and the RSK algorithm have been studied in the mathematics literature for over 50 years, and so we gain access to a large body of techniques which were previously unavailable to researchers studying quantum state learning.

Definition 1.5.2. We will refer to the distribution on Young diagrams given by **D1** or **D2** as the *Schur-Weyl distribution* (after Schur-Weyl duality) and refer to it either by SW^n_ρ or $\mathrm{SW}^n(\alpha)$.

As we saw earlier in the context of distribution **D1**, the probability that a Young diagram λ is sampled from $\mathrm{SW}^n(\alpha)$ is a symmetric polynomial in α. This implies the fact, which is far from obvious given Greene's theorem or the definition of the RSK algorithm, that distribution **D2** is invariant under permuting the α_i's. This surprising fact will play a key role in many of our proofs, as it is often convenient to assume that the α_i's are sorted in either increasing or decreasing order. The proof can be found below as Corollary 3.3.5.



(a) The sorted histogram μ of w.

(b) $\lambda = \text{shRSK}(w)$, the LIS statistics of w. Note that λ is more "top-heavy" than μ, meaning $\lambda \trianglerighteq \mu$.

Figure 1.4: Two types of statistics of the word $w = 54423131423144554251$.

Analyzing the RSK algorithm. Let us now give a taste of the type of analysis involved in learning properties of ρ's spectrum and show how it relates to the classical topic of learning symmetric properties of a distribution. Consider the word $w = 54423131423144554251$. When learning symmetric properties of a distribution, one is given its sorted histogram μ, as in Figure 1.4a. When learning properties of a mixed state's spectrum, one is given its LIS statistics λ, as in Figure 1.4b. Though these two diagrams represent different types of statistics, they are related through *majorization*.

Notation 1.5.3. For $x, y \in \mathbb{R}^d$, we say x *majorizes* y, denoted $x \succ y$, if $\sum_{i=1}^k x_{[i]} \geq \sum_{i=1}^k y_{[i]}$ for all $k \in [d] = \{1, 2, \ldots, d\}$ (recalling Notation 1.3.7), with equality for $k = d$. We also use the traditional notation $\lambda \trianglerighteq \mu$ instead when λ and μ are partitions of n (Young diagrams).

In particular, we have the majorization relationship $\lambda \trianglerighteq \mu$ between the LIS statistics and the sorted histogram. To see this, note that λ and μ are Young diagrams and are therefore already sorted, and so to show majorization we need to show that the partial sums $\lambda_1 + \cdots + \lambda_k$ are always greater than $\mu_1 + \ldots + \mu_k$, for $1 \leq k \leq d$. This follows because w always has k elements which occur with frequencies $\mu_1, \ldots, \mu_k$, and the k subsequences corresponding to just these elements are both disjoint and increasing, and hence $\lambda_1 + \cdots + \lambda_k \geq \mu_1 + \cdots + \mu_k$. As a result, though in the quantum setting we would prefer to have access to the sorted histogram of w, we instead are only given its LIS statistics, which are like a "top-heavy" version of the sorted histogram.

Let $\boldsymbol{w}$ be an n-letter α-random word. Suppose our only goal were to compute α_1, the maximum probability value. If we were given $\boldsymbol{\mu}$, namely $\boldsymbol{w}$'s sorted histogram, then the standard algorithm for this would be to output $\boldsymbol{\mu}_1/n$. By Proposition 1.3.8 this is accurate to within error $\pm\epsilon$ (with high probability) so long as $n = O(1/\epsilon^2)$. However, in distribution **D2** we are only given $\boldsymbol{\lambda}$, namely $\boldsymbol{w}$'s LIS statistics. Still, for lack of a better idea, it is natural to output $\boldsymbol{\lambda}_1/n$ as a guess for α_1. By Schensted's theorem ($\boldsymbol{\lambda}_1 = \text{LIS}(\boldsymbol{w})$), we arrive at the following question:

What is the expected length of the longest increasing subsequence of an n-letter α-random word?

It is perhaps at first surprising that $\mathrm{LIS}(\boldsymbol{w})/n$ does indeed approach α_1 as n approaches infinity. (The rough intuition is that if $\alpha_1 \gg \alpha_2$, then there are so many 1s that the LIS should just take them all.) There is a simple combinatorial proof of this fact (see Proposition 3.7.1 below), and we encourage the reader to try to prove this for themselves first. One of our main results is to provide tight error bounds on the rate of convergence of this estimator. In particular, we show that this estimate is accurate to within $\pm\epsilon$ error (with high probability) so long as $n = O(1/\epsilon^2)$: just as good as if we had the sorted histogram! This is the $k = 1$ case of Theorem 1.4.8.

Suppose instead that we wanted to learn the entire spectrum α. If we had the sorted histogram $\boldsymbol{\mu}$, then the natural strategy would be to output the *empirical distribution*

$$\frac{\boldsymbol{\mu}}{n} := \left(\frac{\boldsymbol{\mu}_1}{n}, \dots, \frac{\boldsymbol{\mu}_d}{n}\right),$$

and it is known that this is an ϵ-accurate estimate in total variation distance when $n = O(d/\epsilon^2)$ (with high probability). If we try the same strategy with the LIS statistics and output $\boldsymbol{\lambda}/n$, then we have the well-known *empirical Young diagram (EYD)* algorithm (commonly known as the *Keyl-Werner* algorithm in the physics literature) mentioned in Section 1.4.

To analyze the EYD algorithm, it is natural to ask whether the estimator $\boldsymbol{\lambda}/n$ has similar moments as α, e.g. whether

$$\left(\frac{\boldsymbol{\lambda}_1}{n}\right)^2 + \ldots + \left(\frac{\boldsymbol{\lambda}_d}{n}\right)^2 \approx \alpha_1^2 + \ldots + \alpha_d^2,$$

up to some small error. In fact, it is possible to show that this second-moment condition is sufficient for $\boldsymbol{\lambda}/n$ to be a good approximator for α. Unfortunately, executing this plan of attack requires being able to compute quantities like the expectation $\mathbf{E}_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}_1^2 + \ldots + \boldsymbol{\lambda}_d^2]$, and to our knowledge no simple formula for this quantity exists. However, as it turns out, there are simple, explicit formulas which exactly compute the expectation of the following related quantity:

$$p_2^*(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_d) = \sum_{i=1}^{d} (\boldsymbol{\lambda}_i - i + \tfrac{1}{2})^2 - (-i + \tfrac{1}{2})^2.$$

(That one can compute the expectation of this quantity follows from results in representation theory: up to normalization, $p_2^*(\boldsymbol{\lambda})$ is the character $\chi_{\boldsymbol{\lambda}}((12))$.) Using this quantity as a proxy for the second moment, we can prove our Corollary 1.4.7, that $n = O(d^2/\epsilon^2)$ copies are sufficient to learn the spectrum.

Considering the approaches we take to solving our various learning and testing problems, our analysis then splits into these two styles: some of the time it uses the combinatorics of longest increasing subsequences of random words and the rest of the time it uses polynomials like $p_2^*(\boldsymbol{\lambda})$ (the so-called *shifted symmetric* polynomials; see Section 3.8) to analyze the moments of a random $\boldsymbol{\lambda}$. Some of our theorems, such as the general-k part of Theorem 1.4.8, require a combination of the two.

We will need the following majorization theorem in the proof of our truncated spectrum estimation and PCA results (Theorems 1.4.8 and 1.4.13). Essentially, it says that if α is more "top-heavy" than β, then the Young diagrams produced by $\mathrm{SW}^n(\alpha)$ are also more top-heavy than those produced by $\mathrm{SW}^n(\beta)$.

Theorem 1.5.4. *Let α, β be probability distributions on $[d]$ with $\beta \succ \alpha$. Then for any $n \in \mathbb{N}$ there is a coupling $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of $\mathrm{SW}^n(\alpha)$ and $\mathrm{SW}^n(\beta)$ such that $\boldsymbol{\mu} \trianglerighteq \boldsymbol{\lambda}$ always.*

The proof is entirely combinatorial, and can be read independently of the quantum content in the rest of the thesis. We note that though a priori the theorem statement seems like it must be true, our proof of it is quite nontrivial.

We will several times use the following elementary majorization inequality:

$$\text{If } c, x, y \in \mathbb{R}^d \text{ are sorted (decreasing) and } x \succ y \text{ then } c \cdot x \geq c \cdot y. \tag{1.9}$$

Conclusion. The take-home message of this thesis is that to understand quantum spectrum estimation, it suffices to understand longest increasing subsequences of random words. Using tools from longest increasing subsequences, we are able to give new and sometimes optimal results for quantum spectrum estimation.[4] Our results for quantum tomography and PCA build further on top of these results and also need new insights in the representation theory of the unitary and general linear groups.

1.6 Outline

The thesis is organized as follows.

- Chapter 2 is a high level introduction to representation theory. It also shows how to use representation theory to design quantum measurements (in particular, weak Schur sampling).

- Chapter 3 is a survey of longest increasing subsequences of random words. It will explain the connection to quantum algorithms and introduce many of the technical tools we will use in this thesis.

- Chapter 4 covers quantum spectrum estimation. It contains the proofs of Theorems 1.4.6 and 1.4.8.

- Chapter 5 covers quantum tomography. It contains the proofs of Theorems 1.4.11 and 1.4.13.

- Chapter 6 covers mixedness testing. It contains the proofs of Theorem 1.4.23.

- Chapter 7 covers uniform distribution testing. It contains the proofs of Theorem 1.4.25.

- Chapter 8 covers rank testing. It contains the proofs of Theorem 1.4.24.

- Chapter 9 contains the proof of Theorem 1.5.4.

[4] In fact, because **D1** = **D2**, one can draw exactly the reverse lesson: to understand longest increasing subsequences of random words, it suffices to understand quantum spectrum estimation. Though at first blush this looks like trying to understand one object via another more complicated object, this approach was carried out by Kuperberg [Kup02], who used quantum tools, such as the quantum central limit theorem, to give new proofs of well-known results about longest increasing subsequences.

- Chapter 10 contains a list of open problems.

The testing chapters, i.e. Chapters 6, 7, and 8, are based on work from [OW15b]. Chapter 9 is based on work from [OW16]. Chapter 4 is also based on work from [OW16], except for Section 4.3 which is based on work from [OW15b]. Finally, Chapter 5 comes from several sources: Section 5.1 is based on unpublished joint work with Akshay Krishnamurthy, Section 5.2 is based on [OW15a], Sections 5.3 and 5.4 are based on [OW16], and Section 5.5 is based on unpublished joint work with Ryan O'Donnell.

Chapter 2

Representation theory

The standard approach for designing quantum state learning algorithms in settings like ours is to exploit the *symmetries* in the problems. By this, we mean that some aspect of the problem is symmetric under some group action. The two key examples of this are the following.

Symmetry under the symmetric group: the inputs to our learning problems are of the form $\rho^{\otimes n}$, i.e. tensor product states which are invariant under permuting the n subsystems.

Symmetry under the unitary group: if our goal is to learn something which depends only on ρ's spectrum, then our algorithm may as well act identically on $\rho^{\otimes n}$ and $(U^\dagger \rho U)^{\otimes n}$, for any unitary matrix U.

The tool we will use—and the subject of this chapter—is *representation theory*, the mathematical field which studies, among other things, symmetries of matrices and vector spaces under group actions. Using representation theory, we will show that there is a basis in which $\rho^{\otimes n}$ has a particularly nice structure and, further, that measuring in this basis (via the so-called weak Schur sampling measurement) is optimal for many of our problems. This approach has been previously used to great success in works such as [ARS88, KW01, HM02, CM06, CHW07].

Due to the above symmetries, we will focus on the representation theory of the following three groups.

Definition 2.0.1. The *symmetric group* $\mathfrak{S}(n)$ is the group of permutations π of the set $[n]$.

Definition 2.0.2. The *general linear group* GL_d is the group of $d \times d$ complex invertible matrices. (More generally, if V is a vector space, then GL_V is the group of invertible linear transformations on V.)

Definition 2.0.3. The *unitary group* $\mathrm{U}(d)$ is the group of $d \times d$ unitary matrices.

The unitary group is a subgroup of the general linear group, and as such the representation theories of the two groups go hand-in-hand. Surprisingly, the representation theories of the symmetric and general linear groups are *also* linked, due to a powerful theorem known as *Schur-Weyl duality*. As these are precisely the two symmetries involved in our learning problems, Schur-Weyl duality plays a central role in the algorithms we consider.

The chapter is organized as follows.

- Section 2.1 gives an introduction to the subject of representation theory tailored to the three groups we are interested in.
- Section 2.2 introduces Young diagrams, a certain combinatorial object which occurs in the representation theories of all three groups we are interested in.
- Section 2.3 covers the representation theory of $\mathfrak{S}(n)$.
- Section 2.4 covers the representation theory of GL_d and $\mathrm{U}(d)$.
- Section 2.5 introduces Schur-Weyl duality, connecting the representation theories of $\mathfrak{S}(n)$ and GL_d.
- Section 2.6 introduces the general methodology we will use to design quantum algorithms based on representation theory.

2.1 Introduction to representation theory

In this section, we will give a high-level overview of representation theory. We will closely follow chapters 3 and 4 in the excellent introduction to representation theory by Steinberg [Ste11], and full proofs of the results in this section can be found there.

Definition 2.1.1. Given a group G, a *complex, finite-dimensional representation* (henceforth, a *representation*) of G is a pair (μ, V), where $\mu : G \to \mathrm{GL}_V$ is a group homomorphism and V is a finite-dimensional complex vector space. In other words, we associate with each $g \in G$ a matrix $\mu(g)$ such that $\mu(g)\mu(h) = \mu(gh)$ for any $g, h \in G$. The dimension of V is called the *dimension* of the representation μ and is denoted $\dim(\mu)$.

Examples:

- For any group G, the *trivial representation* is the one-dimensional representation given by $\mu(g) := [1]$ for each $g \in G$.
- The *standard representation* of GL_d is the d-dimensional representation given by $\mu(M) := M$ for each $M \in \mathrm{GL}_d$.
- The *sign representation* of $\mathfrak{S}(n)$ is the one-dimensional representation given by $\mu(\pi) := [\mathrm{sgn}(\pi)]$ for each $\pi \in \mathfrak{S}(n)$.
- The *determinant representation* of GL_d is the one-dimensional representation given by $\mu(M) := [\det(M)]$ for each $M \in \mathrm{GL}_d$. In fact, for any $z \in \mathbb{Z}$ one has a representation $\mu_z(M) := [\det(M)^z]$.

- For any finite group G, the *(left) regular* representation is the $|G|$-dimensional representation on the vector space $V = \{|g\rangle\}_{g\in G}$ which assigns to each $g \in G$ the permutation matrix $\mu(g)$ acting as $\mu(g)\,|h\rangle = |gh\rangle$ for each $g, h \in G$.

- For the group $\{-1,1\}^n$, where the group operation is defined as component-wise multiplication, there is a one-dimensional representation for each subset $S \subseteq [n]$ given by $\mu(x) = [\chi_S(x)]$, where $\chi_S(x) := \prod_{i\in S} x_i$.

Save for the last example, these representations play an important role in the representation theory of $\mathfrak{S}(n)$ and GL_d.

The two most important representations for us are given in the following definition.

Definition 2.1.2. The groups $\mathfrak{S}(n)$ and GL_d each have a natural action on the space $(\mathbb{C}^d)^{\otimes n}$; the associated representations P and Q (respectively) are defined on the standard basis vectors $|a_1\rangle \otimes |a_2\rangle \otimes \cdots |a_n\rangle$ (for $a_i \in [d]$) via

$$\begin{aligned}\mathrm{P}(\pi)\,|a_1\rangle \otimes |a_2\rangle \otimes \cdots \otimes |a_n\rangle &= |a_{\pi^{-1}(1)}\rangle \otimes |a_{\pi^{-1}(2)}\rangle \otimes \cdots \otimes |a_{\pi^{-1}(n)}\rangle\,,\\ \mathrm{Q}(M)\,|a_1\rangle \otimes |a_2\rangle \otimes \cdots \otimes |a_n\rangle &= (M\,|a_1\rangle) \otimes (M\,|a_2\rangle) \otimes \cdots \otimes (M\,|a_n\rangle).\end{aligned}$$

The representations are then extended to all of $(\mathbb{C}^d)^{\otimes n}$ by linearity.

The relationship between these two representations is the subject of Schur-Weyl duality (Section 2.5).

2.1.1 Decomposing representations

Given a representation, there are a variety of ways of generating new representations. We summarize two of these below.

Definition 2.1.3. If μ is a representation acting on a vector space V, then $M\mu M^{-1}$ is also a representation acting on V, for any invertible matrix $M \in \mathrm{GL}_V$. Two representations μ_1 and μ_2 are said to be *isomorphic* if there is some invertible matrix M such that $M\mu_1 M^{-1} = \mu_2$. In this case we write $\mu_1 \cong \mu_2$.

Definition 2.1.4. The *direct sum* of k representations $(\mu_1, V_1), \ldots, (\mu_k, V_k)$ produces the representation $(\mu, V = V_1 \oplus \cdots \oplus V_k)$ given by block-diagonal matrices:

$$\mu(g) := \begin{bmatrix} \mu_1(g) & 0 & \cdots & 0 \\ 0 & \mu_2(g) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_k(g) \end{bmatrix}$$

for all $g \in G$. Equivalently, we may write

$$\mu(g) := \sum_{i=1}^{k} |i\rangle\langle i| \otimes \mu_i(g).$$

We will also write $\mu = \mu_1 \oplus \ldots \oplus \mu_k$ to denote that μ is the direct sum of $\mu_1, \ldots, \mu_k$. In general, each μ_i may appear multiple (say, m_i) times in the decomposition of μ, which we write as

$$\mu = \bigoplus_{i=1}^{k} m_i \cdot \mu_i := \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{m_i} \mu_i.$$

We will refer to m_i as μ_i's *multiplicity*. Finally, if the μ's and μ_i's are understood from context, we may also write this as

$$V \overset{G}{\cong} \bigoplus_{i=1}^{k} m_i \cdot V_i,$$

where G is the group being represented. (We may also choose to omit G from this expression.) A representation μ is *decomposable* if $\mu \cong \mu_1 \oplus \mu_2$, where μ_1, μ_2 are representations of dimension at least one. Otherwise, μ is *indecomposable*.

A decomposable representation divides the vector space into two orthogonal subspaces on which the representation acts independently.

Definition 2.1.5. If (μ, V) is a representation of G, then an *invariant subspace* is a subspace $U \subseteq V$ such that $\mu(g)U = U$ for all $g \in G$. An invariant subspace is *trivial* if it is either $\emptyset$ or V. We say that μ is *reducible* if it has a nontrivial invariant subspace. Otherwise, it is *irreducible*. For brevity, we will often refer to an irreducible representation as an *irrep*.

Irreducible representations are analogous to prime factors in the world of representation theory. Given a representation of a group, standard tasks involve (i) determining if it is irreducible and (ii) if not, determining how it decomposes as the direct sum of irreducible representations.

Examples:

- The trivial, sign, and determinant representations are trivially irreducible because they are one-dimensional.
- The standard representation is irreducible, because for any two vectors $|v_1\rangle, |v_2\rangle$ there is an invertible matrix M such that $M|v_1\rangle = |v_2\rangle$.
- The regular representation is reducible (if $|G| \geq 2$). This because for every $g \in G$, $\mu(g)$ is a permutation matrix, and so it fixes the vector $\sum_{g \in G} |g\rangle$.
- Consider the subspace of $(\mathbb{C}^d)^{\otimes n}$ spanned by vectors of the form $|v\rangle \otimes \cdots \otimes |v\rangle$. When $n = 1$, this is all of $(\mathbb{C}^d)^{\otimes n}$, but for larger n this is a nontrivial subspace known as the *symmetric subspace*. The symmetric subspace is an invariant subspace of P and Q, and so these representations are reducible when $n > 1$. That it is an invariant subspace for *both* representations is a manifestation of Schur-Weyl duality (Section 2.5).

Note that if a representation is decomposable, then it is reducible, but the reverse is not true in general. For example, the representation μ of the group $\mathbb{C}$ under addition given by

$$\mu(z) = \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix}$$

has $\operatorname{span}\{e_1\}$ as an invariant subspace, and hence is reducible, but $\mu(1)$ is not diagonalizable, and hence μ is not decomposable.

Given that not every reducible representation is decomposable, we would like to know simple conditions that guarantee *complete reducibility*, meaning that the representation can be decomposed into a direct sum of irreducibles. One such condition is that the representation always be a unitary matrix.

Definition 2.1.6. A *unitary representation* of a group G is a representation of the form $\mu : G \to \mathrm{U}(d)$.

Any unitary representation μ is completely reducible, because if V is a nontrivial invariant subspace then so is $V^\perp$. This decomposes μ as $\mu = \mu_1 \oplus \mu_2$, where μ_1, μ_2 are unitary subrepresentations, and one can recursively apply this argument until the subrepresentations have no nontrivial subspaces, and hence are irreducible. The question now becomes for which groups can we guarantee unitary representations, and this question is answered for two of our favorite groups in the following.

Theorem 2.1.7. *If G is a finite group or a unitary group (i.e. $\mathrm{U}(d)$ for some d) then every representation of G is isomorphic to a unitary representation. As a result, every representation of G is completely reducible.*

One consequence is that we may always assume that the irreps of such a G are unitary. The case of finite G is known as Maschke's Theorem and can be found in [Ste11]. The argument for unitary groups is similar and can be found in [Hal15, Proposition 4.36]. Finite groups have other nice properties, such as the following.

Theorem 2.1.8. *If G is a finite group, then the set of all irreps (up to isomorphism), which we denote by $\widehat{G}$, is finite.*

Note that if two representations μ_1 and μ_2 are isomorphic, then this implies that they have an *intertwining operator*, a matrix T such that $T\mu_1 = \mu_2 T$.

Lemma 2.1.9 (Schur's lemma). *Suppose μ_1 and μ_2 are irreducible representations of a group G which have an intertwining operator T. Then either T is invertible (in which case $\mu_1 \cong \mu_2$) or $T = 0$. Furthermore, if $\mu_1 = \mu_2$, then $T = \lambda I$ with $\lambda \in \mathbb{C}$.*

Schur's lemma has a variety of useful consequences, e.g.:

Corollary 2.1.10. *Suppose μ_1 and μ_2 are isomorphic unitary representations. Then they are related by a unitary change of basis, i.e. a unitary matrix U such that $U\mu_1 U^\dagger = \mu_2$.*

Proof. Because μ_1 and μ_2 are isomorphic, they have a nonzero intertwining operator T satisfying

$$T\mu_1 = \mu_2 T. \tag{2.1}$$

Furthermore, for any element $g \in G$,

$$\mu_1(g)T^\dagger = (T\mu_1(g)^\dagger)^\dagger = (T\mu_1(g^{-1}))^\dagger = (\mu_2(g^{-1})T)^\dagger = (\mu_2(g)^\dagger T)^\dagger = T^\dagger \mu_2(g), \tag{2.2}$$

where the second and fourth equalities use that μ_1 and μ_2 are unitary. Together, (2.1) and (2.2) imply that

$$\mu_1 T^\dagger T = T^\dagger \mu_2 T = T^\dagger T \mu_1,$$

meaning that $T^\dagger T$ is an intertwining operator of μ_1 with itself. By Schur's lemma, $T^\dagger T = \lambda I$ for some nonzero λ, and so $T/\sqrt{\lambda}$ is a unitary matrix which gives a unitary change of basis between μ_1 and μ_2. □

If μ is a unitary representation of a finite or unitary group, then it is isomorphic to a direct sum of unitary irreps. Since a direct sum of unitary matrices is unitary, this corollary gives a unitary change-of-basis between the original representation and the direct sum. This is useful for us as unitary rotations are exactly the operations allowed on a quantum computer.

A final way of combining representations involves two representations of different groups.

Definition 2.1.11. Let μ_1 be a representation of the group G_1 and μ_2 be a representation of the group G_2. Then the *tensor product* of μ_1 and μ_2, denoted $\mu_1 \otimes \mu_2$, is the representation defined by

$$(\mu_1 \otimes \mu_2)(g, h) \coloneqq (\mu_1(g)) \otimes (\mu_2(h)),$$

where the right-hand side uses the ordinary matrix tensor product. We have $\dim(\mu_1 \otimes \mu_2) = \dim(\mu_1) \cdot \dim(\mu_2)$.

2.1.2 The regular representation

One of the most basic representations of any finite group is the regular representation. It has the following decomposition into irreps.

Theorem 2.1.12. *Let G be a finite group and let μ_{reg} be its regular representation. Then*

$$\mu_{\text{reg}} \cong \bigoplus_{\mu \in \widehat{G}} \dim(\mu) \cdot \mu.$$

As the dimension of the representation on the left is $|G|$ and the dimension of the representation on the right is $\sum_{\mu \in \widehat{G}} \dim(\mu)^2$, we can conclude that

$$\sum_{\mu \in \widehat{G}} \dim(\mu)^2 = |G|. \tag{2.3}$$

This equality induces a natural probability distribution on the set of irreps.

Definition 2.1.13. For a finite group G, the *Plancherel distribution* is the probability distribution on irreps in which $\mu \in \widehat{G}$ has probability $(\dim \mu)^2/|G|$.

The Plancherel distribution plays a central role in this work.

Given the appealing form of Equation (2.3), it is natural to ask for an interpretation of the matrix elements of the irreps $\mu \in \widehat{G}$ as counting objects in some set of size $|G|$. As it turns out, there is such an interpretation in terms of functions $f : G \to \mathbb{C}$ under the following inner product.

Definition 2.1.14. Let $f, g : G \to \mathbb{C}$. Then we define the inner product $\langle\ ,\ \rangle$ as

$$\langle f, g\rangle := \mathop{\mathbf{E}}_{\boldsymbol{h}\sim G}[f(\boldsymbol{h})\overline{g(\boldsymbol{h})}],$$

where $\boldsymbol{h} \sim G$ is a uniformly random element of G.

The *Schur orthogonality relations* say that if we consider the matrix entries μ_{ij} of the irreps as functions $\mu_{ij} : G \to \mathbb{C}$, then these functions form an orthonormal basis (up to normalization) for the set of functions with domain G.

Theorem 2.1.15 (Schur orthogonality relations)**.** *Let μ and σ be nonisomorphic unitary irreps of a finite group G. Then for any $i, j \in [\dim(\mu)]$ and $k, \ell \in [\dim(\sigma)]$,*

1. $\langle \mu_{ij}, \sigma_{k\ell}\rangle = 0,$

2. $\langle \mu_{ij}, \mu_{k\ell}\rangle = \begin{cases} 1/\dim(\mu) & \textit{if } i = k \textit{ and } j = \ell, \\ 0 & \textit{otherwise.}\end{cases}$

Combined with (2.3), the matrix elements of $\sqrt{d_\mu} \cdot \mu$, for $\mu \in \widehat{G}$, give an orthonormal basis of functions $f : G \to \mathbb{C}$.

One consequence of the orthogonality relations is Theorem 2.1.8 above, that there is only a finite number of nonisomorphic irreps of a finite group.

2.1.3 Characters

Definition 2.1.16. If μ is a representation of a group G, then its *character* is the function $\chi_\mu : G \to \mathbb{C}$ given by $\chi_\mu(g) = \mathrm{tr}(\mu(g))$ for each $g \in G$.

Recall that a conjugacy class of a group is an equivalence class of the group under the equivalence relation $g_1 \sim g_2$ iff $hg_1h^{-1} = g_2$ for some $h \in G$. By the cyclic property of matrix trace, a character χ_μ of a representation is constant on the conjugacy classes of G.

Definition 2.1.17. A *class function* $f : G \to \mathbb{C}$ is any function which is constant on the conjugacy classes of G.

Thus, χ_μ is a class function.

Examples:

- If μ is a representation, then $\mu(e) \cdot \mu(e) = \mu(e^2) = \mu(e)$, and so $\mu(e)$ is a projection matrix. Furthermore, by definition, μ is invertible. As a result, $\mu(e)$ is the identity matrix, in which case
$$\chi_\mu(e) = \dim(\mu). \tag{2.4}$$
- For the one-dimensional representations (trivial, sign, etc.), the character is the value of the matrix entry.
- If μ is the standard representation, then $\chi_\mu(M) = \operatorname{tr}(M)$ for all M.
- If μ is the regular representation, then
$$\chi_\mu(e) = |G| \quad \text{and} \quad \chi_\mu(g) = 0 \text{ for all } e \neq g \in G. \tag{2.5}$$
The first equality is trivial. The second uses $gh = h$ iff $g = e$.
- The $(a_1, \dots, a_n)$-th diagonal entry of $\mathrm{P}(\pi)$ is one iff $a_{\pi(i)} = a_i$ for all $i \in [n]$ and is zero otherwise. There is such a nonzero diagonal entry for each way of assigning a number in $[d]$ to each cycle of π. As a result,
$$\chi_{\mathrm{P}}(\pi) = d^{\ell(\pi)},$$
where $\ell(\pi)$ denotes the number of cycles in π.
- For a matrix $M \in \mathrm{GL}_d$, $\mathrm{Q}(M)$ acts on $(\mathbb{C}^d)^{\otimes n}$ as $M^{\otimes n}$. Hence,
$$\chi_{\mathrm{Q}}(M) = \operatorname{tr}(M^{\otimes n}) = \operatorname{tr}(M)^n.$$

It is easy to check that these are all examples of class functions.

By the cyclic property of matrix trace, isomorphic representations have equivalent characters. On the other hand, nonisormorphic irreducible representations have orthogonal characters.

Theorem 2.1.18 (Character orthogonality relations)**.** *Let μ and σ be irreps of a finite group G. Then*
$$\langle \chi_\mu, \chi_\sigma \rangle = \begin{cases} 1 & \text{if } \mu \cong \sigma, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Because isomorphic representations have equivalent characters, we may assume that μ and σ are unitary irreps. By the definition of trace,
$$\chi_\mu = \sum_{i=1}^{\dim(\mu)} \mu_{i,i}. \quad \text{and} \quad \chi_\sigma = \sum_{i=1}^{\dim(\mu)} \sigma_{i,i}$$

By the Schur orthogonality relations, these functions have inner product zero unless $\mu \cong \sigma$, in which case their inner product is $\dim(\mu)/\dim(\mu) = 1$. □

In fact, not only do the χ_μ functions form an orthonormal set, but they also form a basis.

Theorem 2.1.19. *If G is a finite group, then the functions χ_μ for $\mu \in \widehat{G}$ form an orthonormal basis for the set of class functions on G. As a corollary, the number of nonisomorphic irreps of G is equal to the number of conjugacy classes of G.*

If a representation μ_{red} of a group G is completely reducible, then it can be written as $\mu_{\text{red}} \cong \bigoplus_{\mu \in \widehat{G}} m_\mu \cdot \mu$, for some multiplicities m_μ. To find these multiplicities, it suffices to note that

$$\chi_{\mu_{\text{red}}} = \sum_{\mu \in \widehat{G}} m_\mu \chi_\mu.$$

Then, by Theorem 2.1.18, we have the equality $m_\mu = \langle \mu_{\text{red}}, \mu \rangle$. For example, we can use this to prove Theorem 2.1.12.

Proof of Theorem 2.1.12. For an irrep $\mu \in \widehat{G}$, we can calculate its multiplicity as

$$\mathop{\mathbf{E}}_{\boldsymbol{g} \sim G} \chi_{\mu_{\text{reg}}}(\boldsymbol{g}) \chi_\mu(\boldsymbol{g}) = \frac{1}{|G|} \chi_{\mu_{\text{reg}}}(e) \chi_\mu(e) = \dim(\mu),$$

where the first equality used Equation (2.5) for $\chi_{\mu_{\text{reg}}}$ and (2.4) for χ_μ. □

We now recall some basics of Fourier analysis over an arbitrary finite group G (though we will ultimately only need the case $G = \mathfrak{S}(n)$). For general $f, g : G \to \mathbb{C}$ we define $(f * g)(u) = \mathbf{E}_{\boldsymbol{v} \sim G}[f(\boldsymbol{v}) g(\boldsymbol{v}^{-1} u)]$; this includes a nonstandard normalization by $\frac{1}{|G|}$. For a class function f and $\mu \in \widehat{G}$ we employ the following "Fourier notation": $\widetilde{f}(\mu) = \langle f, \chi_\mu \rangle$. (According to standard notation we would have $\widetilde{f}(\mu) = \frac{1}{|G|} \text{tr}\left(\widehat{f}\right)$). Then Fourier inversion is simply $f = \sum_\mu \widetilde{f}(\mu) \chi_\mu$. Further, if g is another class function we have the formula $\widetilde{f * g}(\mu) = \frac{1}{\dim \mu} \widetilde{f}(\mu) \widetilde{g}(\mu)$.

2.1.4 Branching rules

Given a representaton μ of a group G and a subgroup $H \subseteq G$, μ also trivially serves as a representation of H. When viewed in this way, we use the following notation.

Definition 2.1.20. Given a representation μ of a group G and a subgroup $H \subseteq G$, $\mu{\downarrow_H}$ denotes μ viewed as a representation of H.

Even if μ is an irreducible representation of G, it may not be an irreducible representation of H. For example, if μ is a representation of a group G with identity element e, then we have seen (Section 2.1.3) that $\mu(e)$ is the $\dim(\mu)$-dimensional identity matrix. On the other hand, the trivial group $\{e\}$ is a subgroup of G, and its only irreducible representation is the trivial representation. Together these mean that

$$\mu{\downarrow_{\{e\}}} = (\dim \mu) \cdot \mu_{\text{triv}},$$

by which we mean that μ_{triv} appears with multiplicity $\dim \mu$.

In general, we are interested in how a representation of G, when restricted to the subgroup H, decomposes into a direct sum of irreps for H.

Definition 2.1.21. Given a representation μ of a group G and a subgroup $H \subseteq G$, a *branching rule* is a statement of the form

$$\mu\downarrow_H \cong \bigoplus_{i=1}^{k} m_i \cdot \mu_i,$$

where the μ_i's are irreps of H. In the case when the m_i's are all one (or zero), we say that the branching rule is *multiplicity free*.

We will see examples of branching rules in Sections 2.3.2 and 2.4.2.

2.2 Partitions and Young diagrams

Definition 2.2.1. A *partition* of $n \geq 1$, denoted $\lambda \vdash n$, is a list of nonnegative integers $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$ satisfying $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ and $\lambda_1 + \lambda_2 + \dots + \lambda_k = n$.

Notation 2.2.2. We will use the following notation.

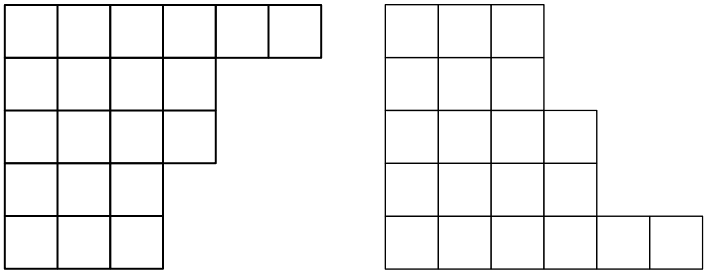
- The *length* of the partition, denoted $\ell(\lambda)$, is the number of nonzero λ_i's in λ.
- The partition's *size* is n, and is also written as $|\lambda|$.
- Two partitions are considered to be equivalent if they only differ in trailing zeros. For example, $(4, 2)$ and $(4, 2, 0, 0)$ are equivalent.
- We write Par to denote the set of all partitions, of any size.
- For $w \in \mathbb{N}^+$ we will use the notation $m_w(\lambda)$ to denote the number of parts i with $\lambda_i = w$.

Finally, at one point we will require the fairly elementary fact (see e.g. [Rom14, (1.15)]) that the number of partitions of n is $2^{O(\sqrt{n})}$ (much more precise asymptotics are known [HR18]).

2.2.1 Young diagrams

It is standard to represent a partition $\lambda \vdash n$ pictorially with a *Young diagram*; i.e., a certain arrangement of n squares, called *cells* or *boxes*. There are several conventions for how to draw Young diagrams: we will define the *English notation*, the *French notation*, the *Russian notation*, and the *Maya notation*.

- In the *English notation*, the Young diagram for $\lambda = (\lambda_1, \dots, \lambda_k)$ is drawn with left-justified rows of cells: λ_1 cells in the top row, λ_2 cells beneath this, λ_3 cells beneath this, etc. As an example, the English notation for $(6, 4, 4, 3, 3)$ is pictured in Figure 2.1a.
- The *French notation* is the reflection English notation across the horizontal axis. The French notation for $(6, 4, 4, 3, 3)$ is pictured in Figure 2.1b. We think of the French diagram as consisting of unit squares sitting in $\mathbb{R}^2_+$, with the bottom-left corner at the origin.



(a) English notation. (b) French notation.

Figure 2.1: Two ways of drawing the partition $\lambda = (6, 4, 4, 3, 3)$.

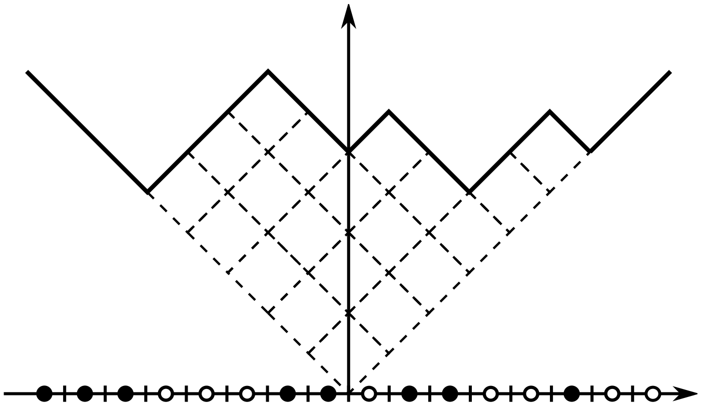


Figure 2.2: The Russian and Maya diagrams for the partition $\lambda = (6, 4, 4, 3, 3)$. The Russian diagram is given by the dashed lines while the Maya diagram is given by the pebbles. The function $\lambda : \mathbb{R} \to \mathbb{R}$ is given by the thick black line and tends towards infinity in both directions.

Definition 2.2.3. Given the English and French notations, it's natural to define the *width* of λ as λ_1, and to refer to $\ell(\lambda)$ as its *height*. We can also define the *conjugate partition* of λ to be the partition $\lambda' \vdash n$ obtained by reflecting the French diagram through the line $y = x$; i.e., exchanging rows and columns. For example, the conjugate of $\lambda = (6, 4, 4, 3, 3)$ is $\lambda' = (5, 5, 5, 3, 1, 1)$. Note that the height of λ is the width of λ', and vice versa; in particular, we sometimes prefer the notation λ'_1 to $\ell(\lambda)$.

- The *Russian notation* is obtained from the French notation by first rotating the diagram 45° counterclockwise about the origin, and then dilating by a factor of $\sqrt{2}$; see Figure 2.2. The purpose of the dilation is so that the corners of the boxes will have integer x- and y-coordinates. The purpose of the rotation is so that conjugation corresponds to reflection in the y-axis and so that the boundary of the diagram forms the graph of a function:

Definition 2.2.4. Given a partition λ drawn in Russian notation, its upper boundary forms the graph of a function with domain $[-\lambda'_1, \lambda_1] \subseteq \mathbb{R}$. We extend this function to have domain all of $\mathbb{R}$ according to the function $x \mapsto |x|$. We will use the notation $\lambda : \mathbb{R} \to \mathbb{R}_+$ for this function, which we remark is a continuous and piecewise linear curve. Any time we write $\lambda(x)$, where λ is a partition and $x \in \mathbb{R}$, we are referring to this curve. See Figure 2.2 for an example.

- Finally, we define the *Maya notation*. It contains no boxes; just a sequence of black and white pebbles. However the Maya notation is typically drawn in conjunction with the Russian notation, with the pebbles being located on the half-integer points $\mathbb{Z} + \frac{1}{2}$ of the x-axis. In the Maya notation, a black pebble is placed at all points directly below a “downward-sloping” segment in λ’s graph, and a white pebble is placed at all points directly below an “upward-sloping” segment. (Thus all sufficiently negative half-integer points have a black pebble and all sufficiently positive half-integer points have a white pebble.) The notation also includes a vertical tick mark to denote the location of the origin. A picture of the Russian and Maya notation for $\lambda = (6, 4, 4, 3, 3)$ appears in Figure 2.2. One can check that the sequence of pebbles uniquely identifies the partition λ. It also uniquely determines the position of the origin mark, in that the number of black pebbles to the right of the origin mark always equals the number of white pebbles to the left of the origin mark. These numbers are both equal to $d(\lambda)$, defined to be the number of cells touching the y-axis in the Russian diagram. We make one more definition:

Definition 2.2.5. Given the Maya diagram of a partition λ, we may define its *modified Frobenius coordinates* to be the half-integer values $a_1^* > a_2^* > \cdots a_d^* > 0$ and $b_1^* > b_2^* > \cdots > b_d^* > 0$ (for $d = d(\lambda)$), where a_i^* is the position of the i-th rightmost black pebble and b_i^* is the negative of the position of the i-th leftmost white pebble. One may check that, equivalently, $a_i^* = \lambda_i - i + \frac{1}{2}$ and $b_i^* = \lambda_i' - i + \frac{1}{2}$. For example, if $\lambda = (6, 4, 4, 3, 3)$, then $a^* = (\frac{11}{2}, \frac{5}{2}, \frac{3}{2})$ and $b^* = (\frac{9}{2}, \frac{7}{2}, \frac{5}{2})$. The coordinates have the property that $\sum_i (a_i^* + b_i^*) = |\lambda|$.

Definition 2.2.6. For a partition λ (drawn either in the English, French, or Russian notation), we often use the symbol “$\square$” to denote a box in λ’s Young diagram. In addition, we will use the following related pieces of notation.

- We write $[\lambda]$ for the set of all boxes in the diagram.

- Each box $\square \in [\lambda]$ is indexed by an ordered pair (i, j), where i is $\square$’s row and j is $\square$’s column in the English notation, counting rows from top-to-bottom and columns from left-to-right.

- We define the *content* of cell $\square$ to be $c(\square) := j - i$. Note that in the Russian diagram, the content of $\square$ is the x-coordinate of its center.

- We also define the *hook length* $h(\square)$ of $\square$ via the French notation: it is the number of cells directly to the right or above $\square$, including $\square$ itself; equivalently, it is $(\lambda_i - j) + (\lambda_j' - i) + 1$.

Definition 2.2.7. Having defined “content” for cells in a Young diagram, we may introduce some convenient notation (essentially from [OO98b]) that generalizes the standard notions of "falling factorial power" and "rising factorial power". First, for $z \in \mathbb{R}$ and $m \in \mathbb{N}$, recall the *falling factorial power*[1]

$$z^{\downarrow m} := z(z-1)(z-2)\cdots(z-m+1)$$

and *rising factorial power*

$$z^{\uparrow m} := z(z+1)(z+2)\cdots(z+m-1).$$

We generalize this notation to the case of an arbitrary partition $\lambda \vdash m$:

$$z^{\downarrow\lambda} := \prod_{\square\in[\lambda]}(z - c(\square)) \quad \text{and} \quad z^{\uparrow\lambda} := \prod_{\square\in[\lambda]}(z + c(\square)).$$

Definition 2.2.8. Given Young diagrams $\mu \vdash n-1$ and $\lambda \vdash n$, we write $\mu \nearrow \lambda$ if λ is a Young diagram formed by adding a box to the end of a row of μ. In other words, $\lambda_i = \mu_i + 1$ for some i and $\lambda_j = \mu_j$ for all $j \neq i$. Similarly, if μ and λ are Young diagrams with $|\mu| \leq |\lambda|$, we write $\mu \precsim \lambda$ if λ is a Young diagram formed by adding some number of boxes to each row of μ such that no two of the newly added boxes are in the same column. In other words, if μ has height d then

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \cdots \geq \lambda_d \geq \mu_d \geq \lambda_{d+1}.$$

2.2.2 Young tableaus

Definition 2.2.9. Let $\mathcal{A}$ be an *alphabet*; i.e., a totally ordered set. Most often we consider $\mathcal{A} = [d]$. A *word* is a finite sequence $(a_1, \ldots, a_n)$ of elements from $\mathcal{A}$. It is *weakly increasing* if $a_1 \leq a_2 \leq \cdots \leq a_n$ and *strongly (or strictly) increasing* if $a_1 < a_2 < \cdots < a_n$. For simplicity, we will often refer to a sequence as *increasing* if it is weakly increasing. If $\mathcal{D}$ is a probability distribution on $\mathcal{A}$ we write $\mathcal{D}^{\otimes n}$ to denote the probability distribution on words of length n given by drawing the letters independently from $\mathcal{D}$.

Definition 2.2.10. Given a word $a \in [d]^n$, there is an associated partition $\lambda \vdash n$ of length at most d called the *sorted type (or histogram)*. It is defined as follows: λ_i is the frequency of the i-th-most frequent letter in a, for $1 \leq i \leq d$. In other words, λ is the histogram of letter frequencies, sorted into nonincreasing order. For example, the sorted type of $(4,1,3,4,4,4,1,4) \in [4]^8$ is $(5,2,1,0) \vdash 8$.

Definition 2.2.11. Let $\lambda \vdash n$, and think of its Young diagram in the English notation. If each cell is filled with an element from some alphabet $\mathcal{A}$, we call the result a *Young tableau of shape* λ. The Young tableau is said to be *semistandard* if its entries are weakly increasing from left-to-right along rows and are strongly increasing from top-to-bottom along columns. If the rows are in fact strongly increasing, the Young tableau is called *standard*. Figure 2.3a gives an example of a standard tableau, and Figure 2.3b gives an example of a semistandard tableau. As shorthand, we sometimes write *SYT* for "standard Young tableau" and *SSYT* for "semistandard Young tableau".

[1] Or Pochhammer symbol, sometimes denoted $(z)_m$ or $z^{\underline{m}}$.

1	2	5	6	10	14
3	4	9	11		
7	8	12	13		
15	17	18			
16	19	20			

(a) A standard tableau.

1	1	1	2	3	3
2	2	3	3		
3	3	4	5		
5	6	6			
6	7	8			

(b) A semistandard tableau with alphabet $[8]$.

Figure 2.3: Two Young tableaus of shape $\lambda = (6, 4, 4, 3, 3)$.

Definition 2.2.12. For reasons we will see in Section 2.3.2, the number of standard Young tableaus of shape $\lambda \vdash n$ over alphabet $[n]$ is denoted $\dim(\lambda)$. It can be computed via the *Hook-Length Formula* of Frame, Robinson, and Thrall [FRT54] (see also [Sag01, Theorem 3.10.2]):

$$\dim(\lambda) = \frac{n!}{\prod_{\square \in [\lambda]} h(\square)}.$$

Definition 2.2.13. For reasons we will see in Section 2.4.2, the number of semistandard Young tableaus of shape $\lambda \vdash n$ over alphabet $[d]$ is denoted $\dim(\mathrm{V}_\lambda^d)$. It can be computed as

$$\dim(\mathrm{V}_\lambda^d) = \frac{(\dim \lambda) d^{\uparrow \lambda}}{|\lambda|!} = \prod_{1 \leq i < j \leq d} \frac{(\lambda_i - \lambda_j) + (j - i)}{j - i}.$$

The first expression is by combining Definition (2.2.12) with [Sta99, Corollary 7.21.4], and the second expression is the Weyl dimension formula [GW09, Equation (7.18)].

2.3 The irreducible representations of the symmetric group

In this section, we will introduce the representation theory of the symmetric group. Proofs of any results in this section can be found in [Sag01].

Our main goal will be to understand the irreducible representations of the symmetric group. To begin, we would like to determine the number of irreps of $\mathfrak{S}(n)$; by Theorem 2.1.19 this is equal to the number of conjugacy classes of $\mathfrak{S}(n)$. The conjugacy class of a permutation $\pi \in \mathfrak{S}(n)$ is determined by its *cycle type*, connecting representation theory to partitions.

Definition 2.3.1. We say that $\pi \in \mathfrak{S}(n)$ has cycle type $\lambda = (\lambda_1, \ldots, \lambda_k) \vdash n$ if π is the product of disjoint cycles of size $\lambda_1, \lambda_2, \ldots, \lambda_k$. (Note that π's length-1 cycles are included.)

Two permutations π_1, π_2 are conjugate iff there is another permutation σ such that $\sigma \pi_1 \sigma^{-1} = \pi_2$, and it can be checked that such a σ exists iff π_1 and π_2 have the same cycle type. As a result, the irreducible representations of $\mathfrak{S}(n)$ are in one-to-one correspondence with partitions $\lambda \vdash n$, and in fact there is a canonical way of associating the two sets.

Theorem 2.3.2. *The irreducible representations of $\mathfrak{S}(n)$ are indexed by partitions $\lambda \vdash n$. We will write κ_λ for the irrep indexed by λ and Sp_λ for the* Specht module, *the vector space κ_λ acts on. We will often abbreviate the character χ_{κ_λ} as χ_λ. We remark that χ_λ is known to take on only rational values; in particular, $\overline{\chi_\lambda} = \chi_\lambda$.*

In Section 2.3.1 below, we will show a natural way to associate the irreps of $\mathfrak{S}(n)$ with the partitions $\lambda \vdash n$. First, let us establish some notation.

Notation 2.3.3. We will use the following notation.

- If π has cycle type λ, then it is standard to write this as $\rho(\pi) = \lambda$. However we will use this notation extremely sparingly (and with warning) so as to preserve the symbol "ρ" for density matrices.

- To prevent using this notation, we adopt the following convention: *whenever a permutation π appears in a place where a partition λ is expected, the meaning is that λ should be the cycle type of π.*

- We also use the following standard notation:

  $$z_\lambda := \prod_{w \geq 1} (w^{m_w(\lambda)} \cdot m_w(\lambda)!).$$

  When $\lambda \vdash n$, the quantity $n!/z_\lambda$ is the number of permutations in $\mathfrak{S}(n)$ of cycle type λ, so z_λ^{-1} represents the probability that a uniformly random permutation in $\mathfrak{S}(n)$ has cycle type λ.

2.3.1 James submodule theorem

The James submodule theorem refers to a particular construction of the irreducible representations of the symmetric group in which partitions λ arise naturally. We will heavily follow the excellent exposition given in [HGG09], where these ideas are used for performing machine learning over permutations. For proofs, see [Sag01, Chapter 2].

Given a permutation $\pi \in \mathfrak{S}(n)$, the λ-representation constructed in the James submodule theorem provide "statistics of shape λ" about π. As an example, given a permutation $\pi \in \mathfrak{S}(6)$, the following is an example of a statistic of shape $(5, 1)$.

$$\text{Does } \pi \left\{ \begin{array}{|c|c|c|c|c|} \hline 1 & 3 & 4 & 5 & 6 \\ \hline \multicolumn{1}{|c|}{2} & \multicolumn{4}{c}{} \\ \cline{1-1} \end{array} \right\} = \left\{ \begin{array}{|c|c|c|c|c|} \hline 1 & 2 & 4 & 5 & 6 \\ \hline \multicolumn{1}{|c|}{3} & \multicolumn{4}{c}{} \\ \cline{1-1} \end{array} \right\} \text{ ?} \tag{2.6}$$

This should be interpreted as asking whether π maps the set $\{1, 3, 4, 5, 6\}$ to the set $\{1, 2, 4, 5, 6\}$ and the set $\{2\}$ to $\{3\}$. In other words, does $\pi(2) = 3$? If one were to ask this question for every pair T_1, T_2 of Young tableaus of shape $(5, 1)$, then one could recover the entire permutation π. However, receiving the answer to just a single question of this form gives very little information about π.

As a more interesting example, the following is an example of a statistic of shape $(3, 2, 1)$.

$$\text{Does } \pi \left\{ \begin{array}{|c|c|c|} \hline 1 & 3 & 5 \\ \hline \multicolumn{1}{|c|}{4} & \multicolumn{1}{c|}{6} & \\ \cline{1-2} \multicolumn{1}{|c|}{2} & \multicolumn{2}{c}{} \\ \cline{1-1} \end{array} \right\} = \left\{ \begin{array}{|c|c|c|} \hline 4 & 5 & 6 \\ \hline \multicolumn{1}{|c|}{1} & \multicolumn{1}{c|}{2} & \\ \cline{1-2} \multicolumn{1}{|c|}{3} & \multicolumn{2}{c}{} \\ \cline{1-1} \end{array} \right\} \text{ ?}$$

In other words, does π map $\{1,3,5\}$ to $\{4,5,6\}$, $\{4,6\}$ to $\{1,2\}$, and $\{2\}$ to $\{3\}$? This statistic provides strictly more information about π than the $(5,1)$-statistic from Equation (2.6). In general, we will think of "top-heavy" statistics as giving less information about π, where we think of λ as being more top-heavy than μ if $\lambda \trianglerighteq \mu$. As examples, the answer to a statistic of type (n) is always "yes", and therefore reveals no information about π, whereas a "yes" answer to a statistic of type (1^n) will determine π completely.

Let us formalize the λ-statistics as follows.

Definition 2.3.4. Let T_1, T_2 be n-box Young tableaus of the same shape which contain each number in $[n]$ exactly once. We call T_1 and T_2 *row equivalent* if the i-th rows of T_1 and T_2 contain the same set of numbers, for each i. A *Young tabloid* $\{T\}$ denotes the equivalence class corresponding to T under this equivalence relation. In other words, a Young tabloid of shape λ is equivalent to a method of partitioning the numbers in $[n]$ into sets of size $\lambda_1, \lambda_2, \ldots$ and assigning the sets to the corresponding rows of λ. Finally, given a permutation $\pi \in \mathfrak{S}(n)$, we define the operation $\pi\{T\}$ as applying π to each of these sets in the natural manner. Equivalently, $\pi\{T\} = \{\pi T\}$, where πT is the Young tableau resulting from applying π box-by-box to T.

Definition 2.3.5. The *permutation representation* corresponding to a Young diagram $\lambda \vdash n$ is the representation τ_λ of $\mathfrak{S}(n)$ with a row and a column for each Young tabloid $\{T\}$ of shape λ defined as

$$(\tau_\lambda(\pi))_{\{T_1\},\{T_2\}} = \begin{cases} 1 & \text{if } \pi\{T_2\} = \{T_1\}, \\ 0 & \text{otherwise.} \end{cases}$$

for each permutation $\pi \in \mathfrak{S}(n)$.

Examples:

- When $\lambda = (n)$, τ_λ is isomorphic to the trivial representation.
- When $\lambda = (n-1,1)$, τ_λ is isomorphic to μ, the dimension-n representation defined as

$$\mu(\pi)_{i,j} = \begin{cases} 1 & \text{if } \pi(j) = i, \\ 0 & \text{otherwise.} \end{cases}$$

- When $\lambda = (1^n)$, τ_λ is isomorphic to the regular representation.

Unfortunately, the permutation representations do not form a set of irreps for $\mathfrak{S}(n)$: for example, as we have seen before, when $n > 1$ the regular representation, and hence $\tau_{(1^n)}$, is reducible. More generally, τ_λ is a permutation matrix, and so it fixes the vector $\sum_{\text{tabloids } \{T\}} |\{T\}\rangle$. As a result, except in the case of $\lambda = (n)$, τ_λ is reducible. However, the *James submodule theorem* states that τ_λ is in fact isomorphic to κ_λ, so long as one first removes the "top-heavy irreps" from it.

Theorem 2.3.6 (James submodule theorem)**.** *There are numbers $K_{\lambda\mu}$ for partitions $\lambda, \mu \vdash n$ called the* Kostka numbers *such that*

$$\tau_\lambda \cong \bigoplus_{\mu \vdash n} K_{\lambda\mu} \kappa_\mu.$$

The Kostka numbers satisfy: (i) $K_{\lambda\lambda} = 1$ and (ii) $K_{\lambda\mu} \neq 0$ iff $\mu \trianglerighteq \lambda$. (In fact, $K_{\lambda\mu}$ has a combinatorial interpretation as the number of Young tableaus of shape λ and sorted type μ.)[2]

This gives a somewhat implicit definition of the irreps of $\mathfrak{S}(n)$. One can also explicitly construct κ_λ and the subspace of the permutation representation that it acts on, and the result is the Specht module; see [Sag01, Chapter 2.3] for details.

Examples

- When $\lambda = (n)$, κ_λ is the trivial representation.
- When $\lambda = (1^n)$, κ_λ is the determinant representation.

2.3.2 Young's orthogonal basis

In this section, we will give an explicit basis for each irrep of the symmetric group which results from the branching rule (see Section 2.1.4) for the symmetric group. For this branching rule, we will consider embedding the group $\mathfrak{S}(n-1)$ into $\mathfrak{S}(n)$ by mapping $\pi \in \mathfrak{S}(n-1)$ to (π, n). Under this mapping, $\mathfrak{S}(n-1)$ becomes a subgroup of $\mathfrak{S}(n)$.

Theorem 2.3.7. *Consider $\pi \in \mathfrak{S}(n-1)$ embedded into $\mathfrak{S}(n)$ through the mapping $\pi \mapsto (\pi, n)$. Given a partition $\lambda \vdash n$, the branching rule for κ_λ restricted to $\mathfrak{S}(n-1)$ is*

$$\kappa_\lambda{\downarrow}_{\mathfrak{S}(n-1)} \cong \bigoplus_{\substack{\mu \vdash n-1 \\ \text{s.t. } \mu \nearrow \lambda}} \kappa_\mu.$$

Note that it is multiplicity-free.

If we recursively apply the branching rule to $\mathfrak{S}(n-1), \mathfrak{S}(n-2), \ldots$, we get that

$$\kappa_\lambda{\downarrow}_{\mathfrak{S}(0)} \cong \bigoplus_{\lambda^{(n-1)} \nearrow \lambda} \quad \bigoplus_{\lambda^{(n-2)} \nearrow \lambda^{(n-1)}} \cdots \bigoplus_{\lambda^{(0)} \nearrow \lambda^{(1)}} \kappa_{\lambda^{(0)}}. \tag{2.7}$$

Note that $\mathfrak{S}(0)$ is a trivial group, and hence $\kappa_{\lambda^{(0)}}$ is the dimension-one trivial representation. As a result, the vector space Sp_λ, when restricted to $\mathfrak{S}(0)$, decomposes into dimension-one subspaces, one corresponding to each chain

$$\emptyset = \lambda^{(0)} \nearrow \lambda^{(1)} \nearrow \cdots \nearrow \lambda^{(n)} = \lambda. \tag{2.8}$$

There is a one-to-one correspondence between chains of this form and standard tableaus T of shape λ. As an example, the chain in (2.8) corresponds to the tableau T which has an "i" in the cell $\lambda^{(i)} \setminus \lambda^{(i-1)}$, for each $i \in [n]$.

Definition 2.3.8. Given $\lambda \vdash n$, the *Young orthogonal basis* for Sp_λ has a unit vector $|T\rangle$ for each standard tableau T of shape λ associated with the corresponding dimension-one subspace in (2.7).

We note that this definition uniquely specifies each basis vector $|T\rangle$ up to phase. The most important consequence for us is the following corollary, which explains the notation $\dim \lambda$.

Corollary 2.3.9. *The dimension of κ_λ is* $\dim(\lambda)$. *As a result, we will henceforth abbreviate* $\dim(\kappa_\lambda)$ *as* $\dim(\lambda)$.

[2] We note, incidentally, that the Kostka numbers are #P-complete to compute [Nar06].

2.4 The irreducible representations of the unitary and general linear groups

In this section, we will introduce the representation theory of the general linear and unitary groups. Proofs of any results in this section can be found in [GW09]. We will limit ourselves to studying *polynomial representations* of these groups.

Definition 2.4.1. A *polynomial representation* (μ, V) of $\mathrm{U}(d)$ is one in which the mapping $U \mapsto \mu(U)$ is polynomial, meaning that every matrix entry $\mu(U)_{ij}$ is a polynomial in the matrix entries of U. One can similarly define the polynomial representations of GL_d.

An example of a non-polynomial representation of GL_d is the representation given by $\mu_z(M) := [\det(M)^z]$ for negative $z \in \mathbb{Z}$. This is a *rational* representation.

Theorem 2.4.2. *The polynomial irreps of* $\mathrm{U}(d)$ *are indexed by partitions* λ *of height at most* d. *We will write them as* $(\pi_\lambda, \mathrm{V}_\lambda^d)$*, where* V_λ^d *is a vector space known as the* Weyl module. *(Throughout this thesis,* π *will also be used for permutations, but it will always be clear from context which of the two is intended.)*

Theorem 2.4.3. *Because* π_λ *is a polynomial irrep of* $\mathrm{U}(d)$*, it is well-defined for any* $d \times d$ *matrix. In this way, the polynomial irreps* $\mathrm{U}(d)$ *extend to form the polynomial irreps of* GL_d*.*

(See [Har05, Chapter 6] or [Wal14, Section 2.1] to see this point discussed further.)

As in the case of the representation theory of the symmetric group, in the representation theory of GL_d (and $\mathrm{U}(d)$) it helps to first understand the conjugacy classes of GL_d. For simplicity, let us restrict our attention to the set of $d \times d$ invertible, diagonalizable matrices $\mathrm{Diag} \subseteq \mathrm{GL}_d$. (We note that Diag is dense in GL_d and hence the behavior of any continuous representation of GL_d is uniquely determined by its behavior on Diag.) Two diagonalizable matrices $D_1, D_2 \in \mathrm{Diag}$ are conjugates of each other (with respect to GL_d) iff they are similar to each other, and for two diagonalizable matrices this is equivalent to them having the same set of d eigenvalues. As a result, the conjugacy classes of Diag (with respect to GL_d) are in one-to-one correspondence with the possible multisets of d eigenvalues $\{\alpha_1, \ldots, \alpha_d\}$.

From this, we can draw the following conclusions:

- For a matrix $D \in \mathrm{Diag}$, the character $\chi_{\pi_\lambda}(D)$ should only depend on D's eigenvalues. This is because D's conjugacy class depends only on its eigenvalues.

- If D has eigenvalues $\alpha_1, \ldots, \alpha_d$, then $\chi_{\pi_\lambda}(D)$ should be a *polynomial* in the α_i's, i.e.

  $$\chi_{\pi_\lambda}(D) = s_\lambda(\alpha_1, \ldots, \alpha_d),$$

  for some d-variate polynomial s_λ. This clearly holds for any diagonal matrix D, because (i) D's eigenvalues are on its diagonal, (ii) π_λ is a polynomial irrep, so its matrix entries are all polynomials in D's matrix entries, and (iii) $\chi_{\pi_\lambda}(D)$ is just the sum of the diagonal entries of $\pi_\lambda(D)$. For more general matrices in Diag, this holds because they are all similar to diagonal matrices.

- The polynomial $s_\lambda(\alpha_1, \dots, \alpha_d)$ should be *symmetric*. This is because there is no intrinstic ordering to a matrix's eigenvalues. (Equivalently, given a diagonal matrix D, one can arbitrarily permute the diagonal entries to form a similar matrix, which should have the same character value.)
- When varying λ over all partitions of height at most d, the s_λ polynomials should form a basis for the set of symmetric polynomials on d variables. (Though we have not shown a formal justification for this, an analogous statement holds in the finite group case (Theorem 2.1.19), and it is natural to hope that it might hold here too.)

It is not immediately clear why it should be natural to index a basis for the set of symmetric polynomials on d variables with Young diagrams of height at most d. In Section 2.4.1, we will show that this is indeed natural, and we will show that the s_λ polynomials are the *Schur polynomials*.

2.4.1 Symmetric polynomials

Definition 2.4.4. A d-variate polynomial $p : \mathbb{C}^d \to \mathbb{C}$ is *symmetric* if $p(x_\pi) = p(x)$ for every $x \in \mathbb{C}^d$, $\pi \in \mathfrak{S}(d)$, where $x_\pi := (x_{\pi(1)}, \dots, x_{\pi(d)})$.

The study of symmetric polynomials is a large field; see [Mac95] and [Sta99, Chapter 7] for standard references. The most basic symmetric polynomials are indexed by natural numbers.

Examples

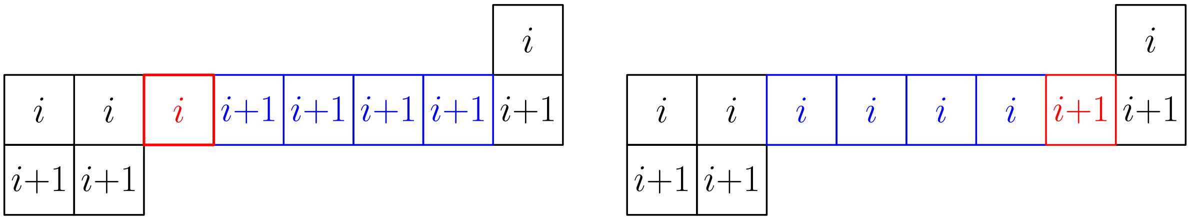
- For $m \in \mathbb{N}$, the m-th *elementary symmetric polynomial* is $e_m(x) = \sum_{i_1 < \cdots < i_m} x_{i_1} \cdots x_{i_m}$.
- For $m \in \mathbb{N}$, the m-th *power sum symmetric polynomial* is $p_m(x) = \sum_{i=1}^d x_i^m$.

More generally, it is useful to index symmetric polynomials by partitions corresponding to the types of the monomials in the polynomials.

- For λ a partition, say that a monomial $x_1^{a_1} \cdots x_d^{a_d}$ has *type* λ if the nonzero a_i's are some permutation of λ. Then the *monomial symmetric function* $m_\lambda(x)$ is the sum over all distinct monomials in the x_i's with type λ. For example, the monomial symmetric function corresponding to $\lambda = (2, 2, 1)$ in the case of $d = 3$ is
$$m_{(2,2,1)}(x_1, x_2, x_3) = x_1^2 x_2^2 x_3 + x_1^2 x_2 x_3^2 + x_1 x_2^2 x_3^2.$$
- For λ a partition, we define $e_\lambda(x) := \prod_{i=1}^{\ell(\lambda)} e_{\lambda_i}(x)$, and similarly for $p_\lambda(x)$.

Any symmetric polynomial will give the same coefficient to each monomial of the same type. Hence, the monomial symmetric functions form a linear basis for the set of all d-variate symmetric polynomials. It can be checked that the e_λ's and p_λ's also form a linear basis this set. For example, we have that
$$e_\lambda(x) = m_\lambda(x) + \{\text{linear combination of } m_\mu(x)\text{'s with } \mu \trianglerighteq \lambda\},$$
and so the fact that the e_λ's form a linear basis can be proved by induction.

For us, the most important set of polynomials are the following.



(a) A row of cells in T which are not paired off. This row consists of one i followed by four $i+1$'s.

(b) The same row in T', which consists of four i's followed by one $i+1$.

Figure 2.4: The involution $T \mapsto T'$, which swap the numbers of i's and $i+1$'s but fixes the remaining cells.

Definition 2.4.5. Let $x_1, \dots, x_d$ be indeterminates, typically standing for real numbers. Given $\lambda \vdash n$, the *Schur polynomial* $s_\lambda(x_1, \dots, x_d)$ is the degree-n homogeneous polynomial defined by $\sum_T x^T$, where the sum is over all semistandard tableaus of shape λ over alphabet $[d]$, and where

$$x^T := \prod_{i=1}^{d} x_i^{(\text{\# of occurrences of letter } i \text{ in } T)} .$$

Their importance comes from the fact that they are the characters of the irreps of the general linear group.

Theorem 2.4.6. *If M is a matrix with eigenvalues $\alpha_1, \dots, \alpha_d$, then $\chi_{\pi_\lambda}(M) = s_\lambda(\alpha_1, \dots, \alpha_d)$.*

We note that from Definition 2.4.5, it is not immediately obvious why the Schur polynomials are even symmetric.

Proposition 2.4.7. *The Schur polynomials s_λ of height at most d are symmetric and form a basis for the set of all d-variate symmetric polynomials.*

Proof. To prove symmetry, we reproduce here the proof from [Sta99, Theorem 7.10.2]. Because $\mathfrak{S}(d)$ is generated by transpositions, it suffices to show that $s_\lambda(x_1, \dots, x_d)$ is symmetric in coordinates i and $i+1$, for any $i \in [d-1]$. Consider a term x^T in the expansion of $s_\lambda(x)$. Some cells in T containing an i have a cell immediately below them containing an $i+1$. These pairs of cells are naturally "paired off". The remaining cells containing an i or an $i+1$ occur in rows containing a certain number r of i's followed by a certain number s of $i+1$'s. Consider the tableau T' in which we replace each such row with a sequence of s i's and r $i+1$'s. See Figure 2.4 for an illustration. Then T' has as many i's as T has $i+1$'s, and vice versa, and the two have the same number of every other letter. As the map $T \mapsto T'$ is an involution, and as $x^{T'}$ also appears as a term in $s_\lambda(x)$, this shows that $s_\lambda(x)$ is symmetric.

To show that Schur polynomials form a basis, note that s_λ contains a monomial of type λ, and every other monomial has a type μ such that $\mu \trianglelefteq \lambda$. As s_λ is symmetric, we can conclude that

$$s_\lambda(x) = m_\lambda(x) + \{\text{linear combination of } m_\mu(x)\text{'s with } \mu \trianglelefteq \lambda\} .$$

The fact that the s_λ's form a basis now follows from induction and the corresponding fact for the m_λ's. □

When $\ell(\lambda) > d$, there are no semistandard tableaus of shape λ over alphabet $[d]$. Thus, the sum $\sum_T x^T$ is the empty sum. This gives us the following fact about Schur polynomials:

Proposition 2.4.8. *Consider the Schur polynomial $s_\lambda(x_1, \dots, x_d)$. If $\ell(\lambda) > d$ then $s_\lambda \equiv 0$.*

Finally, though we will not need this, we note that the Schur polynomials are commonly defined as the ratio of a skew-symmetric polynomial and the Vandermonde determinant (see e.g. [Sta99, Theorem 7.15.1]), as follows.

Theorem 2.4.9. $s_\lambda(x_1, \dots, x_d) = \dfrac{\det\left(x_i^{d+\lambda_j-j}\right)_{ij}}{\det\left(x_i^{d-j}\right)_{ij}}$.

Following Stanley [Sta99, Corollary 7.17.5], we can actually give a definition of the symmetric group characters χ_μ in terms of the power sum and Schur polynomials:

Theorem 2.4.10. *In the context of Fourier analysis over the group $G = \mathfrak{S}(n)$, suppose $\mu \vdash n$ and $x \in \mathbb{C}^d$. Then $p_{(\cdot)}(x) := \pi \mapsto p_\pi(x)$ is a class function, and its Fourier coefficients are given by*

$$\widetilde{p_{(\cdot)}(x)}(\mu) = s_\mu(x).$$

Although this can be taken as an implicit definition of the characters χ_μ, we will more often think of the characters χ_μ as "known" and of Theorem 2.4.10 as letting us express the Schur polynomials in terms of the power sum polynomials.

2.4.2 The Gelfand-Tsetlin basis

In this section, we will give an explicit basis for each irrep of the general linear group which results from the branching rule (see Section 2.1.4) for the general linear group. For this branching rule, we will consider embedding the group $\mathrm{GL}_{d-1} \times \mathrm{GL}_1$ into GL_d by mapping $(M, \alpha) \in \mathrm{GL}_{d-1} \times \mathrm{GL}_1$ as follows:

$$(M, \alpha) \mapsto \begin{bmatrix} M & 0 \\ 0 & \alpha \end{bmatrix}.$$

Under this mapping, $\mathrm{GL}_{d-1} \times \mathrm{GL}_1$ becomes a subgroup of GL_d.

Theorem 2.4.11. *Given a partition λ of height d, the branching rule for π_λ restricted to $\mathrm{GL}_{d-1} \times \mathrm{GL}_1$ is*

$$\pi_\lambda{\downarrow}_{\mathrm{GL}_{d-1}\times\mathrm{GL}_1}(M, \alpha) \cong \bigoplus_{\substack{\mu \text{ of height } d-1 \\ \text{s.t. } \mu \precsim \lambda}} \pi_\mu(M) \cdot \alpha^{|\lambda \setminus \mu|}.$$

Note that it is multiplicity-free.

As in Section 2.3.2, we can recursively apply this branching rule, eventually restricting π_λ to

$$\underbrace{\mathrm{GL}_1 \times \cdots \times \mathrm{GL}_1}_{d \text{ copies}},$$

an element of which is of the form $(\alpha_1, \dots, \alpha_d)$. As a result, V_λ^d decomposes into a set of dimension-one subspaces corresponding to each chain

$$\emptyset = \lambda^{(0)} \precsim \lambda^{(1)} \precsim \cdots \precsim \lambda^{(d)} = \lambda, \tag{2.9}$$

where at every step of the chain, one picks up a multiplicative factor of $\alpha_i^{|\lambda^{(i)} \setminus \lambda^{(i-1)}|}$. There is a one-to-one correspondence between chains of this form and semistandard tableaus T of shape λ and alphabet $[d]$. As an example, the chain in (2.8) corresponds to the tableau T which has an "i" in every cell in the difference $\lambda^{(i)} \setminus \lambda^{(i-1)}$, for each $i \in [d]$. In summary, we get the decomposition

$$\pi_\lambda{\downarrow}_{\mathrm{GL}_1^{\times d}}(\alpha_1, \dots \alpha_d) \cong \bigoplus_{\substack{T \text{ of shape } \lambda, \\ \text{alphabet } [d]}} \alpha^T \cdot \pi_{\lambda^{(0)}}(e). \tag{2.10}$$

Definition 2.4.12. Given λ of height d, the *Gelfand-Tsetlin basis* for V_λ^d has a unit vector $|T\rangle$ (and a corresponding dimension-one subspace called the $|T\rangle$*-weight space*) for each semistandard tableau T of shape λ and alphabet $[d]$. By (2.10), the vector $|T\rangle$ satisfies

$$\langle T| \, \pi_\lambda(\mathrm{diag}(\alpha_1, \dots, \alpha_d)) \, |T\rangle = \alpha^T.$$

When the α_i's are sorted, this expression is maximized for the tableau T which has λ_1 ones in its first row, λ_2 twos in its second row, and so on. The vector $|T\rangle$ for this particular tableau is called the *highest weight vector*, and we will denote it by $|T_\lambda\rangle$.

Let us record a pair of consequences of the GZ basis. The first is that for any matrix of the form $M = \mathrm{diag}(\alpha_1, \dots, \alpha_d)$,

$$\chi_{\pi_\lambda}(M) = \sum_{\substack{T \text{ of shape } \lambda, \\ \text{alphabet } [d]}} \alpha^T = s_\lambda(\alpha),$$

exactly as guaranteed by Theorem 2.4.6. The second consequence is the following corollary.

Corollary 2.4.13. *The dimension of* V_λ^d *is equal to the number of semistandard tableaus of shape* λ *and alphabet* $[d]$. *This explains the notation* $\dim(\mathrm{V}_\lambda^d)$ *for this number. We note the equality*

$$s_\lambda(1^d) = \dim(\mathrm{V}_\lambda^d).$$

2.5 Schur-Weyl duality

Schur-Weyl duality concerns the relationships between the representations $\mathrm{P}(\pi)$ and $\mathrm{Q}(M)$ (Definition 2.1.2), one of the symmetric group and the other of the general linear group. These representations both act on $(\mathbb{C}^d)^{\otimes n}$, and they commute with each other, so we may consider the representation of the product group $\mathfrak{S}(n) \times \mathrm{GL}_d$ given by $\mu(\pi, M) := \mathrm{P}(\pi)\mathrm{Q}(M)$. (That they commute is important for this to be a representation, as

$$\begin{aligned}\mu(\pi_1, M_1)\mu(\pi_2, M_2) &= \mathrm{P}(\pi_1)\mathrm{Q}(M_1)\mathrm{P}(\pi_2)\mathrm{Q}(M_2) \\ &= \mathrm{P}(\pi_1)\mathrm{P}(\pi_2)\mathrm{Q}(M_1)\mathrm{Q}(M_2) = \mathrm{P}(\pi_1\pi_2)\mathrm{Q}(M_1M_2) = \mu(\pi_1\pi_2, M_1M_2).)\end{aligned}$$

The following standard proposition (cf. [Har05]) shows that any time a representation is constructed in this manner, then it decomposes in a particularly nice way.

Proposition 2.5.1. *Suppose (μ_1, V) and (μ_2, V) are representations of two groups G_1 and G_2 which act on the same vector space and commute. Suppose further that μ_1 and μ_2 are completely reducible. Then the representation of the product group $G_1 \times G_2$ given as $\mu(g_1, g_2) := \mu_1(g_1)\mu_2(g_2)$ decomposes as*

$$\mu \cong \bigoplus_{\nu_1 \in \widehat{G}_1, \nu_2 \in \widehat{G}_2} m_{\nu_1,\nu_2} \cdot \nu_1 \otimes \nu_2,$$

for some multiplicities m_{ν_1,ν_2}.

Proof. Because μ_1 is completely reducible, it decomposes into a direct sum of irreps. For simplicity, let us assume that this decomposition occurs in the standard basis, i.e. that

$$\mu_1 = \bigoplus_{\nu_1 \in \widehat{G}_1} m_{\nu_1} \cdot \nu_1 = \bigoplus_{\nu_1 \in \widehat{G}_1} I_{m_{\nu_1}} \otimes \nu_1 = \sum_{\substack{\nu_1 \in \widehat{G}_1 \\ i \in [m_{\nu_1}]}} |\nu_1\rangle\langle\nu_1| \otimes |i\rangle\langle i| \otimes \nu_1, \tag{2.11}$$

for some multiplicities m_{ν_1}. On the other hand, μ_2 will in general look like

$$\mu_2(g_2) = \sum_{\substack{\nu_1, \nu_1' \in \widehat{G}_1 \\ i \in [m_{\nu_1}], j \in [m_{\nu_1'}]}} |\nu_1\rangle\langle\nu_1'| \otimes |i\rangle\langle j| \otimes M_{\nu_1,\nu_1',i,j}(g_2), \tag{2.12}$$

for some $\dim(\nu_1) \times \dim(\nu_1')$ matrices $M_{\nu_1,\nu_1',i,j}(g_2)$. Now we will use the fact that μ_1 and μ_2 commute. Expanding out the equation $\mu_1(g_1)\mu_2(g_2) = \mu_2(g_2)\mu_1(g_1)$ using (2.11) and (2.12), we get that for each ν_1, ν_1', i, j,

$$\nu_1(g_1) \cdot M_{\nu_1,\nu_1',i,j}(g_2) = M_{\nu_1,\nu_1',i,j}(g_2) \cdot \nu_1'(g_1).$$

Thus, $M_{\nu_1,\nu_1',i,j}(g_2)$ is an intertwining operator between ν_1 and ν_1', and Schur's lemma tells us that it is zero unless $\nu_1 = \nu_1'$, in which case it is a multiple of the identity. (This multiple may depend on ν_1, i, j, and g_2.) Hence,

$$\mu_2(g_2) = \sum_{\nu_1 \in \widehat{G}_1} |\nu_1\rangle\langle\nu_1| \otimes M_{\nu_1}(g_2) \otimes I_{\dim(\nu_1)} = \bigoplus_{\nu_1 \in \widehat{G}_1} M_{\nu_1}(g_2) \otimes I_{\dim(\nu_1)}, \tag{2.13}$$

for some matrices $M_{\nu_1}(g_1)$. Multiplying (2.11) with (2.13), we get that

$$\mu(g_1, g_2) = \bigoplus_{\nu_1 \in \widehat{G}_1} M_{\nu_1}(g_2) \otimes \nu_1(g_1).$$

The lemma now follows by noting that from (2.13), $M_{\nu_1}(g_2)$ is a representation of G_2 and hence decomposes into a direct sum of irreps of G_2. □

Given this proposition, it remains in our case to determine the multiplicities. In the cases of both $\mathfrak{S}(n)$ and GL_d, the irreps are indexed by partitions λ, and Schur-Weyl duality states that $\kappa_{\lambda_1} \otimes \pi_{\lambda_2}$ occurs with multiplicity zero unless $\lambda_1 = \lambda_2$, in which case it occurs with multiplicity one.

Theorem 2.5.2 (Schur-Weyl duality)**.** *With respect to the representation μ of $\mathfrak{S}(n) \times \mathrm{GL}_d$, we have the following unitary equivalence.*

$$(\mathbb{C}^d)^{\otimes n} \cong \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} \mathrm{Sp}_\lambda \otimes \mathrm{V}_\lambda^d.$$

A proof of this can be found in [GW09]. The condition that the partition be of size n comes from $\mathfrak{S}(n)$, and the condition that it be of height at most d comes from GL_d.

2.6 Quantum algorithms from representation theory

Let us finally discuss how to use representation theory, and in particular Schur-Weyl duality, to build algorithms for quantum state learning. By Schur-Weyl duality, there is a unitary change-of-basis U_{Schur} on $(\mathbb{C}^d)^{\otimes n}$ such that for any $\pi \in \mathfrak{S}(n)$, $M \in \mathrm{GL}_d$, we have that

$$U_{\mathrm{Schur}} \mathrm{P}(\pi) \mathrm{Q}(M) U_{\mathrm{Schur}}^\dagger = \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} \kappa_\lambda(\pi) \otimes \pi_\lambda(M). \tag{2.14}$$

The matrix U_{Schur} changes the standard basis into something called the *Schur basis*.

Suppose ρ is the state we are trying to learn, and we are given the n copies $\rho^{\otimes n}$. Then applying (2.14) with $\pi = e$ (the identity permutation) and $M = \rho$, and recalling that $\mathrm{Q}(\rho) = \rho^{\otimes n}$, we get that

$$U_{\mathrm{Schur}} \rho^{\otimes n} U_{\mathrm{Schur}}^\dagger = \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} I_{\dim \lambda} \otimes \pi_\lambda(\rho). \tag{2.15}$$

(That we can apply Q to ρ depends on ρ having d nonzero eigenvalues $\alpha_1, \ldots, \alpha_d$. However, if some of these eigenvalues are zero, then we can write ρ as the limit of a sequence of invertible matrices, and by continuity (2.15) holds. We also use here that π_λ is polynomial, and hence is well-defined for any matrix.)

What (2.15) says is that there is a unitary transformation—an allowable quantum mechanical operation—which is independent of ρ and which causes ρ to assume a particularly nice, block diagonal form. By item 1 of Proposition 1.2.7, we may now without loss of generality perform a projective measurement according to these blocks, and we may or may not choose to perform a second measurement following this.

Definition 2.6.1. There are two broad categories of Schur-Weyl-based quantum measurements:

- Write Π_λ for the projector onto the λ-subspace given by Schur-Weyl duality, i.e. the subspace $\mathrm{Sp}_\lambda \otimes \mathrm{V}_\lambda^d$. Then *weak Schur sampling* refers to the projective measurement $\{\Pi_\lambda\}_{\lambda \vdash n, \ell(\lambda) \leq d}$.

- *Strong Schur sampling* refers to the process of first performing weak Schur sampling, receiving a measurement outcome λ, and then performing a subsequent measurement in the subspace corresponding to Π_λ.

Given n copies of a mixed state ρ with spectrum α, weak Schur sampling yields the partition $\lambda \vdash n$ with probability

$$\operatorname{tr}(\Pi_\lambda \rho^{\otimes n}) = \operatorname{tr}(I_{\dim \lambda} \otimes \pi_\lambda(\rho)) = \dim(\lambda) \cdot s_\lambda(\alpha). \tag{2.16}$$

The fact that $\chi_{\pi_\lambda(\rho)} = s_\lambda(\alpha)$ was only established in the case that the α_i's are all nonzero, but again we can appeal to a continuity argument to extend this to all mixed states ρ. Recalling that the eigenvalues of ρ form a probability distribution, we will spend much of this thesis analyzing the following distribution.

Definition 2.6.2. Given a probability distribution $\alpha = (\alpha_1, \ldots, \alpha_d)$ and an integer $n \geq 0$, the *Schur-Weyl distribution* $\mathrm{SW}^n(\alpha)$ is the probability distribution on partitions in which $\lambda \vdash n$ has probability $\dim(\lambda) \cdot s_\lambda(\alpha)$. In the case when α is the uniform distribution, we sometimes write SW^n_d instead of $\mathrm{SW}^n(\alpha)$. In addition, if ρ is a mixed state with spectrum α, we will sometimes write SW^n_ρ for $\mathrm{SW}^n(\alpha)$.

Now we can state the reason why this representation theoretic approach is so powerful: not only is weak Schur sampling without loss of generality, it is often optimal.

Theorem 2.6.3. *Suppose we are interested in computing a property of ρ which depends only on its spectrum α. Then weak Schur sampling is the optimal measurement on $\rho^{\otimes n}$.*

In other words, if we have an algorithm Alg *for computing the property which has failure rate β on any matrix ρ, then there is a similar algorithm for doing so using only weak Schur sampling followed by classical postprocessing.*

Proof. Suppose that prior to giving Alg the input $\rho^{\otimes n}$, we first "averaged" it out by random permutations and unitaries:

$$\rho_{\mathrm{avg}} := \mathop{\mathbf{E}}_{\substack{\boldsymbol{\pi} \sim \mathfrak{S}(n) \\ \boldsymbol{U} \sim \mathrm{U}(d)}} \left[\mathrm{P}(\boldsymbol{\pi})\mathrm{Q}(\boldsymbol{U})\rho^{\otimes n}\mathrm{Q}(\boldsymbol{U}^\dagger)\mathrm{P}(\boldsymbol{\pi}^{-1})\right] = \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} (\boldsymbol{U}\rho\boldsymbol{U}^\dagger)^{\otimes n}.$$

Then ρ_{avg} is an mixture of states with spectrum α, and hence if we feed ρ_{avg} into Alg, it will properly compute the property of α with probability at least $1 - \beta$.

Now we claim that whatever the measurement Alg performs on ρ_{avg}, it is only better for it to measure with respect to the Π_λ projectors. To see this, note that by Equations (2.14) and (2.15),

$$U_{\mathrm{Schur}} \rho_{\mathrm{avg}} U_{\mathrm{Schur}}^\dagger = \bigoplus_{\substack{\lambda \vdash n \\ \ell(\lambda) \leq d}} I_{\dim \lambda} \otimes \underbrace{\mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \left[\pi_\lambda(\boldsymbol{U})\pi_\lambda(\rho)\pi_\lambda(\boldsymbol{U})^\dagger\right]}_{\rho_\lambda}. \tag{2.17}$$

Writing ρ_λ for the matrix given in the expectation, we see that $\pi_\lambda(U)\rho_\lambda\pi_\lambda(U)^\dagger = \rho_\lambda$ for any $U \in \mathrm{U}(d)$ by the properties of the Haar measure. Schur's lemma then tells us that ρ_λ is some multiple of the identity matrix. As a result, ρ_{avg} is block diagonal corresponding to the Π_λ projectors, and so by item 2 of Proposition 1.2.7 the projective measurement with these projectors is optimal.

Finally, we note that WSS when performed on ρ_{avg} gives the same distribution of outcomes as when performed on $\rho^{\otimes n}$. This is because

$$\operatorname{tr}(\Pi_\lambda \rho_{\mathrm{avg}}) = \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \operatorname{tr}((\boldsymbol{U}\rho\boldsymbol{U}^\dagger)^{\otimes n}) = \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \dim(\lambda) \cdot s_\lambda(\alpha) = \dim(\lambda) \cdot s_\lambda(\alpha) = \operatorname{tr}(\Pi_\lambda \rho^{\otimes n}),$$

where we have twice used Equation (2.16). In total, WSS (followed by classical postprocessing) performs at least as well as Alg does on $\rho^{\otimes n}$. □

To our knowledge, this theorem first appeared in [KW01] (cf. [CHW07, Lemma 5] and [MdW13, Lemma 20]).

Chapter 3

Longest increasing subsequences and the RSK algorithm

In this chapter, we focus on two probability distributions on partitions which arise naturally in the study of representation theory. The first is the Plancherel distribution, inspired by the decomposition into irreps of the regular representation.

Definition 3.0.1 (Definition 2.1.13 restated)**.** For a finite group G, the *Plancherel distribution* is the probability distribution on irreps in which $\mu \in \widehat{G}$ has probability $(\dim \mu)^2/|G|$. In the case when G is $\mathfrak{S}(n)$, we denote this distribution by Planch_n.

The second probability distribution is the Schur-Weyl distribution from Definition 2.6.2. As we will see below (Corollary 3.3.4) these two distributions are related; in particular, the Plancherel distribution for $\mathfrak{S}(n)$ is a special case of the Schur-Weyl distribution.

The goal of this chapter is to connect these two distributions to the seemingly-unrelated topic of *longest increasing subsequences of random words.* A key player in this connection is the *Robinson-Schensted-Knuth (RSK) algorithm*, a combinatorial algorithm which takes as input a word w and outputs the "higher-order longest increasing subsequence statistics" of w in the form of a Young diagram λ. As we show below, when the input $\boldsymbol{w}$ is selected from a suitably chosen probability distribution, then the output $\boldsymbol{\lambda}$ of the RSK algorithm will be distributed according to either Planch_n or $\mathrm{SW}^n(\alpha)$.

The outline of this chapter is as follows:

- In Section 3.1 we will introduce Patience sorting, a simple version of the RSK algorithm, and in Section 3.2 we will generalize it to the full RSK algorithm.
- In Section 3.3, we will show that the RSK algorithm, when applied to suitably chosen random words, generates the Plancherel and Schur-Weyl distributions. We then spend Section 3.4 briefly investigating some properties of the Schur-Weyl distribution.
- In Section 3.5, we will survey the history of results about longest increasing subsequences of random permutations. In Section 3.6, we will show how these results have been generalized to study the Plancherel distribution, and in Section 3.7, we will show how these results were further generalized to the Schur-Weyl distribution.

- In Section 3.8, we will introduce one of the main tools for analyzing the Plancherel and Schur-Weyl distributions: Kerov's *algebra of observables*. This is a certain set of polynomials which allow us to analyze the "moments" of random Young diagrams. We will also give a taste here of some of the down-and-dirty work involved when using these polynomials.

3.1 Patience sorting

Patience sorting was originally conceived of as a method of sorting a deck of cards, though it has since found use as an algorithm for computing the longest increasing subsequence of a word. It was originally introduced by Mallows [Mal62], who credits it to A. S. C. Ross, and was rediscovered by Floyd [AD99] and later by Hammersley [Ham72]. Though it is essentially a special case of the RSK algorithm, it was apparently developed independently [Mal63]. See [AD99] for a survey of patience sorting and related topics and [Lan07] for some extensions.

Patience sorting works by first arranging the cards into a sequence of piles (usually, this step alone is referred to as "patience sorting") and then performing a simple postprocessing step on the piles. We will represent the set of piles using a semistandard row tableau, where the number in a given cell represents the top card in that pile.

Definition 3.1.1. A semistandard *row* tableau T is an SSYT whose shape is (ℓ) for some integer ℓ.

Each step of patience sorting is given by the following set of instructions: draw a card (the "current card") from the deck, scan the piles from left to right, and place the current card on top of the first pile whose top card is larger than the current card. If no such pile exists, then create a new pile with the current card at the top. This operation is called *insertion*; in SSYT form, it is given as follows.

Definition 3.1.2. Given a semistandard row tableau T with alphabet $\mathcal{A}$, we *insert* a letter $a \in \mathcal{A}$ into T as follows:

1. Find the leftmost cell of T containing a letter b such that $b > a$. Remove b from the cell and replace it with a.

2. If no such cell exists, then append a new cell containing a to the end of the row.

The letter b is said to have been *bumped*.

In total, then, the algorithm is given as follows.

Definition 3.1.3. Given an n-letter word w with alphabet $\mathcal{A}$, the *patience sorting algorithm* works as follows:

1. Initialize T to be an empty tableau.

2. For $i = 1 \ldots n$, insert w_i into T. Do nothing with the bumped letters.

3. Output T.

Having formed the piles in the way, note that deck's smallest card is on the top of the first pile. Removing this card, the deck's second smallest card is now on the top of the first *two* piles. One can then repeatedly remove the smallest card from the top of the piles to sort the deck. The efficiency of this last step is governed by the number of piles created by patience sorting, and this number was determined by Schensted in [Sch61].

Theorem 3.1.4 (Schensted's theorem)**.** *Let $w \in [d]^n$, and set T to be the result of applying patience sorting to w. Then the length of T is equal to* $\mathrm{LIS}(w)$.

This theorem can be proved easily by a direct analysis; see for example [Rom14, Lemma 1.7]. For a Computer-Science-based approach, consider Algorithm 1, a natural recursive algorithm for computing $\mathrm{LIS}(w)$. (To compute $\mathrm{LIS}(w)$, one simply needs to find the maximum ℓ such that $\texttt{RecursiveLIS}(w, \ell) < \infty$.) This recursive algorithm can be turned into a dynamic program using memoization, and Proposition 3.1.5 shows that patience sorting maintains the data structure used to do this.

```
Function RecursiveLIS($w, \ell$)
    Input  : a word $w \in \mathcal{A}^n$ and a length $\ell \in \mathbb{Z}^{\geq 0}$
    Output: a letter $a \in \mathcal{A}$ such that
            $$a = \min_{\substack{\text{increasing subsequences } s \text{ in } w \\ \text{of length } \ell}} (\text{rightmost letter in } s)$$
    if $n = 0$ then return $\infty$;
    a ← RecursiveLIS($w[1..n-1]$, $\ell - 1$);
    b ← RecursiveLIS($w[1..n-1]$, $\ell$);
    if $a \leq w_n < b$ then return $w_n$;
    else return $b$;
```

Algorithm 1: A recursive algorithm for computing $\mathrm{LIS}(w)$.

Proposition 3.1.5. *Let $w \in [d]^n$, and set T to be the result of applying patience sorting to w. Then the letter a in the $(1, \ell)$-th cell of T is* $\texttt{RecursiveLIS}(w, \ell)$.

The proof of this statement is trivial. From this, Schensted's theorem follows immediately.

The algorithmic complexity of computing $\mathrm{LIS}(w)$ has been studied in various models. In [Fre75], Fredman considered the patience sorting algorithm in which the insertion step is implemented by binary searching through the row tableau T for the letter b. He showed (i) that this algorithm uses $n \log n - n \log\log n + O(n)$ comparisons and (ii) that this number of comparisons is optimal for all algorithms, by reducing the problem of sorting w to the problem of computing $\mathrm{LIS}(w)$. In the RAM model, this algorithm runs in time $O(n \log n)$ in general and time $O(n \log k)$ if $k = \mathrm{LIS}(w)$. Further improvements are possible in the case when w is a permutation of $[n]$. Here, Hunt and Szymanski [HS77] improved the running time of patience sorting to $O(n \log\log n)$ by replacing the binary search step with a priority queue data structure of van Ernde Boas [vEB77] (see also [BS00] for a simplified write-up of this result). Following this, Crochemore and Porat [CP10] showed that a careful implementation of this algorithm would yield a runtime of $O(n \log\log k)$ if $k = \mathrm{LIS}(w)$. Other works have considered the complexity of approximating LIS(w) [SS10], its streaming complexity [LNVZ06, GJKK07, SW07, EJ08, GG10], and its communication complexity [SW07].

3.2 The Robinson-Schensted-Knuth algorithm

The RSK algorithm is a generalization of patience sorting in which the letter "bumped" during an insertion is recursively inserted into the second row of the tableau rather than discarded. This recursive insertion is called *Schensted insertion* (or *row-insertion*).

Definition 3.2.1. Given an SSYT T with alphabet $\mathcal{A}$ and a letter $a \in \mathcal{A}$, *Schensted insertion* is the following procedure:

1. Insert a into the first row of T.

2. If a letter b was bumped, then recursively Schensted insert b into the subtableau of T formed by deleting the first row.

Fact 3.2.2. *If T is an SSYT with alphabet $\mathcal{A}$ and we Schensted insert $a \in \mathcal{A}$ into it, then the resulting tableau is also an SSYT with alphabet $\mathcal{A}$.*

Proof. We will prove this by induction, the base case being that T begins as an SSYT. For the inductive step, assume that the letter recursively inserted into row $i-1$ produced an SSYT T'. Consider the letter a which is inserted into row i of T'. It is easy to see that this row remains weakly increasing, and thus to show that the resulting tabelau is SSYT, it remains to show that the column that a is inserted into remains strongly increasing. Write $b_1 < \ldots < b_m$ for the letters in this column, for some m. Then Schensted insertion replaces b_i with a and leaves the rest unchanged, and so we need only check that $b_{i-1} < a < b_{i+1}$.

Because a bumped b_i, we know that $a < b_i$, and so $a < b_i < b_{i+1}$. For the other inequality, consider the letter a' that was inserted into row $i-1$ in the previous step. The cell it was inserted into was either b_{i-1}'s cell (in which case $a' = b_{i-1}$) or to its right. Otherwise, as this cell originally contained a, the cell below it contains a letter strictly larger than a, but this is a contraction because a is inserted in row i to the right of this column. As a result, we have that $b_{i-1} \leq a'$, and since $a' < a$ we have that $b_{i-1} < a$, completing the proof. □

Definition 3.2.3. Given an n-letter word w with alphabet $\mathcal{A}$, the *Robinson-Schensted-Knuth (RSK) algorithm* works as follows:

1. Initialize P, Q to be empty tableaus.

2. For $i = 1 \ldots n$:

   (a) Schensted insert w_i into P.

   (b) Add a new cell to Q in the same position as the newly added cell in P. Fill it with the number i.

3. Output RSK(w) := (P, Q).

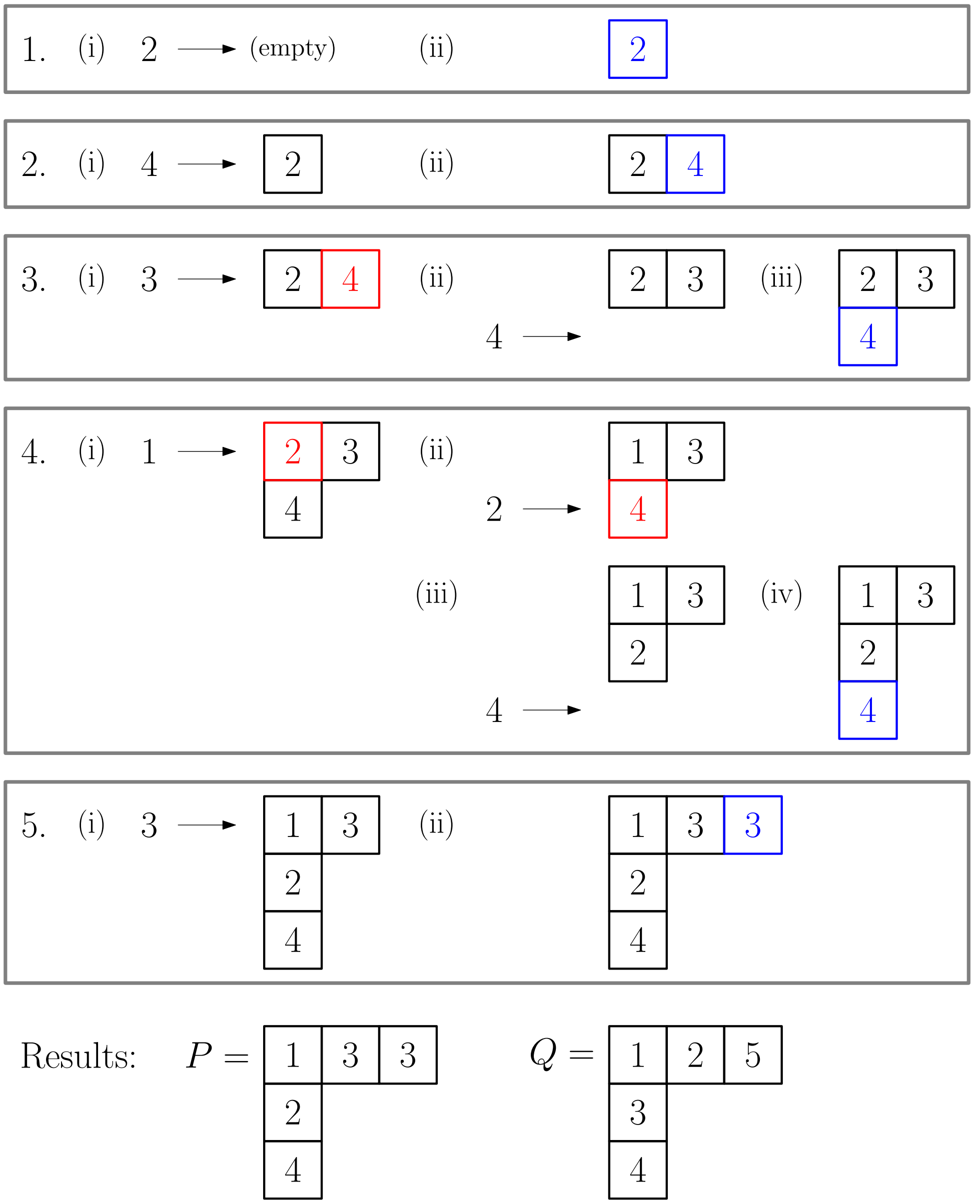

1. (i) 2 (empty) (ii) 2
2. (i) 4 2 (ii) 2 4
3. (i) 3 2 4 (ii) 2 3 4 (iii) 2 3 4
4. (i) 1 2 3 4 (ii) 1 3 2 4 (iii) 1 3 2 4 (iv) 1 3 2 4
5. (i) 3 1 3 2 4 (ii) 1 3 3 2 4
Results: P = 1 3 3 2 4
Q = 1 2 5 3 4

Figure 3.1: The construction of the insertion tableau P from the RSK algorithm applied to the word $w = 24313$. Each large gray box corresponds to a Schensted insertion, each red box corresponds to a bumped letter, and each blue box corresponds to a newly created cell.

P is known as the *insertion* tableau and Q is known as the *recording* tableau. We will write $\mathrm{shRSK}(w)$ for the Young diagram given by the common shape of P and Q.

An example execution of the RSK algorithm is given in Figure 3.1. The RSK algorithm is named after its three inventors: Gilbert de Beauregard Robinson, a mathematician, Craige (Ea Ea) Schensted, a physicist, and Donald Knuth, a computer scientist. The algorithm first appeared in somewhat opaque form in the work of Robinson [Rob38], who only considered the case when w is a permutation. Schensted independently discovered it in [Sch61], extended it to hold for general words w, and discovered the link to longest increasing subsequences (Schensted's theorem). Finally, Knuth [Knu70] gave a further generalization of the algorithm to the case when w is not a word at all but a matrix of nonnegative integers.[1]

The first row of the RSK algorithm carries out the patience sorting algorithm, and so if $\lambda = \mathrm{shRSK}(w)$ then Schensted's theorem says that $\lambda_1 = \mathrm{LIS}(w)$. In 1974, Greene [Gre74] proved the following significant generalization of this fact.

Theorem 3.2.4 (Greene's Theorem)**.** *Let $w \in [d]^n$, and set $(P, Q) = \mathrm{RSK}(w)$. Then for each $k \geq 1$, $\lambda_1 + \ldots + \lambda_k$ equals the length of the longest k disjoint increasing subsequences of w. Similarly, for each $k \geq 1$, $\lambda'_1 + \ldots + \lambda'_k$ equals the length of the longest k disjoint strictly decreasing subsequences of w.*

An example of Greene's Theorem is given in Figure 3.2 (which is Figure 1.3 reprinted). Given any word w, one can always form k increasing subsequences consisting of the top k most frequently occurring letters in w. This yields the following corollary.

Corollary 3.2.5. *Let $\lambda = \mathrm{shRSK}(w)$ and let μ be w's sorted histogram. Then $\lambda \trianglerighteq \mu$.*

Hence, $\mathrm{shRSK}(w)$ can be viewed as a "top-heavy" estimate of w's sorted histogram.

Perhaps the single most important feature of this algorithm is that it establishes a bijection between words w and pairs (P, Q) of insertion and recording tableaus. This bijection is known as the *RSK correspondence*.

Theorem 3.2.6 (RSK correspondence)**.** *The RSK algorithm gives the following two bijections:*

1. *between permutations $\pi \in \mathfrak{S}(n)$ and pairs (P, Q), where P and Q have the same shape $\lambda \vdash n$ and are both SYT.*

2. *between words $w \in [d]^n$ and pairs (P, Q), where P and Q have the same shape $\lambda \vdash n$, P is SSYT with alphabet $[d]$, and Q is SYT.*

Furthermore, the SSYT P in item 2 always has the same multiset of letters as w has.

[1] Though we will only ever use the algorithm in the case when w is a word, and hence Schensted's construction is sufficient, it is common in the literature to still refer to the simpler forms of the algorithm as the RSK algorithm, and we will do so here. That said, some works do distinguish between the RS and the RSK algorithms.

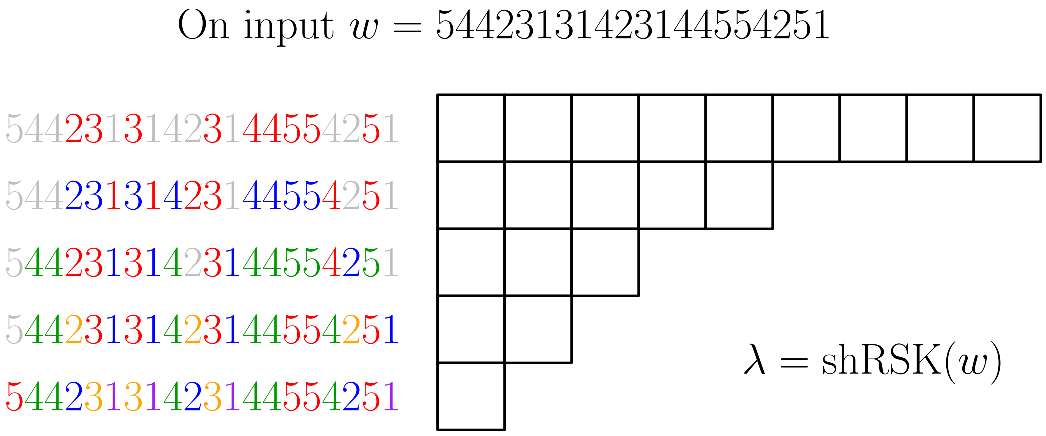


Figure 3.2: An illustration of Greene's theorem. On each row k, we have highlighted the letters in the longest set of k disjoint increasing subsequences. The number of letters highlighted should be equal to $\lambda_1 + \ldots + \lambda_k$, the number of boxes in the first k rows of λ.

One reason this is useful is that it allows us to count the number of words with a certain property by instead counting the number of pairs (P, Q) with a certain property. For example, if we want to count the number of words w with $\mathrm{LIS}(w) \geq \ell$, we need only count the number of pairs (P, Q) whose common shape λ has $\lambda_1 \geq \ell$, and this might look easier given our explicit formulas for counting tableaus, for example Definitions 2.2.12 and 2.2.13. In addition, the RSK correspondence is the reason we are able to relate the distributions **D1** and **D2** from above; for this, see Theorem 3.3.2 below.

Proof of Theorem 3.2.6. We will give the proof of item 2; the proof of item 1 is almost identical. First, we show that RSK does indeed map words to pairs of SSYT and SYT tableaus. Given $w \in [d]^n$, set $(P, Q) := \mathrm{RSK}(w)$. By construction, P and Q have the same shape $\lambda \vdash n$. Furthermore, by Fact 3.2.2, P is always SSYT with alphabet $[d]$. Finally, in the construction of Q each new cell is added southeast of its adjacent cells and contains the largest number yet added, meaning that Q is always SYT.

Now, given a pair (P, Q) we show how to recover the unique word w mapped to it by RSK. We will do so by "reversing" the steps of the RSK algorithm to recover the letter w_n and the insertion and recording tableaus (P', Q') prior to the Schensted insertion of w_n. Applying this reversing procedure recursively will recover the entire word w.

The process of Schensted inserting w into P begins with an insertion followed by a series of alternating bumpings and insertions until a letter is inserted into a new cell at the end of a row. By construction, this is the cell containing the number n in the recording tableau Q. Let a be the letter in the corresponding cell of P, and delete this cell from both P and Q. For a to have been bumped, a letter b must have been inserted into the previous row, and a must have been the first letter in that row larger than b. Hence, b is the last letter in that row which is smaller than a. To "unbump" a, then, remove b from its cell and replace it with a. Continuing this unbumping procedure up the Young tableau, we arrive at the letter which was inserted into the first row, w_n, and we have reverted (P, Q) to (P', Q'). □

3.3 Random words and permutations

In this section, we will consider the following two distributions on random words.

Definition 3.3.1. A *random permutation of length* n is a uniformly random permutation $\boldsymbol{\pi} \in \mathfrak{S}(n)$.

For the purposes of longest increasing subsequences, a random permutation $\boldsymbol{\pi}$ may be viewed as a random word $\boldsymbol{w} \sim \mathsf{Unif}_d^{\otimes n}$ in the limit as $d \to \infty$ (in which case $\boldsymbol{w}$ will have always have n distinct letters). See Corollary 3.3.4 for this intuition formalized. For the same reasons, we may also view a random permutation $\boldsymbol{\pi}$ as a random word $\boldsymbol{w} \sim [0,1]^{\otimes n}$, where each letter of $\boldsymbol{w}$ is a uniformly random element of $[0,1]$.

We are now able to state the main result of this section: using the loose terminology of the introduction, **D1** = **D2**.

Theorem 3.3.2. *Let $\lambda \vdash n$. Let $\boldsymbol{\pi}$ be a random permutation of length n. Then*

$$\mathbf{Pr}[\mathrm{shRSK}(\boldsymbol{\pi}) = \lambda] = \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{Planch}_n}[\boldsymbol{\lambda} = \lambda].$$

Let $\boldsymbol{w}$ be an n-letter α-random word. Then

$$\mathbf{Pr}[\mathrm{shRSK}(\boldsymbol{w}) = \lambda] = \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)}[\boldsymbol{\lambda} = \lambda].$$

Proof. There are $n!$ permutations of length n. By item 1 of the RSK correspondence, the number of permutations $\pi \in \mathfrak{S}(n)$ for which $\mathrm{shRSK}(\pi) = \lambda$ is equal to

$$|\{\text{SYT } T \text{ of shape } \lambda\}|^2 = (\dim \lambda)^2.$$

Hence, $\mathbf{Pr}[\mathrm{shRSK}(\boldsymbol{\pi}) = \lambda] = (\dim \lambda)^2/n!$, as in the Plancherel distribution.

By item 2 of the RSK correspondence, every word $w \in [d]^n$ maps to a unique pair (P, Q) with a common shape $\lambda \vdash n$. Furthermore, Q is SYT and P is SSYT and contains the same multiset of letters as in w. Hence, the probability that $\mathrm{RSK}(\boldsymbol{w}) = (P, Q)$ is α^P. Because RSK gives a bijection, every (P, Q) is mapped to by some word $w \in [d]^n$, and so the probability that $\mathrm{shRSK}(\boldsymbol{w}) = \lambda$ is

$$\sum_{(P,Q) \text{ of shape } \lambda} \alpha^P = \dim(\lambda) \cdot \sum_{P \text{ of shape } \lambda} \alpha^P = \dim(\lambda) \cdot s_\lambda(\alpha),$$

where the sums only include SYTs Q and SSYTs P with alphabet $[d]$. This is the same probability λ is observed in $\mathrm{SW}^n(\alpha)$. □

Recall from Definition 2.6.2 that when $\alpha = \mathsf{Unif}_d$, we may write SW_d^n instead of $\mathrm{SW}^n(\alpha)$. In this case,

$$\mathrm{SW}_d^n(\lambda) = \dim(\lambda) \cdot s_\lambda(\tfrac{1}{d}, \ldots, \tfrac{1}{d}).$$

Definition 2.2.13, Corollary 2.4.13, and the homogeneity of the Schur polynomials give the following well known formula (cf. [CHW07, Equation (26)]).

Proposition 3.3.3. $\mathrm{SW}_d^n(\lambda) = \dfrac{(\dim \lambda)^2}{n!} \cdot \dfrac{d^{\uparrow\lambda}}{d^n}$.

The following corollary motivates the study of the Plancherel distribution. Although it is not the output distribution of any quantum measurement we study[2], it is a special case of the Schur-Weyl distribution SW_d^n.

Corollary 3.3.4. $\mathrm{SW}_d^n \to \mathrm{Planch}_n$ *as* $d \to \infty$.

Proof. We give two proofs of this statement. The first uses the RSK correspondence, whereas the second uses Proposition 3.3.3, as in [CHW07].

Proof 1: Select $\boldsymbol{w} \sim \mathsf{Unif}_d^{\otimes n}$. As $d \to \infty$, the probability that $\boldsymbol{w}$ contains n distinct letters approaches 1. Conditioned on $\boldsymbol{w}$ containing n distinct letters, then $\mathrm{shRSK}(\boldsymbol{w}) \sim \mathrm{Planch}_n$, as (i) the RSK algorithm depends only on the relative magnitudes of the letters in $\boldsymbol{w}$ (and hence we may assume that $\boldsymbol{w}$ is a permutation of n) and (ii) the distribution on $\boldsymbol{w}$ is permutation symmetric, i.e. $\boldsymbol{w}$ and $\boldsymbol{w}_\pi = (\boldsymbol{w}_{\pi(1)}, \ldots, \boldsymbol{w}_{\pi(n)})$ occur with equal probability for any permutation π.

Proof 2: As $d \to \infty$, the ratio $d^{\uparrow\lambda}/d^n \to 1$, and so by Proposition 3.3.3,

$$\mathrm{SW}_d^n(\lambda) \to \frac{(\dim \lambda)^2}{n!},$$

as in the Plancherel distribution. □

The next corollary follows from the fact that the Schur polynomial $s_\lambda(\alpha)$ is symmetric in the α_i's (Proposition 2.4.7).

Corollary 3.3.5. *Given an n-letter α-random word $\boldsymbol{w}$, the distributions of the two random variables* $\mathrm{LIS}(\boldsymbol{w})$ *and* $\mathrm{RSK}(\boldsymbol{w})$ *depend only on the multiset* $\{\alpha_1, \ldots, \alpha_d\}$ *and not on the order of the α_i's. In addition, it follows easily from the RSK algorithm that these random variables depend only on the multiset of* nonzero *α_i's.*

As a result, when working with $\mathrm{SW}^n(\alpha)$ we may assume without loss of generality that α is sorted.

3.4 The Schur-Weyl growth process

In the related *Schur-Weyl growth process*, we imagine the process of growing a Young tableau by Schensted-inserting α-random letters into it one at a time.

Definition 3.4.1. The *Schur-Weyl growth process* is the infinite (random) sequence

$$\emptyset = \boldsymbol{\lambda}^{(0)} \nearrow \boldsymbol{\lambda}^{(1)} \nearrow \boldsymbol{\lambda}^{(2)} \nearrow \boldsymbol{\lambda}^{(3)} \nearrow \cdots$$

where $\boldsymbol{w} \sim \alpha^{\otimes\infty}$ and $\boldsymbol{\lambda}^{(t)} = \mathrm{shRSK}(\boldsymbol{w}[1\,..\,t])$. Note that the marginal distribution on $\boldsymbol{\lambda}^{(n)}$ is given by $\mathrm{SW}^n(\alpha)$.

[2] Though it is a distribution encountered when studying quantum algorithms for Graph Isomorphism [HMR+10].

The Schur-Weyl growth process was studied in, e.g., [O'C03], who noted the following.

Fact 3.4.2. *For any chain $\emptyset = \lambda^{(0)} \nearrow \cdots \nearrow \lambda^{(n)}$,*

$$\mathbf{Pr}[\boldsymbol{\lambda}^{(t)} = \lambda^{(t)} \quad \forall t \leq n] = s_{\lambda^{(n)}}(\alpha).$$

Proof. Let Q^* be the SYT containing, for each $i \in [n]$, the number i in the cell corresponding to $\lambda^{(i)} \setminus \lambda^{(i-1)}$. For a word $w \in [d]^n$, let $(P, Q) = \mathrm{RSK}(w)$. Then $\mathrm{shRSK}(w[1..t]) = \lambda^{(t)}$ for all $t \leq n$ if and only if $Q = Q^*$. By the RSK correspondence, the probability that such a word w is sampled from $\alpha^{\otimes n}$ is

$$\sum_{(P,Q^*) \text{ of shape } \lambda^{(n)}} \alpha^P = \sum_{P \text{ of shape } \lambda^{(n)}} \alpha^P = s_{\lambda^{(n)}}(\alpha). \qquad \square$$

(Together with the fact that $s_\lambda(\alpha)$ is homogeneous of degree $|\lambda|$, this gives yet another alternate definition of the Schur polynomials.) The definition of conditional expectation then immediately yields the following.

Corollary 3.4.3. *For any $i \in [d]$,*

$$\mathbf{Pr}[\boldsymbol{\lambda}^{(n+1)} = \lambda + e_i \mid \boldsymbol{\lambda}^{(n)} = \lambda] = \frac{s_{\lambda+e_i}(\alpha)}{s_\lambda(\alpha)}.$$

(This corollary is correct even when $\lambda+e_i$ is not a valid partition of $n+1$; in this case $s_{\lambda+e_i} \equiv 0$ formally under the determinantal definition.) The above equation is also a probabilistic interpretation of the following special case of *Pieri's rule.*

Fact 3.4.4 (Pieri's rule)**.** $(x_1 + \cdots + x_d)s_\lambda(x_1, \ldots, x_d) = \sum_{i=1}^{d} s_{\lambda+e_i}(x_1, \ldots, x_d)$.

We will need the following consequence of Corollary 3.4.3:

Proposition 3.4.5. *Let $\lambda \vdash n$ and let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then*

$$\left(\frac{s_{\lambda+e_1}(\alpha)}{s_\lambda(\alpha)}, \ldots, \frac{s_{\lambda+e_d}(\alpha)}{s_\lambda(\alpha)}\right) \succ (\alpha_1, \ldots, \alpha_d). \tag{3.1}$$

Proof. Let β be the reversal of α (i.e. $\beta_i = \alpha_{d-i+1}$) and let $(\boldsymbol{\lambda}^{(t)})_{t\geq 0}$ be a Schur-Weyl growth process corresponding to β. By Corollary 3.4.3 and the fact that the Schur polynomials are symmetric, we conclude that the vector on the left of (3.1) is $(p_1, \ldots, p_d)$, where $p_i = \mathbf{Pr}[\boldsymbol{\lambda}^{(n+1)} = \lambda + e_i \mid \boldsymbol{\lambda}^{(n)} = \lambda]$. Now $p_1 + \cdots + p_k$ is the probability, conditioned on $\boldsymbol{\lambda}^{(n)} = \lambda$, that the $(n+1)$-th box in the process enters into one of the first k rows. But this is indeed at least $\alpha_1 + \cdots + \alpha_k = \beta_d + \cdots + \beta_{d-k+1}$, because the latter represents the probability that the $(n+1)$-th letter is $d-k+1$ or higher, and such a letter will always be inserted within the first k rows under RSK. $\square$

3.5 Longest increasing subsequences of random permutations

Over the next three sections, we will use the connection to the RSK algorithm given by Theorem 3.3.2 to understand the distribution $\mathrm{SW}^n(\alpha)$ and its special case, Planch_n. In this section, we will consider only the distribution of $\boldsymbol{\lambda}_1$ when $\boldsymbol{\lambda} \sim \mathrm{Planch}_n$. Then in Section 3.6 we will consider the distribution of the entire shape $\boldsymbol{\lambda}$, and in Section 3.7 we will generalize this to the distribution of the entire shape $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$. Recall from Theorem 3.3.2 and Schensted's Theorem that $\boldsymbol{\lambda}_1$, when $\boldsymbol{\lambda} \sim \mathrm{Planch}_n$, is distributed as $\mathrm{LIS}(\boldsymbol{\pi})$ of a random permutation $\boldsymbol{\pi}$.

The study of longest increasing subsequences of permutations dates back to a paper of Erdős and Szekeres [ES35], who showed the following *worst-case* lower bound on the longest increasing and decreasing subsequences of any permutation.

Theorem 3.5.1 ([ES35]). *If $\pi \in \mathfrak{S}(n)$ and $n \geq rs + 1$ then $\mathrm{LIS}(\pi) > r$ or $\mathrm{LDS}(\pi) > s$. In fact, $\mathrm{LIS}(\pi) \cdot \mathrm{LDS}(\pi) > rs$.*

There are a variety of proofs of this result (see [Ste95] for a collection of them), perhaps the simplest being Seidenberg's [Sei59].

Of course, we are interested not in worst-case behavior, but in the *average*-case LIS behavior of a random permutation. The question of determining this behavior was originally posed by Ulam [Ula61]:

Theorem 3.5.2 (Ulam's problem). *Determine $\ell_n := \mathbf{E}_{\boldsymbol{\pi}\sim\mathfrak{S}(n)} \mathrm{LIS}(\boldsymbol{\pi})$.*

Already, Theorem 3.5.1 gives us a lower bound on ℓ_n: setting $r = s = \sqrt{n-1}$, it tells us that $\mathrm{LIS}(\pi) \cdot \mathrm{LDS}(\pi) \geq n$ for any permutation $\pi \in \mathfrak{S}(n)$. Given (i) that any permutation π and its reverse $\pi_{rev} := (\pi_n, \pi_{n-1}, \ldots, 1)$ occur with equal probability and (ii) the fact that $\mathrm{LIS}(\pi_{rev}) = \mathrm{LDS}(\pi)$ (and vice versa), this tells us that

$$\ell_n = \mathop{\mathbf{E}}_{\boldsymbol{\pi}\in\mathfrak{S}(n)} \left[\frac{1}{2}\mathrm{LIS}(\boldsymbol{\pi}) + \frac{1}{2}\mathrm{LDS}(\boldsymbol{\pi})\right] \geq \mathop{\mathbf{E}}_{\boldsymbol{\pi}\in\mathfrak{S}(n)} \sqrt{\mathrm{LIS}(\boldsymbol{\pi}) \cdot \mathrm{LDS}(\boldsymbol{\pi})} \geq \sqrt{n}, \tag{3.2}$$

where the second step is by the AM-GM inequality. On the other hand, there is an elementary proof (cf. [Rom14, Lemma 1.4]) which gives the upper bound

$$\limsup_{n\to\infty} \frac{\ell_n}{\sqrt{n}} \leq e. \tag{3.3}$$

The result is that, somewhat surprisingly, the average case longest increasing subsequence is not too much longer than the worst case.

In 1968, computer simulations by Baer and Brock suggested that $\ell_n \to 2\sqrt{n}$ as $n \to \infty$ [BB68]. The first step towards confirming this was given by Hammersley, who showed that the limit $\Lambda := \lim_{n\to\infty} \ell_n/\sqrt{n}$ does indeed exist [Ham72]. Following this, Logan and Shepp [LS77] and Vershik and Kerov [VK77] independently proved the Baer and Brock estimates.

Theorem 3.5.3. $\Lambda = 2$.

Following this, independent work by Vershik and Kerov [VK85] and by Pilpel [Pil90] showed that ℓ_n can even be upper-bounded in the non-asymptotic regime.

Theorem 3.5.4. *$\ell_n \leq 2\sqrt{n}$ for all n.*

We will give the elegant proof here as an example of the role the RSK algorithm plays in proving bounds on the LISes of random permutations.

Proof of Theorem 3.5.4. Set $\delta_n := \ell_n - \ell_{n-1}$. Our goal will be to upper bound $\delta_n \leq 1/\sqrt{n}$, from which the theorem will follow, as $\sum_{i=1}^n 1/\sqrt{i} \leq 2\sqrt{n}$. To begin, note that $\ell_{n-1} = \mathbf{E}\,\mathrm{shRSK}(\boldsymbol{\pi}[1..n-1])_1$, for $\boldsymbol{\pi} \sim \mathfrak{S}(n)$. This is because $\boldsymbol{\pi}[1..n-1]$ always has $n-1$ distinct letters in random order. Then

$$\delta_n = \mathop{\mathbf{E}}_{\boldsymbol{\pi}\sim\mathfrak{S}(n)}[\mathrm{shRSK}(\boldsymbol{\pi})_1 - \mathrm{shRSK}(\boldsymbol{\pi}[1..n-1])_1] = \mathop{\mathbf{Pr}}_{\boldsymbol{\pi}\sim\mathfrak{S}(n)}[\text{first-row}(\boldsymbol{\pi})],$$

where first-row$(\boldsymbol{\pi})$ is the event that a box was added to the first row in the last step of RSK on $\boldsymbol{\pi}$. Setting $(\boldsymbol{P}, \boldsymbol{Q}) = \mathrm{RSK}(\boldsymbol{\pi})$, first-row$(\boldsymbol{\pi})$ occurs when $\boldsymbol{Q}$ has an "n" in the rightmost box of its first row. This can occur only when $\mathrm{sh}(\boldsymbol{Q})$ is of the form $\mu + \square$, where $\mu \vdash n-1$ and "$+\,\square$" means adding a box to the first row (as otherwise $\boldsymbol{Q}$ would not be SYT). By the RSK correspondence, the number of permutations $\pi \in \mathfrak{S}(n)$ for which first-row(π) occurs is

$$\sum_{\mu\vdash n-1} \dim(\mu)\dim(\mu+\square).$$

Hence, the probability that a random $\boldsymbol{\pi}$ satisfies first-row$(\boldsymbol{\pi})$ is

$$\sum_{\mu\vdash n-1} \frac{\dim(\mu)\dim(\mu+\square)}{n!} \leq \sqrt{\frac{1}{n}\sum_{\mu\vdash n-1}\frac{\dim(\mu)^2}{(n-1)!}\cdot\sum_{\mu\vdash n-1}\frac{\dim(\mu+\square)^2}{n!}}$$
$$= \sqrt{\frac{1}{n}\sum_{\mu\vdash n-1}\frac{\dim(\mu+\square)^2}{n!}} \leq \sqrt{\frac{1}{n}\sum_{\lambda\vdash n}\frac{\dim(\lambda)^2}{n!}} = \frac{1}{\sqrt{n}},$$

where the first step uses Cauchy-Schwarz and the two equalities follow from the definition of the Plancherel distribution. □

We will reprove and generalize this theorem to random words in our Lemma 4.2.1.

The next step in answering Ulam's problem was to determine the next highest order terms in ℓ_n after $2\sqrt{n}$. Related to this was to determine the limiting distribution of $\mathrm{LIS}(\boldsymbol{\pi})$. In spite of conjectures that it should have standard deviation on the order of $n^{1/4}$, Odlyzko and Rains [OR00] presented numerical evidence that both $2\sqrt{n} - \ell_n$ and $\mathbf{stddev}[\mathrm{LIS}(\boldsymbol{\pi})]$ were on the order of $n^{1/6}$, and further that $\mathrm{LIS}(\boldsymbol{\pi})$ was non-Gaussian in the limit. This was confirmed in the work of Baik, Deift, and Johansson [BDJ99], who connected the limiting distribution of $\mathrm{LIS}(\boldsymbol{\pi})$ to the *Tracy-Widom distribution* from random matrix theory.

Definition 3.5.5. The *Gaussian Unitary Ensemble* GUE_m is the probability distribution over $m \times m$ random Hermitian matrices $\boldsymbol{X}$ with i.i.d. entries in which (i) $\boldsymbol{X}_{k,k} \sim \mathcal{N}(0,1)$ for all $k \in [m]$, and (ii) $\boldsymbol{X}_{j,k} \sim \mathcal{N}(0,1)_{\mathbb{C}}$ and $\boldsymbol{X}_{k,j} = \overline{\boldsymbol{X}_{j,k}}$ for all $j < k \in [m]$. Here $\mathcal{N}(0,1)_{\mathbb{C}}$ refers to the *complex* standard Gaussian, distributed as $\mathcal{N}(0,\frac{1}{2}) + i\mathcal{N}(0,\frac{1}{2})$.

Definition 3.5.6. The *Tracy-Widom distribution* TW is given by the random variable

$$\sqrt{2}m^{1/6} \cdot (\lambda_1(\boldsymbol{X}) - \sqrt{2m}),$$

where $\boldsymbol{X} \sim \mathrm{GUE}_m$, in the limit as $m \to \infty$. Its mean is -1.771 and its variance is 0.813 [TW09].

The Baik, Deift, and Johansson [BDJ99] result is given as follows.

Theorem 3.5.7. *As $n \to \infty$, then the random variable*

$$\frac{\mathrm{LIS}(\boldsymbol{\pi}) - \sqrt{2n}}{n^{1/6}} \to \mathrm{TW}$$

in distribution, where $\boldsymbol{\pi} \sim \mathfrak{S}(n)$. In other words, the two random variables

$$\sqrt{2}m^{1/6} \cdot (\lambda_1(\boldsymbol{X}) - \sqrt{2m}) \qquad \text{and} \qquad \frac{\mathrm{LIS}(\boldsymbol{\pi}) - \sqrt{2n}}{n^{1/6}},$$

where $\boldsymbol{X} \sim \mathrm{GUE}_m$ and $\boldsymbol{\pi} \sim \mathfrak{S}(n)$, converge to each other in distribution as $m, n \to \infty$.

This result essentially gives a complete answer to Ulam's problem, at least in the asymptotic regime.

3.6 RSK of random permutations

Concurrent with, and inspired by, the work on the distribution of $\boldsymbol{\lambda}_1$, for $\boldsymbol{\lambda} \sim \mathrm{Planch}_n$, was work on the distribution of the entire shape $\boldsymbol{\lambda}$. Already, the results on $\boldsymbol{\lambda}_1$ tell us interesting things about the shape of $\boldsymbol{\lambda}$. For example, the fact that $\mathrm{RSK}(\pi) = \mathrm{RSK}(\pi_{rev})'$, where $\pi_{rev} = (\pi_n, \pi_{n-1}, \ldots, \pi_1)$, tells us that $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ have the same distribution, and so it is natural to consider $\boldsymbol{\lambda}$ in the Russian notation. Furthermore, since we expect $\boldsymbol{\lambda}_1 \sim 2\sqrt{n}$, we also expect $\boldsymbol{\lambda}'_1 \sim 2\sqrt{n}$. Therefore, a typical Young diagram will be roughly $2\sqrt{n} \times 2\sqrt{n}$, and so it is natural to define scaled partitions as follows.

Definition 3.6.1. Let $\lambda \vdash n$ and recall Definition 2.2.4. Then $\overline{\lambda} : \mathbb{R} \to \mathbb{R}_+$ is defined as $\overline{\lambda}(x) := \lambda(\sqrt{n} \cdot x)/\sqrt{n}$, for all x.

Logan and Shepp [LS77] and Vershik and Kerov [VK77] independently showed that $\overline{\boldsymbol{\lambda}}$ approaches a limiting shape as $n \to \infty$. This is the so-called "law of large numbers" for the Plancherel distribution.

Theorem 3.6.2 (Law of large numbers)**.** *When $\boldsymbol{\lambda} \sim \mathrm{Planch}_n$ and $n \to \infty$, the function $\overline{\boldsymbol{\lambda}}$ converges to $\Omega(x)$, the curve defined as*

$$\Omega(x) := \begin{cases} \frac{2}{\pi}(x \arcsin \frac{x}{2} + \sqrt{4 - x^2}), & |x| \leq 2, \\ |x| & |x| \geq 2. \end{cases}$$

This "ice cream cone"-shaped function is pictured in Figure 3.3 ($c = 0$ case). We note that Ω is symmetric about the y-axis and that $\Omega(2) = \Omega(-2) = 2$, both as expected. Though this curve is a limiting shape rather than the Russian notation of any Young diagram, it is useful to think of it as a continual analogue of a Young diagram, as per the following definition.

Definition 3.6.3. A *continual diagram* is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying (i) f is 1-Lipschitz and (ii) $f(x) = |x|$ when $|x|$ is sufficiently large.

This definition originates in the paper of [Ker93a]. Note that if λ is a Young diagram, then $\lambda(x)$ is a continual diagram.

As in the case of random permutations, it is now reasonable to ask about the limiting distribution of $\boldsymbol{\lambda}$ for large n, and it is at this point that research on the Plancherel measure splits into two distinct streams. As Okounkov puts it [Oko00], one either cares about the behavior of the limiting distribution in the bulk of the limiting shape, meaning $\overline{\boldsymbol{\lambda}}(x)$ where $x \in (-2, 2)$ is away from the endpoints $x = \pm 2$, or one cares about the behavior near the edge of the limiting shape, when x is close to ± 2. The latter of these involves characterizing the distribution of $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_k$ for small k.

3.6.1 The bulk of the limit shape

The main result on the bulk of the limit shape is Kerov's "central limit theorem" for the Plancherel measure [Ker93b], which characterized the deviation of a random Young diagram from the curve $\Omega(x)$ by a certain Gaussian process. A second proof of this result, also by Kerov, was given in the paper of Ivanov and Olshanski [IO02]. Much of our work is based on the techniques of this paper.

Theorem 3.6.4 (Central limit theorem)**.** *When $\boldsymbol{\lambda} \sim \mathrm{Planch}_n$ and $n \to \infty$, the function $\overline{\boldsymbol{\lambda}}(x)$ "fluctuates" as $\Omega(x) + \frac{2}{\sqrt{n}}\boldsymbol{\Delta}(x)$, where*

$$\boldsymbol{\Delta}(z) = \frac{1}{\pi} \sum_{k=2}^{\infty} \frac{\boldsymbol{\xi}_k}{\sqrt{k}} \sin(k\theta),$$

where $z = 2\cos\theta$ for $0 \leq \theta \leq \pi$ and $\boldsymbol{\xi}_2, \boldsymbol{\xi}_3, \ldots$ are independent standard real Gaussians.

In other words, to compute $\boldsymbol{\Delta}$, multiply each of the sine curves $\sin(k\theta)/\pi\sqrt{k}$, $k \geq 2$, by an independent standard Gaussian and sum the results. Then scale $\boldsymbol{\Delta}$ and add it to Ω.

3.6.2 The edge of the limit shape

Studying the edge of the limit shape involves studying $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_k$, for k small. From the limit shape and the fact that $\boldsymbol{\lambda}_k \leq \boldsymbol{\lambda}_1 \sim 2\sqrt{n}$, we also expect that $\boldsymbol{\lambda}_k \sim 2\sqrt{n}$. Determining the limiting distribution of $\boldsymbol{\lambda}_k$ (and the joint limiting distribution of $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_k)$) was accomplished by proving a natural generalization of Theorem 3.5.7. First, [BDJ00] generalized this theorem to the second row $\boldsymbol{\lambda}_2$, and following this [Joh01] and [BOO00] independently generalized it to the first k rows (see also another proof by [Oko00]).

Theorem 3.6.5. *For any fixed $k \geq 1$, the two k-valued random variables*

$$\left(\sqrt{2}m^{1/6} \cdot (\lambda_i(\boldsymbol{X}) - \sqrt{2m})\right)_{i\in[k]} \qquad \textit{and} \qquad \left(\frac{\boldsymbol{\lambda}_i - \sqrt{2n}}{n^{1/6}}\right)_{i\in[k]},$$

where $\boldsymbol{X} \sim \mathrm{GUE}_m$ and $\boldsymbol{\pi} \sim \mathrm{Planch}_n$, converge to each other in distribution as $m, n \to \infty$.

3.7 RSK of random words

In this section, we can finally discuss the Schur-Weyl distribution. This distribution shares many properties with the Plancherel distribution, such as the relationship with the Tracy-Widom distribution, but also exhibits some basic differences, which partially arise due to the fact that words allow for repeated letters while permutations do not. One example is that the expected LIS of an α-random word has a surprisingly simple form.

Proposition 3.7.1. $\mathbf{E}_{\boldsymbol{w}\sim\alpha^{\otimes n}} \mathrm{LIS}(\boldsymbol{w})/n \to \alpha_1$ *as* $n \to \infty$.

Proof. The lower bound follows from the fact that the all-ones subsequence of any word is always increasing, and hence the longest increasing subsequence is only longer:

$$\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\alpha^{\otimes n}} \mathrm{LIS}(\boldsymbol{w}) \geq \mathop{\mathbf{E}}_{\boldsymbol{w}\sim\alpha^{\otimes n}} [\#\text{ of 1's in } \boldsymbol{w}] = \alpha_1 n.$$

As for the upper bound, we note that any maximal increasing subsequence s in $\boldsymbol{w}$ splits the interval $[1, n]$ into d contiguous subintervals

$$I_1 = [1, n_1], I_2 = [n_1 + 1, n_1 + n_2], \ldots, I_d = [n - n_d + 1, n],$$

so that s contains all the 1's in $\boldsymbol{w}$ which fall in I_1, the $2's$ which fall in I_2, and so forth. The length of the increasing subsequence in $\boldsymbol{w}$ corresponding to this set of intervals is distributed as $\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n$, where $\boldsymbol{X}_i$ is Bernoulli with expectation α_j if $i \in I_j$. Because α_1 is the max of the α_i's, the Chernoff bound states that the probability that $\boldsymbol{X}_1 + \ldots + \boldsymbol{X}_n \geq (\alpha_1 + \epsilon)n$ for some $\epsilon > 0$ is at most $\exp(-2n\epsilon^2)$. Union bounding over all such partitions of $[1, n]$, we get that

$$\mathbf{Pr}[\mathrm{LIS}(\boldsymbol{w}) \geq (\alpha_1 + \epsilon)n] \leq \frac{n^d}{e^{2n\epsilon^2}}.$$

This decays exponentially in n, and from here it is straightforward to derive the statement in the proposition. □

At a high level, this says that the LIS of an α-random word is not much longer than the all-1's subsequence. One might then expect the longest disjoint *pair* of increasing subsequences to not be much longer than the all-1's and all-2's subsequences which, by Greene's Theorem, would tell us that $\boldsymbol{\lambda}_2 \sim \alpha_2 n$. Extending this intuition to all of $\boldsymbol{\lambda}$, we expect that

$$\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\alpha^{\otimes n}} \underline{\boldsymbol{\lambda}} \to \alpha \text{ as } n \to \infty.$$

This fact has been independently proved several times, e.g. in [ARS88] and [KW01]. The earliest reference we know of is by Kerov and Vershik in [KV86] (using a result from [VK81]).

One seeming difference between random words $\boldsymbol{w}$ and random permutations $\boldsymbol{\pi}$ is that in Proposition 3.7.1 we could easily calculate the expected LIS of $\boldsymbol{w}$, whereas calculating the expected LIS of $\boldsymbol{\pi}$ was at one point a famous open problem (Ulam's problem). This difference is deceptive though: just as for random words, we can easily calculate $\mathbf{E}\,\mathrm{LIS}(\boldsymbol{\pi})/n$ as $n \to \infty$: it just happens to be zero. For random permutations, the interesting behavior occurs on the order of $\sqrt{n}$, not n. In fact, LISes of random words also have interesting lower-order behavior around $\sqrt{n}$. For example, we will show below in Theorem 4.2.2 that

$$\mathop{\mathbf{E}}_{\boldsymbol{w} \sim \alpha^{\otimes n}} \mathrm{LIS}(\boldsymbol{w}) \leq \alpha_1 n + 2\sqrt{n}$$

in the case when $\alpha = \mathsf{Unif}_d$ is the uniform distribution. Using this, we may recover Theorem 3.5.4 by taking $d \to \infty$ and applying Corollary 3.3.4. (We believe this statement holds for all α, though we can only prove it by replacing the 2 with 2.83. See Lemma 4.2.1.)

The overall picture of $\mathrm{SW}^n(\alpha)$ is more complex than that of Planch_n, as $\mathrm{SW}^n(\alpha)$ can look fundamentally different for different values of α. For example, if $\boldsymbol{w} \sim \alpha^{\otimes n}$ in the case when $\alpha_1 \gg \alpha_2$, then the longest increasing subsequence of $\boldsymbol{w}$ is relatively easy to construct: there are so many more 1's in $\boldsymbol{w}$ than 2's, 3's, etc., that the longest increasing subsequence in $\boldsymbol{w}$ is practically forced to take almost all of the 1's, along with perhaps a few other letters near the end of $\boldsymbol{w}$. In this case, the distribution of $\mathrm{LIS}(\boldsymbol{w})$ will be very close to $\mathrm{Binomial}(n, \alpha_1)$. On the other hand, if $\alpha_1 = \alpha_2$, then it's not clear whether the longest increasing subsequence should take (almost) all of the 1's, (almost) all of the 2's, or some of the 1's and then some of the 2's, and the situation becomes even more complicated when *all* of the α_i's are the same.

This high-level picture is summarized in the following three cases.

1. If the α_i's are all distinct, then $\mathbf{E}\,\boldsymbol{\lambda}_i/n$ converges to α_i. Furthermore, the fluctuations

$$\sqrt{n}\left(\frac{\boldsymbol{\lambda}_i}{n} - \alpha_i\right)$$

   are Gaussian with covariance $\delta_{ij}\alpha_i - \alpha_i\alpha_j$. This is exactly what one would expect if $\boldsymbol{\lambda}_i$ always were equal to the number of i's in $\boldsymbol{w}$.

2. If $\alpha_1 = \ldots = \alpha_d = \frac{1}{d}$, i.e. $\alpha = \mathsf{Unif}_d$, then $\boldsymbol{\lambda}_1 \sim \frac{n}{d} + 2\sqrt{n}$, $\boldsymbol{\lambda}_d \sim \frac{n}{d} - 2\sqrt{n}$, and the other $\boldsymbol{\lambda}_i$'s interpolate between these two values in a manner reminiscent of the limiting curve for the Plancherel distribution $\Omega(x)$. Furthermore, their limiting distributions are given by the eigenvalues of a suitable Gaussian random matrix.

3. In general, if some of the α_i's are the same and others are different, then we can bucket the α_i's into blocks $B_1, \ldots, B_m \subset [d]$ in which $\alpha_i = \alpha_j$ if and only if i, j fall into the same block B_k. Then across blocks a random $\boldsymbol{\lambda}$ will act as in case 1, but within blocks it will act as in case 2. In other words, given a vector $v \in \mathbb{R}^d$ if we write $v[k] = \sum_{i \in B_k} v_i$, then $\mathbf{E}\,\boldsymbol{\lambda}[k]/n$ converges to $\alpha[k]$. Furthermore, the fluctuations

$$\sqrt{n}\left(\frac{\boldsymbol{\lambda}[k]}{n} - \alpha[k]\right)$$

   are Gaussian with covariance $\delta_{k\ell}\alpha[k] - \alpha[k]\alpha[\ell]$, just as in case 1. On the other hand, if we suppose that $B_k = [i, j]$, then we expect that $\boldsymbol{\lambda}_i \sim \alpha_i n + 2\sqrt{n}$ and $\boldsymbol{\lambda}_j \sim \alpha_j n - 2\sqrt{n}$ (recalling that $\alpha_i = \alpha_j$), with the $\boldsymbol{\lambda}_\ell$'s in the middle interpolating between these two values. Furthermore, their limiting distributions are given by the eigenvalues of a suitable Gaussian random matrix.

This line of work is a continuation of the line of work for the Plancherel distribution pertaining to the "edge of the limit shape" (Section 3.6.2), especially the work of Baik, Deift, and Johansson [BDJ99]. Note that in this line of work for the Schur-Weyl distribution, α is kept fixed while n is allowed to grow unbounded. As a result, there is a fixed number of rows of $\boldsymbol{\lambda}$, and so it is possible to prove convergence for the entire diagram $\boldsymbol{\lambda}$. This contrasts with the case of the Plancherel distribution, where the number of rows grows with n, but one can only prove convergence to the limiting distribution for the first k rows, for any fixed k.

In this work, we are interested in the properties of $\mathrm{SW}^n(\alpha)$ when n is small. It would be natural to try to show that $\mathrm{SW}^n(\alpha)$ converges quickly to its limiting distribution via some sort of Berry-Esseen theorem, and then to use properties of the limiting distribution. Unfortunately, such a Berry-Esseen theorem would necessarily have a convergence rate depending on $\min_{i,j:\alpha_i\neq\alpha_j} |\alpha_i - \alpha_j|$ because the character of the limiting distributions (Gaussian versus Tracy-Widom) depends on whether two α_i's are equal. Hence, this strategy is unlikely to give good small-n estimates. Nevertheless, we have found these limiting distributions useful for providing intuition in the small-n case.

Let us now formalize this high level picture. Item 1 was originally proved in the quantum computing literature in [ARS88] and then proved in the mathematics literature in [HX13] (and reproved in the works of [Buf12], [Mél12], and [FMN13, Equation (55)]). We will trace through the history of items 2 and 3 in the following section.

3.7.1 Convergence to the GUE

We will largely follow the treatment of [HX13] in this section. In [TW01], Tracy and Widom considered the following distribution on random matrices.

Definition 3.7.2. The *traceless GUE*, denoted GUE_d^0, is the probability distribution on $d \times d$ Hermitian matrices $\boldsymbol{X}$ drawn according to the following two-step process: (i) sample $\boldsymbol{Y} \sim \mathrm{GUE}_d$; (ii) output

$$\boldsymbol{X} := \boldsymbol{Y} - \frac{\mathrm{tr}(\boldsymbol{Y})}{d} \cdot I.$$

The next fact characterizes the eigenvalues of the traceless GUE in the limit (cf. [HX13]).

Fact 3.7.3. *Given $\boldsymbol{X} \sim \mathrm{GUE}_d^0$, then as $d \to \infty$,*

$$\left(\frac{\lambda_1(\boldsymbol{X})}{\sqrt{d}}, \ldots, \frac{\lambda_d(\boldsymbol{X})}{\sqrt{d}}\right)$$

converges almost surely to the semicircle law with density $\sqrt{4-x^2}/2\pi$, $-2 \leq x \leq 2$.

The main result of [TW01] was to connect this random matrix ensemble to the *homogeneous random words* problem (i.e. the case when α is the uniform distribution). They showed that as $n \to \infty$ then given a random $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$,

$$\frac{\boldsymbol{\lambda}_1 - n/d}{\sqrt{n/d}} \to \lambda_1(\boldsymbol{X}),$$

in distribution, where $\boldsymbol{X} \sim \mathrm{GUE}_d^0$. They conjectured that this behavior extends to all of the rows as in item 2, which Johansson [Joh01] confirmed in the following theorem. (See also Kuperberg [Kup02] for a quantum-mechanics-inspired proof of this result.)

Theorem 3.7.4. *Let d be fixed. As $n \to \infty$ then for $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$, the random variable*

$$\left(\frac{\boldsymbol{\lambda}_1 - n/d}{\sqrt{n/d}}, \ldots, \frac{\boldsymbol{\lambda}_d - n/d}{\sqrt{n/d}}\right) \to (\lambda_1(\boldsymbol{X}), \ldots, \lambda_d(\boldsymbol{X})) \tag{3.4}$$

in distribution, where $\boldsymbol{X} \sim \mathrm{GUE}_d^0$.

Using Fact 3.7.3, we expect $\lambda_1(\boldsymbol{X}) \approx 2\sqrt{d}$ and $\lambda_d(\boldsymbol{X}) \approx -2\sqrt{d}$. Thus, by Theorem 3.7.4, $\boldsymbol{\lambda}_1 \approx n/d + 2\sqrt{n}$ and $\boldsymbol{\lambda}_d \approx n/d - 2\sqrt{n}$, as guaranteed by Item 2.

Following this, Its, Tracy, and Widom [ITW01] considered the case of nonuniform α_i's (the *inhomogeneous* random word setting). As we have seen before, the distribution $\mathrm{SW}^n(\alpha)$ depends on the degeneracies present in α, i.e. whether distinct α_i's are equal to each other. For this, we will need to establish some notation.

Notation 3.7.5. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution. We will write $d_1 + \cdots + d_m = d$ for the *multiplicities* of the α_i's, meaning that α_1 occurs in α a total of d_1 times, the next largest α_i (i.e. α_{d_1+1}) occurs d_2 times, and so on. In addition, we will write $\alpha^{(k)}$ for the α-value which occurs with multiplicity d_k, and we will write $\alpha[k] := d_k\alpha^{(k)}$ for the sum of the α_i's which occur with multiplicity d_k.

Its, Tracy, and Widom [ITW01] introduced the following generalization of the traceless GUE, which was named by Houdre and Xu [HX13].

Definition 3.7.6. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution. The *generalized traceless GUE*, denoted $\mathrm{GUE}^0(\alpha)$, is the probability distribution on $d \times d$ Hermitian matrices $\boldsymbol{X}$ drawn according to the following process:

- For each $k \in [m]$, draw $\boldsymbol{Y}^{(k)} \sim \mathrm{GUE}_{d_k}$.
- Let $\boldsymbol{Y}$ be the $d \times d$ block-diagonal matrix whose k-th diagonal block is $\boldsymbol{Y}^{(k)}$.
- Output the matrix $\boldsymbol{X}$ defined as

$$\boldsymbol{X}_{i,i} = \begin{cases} \boldsymbol{Y}_{i,i} - \sqrt{\alpha_i} \sum_{j=1}^d \sqrt{\alpha_j} \cdot \boldsymbol{Y}_{j,j} & \text{if } i = j, \\ \boldsymbol{Y}_{i,j} & \text{if } i \neq j. \end{cases} \tag{3.5}$$

Note that $\boldsymbol{X}$, like $\boldsymbol{Y}$, is a block diagonal matrix. We will write $\lambda_1^{(k)}(\boldsymbol{X}), \ldots, \lambda_{d_k}^{(k)}(\boldsymbol{X})$ for the eigenvalues of the k-th block, sorted in decreasing order.

Counterintuitively, the generalized traceless GUE, as defined, is not always a trace zero matrix. However, $\boldsymbol{X}$ does always satisfy the "traceless" condition

$$\sum_{i=1}^d \sqrt{\alpha_i} \cdot \boldsymbol{X}_i = 0$$

which, as we will see, arises naturally in this context. The main result of [ITW01] is a generalization of Equation (3.4) to the case of nonuniform α_i's: as $n \to \infty$ then given a random $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$,

$$\frac{\boldsymbol{\lambda}_1 - \alpha_1 n}{\sqrt{\alpha_1 n}} \to \lambda_1^{(1)}(\boldsymbol{X})$$

in distribution. Following their paper, Houdre and Xu [HX13], in a work originally appearing in 2009, extended this result to apply to the full Young diagram.

Theorem 3.7.7. *Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution. As $n \to \infty$, then for $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, the random variable*

$$\left(\frac{\boldsymbol{\lambda}_1 - \alpha_1 n}{\sqrt{\alpha_1 n}}, \ldots, \frac{\boldsymbol{\lambda}_d - \alpha_d n}{\sqrt{\alpha_d n}}\right) \to \left(\lambda_1^{(1)}(\boldsymbol{X}), \ldots, \lambda_{d_1}^{(1)}(\boldsymbol{X}), \ldots, \lambda_1^{(m)}(\boldsymbol{X}), \ldots, \lambda_{d_m}^{(m)}(\boldsymbol{X})\right)$$

in distribution, where $\boldsymbol{X} \sim \mathrm{GUE}_d^0$.

Example 3.7.8. It is instructive to carry out Theorem 3.7.7 in the case when the α_i's are distinct. In this case, a matrix $\boldsymbol{X} \sim \mathrm{GUE}^0(\alpha)$ can be simulated by drawing d independent Gaussians $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_d \sim \mathcal{N}(0,1)$ and setting

$$\boldsymbol{X}_{i,i} = \boldsymbol{g}_1 - \sqrt{\alpha_i} \sum_{j=1}^{d} \sqrt{\alpha_j} \cdot \boldsymbol{g}_j,$$

for each $i \in [d]$. Because $\boldsymbol{X}$ is block diagonal with blocks of size one, we have that $\lambda_1^{(i)}(\boldsymbol{X}) = \boldsymbol{X}_{i,i}$ for all $i \in [d]$. Thus, by Theorem 3.7.7, as $n \to \infty$,

$$\left(\frac{\boldsymbol{\lambda}_1 - \alpha_1 n}{\sqrt{\alpha_1 n}}, \ldots, \frac{\boldsymbol{\lambda}_d - \alpha_d n}{\sqrt{\alpha_d n}}\right) \to (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$$

in distribution. Equivalently,

$$\left(\frac{\boldsymbol{\lambda}_1 - \alpha_1 n}{\sqrt{n}}, \ldots, \frac{\boldsymbol{\lambda}_d - \alpha_d n}{\sqrt{n}}\right) \to (\sqrt{\alpha_1} \cdot \boldsymbol{X}_1, \ldots, \sqrt{\alpha_d} \cdot \boldsymbol{X}_d)$$

in distribution. The coordinates of the right-hand side are Gaussian with covariance $\delta_{ij}\alpha_i - \alpha_i\alpha_j$. As a result, we recover item 1.

If we write $\boldsymbol{X}^{(k)}$ for the k-th diagonal block of $\boldsymbol{X} \sim \mathrm{GUE}^0(\alpha)$, then

$$\boldsymbol{X}^{(k)} = \boldsymbol{Y}^{(k)} - \sqrt{\alpha^{(k)}} \sum_{\ell=1}^{m} \sqrt{\alpha^{(\ell)}} \mathrm{tr}(\boldsymbol{Y}^{(\ell)}) \cdot I_{d_k \times d_k}$$

by Equation (3.5). The trace of this matrix is

$$\mathrm{tr}(\boldsymbol{X}^{(k)}) = \mathrm{tr}(\boldsymbol{Y}^{(k)}) - d_k \sqrt{\alpha^{(k)}} \sum_{\ell=1}^{m} \sqrt{\alpha^{(\ell)}} \mathrm{tr}(\boldsymbol{Y}^{(\ell)}).$$

It can be checked that the traces of the m diagonal blocks of $\boldsymbol{X}$ are distributed as centered Gaussian random variables with covariance $\delta_{k\ell} d_k - d_k d_\ell \sqrt{\alpha^{(k)}\alpha^{(\ell)}}$. The following theorem, sketched by Meliot in [Mél12, Theorem 4], shows that these Gaussian fluctuations can be decoupled from the GUE fluctuations in the generalized traceless GUE.

Theorem 3.7.9. *Let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a sorted probability distribution. Let $\boldsymbol{X} \sim \mathrm{GUE}^0(\alpha)$, let $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m$ be centered Gaussian random variables with covariance $\delta_{k\ell} d_k - d_k d_\ell \sqrt{\alpha^{(k)}\alpha^{(\ell)}}$, and let $\boldsymbol{Y}^{(k)} \sim \mathrm{GUE}^0_{d_k}$, for each $k \in [m]$. Then we have the following equivalence in distribution:*

$$\lambda_i^{(k)}(\boldsymbol{X}) \overset{d}{=} \frac{\boldsymbol{g}_k}{d_k} + \lambda_i(\boldsymbol{Y}^{(k)}),$$

where the left- and right-hand sides refer to joint random variables ranging over $k \in [m]$ and $i \in [d_k]$. As a result, as $n \to \infty$, then for $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, the random variable

$$\left\{ \frac{\boldsymbol{\lambda}_i^{(k)} - \alpha^{(k)} n}{\sqrt{\alpha^{(k)} n}} \right\}_{k \in [m], i \in [d_k]} \to \left\{ \frac{\boldsymbol{g}_k}{d_k} + \lambda_i(\boldsymbol{Y}^{(k)}) \right\}_{k \in [m], i \in [d_k]} \tag{3.6}$$

in distribution.

We note that Houdre and Xu show a similar, though reversed, statement in [HX13, Proposition 2.6]: given $\boldsymbol{Y}$ and $\boldsymbol{X}$ as in Definition 3.7.6, the eigenvalues of $\boldsymbol{Y}$ are distributed as the eigenvalues of $\boldsymbol{X}$, plus some independent Gaussian random variables $(\boldsymbol{g}_1, \dots, \boldsymbol{g}_d)$ of covariance $\sqrt{\alpha_i \alpha_j}$.

Rewriting Equation (3.6),

$$\frac{\boldsymbol{\lambda}_i^{(k)} - \alpha^{(k)} n}{\sqrt{n}} \to \frac{\sqrt{\alpha^{(k)}}\boldsymbol{g}_k}{d_k} + \sqrt{\alpha^{(k)}}\lambda_i(\boldsymbol{Y}^{(k)}).$$

Hence,

$$\frac{\boldsymbol{\lambda}[k] - \alpha[k] n}{\sqrt{n}} \to \sqrt{\alpha^{(k)}}\boldsymbol{g}_k + \sum_{i=1}^{d_k} \sqrt{\alpha^{(k)}}\lambda_i(\boldsymbol{Y}^{(k)}) = \sqrt{\alpha^{(k)}}\boldsymbol{g}_k.$$

The m random variables $\sqrt{\alpha^{(k)}}\boldsymbol{g}_k$ are centered Gaussians with covariance $\delta_{k\ell}\alpha[k] - \alpha[k]\alpha[\ell]$, and within the k-th block we get GUE-style deviations, as stated in Item 3.

3.7.2 Schur-Weyl for uniform distribution

A parallel line of work has considered the distribution SW^n_d in the case when $d := d(n)$ is a growing function with n (in contrast with item 2, where d is fixed). This line of work is inspired by and extends the work focusing on the "bulk of the limit shape" (Section 3.6.1) for the Plancherel measure. Here, it turns out that the features of a "typical" $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$ depend on the ratio $c := \frac{\sqrt{n}}{d}$.

Biane [Bia01] extended the Plancherel law of large numbers to the Schur-Weyl distribution in the case when c is a fixed constant and $n, d \to \infty$. In this case, for a random $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$, the function $\overline{\boldsymbol{\lambda}}$ will approach a certain limiting curve Ω_c, specified as follows:

Theorem 3.7.10 ([Bia01])**.** *Fix an absolute constant $c > 0$ and assume $n, d \to \infty$ with $\frac{\sqrt{n}}{d} \to c$. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_d}\left[\|\overline{\boldsymbol{\lambda}} - \Omega_c\|_\infty \geq \epsilon\right] \to 0,$$

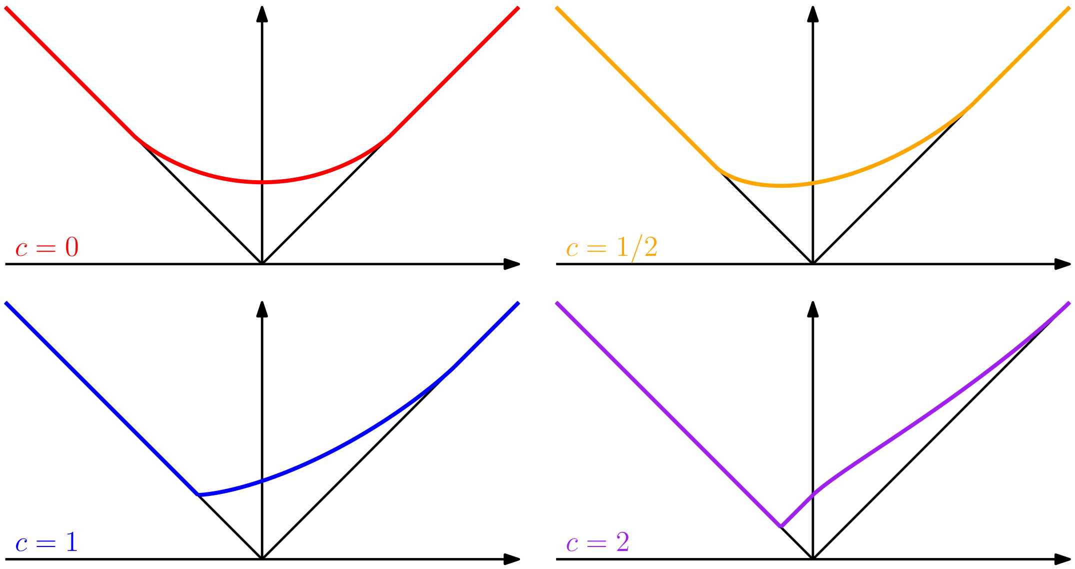


Figure 3.3: The Biane limiting curves Ω_c. The $c = 0$ case corresponds to the function $\Omega(x)$.

where Ω_c is the continual diagram defined as follows:

$$\Omega_0(x) = \Omega(x);$$

$$\Omega_{c\in(0,1)}(x) = \begin{cases} \frac{2}{\pi}\left(x \arcsin(\frac{x+c}{2\sqrt{1+cx}}) + \frac{1}{c}\arccos(\frac{2+cx-c^2}{2\sqrt{1+cx}}) + \frac{\sqrt{4-(x-c)^2}}{2}\right) & \textit{if } |x-c| \leq 2, \\ |x| & \textit{otherwise}; \end{cases}$$

$$\Omega_{c=1}(x) = \begin{cases} \frac{x+1}{2} + \frac{1}{\pi}\left((x-1)\arcsin(\frac{x-1}{2}) + \sqrt{4-(x-1)^2}\right) & \textit{if } |x-1| \leq 2, \\ |x| & \textit{otherwise}; \end{cases}$$

$$\Omega_{c>1}(x) = \begin{cases} x + \frac{2}{c} & \textit{if } x \in (\frac{-1}{c}, c-2) \\ \frac{2}{\pi}\left(x \arcsin(\frac{x+c}{2\sqrt{1+cx}}) + \frac{1}{c}\arccos(\frac{2+cx-c^2}{2\sqrt{1+cx}}) + \frac{\sqrt{4-(x-c)^2}}{2}\right) & \textit{if } |x-c| \leq 2, \\ |x| & \textit{otherwise.} \end{cases}$$

These curves are pictured for various values of c in Figure 3.3 (which we have reproduced from [Mél10a]).

One consequence is these results is that when $n = o(d^2)$, the function $\overline{\boldsymbol{\lambda}}$ converges to the ice cream cone curve $\Omega(x)$ from above. This is a manifestation of the fact that SW^n_d tends to Planch_n as $d \to \infty$, as in Corollary 3.3.4. Indeed, Childs et al. [CHW07] showed that when $n = o(d)$, the two distributions are statistically indistinguishable.

Meliot [Mél10a, Mél10b] has extended Kerov's central limit theorem to the Schur-Weyl distribution, characterizing the fluctuations of $\overline{\boldsymbol{\lambda}}$ around the limiting curves given by Biane. Surprisingly, these fluctuations are identical to Kerov's fluctuations, up to translation in the x-axis.

Theorem 3.7.11. *When $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$ and $n, d \to \infty$ with $\frac{\sqrt{n}}{d} \to c$, the function $\overline{\boldsymbol{\lambda}}(x)$ fluctuates as $\Omega_c(x) + \frac{2}{\sqrt{n}}\boldsymbol{\Delta}(x-c)$, where $\boldsymbol{\Delta}(z)$ is as in Theorem 3.6.4.*

We close this section by recording some simple concentration bounds on the width and length of $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$. They are not as precise as what is suggested by the above limit theorems, but they have the advantage of giving concrete error bounds. We follow a simple line of argument similar to that in [Rom14, Lemma 1.5].

Proposition 3.7.12. *Let $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$. For every $B \in \mathbb{Z}^+$ we have $\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B] \leq \left(\frac{(1+B/d)e^2 n}{B^2}\right)^B$. The same bound holds for $\mathbf{Pr}[\boldsymbol{\lambda}_1' \geq B]$.*

We will typically take $B = \Theta(n/d)$, in which case this bound becomes $\exp(-\Theta(n/d))$.

Proof of Proposition 3.7.12. By Theorem 3.2.4, $\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B]$ (respectively, $\mathbf{Pr}[\boldsymbol{\lambda}_1' \geq B]$) is equal to the probability that a uniformly random word from $[d]^n$ contains a weakly increasing (respectively, strongly increasing) subsequence of length exactly B. As weakly increasing subsequences are more probable than strongly increasing ones, it suffices to bound

$$\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B] \leq \left(\frac{(1+B/d)e^2 n}{B^2}\right)^B.$$

Letting $\boldsymbol{S}$ denote the number of weakly increasing subsequences of length B in a random word we have

$$\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B] \leq \mathbf{E}[\boldsymbol{S}] = \binom{n}{B} \cdot \frac{c}{d^B},$$

where c is the number of words in $[d]^B$ which are weakly increasing. Evidently c also equals the number of "weak d-compositions of B", which [Sta11, Chapter 1.2] is $\binom{d-1+B}{B} \leq \binom{d+B}{B}$. We conclude

$$\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B] \leq \binom{n}{B} \cdot \frac{\binom{d+B}{B}}{d^B} \leq \frac{\left(\frac{en}{B}\right)^B \left(\frac{(1+B/d)ed}{B}\right)^B}{d^B} = \left(\frac{(1+B/d)e^2 n}{B^2}\right)^B,$$

as needed. □

3.8 Polynomial algebras

In Section 2.4.1, we discussed the power sum and Schur polynomials, which are elements of the $\mathbb{C}$-algebra Λ of symmetric polynomials in indeterminates $x_1, x_2, \ldots$.[3] Important to our work will be a closely related polynomial algebra Λ^*, the algebra of *shifted symmetric* polynomials, formally introduced introduced in [OO98b]. This algebra consists of those polynomials which are symmetric in the "shifted" indeterminates $\widetilde{x}_i := x_i - i + c$, where c is any fixed constant. (The definition does not depend on the constant c.) When we view the inputs to the shifted symmetric functions $x_1, x_2, \ldots$ as the values $\lambda_1, \lambda_2, \ldots$ of a partition λ, the result is (isomorphic to) *Kerov's algebra* of polynomial functions on the set of Young diagrams, also known as the *algebra of observables of diagrams*. In a nutshell, the

[3] Strictly speaking, these are *families* of bounded-degree polynomials, one for each number of indeterminates, which are *stable* in the sense that $p_\lambda(x_1, \ldots, x_d, 0) = p_\lambda(x_1, \ldots, x_d)$, and similarly for s_λ. See, e.g., [Mac95] for a formal definition via projective limits.

importance of this algebra is that, on one hand, it still contains polynomials that are similar to "power sums" or "moments" of the λ_i's; and, on the other hand, it is easier to compute their expected value under $\mathrm{SW}^n(\alpha)$ distributions.

We will need to study several families of observables/shifted symmetric polynomials, and their relationships:

Definition 3.8.1. The following polynomials are known to be elements of Λ^*. (We describe the first four as observables of Young diagrams.)

- For $k \geq 1$,
$$p^*_k(\lambda) \coloneqq \sum_{i=1}^{d(\lambda)} \Big((a^*_i)^k - (-b^*_i)^k \Big) = \sum_{i=1}^{\infty} \Big((\lambda_i - i + \tfrac{1}{2})^k - (-i + \tfrac{1}{2})^k \Big).$$
These are the most basic polynomials on Young diagrams, giving the "moments" of the coordinates. For more information on them see [IO02], where they are introduced (in equation (1.4)) under the notation $p_k(\lambda)$. We use the notation $p^*_k(\lambda)$ to distinguish them from the ordinary power sum symmetric polynomials. It is obvious from the second definition above that the p^*_k polynomials are in Λ^*. In fact they are algebraically independent, and they generate Λ^*.

- For $k \geq 0$, the k-th *content sum* polynomial is $c_k(\lambda) \coloneqq \sum_{\square \in [\lambda]} c(\square)^k$. Although these polynomials are quite natural, we will have little occasion to use them. The fact that they are in Λ^* was proven in [KO94].

- For $k \geq 2$,
$$\widetilde{p}_k(\lambda) \coloneqq k(k-1) \int_{-\infty}^{\infty} x^{k-2} \sigma(x)\, dx,$$
where $\sigma(x) \coloneqq \frac{1}{2}(\lambda(x) - |x|)$. These polynomials were introduced and shown to be algebraically independent generators of Λ^* in [IO02, Section 2]. They can shown to be the "moments of the local extrema of $\lambda(x)$", and are also useful for studying continual diagrams. We use them only briefly, to pass between the p^*_k polynomials and $p^\sharp_k$ polynomials defined below.

- For $\lambda \vdash n$ and $\mu \vdash k$, the *central characters* are defined by
$$p^\sharp_\mu(\lambda) = \begin{cases} n^{\downarrow k} \cdot \frac{\chi_\lambda(\mu \cup 1^{n-k})}{\dim(\lambda)} & \text{if } n \geq k, \\ 0 & \text{if } n < k. \end{cases}$$
where $\mu \cup 1^{n-k}$ denotes the partition $(\mu, 1, 1, \dots, 1) \vdash n$. In case $\mu = (k)$ we simply write $p^\sharp_k(\lambda)$. Note that we are somewhat unexpectedly applying the character χ_λ to (an extension of) μ, and not the other way around. The advantage of the $p^\sharp_\mu$ polynomials is that, by virtue of them being characters of the symmetric group (up to some normalizations), their expectations under SW^n_ρ can be easily calculated exactly, as we will see below. A disadvantage is that, by virtue of them being characters of the symmetric group, explicit formulas for them are famously quite complex [Las08, Fér10]

(though in Section 3.8.1 we will mention a formula that allows one to compute $p^\sharp_k$ for small k fairly easily). Wassermann [Was81, III.6] showed that the $p^\sharp_k$ polynomials are in Λ^*, and in fact [VK81, KO94, OO98b] more generally the polynomials $p^\sharp_\mu$ form a *linear* basis of Λ^*.

- For $\mu \vdash k$, the *shifted Schur* polynomial in indeterminates $x_1, \dots, x_d$ is

$$s^*_\mu(x_1, \dots, x_d) = \frac{\det\Big((x_i - i + d)^{\downarrow(d+\mu_j-j)}\Big)_{ij}}{\det\Big((x_i - i + d)^{\downarrow(d-j)}\Big)_{ij}} \quad \text{if } \ell(\mu) \le d, \text{ else } 0.$$

These polynomials are the shifted analogues of the Schur polynomials (cf. Theorem 2.4.9). They were introduced by Okounkov and Olshanski [OO98b], and are similar to the earlier-defined "factorial Schur functions" (see, e.g., [Mac95, I.3.20–21]), but with the advantage that they are *stable*—i.e., $s^*_\mu(x_1, \dots, x_d, 0) = s^*_\mu(x_1, \dots, x_d)$. They arise for us because they can sometimes be used to express the ratio of two Schur functions (see the "Binomial Formula" Theorem 6.3.1). To analyze them, we will use the following "shifted analogue" of Theorem 2.4.10, proved in [OO98b, Theorem 8.1], [IK01, Theorem 9.1] (see also [Mél10b, p.25]):

Theorem 3.8.2. *For $\mu \vdash k$, let us think of the central character polynomial $p^\sharp_\mu$ not as an observable of Young diagrams (applied to $\lambda_1, \dots, \lambda_d$) but as a shifted symmetric polynomial in indeterminates $x_1, \dots, x_d$. In the context of Fourier analysis over the group $G = \mathfrak{S}(k)$, for each fixed $x \in \mathbb{C}^d$ we may think of $p^\sharp_{(\cdot)}(x) \coloneqq \pi \mapsto p^\sharp_\pi(x)$ as a class function. Then its Fourier coefficients are given by*

$$\widetilde{p^\sharp_{(\cdot)}(x)}(\mu) = s^*_\mu(x).$$

(Note that give the determinantal definition of the shifted Schur polynomials, one may alternatively take this Theorem as a definition of the shifted symmetric polynomials $p^\sharp_\mu(x)$.)

As mentioned, the $p^\sharp_\mu$ polynomials are especially important for us as because there is a simple expression for their expectation under any Schur-Weyl distribution. This is the subject of our next proposition.

Proposition 3.8.3. *Let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a probability distribution, and let $\mu \vdash k$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)}[p^\sharp_\mu(\boldsymbol{\lambda})] = n^{\downarrow k} \cdot p_\mu(\alpha_1, \dots, \alpha_d).$$

Proof. It's immediate from the definitions that both sides are 0 if $n < k$, so we assume $n \ge k$. Applying Definition 2.6.2 and the definition of $p^\sharp_\mu$ we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)}[p^\sharp_\mu(\boldsymbol{\lambda})] &= n^{\downarrow k} \cdot \sum_{\lambda \vdash n} s_\lambda(\alpha) \cdot \chi_\lambda(\mu \cup 1^{n-k}) \\ &= n^{\downarrow k} \cdot p_{\mu \cup 1^{n-k}}(\alpha),\end{aligned}$$

where the second equation is from Theorem 2.4.10. But $p_{\mu \cup 1^{n-k}}(\alpha) = p_\mu(\alpha)$, since the two quantities differ only by factors of $p_1(\alpha) = \alpha_1 + \cdots + \alpha_d = 1$. □

Note that in the case of $\alpha_1 = \ldots = \alpha_d = 1/d$, we have that $p_\mu(\alpha) = d^{\ell(\mu)-k}$. This gives us the following important corollary:

Corollary 3.8.4. *Let $\mu \vdash k$. Then* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_d}[p^\sharp_\mu(\boldsymbol{\lambda})] = n^{\downarrow k} \cdot d^{\ell(\mu)-k}$.

Applying Corollary 3.3.4 and the fact that $\ell(\mu) = k$ if $\mu = (k)$ and otherwise $\ell(\mu) < k$, we get the following corollary.

Corollary 3.8.5. *Let $\mu \vdash k$. Then* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{Planch}_n}[p^\sharp_\mu(\boldsymbol{\lambda})] = \begin{cases} n^{\downarrow k} & \text{if } \mu = (k), \\ 0 & \text{otherwise.} \end{cases}$

3.8.1 Working with the $p^\sharp_\mu$ polynomials

As we will be working heavily with the $p^\sharp_\mu$ polynomials, let us describe them further. Computing $p^\sharp_\mu(\lambda)$ is polynomial-time equivalent to computing the character $\chi_\lambda(\mu \cup 1^{n-k})$ because the $\dim(\lambda)$ term in the denominator is easily computed by the Hook-Length Formula (Definition 2.2.12). Unfortunately, computing characters of the symmetric group is known to be #P-complete [Hep94], and even deciding whether a character is nonzero is NP-hard [PP14]. This means that the $p^\sharp_\mu$ polynomials, especially when $|\mu|$ is large, are somewhat inexplicit and rather cumbersome to work with. When $|\mu|$ is small, however, explicit formulas for $p^\sharp_\mu$ can be computed by the following theorem.

Theorem 3.8.6. *The function $p^\sharp_\mu$ is the unique element of Λ^* such that the highest degree term of $p^\sharp_\mu$ is the power sum symmetric polynomial p_μ and*

$$p^\sharp_\mu(\lambda) = 0$$

for all λ such that $|\lambda| < |\mu|$.

This characterization theorem is stated in the talk of Okounkov [Oko08], where he credits it to the papers of [VK81] and [KO94], though we could not find it written explicitly in either of these works. However, it follows immediately from known results by applying Theorems 2.4.10 and 3.8.2 to a similar characterization theorem for shifted Schur functions (which is stated identically except that the highest degree term of s^*_μ is the ordinary Schur function s_μ) given by [OO98b].

As pointed out by Okounkov [Oko08], we can use Theorem 3.8.6 to obtain a relatively simple finitary method for expressing $p^\sharp_\mu$ polynomials in terms of the p^*_j polynomials. The case when $\mu = (k)$ is the most useful for us, so we will begin with it. By [IO02, Proposition 3.4],

$$p^\sharp_k = p^*_k + \Big\{\text{polynomial in } p^*_1, \ldots, p^*_{k-1} \text{ of gradation at most } k-1\Big\}, \tag{3.7}$$

where *gradation* refers to the canonical grading in which $\prod_i p^*_{\lambda_i}$ has gradation $|\lambda|$. Hence, computing $p^\sharp_k$ involves determining the coefficients of this polynomial of gradation at most $k-1$, and the vanishing condition of Theorem 3.8.6 allows us to generate linear equations in these unknowns, which can then be solved for by computer. In particular, we can deduce

$$p^\sharp_1 = p^*_1, \qquad p^\sharp_2 = p^*_2, \qquad p^\sharp_3 = p^*_3 - \tfrac{3}{2}(p^*_1)^2 + \tfrac{5}{4}p^*_1, \qquad p^\sharp_4 = p^*_4 - 4p^*_2p^*_1 + \tfrac{11}{2}p^*_2. \tag{3.8}$$

(Having deduced these equations, Theorem 3.8.6 allows us to verify them easily.) We can of course inductively invert (3.7), deducing that

$$p^*_k = p^\sharp_k + \Big\{\text{polynomial in } p^\sharp_1, \dots, p^\sharp_{k-1} \text{ of gradation at most } k-1\Big\}. \tag{3.9}$$

For example,

$$p^*_1 = p^\sharp_1, \qquad p^*_2 = p^\sharp_2, \qquad p^*_3 = p^\sharp_3 + \tfrac{3}{2}(p^\sharp_1)^2 - \tfrac{5}{4}p^\sharp_1, \qquad p^*_4 = p^\sharp_4 + 4p^\sharp_2 p^\sharp_1 - \tfrac{11}{2}p^\sharp_2. \tag{3.10}$$

This methodology can be extended to general μ via [IO02, Proposition 4.2]:

$$p^\sharp_\mu = p^*_\mu + \Big\{\text{polynomial in } p^*_1, \dots, p^*_{|\mu|-1} \text{ of gradation at most } |\mu|-1\Big\},$$

where $p^*_\mu = \prod_i p^*_{\mu_i}$

Recall that the more general $p^\sharp_\tau$ polynomials (for $\tau \in \mathrm{Par}$) are known to linearly generate the algebra of observables. This means that any product $p^\sharp_{\mu_1} p^\sharp_{\mu_2}$ can be converted to a linear combination of $p^\sharp_\tau$'s. In particular, if we applied this conversion in (3.10) we would get linear expressions for the "low-degree moments of Young diagrams" (i.e., the p^*_j's) in terms of $p^\sharp_\tau$'s; we could then compute the expectation of these, under any Schur-Weyl distribution, using Proposition 3.8.3.

We are therefore interested in the *structure constants* $f^\tau_{\mu_1\mu_2}$ of Λ^* in the basis $\{p^\sharp_\tau\}$; i.e., the numbers such that

$$p^\sharp_{\mu_1} p^\sharp_{\mu_2} = \sum_{\tau \in \mathrm{Par}} f^\tau_{\mu_1\mu_2} p^\sharp_\tau.$$

These were first determined by Ivanov and Kerov [IK01] in terms of the algebra of *partial permutations*. We quote the following formulation from [IO02, Proposition 4.5]:

Proposition 3.8.7. *Let $\tau, \mu_1, \mu_2 \in \mathrm{Par}$. Fix a set R of cardinality $|\tau|$ and a permutation $w : R \to R$ of cycle type τ. Then*

$$f^\tau_{\mu_1\mu_2} = \frac{z_{\mu_1} z_{\mu_2}}{z_\tau} g^\tau_{\mu_1\mu_2},$$

where $g^\tau_{\mu_1\mu_2}$ equals the number of quadruples (R_1, w_1, R_2, w_2) such that:

1. *$R_1 \subseteq R, \quad R_2 \subseteq R, \quad R_1 \cup R_2 = R;$*
2. *$|R_i| = |\mu_i|$ and $w_i : R_i \to R_i$ is a permutation of cycle type μ_i, for $i = 1, 2$;*
3. *$\overline{w}_1 \overline{w}_2 = w$, where $\overline{w}_i : R \to R$ denotes the natural extension of w_i from R_i to the whole of R.*

We present an equivalent formulation we have found to be more convenient. We omit its straightforward combinatorial deduction from Proposition 3.8.7.

Corollary 3.8.8. *Let*

$$C^t_{r_1 r_2} := \frac{r_1! r_2!}{(t-r_1)!(t-r_2)!(r_1+r_2-t)!}$$

if the positive integers r_1, r_2, t *satisfy* $r_1, r_2 \le t \le r_1 + r_2$, *and let* $C^t_{r_1 r_2} := 0$ *otherwise. Then for* $\mu \vdash r_1$, $\nu \vdash r_2$, $\tau \vdash t$,

$$f^\tau_{\mu\nu} = C^t_{r_1 r_2} \cdot \mathop{\mathbf{Pr}}_{\boldsymbol{w}_1, \boldsymbol{w}_2} \left[\overline{\boldsymbol{w}}_1 \overline{\boldsymbol{w}}_2 \text{ has cycle type } \tau \right],$$

where $\boldsymbol{w}_1$ *is a uniformly random permutation on* $\{1, \ldots, r_1\}$ *of cycle type* μ, *and* $\boldsymbol{w}_2$ *is a uniformly random permutation on* $\{t - r_2 + 1, \ldots, t\}$ *of cycle type* ν.

As very simple examples, we can compute

$$(p^\sharp_1)^2 = p^\sharp_{(1,1)} + p^\sharp_1, \qquad p^\sharp_2 p^\sharp_1 = p^\sharp_{(2,1)} + 2p^\sharp_2, \qquad (p^\sharp_2)^2 = p^\sharp_{(2,2)} + 4p^\sharp_3 + 2p^\sharp_{(1,1)}. \tag{3.11}$$

Substituting these into (3.8), we obtain the formulas

$$p^*_1 = p^\sharp_1, \qquad p^*_2 = p^\sharp_2, \qquad p^*_3 = p^\sharp_3 + \tfrac{3}{2} p^\sharp_{(1,1)} + \tfrac{1}{4} p^\sharp_1, \qquad p^*_4 = p^\sharp_4 + 4p^\sharp_{(2,1)} + \tfrac{5}{2} p^\sharp_2, \tag{3.12}$$

which will be useful to us later.

Given the formula for the structure constants, it's not hard to show that

$$p^\sharp_\mu p^\sharp_\nu = p^\sharp_{\mu \cup \nu} + \left\{ \text{linear combination of } p^\sharp_\tau\text{'s with } |\tau| < |\mu \cup \nu| \right\},$$

where $\mu \cup \nu$ denotes the partition formed by joining the parts of μ and ν and sorting them in nonincreasing order (i.e., $m_w(\mu \cup \nu) = m_w(\mu) + m_w(\nu)$). In fact, we will require a stronger statement, based on the following notion introduced in [IK01]:

Definition 3.8.9. For a partition $\lambda \in \mathrm{Par}$, its *weight* is defined to be $\mathrm{wt}(\lambda) = |\lambda| + \ell(\lambda)$.

Now Śniady [Śni06, Corollary 3.8] proved:

Proposition 3.8.10. $p^\sharp_\mu p^\sharp_\nu = p^\sharp_{\mu \cup \nu} + \left\{ \text{linear combination of } p^\sharp_\tau\text{'s with } \mathrm{wt}(\tau) \le \mathrm{wt}(\mu) + \mathrm{wt}(\nu) - 2 \right\}$.

At one point in Chapter 7, we will need explicit bounds on the coefficients of the polynomial that appears in Equation (3.7), and this involves analyzing some expressions that arise in the proof of this equation. The proof in [IO02] is based on an identity from [Was81, III.6] (cf. [IO02, Proposition 3.3]) which uses generating functions:

$$p^\sharp_k = [t^{k+1}] \left\{ -\frac{1}{k} \prod_{j=1}^{k} (1 - (j - \tfrac{1}{2})t) \cdot \exp\left(\sum_{j=1}^{\infty} \frac{p^*_j t^j}{j} (1 - (1 - kt)^{-j}) \right) \right\}.$$

One may rewrite this (cf. [IO02, (3.3)]) as

$$p^\sharp_k = [t^{k+1}] \left\{ -\frac{1}{k} \prod_{j=1}^{k} (1 - (j - \tfrac{1}{2})t) \cdot \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} Q_k(t)^i \right\}, \tag{3.13}$$

where

$$Q_k(t) = \sum_{m=1}^{\infty} Q_{k,m} t^{m+1}, \quad Q_{k,m} = \tfrac{1}{1}\tbinom{m}{0} k^m p^*_1 + \tfrac{1}{2}\tbinom{m}{1} k^{m-1} p^*_2 + \tfrac{1}{3}\tbinom{m}{2} k^{m-2} p^*_3 + \cdots + \tfrac{1}{m}\tbinom{m}{m-1} k p^*_m. \tag{3.14}$$

It follows that in (3.13) we may restrict the sum on i to the range between 0 and $\frac{k+1}{2}$, and in (3.14) we can restrict the sum on m to the range between 1 and k. This gives a separate, though less convenient, method for expressing the $p^\sharp_k$ polynomials in terms of the p^*_j's. Furthermore, one can derive (3.7) from this expression.

Chapter 4

Spectrum estimation

In this section, we consider the following natural algorithm for estimating the spectrum α of a mixed state $\rho \in \mathbb{C}^{d\times d}$.

Definition 4.0.1 (Empirical Young diagram algorithm)**.** Given $\rho^{\otimes n}$:

1. Sample $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$.
2. Output $\dfrac{\boldsymbol{\lambda}}{n} = \left(\dfrac{\boldsymbol{\lambda}_1}{n}, \dots, \dfrac{\boldsymbol{\lambda}_d}{n}\right)$.

We will sometimes write $\underline{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda}/n$.

This algorithm was first proposed by Alicki, Rudnicki, and Sadowski [ARS88] who not only showed that $\boldsymbol{\lambda}/n \to \alpha$ almost surely as $n \to \infty$, but also proved a Gaussian central limit theorem for $\boldsymbol{\lambda}/n$. Later, it was independently suggested and analyzed by Keyl and Werner [KW01]. In a non-quantum context, the fact that $\boldsymbol{\lambda}/n \to \alpha$ almost surely was, to our knowledge, first shown by Kerov and Vershik in [KV86]. We note that though this is a natural algorithm for spectrum estimation, it is *not* in general an unbiased estimator for α (see Lemma 4.1.2 below).

Our first result upper bounds the expected ℓ_2^2 error of the EYD algorithm, proving Theorem 1.4.6.

Theorem 4.0.2 (Theorem 1.4.6 restated)**.** $\mathop{\mathbf{E}}\limits_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \|\underline{\boldsymbol{\lambda}} - \alpha\|_2^2 \leq \dfrac{d}{n}$.

Corollary 1.3.6 follows as an immediate consequence, giving an $n = O(d^2/\epsilon^2)$ bound for spectrum estimation. Previously, the best bound for spectrum estimation was $= O(d^2/\epsilon \cdot \log(d/\epsilon))$ [HM02, CM06].

We will also study the following natural modification of the EYD algorithm for solving truncated spectrum estimation.

Definition 4.0.3 (Truncated EYD algorithm)**.** Given $\rho^{\otimes n}$ and an integer $k \in [d]$:

1. Sample $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$.
2. Output $\left(\dfrac{\boldsymbol{\lambda}_1}{n}, \dots, \dfrac{\boldsymbol{\lambda}_k}{n}\right)$.

Our next result upper bounds the expected trace distance error of the truncated EYD algorithm, proving Theorem 1.4.8.

Theorem 4.0.4 (Theorem 1.4.8 restated)**.** $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} d_{\mathrm{TV}}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha) \leq \dfrac{1.92\,k + .5}{\sqrt{n}}$.

We complement our upper bounds with a matching lower bound for the EYD algorithm, proving Theorem 1.4.10.

Theorem 4.0.5 (Theorem 1.4.10 restated)**.** *If $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state, then the EYD algorithm* fails *to give an ϵ-accurate estimate in total variation distance with high probability unless $\Omega(d^2/\epsilon^2)$ copies are used.*

In general, the best lower bound for spectrum estimation for all algorithms is $n = \Omega(d/\epsilon^2)$, which follows from our Theorem 1.4.23. It is an interesting open question to determine the copy complexity of spectrum estimation.

This chapter is organized as follows:

- Section 4.1 gives the proof of Theorem 4.0.2.
- Section 4.2 gives the proof of Theorem 4.0.4.
- Section 4.3 gives the proof of Theorem 4.0.5.

4.1 Spectrum estimation

We give two lemmas and then the proof of Theorem 4.0.2.

Lemma 4.1.1. *Let $\alpha \in \mathbb{R}^d$ be a probability distribution. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} \sum_{i=1}^d \boldsymbol{\lambda}_i^2 \leq \sum_{i=1}^d (n\alpha_i)^2 + dn.$$

Proof. Define the polynomial function

$$p_2^*(\lambda) = \sum_{i=1}^{\ell(\lambda)} \left((\lambda_i - i + \tfrac{1}{2})^2 - (-i + \tfrac{1}{2})^2\right).$$

By Proposition 2.34 and equation (12) of [OW15b], $\mathbf{E}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)}[p_2^*(\boldsymbol{\lambda})] = n(n-1)\cdot \sum_{i=1}^d \alpha_i^2$. Hence,

$$\mathbf{E}\sum_{i=1}^d \boldsymbol{\lambda}_i^2 = \mathbf{E}\left[p_2^*(\boldsymbol{\lambda}) + \sum_{i=1}^d (2i-1)\boldsymbol{\lambda}_i\right] \leq \mathbf{E}\, p_2^*(\boldsymbol{\lambda}) + \sum_{i=1}^d (2i-1)(n/d) \leq n^2 \cdot \sum_{i=1}^d \alpha_i^2 + dn.$$

Here the first inequality used inequality (1.9) and $\boldsymbol{\lambda} \succ (n/d, \ldots, n/d)$. □

Lemma 4.1.2. *Let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, where $\alpha \in \mathbb{R}^d$ is a sorted probability distribution. Then $(\mathbf{E}\,\boldsymbol{\lambda}_1, \ldots, \mathbf{E}\,\boldsymbol{\lambda}_d) \succ (\alpha_1 n, \ldots, \alpha_d n)$.*

Proof. Let $\boldsymbol{w} \sim \alpha^{\otimes n}$, so $\boldsymbol{\lambda}$ is distributed as $\mathrm{shRSK}(\boldsymbol{w})$. The proof is completed by linearity of expectation applied to the fact that $(\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_d) \succ (\#_1\boldsymbol{w}, \dots, \#_d\boldsymbol{w})$ always, where $\#_k\boldsymbol{w}$ denotes the number of times letter k appears in $\boldsymbol{w}$. In turn this fact holds by Greene's Theorem: we can form k disjoint increasing subsequences in $\boldsymbol{w}$ by taking all its 1's, all its 2's, $\dots$, all its k's. □

Proof of Theorem 4.0.2. We have

$$\begin{aligned}
n^2 \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} \|\underline{\boldsymbol{\lambda}} - \alpha\|_2^2 &= \mathbf{E} \sum_{i=1}^{d} (\boldsymbol{\lambda}_i - \alpha_i n)^2 = \mathbf{E} \sum_{i=1}^{d} (\boldsymbol{\lambda}_i^2 + (\alpha_i n)^2) - 2 \sum_{i=1}^{d} (\alpha_i n) \cdot \mathbf{E}\, \boldsymbol{\lambda}_i \\
&\le dn + 2 \sum_{i=1}^{d} (\alpha_i n)^2 - 2 \sum_{i=1}^{d} (\alpha_i n) \cdot \mathbf{E}\, \boldsymbol{\lambda}_i \le dn + 2 \sum_{i=1}^{d} (\alpha_i n)^2 - 2 \sum_{i=1}^{d} (\alpha_i n) \cdot (\alpha_i n) = dn,
\end{aligned}$$

where the first inequality used Lemma 4.1.1 and the second used Lemma 4.1.2 and inequality (1.9) (recall that the coefficients $\alpha_i n$ are decreasing). Dividing by n^2 completes the proof. □

4.2 Truncated spectrum estimation

The key lemma involved in the proof of Theorem 4.0.4 is the following:

Lemma 4.2.1. *Let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then for any $k \in [d]$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} \sum_{i=1}^{k} \boldsymbol{\lambda}_i \le \sum_{i=1}^{k} \alpha_i n + 2\sqrt{2}k\sqrt{n}.$$

We remark that it is easy to *lower*-bound this expectation by $\sum_{i=1}^{k} \alpha_i n$ via Lemma 4.1.2. We now show how to deduce Theorem 4.0.4 from Lemma 4.2.1. Then in Section 4.2.1 we prove the lemma.

Proof of Theorem 4.0.4. Let $\boldsymbol{w} \sim \alpha^{\otimes n}$, let $\mathrm{RSK}(\boldsymbol{w}) = (\boldsymbol{P}, \boldsymbol{Q})$, and let $\boldsymbol{\lambda} = \mathrm{sh}(\boldsymbol{P})$, so $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$. Write $\boldsymbol{w}'$ for the string formed from $\boldsymbol{w}$ by deleting all letters bigger than k. Then it is a basic property of the RSK algorithm that $\mathrm{RSK}(\boldsymbol{w}')$ produces the insertion tableau $\boldsymbol{P}'$ formed from $\boldsymbol{P}$ by deleting all boxes with labels bigger than k. Thus $\boldsymbol{\lambda}' = \mathrm{sh}(\boldsymbol{P}') = \mathrm{shRSK}(\boldsymbol{w}')$. Denoting $\alpha_{[k]} = \alpha_1 + \cdots + \alpha_k$, we have $\boldsymbol{\lambda}' \sim \mathrm{SW}^{\boldsymbol{m}}(\alpha')$, where $\boldsymbol{m} \sim \mathrm{Binomial}(n, \alpha_{[k]})$ and α' denotes α conditioned on the first k letters; i.e., $\alpha' = (\alpha_i/\alpha_{[k]})_{i=1}^{k}$. Now by the triangle inequality,

$$2n \cdot \mathbf{E}\, d_{\mathrm{TV}}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha) = \mathbf{E} \sum_{i=1}^{k} |\boldsymbol{\lambda}_i - \alpha_i n| \le \mathbf{E} \sum_{i=1}^{k} (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}'_i) + \mathbf{E} \sum_{i=1}^{k} |\boldsymbol{\lambda}'_i - \alpha'_i \boldsymbol{m}| + \sum_{i=1}^{k} |\alpha'_i \boldsymbol{m} - \alpha_i n| \,. \tag{4.1}$$

The first quantity in (4.1) is at most $2\sqrt{2}k\sqrt{n}$, using Lemma 4.2.1 and the fact that $\mathbf{E}[\sum_{i=1}^k \boldsymbol{\lambda}'_i] = \mathbf{E}[\boldsymbol{m}] = \sum_{i=1}^k \alpha_i n$. The second quantity in (4.1) is at most $k\sqrt{n}$ using Theorem 4.0.2:

$$\mathbf{E}\sum_{i=1}^k |\boldsymbol{\lambda}'_i - \alpha'_i \boldsymbol{m}| = \mathop{\mathbf{E}}_{\boldsymbol{m}} \boldsymbol{m} \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}'} \|\underline{\boldsymbol{\lambda}}' - \alpha'\|_1 \le \mathop{\mathbf{E}}_{\boldsymbol{m}} \boldsymbol{m}\sqrt{k}\sqrt{\mathop{\mathbf{E}}_{\boldsymbol{\lambda}'} \|\underline{\boldsymbol{\lambda}}' - \alpha'\|_2^2} \le k \mathop{\mathbf{E}}_{\boldsymbol{m}} \sqrt{\boldsymbol{m}} \le k\sqrt{n}.$$

And the third quantity in (4.1) is at most $\sqrt{n}$:

$$\mathop{\mathbf{E}}_{\boldsymbol{m}} \sum_{i=1}^k |\alpha'_i \boldsymbol{m} - \alpha_i n| = \mathop{\mathbf{E}}_{\boldsymbol{m}} \sum_{i=1}^k \tfrac{\alpha_i}{\alpha_{[k]}} \left|\boldsymbol{m} - \alpha_{[k]} n\right| = \mathop{\mathbf{E}}_{\boldsymbol{m}} |\boldsymbol{m} - \alpha_{[k]} n| \le \mathbf{stddev}(\boldsymbol{m}) \le \sqrt{n}.$$

Thus $2n \cdot \mathbf{E}\, d_{\mathrm{TV}}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha) \le ((2\sqrt{2}+1)k+1)\sqrt{n}$, and dividing by $2n$ completes the proof. □

4.2.1 Proof of Lemma 4.2.1

Our proof of Lemma 4.2.1 is essentially by reduction to the case when α is the uniform distribution and $k = 1$. We thus begin by analyzing the uniform distribution.

The uniform distribution case

In this subsection we will use the abbreviation $(1/d)$ for the uniform distribution $(1/d, \dots, 1/d)$ on $[d]$. Our goal is the following fact, which is of independent interest:

Theorem 4.2.2. $\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(1/d)} \boldsymbol{\lambda}_1 \le n/d + 2\sqrt{n}$.

We remark that Theorem 4.2.2 implies Lemma 4.2.1 (with a slightly better constant) in the case of $\alpha = (1/d, \dots, 1/d)$, since of course $\boldsymbol{\lambda}_i \le \boldsymbol{\lambda}_1$ for all $i \in [k]$. Also, by taking $d \to \infty$ we recover the well known fact that $\mathbf{E}\,\boldsymbol{\lambda}_1 \le 2\sqrt{n}$ when $\boldsymbol{\lambda}$ has the Plancherel distribution. Indeed, our proof of Theorem 4.2.2 extends the original proof of this fact by Vershik and Kerov [VK85], which we presented in Section 3.5.4 (cf. the exposition in [Rom14]).

Proof. Consider the Schur-Weyl growth process under the uniform distribution $(1/d, \dots, 1/d)$ on $[d]$. For $m \ge 1$ we define

$$\delta_m = \mathbf{E}[\boldsymbol{\lambda}_1^{(m)} - \boldsymbol{\lambda}_1^{(m-1)}] = \mathbf{Pr}[\text{the } m\text{-th box enters into the 1st row}] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^{m-1}(1/d)} \frac{s_{\boldsymbol{\lambda}+e_1}(1/d)}{s_{\boldsymbol{\lambda}}(1/d)},$$

where we used Corollary 3.4.3. By Cauchy–Schwarz,

$$\begin{aligned}
\delta_m^2 &\le \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^{m-1}(1/d)} \left(\frac{s_{\boldsymbol{\lambda}+e_1}(1/d)}{s_{\boldsymbol{\lambda}}(1/d)}\right)^2 = \sum_{\lambda \vdash m-1} \dim(\lambda) s_\lambda(1/d) \cdot \left(\frac{s_{\lambda+e_1}(1/d)}{s_\lambda(1/d)}\right)^2 \\
&= \sum_{\lambda \vdash m-1} \dim(\lambda) s_{\lambda+e_1}(1/d) \cdot \left(\frac{s_{\lambda+e_1}(1/d)}{s_\lambda(1/d)}\right) = \sum_{\lambda \vdash m-1} \dim(\lambda+e_1) s_{\lambda+e_1}(1/d) \cdot \left(\frac{d+\lambda_1}{dm}\right) \qquad (4.2) \\
&\le \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^m(1/d)} \left(\frac{d+\boldsymbol{\lambda}_1}{dm}\right) = \left(\frac{d+\delta_1+\ldots+\delta_m}{dm}\right),
\end{aligned}$$

where the ratio in (4.2) was computed using the first formula of Definition 2.2.13 (and the homogeneity of Schur polynomials). Thus we have established the following recurrence:

$$\delta_m \le \frac{1}{\sqrt{dm}}\sqrt{d+\delta_1+\cdots+\delta_m}. \tag{4.3}$$

We will now show by induction that $\delta_m \le \frac{1}{d} + \frac{1}{\sqrt{m}}$ for all $m \ge 1$. Note that this will complete the proof, by summing over $m \in [n]$. The base case, $m = 1$, is immediate since $\delta_1 = 1$. For general $m > 1$, think of $\delta_1, \dots, \delta_{m-1}$ as fixed and δ_m as variable. Now if δ_m satisfies (4.3), it is bounded above by the (positive) solution δ^* of

$$\delta = \frac{1}{\sqrt{dm}}\sqrt{c+\delta}, \qquad \text{where } c = d+\delta_1+\cdots+\delta_{m-1}.$$

Note that if $\delta > 0$ satisfies

$$\delta \ge \frac{1}{\sqrt{dm}}\sqrt{c+\delta} \tag{4.4}$$

then it must be that $\delta \ge \delta^* \ge \delta_m$. Thus it suffices to show that (4.4) holds for $\delta = \frac{1}{d} + \frac{1}{\sqrt{m}}$. But indeed,

$$\frac{1}{\sqrt{dm}}\sqrt{c+\frac{1}{d}+\frac{1}{\sqrt{m}}} = \frac{1}{\sqrt{dm}}\sqrt{d+\delta_1+\cdots+\delta_{m-1}+\frac{1}{d}+\frac{1}{\sqrt{m}}}$$

$$\le \frac{1}{\sqrt{dm}}\sqrt{d+\sum_{i=1}^{m}\left(\frac{1}{d}+\frac{1}{\sqrt{i}}\right)} \le \frac{1}{\sqrt{dm}}\sqrt{d+\frac{m}{d}+2\sqrt{m}} = \frac{1}{\sqrt{dm}}\left(\sqrt{d}+\sqrt{\frac{m}{d}}\right) = \frac{1}{d}+\frac{1}{\sqrt{m}},$$

where the first inequality used induction. The proof is complete. □

Reduction to the uniform case

Proof of Lemma 4.2.1. Given the sorted distribution α on $[d]$, let β be the sorted probability distribution on $[d]$ defined, for an appropriate value of m, as

$$\beta_1 = \alpha_1, \dots, \beta_k = \alpha_k, \quad \beta_{k+1} = \ldots = \beta_m = \alpha_{k+1} > \beta_{m+1} \ge 0, \quad \beta_{m+2} = \ldots = \beta_d = 0.$$

In other words, β agrees with α on the first k letters and is otherwise uniform, except for possibly a small "bump" at β_{m+1}. By construction we have $\beta \succ \alpha$. Thus it follows from our coupling result, Theorem 1.5.4, that

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} \sum_{i=1}^{k} \boldsymbol{\lambda}_i \le \mathop{\mathbf{E}}_{\boldsymbol{\mu}\sim \mathrm{SW}^n(\beta)} \sum_{i=1}^{k} \boldsymbol{\mu}_i,$$

and hence it suffices to prove the lemma for β in place of α. Observe that β can be expressed as a mixture

$$\beta = p_1 \cdot \mathcal{D}_1 + p_2 \cdot \mathcal{D}_2 + p_3 \cdot \mathcal{D}_3, \tag{4.5}$$

of a certain distribution $\mathcal{D}_1$ supported on $[k]$, the uniform distribution $\mathcal{D}_2$ on $[m]$, and the uniform distribution $\mathcal{D}_3$ on $[m+1]$. We may therefore think of a draw $\boldsymbol{\mu} \sim \mathrm{SW}^n(\beta)$ occurring as follows. First, $[n]$ is partitioned into three subsets $\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3$ by including each $i \in [n]$ into $\boldsymbol{I}_j$ independently with probability p_j. Next we draw strings $\boldsymbol{w}^{(j)} \sim \mathcal{D}_j^{\otimes \boldsymbol{I}_j}$ independently for $j \in [3]$. Finally, we let $\boldsymbol{w} = (\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \boldsymbol{w}^{(3)}) \in [d]^n$ be the natural composite string and define $\boldsymbol{\mu} = \mathrm{shRSK}(\boldsymbol{w})$. Let us also write $\boldsymbol{\mu}^{(j)} = \mathrm{shRSK}(\boldsymbol{w}^{(j)})$ for $j \in [3]$. We now claim that

$$\sum_{i=1}^{k} \boldsymbol{\mu}_i \leq \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(1)} + \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(2)} + \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(3)}$$

always holds. Indeed, this follows from Greene's Theorem: the left-hand side is $|\boldsymbol{s}|$, where $\boldsymbol{s} \in [d]^n$ is a maximum-length disjoint union of k increasing subsequences in $\boldsymbol{w}$; the projection of $\boldsymbol{s}^{(j)}$ onto coordinates $\boldsymbol{I}_j$ is a disjoint union of k increasing subsequences in $\boldsymbol{w}^{(j)}$ and hence the right-hand side is at least $|\boldsymbol{s}^{(1)}| + |\boldsymbol{s}^{(2)}| + |\boldsymbol{s}^{(3)}| = |\boldsymbol{s}|$. Thus to complete the proof of the lemma, it suffices to show

$$\mathbf{E} \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(1)} + \mathbf{E} \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(2)} + \mathbf{E} \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(3)} \leq \sum_{i=1}^{k} \alpha_i n + 2\sqrt{2}\, k\sqrt{n}. \tag{4.6}$$

Since $\mathcal{D}_1$ is supported on $[k]$, the first expectation above is equal to $\mathbf{E}[|\boldsymbol{w}^{(1)}|] = p_1 n$. By (the remark just after) Theorem 4.2.2, we can bound the second expectation as

$$\mathbf{E} \sum_{i=1}^{k} \boldsymbol{\mu}_i^{(2)} \leq k\, \mathbf{E}\, \boldsymbol{\mu}_1^{(2)} \leq k\, \mathbf{E}\, |\boldsymbol{w}^{(2)}|/m + 2k\, \mathbf{E} \sqrt{|\boldsymbol{w}^{(2)}|} \leq k(p_2 n)/m + 2k\sqrt{p_2 n}.$$

Similarly the third expectation in (4.6) is bounded by $k(p_3 n)/(m+1) + 2k\sqrt{p_3 n}$. Using $\sqrt{p_2} + \sqrt{p_3} \leq \sqrt{2}$, we have upper-bounded the left-hand side of (4.6) by

$$(p_1 + p_2 \tfrac{k}{m} + p_3 \tfrac{k}{m+1})n + 2\sqrt{2}\, k\sqrt{n} = \left(\sum_{i=1}^{k} \beta_i\right) n + 2\sqrt{2}\, k\sqrt{n},$$

as required. □

4.3 The lower bound

In this section, we prove Theorem 4.0.5. Since ρ is the maximally mixed state, its spectrum is the uniform distribution Unif_d, and so our goal is to show that the $\underline{\boldsymbol{\lambda}}$ output by the EYD algorithm is ϵ-far from Unif_d with constant probability unless n is sufficiently large.

Theorem 4.3.1. *There is a $\delta > 0$ such that for sufficiently small values of ϵ,*

$$\Pr_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}[d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) > \epsilon] \geq \delta$$

unless $n = \Omega(d^2/\epsilon^2)$.

We will split the lower bound into two cases.

Theorem 4.3.2. *For every constant $C > 0$, there are constants $\delta, \epsilon > 0$ such that*

$$\Pr_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}[d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) > \epsilon] \geq \delta$$

when $n < C \cdot d^2$ and d is sufficiently large.

Theorem 4.3.3. *There are absolute constants $C > 0$ and $0 < \delta < 1$ such that*

$$\Pr_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}[d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) > \epsilon] \geq \delta$$

when $n \geq C \cdot d^2$, unless $n = \Omega(d^2/\epsilon^2)$.

To prove Theorem 4.3.1, let C and δ_1 be the constants in Theorem 4.3.3. Apply Theorem 4.3.2 with the value of C, and let δ_2 and ϵ_0 be the resulting constants. Set $\delta := \min\{\delta_1, \delta_2\}$. Then we see that for all $\epsilon \leq \epsilon_0$,

$$\Pr_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}[d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d)) > \epsilon] \geq \delta$$

unless $n = \Omega(d^2/\epsilon^2)$, giving Theorem 4.3.1.

Theorem 4.3.2 might look unnecessary, as Theorem 4.3.3 already proves the lower bound for sufficiently large values of n (i.e., $n \geq C \cdot d^2$), and intuitively having fewer copies of ρ shouldn't improve the performance of the EYD algorithm. However, this intuition, though it may be true in some approximate sense, is false in general: there are regimes of state estimation where the performance of the EYD algorithm does *not* increase monotonically with the value of n. For example, if n is a multiple of d, then when $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$, $\underline{\boldsymbol{\lambda}}$ will equal Unif_d with some nonzero probability. On the other hand, a random $\boldsymbol{\lambda} \sim \mathrm{SW}_d^{n+1}$ will never be uniform, because $n+1$ is not a multiple of d. Thus, decreasing the value of n can sometimes help (according to some performance metrics), and this shows why we need Theorem 4.3.2 to supplement Theorem 4.3.3.

The proof of Theorem 4.3.2 is quite technical, and we defer it to Section 4.3.1. Our proof of Theorem 4.3.3 is simpler and appears below. It is a good illustration of the basic technique of using polynomial functions on Young diagrams. The intuition behind the proof is as follows: By the (traceless) Gaussian Unitary Ensemble fluctuations predicted in [ITW01], we expect that for $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$, the empirical distribution $\underline{\boldsymbol{\lambda}}$ will deviate from Unif_d by roughly $\Theta(1/\sqrt{n})$ in each coordinate. This will yield total variation distance $\Theta(d/\sqrt{n})$, necessitating $n \geq \Omega(d^2/\epsilon^2)$ to achieve $d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) \leq \epsilon$. Actually analyzing the precise rate of convergence to Gaussian fluctuations in terms of n is difficult, and is overkill anyway; instead, we use the Fourth Moment Method to lower bound the fluctuations.

Proof of Theorem 4.3.3. Our goal is to show that for $n \geq 10^{10}d^2$, with 1% probability over a random $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$, at least $\frac{d}{200}$ coordinates $i \in [d]$ satisfy

$$\left|\boldsymbol{\lambda}_i - \frac{n}{d}\right| \geq \frac{\sqrt{n}}{1000}.$$

When this event occurs,

$$d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) = \frac{1}{2} \cdot \sum_{i=1}^{d} \left| \frac{\boldsymbol{\lambda}_i}{n} - \frac{1}{d} \right| = \frac{1}{2} \cdot \sum_{i=1}^{d} \frac{1}{n} \cdot \left| \boldsymbol{\lambda}_i - \frac{n}{d} \right|$$
$$\geq \frac{1}{2} \cdot \frac{d}{200} \cdot \frac{1}{n} \cdot \frac{\sqrt{n}}{1000} = \frac{1}{400000} \cdot \frac{d}{\sqrt{n}},$$

which is bigger than ϵ unless $n = \Omega(d^2/\epsilon^2)$. Showing this will prove Theorem 4.3.3 with the parameters $C = 10^{10}$ and $\delta = .01$.

To begin, let us define a family of polynomials.

Definition 4.3.4. Given $k \geq 1$ and $c \in \mathbb{R}$, we define $p^*_{k,c}(\lambda) := \sum_{i=1}^{\infty} (\lambda_i - i - c)^k - (-i-c)^k$. This generalizes the definition of the p^*_k polynomials, as $p^*_{k,-\frac{1}{2}} = p^*_k$.

Fact 4.3.5. *Let $c \in \mathbb{R}$. Then*

- $p^*_{2,c} = (-2c-1)p^\sharp_1 + p^\sharp_2$, *and*
- $p^*_{4,c} = (-4c^3 - 6c^2 - 4c - 1)p^\sharp_1 + (6c^2 + 6c + 4)p^\sharp_2 + (-6c-3)p^\sharp_{(1,1)} + (-4c-2)p^\sharp_3 + 4p^\sharp_{(2,1)} + p^\sharp_4.$

Proof. By explicit computation, one can check that

$$p^*_{2,c} = 2(-c - \tfrac{1}{2})p^*_1 + p^*_2, \qquad p^*_{4,c} = 4(-c - \tfrac{1}{2})^3 p^*_1 + 6(-c - \tfrac{1}{2})^2 p^*_2 + 4(-c - \tfrac{1}{2})p^*_3 + p^*_4.$$

(Indeed, it's not hard to show that in general, $p^*_{k,c} = \sum_{j=1}^{k} \binom{k}{j} (-c - \tfrac{1}{2})^{k-j} p^*_j$.) The claim now follows from (3.12). □

For any c, these formulas allow us to compute the expected value of $p^*_{2,c}$ and $p^*_{4,c}$ over a random $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$, by using Corollary 3.8.4. Furthermore, for any k and d, $\sum_{i=1}^{d} (-i-c)^k$ is a constant which doesn't depend on $\boldsymbol{\lambda}$. Combining these two facts allows us to compute average value over a random $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$ of $\sum_{i=1}^{d} (\boldsymbol{\lambda}_i - i - c)^k$, for $k = 2, 4$. In particular, we are interested in computing this expectation when $c = \frac{n}{d}$. Write $\boldsymbol{L}_i := \boldsymbol{\lambda}_i - i - \frac{n}{d}$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_d} \left[\sum_{i=1}^{d} \boldsymbol{L}_i^2 \right] = -\tfrac{n}{d} + nd + \tfrac{d^3}{3} + \tfrac{d^2}{2} + \tfrac{d}{6} \geq -\tfrac{n}{d} + nd \geq \tfrac{3nd}{4}, \tag{4.7}$$

where in the last step we used the fact that $n/d \leq nd/4$ because $d \geq 2$.

Similarly, as $n \geq 10^{10} d^2 \geq d^2$, we can use the bound

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_d} \left[\sum_{i=1}^{d} \boldsymbol{L}_i^4 \right] = 2n - \tfrac{d}{30} - \tfrac{4n}{d^2} - \tfrac{6n}{d^3} + 2nd^2 + \tfrac{d^5}{5} + \tfrac{d^3}{3} + \tfrac{3n^2}{d^3} + \tfrac{d^4}{2} + nd^3 + 2n^2 d + nd - \tfrac{5n^2}{d} + \tfrac{4n}{d}$$
$$\leq 2n + 2nd^2 + \tfrac{d^5}{5} + \tfrac{d^3}{3} + \tfrac{3n^2}{d^3} + \tfrac{d^4}{2} + nd^3 + 2n^2 d + nd + \tfrac{4n}{d} \leq 6n^2 d,$$

where the last step only uses trivial bounds involving the facts $n \geq d^2$ and $d \geq 2$.

For a fixed λ, let $\mathcal{L}(\lambda) := \{i \in [d] \mid |L_i| \geq 5\sqrt{n}\}$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}\left[\sum_{i\in\mathcal{L}(\boldsymbol{\lambda})} \boldsymbol{L}_i^2\right] \leq \frac{1}{25n}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}\left[\sum_{i\in\mathcal{L}(\boldsymbol{\lambda})} \boldsymbol{L}_i^4\right] \leq \frac{1}{25n}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}\left[\sum_{i=1}^{d} \boldsymbol{L}_i^4\right] \leq \frac{nd}{4}.$$

Thus, by (4.7),

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}\left[\sum_{i\in[d]\setminus\mathcal{L}(\boldsymbol{\lambda})} \boldsymbol{L}_i^2\right] \geq \frac{nd}{2}.$$

Now define

$$\mathcal{M}(\lambda) := \left\{i \in [d] \;\middle|\; \frac{\sqrt{n}}{200} \leq |L_i| < 5\sqrt{n}\right\},$$

and let $\mathcal{E}$ be the event that $|\mathcal{M}(\lambda)| \geq d/200$. We claim that $p = \mathbf{Pr}[\mathcal{E}] \geq 1/100$. This is because if $p < 1/100$, then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}\left[\sum_{i\in[d]\setminus\mathcal{L}(\boldsymbol{\lambda})} \boldsymbol{L}_i^2\right] \leq p\cdot 25nd + (1-p)\cdot\left(\frac{25nd}{200} + \left(1-\frac{1}{200}\right)\cdot\frac{nd}{200^2}\right) < \frac{nd}{2},$$

which is a contradiction.

Now let us use the assumption that $n \geq 10^{10}d^2$. Consider any coordinate $i \in [d]$ satisfying

$$|\boldsymbol{L}_i| = \left|\boldsymbol{\lambda}_i - i - \frac{n}{d}\right| \geq \frac{\sqrt{n}}{200}.$$

By our assumption that $n \geq 10^{10}d^2$, this implies that

$$\left|\boldsymbol{\lambda}_i - \frac{n}{d}\right| \geq \frac{\sqrt{n}}{1000}.$$

As a result, when $\mathcal{E}$ holds, which happens with at least 1% probability, there are at least $\frac{d}{200}$ coordinates $i \in [d]$ such that

$$\left|\boldsymbol{\lambda}_i - \frac{n}{d}\right| \geq \frac{\sqrt{n}}{1000}.$$

This completes the proof. □

4.3.1 The EYD lower bound (continued)

In this section, we prove Theorem 4.3.2.

Theorem 4.3.2 restated. *For every constant $C > 0$, there are constants $\delta, \epsilon > 0$ such that*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_d^n}[d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) > \epsilon] \geq \delta$$

when $n \leq Cd^2$ and d is sufficiently large.

Proof. To prove Theorem 4.3.2, we show, at a high level, that when $n \leq Cd^2$, Biane's law of large numbers kicks in and $\overline{\boldsymbol{\lambda}}$ approaches the limiting curve Ω_θ, for $\theta := \frac{\sqrt{n}}{d}$. Each of these curves is constantly far from the curve produced by the uniform partition, and the lower bound follows. However, carrying out this proof involves some subtle argumentation and splitting of hairs which we will go into.

There is one regime where $\overline{\boldsymbol{\lambda}}$ certainly does not approach Ω_θ: when n is a fixed value independent of the value of d, then $\overline{\boldsymbol{\lambda}}$ will be always be constantly far from Ω_θ. However, we can rule this case out by noting that when n is too small as a function of d, then any $\lambda = (\lambda_1, \ldots, \lambda_d)$ with n boxes will have most of its λ_i's zero, and so $\underline{\lambda}$ will be far from uniform. In particular, when $n = o(d)$, then we have that $d_{\text{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) \to 1$ as $d \to \infty$. As a result, for sufficiently large d we can immediately assume that $n \geq f(d)$, where $f(d)$ is any function which is both $\omega_d(1)$ and $o(d)$. For concreteness, we will take $f(d) := \sqrt{d}$.

We are now in the regime where Biane's law of large numbers holds. Theorem 3.7.10 tells us that if $\frac{\sqrt{n}}{d} \sim c$ for c some absolute constant, then there is some constant $d(c) > 0$ such that for a random $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$, $\overline{\boldsymbol{\lambda}}$ is ϵ-close (in L^∞ distance) to Ω_c whenever $d \geq d(c)$. The main difficulty we have in applying Biane's law of large numbers directly is that the function $d(c)$ is left unspecified and, for example, could be wildly different even for two close values of c. This is problematic in our case, because for each value of d, the ratio $\theta = \frac{\sqrt{n}}{d}$ may be any real number in the interval $[\sqrt{f(d)}/d, \sqrt{C}]$, and so θ may jump around and never converge to a fixed value c. In particular, an adversary could potentially choose n (and therefore θ) as a function of d cleverly so that for each d, we have that $d < d(\theta)$, and so Biane's law of large numbers never applies. Though seemingly unlikely, this possibility is not ruled out by the statements of known theorems.

Our goal now is to show that the convergence to the limiting shapes guaranteed by Biane's theorem happens at roughly the same rate for all values of θ in our interval. First we will need a definition.

Definition 4.3.6. Given continual diagrams $f, g : \mathbb{R} \to \mathbb{R}$, the L^1 *distance* between them is

$$d_1(f, g) := \int_{\mathbb{R}} |f(x) - g(x)| \, \mathrm{d}x.$$

This defines a metric on the set of continual diagrams, and it is well-defined because $f(x) - g(x) = 0$ whenever $|x|$ is sufficiently large. If λ, μ are both partitions of n, then $d_1(\overline{\lambda}, \overline{\mu}) = 4 \cdot d_{\text{TV}}(\underline{\lambda}, \underline{\mu})$.

We will prove the following result:

Theorem 4.3.7. *Let $C > 0$ be an absolute constant, and let $f(d) : \mathbb{N} \to \mathbb{N}$ be $\omega_d(1)$. Then for any constant $0 < \delta < 1$, if $f(d) \leq n \leq Cd^2$, then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_d} \left[d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta) \geq \delta \right] \leq \delta,$$

for sufficiently large d, where $\theta = \frac{\sqrt{n}}{d}$.

Let us now complete the argument assuming Theorem 4.3.7. For $\kappa > 0$, define the following continual diagram:

$$\overline{\mathsf{unif}_\kappa}(x) := \begin{cases} x + \frac{2}{\kappa} & \text{if } x \in (-\frac{1}{\kappa}, \kappa - \frac{1}{\kappa}] \\ -x + 2\kappa & \text{if } x \in (\kappa - \frac{1}{\kappa}, \kappa), \\ |x| & \text{otherwise.} \end{cases} \tag{4.8}$$

To see how such a function arises, consider the uniform "partition" $\left(\frac{n}{d}, \ldots, \frac{n}{d}\right)$ ("partition" being in quotation marks because $\frac{n}{d}$ may not be integral). Drawing this in the French notation gives a rectangle of width $\frac{n}{d}$ and height d whose bottom-left corner is the origin. Drawing this in the Russian notation and dilating by a factor of $1/\sqrt{n}$ therefore gives the curve $\overline{\mathsf{unif}_\theta}(x)$. One consequence of this is that if λ is a partition of n, then $d_1(\overline{\lambda}, \overline{\mathsf{unif}_\theta}) = 4 \cdot d_{\mathrm{TV}}(\underline{\lambda}, \mathsf{Unif}_d,)$.

Define the function $\Delta : (0, \sqrt{C}] \to \mathbb{R}^{\geq 0}$ by $\Delta(\kappa) := d_1(\overline{\mathsf{unif}_\kappa}, \Omega_\kappa)$. When $\kappa < .3$, $\Delta(\kappa) > .5$ for all c. This is because $\Omega_\kappa(x) = -x$ for all $x \leq -2$ regardless of κ, whereas $\overline{\mathsf{unif}_\kappa}(x) = -x + 2\kappa$ in $(\kappa - \frac{1}{\kappa}, -2]$. Because $\kappa < .3$,

$$d_1(\overline{\mathsf{unif}_\kappa}, \Omega_\kappa) = \int_{\mathbb{R}} \left|\overline{\mathsf{unif}_\kappa}(x) - \Omega_\kappa(x)\right| \mathrm{d}x \geq 2\kappa \cdot \left(\tfrac{1}{\kappa} - 2 - \kappa\right) \geq 0.5.$$

Now, let us lower-bound $\Delta(\kappa)$ when $\kappa \geq .3$. Write I for the interval $[.3, \sqrt{C}]$. (If $.3 > \sqrt{C}$ then this step can be skipped.) To begin, we note that $\Delta(\kappa)$ is continuous on I. By comparing (4.8) with Theorem 3.7.10, it is easy to see that $\Delta(\kappa) > 0$ for all $\kappa > 0$. We can now apply the extreme value theorem, which implies that Δ achieves its minimum on I at some fixed point $\kappa^* \in I$. We therefore have that $\Delta(\kappa) \geq \Delta(\kappa^*) > 0$ for all $\kappa \in I$.

Combining the last two paragraphs, we now know that there is some value

$$\delta := \min\{0.5, \Delta(\kappa^*)\} > 0$$

such that $\Delta(\kappa) > \delta$ for all $\kappa \in (0, \sqrt{C}]$. Crucially, δ is an absolute constant which depends only on the constant C and is independent of n and d. Now, let us apply Theorem 4.3.7 with the values $f(d) = \sqrt{d}$, C, and $\frac{\delta}{2}$. Then with probability at least $1 - \frac{\delta}{2}$, $d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta) < \frac{\delta}{2}$. When this occurs,

$$d_{\mathrm{TV}}(\underline{\boldsymbol{\lambda}}, \mathsf{Unif}_d) = \frac{1}{4} d_1(\overline{\boldsymbol{\lambda}}, \overline{\mathsf{unif}_\theta}) \geq \frac{1}{4}\left(d_1(\Omega_\theta, \overline{\mathsf{unif}_\theta}) - d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta)\right) \geq \frac{\delta}{8},$$

where the second step follows from the triangle inequality, and the third step uses the fact that $d_1(\Omega_\theta, \overline{\mathsf{unif}_\theta}) = \Delta(\theta) \geq \delta$. This proves the theorem with the parameters $1 - \frac{\delta}{2}$ and $\frac{\delta}{8}$. □

It remains to prove Theorem 4.3.7, and this is done in the next subsection.

Proof of Theorem 4.3.7

Our goal is to give a rate of convergence of $\overline{\boldsymbol{\lambda}}$ to Ω_θ which depends only on d and is independent of n. To do this, we will show that standard law of large numbers arguments give convergence rates of this form. Biane's [Bia01] proof of the law of large numbers for the Schur-Weyl distribution does not use Kerov's algebra of observables. Instead, we will follow the proof of the law of large numbers (second form) for the Plancherel distribution in [IO02, Theorem 5.5] and use results from [Mél10a] to extend this proof to the Schur-Weyl distribution. We emphasize that our proof contains no ideas not already found in [IO02, Mél10a], and that our goal is just to show that proper bookkeeping of their arguments yields our Theorem 4.3.7. (Finally, we note that Meliot [Mél10a] also sketches a proof the law of large numbers for the Schur-Weyl distribution using Kerov's algebra of observables at the beginning of his Section 3.)

Write $\Delta_{\boldsymbol{\lambda}}(x) := \overline{\boldsymbol{\lambda}}(x) - \Omega_\theta(x)$. Because $\overline{\boldsymbol{\lambda}}$ and Ω_θ are both continual diagrams, we know that $\Delta_{\boldsymbol{\lambda}}$ is supported (i.e., nonzero) on a finite interval. We will need a stronger property, which is that the width of this interval does not grow with d (or, equivalently, with n). To show this, note that $\Delta_{\boldsymbol{\lambda}}(x)$ is zero when both $\Omega_\theta(x) = |x|$ and $\overline{\boldsymbol{\lambda}}(x) = |x|$. For the first of these, we can consult Theorem 3.7.10 and see that $\Omega_\theta(x) = |x|$ outside the interval $[-2, \theta+2]$. On the other hand, $\overline{\boldsymbol{\lambda}}(x)$ does not equal $|x|$ outside a constant-width interval for all $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$. (For example, with nonzero probability $\boldsymbol{\lambda} = (n)$, in which case $\overline{\boldsymbol{\lambda}}(x) = |x|$ only outside the interval $(-1/\sqrt{n}, \sqrt{n})$.) However, the next proposition shows that our desired property occurs with high probability.

Proposition 4.3.8. *With probability* $1 - \frac{\delta}{2}$, $\overline{\boldsymbol{\lambda}}(x) \neq |x|$ *only on an interval of width* $w = O_\delta(1)$.

Proof. We will show that $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}'_1 \leq \beta\sqrt{n}$, each with probability $1-\delta/4$, for some constant β which depends only on δ (and C). The proposition will then follow from the union bound, as $\overline{\boldsymbol{\lambda}} = |x|$ outside the interval $[-\boldsymbol{\lambda}_1/\sqrt{n}, \boldsymbol{\lambda}_1/\sqrt{n}]$. By Proposition 3.7.12,

$$\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq \beta\sqrt{n}], \mathbf{Pr}[\boldsymbol{\lambda}'_1 \geq \beta\sqrt{n}] \leq \left(\frac{(1+\beta\theta)e^2}{\beta^2}\right)^{\beta\sqrt{n}} \leq \frac{(1+\beta\theta)e^2}{\beta^2} \leq \frac{(1+\beta\sqrt{C})e^2}{\beta^2}.$$

This can be made less than $\delta/4$ by choosing β to be a sufficiently large function of C and δ. □

Let I' be the constant-width interval guaranteed by Proposition 4.3.8. Clearly, I' contains the point zero. Thus, if we define

$$I := [-2, \theta+2] \cup I'$$

then this is a single interval of width $w = O_\delta(1)$. This motivates the following definition:

Definition 4.3.9. We say that λ is *usual* if Δ_λ is supported on I. By the previous discussion, a random $\boldsymbol{\lambda}$ is usual with probability $1 - \delta/2$.

Let us condition $\boldsymbol{\lambda}$ on it being usual, and let us suppose that $d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta) \geq \delta$. Then there is some point $x \in I$ such that $|\Delta_{\boldsymbol{\lambda}}(x)| \geq \frac{\delta}{w}$. Now we will use the fact that Ω_θ and $\overline{\boldsymbol{\lambda}}$ are continual diagrams, which implies that they are both 1-Lipschitz, and therefore $\Delta_{\boldsymbol{\lambda}}$ is 2-Lipschitz. Then if we consider the subinterval $I_x \subseteq I$ defined as $I_x := [x - \frac{\delta}{4w}, x + \frac{\delta}{4w}]$, this Lipschitz property implies that $|\Delta_{\boldsymbol{\lambda}}(y)| \geq \frac{\delta}{2w}$ for all $y \in I_x$. (That I_x is contained in I follows from the fact that $\Delta_{\boldsymbol{\lambda}}$ is nonzero on I_x and $\boldsymbol{\lambda}$ is usual.) We note that the width of I_x is $\frac{\delta}{2w}$.

Let $\mathcal{J}$ be a set of $\lceil \frac{4w^2}{\delta} \rceil$ closed intervals of width $\frac{\delta}{4w}$ which cover I. These intervals are chosen to have half the width of I_x, the result being that there is some interval $J^* \in \mathcal{J}$ which is completely contained in I_x. For each interval $J \in \mathcal{J}$, let $\Psi_J : \mathbb{R} \to \mathbb{R}^{\geq 0}$ be a continuous function supported on J which satisfies $\int \Psi_J(y)dy = 1$ (such functions are known to exist; e.g., bump functions). Then

$$\left|\int_{-\infty}^{\infty} \Delta_{\boldsymbol{\lambda}}(y)\Psi_{J^*}(y)dy\right| \geq \min_{y \in I_x} |\Delta_{\boldsymbol{\lambda}}(y)| \cdot \int_{-\infty}^{\infty} \Psi_{J^*}(y)dy \geq \frac{\delta}{2w}.$$

By the Weierstrass approximation theorem, we can approximate each Ψ_J with a polynomial function $\widetilde{\Psi}_J$ such that for each $x \in I$, $|\Psi_J(x) - \widetilde{\Psi}_J(x)| \leq \frac{\delta}{8w^3}$. (Outside of I, $\widetilde{\Psi}_J$ can—and will—be an arbitrarily bad approximator for Ψ_J.) Because $\Delta_{\boldsymbol{\lambda}}$ is 2-Lipschitz and $\boldsymbol{\lambda}$ is usual, $|\Delta_{\boldsymbol{\lambda}}(x)| \leq 2w$ for all $x \in I$ and is zero everywhere else. As a result, for the interval J^*,

$$\left|\int_{-\infty}^{\infty} \Delta_{\boldsymbol{\lambda}}(y)\widetilde{\Psi}_{J^*}(y)dy\right| \geq \left|\int_{-\infty}^{\infty} \Delta_{\boldsymbol{\lambda}}(y)\Psi_{J^*}(y)dy\right| - \left|\int_{-\infty}^{\infty} \Delta_{\boldsymbol{\lambda}}(y)\left(\Psi_{J^*}(y) - \widetilde{\Psi}_{J^*}(y)\right)dy\right| \geq \frac{\delta}{4w}.$$

The first inequality uses the triangle inequality, and the second inequality uses crucially the fact that $\Delta_{\boldsymbol{\lambda}}$ is zero outside I.

In summary, we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}\left[d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta) \geq \delta\right] \leq \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda} \sim \mathrm{SW}_d^n}\left[\exists J \in \mathcal{J} : \left|\int_{-\infty}^{\infty} \Delta_{\boldsymbol{\lambda}}(y)\widetilde{\Psi}_J(y)dy\right| \geq \frac{\delta}{4w}\right] + \frac{\delta}{2}, \tag{4.9}$$

where the $\delta/2$ comes from the event that $\boldsymbol{\lambda}$ is not usual. We will therefore show that $\left|\int \Delta_{\boldsymbol{\lambda}}(y)\widetilde{\Psi}_J(y)dy\right|$ is at most $\frac{\delta}{4w}$ for all $J \in \mathcal{J}$ with probability at least $1 - \frac{\delta}{2}$. By the union bound, it suffices to show that for each $J \in \mathcal{J}$, $\left|\int \Delta_{\boldsymbol{\lambda}}(y)\widetilde{\Psi}_J(y)dy\right| \leq \frac{\delta}{4w}$ with probability at least $1 - \frac{\delta}{2\cdot|\mathcal{J}|}$.

Let m be the maximum degree of the $\widetilde{\Psi}_J$ functions, for all $J \in \mathcal{J}$. Fix an interval $J \in \mathcal{J}$. Then we can write

$$\widetilde{\Psi}_J(x) = \sum_{k=0}^{m} a_J^{(k)} x^k \quad \text{and} \quad \int_{-\infty}^{\infty} \Delta_\lambda(y)\widetilde{\Psi}_J(y)dy = \sum_{k=0}^{m} a_J^{(k)} \int_{-\infty}^{\infty} x^k \Delta_\lambda(x)dx, \tag{4.10}$$

where the $a_J^{(k)}$'s are constants. The following proposition, found in [Mél10a, Lemma 7], gives a nice expression for the integrals on the right-hand side.

Proposition 4.3.10. *Let $k \geq 1$. Then*

$$\int_{-\infty}^{\infty} x^k \Delta_\lambda(x)dx = \frac{2 \cdot \widetilde{q}_{k+1}(\lambda)}{(k+1)\sqrt{n}},$$

where $\widetilde{q}_k(\lambda)$ is the quantity defined as

$$\widetilde{q}_k(\lambda) := \frac{\widetilde{p}_{k+1}(\lambda)}{(k+1)n^{k/2}} - \sum_{\ell=1}^{\lfloor \frac{k+1}{2} \rfloor} \frac{k^{\downarrow 2\ell-1}}{(k+1-\ell)\ell!(\ell-1)!} \cdot \frac{n^{k/2+1-\ell}}{d^{k+1-2\ell}}.$$

The key fact we will use is that we can upper bound the right-hand side of Equation (4.10) by a quantity which decays with d, independent of the value of n. This is the subject of the following lemma.

Lemma 4.3.11. *The random variable* $\left|\frac{\widetilde{q}_k(\boldsymbol{\lambda})}{\sqrt{n}}\right|$, *for* $\boldsymbol{\lambda} \sim \mathrm{SW}_d^n$, *has mean* $o_d(1)$, *for all* $f(d) \leq n \leq Cd^2$.

Applying Proposition 4.3.10 and Lemma 4.3.11 to Equation (4.10), we see that

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left|\int \Delta_{\boldsymbol{\lambda}}(x)\widetilde{\Psi}_J(x)dx\right|$$

is $o_d(1)$. We may take d large enough to make this quantity arbitrarily small. Thus, select d_J so that for all $d \geq d_J$, this expectation is at most $\frac{\delta^2}{8w\cdot|\mathcal{J}|}$. Then by Markov's inequality, $\left|\int \Delta_{\boldsymbol{\lambda}}(x)\widetilde{\Psi}_J(x)dx\right| \leq \frac{\delta}{4w}$ with probability at least $1 - \frac{\delta}{2\cdot|\mathcal{J}|}$. If we set d_0 to be the max of d_J over all $J \in \mathcal{J}$, then by Equation (4.9), $\mathbf{Pr}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[d_1(\overline{\boldsymbol{\lambda}}, \Omega_\theta) \geq \delta\right] \leq \delta$ so long as $d \geq d_0$, and we are done.

Now we turn to the proof of Lemma 4.3.11.

Proof of Lemma 4.3.11. Define

$$X_k(\lambda) \coloneqq \sum_{\mu:\mathrm{wt}(\mu)=k} \frac{k^{\downarrow\ell(\mu)}}{m(\mu)} \cdot p_\mu^\sharp(\lambda)$$

and

$$q_k^\sharp(\lambda) \coloneqq \frac{X_{k+1}(\lambda)}{(k+1)n^{k/2}} - \sum_{\ell=1}^{\lfloor\frac{k+1}{2}\rfloor} \frac{k^{\downarrow 2\ell-1}}{(k+1-\ell)\ell!(\ell-1)!} \cdot \frac{n^{k/2+1-\ell}}{d^{k+1-2\ell}}. \tag{4.11}$$

Then by Proposition 7.2.5, $\widetilde{q}_k(\lambda)$ and $q_k^\sharp(\lambda)$ differ from each other by $n^{-k/2}$ times an observable $\mathcal{O}(\lambda)$ of weight k. Thus,

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left|\frac{\widetilde{q}_k(\boldsymbol{\lambda})}{\sqrt{n}}\right| \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left|\frac{q_k^\sharp(\boldsymbol{\lambda})}{\sqrt{n}}\right| + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left|\frac{\mathcal{O}(\boldsymbol{\lambda})}{n^{(k+1)/2}}\right|.$$

By Cauchy–Schwarz, $\mathbf{E}\,|\mathcal{O}(\boldsymbol{\lambda})/n^{(k+1)/2}| \leq \sqrt{\mathbf{E}\,\mathcal{O}(\boldsymbol{\lambda})^2/n^{k+1}}$. Because $\mathcal{O}$ has weight k, $\mathcal{O}^2$ has weight $2k$. As a result, we can use the next proposition to bound the contribution from this term by $o_d(1)$.

Proposition 4.3.12. *Let* $\mathcal{O}(\lambda)$ *be an observable of weight at most* $2k$. *Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[\frac{\mathcal{O}(\boldsymbol{\lambda})}{n^{k+1}}\right] = o_d(1).$$

Proof. As in the proof of Lemma 7.2.7, this reduces to showing that $\mathbf{E}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[p_\mu^\sharp(\boldsymbol{\lambda})/n^{k+1}\right] = o_d(1)$, where μ is a partition of weight $2k$, i.e. $|\mu| + \ell(\mu) \leq 2k$. By Corollary 3.8.4,

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[\frac{p_\mu^\sharp(\boldsymbol{\lambda})}{n^{k+1}}\right] = \frac{n^{\downarrow|\mu|}}{n^{k+1}} \cdot \frac{d^{\ell(\mu)}}{d^{|\mu|}} \leq \frac{n^{|\mu|}}{n^{k+1}} \cdot \frac{d^{\ell(\mu)}}{d^{|\mu|}} = \frac{n^{|\mu|}}{n^{k+1}} \cdot \frac{d^{\mathrm{wt}(\mu)}}{d^{2|\mu|}}.$$

If $|\mu| < k+1$, then this expression is at most $1/n$, which is $o_d(1)$ because $n \geq f(d) = \omega_d(1)$. On the other hand, if $|\mu| \geq k+1$, then for all $n \leq Cd^2$ this expression is at most

$$\frac{(Cd^2)^{|\mu|}}{(Cd^2)^{k+1}} \cdot \frac{d^{\mathrm{wt}(\mu)}}{d^{2|\mu|}} \leq C^{|\mu|-(k+1)} \cdot \frac{d^{\mathrm{wt}(\mu)}}{d^{2(k+1)}},$$

which is $o_d(1)$ as $\mathrm{wt}(\mu) \leq 2k$. □

It remains to bound $\mathbf{E}\,|q_k^\sharp(\boldsymbol{\lambda})/\sqrt{n}|$ by $o_d(1)$. First, we will show that $q_k^\sharp(\boldsymbol{\lambda})$ can be viewed as (approximately) computing the deviation of a certain random variable from its mean. To do this, let us compute the mean of the first term on the right-hand side of Equation (4.11).

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n} \frac{X_{k+1}(\boldsymbol{\lambda})}{(k+1)n^{k/2}} &= \frac{1}{(k+1)n^{k/2}} \cdot \sum_{\mu:\mathrm{wt}(\mu)=k+1} \frac{(k+1)^{\downarrow \ell(\mu)}}{m(\mu)} \cdot \frac{n^{\downarrow|\mu|}}{d^{|\mu|-\ell(\mu)}} \\
&= \frac{1}{(k+1)n^{k/2}} \cdot \sum_{\ell=1}^{\lfloor \frac{k+1}{2} \rfloor} \frac{(k+1)^{\downarrow \ell} n^{\downarrow k+1-\ell}}{d^{k+1-2\ell}} \sum_{\mu:\mathrm{wt}(\mu)=k+1} \frac{1}{m(\mu)} \\
&= \frac{1}{(k+1)n^{k/2}} \cdot \sum_{\ell=1}^{\lfloor \frac{k+1}{2} \rfloor} \frac{(k+1)^{\downarrow \ell} n^{\downarrow k+1-\ell}}{d^{k+1-2\ell}} \cdot \frac{1}{\ell!}\binom{k-\ell}{\ell-1} \\
&= \sum_{\ell=1}^{\lfloor \frac{k+1}{2} \rfloor} \frac{k^{\downarrow 2\ell-1}}{(k+1-\ell)\ell!(\ell-1)!} \cdot \frac{n^{\downarrow k+1-\ell}}{n^{k/2}\cdot d^{k+1-2\ell}},
\end{aligned}$$

where the third equality follows from [Mél10a, Lemma 11]. As a result, the difference

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n} \frac{X_{k+1}(\boldsymbol{\lambda})}{(k+1)n^{k/2}} - \sum_{\ell=1}^{\lfloor \frac{k+1}{2} \rfloor} \frac{k^{\downarrow 2\ell-1}}{(k+1-\ell)\ell!(\ell-1)!} \cdot \frac{n^{k/2+1-\ell}}{d^{k+1-2\ell}}$$

can be written as a sum over terms of the form $a\cdot n^b/d^{k+1-2\ell}$, where a is a constant coefficient, $1 \leq b \leq k/2 - \ell$, and $1 \leq \ell \leq \lfloor \frac{k+1}{2} \rfloor$. Given that $n \leq Cd^2$, each of these terms if $\pm o_d(1)$. Thus, if we set

$$q_k(\lambda) := \frac{X_{k+1}(\lambda)}{(k+1)n^{k/2}} - \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n} \frac{X_{k+1}(\boldsymbol{\lambda})}{(k+1)n^{k/2}},$$

then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n} \left| \frac{q_k^\sharp(\boldsymbol{\lambda})}{\sqrt{n}} \right| \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n} \left| \frac{q_k(\boldsymbol{\lambda})}{\sqrt{n}} \right| + o_1(d).$$

Finally, we show that $\mathbf{E}\,|q_k(\boldsymbol{\lambda})/\sqrt{n}| = o_d(1)$. By Cauchy–Schwarz,

$$\mathbf{E} \left| \frac{q_k(\boldsymbol{\lambda})}{\sqrt{n}} \right| \leq \sqrt{\mathbf{E} \left(\frac{q_k(\boldsymbol{\lambda})}{\sqrt{n}} \right)^2},$$

so it suffices to show that $\mathbf{E}\,(q_k(\boldsymbol{\lambda})/\sqrt{n})^2 = o_d(1)$. This expectation is simply the variance of the random variable $X_{k+1}(\boldsymbol{\lambda})/(k+1)n^{(k+1)/2}$, which itself is a weighted sum of a constant number of random variables of the form $p^\sharp_\mu(\boldsymbol{\lambda})/n^{(k+1)/2}$, where $\mathrm{wt}(\mu) = k+1$. An easy application of Cauchy–Schwarz shows that the variance of a weighted sum of a constant number of random variables is $o_d(1)$ if the variance of each random variables is $o_d(1)$. Thus, we will show that $\mathbf{Var}[p^\sharp_\mu(\boldsymbol{\lambda})/n^{(k+1)/2}] = o_d(1)$ for all $\mathrm{wt}(\mu) = k+1$.

Fix a partition μ of weight $k+1$. Then

$$\mathbf{Var}\left[\frac{p^\sharp_\mu(\boldsymbol{\lambda})}{n^{(k+1)/2}}\right] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[\frac{1}{n^{(k+1)/2}}\left(p^\sharp_\mu(\boldsymbol{\lambda})p^\sharp_\mu(\boldsymbol{\lambda}) - \mathbf{E}[p^\sharp_\mu]^2\right)\right]$$

By Proposition 3.8.10, $p^\sharp_\mu(\lambda)\cdot p^\sharp_\mu(\lambda) = p^\sharp_{\mu\cup\mu}(\lambda) + \mathcal{O}(\lambda)$, where $\mathcal{O}(\lambda)$ is an observable of weight at most $2\cdot\mathrm{wt}(p^\sharp_\mu) - 2 = 2k$. Then

$$\mathbf{Var}\left[\frac{p^\sharp_\mu(\boldsymbol{\lambda})}{n^{(k+1)/2}}\right] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[\frac{1}{n^{k+1}}\cdot\left(p^\sharp_{\mu\cup\mu}(\boldsymbol{\lambda}) - \mathbf{E}[p^\sharp_\mu]^2\right)\right] + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[\frac{1}{n^{k+1}}\cdot\mathcal{O}(\boldsymbol{\lambda})\right].$$

The second term is $\pm o_d(1)$ by Proposition 4.3.12. As for the first term, Corollary 3.8.4, shows that it equals

$$\frac{1}{n^{k+1}}\cdot\left(n^{\downarrow 2|\mu|}d^{2\ell(\mu)-2|\mu|} - n^{\downarrow|\mu|}n^{\downarrow|\mu|}d^{2\ell(\mu)-2|\mu|}\right) = \frac{1}{d^{4|\mu|-2(k+1)}}\cdot\left(\frac{n^{\downarrow 2|\mu|} - (n^{\downarrow|\mu|})^2}{n^{k+1}}\right), \tag{4.12}$$

where we used the fact that $\ell(\mu) = \mathrm{wt}(\mu) - |\mu| = k+1-|\mu|$. The highest-degree term of both $n^{\downarrow 2|\mu|}$ and $(n^{\downarrow|\mu|})^2$ is $n^{2|\mu|}$, so we can write

$$(4.12) = \frac{1}{d^{4|\mu|-2(k+1)}}\cdot\sum_{b=-(k+1)}^{2|\mu|-(k+2)}\alpha_b\cdot n^b$$

for some constants α_b. When $b<0$, $n^b/d^{4|\mu|-2k-2}\leq 1/n$, which is $o_d(1)$ because $n\geq f(d) = \omega_d(1)$. On the other hand, when $b\geq 0$, then this term is $o_d(1)$ because $n\leq Cd^2$. □

Chapter 5

Quantum tomography

In this chapter, we prove our quantum tomography and principal component analysis (PCA) results. Quantum tomography is a fundamental problem with a vast field of research devoted to it, and the research done varies wildly in its focus on issues such as current-day practicality and error metrics. Many of these issues are beyond the scope of this thesis, but more details can be found in the thesis of [Hua12]. Our focus here is on fully entangled quantum measurements—those which may one day be implemented on a scalable quantum computer—for which we can prove optimal trace distance error rates.

The "textbook" tomography algorithm [NC10, page 389] is the *Pauli basis measurement* tomography algorithm. This algorithm uses a certain set of d^2 $d \times d$ matrices called the *Pauli matrices* which form a linear basis for the set of all $d \times d$ matrices. The goal of this tomography algorithm is to learn the coefficients of the unknown $\rho \in \mathbb{C}^{d \times d}$ in this basis, thereby recovering ρ. Estimating each coefficient to sufficient accuracy requires $O(d^2/\epsilon^2)$ copies, and as there are d^2 matrices in total, learning ρ takes $n = O(d^4/\epsilon^2)$ copies in total (see [FGLE12, Footnote 2] for details). Though this has an extra factor of d^2 compared to our bounds, the algorithm has more current-day practicality: it uses nonadaptive measurements, and the measurement projectors are of a particularly nice form.

This was the best known upper bound for tomography until the 2014 work of [KRT14] gave an algorithm using $n = O(d^3/\epsilon^2)$ copies. Though it also uses nonadaptive measurements, its measurements are more-difficult-to-implement POVMs. We will discuss this algorithm and give a related, though simpler, algorithm achieving the same bound in Section 5.1.

In this chapter, we focus on two related algorithms which use fully entangled measurements. Both begin with the weak Schur sampling measurement and then follow up with a measurement in the irrep space that the state collapses to. The first is an algorithm due to Keyl [Key06] which uses the *highest weight vector* (Definition 2.4.12). The second is an algorithm due to Haah et al.[HHJ+16] which is inspired by the *pretty good measurement (PGM)* from quantum hypothesis testing [HW94]. For both, we are able to prove the exact same copy complexity bounds.

Theorem 5.0.1 (Theorem 1.4.11 restated)**.** *Given n copies of a mixed state $\rho \in \mathbb{C}^{d \times d}$, let $\hat{\boldsymbol{\rho}}$ be the random output of either Keyl's algorithm or the PGM tomography algorithm. Then*

$$\mathbf{E}\, \|\hat{\boldsymbol{\rho}} - \rho\|_F^2 \leq \frac{4d - 3}{n}.$$

Corollary 1.4.12 follows as an immediate consequence, showing that $n = O(d^2/\epsilon^2)$ copies are sufficient for tomography.

In addition, we show that a natural truncated version of Keyl's algorithm generalizes to the case when ρ is approximately low rank.

Theorem 5.0.2 (Theorem 1.4.13 restated)**.** *Given n copies of a mixed state $\rho \in \mathbb{C}^{d\times d}$, let $\hat{\boldsymbol{\rho}}$ be the rank k random output of the truncated version of Keyl's algorithm. Then*

$$\mathbf{E}\, \|\hat{\boldsymbol{\rho}} - \rho\|_1 \leq \alpha_{k+1} + \cdots + \alpha_d + 6\sqrt{\frac{kd}{n}}.$$

Thus, quantum PCA can be solved with $n = O(kd/\epsilon^2)$ copies.

Finally, we prove a lower bound showing that the dependence of Theorem 5.0.2 on k and d is optimal.

Theorem 5.0.3 (Theorem 1.4.14 restated)**.** *There exists an absolute constant $\epsilon_0 > 0$ such that the following holds: suppose with n copies of an unknown rank-r $\rho \in \mathbb{C}^{d\times d}$ it is possible (with constant probability) to produce an estimate $\hat{\rho} \in \mathbb{C}^{d\times d}$ such that $d_{\mathrm{tr}}(\rho, \hat{\rho}) \leq \epsilon_0$. Then $n \geq \Omega(rd)$.*

The chapter is organized as follows:

- Section 5.1 covers nonadaptive tomography and gives a new $n = O(d^3/\epsilon^2)$ algorithm.
- Section 5.2 covers the PGM-based tomography algorithm and proves the PGM half of Theorem 5.0.1.
- Section 5.3 covers Keyl's tomography algorithm and proves the corresponding half of Theorem 5.0.1.
- Section 5.4 proves Theorem 5.0.2, the PCA result for Keyl's algorithm.
- Section 5.5 proves Theorem 5.0.3, the tomography lower bound.

5.1 Tomography with unentangled measurements

In this section, we consider the following nonadaptive tomography algorithm. This algorithm and its analysis are joint work with Akshay Krishnamurthy.

Definition 5.1.1 (Random basis tomography algorithm)**.** Given n copies of ρ,

1. Measure each copy of ρ with the POVM $\{d\,|v\rangle\,\langle v|\,dv\}$.
2. For each $i \in [n]$, if $|\boldsymbol{v}_i\rangle$ is the i-th measurement outcome, set $\boldsymbol{\rho}_i := (d+1)\,|\boldsymbol{v}_i\rangle\,\langle \boldsymbol{v}_i| - I$.
3. Output $\hat{\boldsymbol{\rho}} := \mathrm{avg}_{i\in[n]}\{\boldsymbol{\rho}_i\}$.

Here, dv denotes the uniform measure on unit vectors $|v\rangle \in \mathbb{C}^d$.

Proposition 5.1.2. *Both $\hat{\boldsymbol{\rho}}$ and the $\boldsymbol{\rho}_i$'s are unbiased estimators for ρ, meaning $\mathbf{E}\,\hat{\boldsymbol{\rho}} = \mathbf{E}\,\boldsymbol{\rho}_i = \rho$.*

Proof. As $\hat{\boldsymbol{\rho}}$ is the average of the $\boldsymbol{\rho}_i$'s, it suffices to prove the proposition for a fixed $\boldsymbol{\rho}_i$. We begin by analyzing the i-th POVM outcome.

$$\mathbf{E}\,|\boldsymbol{v}_i\rangle\langle\boldsymbol{v}_i| = d\int_v |v\rangle\langle v|\cdot\operatorname{tr}(|v\rangle\langle v|\,\rho)dv = d\int_v \operatorname{tr}_B\left((|v\rangle\langle v|\otimes|v\rangle\langle v|)(I\otimes\rho)\right)dv$$

$$= d\cdot\operatorname{tr}_B\left(\left(\int_v |v\rangle\langle v|\otimes|v\rangle\langle v|\,dv\right)(I\otimes\rho)\right) = \frac{1}{d+1}\cdot\operatorname{tr}_B((I\otimes I+\mathrm{SWAP})(I\otimes\rho)) = \frac{\rho+I}{d+1},$$

where the fourth equality uses [Har13, Proposition 6]. The proposition follows after rearranging. □

Theorem 5.1.3. $\mathbf{E}\,\|\hat{\boldsymbol{\rho}}-\rho\|_F^2 \le \dfrac{d^2+d-1}{n}$ *and* $\mathbf{E}\,d_{\mathrm{tr}}(\hat{\boldsymbol{\rho}},\rho) \le \sqrt{\dfrac{d^3+d^2-d}{4n}}$.

Proof. Using the fact that the variance of the sum of independent random variables is equal to the sum of the variances,

$$\mathbf{E}\,\|\hat{\boldsymbol{\rho}}-\rho\|_F^2 = \frac{1}{n^2}\mathbf{Var}\left[\sum \boldsymbol{\rho}_i\right] = \frac{1}{n^2}\sum_i \mathbf{Var}[\boldsymbol{\rho}_i] = \frac{1}{n^2}\sum_i \mathbf{E}\,\|\boldsymbol{\rho}_i-\rho\|_F^2. \tag{5.1}$$

For any $i\in[n]$, $\mathbf{E}\,\|\boldsymbol{\rho}_i-\rho\|_F^2 = \mathbf{E}\operatorname{tr}(\boldsymbol{\rho}_i^2)-\operatorname{tr}(\rho^2)\le\mathbf{E}\operatorname{tr}(\boldsymbol{\rho}_i^2)$. By construction, $\boldsymbol{\rho}_i$ always has the eigenvalue d with multiplicity 1 and the eigenvalue -1 with multiplicity $d-1$. Hence $\operatorname{tr}(\boldsymbol{\rho}_i^2) = d^2+d-1$. Plugging this into Equation (5.1) and summing over all $i\in[n]$ yields the Frobenius bound in the theorem. The trace distance bound then follows from Cauchy-Schwartz. □

The paper of Kueng, Rauhut, and Terstiege [KRT14] considers a similar tomography scheme. They begin with the same measurement as in the random basis tomography algorithm, but rather than averaging together the empirical outcomes, they use low rank matrix recovery techniques to infer the value of ρ. They prove similar bounds as our Theorem 5.1.3 and also give improved bounds in the case when ρ is low rank.

Theorem 5.1.4 ([KRT14])**.** *If $\rho\in\mathbb{C}^{d\times d}$ is rank k, there is a tomography algorithm using nonadaptive, unentangled measurements which outputs an estimate $\hat{\boldsymbol{\rho}}$ satisfying*

$$\|\hat{\boldsymbol{\rho}}-\rho\|_F \le O\left(\sqrt{\frac{kd}{n}}\right) \quad \text{and} \quad d_{\mathrm{tr}}(\hat{\boldsymbol{\rho}},\rho)\le O\left(\sqrt{\frac{k^2d}{n}}\right)$$

with high probability. Thus, trace distance tomography is possible with $n = O(k^2d/\epsilon^2)$ copies.

An exposition of their result more in the style of this thesis can be found in the second appendix of [HHJ$^+$16]. Finally, at QIP 2016 Jeongwan Haah announced [Haa16] that he and his coauthors [HHJ$^+$16] had proven a matching lower bound, showing that $n = \Omega(k^2d/\epsilon^2)$ copies of ρ are necessary to learn a rank-k density matrix using only nonadaptive, unentangled measurements. Compared with our Theorems 5.0.1 and 5.0.2, this shows that entangled measurements can achieve strictly better copy complexity than nonadaptive measurements. We note that their lower bound has yet to appear in print.

5.2 The pretty good measurement

In this section, we will consider the first of the two tomography algorithms introduced in [HHJ+16].

Definition 5.2.1 (PGM tomography algorithm)**.** Given n copies of ρ,

1. Perform weak Schur sampling on ρ, resulting in a random $\boldsymbol{\lambda}$. Discard the permutation irrep register. Then ρ collapses to $\pi_{\boldsymbol{\lambda}}(\rho)/s_{\boldsymbol{\lambda}}(\alpha)$.
2. Measure within the space $\mathrm{V}^d_{\boldsymbol{\lambda}}$ using the POVM with elements
$$\frac{\dim(\mathrm{V}^d_{\boldsymbol{\lambda}})}{s_{\boldsymbol{\lambda}}(\underline{\boldsymbol{\lambda}})}\pi_{\boldsymbol{\lambda}}(U\mathrm{diag}(\underline{\boldsymbol{\lambda}})U^\dagger)\mathrm{d}U$$
for each $U \in \mathrm{U}(d)$.
3. Output $\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger$.

That the POVM in step 2 is valid follows by averaging over all $U \in \mathrm{U}(d)$ and applying Schur's lemma. The weight the POVM gives a particular $U \in U(d)$ is

$$\begin{aligned}\frac{\dim(\mathrm{V}^d_\lambda)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)}\mathrm{tr}(\pi_\lambda(\rho)\pi_\lambda(U\mathrm{diag}(\underline{\lambda})U^\dagger))\mathrm{d}U &= \frac{\dim(\mathrm{V}^d_\lambda)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)}\mathrm{tr}(\pi_\lambda(\rho U\mathrm{diag}(\underline{\lambda})U^\dagger))\mathrm{d}U\\ &= \frac{\dim(\mathrm{V}^d_\lambda)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)}s_\lambda(\rho U\mathrm{diag}(\underline{\lambda})U^\dagger)\mathrm{d}U.\end{aligned}$$

Integrating this quantity over the unitary group should yield 1, which (essentially) proves the following well-known equation from reprsentation theory.

$$\int_U s_\lambda(AUBU^\dagger)\mathrm{d}U = \frac{s_\lambda(A)s_\lambda(B)}{\dim(\mathrm{V}^d_\lambda)}, \tag{5.2}$$

where here the $s_\lambda(\cdot)$'s are applied to the eigenvalues of their arguments.

Haah et al. show the following result [HHJ+16].

Theorem 5.2.2 ([HHJ+16])**.** *The PGM tomography algorithm estimates an unknown mixed state $\rho \in \mathbb{C}^{d\times d}$ to error ϵ in infidelity using $n = O(d^2/\epsilon \cdot \log(d/\epsilon))$ copies, or to error ϵ in trace distance using $n = O(d^2/\epsilon^2 \cdot \log(d/\epsilon))$ copies.*

(For the definition of infidelity, see [HHJ+16].)

We show the following result.

Theorem 5.2.3. $\displaystyle\mathop{\mathbf{E}}_{\substack{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)\\ \boldsymbol{U}\sim\mathrm{PGM}_{\boldsymbol{\lambda}}(\rho)}} \|\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2 \leq \frac{4d-3}{n}.$

This proves the PGM part of our Theorem 5.0.1, and it shows that the Haah et al. trace distance bound in Theorem 5.2.2 holds with $n = O(d^2/\epsilon^2)$.

Proof of Theorem 5.2.3. Throughout the proof we assume $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$ and $\boldsymbol{U} \sim \mathrm{PGM}_{\boldsymbol{\lambda}}(\rho)$. We have

$$n^2 \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2 = \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \left[\sum_{i=1}^d (\alpha_i n)^2 + \sum_{i=1}^d \boldsymbol{\lambda}_i^2 - 2n^2 \cdot \mathrm{tr}(\rho\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger)\right]. \tag{5.3}$$

Let's analyze the cross-term for a fixed λ:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{U}} \mathrm{tr}(\rho\boldsymbol{U}\mathrm{diag}(\underline{\lambda})\boldsymbol{U}^\dagger) &= \frac{\dim(\mathrm{V}_\lambda^d)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)} \int_U \mathrm{tr}(\rho U\mathrm{diag}(\underline{\lambda})U^\dagger)s_\lambda(\rho U\mathrm{diag}(\underline{\lambda})U^\dagger)\mathrm{d}U \\
&= \frac{\dim(\mathrm{V}_\lambda^d)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)} \int_U \sum_{i=1}^d s_{\lambda+e_i}(\rho U\mathrm{diag}(\underline{\lambda})U^\dagger)\mathrm{d}U && \text{(Pieri)} \\
&= \frac{\dim(\mathrm{V}_\lambda^d)}{s_\lambda(\underline{\lambda})s_\lambda(\alpha)} \sum_{i=1}^d \frac{s_{\lambda+e_i}(\alpha)s_{\lambda+e_i}(\underline{\lambda})}{\dim(\mathrm{V}_{\lambda+e_i}^d)}, && \text{(Equation (5.2))} \\
&= \sum_{i=1}^d \frac{\Phi_{\lambda+e_i}(\alpha)}{\Phi_\lambda(\alpha)} \cdot \frac{s_{\lambda+e_i}(\underline{\lambda})}{s_\lambda(\underline{\lambda})} \geq \sum_{i=1}^d \frac{\Phi_{\lambda+e_i}(\alpha)}{\Phi_\lambda(\alpha)} \cdot \left(\frac{\lambda_i}{n}\right).
\end{aligned}$$

Here this last step uses three facts: (i) that the $\Phi_{\lambda+e_i}(\alpha)$'s form a *decreasing* sequence (by a recent result of Suvrit Sra [Sra15]), (ii) Proposition 3.4.5 (applied to $s_{\lambda+e_i}(\underline{\lambda})/s_\lambda(\underline{\lambda})$), and (iii) equation 1.9, i.e. the elementary majorization inequality.

Plugging this into (5.3), we see that

$$\begin{aligned}
n^2 \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2 &\leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \left[\sum_{i=1}^d (\alpha_i n)^2 + \sum_{i=1}^d \boldsymbol{\lambda}_i^2 - 2n \cdot \sum_{i=1}^d \frac{\Phi_{\lambda+e_i}(\alpha)}{\Phi_\lambda(\alpha)} \cdot \boldsymbol{\lambda}_i\right] && (5.4) \\
&\leq dn + 2\sum_{i=1}^d (\alpha_i n)^2 - 2n \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \frac{\Phi_{\lambda+e_i}(\alpha)}{\Phi_\lambda(\alpha)} \cdot \boldsymbol{\lambda}_i, && (5.5)
\end{aligned}$$

using Lemma 4.1.1. Focusing on the last term,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\lambda} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{\Phi_{\boldsymbol{\lambda}+e_i}(\alpha)}{\Phi_{\boldsymbol{\lambda}}(\alpha)} = \mathop{\mathbf{E}}_{\lambda} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}}(1,\ldots,1)}{s_{\boldsymbol{\lambda}+e_i}(1,\ldots,1)} \\
&\geq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \left(2 - \frac{s_{\boldsymbol{\lambda}+e_i}(1,\ldots,1)}{s_{\boldsymbol{\lambda}}(1,\ldots,1)}\right) = 2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} - \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}+e_i}(1,\ldots,1)}{s_{\boldsymbol{\lambda}}(1,\ldots,1)},
\end{aligned} \tag{5.6}$$

where we used $r \geq 2 - \frac{1}{r}$ for $r > 0$. We lower-bound the first term in (5.6) by first using the inequality (1.9) and Proposition 3.4.5, and then using inequality (1.9) and Lemma 4.1.2 (as in the proof of Theorem 4.0.2):

$$2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \geq 2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^d \boldsymbol{\lambda}_i \alpha_i \geq 2n \sum_{i=1}^d \alpha_i^2. \tag{5.7}$$

As for the second term in (5.6), we use the definition of the Schur Weyl distribution and the first formula in Definition 2.2.13 to compute

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^{d} \boldsymbol{\lambda}_i \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}+e_i}(1,\dots,1)}{s_{\boldsymbol{\lambda}}(1,\dots,1)} &= \sum_{i=1}^{d} \sum_{\lambda \vdash n} \dim(\lambda) s_\lambda(\alpha) \cdot \lambda_i \cdot \frac{s_{\lambda+e_i}(\alpha)}{s_\lambda(\alpha)} \frac{\dim(\lambda+e_i)(d+\lambda_i-i+1)}{\dim(\lambda)(n+1)} \\
&= \sum_{i=1}^{d} \sum_{\lambda \vdash n} \dim(\lambda+e_i) s_{\lambda+e_i}(\alpha) \cdot \frac{\lambda_i(d-i+\lambda_i+1)}{n+1} \\
&\le \sum_{i=1}^{d} \mathop{\mathbf{E}}_{\boldsymbol{\lambda}' \sim \mathrm{SW}^{n+1}(\alpha)} \frac{(\boldsymbol{\lambda}'_i - 1)(d-i+\boldsymbol{\lambda}'_i)}{n+1} \\
&\le \frac{1}{n+1}\left(\mathop{\mathbf{E}}_{\boldsymbol{\lambda}' \sim \mathrm{SW}^{n+1}(\alpha)} \sum_{i=1}^{d} (\boldsymbol{\lambda}'_i)^2 + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}' \sim \mathrm{SW}^{n+1}(\alpha)} \sum_{i=1}^{d} (d-i-1)\boldsymbol{\lambda}'_i \right) \\
&\le \frac{1}{n+1}\left((n+1)n \sum_{i=1}^{d} \alpha_i^2 + \sum_{i=1}^{d} (d+i-2)((n+1)/d) \right) \\
&= n \sum_{i=1}^{d} \alpha_i^2 + \frac{3}{2}d - \frac{3}{2} \qquad (5.8)
\end{aligned}$$

where the last inequality is deduced exactly as in the proof of Lemma 4.1.1. Finally, combining (5.3)–(5.8) we get

$$n^2 \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}, \boldsymbol{U}} \|\boldsymbol{U} \mathrm{diag}(\underline{\boldsymbol{\lambda}}) \boldsymbol{U}^\dagger - \rho\|_F^2 \le 4dn - 3n.$$

Dividing both sides by n^2 completes the proof. □

5.3 Keyl's algorithm

In this section we analyze the tomography algorithm proposed by Keyl [Key06] based on projection to the highest weight vector $|T_\lambda\rangle$.

Definition 5.3.1 (Keyl's algorithm)**.** Given n copies of ρ,

1. Perform weak Schur sampling on ρ, resulting in a random $\boldsymbol{\lambda}$. Discard the permutation irrep register. Then ρ collapses to $\pi_{\boldsymbol{\lambda}}(\rho)/s_{\boldsymbol{\lambda}}(\alpha)$.

2. Measure within the space $\mathrm{V}^d_{\boldsymbol{\lambda}}$ using the POVM with elements

$$\pi_\lambda(U)\,|T_\lambda\rangle\,\langle T_\lambda|\,\pi_\lambda(U)^\dagger \cdot \dim(\mathrm{V}^d_\lambda)\, dU$$

for each $U \in \mathrm{U}(d)$.

3. Output $\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger$.

(To see that this is indeed a POVM — i.e., that

$$M \coloneqq \int \pi_\lambda(U) \left|T_\lambda\right\rangle \left\langle T_\lambda\right| \pi_\lambda(U)^\dagger \cdot \dim(\mathrm{V}_\lambda^d)\, dU,$$

is the identity matrix, — first note that the translation invariance of Haar measure implies $\pi_\lambda(V) M \pi_\lambda(V)^\dagger = M$ for any $V \in \mathrm{U}(d)$. Thinking of π_λ as an irreducible representation of the unitary group, Schur's lemma implies M must be a scalar matrix. Taking traces shows M is the identity.)

We write $\mathrm{K}_\lambda(\rho)$ for the probability distribution on $\mathrm{U}(d)$ associated to this POVM; its density with respect to the Haar measure is therefore

$$\mathrm{tr}\left(\pi_\lambda(\tfrac{1}{s_\lambda(\alpha)}\rho)\pi_\lambda(U) \left|T_\lambda\right\rangle \left\langle T_\lambda\right| \pi_\lambda(U)^\dagger \cdot \dim(\mathrm{V}_\lambda^d)\right) = \Phi_\lambda(\alpha)^{-1} \cdot \left\langle T_\lambda\right| \pi_\lambda(U^\dagger \rho U) \left|T_\lambda\right\rangle. \tag{5.9}$$

Supposing the outcome of the measurement is U, Keyl's final estimate for ρ is $\hat{\rho} = U\mathrm{diag}(\underline{\lambda})U^\dagger$. Thus the expected Frobenius-squared error of Keyl's tomography algorithm is precisely

$$\mathop{\mathbf{E}}_{\substack{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)\\ \boldsymbol{U}\sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)}} \|\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2.$$

Theorem 5.0.1, which we prove in this section, bounds the above quantity by $\frac{4d-3}{n}$.

5.3.1 Integration formulas

Notation 5.3.2. Let $Z \in \mathbb{C}^{d\times d}$ and let λ be a partition of length at most d. The *generalized power function* Δ_λ is defined by

$$\Delta_\lambda(Z) = \prod_{k=1}^d \mathrm{pm}_k(Z)^{\lambda_k - \lambda_{k+1}},$$

where $\mathrm{pm}_k(Z)$ denotes the k-th principal minor of Z (and $\lambda_{d+1} = 0$).

As noted by Keyl [Key06, equation (141)], when Z is positive semidefinite we have $\left\langle T_\lambda\right| \pi_\lambda(Z) \left|T_\lambda\right\rangle = \Delta_\lambda(Z)$; this follows by writing $Z = LL^\dagger$ for $L = (L_{ij})$ lower triangular with nonnegative diagonal and using the fact that $\Delta_\lambda(Z) = \Delta_\lambda(L^\dagger)^2 = \prod_{k=1}^d L_{kk}^{2\lambda_k}$. Putting this into (5.9) we have an alternate definition for the distribution $\mathrm{K}_\lambda(\rho)$:

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{K}_\lambda(\rho)} f(\boldsymbol{U}) = \Phi_\lambda(\alpha)^{-1} \mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{U}(d)} \left[f(\boldsymbol{U}) \cdot \Delta_\lambda(\boldsymbol{U}^\dagger \rho \boldsymbol{U})\right], \tag{5.10}$$

where $\boldsymbol{U} \sim \mathrm{U}(d)$ denotes that $\boldsymbol{U}$ has the Haar measure. For example, taking $f \equiv 1$ yields the identity

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{U}(d)} \Delta_\lambda(\boldsymbol{U}^\dagger \rho \boldsymbol{U}) = \Phi_\lambda(\alpha); \tag{5.11}$$

this expresses the fact that the spherical polynomial of weight λ for $\mathrm{GL}_d/\mathrm{U}(d)$ is precisely the normalized Schur polynomial (see, e.g., [Far15]). For a further example, taking $f(U) = \Delta_\mu(U^\dagger \rho U)$ and using the fact that $\Delta_\lambda \cdot \Delta_\mu = \Delta_{\lambda+\mu}$, we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{K}_\lambda(\rho)} \Delta_\mu(\boldsymbol{U}^\dagger \rho \boldsymbol{U}) = \frac{\Phi_{\lambda+\mu}(\alpha)}{\Phi_\lambda(\alpha)}; \quad \text{in particular,} \mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{K}_\lambda(\rho)} (\boldsymbol{U}^\dagger \rho \boldsymbol{U})_{1,1} = \frac{\Phi_{\lambda+e_1}(\alpha)}{\Phi_\lambda(\alpha)}. \tag{5.12}$$

For our proof of Theorem 5.0.1, we will need to develop and analyze a more general formula for the expected diagonal entry $\mathbf{E}(\boldsymbol{U}^\dagger\rho\boldsymbol{U})_{k,k}$. We begin with some lemmas.

Definition 5.3.3. For λ a partition and m a positive integer we define the following partition of height (at most) m:

$$\lambda^{[m]} = (\lambda_1 - \lambda_{m+1}, \dots, \lambda_m - \lambda_{m+1}).$$

We also define the following "complementary" partition $\lambda_{[m]}$ satisfying $\lambda = \lambda^{[m]} + \lambda_{[m]}$:

$$(\lambda_{[m]})_i = \begin{cases} \lambda_{m+1} & i \leq m, \\ \lambda_i & i \geq m+1. \end{cases}$$

Lemma 5.3.4. *Let $\rho \in \mathbb{C}^{d\times d}$ be a density matrix with spectrum α and let $\lambda \vdash n$ have height at most d. Let $m \in [d]$ and let f_m be an m-variate symmetric polynomial. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{K}_\lambda(\rho)} f_m(\boldsymbol{\beta}) = \Phi_\lambda(\alpha)^{-1} \cdot \mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{U}(d)} \left[f_m(\boldsymbol{\beta}) \cdot \Phi_{\lambda^{[m]}}(\boldsymbol{\beta}) \cdot \Delta_{\lambda_{[m]}}(\boldsymbol{U}^\dagger\rho\boldsymbol{U}) \right],$$

where we write $\boldsymbol{\beta} = \mathrm{spec}_m(\boldsymbol{U}^\dagger\rho\boldsymbol{U})$ for the spectrum of the top-left $m\times m$ submatrix of $\boldsymbol{U}^\dagger\rho\boldsymbol{U}$.

Proof. Let $\boldsymbol{V} \sim \mathrm{U}(m)$ and write $\overline{\boldsymbol{V}} = \boldsymbol{V} \oplus I$, where I is the $(d-m)$-dimensional identity matrix. By translation-invariance of Haar measure we have $\boldsymbol{U}\overline{\boldsymbol{V}} \sim \mathrm{U}(d)$, and hence from (5.10),

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{K}_\lambda(\rho)} f_m(\boldsymbol{\beta}) = \Phi_\lambda(\alpha)^{-1} \mathop{\mathbf{E}}_{\boldsymbol{U}\sim \mathrm{U}(d), \boldsymbol{V}\sim\mathrm{U}(m)} \left[f_m(\mathrm{spec}_m(\overline{\boldsymbol{V}}^\dagger\boldsymbol{U}^\dagger\rho\boldsymbol{U}\overline{\boldsymbol{V}})) \cdot \Delta_\lambda(\overline{\boldsymbol{V}}^\dagger\boldsymbol{U}^\dagger\rho\boldsymbol{U}\overline{\boldsymbol{V}}) \right]. \tag{5.13}$$

Note that conjugating a matrix by $\overline{\boldsymbol{V}}$ does not change the spectrum of its upper-left $k\times k$ block for any $k \geq m$. Thus $\mathrm{spec}_m(\overline{\boldsymbol{V}}^\dagger\boldsymbol{U}^\dagger\rho\boldsymbol{U}\overline{\boldsymbol{V}})$ is identical to β, and $\mathrm{pm}_k(\overline{\boldsymbol{V}}^\dagger\boldsymbol{U}^\dagger\rho\boldsymbol{U}\overline{\boldsymbol{V}}) = \mathrm{pm}_k(\boldsymbol{U}^\dagger\rho\boldsymbol{U})$ for all $k \geq m$. Thus using $\Delta_\lambda = \Delta_{\lambda_{[m]}} \cdot \Delta_{\lambda^{[m]}}$ we have

$$(5.13) = \Phi_\lambda(\alpha)^{-1} \mathop{\mathbf{E}}_{\boldsymbol{U}\sim\mathrm{U}(d)} \left[f_m(\boldsymbol{\beta}) \cdot \Delta_{\lambda_{[m]}}(\boldsymbol{U}^\dagger\rho\boldsymbol{U}) \cdot \mathop{\mathbf{E}}_{\boldsymbol{V}\sim\mathrm{U}(m)} \left[\Delta_{\lambda^{[m]}}(\overline{\boldsymbol{V}}^\dagger\boldsymbol{U}^\dagger\rho\boldsymbol{U}\overline{\boldsymbol{V}}) \right] \right].$$

But the inner expectation equals $\Phi_{\lambda^{[m]}}(\boldsymbol{\beta})$ by (5.11), completing the proof. □

Lemma 5.3.5. *In the setting of Lemma 5.3.4,*

$$\mathop{\mathbf{E}}_{\boldsymbol{U}\sim\mathrm{K}_\lambda(\rho)} \mathop{\mathrm{avg}}_{i=1}^{m} \left\{ (\boldsymbol{U}^\dagger\rho\boldsymbol{U})_{i,i} \right\} = \sum_{i=1}^{m} \frac{s_{\lambda^{[m]}+e_i}(1/m)}{s_{\lambda^{[m]}}(1/m)} \cdot \frac{\Phi_{\lambda+e_i}(\alpha)}{\Phi_\lambda(\alpha)}, \tag{5.14}$$

where $1/m$ abbreviates $1/m, \dots, 1/m$ (repeated m times).

Remark 5.3.6. The right-hand side of (5.14) is also a weighted average — of the quantities $\Phi_{\lambda+e_i}(\alpha)/\Phi_\lambda(\alpha)$ — by virtue of Fact 3.4.2. The lemma also generalizes (5.12), as $s_{\lambda^{[1]}+e_1}(1)/s_{\lambda^{[1]}}(1)$ is simply 1.

Proof. On the left-hand side of (5.14) we have $\frac{1}{m}$ times the expected trace of the upper-left $m \times m$ submatrix of $\boldsymbol{U}^\dagger \rho \boldsymbol{U}$. So by applying Lemma 5.3.4 with $f_m(\beta) = \frac{1}{m}(\beta_1 + \cdots + \beta_m)$, it is equal to

$$\begin{aligned}
&\Phi_\lambda(\alpha)^{-1} \cdot \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \left[\frac{1}{m}(\boldsymbol{\beta}_1 + \cdots + \boldsymbol{\beta}_m) \cdot \frac{s_{\lambda^{[m]}}(\boldsymbol{\beta})}{s_{\lambda^{[m]}}(1, \dots, 1)} \cdot \Delta_{\lambda_{[m]}}(\boldsymbol{U}^\dagger \rho \boldsymbol{U}) \right] \\
&= \Phi_\lambda(\alpha)^{-1} \cdot \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \left[\frac{1}{m} \sum_{i=1}^{m} \frac{s_{\lambda^{[m]}+e_i}(\boldsymbol{\beta})}{s_{\lambda^{[m]}}(1, \dots, 1)} \cdot \Delta_{\lambda_{[m]}}(\boldsymbol{U}^\dagger \rho \boldsymbol{U}) \right] \qquad \text{(by Pieri's rule)} \\
&= \Phi_\lambda(\alpha)^{-1} \cdot \sum_{i=1}^{m} \frac{s_{\lambda^{[m]}+e_i}(1, \dots, 1)}{m \cdot s_{\lambda^{[m]}}(1, \dots, 1)} \cdot \mathop{\mathbf{E}}_{\boldsymbol{U} \sim \mathrm{U}(d)} \left[\Phi_{\lambda^{[m]}+e_i}(\boldsymbol{\beta}) \cdot \Delta_{\lambda_{[m]}}(\boldsymbol{U}^\dagger \rho \boldsymbol{U}) \right] \\
&= \Phi_\lambda(\alpha)^{-1} \cdot \sum_{i=1}^{m} \frac{s_{\lambda^{[m]}+e_i}(1, \dots, 1)}{m \cdot s_{\lambda^{[m]}}(1, \dots, 1)} \cdot \Phi_{\lambda+e_i}(\alpha),
\end{aligned}$$

where in the last step we used Lemma 5.3.4 again, with $f_m \equiv 1$ and $\lambda + e_i$ in place of λ. But this is equal to the right-hand side of (5.14), using the homogeneity of Schur polynomials. □

Lemma 5.3.7. *Assume the setting of Lemma 5.3.4. Then $\eta_i \coloneqq \mathbf{E}_{\boldsymbol{U} \sim \mathrm{K}_\lambda(\rho)}(\boldsymbol{U}^\dagger \rho \boldsymbol{U})_{m,m}$ is a convex combination of the quantities $R_i \coloneqq \Phi_{\lambda+e_i}(\alpha)/\Phi_\lambda(\alpha)$, $1 \le i \le m$.*[1]

Proof. This is clear for $m = 1$. For $m > 1$, Remark 5.3.6 implies

$$\operatorname*{avg}_{i=1}^{m}\{\eta_i\} = p_1 R_1 + \cdots + p_m R_m, \qquad \operatorname*{avg}_{i=1}^{m-1}\{\eta_i\} = q_1 R_1 + \cdots + q_m R_m,$$

where $p_1 + \cdots + p_m = q_1 + \cdots + q_m = 1$ and $q_m = 0$. Thus $\eta_i = \sum_{i=1}^m r_i R_i$, where $r_i = (mp_i - (m-1)q_i)$, and evidently $\sum_{i=1}^m r_i = m - (m-1) = 1$. It remains to verify that each $r_i \ge 0$. This is obvious for $i = m$; for $i < m$, we must check that

$$\frac{s_{\lambda^{[m]}+e_i}(1, \dots, 1)}{s_{\lambda^{[m]}}(1, \dots, 1)} \ge \frac{s_{\lambda^{[m-1]}+e_i}(1, \dots, 1)}{s_{\lambda^{[m-1]}}(1, \dots, 1)}. \tag{5.15}$$

Using the Weyl dimension formula from Definition 2.2.13, one may explicitly compute that the ratio of the left side of (5.15) to the right side is precisely $1 + \frac{1}{(\lambda_i - \lambda_m) + (m-i)} \ge 1$. This completes the proof. □

We will in fact only need the following corollary:

Corollary 5.3.8. *Let $\rho \in \mathbb{C}^{d \times d}$ be a density matrix with spectrum α and let $\lambda \vdash n$ have height at most d. Then $\mathbf{E}_{\boldsymbol{U} \sim \mathrm{K}_\lambda(\rho)}(\boldsymbol{U}^\dagger \rho \boldsymbol{U})_{m,m} \ge \Phi_{\lambda+e_m}(\alpha)/\Phi_\lambda(\alpha)$ for every $m \in [d]$,*

Proof. This is immediate from Lemma 5.3.7 and the fact that $\Phi_{\lambda+e_i}(\alpha) \ge \Phi_{\lambda+e_m}(\alpha)$ whenever $i < m$ (assuming $\lambda + e_i$ is a valid partition). This latter fact was recently proved by Sra [Sra15], verifying a conjecture of Cuttler et al. [CGS11]. □

[1] To be careful, we may exclude all those i for which $\lambda + e_i$ is an invalid partition and thus $R_i = 0$.

5.3.2 Proof of Theorem 5.0.1

Throughout the proof we assume $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$ and $\boldsymbol{U} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)$. We have

$$n^2 \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2 = n^2 \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{U}^\dagger \rho \boldsymbol{U}\|_F^2$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^{d} \boldsymbol{\lambda}_i^2 + \sum_{i=1}^{d} (\alpha_i n)^2 - 2n \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \sum_{i=1}^{d} \boldsymbol{\lambda}_i (\boldsymbol{U}^\dagger \rho \boldsymbol{U})_{i,i} \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^{d} \boldsymbol{\lambda}_i^2 + \sum_{i=1}^{d} (\alpha_i n)^2 - 2n \mathop{\mathbf{E}}_{\lambda} \sum_{i=1}^{d} \boldsymbol{\lambda}_i \frac{\Phi_{\boldsymbol{\lambda}+e_i}(\alpha)}{\Phi_{\boldsymbol{\lambda}}(\alpha)},$$

using Corollary 5.3.8. This equation is equal to Equation (5.4) which was shown to be at most $4dn - 3n$ in the proof of Theorem 5.2.3. Dividing by n^2 gives the desired Frobenius-squared bound. □

5.4 Principal component analysis

In this section we analyze a straightforward modification to Keyl's tomography algorithm that allows us to perform principal component analysis on an unknown density matrix $\rho \in \mathbb{C}^{d \times d}$. The PCA algorithm is the same as Keyl's algorithm, except that having measured $\boldsymbol{\lambda}$ and $\boldsymbol{U}$, it outputs the rank-k matrix $\boldsymbol{U}\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger$ rather than the potentially full-rank matrix $\boldsymbol{U}\mathrm{diag}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger$. (Here we use the notation $\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})$ for the $d \times d$ matrix $\mathrm{diag}(\underline{\boldsymbol{\lambda}}_1, \ldots, \underline{\boldsymbol{\lambda}}_k, 0, \ldots, 0)$.) Thus the expected Frobenius-squared error of Keyl's tomography algorithm is precisely

$$\mathop{\mathbf{E}}_{\substack{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha) \\ \boldsymbol{U} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)}} \|\boldsymbol{U}\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_F^2.$$

Theorem 5.0.2 bounds the above quantity by $\alpha_{k+1} + \ldots + \alpha_d + 6\sqrt{kd/n}$.

Before giving the proof of Theorem 5.0.2, let us show why the case of Frobenius-norm PCA appears to be less interesting than the case of trace-distance PCA. The goal for Frobenius PCA would be to output a rank-k matrix $\hat{\rho}$ satisfying

$$\|\hat{\rho} - \rho\|_F \leq \sqrt{\alpha_{k+1}^2 + \ldots + \alpha_d^2} + \epsilon,$$

with high probability, while trying to minimize the number of copies n as a function of k, d, and ϵ. However, even when ρ is guaranteed to be of rank 1, it is likely that any algorithm will require $n = \Omega(d/\epsilon^2)$ copies to output an ϵ-accurate rank-1 approximator $\hat{\rho}$. This is because such an approximator will satisfy $\|\hat{\rho} - \rho\|_1 \leq \sqrt{2} \cdot \|\hat{\rho} - \rho\|_F = O(\epsilon)$, and it is likely that $n = \Omega(d/\epsilon^2)$ copies of ρ are required for such a guarantee (see, for example, the lower bounds of [HHJ+16], which show that $n = \Omega(\frac{d}{\epsilon^2 \log(d/\epsilon)})$ copies are necessary for tomography of rank-1 states.). Thus, even in the simplest case of rank-1 PCA of rank-1 states, we probably cannot improve on the $n = O(d/\epsilon^2)$ copy complexity for full tomography given by Theorem 5.0.2.

Now we prove Theorem 5.0.2. We note that the proof shares many of its steps with the proof of Theorem 5.0.1.

Proof of Theorem 5.0.2. Throughout the proof we assume $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$ and $\boldsymbol{U} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)$. We write $\boldsymbol{R}$ for the lower-right $(d-k)\times(d-k)$ submatrix of $\boldsymbol{U}^\dagger \rho \boldsymbol{U}$ and we write $\boldsymbol{\Gamma} = \boldsymbol{U}^\dagger\rho\boldsymbol{U} - \boldsymbol{R}$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_1 = \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{U}^\dagger\rho\boldsymbol{U}\|_1 \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{\Gamma}\|_1 + \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{R}\|_1. \tag{5.16}$$

We can upper-bound the first term in (5.16) using

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{\Gamma}\|_1 \leq \sqrt{2k} \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{\Gamma}\|_F \leq \sqrt{2k} \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\mathrm{diag}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{U}^\dagger\rho\boldsymbol{U}\|_F \leq \sqrt{\frac{8kd}{n}}. \tag{5.17}$$

The first inequality is Cauchy–Schwarz together with the fact that $\mathrm{rank}(\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{\Gamma}) \leq 2k$ (since the matrix is nonzero only in its first k rows and columns). The second inequality uses that $\mathrm{diag}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{U}^\dagger\rho\boldsymbol{U}$ is formed from $\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{\Gamma}$ by adding a matrix, $\mathrm{diag}(\underline{\boldsymbol{\lambda}}) - \mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}}) - \boldsymbol{R}$, of disjoint support; this can only increase the squared Frobenius norm (sum of squares of entries). Finally, the third inequality uses Theorem 5.0.1. To analyze the second term in (5.16), we note that $\boldsymbol{R}$ is a principal submatrix of $\boldsymbol{U}^\dagger\rho\boldsymbol{U}$, and so it is positive semidefinite. As a result,

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{R}\|_1 = \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \mathrm{tr}(\boldsymbol{R}) = 1 - \mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \mathrm{tr}(\boldsymbol{\Gamma}). \tag{5.18}$$

By Corollary 5.3.8,

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \mathrm{tr}(\boldsymbol{\Gamma}) &= \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \mathop{\mathbf{E}}_{\boldsymbol{U}} (\boldsymbol{U}^\dagger\rho\boldsymbol{U})_{i,i} \geq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{\Phi_{\boldsymbol{\lambda}+e_i}(\alpha)}{\Phi_{\boldsymbol{\lambda}}(\alpha)} = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}}(1,\dots,1)}{s_{\boldsymbol{\lambda}+e_i}(1,\dots,1)} \\ &\geq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \left(2 - \frac{s_{\boldsymbol{\lambda}+e_i}(1,\dots,1)}{s_{\boldsymbol{\lambda}}(1,\dots,1)}\right) = 2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} - \mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}+e_i}(1,\dots,1)}{s_{\boldsymbol{\lambda}}(1,\dots,1)},\end{aligned} \tag{5.19}$$

where we used $r \geq 2 - \frac{1}{r}$ for $r > 0$. The first term here is lower-bounded using Proposition 3.4.5:

$$2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^k \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \geq 2\sum_{i=1}^k \alpha_i. \tag{5.20}$$

As for the second term in (5.19), we use the definition of the Schur-Weyl distribution and the first formula in Definition 2.2.13 to compute

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\lambda}} \sum_{i=1}^{k} \frac{s_{\boldsymbol{\lambda}+e_i}(\alpha)}{s_{\boldsymbol{\lambda}}(\alpha)} \frac{s_{\boldsymbol{\lambda}+e_i}(1,\dots,1)}{s_{\boldsymbol{\lambda}}(1,\dots,1)} &= \sum_{i=1}^{k} \sum_{\lambda \vdash n} \dim(\lambda) s_\lambda(\alpha) \cdot \frac{s_{\lambda+e_i}(\alpha)}{s_\lambda(\alpha)} \frac{\dim(\lambda+e_i)(d+\lambda_i-i+1)}{\dim(\lambda)(n+1)} \\
&= \sum_{i=1}^{k} \sum_{\lambda \vdash n} \dim(\lambda+e_i) s_{\lambda+e_i}(\alpha) \cdot \frac{(d-i+\lambda_i+1)}{n+1} \\
&\leq \sum_{i=1}^{k} \mathop{\mathbf{E}}_{\boldsymbol{\lambda}' \sim \mathrm{SW}^{n+1}(\alpha)} \frac{(d-i+\boldsymbol{\lambda}'_i)}{n+1} \\
&\leq \frac{1}{n+1} \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}' \sim \mathrm{SW}^{n+1}(\alpha)} \sum_{i=1}^{k} \boldsymbol{\lambda}'_i + \frac{kd}{n} \\
&\leq \sum_{i=1}^{k} \alpha_i + \frac{2\sqrt{2}k}{\sqrt{n}} + \frac{kd}{n}, \qquad (5.21)
\end{aligned}$$

where the last step is by Lemma 4.2.1. Combining (5.16)–(5.21) we get

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_1 \leq \left(1 - \sum_{i=1}^{k} \alpha_i\right) + \sqrt{\frac{8kd}{n}} + \frac{2\sqrt{2}k}{\sqrt{n}} + \frac{kd}{n} \leq \sum_{i=k+1}^{d} \alpha_i + \sqrt{\frac{32kd}{n}} + \frac{kd}{n},$$

where the second inequality used $k \leq \sqrt{kd}$. Finally, as the expectation is also trivially upper-bounded by 2, we may use $6\sqrt{r} \geq \min(2, \sqrt{32r} + r)$ (which holds for all $r \geq 0$) to conclude

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{U}} \|\boldsymbol{U}\mathrm{diag}^{(k)}(\underline{\boldsymbol{\lambda}})\boldsymbol{U}^\dagger - \rho\|_1 \leq \sum_{i=k+1}^{d} \alpha_i + 6\sqrt{\frac{kd}{n}}. \qquad \square$$

5.5 A lower bound

We now prove Theorem 5.0.3. Similar to the lower bound proofs from [HHJ+16], we choose a hard ensemble of mixed states for tomography and perform a careful analysis of their Holevo information.

Proof of Theorem 5.0.3. From [FGLE12, Lemma 5], there exists a set of rank-r projectors $\rho_1, \dots, \rho_s$ for which $s \geq 2^{\Omega(rd)}$ and $d_{\mathrm{tr}}(\rho_i, \rho_j) \geq \epsilon_0$ for $i \neq j$, where ϵ_0 is some absolute constant. Consider the communication scenario in which Angela selects a message $\boldsymbol{m} \in [s]$ uniformly at random, encodes it in the state $\rho_{\boldsymbol{m}}^{\otimes n}$, and then Bob measures the state $\rho_{\boldsymbol{m}}^{\otimes n}$ to produce the message $\widetilde{\boldsymbol{m}}$. Supposing that the tomography algorithm in the theorem statement succeeds with probability $(1-\eta)$, then Bob can successfully decode Angela's message with probability $(1-\eta)$. Thus, by Fano's inequality (cf. [HHJ+16, Equation (21)])

$$I(\boldsymbol{m}; \widetilde{\boldsymbol{m}}) \geq (1-\eta)\log(s) \geq \Omega(rd).$$

If we write $\mathcal{E}$ for the ensemble $\{1/s, \rho_i\}_{i \in [s]}$, then by Holevo's theorem, $I(\boldsymbol{m}; \widetilde{\boldsymbol{m}}) \leq \chi(\mathcal{E})$, where $\chi(\mathcal{E})$ is the Holevo information. Writing avg for the averaged state $\mathbf{E}_{\boldsymbol{m}}[\rho_{\boldsymbol{m}}^{\otimes n}]$, we have that

$$\chi(\mathcal{E}) = H\left(\mathop{\mathbf{E}}_{\boldsymbol{m}}[\rho_{\boldsymbol{m}}^{\otimes n}]\right) - \mathop{\mathbf{E}}_{\boldsymbol{m}}\left[H(\rho_{\boldsymbol{m}}^{\otimes n})\right] = H(\mathsf{avg}) - \mathop{\mathbf{E}}_{\boldsymbol{m}}\left[H(\rho_{\boldsymbol{m}}^{\otimes n})\right] = H(\mathsf{avg}) - n\log r,$$

where the last equality uses the fact that every ρ_i is a rank-r projector. By Schur-Weyl duality,

$$\rho_i^{\otimes n} \cong \bigoplus_{\substack{\lambda\vdash n\\ \ell(\lambda)\le d}} \pi_\lambda(\rho_i)\otimes I_\lambda \cong \bigoplus_{\substack{\lambda\vdash n\\ \ell(\lambda)\le r}} \pi_\lambda(\rho_i)\otimes I_\lambda,$$

where the second step uses the fact that ρ_i is a rank-r projector. Here I_λ denotes the $\dim(\lambda)\times\dim(\lambda)$ identity matrix acting on the Specht module Sp_λ. Each $\rho_i^{\otimes n}$ is supported only on the subspace of the Schur basis corresponding to λ's of height at most r, and so avg is also supported only on the subspace of the Schur basis corresponding to λ's of height at most r. At worst, avg is maximally mixed on this subspace, meaning that

$$\chi(\mathcal{E}) = H(\mathsf{avg}) - n\log r \le \log\dim(\lambda\vdash n, \ell(\lambda)\le r) - n\log r.$$

Write Par_r^n for the set of partitions $\lambda\vdash n$ of height at most r. Our goal is to bound the sum

$$\sum_{\lambda\in\mathrm{Par}_r^n} \dim(\lambda)\dim(\mathrm{V}_\lambda^d),$$

and we will do so with the following two facts:

- First, $\sum_{\lambda\in\mathrm{Par}_r^n}\dim(\lambda)\le r^n$. This follows from Schur-Weyl duality applied to $(\mathbb{C}^r)^{\otimes n}$:

  $$(\mathbb{C}^r)^{\otimes n} \cong \bigoplus_{\substack{\lambda\vdash n\\ \ell(\lambda)\le r}} \mathrm{Sp}_\lambda\otimes \mathrm{V}_\lambda^r.$$

  Ignoring the Weyl modules, the dimensionality of the right-hand side is at least $\sum_{\lambda\in\mathrm{Par}_r^n}\dim(\lambda)$, which is at most the dimensionality of the left-hand side, r^n.

- Write $\overline{H}(x) = (1+x)H(\frac{1}{1+x})$. Then for any $\lambda\in\mathrm{Par}_r^n$, $\dim(\mathrm{V}_\lambda^d)\le 2^{r(d-1)\cdot\overline{H}(n/r(d-1))}$. This is because $\dim(\mathrm{V}_\lambda^d)$ is equal to the number of semistandard tableaus of shape λ and alphabet $[d]$. Given such a tableau, each row i contains a weakly increasing sequence of length λ_i, and the total number of such sequences is at most

  $$\binom{\lambda_i+d-1}{d-1} \le 2^{(\lambda_i+d-1)H((d-1)/(\lambda_i+d-1))} = 2^{(d-1)\overline{H}(\lambda_i/(d-1))}.$$

  Thus,

  $$\begin{aligned}\dim(\mathrm{V}_\lambda^d) \le 2^{(d-1)\sum_{i=1}^r \overline{H}(\lambda_i/(d-1))} &= 2^{r(d-1)\,\mathrm{avg}_i\{\overline{H}(\lambda_i/(d-1))\}}\\ &\le 2^{r(d-1)\overline{H}(\mathrm{avg}_i\{\lambda_i/(d-1)\})} = 2^{r(d-1)\cdot\overline{H}(n/r(d-1))},\end{aligned}$$

  where the second inequality follows from Jensen's inequality applied to the concave $\overline{H}(\cdot)$.

In summary,

$$\sum_{\lambda \in \mathrm{Par}_r^n} \dim(\lambda) \dim(\mathrm{V}_\lambda^d) \leq \sum_{\lambda \in \mathrm{Par}_r^n} \dim(\lambda) \cdot 2^{r(d-1) \cdot \overline{H}(n/r(d-1))} \leq r^n \cdot 2^{r(d-1) \cdot \overline{H}(n/r(d-1))}.$$

Thus, the Holevo information is bounded by

$$\chi(\mathcal{E}) \leq \log\left(r^n \cdot 2^{r(d-1) \cdot \overline{H}(n/r(d-1))}\right) - n \log r = r(d-1) \cdot \overline{H}(n/r(d-1)),$$

and so we have

$$r(d-1) \cdot \overline{H}(n/r(d-1)) \geq \chi(\mathcal{E}) \geq I(\boldsymbol{m}; \widetilde{\boldsymbol{m}}) \geq \Omega(rd).$$

Dividing both sides by rd, we require that $\overline{H}(n/r(d-1)) \geq \Omega(1)$. Inspecting $\overline{H}(\cdot)$, it is clear that this is satisfied only if $n = \Omega(rd)$. □

Chapter 6

A quantum Paninski theorem

In this section, we prove Theorem 1.4.23, that $\Theta(d/\epsilon^2)$ copies are necessary and sufficient to test whether or not a given state $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state, i.e., has spectrum $(\frac{1}{d}, \dots, \frac{1}{d})$.

6.1 The upper bound

The upper bound for Theorem 1.4.23 will follow from our analysis of the following simple algorithm.

Definition 6.1.1.]Mixedness Tester] Given $\rho^{\otimes n}$, where ρ is d-dimensional:

1. Sample $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$.
2. Accept if $p^\sharp_2(\boldsymbol{\lambda}) \leq \left(1 + \frac{\epsilon^2}{2}\right) \cdot \frac{n(n-1)}{d}$. Reject otherwise.

We remark that the tester Childs et al. [CHW07] used to distinguish the maximally mixed states of dimension $\frac{d}{2}$ and d also depended only on the magnitude of $p^\sharp_2(\boldsymbol{\lambda}) = 2c_1(\boldsymbol{\lambda})$; see [CHW07, equations (49), (50)].

Theorem 6.1.2. *The Mixedness Tester can test whether a state $\rho \in \mathbb{C}^{d\times d}$ is the maximally mixed state using $n = O(d/\epsilon^2)$ copies of ρ.*

Proof. We will run the Mixedness Tester with $n = 100d/\epsilon^2$. Both the "completeness" and the "soundness" analysis will require the last identity from (3.11), namely

$$(p^\sharp_2)^2 = p^\sharp_{(2,2)} + 4p^\sharp_3 + 2p^\sharp_{(1,1)}. \tag{6.1}$$

Completeness. Suppose first that ρ is the maximally mixed state, so that in fact $\boldsymbol{\lambda} \sim \mathrm{SW}^n_d$. We compute the mean and variance of $p^\sharp_2(\boldsymbol{\lambda})$ using (6.1) and Corollary 3.8.4:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}[p^\sharp_2(\boldsymbol{\lambda})] = \frac{n(n-1)}{d}, \tag{6.2}$$

$$\mathop{\mathbf{Var}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[p^\sharp_2(\boldsymbol{\lambda})\right] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[p^\sharp_2(\boldsymbol{\lambda})^2\right] - \left(\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}[p^\sharp_2(\boldsymbol{\lambda})]\right)^2 = \frac{2n(n-1)(d^2-1)}{d^2} \leq 2n(n-1). \tag{6.3}$$

Thus by Chebyshev's inequality,

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[p_2^\sharp(\boldsymbol{\lambda}) > \left(1+\frac{\epsilon^2}{2}\right)\cdot\frac{n(n-1)}{d}\right] \leq \frac{8d^2}{n(n-1)\epsilon^4} \leq \frac{1}{3},$$

by our choice of n. Thus indeed when ρ is the maximally mixed state, the Mixedness Tester accepts with probability at least $2/3$.

Soundness. Suppose now that ρ is a density matrix whose spectrum $\eta = (\eta_1, \dots, \eta_d)$ satisfies $d_{\mathrm{TV}}^{\mathrm{sym}}(\eta, \mathsf{Unif}_d) \geq \epsilon$. Writing $\eta_i = \frac{1}{d} + \Delta_i$, this means that

$$\epsilon \leq \frac{1}{2}\cdot\sum_{i=1}^d |\Delta_i| \leq \frac{1}{2}\sqrt{d\cdot\sum_{i=1}^d \Delta_i^2},$$

using Cauchy–Schwarz; hence

$$\sum_{i=1}^d \Delta_i^2 \geq \frac{4\epsilon^2}{d}. \tag{6.4}$$

Using (6.1) and Proposition 3.8.3, we can calculate the difference between the mean of $p_2^\sharp(\boldsymbol{\lambda})$ and the cutoff used by the Mixedness Tester as

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_\rho^n}\left[p_2^\sharp(\boldsymbol{\lambda})\right] - \frac{n(n-1)}{d}\cdot\left(1+\frac{\epsilon^2}{2}\right) &= n(n-1)\left(\sum_{i=1}^d \eta_i^2 - \frac{1}{d}\left(1+\frac{\epsilon^2}{2}\right)\right). \\
&= n(n-1)\left(\sum_{i=1}^d \Delta_i^2 - \frac{\epsilon^2}{2d}\right) \\
&\geq \frac{n(n-1)}{2}\sum_{i=1}^d \Delta_i^2,
\end{aligned}$$

where the last line follows from (6.4). Similarly, we can calculate the variance of $p_2^\sharp(\boldsymbol{\lambda})$ as

$$\begin{aligned}
\mathop{\mathbf{Var}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_\rho^n}\left[p_2^\sharp(\boldsymbol{\lambda})\right] &= n(n-1)\left(2 + 4n\left(\sum \eta_i^3 - \left(\sum \eta_i^2\right)^2\right) + 6\left(\sum \eta_i^2\right)^2 - 8\sum \eta_i^3\right) \\
&\leq n(n-1)\left(8 + 4n\left(\sum \eta_i^3 - \left(\sum \eta_i^2\right)^2\right)\right) \\
&= n(n-1)\left(8 + 4n\left(\frac{1}{d}\sum \Delta_i^2 + \sum \Delta_i^3 - \left(\sum \Delta_i^2\right)^2\right)\right) \\
&\leq n(n-1)\left(8 + 8n\left(\sum \Delta_i^2\right)\right).
\end{aligned}$$

Applying Chebyshev's inequality gives us

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_\rho^n}\left[p_{(2)}^\sharp(\boldsymbol{\lambda}) < \left(1+\frac{\epsilon^2}{2}\right)\cdot\frac{n(n-1)}{d}\right] &\leq \frac{1}{n(n-1)\left(\sum_{i=1}^d \Delta_i^2\right)^2}\cdot\left(32 + 32n\left(\sum_{i=1}^d \Delta_i^2\right)\right) \\
&\leq \frac{4}{n^2\left(\epsilon^2/d\right)^2} + \frac{16}{n\left(\epsilon^2/d\right)},
\end{aligned}$$

where the second step follows from (6.4). By our choice of n, this is at most $1/3$. Thus, when ρ is ϵ-far from the maximally mixed state, the Mixedness Tester rejects with probability at least $2/3$, as required. □

6.2 The lower bound: overview

For almost all of the lower bound proof we will assume d is even. In the end we will indicate how to obtain the lower bound when d is odd. For $0 \leq \epsilon \leq \frac{1}{2}$, let P_d^ϵ denote the probability distribution on $[d]$ in which

$$\mathsf{P}_d^\epsilon(j) = \frac{1 + (-1)^{j-1} 2\epsilon}{d}.$$

This is essentially the same probability distribution that Paninski [Pan08] studies in his lower bound. As usual, we also identify P_d^ϵ with the diagonal density matrix having these entries; i.e.,

$$\mathsf{P}_d^\epsilon = \operatorname{diag}\left(\frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \dots, \frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}\right).$$

Note that $d_{\mathrm{TV}}^{\mathrm{sym}}(\mathsf{P}_d^\epsilon, \mathsf{Unif}_d) = \epsilon$. We also remark that when $\epsilon = \frac{1}{2}$, the distribution P_d^ϵ is the uniform distribution on $\frac{d}{2}$ elements (the odd-numbered ones). As in [Pan08], it proves to be most convenient to study the chi-squared distance between $\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}$ and SW^n_d; our main theorem is the following:

Theorem 6.2.1. $d_{\chi^2}(\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}, \mathrm{SW}^n_d) \leq \exp((4n\epsilon^2/d)^2) - 1.$

Since this distance is small unless $n = \Omega(d/\epsilon^2)$, our lower bound is complete. More precisely:

Corollary 6.2.2. *For even d, testing whether a d-dimensional mixed state ρ has the the property of being the maximally mixed requires $n \geq .15d/\epsilon^2$ copies.*

Proof. In light of Theorem 2.6.3 we know that any ϵ-tester may as well make its testing decision based on a draw $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$. Since $d_{\mathrm{TV}}^{\mathrm{sym}}(\mathsf{P}_d^\epsilon, \mathsf{Unif}_d) = \epsilon$, the tester must be able to distinguish a draw from $\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}$ and a draw from SW^n_d with probability advantage $1/3$; this is possible if and only if $d_{\mathrm{TV}}(\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}, \mathrm{SW}^n_d) \geq 1/3$. But

$$d_{\mathrm{TV}}(\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}, \mathrm{SW}^n_d) \leq \frac{1}{2}\sqrt{d_{\chi^2}(\mathrm{SW}^n_{\mathsf{P}_d^\epsilon}, \mathrm{SW}^n_d)} \leq \frac{1}{2}\sqrt{\exp((4n\epsilon^2/d)^2) - 1} < 1/3.$$

if $n < .15d/\epsilon^2$. □

We remark that by taking $\epsilon = \frac{1}{2}$ we exactly recover the lower bound from Theorem 1.4.20 due to Childs et al. [CHW07].

There are two major steps in the proof of Theorem 6.2.1. The first major step will be proving the following formula:

Theorem 6.2.3. *Let $x \in \mathbb{R}^d$ satisfy $x_1 + \cdots + x_d = 0$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[\left(\frac{s_{\boldsymbol{\lambda}}(1+x_1,\dots,1+x_d)}{s_{\boldsymbol{\lambda}}(1,\dots,1)}-1\right)^2\right] = \sum_{\substack{\mu\in\mathrm{Par}\\0<\ell(\mu)\le d}} \frac{s_\mu(x)^2}{d^{\uparrow\mu}\cdot d^{|\mu|}}\cdot n^{\downarrow|\mu|}.$$

(The sum has only finitely many terms since $n^{\downarrow|\mu|} = 0$ when $|\mu| > n$.)

Once the above theorem is established, the following consequence is essentially immediate:

Corollary 6.2.4. *Let $x \in \mathbb{R}^d$ satisfy $x_1 + \cdots + x_d = 0$ and $x_i \ge -1$ for all i. We write $\mathcal{Q}_x$ for the probability distribution on $[d]$ in which i has probability $\frac{1+x_i}{d}$. Then*

$$d_{\chi^2}(\mathrm{SW}_{\mathcal{Q}_x}^n, \mathrm{SW}_d^n) = \sum_{\substack{\mu\in\mathrm{Par}\\0<\ell(\mu)\le d}} \frac{s_\mu(x)^2}{d^{\uparrow\mu}\cdot d^{|\mu|}}\cdot n^{\downarrow|\mu|}.$$

Proof. By definition, $d_{\chi^2}(\mathrm{SW}_{\mathcal{Q}_x}^n, \mathrm{SW}_d^n)$ is equal to

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[\left(\frac{\mathrm{SW}_{\mathcal{Q}_x}^n(\boldsymbol{\lambda})}{\mathrm{SW}_d^n(\boldsymbol{\lambda})}-1\right)^2\right] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_d^n}\left[\left(\frac{s_{\boldsymbol{\lambda}}(\frac{1+x_1}{d},\dots,\frac{1+x_d}{d})\dim(\boldsymbol{\lambda})}{s_{\boldsymbol{\lambda}}(\frac{1}{d},\dots,\frac{1}{d})\dim(\boldsymbol{\lambda})}-1\right)^2\right],$$

where we used the definition of the Schur Weyl distribution. In turn, this equals the quantity on the left in Theorem 6.2.3 after canceling the common factor of $d^{-|\boldsymbol{\lambda}|}\dim\boldsymbol{\lambda}$ in the fraction (recall the homogeneity of the Schur polynomials). □

Let us sketch the intuition of the proof once Theorem 6.2.3 is established. We are ultimately interested in the case $x = 2\epsilon\cdot c$, where $\epsilon > 0$ is thought of as "small" and $c \in \mathbb{R}^d$ satisfies $c_1 + \cdots + c_d = 0$; specifically, $c = c_\pm := (+1,-1,+1,-1,\dots,+1,-1)$. For simplicity, let us write ϵ instead of 2ϵ. Since s_μ is homogeneous of degree $|\mu|$, this means $s_\mu(x)^2 = s_\mu(c)^2\epsilon^{2|\mu|}$. For the sake of intuition, let us consider the summands in Theorem 6.2.3 when $|\mu| = k$ is "small"; i.e., the coefficients on ϵ^{2k}. For $k = 1$ we have only $\mu = (1)$, and the associated summand actually drops out: this is because $s_{(1)}(x) = x_1 + \cdots + x_d = 0$. For $k \ge 2$, the term $n^{\downarrow|\mu|}$ is asymptotically n^k and the denominator $d^{|\mu|}\cdot d^{\uparrow\mu}$ is asymptotically d^{2k}. It remains to analyze $s_\mu(c_\pm)$. This is the second major step in the proof of Theorem 6.2.1: in Section 6.4 we establish an exact formula for it. Naively one might expect $|s_\mu(c_\pm)|$ to scale like d^k when $|\mu| = k$; however, as we will see it scales only like $d^{k/2}$ (and will in fact be 0 whenever k is odd). Thus the summands with $|\mu| = k$ small scale asymptotically as $n^k\cdot\frac{\epsilon^{2k}}{d^k}$, whence we get that $d_{\chi^2}(\mathrm{SW}_{\mathcal{Q}_{\epsilon\cdot c_\pm}}^n, \mathrm{SW}_d^n)$ is small if $n \ll \frac{d}{\epsilon^2}$.

6.3 Proof of Theorem 6.2.3

To analyze the quantity in Theorem 6.2.3 we will require the so-called Binomial Formula. (It generalizes the "usual" Binomial Formula, viz. $(1+x)^\ell = \sum_{m\ge0} x^m\ell^{\downarrow m}/m!$, in the case $d = 1$.)

Theorem 6.3.1. *The following polynomial identity holds:*

$$\frac{s_\lambda(1+x_1,\dots,1+x_d)}{s_\lambda(1,\dots,1)} = \sum_{\substack{\mu\in\mathrm{Par}\\ \ell(\mu)\le d}} \frac{s_\mu(x)}{d^{\uparrow\mu}}\cdot s^*_\mu(\lambda).$$

(The sum is actually finite since we may include the restriction $\mu\subseteq\lambda$ due to the factor $s^*_\mu(\lambda)$.)

In this form with the shifted Schur polynomials, the result appears in Okounkov and Olshanski's work [OO98b, Theorem 5.1] (see also [OO98a]). In a form involving factorial Schur polynomials it dates back to Lascoux [Las78]; see [Mac95, Example I.3.10].

The $\mu=\emptyset$ summand in Theorem 6.3.1 is always equal to 1; it follows that the quantity on the left of Theorem 6.2.3 is

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[\left(\sum_{0<\ell(\mu)\le d}\frac{s_\mu(x)}{d^{\uparrow\mu}}\cdot s^*_\mu(\boldsymbol{\lambda})\right)^2\right] = \sum_{0<\ell(\mu),\ell(\nu)\le d}\frac{s_\mu(x)s_\nu(x)}{d^{\uparrow\mu}d^{\uparrow\nu}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right].$$

Therefore proving Theorem 6.2.3 reduces to proving

$$x_1+\cdots+x_d=0 \implies \sum_{0<\ell(\mu),\ell(\nu)\le d}\frac{s_\mu(x)s_\nu(x)}{d^{\uparrow\mu}d^{\uparrow\nu}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right] = \sum_{0<\ell(\mu)\le d}\frac{s_\mu(x)^2}{d^{\uparrow\mu}\cdot d^{|\mu|}}\cdot n^{\downarrow|\mu|}. \tag{6.5}$$

This is the main difficult step of the proof; the surprising aspect here is that we only get a contribution on the order of n^k from the terms with $|\mu|=k$, whereas naively one would expect n^{2k}. Showing that the $n^{k+1},n^{k+2},\dots,n^{2k}$ contributions "drop out" is the essence of the proof.

In aid of proving (6.5), it's tempting to guess that $\mathbf{E}[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})] = 1_{\{\mu=\nu\}}\cdot\frac{d^{\uparrow\mu}}{d^{|\mu|}}\cdot n^{\downarrow|\mu|}$; however such a statement is false. Instead, what *is* true is the following:

Theorem 6.3.2. *Let $x\in\mathbb{R}^d$ satisfy $x_1+\cdots+x_d=0$ and let $\mu\in\mathrm{Par}$ satisfy $|\mu|=r_1$ and $0<\ell(\mu)\le d$. Assume $r_2\ge r_1$. Then*

$$\sum_{\substack{|\nu|=r_2\\ \ell(\nu)\le d}}\frac{s_\nu(x)}{d^{\uparrow\nu}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right] = 1_{\{r_2=r_1\}}\cdot\frac{s_\mu(x)}{d^{|\mu|}}\cdot n^{\downarrow|\mu|}.$$

To deduce (6.5) from Theorem 6.3.2, simply write

$$\sum_{0<\ell(\mu),\ell(\nu)\le d}\frac{s_\mu(x)s_\nu(x)}{d^{\uparrow\mu}d^{\uparrow\nu}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right] = \sum_{r_1,r_2>0}\sum_{\substack{|\mu|=r_1\\ \ell(\mu)\le d}}\sum_{\substack{|\nu|=r_2\\ \ell(\nu)\le d}}\frac{s_\mu(x)s_\nu(x)}{d^{\uparrow\mu}d^{\uparrow\nu}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right].$$

Then use Theorem 6.3.2 when $r_2\ge r_1$ and use it with the roles of μ and ν reversed when $r_2<r_1$.

As for the proof of Theorem 6.3.2 itself, the first step is to compute the expected product of the shifted Schur polynomials. One possible approach for this might be to use the Littlewood–Richardson rule for factorial Schur functions (see [MS99, Proposition 4.2] or [Mol09, Corollary 3.3]) to write $s^*_\mu s^*_\nu$ as a linear combination of s^*_τ polynomials. Unfortunately, these Littlewood–Richardson coefficients seem somewhat difficult to work with. Instead, we will expand the shifted Schur polynomials in terms of the central characters and then multiply them via the known structure constants. We do this in the below lemma, carried out for a generic Schur-Weyl distribution. In this lemma, $\mathfrak{S}(R)$ denotes the symmetric group acting on the finite set R.

Lemma 6.3.3. *Let $q = (q_1, \dots, q_d)$ be a probability distribution on $[d]$ and let $\mu \vdash r_1$, $\nu \vdash r_2$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n_q}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right] = \sum_{t=r_1\vee r_2}^{r_1+r_2} C^t_{r_1r_2}\cdot n^{\downarrow t}\cdot \mathop{\mathbf{E}}_{\substack{\boldsymbol{w}_1\sim\mathfrak{S}(R_1)\\ \boldsymbol{w}_2\sim\mathfrak{S}(R_2)}}\left[\chi_\mu(\boldsymbol{w}_1)\chi_\nu(\boldsymbol{w}_2)p_{\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2}(q)\right].$$

Here, for each choice of t, we let R_1, R_2 denote (arbitrary but fixed) subsets of $[t]$ having cardinality r_1, r_2, respectively, with $R_1 \cup R_2 = [t]$. (E.g., $R_1 = \{1, \dots, r_1\}$, $R_2 = \{t - r_2 + 1, \dots, t\}$.) Also, $\overline{\boldsymbol{w}}_1$ denotes the extension of $\boldsymbol{w}_1$ to $\mathfrak{S}(t)$ formed by letting $\overline{\boldsymbol{w}}_1$ fix each element of $[t] \setminus R_1$; similarly for $\overline{\boldsymbol{w}}_2$.

Proof. Recall the notation $\rho(w)$ from Definition 2.3.1 used denote the cycle type of a permutation w. In this proof, we also use the following notation: We write $\boldsymbol{\rho} \sim \mathfrak{S}(r)$ to denote that $\boldsymbol{\rho}$ is a random partition of r formed by first choosing $\boldsymbol{w} \sim \mathfrak{S}(r)$ uniformly and then taking $\boldsymbol{\rho} = \rho(\boldsymbol{w})$.

Using Theorem 3.8.2 for the first equality below, and Corollary 3.8.8 for the third equality, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n_q}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right]\\
&= \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n_q}\left[\mathop{\mathbf{E}}_{\boldsymbol{\rho}_1\sim\mathfrak{S}(r_1)}[\chi_\mu(\boldsymbol{\rho}_1)\cdot p^\sharp_{\boldsymbol{\rho}_1}(\boldsymbol{\lambda})]\cdot \mathop{\mathbf{E}}_{\boldsymbol{\rho}_2\sim\mathfrak{S}(r_2)}[\chi_\nu(\boldsymbol{\rho}_2)\cdot p^\sharp_{\boldsymbol{\rho}_2}(\boldsymbol{\lambda})]\right]\\
&= \mathop{\mathbf{E}}_{\substack{\boldsymbol{\rho}_1\sim\mathfrak{S}(r_1)\\ \boldsymbol{\rho}_2\sim\mathfrak{S}(r_2)}}\left[\chi_\mu(\boldsymbol{\rho}_1)\chi_\nu(\boldsymbol{\rho}_2)\cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n_q}\left[p^\sharp_{\boldsymbol{\rho}_1}(\boldsymbol{\lambda})\cdot p^\sharp_{\boldsymbol{\rho}_2}(\boldsymbol{\lambda})\right]\right]\\
&= \mathop{\mathbf{E}}_{\substack{\boldsymbol{\rho}_1\sim\mathfrak{S}(r_1)\\ \boldsymbol{\rho}_2\sim\mathfrak{S}(r_2)}}\left[\chi_\mu(\boldsymbol{\rho}_1)\chi_\nu(\boldsymbol{\rho}_2)\cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n_q}\left[\sum_{t=r_1\vee r_2}^{r_1+r_2}\sum_{\tau\vdash t} C^t_{r_1r_2}\cdot \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{w}_1\sim\mathfrak{S}(R_1)|\boldsymbol{\rho}_1\\ \boldsymbol{w}_2\sim\mathfrak{S}(R_2)|\boldsymbol{\rho}_2}}[\rho(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2)=\tau]\cdot p^\sharp_\tau(\boldsymbol{\lambda})\right]\right],
\end{aligned}$$

where here $\boldsymbol{w}_i$ is chosen to be a uniformly random permutation on R_i (as in the lemma's statement), *conditioned on* having cycle type $\boldsymbol{\rho}_i$. By Proposition 3.8.3 the above equals

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\substack{\boldsymbol{\rho}_1\sim\mathfrak{S}(r_1)\\ \boldsymbol{\rho}_2\sim\mathfrak{S}(r_2)}}\left[\chi_\mu(\boldsymbol{\rho}_1)\chi_\nu(\boldsymbol{\rho}_2)\cdot \sum_{t=r_1\vee r_2}^{r_1+r_2}\sum_{\tau\vdash t} C^t_{r_1r_2}\cdot \mathop{\mathbf{Pr}}_{\substack{\boldsymbol{w}_1\sim\mathfrak{S}(R_1)|\boldsymbol{\rho}_1\\ \boldsymbol{w}_2\sim\mathfrak{S}(R_2)|\boldsymbol{\rho}_2}}[\rho(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2)=\tau]\cdot n^{\downarrow t}\cdot p_\tau(q)\right]\\
&= \sum_{t=r_1\vee r_2}^{r_1+r_2} C^t_{r_1r_2}\cdot n^{\downarrow t}\cdot \mathop{\mathbf{E}}_{\substack{\boldsymbol{\rho}_1\sim\mathfrak{S}(r_1),\ \boldsymbol{\rho}_2\sim\mathfrak{S}(r_2)\\ \boldsymbol{w}_1\sim\mathfrak{S}(R_1)|\boldsymbol{\rho}_1\\ \boldsymbol{w}_2\sim\mathfrak{S}(R_2)|\boldsymbol{\rho}_2}}\left[\chi_\mu(\boldsymbol{\rho}_1)\chi_\nu(\boldsymbol{\rho}_2)\cdot \sum_{\tau\vdash t} 1_{\{\rho(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2)=\tau\}}\cdot p_\tau(q)\right]
\end{aligned}$$

The summation on the inside here simply equals $p_{\rho(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2)}(q)$; we may also replace $\chi_\mu(\boldsymbol{\rho}_1)$ with $\chi_\mu(\boldsymbol{w}_1)$, and similarly for $\chi_\nu(\boldsymbol{\rho}_2)$. Thus to complete the proof it remains to show that $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ have the same distribution as in the statement of the lemma. But this is clear: if we first pick a random permutation of r_i symbols, then take its cycle type, then set $\boldsymbol{w}_i$ to be a random permutation of r_i symbols of this cycle type, this is the same as simply taking $\boldsymbol{w}_i$ to be a uniformly random permutation of r_i symbols. □

We will also require the following Fourier-theoretic lemma:

Lemma 6.3.4. *For $u \in \mathfrak{S}(r)$, $\nu \vdash r$, and $d \in \mathbb{Z}^+$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\mathfrak{S}(r)}[\chi_\nu(\boldsymbol{w}) \cdot d^{\ell(u\boldsymbol{w})}] = \frac{\chi_\nu(u)d^{\uparrow\nu}}{r!}.$$

Proof. Define the class function e on $\mathfrak{S}(r)$ by

$$e(v) = p_v(\underbrace{1,\dots,1}_{d\text{ entries}}) = d^{\ell(v)}.$$

Since $\chi_\nu(\boldsymbol{w}) = \chi_\nu(\boldsymbol{w}^{-1})$ because χ_ν is a class function, the quantity on the left in the proposition's statement is

$$\mathop{\mathbf{E}}_{\boldsymbol{w}\sim\mathfrak{S}(r)}[\chi_\nu(\boldsymbol{w}^{-1}) \cdot d^{\ell(u\boldsymbol{w})}] = \mathop{\mathbf{E}}_{\boldsymbol{v}\sim\mathfrak{S}(r)}[\chi_\nu(\boldsymbol{v}^{-1}u) \cdot d^{\ell(\boldsymbol{v})}] = (e * \chi_\nu)(u) = \sum_{\mu\vdash r} \widetilde{e * \chi_\nu}(\mu)\chi_\mu(u)$$

$$= \sum_{\mu\vdash r} \tfrac{1}{\dim\mu}\widetilde{e}(\mu)\widetilde{\chi_\nu}(\mu)\chi_\mu(u) = \tfrac{1}{\dim\nu}\widetilde{e}(\nu)\chi_\nu(u) = \tfrac{1}{\dim\nu}s_\nu(1,\dots,1)\chi_\nu(u) = \frac{\chi_\nu(u)d^{\uparrow\nu}}{r!},$$

the last equality being Definition 2.2.13 □

We can now complete the proof of Theorem 6.3.2 (and therefore also Theorem 6.2.3):

Proof of Theorem 6.3.2. We will use Lemma 6.3.3 in the case of SW^n_d, i.e., $q = (\frac{1}{d}, \dots, \frac{1}{d})$; in this case, for $\tau \vdash t$ we have $p_\tau(q) = d^{\ell(\tau)-t}$. We thereby obtain

$$\sum_{\substack{|\nu|=r_2\\ \ell(\nu)\leq d}} \frac{s_\nu(x)}{d^{\uparrow\nu}} \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_d}\left[s^*_\mu(\boldsymbol{\lambda})s^*_\nu(\boldsymbol{\lambda})\right]$$

$$= \sum_{t=r_2}^{r_1+r_2} C^t_{r_1r_2} \cdot \frac{n^{\downarrow t}}{d^t} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1\sim\mathfrak{S}(R_1)}\Bigg[\chi_\mu(\boldsymbol{w}_1) \cdot \sum_{\substack{|\nu|=r_2\\ \ell(\nu)\leq d}} \frac{s_\nu(x)}{d^{\uparrow\nu}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_2\sim\mathfrak{S}(R_2)}\left[\chi_\nu(\boldsymbol{w}_2)d^{\ell(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2)}\right]\Bigg]. \tag{6.6}$$

(Here we are using the convention $\ell(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2) = \ell(\rho(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2))$.) We now would like to analyze the number of cycles of $\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2$ within $\mathfrak{S}(t)$. In $\overline{\boldsymbol{w}}_1$'s cycle decomposition, there are some cycles that act *only* on elements of $R_1 \setminus R_2$. Let's write $\ell^{\setminus}(\boldsymbol{w}_1)$ for the number of such cycles, and let's define $\overline{\boldsymbol{w}}_1^{\cap} \in \mathfrak{S}(t)$ to be $\overline{\boldsymbol{w}}_1$ with those cycles deleted. Thus

$$\ell(\overline{\boldsymbol{w}}_1\overline{\boldsymbol{w}}_2) = \ell^{\setminus}(\boldsymbol{w}_1) + \ell(\overline{\boldsymbol{w}}_1^{\cap} \cdot \overline{\boldsymbol{w}}_2).$$

Next, let $\boldsymbol{w}_1^\perp$ denote the permutation obtained by deleting every element of $R_1 \setminus R_2$ from the cycle decomposition of $\overline{\boldsymbol{w}}_1^\cap$. Though $\boldsymbol{w}_1^\perp$ acts only on $R_1 \cap R_2$, we will view it as an element of $\mathfrak{S}(R_2)$. Although we don't have $\boldsymbol{w}_1^\perp \cdot \boldsymbol{w}_2 = \overline{\boldsymbol{w}}_1^\cap \cdot \overline{\boldsymbol{w}}_2$, it's not too hard to see that

$$\ell(\overline{\boldsymbol{w}}_1^\cap \cdot \overline{\boldsymbol{w}}_2) = \ell(\boldsymbol{w}_1^\perp \cdot \boldsymbol{w}_2).$$

Thus we obtain

$$(6.6) = \sum_{t=r_2}^{r_1+r_2} C_{r_1 r_2}^t \cdot \frac{n^{\downarrow t}}{d^t} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(R_1)} \left[\chi_\mu(\boldsymbol{w}_1) d^{\ell^\setminus(\boldsymbol{w}_1)} \cdot \sum_{\substack{|\nu|=r_2 \\ \ell(\nu)\le d}} \frac{s_\nu(x)}{d^{\uparrow\nu}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_2 \sim \mathfrak{S}(R_2)} \left[\chi_\nu(\boldsymbol{w}_2) d^{\ell(\boldsymbol{w}_1^\perp \cdot \boldsymbol{w}_2)} \right] \right].$$

Applying Lemma 6.3.4, we deduce

$$(6.6) = \sum_{t=r_2}^{r_1+r_2} C_{r_1 r_2}^t \cdot \frac{n^{\downarrow t}}{d^t} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(R_1)} \left[\chi_\mu(\boldsymbol{w}_1) d^{\ell^\setminus(\boldsymbol{w}_1)} \cdot \frac{1}{r_2!} \sum_{\substack{|\nu|=r_2 \\ \ell(\nu)\le d}} s_\nu(x) \chi_\nu(\boldsymbol{w}_1^\perp) \right].$$

Notice that we may extend the summation over ν to include $\ell(\nu) > d$ as well: since x has d coordinates, $s_\nu(x) = 0$ anyway when $\ell(\nu) > d$ by Proposition 2.4.8. Having done this, we replace $s_\nu(x)$ with $\mathbf{E}_{\boldsymbol{v} \sim \mathfrak{S}(r_2)}[\chi_\nu(\boldsymbol{v}) p_{\boldsymbol{v}}(x)]$, obtaining

$$\begin{aligned}(6.6) &= \sum_{t=r_2}^{r_1+r_2} C_{r_1 r_2}^t \cdot \frac{n^{\downarrow t}}{d^t} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(R_1)} \left[\chi_\mu(\boldsymbol{w}_1) d^{\ell^\setminus(\boldsymbol{w}_1)} \cdot \frac{1}{r_2!} \sum_{|\nu|=r_2} \mathop{\mathbf{E}}_{\boldsymbol{v} \sim \mathfrak{S}(r_2)} [\chi_\nu(\boldsymbol{v}) \cdot p_{\boldsymbol{v}}(x)] \chi_\nu(\boldsymbol{w}_1^\perp) \right] \\ &= \sum_{t=r_2}^{r_1+r_2} \frac{C_{r_1 r_2}^t}{r_2!} \cdot \frac{n^{\downarrow t}}{d^t} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(R_1)} \left[\chi_\mu(\boldsymbol{w}_1) d^{\ell^\setminus(\boldsymbol{w}_1)} \cdot \mathop{\mathbf{E}}_{\boldsymbol{v} \sim \mathfrak{S}(r_2)} \left[p_{\boldsymbol{v}}(x) \cdot \sum_{|\nu|=r_2} \chi_\nu(\boldsymbol{v}) \chi_\nu(\boldsymbol{w}_1^\perp) \right] \right].\end{aligned}$$

We claim that the inner expectation is 0 in most cases. First, $p_{\boldsymbol{v}}(x)$ vanishes whenever $\boldsymbol{v}$ has a fixed point, since $p_1(x) = x_1 + \cdots + x_d = 0$ by assumption. Next, suppose that $\boldsymbol{v}$ has no fixed points. By the orthogonality relations of representation theory, the innermost sum vanishes unless $\boldsymbol{v}$ and $\boldsymbol{w}_1^\perp$ are conjugate. Since $\boldsymbol{w}_1^\perp \in \mathfrak{S}(R_2)$ acts only on $\mathbb{R}_1 \cap R_2$, it *must* have a fixed point (and therefore not be conjugate to $\boldsymbol{v}$) unless $R_2 \setminus R_1 = \emptyset$. Since $r_2 \ge r_1$, this can only happen if $|\mu| = r_1 = r_2 = t$. We conclude that the inner expectation can only be nonzero in case $|\mu| = r_1 = r_2 = t$. In this case we have $C_{r_1 r_2}^t = r_2!$ and $\ell^\setminus(\boldsymbol{w}_1) = 0$, whence

$$(6.6) = 1_{\{r_2=r_1\}} \cdot \frac{n^{\downarrow r_1}}{d^{r_1}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(r_1)} \left[\chi_\mu(\boldsymbol{w}_1) \cdot \mathop{\mathbf{E}}_{\boldsymbol{v} \sim \mathfrak{S}(r_1)} \left[p_{\boldsymbol{v}}(x) \cdot \sum_{|\nu|=r_1} \chi_\nu(\boldsymbol{v}) \chi_\nu(\boldsymbol{w}_1^\perp) \right] \right].$$

Once again, the summation is 0 if $\boldsymbol{v}$ and $\boldsymbol{w}_1$ are not conjugate; otherwise it equals $z_{\rho(\boldsymbol{w}_1)}$. Further, having chosen $\boldsymbol{w}_1$, the probability that $\boldsymbol{v}$ is conjugate to $\boldsymbol{w}_1$ is precisely $z_{\rho(\boldsymbol{w}_1)}^{-1}$. Thus these factors cancel and we obtain

$$(6.6) = 1_{\{r_2=r_1\}} \cdot \frac{n^{\downarrow r_1}}{d^{r_1}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{w}_1 \sim \mathfrak{S}(r_1)} [\chi_\mu(\boldsymbol{w}_1) \cdot p_{\boldsymbol{w}_1}(x)] = 1_{\{r_2=r_1\}} \cdot \frac{n^{\downarrow r_1}}{d^{r_1}} \cdot s_\mu(x),$$

completing the proof. □

6.4 A formula for $s_\mu(+1, -1, +1, -1, \dots)$

For this formula we will need to recall the notion of the 2-quotient of a partition. This definition essentially encodes the ways in which a partition can be tiled by dominoes.

Definition 6.4.1. Given a partition μ, a 2-*hook* in $[\mu]$ is a hook of length 2; i.e., a domino whose removal from $[\mu]$ results in a valid Young diagram.

Definition 6.4.2. A partition μ is said to be *balanced* (or to have an *empty* 2-*core*) if $[\mu]$ can be reduced to the empty diagram by successive removal of 2-hooks.

Definition 6.4.3. Given a partition μ we write $[\mu]_{\text{even}}$ (respectively, $[\mu]_{\text{odd}}$) for the set of boxes $\square \in [\mu]$ with even (respectively, odd) content $c(\square)$.

Remark 6.4.4. It's obvious from Definition 6.4.2 that if $\mu \vdash k$ is balanced then $|[\mu]_{\text{even}}| = k/2$. In fact, the converse also holds (this follows from, e.g., [JK81, Theorem 2.7.41]).

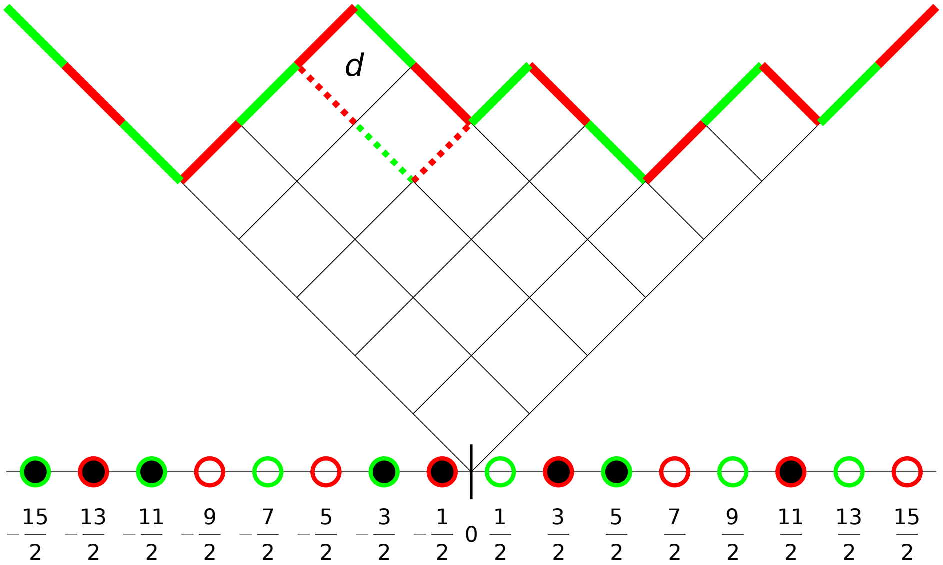

d
$-\frac{15}{2}$ $-\frac{13}{2}$ $-\frac{11}{2}$ $-\frac{9}{2}$ $-\frac{7}{2}$ $-\frac{5}{2}$ $-\frac{3}{2}$ $-\frac{1}{2}$ 0 $\frac{1}{2}$ $\frac{3}{2}$ $\frac{5}{2}$ $\frac{7}{2}$ $\frac{9}{2}$ $\frac{11}{2}$ $\frac{13}{2}$ $\frac{15}{2}$

Figure 6.1: The Russian and Maya diagrams for $\mu = (6, 4, 4, 3, 3) \vdash 20$. The segments and pebbles corresponding to the 2-quotient pair are colored green and red. The dashed lines outline a 2-hook that could be removed; d is the square in this 2-hook with even content (namely, -2).

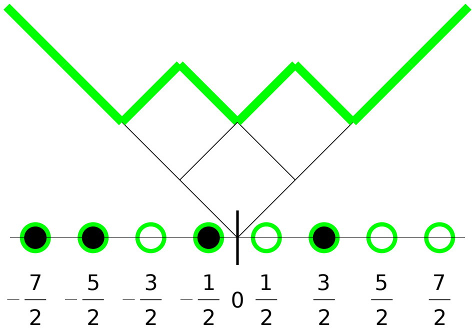

$-\frac{7}{2}$ $-\frac{5}{2}$ $-\frac{3}{2}$ $-\frac{1}{2}$ 0 $\frac{1}{2}$ $\frac{3}{2}$ $\frac{5}{2}$ $\frac{7}{2}$

Figure 6.2: The diagram for 2-quotient partition $\mu^{(0)} = (2, 1) \vdash 3$.

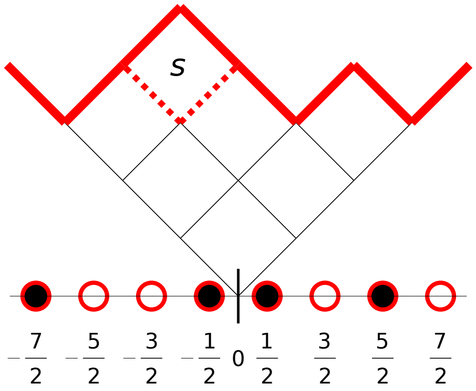


Figure 6.3: The diagram for 2-quotient partition $\mu^{(1)} = (3, 2, 2) \vdash 7$. The 1-hook square s (with content -1) is associated to the 2-hook in Figure 6.1 that contains square d.

Definition 6.4.5. Let μ be a partition. From the Maya diagram for $[\mu]$, form two new Maya diagrams by taking the two alternating sequences of pebbles. More precisely, for $b \in \{0, 1\}$, let $\mu^{(b)}$ denote the partition whose Maya diagram is formed by the pebbles at positions $2z + (-1)^b \frac{1}{2}$, $z \in \mathbb{Z}$. (See Figure 6.1, in which $b = 0$ is associated to green and $b = 1$ is associated to red.) The pair $(\mu^{(0)}, \mu^{(1)})$ is called the 2-*quotient* of μ. (See Figures 6.2, 6.3 respectively.)

Remark 6.4.6. Note that when the Maya diagrams for $\mu^{(0)}$, $\mu^{(1)}$ are formed, each of the two origin mark positions may need to be adjusted from the former origin mark position coming from μ's origin mark. It is a fact (see, e.g., [RZ12, Section 2.1]) that μ is balanced if and only if *neither* origin mark position must be adjusted.

Fact 6.4.7. *A 2-hook in $[\mu]$ naturally corresponds to a sequence of three pebbles in $[\mu]$'s Maya diagram of the form (white, $*$, black). (See the dashed domino containing the label d in Figure 6.1.) In turn, this corresponds to a "1-hook" in one of $\mu^{(0)}, \mu^{(1)}$; i.e., a square on the rim whose removal leaves a valid Young diagram (see the square labeled s in Figure 6.3). Removal of the 2-hook from $[\mu]$ corresponds to replacing the sequence (white, $*$, black) by (black, $*$, white). (One thinks of the "filled" black pebble as jumping two positions to the left, onto the "empty" white pebble.) In turn, this corresponds to removing the associated 1-hook from either $\mu^{(0)}$ or $\mu^{(1)}$.*

We will require the following lemma. It is likely to be known; however we were unable to find its statement in the literature. The analogous lemma for hook lengths is well known (see, e.g., [RZ12, Lemma 2.1.ii]).

Lemma 6.4.8. *Let $\mu \vdash k$ be a balanced partition with 2-quotient $(\mu^{(0)}, \mu^{(1)})$. Then the multiset $\{c(\square) : \square \in [\mu^{(0)}], \square \in [\mu^{(1)}]\}$ is equal to the multiset $\{\frac{1}{2}c(\square) : \square \in [\mu]_{\text{even}}\}$.*

Proof. The statement is proved by induction on the deconstruction of μ from 2-hooks, with the base case being $\mu = \emptyset$. We rely on the fact that since μ is balanced, the Maya diagrams of $\mu^{(0)}$ and $\mu^{(1)}$ can be seen alternating within the Maya diagram for μ, with all three origin markers "lining up" (see Remark 6.4.6). By way of induction, suppose we consider the removal of some 2-hook D from $[\mu]$. This corresponds (see Fact 6.4.7) to removing a 1-hook (square) s from $\mu^{(b)}$, for some $b \in \{0,1\}$. Exactly one of D's two squares is in $[\mu]_{\text{even}}$; call that square d. (See Figures 6.1, 6.3 for illustration.) By induction, it suffices to show that $\frac{1}{2}c(d) = c(s)$. But this is easily seen from the combination of the Russian and Maya diagrams, as the content of a square is simply the horizontal displacement of its center. □

We are now ready to establish a formula for $s_\mu(+1,-1,+1,-1,\dots)$.

Theorem 6.4.9. *Let $\mu \vdash k$ and let d be even. Then*

$$s_\mu(\underbrace{+1,-1,+1,-1,\dots}_{d \text{ entries}}) = \begin{cases} 0 & \text{if } \mu \text{ is not balanced,} \\ \chi_\mu(\underbrace{2,2,\dots,2}_{k/2 \text{ entries}}) \cdot \dfrac{1}{k!!} \cdot (d^{\uparrow[\mu]_{\text{even}}}) & \text{if } \mu \text{ is balanced.} \end{cases}$$

Proof. The first part of the proof relies on a formula from [RSW04, Theorem 4.3], specialized to the case of "t" $= 2$:

$$\begin{aligned} &s_\mu(\underbrace{+1,-1,+1,-1,\dots}_{d \text{ entries}}) \\ &\quad = \begin{cases} 0 & \text{if } \mu \text{ is not balanced,} \\ \operatorname{sgn}(\chi_\mu(\underbrace{2,2,\dots,2}_{k/2 \text{ entries}})) \cdot s_{\mu^{(0)}}(\underbrace{1,1,\dots,1}_{d/2 \text{ entries}}) \cdot s_{\mu^{(1)}}(\underbrace{1,1,\dots,1}_{d/2 \text{ entries}}) & \text{if } \mu \text{ is balanced,} \end{cases} \end{aligned}$$

where $(\mu^{(0)}, \mu^{(1)})$ is the 2-quotient of μ. Thus it suffices to show

$$s_{\mu^{(0)}}(1,1,\dots,1) \cdot s_{\mu^{(1)}}(1,1,\dots,1) = \frac{|\chi_\mu(2,2,\dots,2)| \cdot (d^{\uparrow[\mu]_{\text{even}}})}{(k/2)! \cdot 2^{k/2}} \tag{6.7}$$

assuming μ is balanced. Applying Definition 2.2.13, the left-hand side of (6.7) is

$$\frac{(\frac{d}{2}^{\uparrow\mu^{(0)}}) \cdot (\frac{d}{2}^{\uparrow\mu^{(1)}}) \cdot \dim \mu^{(0)} \cdot \dim \mu^{(1)}}{|\mu^{(0)}|! \cdot |\mu^{(1)}|!}.$$

Next, we appeal to [RZ12, formula (2.2)], which states

$$\chi_\mu(2,2,\dots,2) = \sigma_\mu \cdot \binom{|\mu|/2}{|\mu^{(0)}|, |\mu^{(1)}|} \cdot \dim \mu^{(0)} \cdot \dim \mu^{(1)},$$

where $\sigma_\mu \in \{\pm 1\}$ is a certain sign. Thus to verify (6.7) it remains to show

$$(\tfrac{d}{2}^{\uparrow\mu^{(0)}}) \cdot (\tfrac{d}{2}^{\uparrow\mu^{(1)}}) = \frac{d^{\uparrow[\mu]_{\text{even}}}}{2^{k/2}}. \tag{6.8}$$

But this follows immediately from Lemma 6.4.8. □

6.5 Wrapping up the lower bound

In this section we complete the proof of Theorem 6.2.1. We begin by applying Corollary 6.2.4 with $x = (+2\epsilon, -2\epsilon, +2\epsilon, -2\epsilon, \dots)$. Using Theorem 6.4.9 and the homogeneity of Schur polynomials, we obtain the following after a few manipulations:

Theorem 6.5.1. *For d even and $0 \leq \epsilon \leq \frac{1}{2}$,*

$$d_{\chi^2}(\mathrm{SW}^n_{\mathsf{P}^\epsilon_d}, \mathrm{SW}^n_d) = \sum_{k=2,4,6,\dots} n^{\downarrow k}(2\epsilon)^{2k} d^{-k} \cdot \Bigl(\frac{1}{k!!^2} \sum_{\substack{\mu \vdash k \text{ balanced} \\ 0<\ell(\mu)\leq d}} \chi_\mu(2,\dots,2)^2 \cdot \frac{d^{\uparrow[\mu]_{\text{even}}}}{d^{\uparrow[\mu]_{\text{odd}}}}\Bigr). \tag{6.9}$$

To estimate this quantity we will use the following very crude bound:

Proposition 6.5.2. *Let $d \in \mathbb{Z}^+$ and let $\mu \vdash k$ be balanced, with $0 < \ell(\mu) \leq d$. Then*

$$\frac{d^{\uparrow[\mu]_{\text{even}}}}{d^{\uparrow[\mu]_{\text{odd}}}} \leq 2^{k/2}. \tag{6.10}$$

Proof. Fix any domino-tiling for μ. Each of the $k/2$ dominoes contains one cell of even content c_e and one cell of odd content c_o, with $|c_e - c_o| = 1$. Thus each contributes a factor of $\frac{d+c_e}{d+c_o} \leq \frac{2}{1} = 2$ to $(d^{\uparrow[\mu]_{\text{even}}})/(d^{\uparrow[\mu]_{\text{odd}}})$. □

By character orthogonality relations we also have

$$\sum_{\substack{\mu \vdash k \text{ balanced} \\ 0<\ell(\mu)\leq d}} \chi_\mu(2,\dots,2)^2 \leq \sum_{\mu \vdash k} \chi_\mu(2,\dots,2)^2 = z_{(2,\dots,2)} = k!!. \tag{6.11}$$

Combining (6.10), (6.11), we get that the parenthesized expression in (6.9) is at most $2^{k/2}/k!! = 1/(k/2)!$. Using also $n^{\downarrow k} \leq n^k$, the right-hand side of (6.9) is thus bounded by

$$\sum_{k=2,4,6,\dots} n^k (2\epsilon)^{2k} d^{-k}/(k/2)! = \exp((4n\epsilon^2/d)^2) - 1,$$

completing the proof of Theorem 6.2.1.

We end by indicating how to obtain the testing lower bound in the case when $d \geq 3$ is odd. In this case we define P^ϵ_d to be $(\frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \dots, \frac{1+2\epsilon}{d}, \frac{1-2\epsilon}{d}, \frac{1}{d})$. This distribution has $d^{\text{sym}}_{\text{TV}}(\mathsf{P}^\epsilon_d, \mathsf{Unif}_d) = \frac{d-1}{d}\epsilon \geq \frac{2}{3}\epsilon$; since this differs from ϵ only by a constant factor, the lower bound of $\Omega(d/\epsilon^2)$ is not affected. Now Corollary 6.2.4 is applied with $x = (+2\epsilon, -2\epsilon, \dots, +2\epsilon, -2\epsilon, 0)$. By stability of the shifted Schur polynomials we have

$$s_\mu(+1, -1, \dots, +1, -1, 0) = s_\mu(+1, -1, \dots, +1, -1),$$

where there are $d-1$ entries in the latter. Now we get $\chi_\mu(2, 2, \dots, 2) \cdot \frac{1}{k!!} \cdot (d-1)^{\uparrow[\mu]_{\text{even}}}$ out of Theorem 6.4.9, and we can simply upper-bound $(d-1)$ by d and proceed with the remainder of the proof.

Chapter 7

Hardness of distinguishing uniform distributions

In this section, we prove Theorem 1.4.25, namely that $O(r^2/\Delta)$ copies are sufficient to distinguish between the cases when ρ's spectrum is uniform on either r or $r+\Delta$ eigenvalues ($1 \leq \Delta \leq r$), and that $\widetilde{\Omega}(r^2/\Delta)$ copies are necessary. To be more precise, our lower bound on the number of copies n will be

$$n \geq r^{2-O(1/\log^{.33} r)}/\Delta. \tag{7.1}$$

7.1 The upper bound

The proof of the upper bound is quite similar to that of Theorem 6.1.2 for the Mixedness Tester. We employ the following tester:

Definition 7.1.1 (Uniform Distribution Distinguisher)**.** Given $\rho^{\otimes n}$:

1. Sample $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$.

2. Accept if $p^\sharp_2(\boldsymbol{\lambda}) \leq e := n(n-1) \cdot \frac{1}{2}\left(\frac{1}{r} + \frac{1}{r+\Delta}\right)$. Reject otherwise.

As for the analysis, from Equations (6.2) and (6.3):

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_m}\left[p^\sharp_2(\boldsymbol{\lambda})\right] = \frac{n(n-1)}{m}, \quad \text{and} \quad \mathop{\mathbf{Var}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_m}\left[p^\sharp_2(\boldsymbol{\lambda})\right] \leq 2n(n-1).$$

We see that the variance is the same whether $m = r$ or $m = r + \Delta$; only the expectation is different, and the tester's acceptance cutoff e is precisely the midway point between the two expectations. If $m = r$, then Chebyshev's inequality implies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_m}\left[p^\sharp_2(\boldsymbol{\lambda}) \geq e\right] \leq \frac{8r^2(r+\Delta)^2}{n(n-1)\Delta^2} \leq \frac{32r^4}{(n-1)^2\Delta^2},$$

and we have the same upper bound by Chebyshev for $\mathbf{Pr}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_m}\left[p^\sharp_2(\boldsymbol{\lambda}) \leq e\right]$ when $m = r+\Delta$. This upper bound is at most $1/3$ provided $n \geq 4\sqrt{6} \cdot \frac{r^2}{\Delta} + 1$, completing the proof of the upper bound in Theorem 1.4.25.

The end of Section 8.1 gives a different $O(r^2)$-copy tester (the "Rank Tester") for the r-versus-$(r+1)$ case. In this case it's superior to the Uniform Distribution Distinguisher in that it has one-sided error (i.e., it never rejects in the rank-r case).

7.2 The lower bound

The bulk our work for the lower bound will be devoted to the case of $\Delta = 1$. The extension to larger Δ is very tedious and will be dealt with in Section 7.3. So let $r \in \mathbb{Z}^+$ be a parameter which we think of as tending to infinity, and for brevity let $r_+ = r+1$. Our task is to show that the distributions SW^n_r and $\mathrm{SW}^n_{r_+}$ are very close in total variation distance unless $n \geq \widetilde{\Omega}(r^2)$. For notational convenience we will write

$$n = \frac{r^2}{\omega^2}$$

and seek to show that SW^n_r and $\mathrm{SW}^n_{r_+}$ are close once ω is sufficiently large as a function of r. Ultimately we will select $\omega = \exp(\Theta(\log^{.67} r))$. For now, though, let's keep ω general, subjecting it only to the following assumption:

$$200 \leq \omega \leq \sqrt{r}. \tag{7.2}$$

7.2.1 Initial approximations

It proves more convenient to study the Kullback–Leibler divergence between SW^n_r and $\mathrm{SW}^n_{r_+}$:

$$\begin{aligned} d_{\mathrm{KL}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r_+}) &= \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_r}\left[\ln\left(\frac{\mathrm{SW}^n_r[\boldsymbol{\lambda}]}{\mathrm{SW}^n_{r_+}[\boldsymbol{\lambda}]}\right)\right] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_r}\left[\ln\left(\frac{r_+^n}{r^n} \cdot \frac{r^{\uparrow \boldsymbol{\lambda}}}{r_+^{\uparrow \boldsymbol{\lambda}}}\right)\right] \\ &= n \ln\left(\frac{r_+}{r}\right) + \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_r}\left[\ln\left(\frac{\prod_{\square \in [\boldsymbol{\lambda}]}(r + c(\square))}{\prod_{\square \in [\boldsymbol{\lambda}]}(r_+ + c(\square))}\right)\right], \end{aligned} \tag{7.3}$$

where the second equality used Proposition 3.3.3. (We remark that the logarithms above are always finite since $\mathrm{supp}(\mathrm{SW}^n_r) \subseteq \mathrm{supp}(\mathrm{SW}^n_{r_+})$.)

Recalling that $r_+ = r+1$, it is very easy to verify (cf. [Mac95, Exercise I.1.11], [CGS04, Section 2.5]) that the large fraction inside the inner logarithm of (7.3) is equal to

$$\prod_{i=1}^{\ell(\boldsymbol{\lambda})} \frac{r-(i-1)}{r-(i-1-\boldsymbol{\lambda}_i)} = \Phi(-(r+\tfrac{1}{2}); \boldsymbol{\lambda}),$$

where Φ denotes a generating function for the modified Frobenius coordinates, defined in [IO02] and similar to the "Frobenius function" from [Las08, CSST10]. Proposition 1.2 in [IO02] observes that

$$\Phi(z; \lambda) = \prod_i \frac{z + b_i^*}{z - a_i^*},$$

where the a_i^*'s and b_i^*'s are the modified Frobenius coordinates of λ; as a consequence, Proposition 1.4 in [IO02] states that

$$\ln \Phi(z;\lambda) = \sum_{k=1}^{\infty} \frac{p_k^*(\lambda)}{k} z^{-k}. \tag{7.4}$$

However we cannot immediately take $z = -(r+\frac{1}{2})$ and conclude

$$(7.3) \overset{?}{=} n \ln\left(1+\frac{1}{r}\right) + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[\sum_{k=1}^{\infty} \frac{(-1)^k p_k^*(\boldsymbol{\lambda})}{k(r+\frac{1}{2})^k}\right] \tag{7.5}$$

because (7.4) is merely a formal identity of generating functions and does not hold for all real z. More specifically, it's necessary that the Taylor series for $\ln(1+b_i/z)$ and $\ln(1-a_i/z)$ converge, which happens provided $|b_i/(r+\frac{1}{2})|, |a_i/(r+\frac{1}{2})| \leq 1$. These conditions are equivalent to $\ell(\boldsymbol{\lambda}) = \boldsymbol{\lambda}_1' \leq r+1$ and $\boldsymbol{\lambda}_1 \leq r+1$. The first condition is automatic, since $\boldsymbol{\lambda} \sim \mathrm{SW}_r^n$. The second condition does not always hold; however, we will show (see Lemma 7.2.2 below) that it holds with overwhelming probability when $n \ll r^2$. Indeed the "central limit theorems" for the Schur-Weyl distributions suggest that both $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_1'$ will almost always be $O(\sqrt{n}) = O(\frac{r}{\omega})$. Let us therefore make a definition:

Definition 7.2.1. We say that $\lambda \vdash n$ is *usual* if $\lambda_1, \lambda_1' \leq \frac{10}{\omega} r$. Since we are assuming $\omega \geq 200$, usual λ's satisfy $\lambda_1, \lambda_1' \leq \frac{1}{20} r \leq r+1$.

Thus when λ is usual we may apply (7.5). Since the quantity inside the expectation in (7.3) is clearly always negative, we may write

$$\begin{aligned} d_{\mathrm{KL}}(\mathrm{SW}_r^n, \mathrm{SW}_{r_+}^n) = (7.3) &\leq n\ln\left(1+\frac{1}{r}\right) + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[1_{\{\boldsymbol{\lambda}\text{ usual}\}} \cdot \ln\left(\frac{\prod_{\square\in[\boldsymbol{\lambda}]}(r+c(\square))}{\prod_{\square\in[\boldsymbol{\lambda}]}(r_+ + c(\square))}\right)\right] \\ &= n\ln\left(1+\frac{1}{r}\right) + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[1_{\{\boldsymbol{\lambda}\text{ usual}\}} \cdot \sum_{k=1}^{\infty} \frac{(-1)^k p_k^*(\boldsymbol{\lambda})}{k(r+\frac{1}{2})^k}\right] \\ &= n\ln\left(1+\frac{1}{r}\right) - \frac{1}{r+\frac{1}{2}} \cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[1_{\{\boldsymbol{\lambda}\text{ usual}\}} \cdot p_1^*(\boldsymbol{\lambda})\right] \qquad (7.6) \\ &\qquad + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[1_{\{\boldsymbol{\lambda}\text{ usual}\}} \cdot \sum_{k=2}^{\infty} \frac{(-1)^k p_k^*(\boldsymbol{\lambda})}{k(r+\frac{1}{2})^k}\right]. \qquad (7.7) \end{aligned}$$

Recall that $p_1^*(\lambda)$ is simply $|\lambda|$; thus the expectation in (7.6) is simply $n\,\mathbf{Pr}[\boldsymbol{\lambda}\text{ usual}]$. As Lemma 7.2.2 below shows, $\mathbf{Pr}[\boldsymbol{\lambda}\text{ usual}] = 1-\delta$ for $\delta \lll \frac{1}{60r^2}$. Thus:

$$(7.6) = n\left(\ln\left(1+\frac{1}{r}\right) - \frac{1}{r+\frac{1}{2}} + \frac{\delta}{r+\frac{1}{2}}\right) \leq n\left(\frac{1}{12r^3} + \frac{1/(60r^2)}{r+\frac{1}{2}}\right) \leq \frac{n}{10r^3} = \frac{1}{10\omega^2 r}. \tag{7.8}$$

Lemma 7.2.2. *Let* $\boldsymbol{\lambda} \sim \mathrm{SW}_r^n$. *Then* $\mathbf{Pr}[\boldsymbol{\lambda}\text{ unusual}] \leq 2^{-20r/\omega}$.

Proof. Write $B = \lceil \frac{10}{\omega} r \rceil$. By Proposition 3.7.12 and the fact that $B \leq r$,

$$\mathbf{Pr}[\boldsymbol{\lambda}_1 \geq B], \mathbf{Pr}[\boldsymbol{\lambda}'_1 \geq B] \leq \left(\frac{2e^2 n}{B^2}\right)^B \leq \left(\frac{2e^2}{100}\right)^{10r/\omega} \leq 2^{-1-20r/\omega}.$$

The lemma now follows from the union bound. □

Turning to (7.7), let's write

$$L^*_C(\lambda) \coloneqq \sum_{k=2}^{C} \frac{(-1)^k p^*_k(\lambda)}{k(r+\frac{1}{2})^k},$$

recalling that $L^*_\infty(\lambda)$ is definitely convergent if λ is usual. The infinite sum in (7.7) is inconvenient, as is the $+\frac{1}{2}$ in the denominator. We clean these issues up with the following lemma:

Lemma 7.2.3. *Assuming $\lambda \vdash n$ is usual, if*

$$C \geq \frac{3\log(10r)}{\log(\omega/10)},$$

it follows that

$$|L^*_\infty(\lambda) - L_C(\lambda)| \leq \frac{201}{\omega^3},$$

where $L_C(\lambda)$ denotes the same quantity as $L^*_C(\lambda)$ except with no $+\frac{1}{2}$ in the denominator.

Proof. For any $\lambda \vdash n$ (not necessarily usual), we have the crude bound $|p^*_k(\lambda)| \leq 2\sqrt{n}B^k$ whenever $\lambda_1, \lambda'_1 \leq B$. This is because each modified Frobenius coordinate a^*_i or b^*_i (of which there are at most $\sqrt{n}$ each) is at most B. For *usual* λ we may take $B = \frac{10}{\omega} r$. Thus we have

$$|L^*_\infty(\lambda) - L^*_C(\lambda)| \leq \sum_{k=C+1}^{\infty} \frac{|p^*_k(\lambda)|}{k(r+\frac{1}{2})^k} \leq \sum_{k=C+1}^{\infty} \frac{2\frac{r}{\omega}(10\frac{r}{\omega})^k}{kr^k} \leq 2r \sum_{k=C+1}^{\infty} \left(\frac{10}{\omega}\right)^k \leq 4r\left(\frac{10}{\omega}\right)^C \leq \frac{1}{250r^2},$$

where the last inequality used the assumption about C (and the second-to-last inequality used $\omega \geq 200$ in a crude way). Further,

$$|L^*_C(\lambda) - L_C(\lambda)| \leq \sum_{k=2}^{C} \frac{|p^*_k(\lambda)|}{k} \left(\frac{1}{r^k} - \frac{1}{(r+\frac{1}{2})^k}\right) \leq \sum_{k=2}^{C} \frac{2\frac{r}{\omega}(10\frac{r}{\omega})^k}{k}\left(\frac{k}{2r^{k+1}}\right) = \frac{1}{\omega}\sum_{k=2}^{C}\left(\frac{10}{\omega}\right)^k \leq \frac{200}{\omega^3}.$$

Finally, $\frac{200}{\omega^3} + \frac{1}{250r^2} \leq \frac{201}{\omega^3}$ by our assumption (7.2) that $\omega \leq \sqrt{r}$. □

Let us use this lemma in (7.7), and also apply (7.8) in (7.6). Assuming the lemma's hypotheses, we obtain

$$\begin{aligned} d_{\mathrm{KL}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r_+}) &\leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}\left[1_{\{\boldsymbol{\lambda}\text{ usual}\}} \cdot L_C(\boldsymbol{\lambda})\right] + \tfrac{1}{10\omega^2 r} + \tfrac{201}{\omega^3} \\ &\leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}[L_C(\boldsymbol{\lambda})] - \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}\left[1_{\{\boldsymbol{\lambda}\text{ unusual}\}} \cdot L_C(\boldsymbol{\lambda})\right] + \tfrac{202}{\omega^3}. \end{aligned}$$

We can use Cauchy–Schwarz to bound

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_r^n}\left[1_{\{\boldsymbol{\lambda}\text{ unusual}\}}\cdot L_C(\boldsymbol{\lambda})\right]\right| \leq \sqrt{\mathbf{E}[1^2_{\{\boldsymbol{\lambda}\text{ unusual}\}}]}\sqrt{\mathbf{E}[L_C(\boldsymbol{\lambda})^2]} \leq 2^{-10r/\omega}\sqrt{\mathbf{E}[L_C(\boldsymbol{\lambda})^2]}, \quad (7.9)$$

where the last inequality used Lemma 7.2.2. Finally, we can afford to use an extraordinarily crude bound on $\mathbf{E}[L_C(\boldsymbol{\lambda})^2]$:

$$\mathbf{E}[L_C(\boldsymbol{\lambda})^2] \leq C\sum_{k=2}^{C}\mathbf{E}[p_k^*(\boldsymbol{\lambda})^2] \leq C\sum_{k=2}^{C}(2\sqrt{n}n^k)^2 \leq n^{3C} \leq r^{6C},$$

where the second inequality used the crude bound on $|p_k^*(\lambda)|$ from the proof of Lemma 7.2.3. (In fact, in Section 7.3 we will actually show that this quantity is quite tiny.) If we now make the very weak assumption that $C \leq \frac{3r}{\omega\log r}$, we may conclude (7.9) $\leq 2^{-r/\omega} \ll \frac{1}{\omega^3}$.

Now we can summarize all of the preparatory work we have done so far:

Proposition 7.2.4. *Assuming* $\frac{3\log(10r)}{\log(\omega/10)} \leq C \leq \frac{3r}{\omega\log r}$, *for* $\boldsymbol{\lambda}\sim\mathrm{SW}_r^n$ *we have*

$$d_{\mathrm{KL}}(\mathrm{SW}_r^n, \mathrm{SW}_{r_+}^n) \leq \mathbf{E}\left[L_C(\boldsymbol{\lambda})\right] + \tfrac{203}{\omega^3},$$

where

$$L_C(\lambda) := \sum_{k=2}^{C}\frac{(-1)^k p_k^*(\lambda)}{kr^k}. \quad (7.10)$$

(It is straightforward to check using (7.2) that the range of values for C is nonempty.)

We now come to the main task: showing that $\mathbf{E}[L_C(\boldsymbol{\lambda})]$ is small.

7.2.2 Passing to the $p^\sharp$ polynomials

In this section and the following one, we will use the notation

$$\mathrm{fact}(\mu) = \prod_{w\geq 1} m_w(\mu)!$$

where, recall, $m_w(\mu)$ is the number of parts of μ equal to w.

The following proposition is essentially immediate from known formulas:

Proposition 7.2.5. *For any $k \in \mathbb{Z}^+$, we have the following identity on observables:*

$$p_k^* = \sum_{\mu\ :\ \mathrm{wt}(\mu)=k+1}\frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)}p_\mu^\sharp + \mathcal{O}_k,$$

where $\mathcal{O}_k$ is an observable with $\mathrm{wt}(\mathcal{O}_k) \leq k$. *More precisely,*

$$\mathcal{O}_k = \sum_{\mu\ :\ \mathrm{wt}(\mu)\leq k} c_{k,\mu}p_\mu^\sharp$$

for some rational coefficients $c_{k,\mu}$.

Proof. From [IO02, Corollary 2.8] we have

$$p_k^* = \frac{1}{k+1} \cdot \widetilde{p}_{k+1} + \Big\{\text{a linear combination of } \widetilde{p}_k, \dots, \widetilde{p}_2\Big\}.$$

From [IO02, Corollary 3.7] (cf. [Mél10b, Lemma 10.10]) we have

$$\widetilde{p}_{k+1} = \sum_{\mu\ :\ \mathrm{wt}(\mu)=k+1} \frac{(k+1)^{\downarrow \ell(\mu)}}{\mathrm{fact}(\mu)} \prod_{i\geq 1} (p_i^\sharp)^{m_i(\mu)}.$$

The result is now easily deduced from Proposition 3.8.10. □

Substituting the above result into (7.10) yields:

$$L_C(\lambda) = \sum_{k=2}^{C} \frac{(-1)^k}{kr^k} \cdot \sum_{\mathrm{wt}(\mu)=k+1} \frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)} p_\mu^\sharp(\lambda) + \sum_{k=2}^{C} \frac{(-1)^k \mathcal{O}_k(\lambda)}{kr^k}. \tag{7.11}$$

Taking the expectation over $\boldsymbol{\lambda} \sim \mathrm{SW}_r^n$, and using Corollary 3.8.4 to evaluate the expectation of $p_\mu^\sharp$, we obtain:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}_r^n}[L_C(\boldsymbol{\lambda})] = \sum_{k=2}^{C} \frac{(-1)^k}{kr^k} \cdot \sum_{\mathrm{wt}(\mu)=k+1} \frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)} n^{\downarrow|\mu|} r^{\ell(\mu)-|\mu|} \tag{7.12}$$

$$+ \sum_{k=2}^{C} \frac{(-1)^k\, \mathbf{E}_{\boldsymbol{\lambda}\sim\mathrm{SW}_r^n}[\mathcal{O}_k(\boldsymbol{\lambda})]}{kr^k}. \tag{7.13}$$

We will show in Lemmas 7.2.7, 7.2.8 below that the "error term" (7.13) is small assuming $n \ll r^2$. Thus we focus on (7.12).

7.2.3 Showing the "main term" is small: some intuition

Before diving into manipulations, let's take a high-level look at the contributions to (7.12) from $k = 2, 3, 4, 5, \dots$, focusing on the powers of n and r. First consider the case of $k = 2$. Here there is only one μ with $\mathrm{wt}(\mu) = 3$, namely $\mu = (2)$, which has $|\mu| = 2$ and $\ell(\mu) = 1$. Thus from $k = 2$ we pick up a factor on the order of $\frac{n^2}{r^3}$; more precisely, $\frac{n^{\downarrow 2}}{2r^3}$. This looks rather bad from the point of view of proving a quadratic lower bound for n: the term $\frac{n^{\downarrow 2}}{2r^3}$ is not small unless $n \ll r^{3/2}$. The main surprise in our proof is that this term will be exactly canceled by "lower-degree" contributions from larger k.

To see an example of this, consider the $k = 3$ contribution in (7.12). Here there are two μ's with $\mathrm{wt}(\mu) = 4$, namely $\mu = (3)$ and $\mu = (1, 1)$. The first gives a contribution on the order of $\frac{n^3}{r^5}$; more precisely, $-\frac{n^{\downarrow 3}}{3r^5}$. The second gives a contribution of $-\frac{n^{\downarrow 2}}{2r^3}$, thereby precisely canceling the $k = 2$ term. Thus we are left (so far) with $-\frac{n^{\downarrow 3}}{3r^5}$, which is small if $n \ll r^{5/3}$. This is still far from a quadratic bound, but it's better than the $r^{3/2}$ bound we were faced with previously.

In turn, the $-\frac{n^{\downarrow 3}}{3r^5}$ contribution will be canceled by a certain $k = 3$ term, namely $\frac{n^{\downarrow 3}}{r^5}$ from $\mu = (2, 1)$, together with a certain $k = 4$ term, namely $\frac{2n^{\downarrow 3}}{3r^5}$ from $\mu = (1, 1, 1)$. Indeed, if we sum up through $k = 6$, the total contribution is $-\frac{5n^{\downarrow 4}}{r^7} - \frac{n^{\downarrow 5}}{5r^9}$, which is small if $n \ll r^{7/4}$. This gets us still closer to a quadratic bound.

In fact, looking carefully at small partitions suggests that perfect cancelation is achieved if we group contributions according to $|\mu|$. This proves to be the case, as we will show below. In the end (7.12) does not precisely vanish because for $m > C/2$, not all μ's with $|\mu| = m$ appear in (7.12). However the "leftover contributions" are of the shape $r(\frac{n}{r^2})^k$ for $k > C/2$, a quantity we can ensure is small by taking ω and C large enough. (There is a tradeoff involved preventing us from taking C too large: our "error bound" (7.13) increases with C.)

7.2.4 Proof that the "main term" is small

Although (7.12) has a double summation, the summed quantity is simply counted exactly once for each μ with $3 \leq \mathrm{wt}(\mu) \leq C+1$. As suggested above, let us rearrange the summation according to $|\mu|$. We will use the notation $s = |\mu|-1$ and $h = \ell(\mu)-1$, so that $\mathrm{wt}(\mu) = s+h+2$ (i.e., $k = s+h+1$) and $\mathrm{wt}(\mu) \leq C+1 \iff h \leq C-1-s$:

$$\begin{aligned}(7.12) &= \sum_{s=1}^{C-1} \sum_{h=0}^{\min(s,C-1-s)} \sum_{\substack{\mu \vdash s+1 \\ \ell(\mu)=h+1}} \frac{(-1)^{s+h+1}}{(s+h+1)r^{s+h+1}} \frac{(s+h+1)^{\downarrow h}}{\mathrm{fact}(\mu)} n^{\downarrow(s+1)} r^{h-s} \\ &= \sum_{s=1}^{C-1} (-1)^{s+1} \cdot \frac{n^{\downarrow(s+1)}}{r^{2s+1}} \sum_{h=0}^{\min(s,C-1-s)} (-1)^h (s+h)^{\downarrow(h-1)} \sum_{\substack{\mu \vdash s+1 \\ \ell(\mu)=h+1}} \frac{1}{\mathrm{fact}(\mu)}.\end{aligned}$$

(We remark that we switched from $r+\frac{1}{2}$ to r in Lemma 7.2.3 so as to obtain nice cancelations on r here. We also recall the convention $m^{\downarrow(-1)} = \frac{1}{m+1}$.) It is not hard to show (see, e.g., [Mél10a, Lemma 11]) that

$$\sum_{\substack{\mu \vdash s+1 \\ \ell(\mu)=h+1}} \frac{1}{\mathrm{fact}(\mu)} = \frac{1}{(h+1)!}\binom{s}{h}.$$

Substituting this into the above, and also using $(s+h)^{\downarrow(h-1)} = \frac{(s+h)!}{(s+1)!}$, we get

$$\begin{aligned}(7.12) &= \sum_{s=1}^{C-1} (-1)^{s+1} \cdot \frac{n^{\downarrow(s+1)}}{r^{2s+1}} \sum_{h=0}^{\min(s,C-1-s)} (-1)^h \frac{(s+h)!}{(s+1)!(h+1)!}\binom{s}{h} \\ &= \sum_{s=1}^{C-1} \frac{(-1)^{s+1}}{s+1} \cdot \frac{n^{\downarrow(s+1)}}{r^{2s+1}} \sum_{h=0}^{\min(s,C-1-s)} \frac{(-1)^h}{h+1}\binom{s+h}{h}\binom{s}{h}.\end{aligned}$$

We now obtain the promised cancelation. Specifically, it is a known combinatorial identity (see, e.g., [GKP94, page 182]) that for all $s \in \mathbb{Z}^+$, the inner summation equals 0 provided h ranges all the way up to s. In other words, all contributions from $s \leq \frac{C-1}{2}$ vanish. For larger s, it's not hard to bound the inner "partial sum" crudely by, say, 9^s in absolute value. We therefore finally conclude:

$$|(7.12)| \leq \sum_{\frac{C}{2} \leq s \leq C-1} \frac{1}{s+1} \cdot \frac{n^{\downarrow(s+1)}}{r^{2s+1}} \cdot 9^s \leq \frac{n}{r} \sum_{s \geq \frac{C}{2}} \left(\frac{9n}{r^2}\right)^s = \frac{r}{\omega^2} \sum_{s \geq \frac{C}{2}} \left(\frac{9}{\omega^2}\right)^s \leq r\left(\frac{3}{\omega}\right)^C. \quad (7.14)$$

7.2.5 Bounding the "error term"

In this section we bound the "error term" (7.13), using the following lemma:

Lemma 7.2.6. *Suppose $n = \frac{r^2}{\omega^2}$. Then $0 \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}[p^\sharp_\mu(\boldsymbol{\lambda})] \leq r^{\mathrm{wt}(\mu)} \cdot (1/\omega^2)^{|\mu|}$.*

Proof. By Corollary 3.8.4, $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[p^\sharp_\mu(\boldsymbol{\lambda})\right] = n^{\downarrow|\mu|} r^{\ell(\mu)-|\mu|} \leq n^{|\mu|} r^{\mathrm{wt}(\mu)-2|\mu|} = r^{\mathrm{wt}(\mu)} \cdot (1/\omega^2)^{|\mu|}$. □

We will first use this lemma to bound (7.13) in a "soft" way, thinking of C as an absolute universal constant. This is enough to get a testing lower bound like $n \geq \Omega_\delta(r^{2-\delta})$ for every $\delta > 0$. Subsequently we do some technical work (which the uninterested reader may skip) to get a more explicit lower bound.

Lemma 7.2.7. *For all $C \geq 2$ there is a constant A_C such that $|(7.13)| \leq A_C \cdot \frac{1}{\omega^2}$.*

Proof. It suffices to show that for all $k \geq 2$ there is a constant A'_k such that

$$\frac{\mathbf{E}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}[\mathcal{O}_k(\boldsymbol{\lambda})]}{r^k} \leq A'_k \cdot \frac{1}{\omega^2}.$$

But recalling Proposition 7.2.5, the left-hand side is

$$\sum_{\mu\ :\ \mathrm{wt}(\mu)\leq k} c_{k,\mu} \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}_r^n}\left[\frac{p^\sharp_\mu(\boldsymbol{\lambda})}{r^k}\right],$$

and each expectation here is at most $(\frac{1}{\omega^2})^{|\mu|} \leq \frac{1}{\omega^2}$ by Lemma 7.2.6. This completes the proof. □

Lemma 7.2.8. *In fact, the constants A_C from Lemma 7.2.7 satisfy $A_C \leq 2^{O(C^2 \log C)}$.*

Proof. The proof involves some tedious analysis using the results of Section 3.8.1. It suffices to show that

$$\sum_{\mu:\mathrm{wt}(\mu)\leq k} |c_{k,\mu}| \leq 2^{O(k^2 \log k)}, \tag{7.15}$$

where, recall, the coefficients $c_{k,\mu}$ are defined by

$$p^*_k = \sum_{\mu\ :\ \mathrm{wt}(\mu)=k+1} \frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)} p^\sharp_\mu + \sum_{\mu\ :\ \mathrm{wt}(\mu)\leq k} c_{k,\mu} p^\sharp_\mu. \tag{7.16}$$

Let us return to the relationship between the p^* and $p^\sharp$ polynomials described in Section 3.8.1. Specifically, we'll need identities (3.13), (3.14), which express each $p^\sharp_k$ as a polynomial in $p^*_1, \dots, p^*_k$ via the power series $Q_k(t)$.

Given any polynomial R in indeterminates $p_1, \dots, p_k$ (either p^*'s or $p^\sharp$'s), write $\|R\|$ for the sum of the absolute values of R's coefficients. This is a submultiplicative norm. Observe from (3.14) that $\|Q_{k,m}\| \leq (k+1)^{m+1}$ (indeed, one may show it's precisely $\frac{(k+1)^{m+1}-k^{m+1}-1}{m+1}$). Thus the coefficient on t^s in $Q_k(t)^i$ is a polynomial in $p^*_1, \dots, p^*_k$ of norm at most $O(k)^s$. Hence the same is true for the coefficient on t^s in the expression $\sum_{i=0}^\infty \frac{(-1)^i}{i!} Q_k(t)^i$ from (3.14). As the coefficient on each power of t in $\prod_{j=1}^k (1-(j-\frac{1}{2})t)$ is a number of magnitude at most $(k-\frac{1}{2})^k$, we finally deduce that the relationship (3.7) can be expressed more quantitatively as

$$p_k^\sharp = p_k^* + R_k(p_1^*, \ldots, p_{k-1}^*), \quad \text{where } 1 + \|R_k\| \le \exp(bk\log k), \quad b \text{ a universal constant.}$$

We inductively invert this relationship as in (3.9), writing

$$p_k^* = S_k(p_1^\sharp, \ldots, p_k^\sharp), \quad \text{where } S_k = p_k^\sharp + \Big\{\text{polynomial in } p_1^\sharp, \ldots, p_{k-1}^\sharp \text{ of gradation at most } k-1\Big\}. \tag{7.17}$$

If we let $s(k) = \|S_k\|$, using convexity of $\exp(bk\log k)$ we get the inductive bound

$$s(k) \le \exp(bk\log k)s(k-1),$$

leading to the bound $s(k) \le \exp(O(k^2 \log k))$. This is nearly enough to complete the proof; the only issue is that in (7.17) we have a polynomial in the $p_j^\sharp$'s, whereas in (7.16) we have the products of $p_j^\sharp$'s expanded out into linear combinations of $p_\mu^\sharp$'s. However Lemma 7.2.9 below, which crudely bounds the magnitude of the structure constants for the $p^\sharp$'s, shows that each monomial $\prod_i p_{\lambda_i}^\sharp$ with gradation $|\lambda| = w$ can be replaced by a linear polynomial in $p_\mu^\sharp$'s (with $|\mu| \le w$) wherein each coefficient has magnitude at most $4^{w^2 \log w}$. Since w is always bounded by $k-1$ and since there are at most $2^{O(\sqrt{k})} \ll \exp(O(k^2 \log k))$ partitions μ with $|\mu| \le k$, we conclude that each of these linear polynomials has norm at most $\exp(O(k^2 \log k))$. Thus making these replacements in S_k only increases its norm by another multiplicative factor of $\exp(O(k^2 \log k))$. The proof is complete. □

Lemma 7.2.9. *Let $\lambda \vdash w$, and suppose $\displaystyle\prod_{i=1}^{\ell(\lambda)} p_{\lambda_i}^\sharp = \sum_\mu c_\mu p_\mu^\sharp$ within Λ^*. Then $|c_\mu| \le 4^{w^2 \log w}$ for all μ.*

Proof. The proof is an induction on $\ell = \ell(\lambda)$, the base case of $\ell = 1$ being trivial. Now for general λ with $\lambda_\ell = k$ we have

$$\prod_{i=1}^{\ell} p_{\lambda_i}^\sharp = \left(\prod_{i=1}^{\ell-1} p_{\lambda_i}^\sharp\right) \cdot p_k^\sharp = \left(\sum_\mu d_\mu p_\mu^\sharp\right) \cdot p_k^\sharp = \sum_\mu d_\mu \sum_\tau f_{\mu k}^\tau p_\tau^\sharp = \sum_\tau p_\tau^\sharp \sum_\mu d_\mu f_{\mu k}^\tau, \tag{7.18}$$

where each $|d_\mu|$ is at most $4^{(w-k)^2 \log(w-k)} \le 4^{(w-1)^2 \log(w)}$ by induction. By Corollary 3.8.8, the structure constants $f_{\mu k}^\tau$ satisfy $|f_{\mu k}^\tau| \le |C_{|\mu| k}^{|\tau|}| \le |\mu|!k! \le w^w$. Since the number of partitions of $(w-k)$ is trivially at most w^w, the coefficient on $p_\tau^\sharp$ in (7.18) has magnitude at most

$$\sum_\mu |d_\mu f_{\mu k}^\tau| \le w^{2w} \cdot \max_\mu |d_\mu| \le w^{2w} \cdot 4^{(w-1)^2 \log(w)} \le 4^{w^2 \log w},$$

completing the induction. □

7.2.6 Combining the bounds

Combining (7.14), and Lemmas 7.2.7, 7.2.8, we get that under the hypotheses of Proposition 7.2.4,

$$d_{\mathrm{KL}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r_+}) \leq r\left(\frac{3}{\omega}\right)^C + \exp(O(C^2 \log C)) \cdot \frac{1}{\omega^2} + \tfrac{203}{\omega^3} \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}. \tag{7.19}$$

In the above we used $r\left(\frac{3}{\omega}\right)^C \leq r\left(\frac{10}{\omega}\right)^C \leq r\left(\frac{1}{10r}\right)^3 \leq \frac{1}{\omega^3}$, the second inequality here following from the assumed lower bound on C. It's now evident that we should take C as small as we can; in particular, to equal $\lceil 3\frac{\log(10r)}{\log(\omega/10)}\rceil$. We conclude:

Theorem 7.2.10. *For any $200 \leq \omega \leq \sqrt{r}$, if $n = \frac{r^2}{\omega^2}$ then*

$$d_{\mathrm{KL}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r+1}) \leq \exp(O((\log r)/(\log \omega))^{2.01}) \cdot \omega^{-2}.$$

In particular, for $\omega = \exp(O(\log^{.67} r))$ and hence $n = r^{2-O(1/\log^{.33} r)}$, the above bound is $o_r(1)$.

By Pinsker's inequality we may conclude also that $d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r_+}) \leq o_r(1)$ unless $n = r^{2-O(1/\log^{.33} r)} = \widetilde{\Omega}(r^2)$. This completes the proof of the rank-r versus rank-$(r+1)$ testing lower bound; in particular, the more precise bound (7.1) in the case $\Delta = 1$.

7.3 Extension to $\Delta > 1$

Let us henceforth fix the parameter $C = \lceil 3\frac{\log(10r)}{\log(\omega/10)}\rceil$. To recap the preceding section we saw that

$$|\mathbf{E}[L_C(\boldsymbol{\lambda})]| \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}, \quad \text{and hence} \quad d_{\mathrm{KL}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r+1}) \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}. \tag{7.20}$$

If we apply Pinsker's inequality to the latter bound we obtain

$$d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r+1}) \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega}.$$

The key to getting a good lower bound when $\Delta > 1$ is to show that Pinsker's inequality is not sharp in our setting, and in fact the following is true:

Theorem 7.3.1. *For any $200 \leq \omega \leq \sqrt{r}$, if $n = \frac{r^2}{\omega^2}$ then*

$$d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r+1}) \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}.$$

From this we can obtain the testing bound (7.1) for rank-r versus rank-$(r+\Delta)$ (where $1 \leq \Delta \leq r$) simply by using the triangle inequality. Specifically, given $r \leq r' \leq 2r$ and n, define $\omega_{r'}$ by $n = \frac{(r')^2}{\omega^2_{r'}}$. Applying Theorem 7.3.1 for each r', we get

$$d_{\mathrm{TV}}(\mathrm{SW}^n_{r'}, \mathrm{SW}^n_{r'+1}) \leq \exp(O((\log r')/(\log \omega_{r'}))^{2.01}) \cdot \frac{1}{\omega^2_{r'}} \quad \text{for all } r \leq r' < 2r.$$

But $\omega_{r'}$ is within a factor of 2 of ω_r for all $r \leq r' \leq 2r$; thus by adjusting the constant in the $O(\cdot)$, the above holds with ω_r in place of $\omega_{r'}$. Applying the triangle inequality, we get

$$d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r'+\Delta}) \leq \exp(O((\log r)/(\log \omega_r))^{2.01}) \cdot \frac{1}{\omega_r^2} \cdot \Delta.$$

Again, taking $\omega_r = \exp(O(\log^{.67} r))$, we get

$$d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r'+\Delta}) \leq \frac{n}{r^{2-O(1/\log^{.33} r)}} \cdot \Delta,$$

and this completes the proof of the rank-testing lower bound (7.1).

Thus it remains to prove Theorem 7.3.1. The main result we need for this is the following:

Theorem 7.3.2. $\underset{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}{\mathbf{Var}}[L_C(\boldsymbol{\lambda})] \leq \exp(O(C^{2.01})) \cdot \dfrac{1}{\omega^4}.$

To prove Theorem 7.3.2 we will employ the following lemma:

Lemma 7.3.3. *Let μ be a partition with* $\mathrm{wt}(\mu) = k \geq 2$. *Then*

$$\underset{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}{\mathbf{Var}}[p^\sharp_\mu(\boldsymbol{\lambda})] \leq \exp(O(k^2 \log k)) \cdot r^{2k-2} \cdot (1/\omega^4).$$

Proof. If $|\mu| = 1$ then $p^\sharp_\mu(\lambda) = n$ which has variance 0. Thus we may assume $|\mu| \geq 2$ and hence $k \geq 3$. Using Proposition 3.8.10,

$$\mathbf{Var}[p^\sharp_\mu(\boldsymbol{\lambda})] = \mathbf{E}[p^\sharp_\mu(\boldsymbol{\lambda})^2] - \mathbf{E}[p^\sharp_\mu(\boldsymbol{\lambda})]^2 = \mathbf{E}[p^\sharp_{\mu\cup\mu}(\boldsymbol{\lambda})] - \mathbf{E}[p^\sharp_\mu(\boldsymbol{\lambda})]^2 + \mathbf{E}[q_\mu(\boldsymbol{\lambda})] \tag{7.21}$$

where $q_\mu(\lambda)$ is a certain linear combination of $p^\sharp_\nu$ polynomials, each of weight at most $2k-2$. Regarding the first two quantities here, Corollary 3.8.4 tells us that

$$\mathbf{E}[p^\sharp_{\mu\cup\mu}(\boldsymbol{\lambda})] - \mathbf{E}[p^\sharp_\mu(\boldsymbol{\lambda})]^2 = n^{\downarrow(2|\mu|)} r^{2\ell(\mu)-2|\mu|} - (n^{\downarrow|\mu|} r^{\ell(\mu)-|\mu|})^2 = r^{2\ell(\mu)-2|\mu|}(n^{\downarrow(2|\mu|)} - (n^{\downarrow|\mu|})^2),$$

which is evidently nonpositive. Thus it suffices to prove the upper bound

$$|\mathbf{E}[q_\mu(\boldsymbol{\lambda})]| \leq \exp(O(k^2 \log k)) \cdot r^{2k-2} \cdot (1/\omega^4). \tag{7.22}$$

By Lemma 7.2.9, the coefficients on the $p^\sharp_\nu$'s in the linear combination $q_\mu(\lambda)$ each have magnitude at most $\exp(O(k^2 \log k))$, and there are at most $2^{O(\sqrt{k})}$ of them. Thus (7.22) follows provided we can show $\mathbf{E}[p^\sharp_\nu(\boldsymbol{\lambda})] \leq r^{2k-2}/\omega^4$ for all ν of weight at most $2k-2$. This is immediate from Lemma 7.2.6 for all $\nu \neq (1)$, and when $\nu = (1)$ it still holds: Lemma 7.2.6 gives us the bound $r^2/\omega^2 \leq r^3/\omega^4 \leq r^{2k-2}/\omega^4$, the first inequality using $\omega \leq \sqrt{r}$ and the second using $k \geq 3$. □

We can now prove Theorem 7.3.2.

Proof of Theorem 7.3.2. Recall identity (7.11):

$$L_C(\lambda) = \sum_{k=2}^{C} \frac{(-1)^k}{kr^k} \cdot \sum_{\mathrm{wt}(\mu)=k+1} \frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)} p^\sharp_\mu(\lambda) + \sum_{k=2}^{C} \frac{(-1)^k \mathcal{O}_k(\lambda)}{kr^k}.$$

We claim that for each $2 \leq k \leq C$,

$$\mathbf{Var}\left[\frac{(-1)^k \mathcal{O}_k(\boldsymbol{\lambda})}{kr^k}\right] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}, \tag{7.23}$$

and that furthermore for each μ of weight $k+1$ we have

$$\mathbf{Var}\left[\frac{(-1)^k}{kr^k} \cdot \frac{k^{\downarrow(\ell(\mu)-1)}}{\mathrm{fact}(\mu)} p^\sharp_\mu(\boldsymbol{\lambda})\right] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}. \tag{7.24}$$

This is sufficient to complete the proof, as in general

$$\mathbf{Var}[\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_m] \leq m(\mathbf{Var}[\boldsymbol{X}_1] + \cdots + \mathbf{Var}[\boldsymbol{X}_m]); \tag{7.25}$$

in our particular case we have only $m = \exp(O(\sqrt{C}))$ summands, and this factor can be absorbed into the target variance bound of $\exp(O(C^{2.01})) \cdot (1/\omega^4)$. To verify (7.23), first recall that each $\mathcal{O}_k(\boldsymbol{\lambda})$ is a linear combination of $p^\sharp_\nu(\lambda)$'s for ν of weight at most $k \leq C$; further, the sum of the absolute value of the coefficients is at most $\exp(O(C^{2.01}))$ (see (7.15)). Using (7.25) again, it therefore suffices to check that

$$\mathbf{Var}\left[\frac{p^\sharp_\nu(\boldsymbol{\lambda})}{r^k}\right] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}$$

when $\mathrm{wt}(\nu) \leq k \leq C$. By Lemma 7.3.3 this is true, with a factor of r^{-2} to spare.

To verify (7.24), we may ignore the factor $\frac{(-1)^k}{k \cdot \mathrm{fact}(\mu)}$, and also ignore the factor $k^{\downarrow(\ell(\mu)-1)}$ as it contributes at most a multiplicative $C^C \ll \exp(O(C^{2.01}))$. Thus it suffices to show $\mathbf{Var}[p^\sharp_\mu(\boldsymbol{\lambda})/r^k] \leq \exp(O(C^{2.01}))/\omega^4$ for μ of weight $k+1$ (and $k \leq C$). But this is immediate from Lemma 7.3.3. □

We now work towards the proof of Theorem 7.3.1. Adding Theorem 7.3.2 and the square of (7.20) we obtain

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_r}[L_C(\boldsymbol{\lambda})^2] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}. \tag{7.26}$$

We would now like to similarly claim that

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n_{r_+}}[L^+_C(\boldsymbol{\lambda})^2] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}, \tag{7.27}$$

where we are writing

$$L^+_C(\lambda) \coloneqq \sum_{k=2}^{C} \frac{(-1)^k p^*_k(\lambda)}{k(r+1)^k}.$$

To obtain this, it suffices to repeat all of the arguments beginning with Section 7.2.2 until this point; the only thing that really changes is that $\omega = \omega_r$ needs to be replaced with ω_{r+1}, but this has a negligible effect on the bounds (and indeed usually very slightly improves them).

Next, we claim that Lemma 7.2.3 continues to hold if we replace $L_C(\lambda)$ with the analogous $L^+_C(\lambda)$. The key change to the proof comes in the last main inequality, where we need to observe that the

$$\left(\frac{1}{r^k} - \frac{1}{(r+\frac{1}{2})^k}\right) \leq \frac{k}{2r^{k+1}}$$

continues to hold if the left-hand side is replaced with

$$\left(\frac{1}{(r+\frac{1}{2})^k} - \frac{1}{(r+1)^k}\right).$$

We need one more definition for the proof of Theorem 7.3.1.

Definition 7.3.4. Say that $\lambda \vdash n$ is *usual*$^+$ if it is usual and if furthermore $|L^*_\infty(\lambda)| \leq 2$.

Lemma 7.3.5. *Both for* $\boldsymbol{\lambda} \sim \mathrm{SW}^n_r$ *and* $\boldsymbol{\lambda} \sim \mathrm{SW}^n_{r_+}$ *it holds that*

$$\mathbf{Pr}[\boldsymbol{\lambda} \text{ not usual}^+] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}.$$

Proof. For $\boldsymbol{\lambda} \sim \mathrm{SW}^n_r$, Lemma 7.2.2 tells us that

$$\mathbf{Pr}[\boldsymbol{\lambda} \text{ not usual}] \leq 2^{-20r/\omega} \leq 2^{-\Omega(\sqrt{r})} \ll \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}$$

and it's easy to check that this is also true with plenty of room to spare for $\boldsymbol{\lambda} \sim \mathrm{SW}^n_{r_+}$. Thus it suffices to verify for both distributions on $\boldsymbol{\lambda}$ that the probability of $|L^*_\infty(\lambda)| \leq 2$ satisfies the same upper bound. By applying Markov's inequality to (7.26), (7.27) we get

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}[L_C(\boldsymbol{\lambda})^2 \geq 1], \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}[L^+_C(\boldsymbol{\lambda})^2 \geq 1] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^4}.$$

Finally, when $\boldsymbol{\lambda}$ is usual and $|L_C(\boldsymbol{\lambda})^2| \not\geq 1$, it follows that necessarily $|L^*_\infty(\boldsymbol{\lambda})| \leq 2$, in light of Lemma 7.2.3 and the fact that $\frac{201}{\omega^3} \leq 1$. As noted earlier, the r_+-analogue of Lemma 7.2.3 holds, and hence we may draw the same conclusion concerning $L^+_C(\boldsymbol{\lambda})^2$. □

Finally we are ready to complete the proof of Theorem 7.3.1. We begin with

$$d_{\mathrm{TV}}(\mathrm{SW}^n_r, \mathrm{SW}^n_{r_+}) \leq \frac{1}{2} \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_r}[\boldsymbol{\lambda} \text{ not usual}^+] + \frac{1}{2} \mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}[\boldsymbol{\lambda} \text{ not usual}^+] \\ + \frac{1}{2} \sum_{\text{usual}^+ \ \lambda} \left|\mathrm{SW}^n_{r_+}[\lambda] - \mathrm{SW}^n_r[\lambda]\right|.$$

We can bound the first two terms above using Lemma 7.3.5. Indeed there is room to spare, as the bound we get is the square of what we can tolerate. Thus it remains to bound the third term by $\exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}$. For it we use

$$\begin{aligned}\sum_{\text{usual}^+\ \lambda} \left|\mathrm{SW}^n_{r_+}[\lambda] - \mathrm{SW}^n_r[\lambda]\right| &= \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}\left[1_{\{\boldsymbol{\lambda}\text{ usual}^+\}} \cdot \left|1 - \frac{\mathrm{SW}^n_r[\boldsymbol{\lambda}]}{\mathrm{SW}^n_{r_+}[\boldsymbol{\lambda}]}\right|\right] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}\left[1_{\{\boldsymbol{\lambda}\text{ usual}^+\}} \cdot \left|1 - \exp(u(\boldsymbol{\lambda}))\right|\right] \qquad (7.28)\end{aligned}$$

where

$$u(\boldsymbol{\lambda}) = \ln\left(\frac{\mathrm{SW}^n_r[\lambda]}{\mathrm{SW}^n_{r_+}[\lambda]}\right) = n\ln\left(1+\frac{1}{r}\right) - \frac{n}{r+\frac{1}{2}} + L^*_\infty(\boldsymbol{\lambda}), \qquad (7.29)$$

the last equality holding from (7.5) (see also the sentence after (7.7)) under the assumption that $\boldsymbol{\lambda}$ is usual (which we can indeed assume, since we're multiplying against $1_{\{\boldsymbol{\lambda}\text{ usual}^+\}}$). As we noted after (7.7), the first two quantities in (7.29) sum to a positive quantity not exceeding $\frac{n}{12r^3} \leq \frac{1}{\omega^2}$. Furthermore, because of the presence of the usual$^+$-indicator in (7.28) we may assume in analyzing (7.29) that $|L^*_\infty(\boldsymbol{\lambda})| \leq 2$. Thus we may use the bound $u(\boldsymbol{\lambda}) \leq 2 + \frac{1}{\omega^2} \leq 2.01$. Since $|1-\exp(u)| \leq 4|u|$ for $u \in [-2.01, 2.01]$, we may conclude that

$$(7.28) \leq 4 \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}\left[1_{\{\boldsymbol{\lambda}\text{ usual}^+\}} \cdot \left(\frac{1}{\omega^2} + |L^*_\infty(\boldsymbol{\lambda})|\right)\right].$$

Thus to complete the proof of Theorem 7.3.1 it remains to show

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}[|L^*_\infty(\boldsymbol{\lambda})|] \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2}.$$

By the r_+-analogue of Lemma 7.2.3, it suffices to prove this with $L^+_C(\boldsymbol{\lambda})$ in place of $L^*_\infty(\boldsymbol{\lambda})$, because $201/\omega^3 \ll \exp(O(C^{2.01}))/\omega^2$. But finally

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}[|L^+_C(\boldsymbol{\lambda})|] \leq \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n_{r_+}}[L^+_C(\boldsymbol{\lambda})^2]} \leq \exp(O(C^{2.01})) \cdot \frac{1}{\omega^2},$$

using Cauchy–Schwarz and (7.27). The proof of Theorem 7.3.1—and hence also the testing lower bound (7.1)—is therefore complete.

Chapter 8

Quantum rank testing

8.1 Testers with one-sided error

In this section, we prove the first part of Theorem 1.4.24, that $\Theta(r^2/\epsilon)$ copies are necessary and sufficient to test whether or not a state has rank r with one-sided error. We will show this by analyzing the following algorithm.

Definition 8.1.1 (Rank Tester)**.** Given $\rho^{\otimes n}$,

1. Sample $\boldsymbol{\lambda} \sim \mathrm{SW}^n_\rho$.

2. Accept if $\ell(\boldsymbol{\lambda}) \leq r$. Reject otherwise.

Write α for ρ's sorted spectrum, and let $\boldsymbol{w} \sim \alpha^{\otimes n}$. By Theorem 3.2.4, $\ell(\boldsymbol{\lambda}) = \mathrm{LDS}(\boldsymbol{w})$.

The key property we will need of the Rank Tester is the following:

Proposition 8.1.2. *The Rank Tester is the optimal algorithm for testing whether or not a state has rank r with one-sided error.*

Proof. To show this, we need to show (i) that every λ satisfying $\ell(\lambda) \leq r$ occurs with nonzero probability in SW^n_ρ for some ρ of rank r and (ii) that no λ satisfying $\ell(\lambda) > r$ occurs in SW^n_ρ for any ρ of rank r. The first follows because if ρ has r nonzero eigenvalues, then the word

$$w := \underbrace{r, \ldots, r}_{\lambda_r \text{ letters}}, \underbrace{(r-1), \ldots, (r-1)}_{\lambda_{r-1} \text{ letters}}, \ldots, \underbrace{1, \ldots, 1}_{\lambda_1 \text{ letters}}$$

occurs in $\mathcal{D}^{\otimes n}$ with nonzero probability. It is easy to check that $\lambda = \mathrm{RSK}(w)$.

To show that (ii) holds, if ρ is rank r, then α has at most r nonzero entries. Thus, any word w in the support of $\alpha^{\otimes n}$ will always satisfy $\mathrm{LDS}(w) \leq r$ because w will contain at most r distinct letters. As $\ell(\lambda) = \mathrm{LDS}(w)$, we are done. □

As a result of Proposition 8.1.2, Theorem 1.4.24 follows from the following lemma.

Lemma 8.1.3. *The Rank Tester tests whether or not a state has rank r with $\Theta(r^2/\epsilon)$ copies.*

Proof. If ρ is ϵ-far from having rank r, then α is ϵ-far in TV distance from having support size r. Thus, we can show the lemma by showing the following two facts about probability distributions.

(i) For every probability distribution α which is ϵ-far from having support size r, a random word $\boldsymbol{w} \sim \alpha^{\otimes n}$ satisfies $\mathrm{LDS}(\boldsymbol{w}) \geq r+1$ with probability at least $2/3$ for some $n = O(r^2/\epsilon)$.

(ii) There exists an integer d and a probability distribution α which is ϵ-far from having support size r such that, for a random word $\boldsymbol{w} \sim \alpha^{\otimes n}$, $\mathrm{LDS}(\boldsymbol{w}) \leq r$ with probability greater than $1/3$ whenever $n = o(r^2/\epsilon)$.

Proof of statement (i): We will need the following concentration bound for sums of geometric random variables.

Proposition 8.1.4 ([Bro]). *Write $X = X_1 + \ldots + X_n$, where the X_i's are i.i.d. geometric random variables with expectation μ. For any $k > 1$, $\mathbf{Pr}[X > kn\mu] \leq \exp\left(-\frac{1}{2}kn(1-1/k)^2\right)$.*

We note that Proposition 8.1.4 also holds with the weaker hypothesis that the X_i's are independent (and not necessarily identically distributed), each with expectation at most μ.

Recall that $\alpha_1 \geq \ldots \geq \alpha_d$. We will split into two cases, handled below: (1) $\alpha_{r+1} \geq \epsilon/4r$ and (2) $\alpha_{r+1} < \epsilon/4r$.

(1) Because the probabilities are sorted, $\alpha_1, \ldots, \alpha_{r+1} \geq \epsilon/4r$. For the infinite random word $\boldsymbol{w} \sim \alpha^{\otimes\infty}$, consider the number of letters one has to traverse through before finding $(r+1), r, \ldots, 1$ as a subsequence. This number is distributed as $\boldsymbol{X} = \boldsymbol{X}_{r+1} + \ldots + \boldsymbol{X}_1$, where $\boldsymbol{X}_i$ is a geometric random variable with success probability α_i.

By assumption, $\alpha_i \geq \epsilon/4r$, and therefore $\mathbf{E}\,\boldsymbol{X}_i \leq 4r/\epsilon$, for each $i \in [r+1]$. By Proposition 8.1.4, $\boldsymbol{X}$ is at most $24r^2/\epsilon$ with probability at least $2/3$. Thus, if $n = 24r^2/\epsilon$, then $\boldsymbol{w} \sim \alpha^{\otimes n}$ has a strictly decreasing subsequence of size $r+1$ with high probability.

(2) Because the probabilities are sorted, $\alpha_{r+1}, \ldots, \alpha_d < \epsilon/4r$. Place the letters from $\{r+1, \ldots, d\}$ into buckets as follows: starting from letter $(r+1)$ and proceeding in order, add each letter to the current bucket until it contains at least $\epsilon/4r$ weight. At this point, move to the next bucket and repeat this process starting with the current letter until all letters have been bucketed.

Because these letters have weight $\leq \epsilon/4r$, each bucket has total weight in the interval $[\epsilon/4r, \epsilon/2r)$ (except possibly the final bucket). There must be at least $2r+1$ buckets with nonzero total weight, as otherwise $\alpha_{r+1} + \ldots + \alpha_d < \epsilon$, contradicting the fact that α is ϵ-far from having support size r. This gives us at least $2r \geq r+1$ buckets each of which contains at least $\epsilon/4r$ total weight.

Now we can use an argument similar to case (1) to show that when $n = 24r^2/\epsilon$, a random $\boldsymbol{w} \sim \alpha^{\otimes n}$ will with probability $\geq 2/3$ have a strictly decreasing subsequence in which the first letter comes from bucket $r+1$, the second letter comes from bucket r, and so on (ending in a letter from the first bucket). This is a strictly decreasing subsequence of size $r+1$.

Proof of statement (ii): For $d \gg r$, define the probability distribution

$$\alpha = \left(1 - 2\epsilon, \frac{2\epsilon}{d-1}, \dots, \frac{2\epsilon}{d-1}\right).$$

Because $d \gg r$, α is ϵ-far from having support size r. For a string $w \in [d]^n$, let $\widetilde{w}$ be the substring of w formed by deleting all occurrences of the letter "1" from w. It is easy to see that $\mathrm{LDS}(\widetilde{w}) \leq \mathrm{LDS}(w) \leq \mathrm{LDS}(\widetilde{w}) + 1$.

For a randomly drawn $\boldsymbol{w} \sim \alpha^{\otimes n}$, let us condition on $\widetilde{\boldsymbol{w}}$ having a certain fixed length m. The value of $\mathrm{LDS}(\widetilde{\boldsymbol{w}})$ is distributed as the length of the longest decreasing subsequence in a uniformly random word drawn from $[d-1]^m$. By Theorem 3.2.4, this is distributed as $\boldsymbol{\lambda}_1'$ for $\boldsymbol{\lambda} \sim \mathrm{SW}_{d-1}^m$. Setting $B = \lceil 100\sqrt{m}\rceil$, let us show that $\mathbf{Pr}[\boldsymbol{\lambda}_1' \geq B]$ is small. If $B \geq d$, then surely $\boldsymbol{\lambda}_1' < B$ always, as $\boldsymbol{\lambda} \sim \mathrm{SW}_{d-1}^m$ will always have height at most $d-1$. On the other hand, if $B < d$, then by Proposition 3.7.12,

$$\mathbf{Pr}[\boldsymbol{\lambda}_1' \geq B] \leq \left(\frac{2e^2 m}{B^2}\right)^B \leq \frac{2e^2}{10000}.$$

In summary, conditioned on $\widetilde{\boldsymbol{w}}$ having a certain fixed length m, $\mathrm{LDS}(\widetilde{\boldsymbol{w}}) \leq O(\sqrt{m})$ with all but the above probability.

In expectation, for a random $\boldsymbol{w} \sim \alpha^{\otimes n}$, $\widetilde{\boldsymbol{w}}$ has length $2\epsilon d$. By Markov's inequality, the probability that the length of $\widetilde{\boldsymbol{w}}$ is greater than $200\epsilon d$ is at most $1/100$. Conditioned on the length of $\widetilde{\boldsymbol{w}}$ being at most $200\epsilon d$, the above paragraph tells us that $\mathrm{LDS}(\widetilde{\boldsymbol{w}}) \leq O(\sqrt{\epsilon d})$ with probability $1 - 2e^2/10000$. Thus, when $\boldsymbol{w} \sim \alpha^{\otimes n}$, we have with probability greater than $1/3$ that $\mathrm{LDS}(\boldsymbol{w}) \leq O(\sqrt{\epsilon d})$, which is $o(r)$ unless $d = \Omega(r^2/\epsilon)$. □

For our last result of this section, we will show that the copy complexity of the Rank Tester can be improved in certain interesting cases. In particular, the Rank Tester matches the upper bound of the Uniform Distribution Distinguisher from Definition 7.1.1 for the case of r v. $r+1$, and does so with one-sided error.

Proposition 8.1.5. *The Rank Tester can distinguish between the case when ρ's spectrum is uniform on either r or $r+1$ eigenvalues with $O(r^2)$ copies of ρ.*

Proof. If ρ's spectrum is uniform on r eigenvalues, then it is rank r and so the Rank Tester never rejects. Thus, we need only show that the Rank Tester rejects with probability $\geq 2/3$ when ρ's spectrum is uniform on $r+1$ eigenvalues for some $n = O(r^2)$. We will follow the analysis in the proof of statement (i) above and show that a random word $\boldsymbol{w} \sim \mathsf{Unif}_{r+1}^{\otimes n}$ has $\mathrm{LDS}(\boldsymbol{w}) = r+1$ with high probability. The gain will come from the fact that $\mathsf{Unif}_{r+1} = (1/(r+1), \dots, 1/(r+1))$.

For the infinite random word $\boldsymbol{w} \sim \mathsf{Unif}_{r+1}^{\otimes\infty}$, consider the number of letters one has to traverse through before finding $(r+1), r, \dots, 1$ as a subsequence. This number is distributed as $\boldsymbol{X} = \boldsymbol{X}_{r+1} + \dots + \boldsymbol{X}_1$, where $\boldsymbol{X}_i$ is a geometric random variable with success probability $1/(r+1)$ and expectation $r+1$. By Proposition 8.1.4, $\boldsymbol{X}$ is at most $6r^2$ with probability at least $2/3$. Thus, if $n = 6r^2$, then $\boldsymbol{w} \sim \mathsf{Unif}_{r+1}^{\otimes n}$ has a strictly decreasing subsequence of size $r+1$ with high probability. □

8.2 A lower bound for testers with two-sided error

In this section, we prove the second part of Theorem 1.4.24, that $\Omega(r/\epsilon)$ copies are necessary to test whether or not a state has rank r with two-sided error.

Proof. Let $d \gg r$. In this proof, we will take the viewpoint of a density matrix as a probability distribution over pure states. Let ρ and σ be maximally mixed on subspaces of dimension $(r-1)$ and $(d-1)$, respectively. Consider the following process for generating a product state $|\Psi\rangle = |\Psi_1\rangle \otimes \cdots \otimes |\Psi_n\rangle$:

1. Let $x \in \{0,1\}^n_{2\epsilon}$ be a uniformly random 2ϵ-biased string, meaning each coordinate is selected independently according to $\mathbf{Pr}[x_i = 1] = 2\epsilon$.
2. For each $i \in [n]$ such that $x_i = 0$, set $|\Psi_1\rangle \coloneqq |d\rangle$.
3. Let b be an arbitrary $\{0,1\}$-bit. For each $i \in [n]$ such that $x_i = 1$,
    (a) if $b = 0$, then set $|\Psi_i\rangle$ to be a state vector sampled from ρ.
    (b) if $b = 1$, then set $|\Psi_i\rangle$ to be a state vector sampled from σ.

If b is 0, then the mixed state output by this procedure has spectrum $(1-2\epsilon, \frac{2\epsilon}{r-1}, \ldots, \frac{2\epsilon}{r-1})$, which is rank r. On the other hand, if b is 1, then the mixed state output by this procedure has spectrum $(1-2\epsilon, \frac{2\epsilon}{d-1}, \ldots, \frac{2\epsilon}{d-1})$, which because $d \gg r$ is ϵ-far from having rank r.

Let us consider the choice of x in the first step, and set $\mathrm{wt}(\boldsymbol{x})$ to be the number of 1's in x. In expectation, $\mathrm{wt}(\boldsymbol{x})$ will be $2\epsilon n$, and so by Markov's inequality $\mathrm{wt}(\boldsymbol{x})$ will be at most $200\epsilon n$ with probability at least $99/100$. There must exist an x with $\mathrm{wt}(x) \leq 200\epsilon n$ conditioned on which the algorithm succeeds with probability at least $3/5$, as otherwise it will succeed in total with probability at most $1/100 + 99/100 \cdot 3/5 < 2/3$.

Fix any such x. The job of the algorithm is reduced to distinguishing between the cases when those $|\Psi_i\rangle$'s for which $x_i = 1$ came from ρ which is maximally mixed on a subspace of dimension $(r-1)$ (when $b = 0$) or from σ which is maximally mixed on a subspace of dimension $(d-1)$ (when $b = 1$). Because $d \gg r$, we have by Theorem 1.4.20 that this requires at least $\Omega(r)$ copies to succeed with probability at least $3/5$. Thus, we must have $200\epsilon n \geq \Omega(r)$, in which case $n = \Omega(r/\epsilon)$. □

Chapter 9

Majorization for the RSK algorithm

In this section we prove Theorem 1.5.4. The key to the proof will be the following strengthened version of the $d = 2$ case, which we believe is of independent interest.

Theorem 9.0.1. *Let $0 \leq p, q \leq 1$ satisfy $|q - \frac{1}{2}| \geq |p - \frac{1}{2}|$; in other words, the q-biased probability distribution $(q, 1-q)$ on $\{1, 2\}$ is "more extreme" than the p-biased distribution $(p, 1-p)$. Then for any $n \in \mathbb{N}$ there is a coupling $(\boldsymbol{w}, \boldsymbol{x})$ of the p-biased distribution on $\{1, 2\}^n$ and the q-biased distribution on $\{1, 2\}^n$ such that for all $1 \leq i \leq j \leq n$ we have $\mathrm{LIS}(\boldsymbol{x}[i \mathinner{..} j]) \geq \mathrm{LIS}(\boldsymbol{w}[i \mathinner{..} j])$ always.*

We now show how to prove Theorem 1.5.4 given Theorem 9.0.1. Then in the following sections we will prove Theorem 9.0.1.

Proof of Theorem 1.5.4 given Theorem 9.0.1. A classic result of Muirhead [Mui02] (see also [MOA11, B.1 Lemma]) says that $\beta \succ \alpha$ implies there is a sequence $\beta = \gamma_0 \succ \gamma_1 \succ \cdots \succ \gamma_t = \alpha$ such γ_i and γ_{i+1} differ in at most 2 coordinates. Since the $\trianglerighteq$ relation is transitive, by composing couplings it suffices to assume that α and β themselves differ in at most two coordinates. Since the Schur-Weyl distribution is symmetric with respect to permutations of $[d]$, we may assume that these two coordinates are 1 and 2. Thus we may assume $\alpha = (\alpha_1, \alpha_2, \beta_3, \beta_4, \ldots, \beta_d)$, where $\alpha_1 + \alpha_2 = \beta_1 + \beta_2$ and α_1, α_2 are between β_1, β_2.

We now define the coupling $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ as follows: We first choose a string $\boldsymbol{z} \in (\{*\} \cup \{3, 4, \ldots, d\})^n$ according to the product distribution in which symbol j has probability β_j for $j \geq 3$ and symbol $*$ has the remaining probability $\beta_1 + \beta_2$. Let $\boldsymbol{n}_*$ denote the number of $*$'s in $\boldsymbol{z}$. Next, we use Theorem 9.0.1 to choose coupled strings $(\boldsymbol{w}, \boldsymbol{x})$ with the p-biased distribution on $\{1, 2\}^{\boldsymbol{n}_*}$ and the q-biased distribution on $\{1, 2\}^{\boldsymbol{n}_*}$ (respectively), where $p = \frac{\alpha_1}{\beta_1 + \beta_2}$ and $q = \frac{\beta_1}{\beta_1 + \beta_2}$. Note indeed that $|q - \frac{1}{2}| \geq |p - \frac{1}{2}|$, and hence $\mathrm{LIS}(\boldsymbol{x}[i \mathinner{..} j]) \geq \mathrm{LIS}(\boldsymbol{w}[i \mathinner{..} j])$ for all $1 \leq i \leq \boldsymbol{n}_*$. Now let "$\boldsymbol{z} \cup \boldsymbol{w}$" denote the string in $[d]^n$ obtained by filling in the $*$'s in $\boldsymbol{z}$ with the symbols from $\boldsymbol{w}$, in the natural left-to-right order; similarly define "$\boldsymbol{z} \cup \boldsymbol{x}$". Note that $\boldsymbol{z} \cup \boldsymbol{w}$ is distributed according to the product distribution $\alpha^{\otimes n}$ and likewise for $\boldsymbol{z} \cup \boldsymbol{x}$ and $\beta^{\otimes n}$. Our final coupling is now obtained by taking $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{z} \cup \boldsymbol{w})$ and $\boldsymbol{\mu} = \mathrm{shRSK}(\boldsymbol{z} \cup \boldsymbol{x})$. We need to show that $\boldsymbol{\mu} \trianglerighteq \boldsymbol{\lambda}$ always.

By Greene's Theorem, it suffices to show that if $s_1, \ldots, s_k$ are disjoint increasing subsequences in $\boldsymbol{z} \cup \boldsymbol{w}$ of total length S, we can find k disjoint increasing subsequences $s'_1, \ldots, s'_k$ in $\boldsymbol{z} \cup \boldsymbol{x}$ of total length at least S. We first dispose of some simple cases. If none of $s_1, \ldots, s_k$

contains any 1's or 2's, then we may take $s'_i = s_i$ for $i \in [k]$, since these subsequences all still appear in $\boldsymbol{z} \cup \boldsymbol{x}$. The case when exactly one of $s_1, \dots, s_k$ contains any 1's or 2's is also easy. Without loss of generality, say that s_k is the only subsequence containing 1's and 2's. We may partition it as (t, u), where t is a subsequence of $\boldsymbol{w}$ and u is a subsequence of the non-$*$'s in $\boldsymbol{z}$ that follow $\boldsymbol{w}$. Now let t' be the longest increasing subsequence in $\boldsymbol{x}$. As t is an increasing subsequence of $\boldsymbol{w}$, we know that t' is at least as long as t. Further, (t', u) is an increasing subsequence in $\boldsymbol{z} \cup \boldsymbol{x}$. Thus we may take $s'_i = s_i$ for $i < k$, and $s'_k = (t', u)$.

We now come to the main case, when at least two of $s_1, \dots, s_k$ contain 1's and/or 2's. Let's first look at the position $j \in [n]$ of the rightmost 1 or 2 among $s_1, \dots, s_k$. Without loss of generality, assume it occurs in s_k. Next, look at the position $i \in [n]$ of the rightmost 1 or 2 among $s_1, \dots, s_{k-1}$. Without loss of generality, assume it occurs in s_{k-1}. We will now modify the subsequences $s_1, \dots, s_k$ as follows:

- all 1's and 2's are deleted from $s_1, \dots, s_{k-2}$ (note that these all occur prior to position i);
- s_{k-1} is changed to consist of all the 2's within $(\boldsymbol{z} \cup \boldsymbol{w})[1 \mathinner{..} i]$;
- the portion of s_k to the right of position i is unchanged, but the preceding portion is changed to consist of all the 1's within $(\boldsymbol{z} \cup \boldsymbol{w})[1 \mathinner{..} i]$.

It is easy to see that the new $s_1, \dots, s_k$ remain disjoint subsequences of $\boldsymbol{z} \cup \boldsymbol{w}$, with total length at least S. We may also assume that the portion of s_k between positions $i + 1$ and j consists of a longest increasing subsequence of $\boldsymbol{w}$.

Since the subsequences $s_1, \dots, s_{k-2}$ don't contain any 1's or 2's, they still appear in $\boldsymbol{z} \cup \boldsymbol{x}$, and we may take these as our $s'_1, \dots, s'_{k-2}$. We will also define s'_{k-1} to consist of all 2's within $(\boldsymbol{z} \cup \boldsymbol{x})[1 \mathinner{..} i]$. Finally, we will define s'_k to consist of all 1's within $(\boldsymbol{z} \cup \boldsymbol{z})[1 \mathinner{..} i]$, followed by the longest increasing subsequence of $\boldsymbol{x}$ occurring within positions $(i + 1) \mathinner{..} j$ in $\boldsymbol{z} \cup \boldsymbol{x}$, followed by the portion of s_k to the right of position j (which does not contain any 1's or 2's and hence is still in $\boldsymbol{z} \cup \boldsymbol{x}$). It is clear that $s'_1, \dots, s'_k$ are indeed disjoint increasing subsequences of $\boldsymbol{z} \cup \boldsymbol{x}$. Their total length is the sum of four quantities:

- the total length of $s_1, \dots, s_{k-2}$;
- the total number of 1's and 2's within $(\boldsymbol{z} \cup \boldsymbol{x})[1 \mathinner{..} i]$;
- the length of the longest increasing subsequence of $\boldsymbol{x}$ occurring within positions $(i + 1) \mathinner{..} j$ in $\boldsymbol{z} \cup \boldsymbol{x}$;
- the length of the portion of s_k to the right of position j.

By the coupling property of $(\boldsymbol{w}, \boldsymbol{x})$, the third quantity above is at least the length of the longest increasing subsequence of $\boldsymbol{w}$ occurring within positions $(i + 1) \mathinner{..} j$ in $\boldsymbol{z} \cup \boldsymbol{w}$. But this precisely shows that the total length of $s'_1, \dots, s'_k$ is at least that of $s_1, \dots, s_k$, as desired. □

9.1 Substring-LIS-dominance: RSK and Dyck paths

In this section we make some preparatory definitions and observations toward proving Theorem 9.0.1. We begin by codifying the key property therein.

Definition 9.1.1. Let $w, w' \in \mathcal{A}^n$ be strings of equal length. We say w' *substring-LIS-dominates* w, notated $w' \unrhd\!\!\gg w$, if $\mathrm{LIS}(w'[i\,..\,j]) \geq \mathrm{LIS}(w[i\,..\,j])$ for all $1 \leq i \leq j \leq n$. (Thus the coupling in Theorem 9.0.1 satisfies $\boldsymbol{w} \unrhd\!\!\gg \boldsymbol{v}$ always.) The relation $\unrhd\!\!\gg$ is reflexive and transitive. If we have the substring-LIS-dominance condition just for $i = 1$ we say that w' *prefix-LIS-dominates* w. If we have it just for $j = n$ we say that w' *suffix-LIS-dominates* w.

Definition 9.1.2. For a string $w \in \mathcal{A}^n$ we write $\mathrm{behead}(w)$ for $w[2\,..\,n]$ and $\mathrm{curtail}(w)$ for $w[1\,..\,n-1]$.

Remark 9.1.3. We may equivalently define substring-LIS-dominance recursively, as follows. If w' and w have length 0 then $w' \unrhd\!\!\gg w$. If w' and w have length $n > 0$, then $w' \unrhd\!\!\gg w$ if and only if $\mathrm{LIS}(w') \geq \mathrm{LIS}(w)$ and $\mathrm{behead}(w') \unrhd\!\!\gg \mathrm{behead}(w)$ and $\mathrm{curtail}(w') \unrhd\!\!\gg \mathrm{curtail}(w)$. By omitting the second/third condition we get a recursive definition of prefix/suffix-LIS-dominance.

Definition 9.1.4. Let Q be a (nonempty) standard Young tableau. We define $\mathrm{curtail}(Q)$ to be the standard Young tableau obtained by deleting the box with maximum label from Q.

The following fact is immediate from the definition of the RSK correspondence:

Proposition 9.1.5. *Let $w \in \mathcal{A}^n$ be a nonempty string. Suppose $\mathrm{RSK}(w) = (P, Q)$ and $\mathrm{RSK}(\mathrm{curtail}(w)) = (P', Q')$. Then $Q' = \mathrm{curtail}(Q)$.*

The analogous fact for beheading is more complicated.

Definition 9.1.6. Let Q be a (nonempty) standard Young tableau. We define $\mathrm{behead}(Q)$ to be the standard Young tableau obtained by deleting the top-left box of Q, sliding the hole outside of the tableau according to jeu de taquin (see, e.g., [Ful97, Sag01]), and then decreasing all entries by 1. (The more traditional notation for $\mathrm{behead}(Q)$ is $\Delta(Q)$.)

The following fact is due to [Sch63]; see [Sag01, Proposition 3.9.3] for an explicit proof.[1]

Proposition 9.1.7. *Let $w \in \mathcal{A}^n$ be a nonempty string. Suppose $\mathrm{RSK}(w) = (P, Q)$ and $\mathrm{RSK}(\mathrm{behead}(w)) = (P', Q')$. Then $Q' = \mathrm{behead}(Q)$.*

Proposition 9.1.8. *Let $w, w' \in \mathcal{A}^n$ be strings of equal length and write $\mathrm{RSK}(w) = (P, Q)$, $\mathrm{RSK}(w') = (P', Q')$. Then whether or not $w' \unrhd\!\!\gg w$ can be determined just from the recording tableaus Q' and Q.*

Proof. This follows from the recursive definition of $\unrhd\!\!\gg$ given in Remark 9.1.3: whether $\mathrm{LIS}(w') \geq \mathrm{LIS}(w)$ can be determined by checking whether the first row of Q' is at least as long as the first row of Q; the recursive checks can then be performed with the aid of Propositions 9.1.5, 9.1.7. □

Definition 9.1.9. In light of Proposition 9.1.8 we may define the relation $\unrhd\!\!\gg$ on standard Young tableaus.

[1] Technically, therein it is proved only for strings with distinct letters. One can recover the result for general strings in the standard manner; if the letters w_i and w_j are equal we break the tie by using the order relation on i, j. See also [vL13, Lemma].

Remark 9.1.10. The simplicity of Proposition 9.1.5 implies that it is very easy to tell, given $w, w' \in \mathcal{A}^n$ with recording tableaus Q and Q', whether w' suffix-LIS-dominates w. One only needs to check whether $Q'_{1j} \le Q_{1j}$ for all $j \ge 1$ (treating empty entries as ∞). On the other hand, it is not particularly easy to tell from Q' and Q whether w' prefix-LIS-dominates w; one seems to need to execute all of the jeu de taquin slides.

We henceforth focus attention on alphabets of size 2. Under RSK, these yield standard Young tableaus with at most 2-rows. (For brevity, we henceforth call these *2-row Young tableaus*, even when they have fewer than 2 rows.) In turn, 2-row Young tableaus can be identified with Dyck paths (also known as ballot sequences).

Definition 9.1.11. We define a *Dyck path of length n* to be a path in the xy-plane that starts from $(0, 0)$, takes n steps of the form $(+1, +1)$ (an *upstep*) or $(+1, -1)$ (a *downstep*), and never passes below the x-axis. We say that the *height* of a step s, written $\mathrm{ht}(s)$, is the y-coordinate of its endpoint; the *(final) height* of a Dyck path W, written $\mathrm{ht}(W)$, is the height of its last step. We do *not* require the final height of a path to be 0; if it is we call the path *complete*, and otherwise we call it *incomplete*. A *return* refers to a point where the path returns to the x-axis; i.e., to the end of a step of height 0. An *arch* refers to a minimal complete subpath of a Dyck path; i.e., a subpath between two consecutive returns (or between the origin and the first return).

Definition 9.1.12. We identify each 2-row standard Young tableau Q of size n with a Dyck path W of length n. The identification is the standard one: reading off the entries of Q from 1 to n, we add an upstep to W when the entry is in the first row and a downstep when it is in the second row. The fact that this produces a Dyck path (i.e., the path does not pass below the x-axis) follows from the standard Young tableau property. Note that the final height of W is the difference in length between Q's two rows. We also naturally extend the terminology "return" to 2-row standard Young tableaus Q: a *return* is a second-row box labeled $2j$ such that boxes in Q labeled $1, \dots, 2j$ form a rectangular $2 \times j$ standard Young tableau.

Definition 9.1.13. In light of Definition 9.1.9 and the above identification, we may define the relation $\ggg$ on Dyck paths.

Of course, we want to see how beheading and curtailment apply to Dyck paths. The following fact is immediate:

Proposition 9.1.14. *If W is the Dyck path corresponding to a nonempty 2-row standard Young tableau Q, then the Dyck path W' corresponding to* $\mathrm{curtail}(Q)$ *is formed from W by deleting its last segment. We write $W' = \mathrm{curtail}(W)$ for this new path.*

Again, the case of beheading is more complicated. We first make some definitions.

Definition 9.1.15. *Raising* refers to converting a downstep in a Dyck path to an upstep; note that this increases the Dyck path's height by 2. Conversely, *lowering* refers to converting an upstep to a downstep. Generally, we only allow lowering when the result is still a Dyck path; i.e., never passes below the x-axis.

Proposition 9.1.16. *Let Q be a nonempty 2-row standard Young tableau, with corresponding Dyck path W. Let W' be the Dyck path corresponding to* $\mathrm{behead}(Q)$*. Then W' is formed from W as follows: First, the initial step of W is deleted (and the origin is shifted to the new initial point). If W had no returns then the operation is complete and W' is the resulting Dyck path. Otherwise, if W had at least one return, then in the new path W' that step (which currently goes below the x-axis) is raised. In either case, we write $W' = \mathrm{behead}(W)$ for the resulting path.*

Proof. We use Definitions 9.1.6 and 9.1.12. Deleting the top-left box of Q corresponds to deleting the first step of W, and decreasing all entries in Q by 1 corresponds to shifting the origin in W. Consider now the jeu de taquin slide in Q. The empty box stays in the first row until it first reaches a position j such that $Q_{1,j+1} > Q_{2,j}$ — if such a position exists. Such a position does exist if and only if Q contains a return (with box $(2, j)$ being the first such return). If Q (equivalently, W) has no return then the empty box slides out of the first row of Q, and indeed this corresponds to making no further changes to W. If Q has its first return at box $(2, j)$, this means the jeu de taquin will slide up the box labeled $2j$ (corresponding to raising the first return step in W); then all remaining slides will be in the bottom row of Q, corresponding to no further changes to W. □

Remark 9.1.17. Similar to Remark 9.1.10, it is easily to "visually" check the suffix-LIS-domination relation for Dyck paths: W' suffix-LIS-dominates W if and only if W' is at least as high as W throughout the length of both paths. On the other hand, checking the full substring-LIS-domination relation is more involved; we have $W' \mathrel{\rhd\!\!\!\gg} W$ if and only if for any number of simultaneous beheadings to W' and W, the former path always stays at least as high as the latter.

Finally, we will require the following definition:

Definition 9.1.18. A *hinged range* is a sequence $(R_0, s_1, R_1, s_2, R_2, \ldots, s_k, R_k)$ (with $k \geq 0$), where each s_i is a step (upstep or downstep) called a *hinge* and each R_i is a Dyck path (possibly of length 0) called a *range*. The "internal ranges" $R_1, \ldots, R_{k-1}$ are required to be complete Dyck paths; the "external ranges" R_0 and R_k may be incomplete.

We may identify the hinged range with the path formed by concatenating its components; note that this need not be a Dyck path, as it may pass below the origin.

If H is a hinged range and H' is formed by raising zero or more of its hinges (i.e., converting downstep hinges to upsteps), we say that H' is a *raising* of H or, equivalently, that H is a *lowering* of H'. We call a hinged range *fully lowered* (respectively, *fully raised*) if all its hinges are downsteps (respectively, upsteps).

9.2 A bijection on Dyck paths

Theorem 9.2.1. *Fix integers $n \geq 2$ and $1 \leq \lambda_2 \leq \lfloor \frac{n}{2} \rfloor$. Define*

$$\begin{aligned} \mathcal{W} = \big\{ (W, s_1) : {} & W \text{ is a length-}n \text{ Dyck path with exactly } \lambda_2 \text{ downsteps;} \\ & s_1 \text{ is a downstep in } W \big\} \end{aligned}$$

and

$$\mathcal{W}' = \bigcup_{k=1}^{\lambda_2} \big\{ (W', s_1') : \begin{array}{l} W' \text{ is a length-}n \text{ Dyck path with exactly } \lambda_2 - k \text{ downsteps;} \\ s_1' \text{ is an upstep in } W' \text{ with } k+1 \le \mathrm{ht}(s_1') \le \mathrm{ht}(W') - k + 1; \\ s_1' \text{ is the rightmost upstep in } W' \text{ of its height} \big\}. \end{array}$$

Then there is an explicit bijection $f : \mathcal{W} \to \mathcal{W}'$ such that whenever $f(W, s_1) = (W', s_1')$ it holds that $W' \ggg W$.

Remark 9.2.2. Each length-n Dyck path with exactly λ_2 downsteps occurs exactly λ_2 times in $\mathcal{W}$. Each length-n Dyck path with strictly fewer than λ_2 downsteps occurs exactly $n - 2\lambda_2 + 1$ times in $\mathcal{W}'$.

Proof of Theorem 9.2.1. Given any $(W, s_1) \in \mathcal{W}$, we define f's value on it as follows. Let s_2 be the first downstep following s_1 in W having height $\mathrm{ht}(s_1) - 1$; let s_3 be the first downstep following s_2 in W following s_2 having height $\mathrm{ht}(s_2) - 1$; etc., until reaching downstep s_k having no subsequent downstep of smaller height. Now decompose W as a (fully lowered) hinged range $H = (R_0, s_1, R_1, \ldots, s_k, R_k)$. Let $H' = (R_0', s_1', R_1', \ldots, s_k', R_k')$ be the fully raised version of H (where each R_j' is just R_j and each s_j' is an upstep). Then $f(W, s_k)$ is defined to be (W', s_1'), where W' is the Dyck path corresponding to H'.

First we check that indeed $(W', s_1') \in \mathcal{W}'$. As W' is formed from W by k raisings, it has exactly $\lambda_2 - k$ downsteps. Since $\mathrm{ht}(s_k) \ge 0$ it follows that $\mathrm{ht}(s_1) \ge k - 1$ and hence $\mathrm{ht}(s_1') \ge k + 1$. On the other hand, $\mathrm{ht}(s_1') + (k-1) = \mathrm{ht}(s_k') \le \mathrm{ht}(W')$ and so $\mathrm{ht}(s_1') \le \mathrm{ht}(W') - k + 1$. Finally, s_1' is the rightmost upstep in W' of its height because H' is fully raised.

To show that f is a bijection, we will define the function $g : \mathcal{W}' \to \mathcal{W}$ that will evidently be f's inverse. Given any $(W', s_1') \in \mathcal{W}$, with W' having exactly $\lambda_2 - k$ downsteps, we define g's value on it as follows. Let s_2' be the *last* (rightmost) upstep following s_1' in W' having height $\mathrm{ht}(s_1') + 1$; let s_3' be the last upstep following s_2' in W' having height $\mathrm{ht}(s_2') + 1$; etc., until s_k' is defined. That this s_k' indeed exists follows from the fact that $\mathrm{ht}(s_1') \le \mathrm{ht}(W') - k + 1$. Now decompose W' as a (fully raised) hinged range $H' = (R_0', s_1', R_1', \ldots, s_k', R_k')$. The fact that R_k' is a Dyck path (i.e., does not pass below its starting height) again follows from the fact that $\mathrm{ht}(s_k') = \mathrm{ht}(s_1') + k - 1 \le \mathrm{ht}(W')$. Finally, let $H = (R_0, s_1, R_1, \ldots, s_k, R_k)$ be the fully lowered version of H', and W the corresponding path. As W has exactly λ_2 downsteps, we may define $g(W', s_1') = (W, s_1)$ provided W is indeed a Dyck path. But this is the case, because the lowest point of W occurs at the endpoint of s_k, and $\mathrm{ht}(s_k) = \mathrm{ht}(s_1) - k + 1 = \mathrm{ht}(s_1') - 2 - k + 1 = \mathrm{ht}(s_1') - k - 1 \ge 0$ since $\mathrm{ht}(s_1') \ge k + 1$.

It is fairly evident that f and g are inverses. The essential thing to check is that the sequence $s_1, \ldots, s_k$ determined from s_1 when computing $f(W, s_1)$ is "the same" (up to raising/lowering) as the sequence $s_1', \ldots, s_{k'}'$ determined from s_1' in computing $g(W', s_1')$, and vice versa. The fact that the sequences have the same *length* follows, in the $g \circ f = id$ case, from the fact that $\mathrm{ht}(W') = \mathrm{ht}(W) + 2k$; it follows, in the $f \circ g = id$ case, from the fact that R_k' is a Dyck path. The fact that the hinges have the same identity is evident from the nature of fully raising/lowering hinged ranges.

It remains to show that if $f(W, s_1) = (W', s_1')$ then $W' \ggg W$. Referring to Remark 9.1.17, we need to show that if W' and W are both simultaneously beheaded some number of times b, then in the resulting paths, W' is at least as high as W throughout their lengths. In turn, this is implied by the following more general statement:

Claim 9.2.3. *After b beheadings, W' and W may be expressed as hinged ranges $H' = (R_0, s'_1, R_1, \dots, s'_k, R_k)$ and $H = (R_0, s_1, R_1, \dots, s_k, R_k)$ (respectively) such that H' is the fully raised version of H (i.e., each s'_j is an upstep).*

(Note that we do not necessarily claim that H is the fully lowered version of H'.)

The claim can be proved by induction on b. The base case $b = 0$ follows by definition of f. Throughout the induction we may assume that the common initial Dyck path R_0 is nonempty, as otherwise s_1 must be an upstep, in which case we can redefine the common initial Dyck path of W and W' to be $(s_1, R_1) = (s'_1, R_1)$.

We now show the inductive step. Assume W' and W are nonempty paths as in the claim's statement, with R_0 nonempty. Suppose now that W' and W are simultaneously beheaded. The first step of W' and W (an upstep belonging to R_0) is thus deleted, and the origin shifted. If R_0 contained a downstep to height 0 then the first such downstep is raised in both behead(W') and behead(W) and the inductive claim is maintained. Otherwise, suppose R_0 contained no downsteps to height 0. It follows immediately that W' originally had no returns to height 0 at all; hence the beheading of W' is completed by the deletion of its first step. It may also be that W had no returns to height 0 at all; then the beheading of W is also completed by the deletion of its first step and the induction hypothesis is clearly maintained. On the other hand, W *may* have had some downsteps to 0 within $(s_1, R_1, \dots, s_k, R_k)$. In this case, the first (leftmost) such downstep must occur at one of the hinges s_j, and the beheading of W is completed by raising this hinge. The inductive hypothesis is therefore again maintained. This completes the induction. □

We derive an immediate corollary, after introducing a bit of notation:

Definition 9.2.4. We write $\mathrm{SYT}_n(=\lambda_2)$ (respectively, $\mathrm{SYT}_n(\leq\lambda_2)$) for the set of 2-row standard Young tableaus of size n with exactly (respectively, at most) λ_2 boxes in the second row.

Corollary 9.2.5. *For any integers $n \geq 2$ and $0 \leq \lambda_2 \leq \lfloor \frac{n}{2} \rfloor$, there is a coupling $(\boldsymbol{Q}, \boldsymbol{Q}')$ of the uniform distribution on $\mathrm{SYT}_n(=\lambda_2)$ and the uniform distribution on $\mathrm{SYT}_n(\leq\lambda_2 - 1)$ such that $\boldsymbol{Q}' \rhd\!\!\gg \boldsymbol{Q}$ always.*

Proof. Let $(\boldsymbol{W}, \boldsymbol{s}_1)$ be drawn uniformly at random from the set $\mathcal{W}$ defined in Theorem 9.2.1, and let $(\boldsymbol{W}', \boldsymbol{s}'_1) = f(\boldsymbol{W}, \boldsymbol{s}_1)$. Let $\boldsymbol{Q} \in \mathrm{SYT}_n(=\lambda_2)$, $\boldsymbol{Q}' \in \mathrm{SYT}_n(\leq\lambda_2 - 1)$ be the 2-row standard Young tableaus identified with $\boldsymbol{W}$, $\boldsymbol{W}'$ (respectively). Then Theorem 9.2.1 tells us that $\boldsymbol{Q}' \rhd\!\!\gg \boldsymbol{Q}$ always, and Remark 9.2.2 tells us that $\boldsymbol{Q}$ and $\boldsymbol{Q}'$ are each uniformly distributed. □

Corollary 9.2.6. *For any integers $n \geq 0$ and $0 \leq \lambda'_2 \leq \lambda_2 \leq \lfloor \frac{n}{2} \rfloor$, there is a coupling $(\boldsymbol{Q}, \boldsymbol{Q}')$ of the uniform distribution on $\mathrm{SYT}_n(\leq\lambda_2)$ and the uniform distribution on $\mathrm{SYT}_n(\leq\lambda'_2)$ such that $\boldsymbol{Q}' \rhd\!\!\gg \boldsymbol{Q}$ always.*

Proof. The cases $n < 2$ and $\lambda'_2 = \lambda_2$ are trivial, so we may assume $n \geq 2$ and $0 \leq \lambda'_2 < \lambda_2 \leq \lfloor \frac{n}{2} \rfloor$. By composing couplings and using transitivity of $\rhd\!\!\gg$, it suffices to treat the case

$\lambda'_2 = \lambda_2 - 1$. But the uniform distribution on $\mathrm{SYT}_n(\leq\lambda_2)$ is a mixture of (a) the uniform distribution on $\mathrm{SYT}_n(=\lambda_2)$, (b) the uniform distribution on $\mathrm{SYT}_n(\leq\lambda_2 - 1)$; and these can be coupled to $\mathrm{SYT}_n(\leq\lambda_2-1)$ under the $\ggg$ relation using (a) Corollary 9.2.5, (b) the identity coupling. □

Before giving the next corollary, we have a definition.

Definition 9.2.7. Let $\mathcal{A}$ be any 2-letter alphabet. We write $\mathcal{A}^n_k$ for the set of length-n strings over $\mathcal{A}$ with exactly k copies of the larger letter, and we write $\mathcal{A}^n_{k,n-k} = \mathcal{A}^n_k \cup \mathcal{A}^n_{n-k}$.

Corollary 9.2.8. *For $\mathcal{A}$ a 2-letter alphabet and integers $0 \leq k' \leq k \leq \lfloor \frac{n}{2} \rfloor$, there is a coupling $(\boldsymbol{w}, \boldsymbol{w}')$ of the uniform distribution on $\mathcal{A}^n_{k,n-k}$ and the uniform distribution on $\mathcal{A}^n_{k',n-k'}$ such that $\boldsymbol{w}' \ggg \boldsymbol{w}$ always.*

Proof. We first recall that if $\boldsymbol{x} \sim \mathcal{A}^n_k$ is uniformly random and $(\boldsymbol{P}, \boldsymbol{Q}) = \mathrm{RSK}(\boldsymbol{x})$, then the recording tableau $\boldsymbol{Q}$ is uniformly random on $\mathrm{SYT}_n(\leq k)$. This is because for each possible recording tableau $Q \in \mathrm{SYT}_n(\leq k)$ there is a unique insertion tableau P of the same shape as Q having exactly k boxes labeled with the larger letter of $\mathcal{A}$. (Specifically, if $P \vdash (\lambda_1, \lambda_2)$, then the last $k - \lambda_2$ boxes of P's first row, and all of the boxes of P's second row, are labeled with A's larger letter.) It follows that the same is true if $\boldsymbol{x} \sim \mathcal{A}^n_{k,n-k}$ is uniformly random. But now the desired coupling follows from Corollary 9.2.6 (recalling Definition 9.1.9). □

In fact, Corollary 9.2.8 is fundamentally stronger than our desired Theorem 9.0.1, as we now show:

Proof of Theorem 9.0.1. For $r \in [0,1]$, suppose we draw an r-biased string $\boldsymbol{y} \in \{1,2\}^n$ and define the random variable $\boldsymbol{j}$ such that $\boldsymbol{y} \in \{1,2\}^n_{\boldsymbol{j},n-\boldsymbol{j}}$. (Note that given $\boldsymbol{j}$, the string $\boldsymbol{y}$ is uniformly distributed on $\{1,2\}^n_{\boldsymbol{j},n-\boldsymbol{j}}$.) Write $L_r(\ell)$ for the cumulative distribution function of $\boldsymbol{j}$; i.e., $L_r(\ell) = \mathbf{Pr}[\boldsymbol{y} \in \cup_{j \leq \ell}\{1,2\}^n_{j,n-j}]$, where $\boldsymbol{y}$ is r-biased.

Claim: $L_q(\ell) \geq L_p(\ell)$ for all $0 \leq \ell \leq \lfloor \frac{n}{2} \rfloor$.

Before proving the claim, let us show how it is used to complete the proof of Theorem 9.0.1. We define the required coupling $(\boldsymbol{w}, \boldsymbol{x})$ of p-biased and q-biased distributions as follows: First we choose $\boldsymbol{\theta} \in [0,1]$ uniformly at random. Next we define $\boldsymbol{k}$ (respectively, $\boldsymbol{k}'$) to be the least integer such that $L_p(\boldsymbol{k}) \geq \boldsymbol{\theta}$ (respectively, $L_q(\boldsymbol{k}') \geq \boldsymbol{\theta}$); from the claim it follows that $\boldsymbol{k}' \leq \boldsymbol{k}$ always. Finally, we let $(\boldsymbol{w}, \boldsymbol{x})$ be drawn from the coupling on $\{1,2\}^n_{\boldsymbol{k},n-\boldsymbol{k}}$ and $\{1,2\}^n_{\boldsymbol{k}',n-\boldsymbol{k}'}$ specified in Corollary 9.2.8. Then as required, we have that $\boldsymbol{x}' \ggg \boldsymbol{w}$ always, and that $\boldsymbol{w}$ has the p-biased distribution and $\boldsymbol{x}$ has the q-biased distribution.

It therefore remains to prove the claim. We may exclude the trivial cases $\ell = \frac{n}{2}$ or $q \in \{0,1\}$, where $L_q(\ell) = 1$. Also, since $L_r(\ell) = L_{1-r}(\ell)$ by symmetry, we may assume $0 < q \leq p \leq \frac{1}{2}$. Thus it suffices to show that $\frac{d}{dr}L_r(\ell) \leq 0$ for $0 < r \leq \frac{1}{2}$. Letting $\boldsymbol{h}$ denote the "Hamming weight" (number of 2's) in an r-biased random string on $\{1,2\}^n$, we have

$$L_r(\ell) = \mathbf{Pr}[\boldsymbol{h} \leq \ell] + \mathbf{Pr}[\boldsymbol{h} \geq n - \ell] = 1 - \mathbf{Pr}[\boldsymbol{h} > \ell] + \mathbf{Pr}[\boldsymbol{h} > n - \ell - 1]$$

$$\Rightarrow \frac{d}{dr}L_r(\ell) = -\frac{d}{dr}\,\mathbf{Pr}[\boldsymbol{h} > \ell] + \frac{d}{dr}\,\mathbf{Pr}[\boldsymbol{h} > n - 1 - \ell].$$

(The first equality used $\ell < \frac{n}{2}$.) But it is a basic fact that $\frac{d}{dr}\,\mathbf{Pr}[\boldsymbol{h} > t] = n\binom{n-1}{t} r^t (1-r)^{n-1-t}$. Thus

$$\frac{d}{dr} L_r(\ell) = n\binom{n-1}{\ell}\left(-r^{\ell}(1-r)^{n-1-\ell} + r^{n-1-\ell}(1-r)^{\ell}\right),$$

and we may verify this is indeed nonpositive:

$$-r^{\ell}(1-r)^{n-1-\ell} + r^{n-1-\ell}(1-r)^{\ell} \le 0 \iff 1 \le \left(\tfrac{1-r}{r}\right)^{n-1-2\ell},$$

which is true since $0 < r \le \frac{1}{2}$ and $n - 1 - 2\ell \ge 0$ (using $\ell < \frac{n}{2}$ again). □

Chapter 10

Open problems

10.1 Identity testing

What is the copy complexity of identity testing, i.e. of testing whether $\rho = \sigma$ for a known $\sigma \in \mathbb{C}^{d \times d}$. In the classical setting, it is known that the sample complexity of testing whether an unknown distribution α equals a known distribution β is maximized when β is the uniform distribution. Analogously, we expect that the copy complexity of testing $\rho = \sigma$ is maximized in the case of σ being the maximally mixed state, which is $n = \Theta(d/\epsilon^2)$ copies by Theorem 1.4.23. In the classical setting, several algorithms for distribution identity testing work by reduction to the case of uniformity testing (e.g. [BFF+01, DK16]), but these reductions do not appear to translate to the quantum setting. Other distribution identity testing algorithms use modified chi-squared tests [VV14, ADK15], and it is possible that something in this vein might be appropriate in the quantum setting as well.

10.2 Spectrum estimation

What is the copy complexity of learning ρ's spectrum α? We know that $n = O(d^2/\epsilon^2)$ copies are sufficient by Corollary 1.4.7 and that $n = \Omega(d/\epsilon^2)$ copies are necessary by Theorem 1.4.23. Furthermore, we know that the EYD algorithm requires $n = \Omega(d^2/\epsilon^2)$ copies by Theorem 1.4.10. The classical analogue of this problem is to learn the multiset of values $\{\alpha_1, \ldots, \alpha_d\}$; in other words, to output an estimate $\hat{\alpha}$ which is ϵ-close to α in the $d_{\mathrm{TV}}^{\mathrm{sym}}$ distance. By Corollary 1.3.6, this can be done with $n = O(d/\epsilon^2)$ samples. As for lower bounds, for fixed ϵ this is known to require $n = \Omega(d/\log(d))$ samples [VV11a]. We expect that there should be a way to prove a quantum version of this lower bound, showing that spectrum estimation requires $\Omega(d^2/\log(d))$ copies, though we also believe that this will be difficult to prove.

10.3 Graph isomorphism

A somewhat remarkable fact is that the standard setup for quantum algorithm for graph isomorphism has many similarities with our setup for quantum state learning. Here one is given many copies of a “coset state” ρ and is asked to determine if ρ encodes a pair of isomorphic graphs or not. Given a single copy of ρ, one is able to perform a measurement called “weak Fourier sampling”, analogous to our weak Schur sampling measurement, and to follow this measurement with a “strong Fourier sampling” measurement, analogous to our strong Scnur sampling measurement [HMR+10]. Weak Fourier sampling produces a Young diagram $\boldsymbol{\lambda}$ distributed according to Planch_n or a related distribution. It is natural to ask whether the connection between Planch_n and longest increasing subsequences, which does not appear to have been previously observed in this literature, can help give new upper or lower bounds here. For example, it is possible to reprove prior lower bounds in this area, e.g. [HRTS03], using the techniques in this thesis. We hope that our techniques might also lead to new quantum algorithms for graph isomorphism.

10.4 Miscellaneous

(1) **More spectrum learning questions:** What is the copy complexity of approximating the rank and von Neumann entropy?

(2) **Diagonality testing:** How many copies are needed to test whether ρ is diagonal in the standard basis?

(3) **Equivalence testing:** How do our techniques generalize to testing properties of two unknown mixed states? For example, how many copies are needed to test whether $\rho = \sigma$ when both ρ and σ are unknown? The answer to the corresponding classical question is given in [BFR+00].

(4) **Separability testing:** This is an open question from [MdW13]. Given a bipartite quantum state ρ, what is the copy complexity of testing whether ρ is separable?

(5) **Necessity of entangled measurements:** Do entangled measurements offer provable advantages over unentangled measurements? E.g. can one show that unentangled measurements require $\omega(d)$ copies to test mixedness?

(6) **Necessity of Schur sampling:** Is it possible to replicate any of the tight Schur-sampling-based lower/upper bounds with arguments that do not reference Schur sampling?

(7) **Distribution of the p_2^* statistic:** The statistic $p_2^*(\lambda)$ occupies a special place in our work; it appears in two of our property testing algorithms as well as in the analysis of the EYD algorithm. It is known [Ker93b, IO02] that when $\boldsymbol{\lambda}$ is sampled from the Plancherel distribution, $p_2^*(\boldsymbol{\lambda})$ is distributed as a Gaussian. Furthermore, explicit convergence rates of $p_2^*(\boldsymbol{\lambda})$ to the Gaussian distribution have been shown in [Ful05, Ful06b, SS06, Ful06a] via Stein’s-method-based arguments. Does a similar limiting statement (with explicit error bounds) hold when $\boldsymbol{\lambda} \sim \mathrm{SW}_\rho^n$ for an arbitrary mixed state ρ?

(8) **Connection to classical distribution estimation:** Is there a strong connection to classical distribution estimation, for example one that would allow us to take classical

property testing algorithms and apply them in a black-box manner to get tight quantum property testing algorithms?

(9) **Poissonization:** A central technique in distribution estimation [RRSS09, Val08] is Poissonization, in which one replace the number of samples n with a number of samples which is distributed as a Poisson random variable with parameter n. A similar Poissonization trick has found application in various limiting statements for Plancherel and Schur-Weyl distributions [BDJ99, Xu08]. Is Poissonization helpful for mixed state learning?

Bibliography

[AD99] David Aldous and Persi Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bulletin of the American Mathematical Society*, 36(4):413–432, 1999. 3.1

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam C Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems*, pages 3577–3598, 2015. 10.1

[ARS88] Robert Alicki, Sławomir Rudnicki, and Sławomir Sadowski. Symmetry properties of product states for the system of N n-level atoms. *Journal of mathematical physics*, 29(5):1158–1162, 1988. (document), 1.4.1, 1.5, 2, 3.7, 3.7, 4

[BAH$^+$16] Michael Beverland, Gorjan Alagic, Jeongwan Haah, Gretchen Campbell, Ana Maria Rey, and Alexey Gorshhkov. Implementing a quantum algorithm for spectrum estimation with alkaline earth atoms. In *19th Conference on Quantum Information Processing*, 2016. QIP 2016. 1.4.1

[Bat01] Tuğkan Batu. *Testing properties of distributions*. PhD thesis, Cornell University, 2001. 1.3.2, 1.5

[BB68] Robert Baer and Paul Brock. Natural sorting over permutation spaces. *Mathematics of Computation*, 22(102):385–410, 1968. 3.5

[BDJ99] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *Journal of the American Mathematical Society*, 12(4):1119–1178, 1999. 3.5, 3.5, 3.7, (9)

[BDJ00] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the second row of a Young diagram under Plancherel measure. *Geometric & Functional Analysis*, 10(4):702–731, 2000. 3.6.2

[BFF$^+$01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001. 1.3.2, 10.1

[BFR$^+$00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual*

IEEE Symposium on Foundations of Computer Science, pages 259–269, 2000. (document), 1.3.2, 1.3.2, (3)

[BFR⁺13] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4, 2013. 1.3.2

[Bia01] Philippe Biane. Approximate factorization and concentration for characters of symmetric groups. *International Mathematics Research Notices*, 2001(4):179–192, 2001. 3.7.2, 3.7.10, 4.3.1

[BOO00] Alexei Borodin, Andrei Okounkov, and Grigori Olshanski. Asymptotics of plancherel measures for symmetric groups. *Journal of the American Mathematical Society*, 13(3):481–515, 2000. 3.6.2

[Bro] Daniel Brown. How I wasted too long finding a concentration inequality for sums of geometric variables. Found at https://cs.uwaterloo.ca/~browndg/negbin.pdf. 8.1.4

[BS00] Sergei Bespamyatnikh and Michael Segal. Enumerating longest increasing subsequences and patience sorting. *Information Processing Letters*, 76(1):7–11, 2000. 6

[Buf12] Alexey Bufetov. A central limit theorem for extremal characters of the infinite symmetric group. *Functional Analysis and Its Applications*, 46(2):83–93, 2012. 3.7

[Can16] Clément Canonne. A survey on distribution testing: Your data is big. but is it blue?, 2016. http://www.cs.columbia.edu/~ccanonne/files/misc/2015-survey-distributions.pdf. 1.3.2

[CEM99] Juan Cirac, Artur Ekert, and Chiara Macchiavello. Optimal purification of single qubits. *Physical review letters*, 82(21):4344, 1999. 1.5

[CGS04] Sylvie Corteel, Alain Goupil, and Gilles Schaeffer. Content evaluation and class symmetric functions. *Advances in Mathematics*, 188(2):315–336, 2004. 7.2.1

[CGS11] Allison Cuttler, Curtis Greene, and Mark Skandera. Inequalities for symmetric means. *European Journal of Combinatorics*, 32(6):745–761, 2011. 5.3.1

[CHW07] Andrew Childs, Aram Harrow, and Paweł Wocjan. Weak Fourier-Schur sampling, the hidden subgroup problem, and the quantum collision problem. In *24th Annual Symposium on Theoretical Aspects of Computer Science*, pages 598–609, 2007. 1.4.2, 1.5, 2, 2.6, 3.3, 3.3, 3.7.2, 6.1, 6.2

[CM06] Matthias Christandl and Graeme Mitchison. The spectra of quantum states and the Kronecker coefficients of the symmetric group. *Communications in mathematical physics*, 261(3):789–797, 2006. 1.4.1, 1.5, 2, 4

[CP10] Maxime Crochemore and Ely Porat. Fast computation of a longest increasing subsequence and application. *Information and computation*, 208(9):1054–1059, 2010. 6

[CSST10] Tullio Ceccherini-Silberstein, Fabio Scarabotti, and Filippo Tolli. *Representation theory of the symmetric groups: the Okounkov-Vershik approach, character formulas, and partition algebras.* Cambridge University Press, 2010. 7.2.1

[DF09] Persi Diaconis and Jason Fulman. Carries, shuffling, and symmetric functions. *Advances in Applied Mathematics*, 43(2):176–196, 2009. 1.4.2

[Dia14] Ilias Diakonikolas. Beyond histograms: structure and distribution estimation. Found at http://www.iliasdiakonikolas.org/stoc14-workshop/diakonikolas.pdf, 2014. 1.3.1

[DK16] Ilias Diakonikolas and Daniel Kane. A new approach for testing properties of discrete distributions, 2016. 10.1

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer, 2001. 1.3.1

[EJ08] Funda Ergün and Hossein Jowhari. On distance to monotonicity and longest increasing subsequence of a data stream. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 730–736, 2008. 6

[ES35] Paul Erdős and George Szekeres. A combinatorial problem in geometry. *Compositio Mathematica*, 2:463–470, 1935. 3.5, 3.5.1

[Far15] Jacques Faraut. Rayleigh theorem, projection of orbital measures and spline functions. *Advances in Pure and Applied Mathematics*, 2015. 5.3.1

[Fér10] Valentin Féray. Stanley's formula for characters of the symmetric group. *Annals of Combinatorics*, 13(4):453–461, 2010. 3.8.1

[FGLE12] Steven Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012. 1.4.1, 1.4.1, 1.4.1, 5, 5.5

[FMN13] Valentin Féray, Pierre-Loïc Méliot, and Ashkan Nikeghbali. Mod-ϕ convergence I: Normality zones and precise deviations. Technical report, arXiv:1304.2934, 2013. 3.7

[Fre75] Michael L Fredman. On computing the length of longest increasing subsequences. *Discrete Mathematics*, 11(1):29–35, 1975. 6

[FRT54] James Frame, Gilbert Robinson, and Robert Thrall. The hook graphs of the symmetric group. *Canadian Journal of Mathematics*, 6:316–324, 1954. 2.2.12

[Ful97] William Fulton. *Young tableaux: with applications to representation theory and geometry.* Cambridge University Press, 1997. 9.1.6

[Ful05] Jason Fulman. Stein's method and Plancherel measure of the symmetric group. *Transactions of the American Mathematical Society*, 357(2):555–570, 2005. (7)

[Ful06a] Jason Fulman. An inductive proof of the Berry-Esseen theorem for character ratios. *Annals of Combinatorics*, 10(3):319–332, 2006. (7)

[Ful06b] Jason Fulman. Martingales and character ratios. *Transactions of the American Mathematical Society*, 358(10):4533–4552, 2006. (7)

[GG10] Anna Gál and Parikshit Gopalan. Lower bounds on streaming algorithms for approximating the length of the longest increasing subsequence. *SIAM Journal on Computing*, 39(8):3463–3479, 2010. 6

[GJKK07] Parikshit Gopalan, Thathachar Jayram, Robert Krauthgamer, and Ravi Kumar. Estimating the sortedness of a data stream. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 318–327, 2007. 6

[GKP94] Ronald Graham, Donald Knuth, and Oren Patashnik. *Concrete mathematics: a foundation for computer science*. Addison–Wesley, second edition, 1994. 7.2.4

[Gog99] Dave Goggin. The Schensted algorithm demo (1.0.2), 1999. `http://www.math.uconn.edu/~troby/Goggin/BumpingAlg.html`. 1.5

[GR11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011. 1.3.2, 1.3.2

[Gre74] Curtis Greene. An extension of Schensted's theorem. *Advances in Mathematics*, 14:254–265, 1974. 1.5, 3.2

[GT09] Otfried Gühne and Géza Tóth. Entanglement detection. *Physics Reports*, 474(1):1–75, 2009. 1

[GW09] Roe Goodman and Nolan Wallach. *Symmetry, representations, and invariants*. Springer, 2009. 2.2.13, 2.4, 2.5

[Haa16] Jeongwan Haah. Sample-optimal tomography of quantum states, 2016. `https://www.youtube.com/watch?v=1biogtHaMxw`. 1.4.1, 5.1

[Hal15] Brian Hall. *Lie groups, Lie algebras, and representations: an elementary introduction*. Springer, 2015. 2.1.1

[Ham72] John Hammersley. A few seedlings of research. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 345–394, 1972. 3.1, 3.5

[Har05] Aram Harrow. *Applications of coherent classical communication and the Schur transform to quantum information theory*. PhD thesis, Massachusetts Institute of Technology, 2005. 2.4, 2.5

[Har13] Aram Harrow. The church of the symmetric subspace. Technical report, arXiv:1308.6595, 2013. 5.1

[Har15] Aram Harrow, 2015. http://dabacon.org/pontiff/?p=10785. 1

[HE02] Paweł Horodecki and Artur Ekert. Method for direct detection of quantum entanglement. *Physical review letters*, 89(12):127902, 2002. 1

[Hep94] Charles Hepler. *On the complexity of computing characters of finite groups*. PhD thesis, University of Calgary, 1994. 3.8.1

[HGG09] Jonathan Huang, Carlos Guestrin, and Leonidas Guibas. Fourier theoretic probabilistic inference over permutations. *The Journal of Machine Learning Research*, 10:997–1070, 2009. 2.3.1

[HHJ$^+$16] Jeongwan Haah, Aram Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, August 2016. Preprint. (document), 1, 1.4.1, 1.4.1, 5, 5.1, 5.2, 5.2, 5.2.2, 5.2, 5.4, 5.5

[HHR$^+$05] Hartmut Häffner, Wolfgang Hänsel, Christian Roos, Jan Benhelm, Michael Chwalla, Timo Körber, Umakant Rapol, Mark Riebe, Piet Schmidt, Christoph Becher, Otfried Günhe, Wolfgang Dür, and Rainer Blatt. Scalable multiparticle entanglement of trapped ions. *Nature*, 438(7068):643–646, 2005. 1, 1.2

[HJ13] Roger Horn and Charles Johnson. *Matrix analysis*. Cambridge University Press, 2nd edition, 2013. 1.4.1

[HM02] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, 2002. 1.4.1, 1.5, 2, 4

[HMR$^+$10] Sean Hallgren, Cristopher Moore, Martin Rötteler, Alexander Russell, and Pranab Sen. Limitations of quantum coset states for graph isomorphism. *Journal of the ACM (JACM)*, 57(6):34, 2010. 2, 10.3

[HR18] G. H. Hardy and Srinivasa Ramanujan. Asymptotic formulae in combinatory analysis. *Proceddings of the London Mathematical Society*, 2(17):75–115, 1918. 2.2

[HRTS03] Sean Hallgren, Alexander Russell, and Amnon Ta-Shma. The hidden subgroup problem and quantum computation using group representations. *SIAM Journal on Computing*, 32(4):916–934, 2003. 10.3

[HS77] James Hunt and Thomas Szymanski. A fast algorithm for computing longest common subsequences. *Communications of the ACM*, 20(5):350–353, 1977. 6

[Hua12] Zhu Huangjun. *Quantum State Estimation and Symmetric Informationally Complete POMs*. PhD thesis, National University of Singapore, 2012. 1, 5

[HW94] Paul Hausladen and William Wootters. A 'pretty good' measurement for distinguishing quantum states. *Journal of Modern Optics*, 41(12):2385–2390, 1994. 5

[HX13] Christian Houdré and Hua Xu. On the limiting shape of Young diagrams associated with inhomogeneous random words. In *High Dimensional Probability VI*, volume 66 of *Progress in Probability*, pages 277–302. Springer Basel, 2013. 3.7, 3.7.1, 3.7.1, 3.7.1, 3.7.1, 3.7.1

[IK01] Vladimir Ivanov and Sergei Kerov. The algebra of conjugacy classes in symmetric groups and partial permutations. *Journal of Mathematical Sciences*, 107(5):4212–4230, 2001. 3.8.1, 3.8.1, 3.8.1

[IO02] Vladimir Ivanov and Grigori Olshanski. Kerov's central limit theorem for the Plancherel measure on Young diagrams. In *Symmetric functions 2001: surveys of developments and perspectives*, pages 93–151. Springer, 2002. 3.6.1, 3.8.1, 3.8.1, 3.8.1, 3.8.1, 4.3.1, 7.2.1, 7.2.2, (7)

[ITW01] Alexander Its, Craig Tracy, and Harold Widom. Random words, Toeplitz determinants and integrable systems I. In *Random Matrices and their Applications*, pages 245–258. Cambridge University Press, 2001. 3.7.1, 3.7.1, 3.7.1, 4.3

[JK81] Gordon James and Adalbert Kerber. *The representation theory of the symmetric group*. Addison–Wesley, 1981. 6.4.4

[Joh01] Kurt Johansson. Discrete orthogonal polynomial ensembles and the Plancherel measure. *Annals of Mathematics*, 153(1):259–296, 2001. 3.6.2, 3.7.1

[Ker93a] Sergei Kerov. The asymptotics of root separation for orthogonal polynomials. *Algebra i Analiz*, 5(5):68–86, 1993. 3.6

[Ker93b] Sergei Kerov. Gaussian limit for the Plancherel measure of the symmetric group. *Comptes Rendus de l'Académie des Sciences, Série 1*, 316:303–308, 1993. 3.6.1, (7)

[Key06] Michael Keyl. Quantum state estimation and large deviations. *Reviews in Mathematical Physics*, 18(01):19–60, 2006. 1.4.1, 5, 5.3, 5.3.1

[Knu70] Donald Knuth. Permutations, matrices, and generalized Young tableaux. *Pacific Journal of Mathematics*, 34(3):709–727, 1970. 1.5, 3.2

[KO94] Sergei Kerov and Grigori Olshanski. Polynomial functions on the set of Young diagrams. *Comptes Rendus de l'Académie des Sciences, Série 1*, 319(2):121–126, 1994. 3.8.1, 3.8.1

[KRT14] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. Technical report, arXiv:1410.6913, 2014. 1.4.1, 1.4.1, 5, 5.1, 5.1.4

[Kup02] Greg Kuperberg. Random words, quantum statistics, central limits, random matrices. *Methods and Applications of Analysis*, 9(1):99–118, 2002. 4, 3.7.1

[KV86] Sergei Kerov and Anatoly Vershik. The characters of the infinite symmetric group and probability properties of the Robinson-Schensted-Knuth algorithm. *SIAM Journal on Algebraic Discrete Methods*, 7(1):116–124, 1986. 3.7, 4

[KW01] Michael Keyl and Reinhard Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5):052311, 2001. (document), 1.4.1, 1.5, 2, 2.6, 3.7, 4

[Lan07] Isaiah Lankham. *Patience Sorting and Its Generalizations.* PhD thesis, University of California, Davis, 2007. 3.1

[Las78] Alain Lascoux. Classes de chern d'un produit tensoriel. *Comptes Rendus de l'Académie des Sciences, Série 1*, 286:385–387, 1978. 6.3

[Las08] Michel Lassalle. An explicit formula for the characters of the symmetric group. *Mathematische Annalen*, 340(2):383–405, 2008. 3.8.1, 7.2.1

[LNVZ06] David Liben-Nowell, Erik Vee, and An Zhu. Finding longest increasing and common subsequences in streaming data. *Journal of Combinatorial Optimization*, 11(2):155–175, 2006. 6

[LS77] Benjamin Logan and Larry Shepp. A variational problem for random Young tableaux. *Advances in Mathematics*, 26(2):206–222, 1977. 3.5, 3.6

[Mac95] Ian Macdonald. *Symmetric functions and Hall polynomials.* Oxford University Press, 1995. 2.4.1, 3, 3.8.1, 6.3, 7.2.1

[Mal62] Colin Mallows. Problem 62-2, patience sorting. *SIAM Review*, 4(2):148–149, 1962. 3.1

[Mal63] Colin Mallows. Problem 62-2, patience sorting. *SIAM review*, 5(4):375–376, 1963. 3.1

[MdW13] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. Technical report, arXiv:1310.2035, 2013. 1.4.2, 1.4.2, 2.6, (4)

[Mél10a] Pierre-Loïc Méliot. Kerov's central limit theorem for Schur-Weyl measures of parameter 1/2. Technical report, arXiv:1009.4034, 2010. 3.7.2, 3.7.2, 4.3.1, 4.3.1, 4.3.1, 7.2.4

[Mél10b] Pierre-Loïc Méliot. *Partitions aléatoires et théorie asymptotique des groupes symétriques, des algèbres d'Hecke et des groupes de Chevalley finis.* PhD thesis, University Paris-Est Marne-la-Vallée, 2010. 3.7.2, 3.8.1, 7.2.2

[Mél12] Pierre-Loïc Méliot. Fluctuations of central measures on partitions. In *24th International Conference on Formal Power Series and Algebraic Combinatorics*, pages 385–396, 2012. 3.7, 3.7.1

[MHS$^+$12] Xiao-Song Ma, Thomas Herbst, Thomas Scheidl, Daqing Wang, Sebastian Kropatschek, William Naylor, Bernhard Wittmann, Alexandra Mech, Johannes Kofler, Elena Anisimova, Vadim Makarov, Thomas Jennewein, Rupert Ursin, and Anton Zeilinger. Quantum teleportation over 143 kilometres using active feed-forward. *Nature*, 489(7415):269–273, 2012. 1, 1.2

[MOA11] Albert W Marshall, Ingram Olkin, and Barry Arnold. *Inequalities: theory of majorization and its applications*. Springer Series in Statistics, 2011. 9

[Mol09] Alexander Molev. Littlewood-Richardson polynomials. *Journal of Algebra*, 321(11):3450–3468, 2009. 6.3

[Mon09] Ashley Montanaro. Symmetric functions of qubits in an unknown basis. *Physical Review A*, 79(6):062316, 2009. 1.5

[Mon14] Ashley Montanaro. Personal communication, 2014. 1.4.1

[MS99] Alexander Molev and Bruce Sagan. A Littlewood-Richardson rule for factorial Schur functions. *Transactions of the American Mathematical Society*, 351(11):4429–4443, 1999. 6.3

[Mui02] Robert Muirhead. Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters. *Proceedings of the Edinburgh Mathematical Society*, 21:144–162, 1902. 9

[Nar06] Hariharan Narayanan. On the complexity of computing Kostka numbers and Littlewood-Richardson coefficients. *Journal of Algebraic Combinatorics*, 24(3):347–354, 2006. 2

[NC10] Michael Nielsen and Isaac Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010. 1.2.1, 1.4, 1.4.1, 5

[O'C03] Neil O'Connell. Conditioned random walks and the RSK correspondence. *Journal of Physics A: Mathematical and General*, 36(12):3049, 2003. 3.4

[O'D16] Ryan O'Donnell. Random words, longest increasing subsequences, and quantum pca, 2016. https://www.youtube.com/watch?v=qrli_ZgM4cM. 1.5

[Oko00] Andrei Okounkov. Random matrices and random permutations. *International Mathematics Research Notices*, 2000(20):1043–1095, 2000. 3.6, 3.6.2

[Oko08] Andrei Okounkov. Characters of Symmetric Groups, 2008. 3.8.1

[OO98a] Andrei Okounkov and Grigori Olshanski. Asymptotics of Jack polynomials as the number of variables goes to infinity. *International Mathematics Research Notices*, 13:641–682, 1998. 6.3

[OO98b] Andrei Okounkov and Grigori Olshanski. Shifted Schur functions. *St. Petersburg Mathematical Journal*, 9(2):239–300, 1998. 2.2.7, 3.8, 3.8.1, 3.8.1, 6.3

[OR00] Andrew Odlyzko and Eric Rains. On longest increasing subsequences in random permutations. *Contemporary Mathematics*, 251:439–452, 2000. 3.5

[OW15a] Ryan O'Donnell and John Wright. A note on the Haah et al. tomography algorithm, 2015. `http://www.cs.cmu.edu/~jswright`. 1.4.1, 1.6

[OW15b] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, 2015. 1.6, 4.1

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 2016. *To appear*, QIP 2016. 1.4.1, 1.6

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. (document), 1.3.2, 1.3.14, 6.2

[Pil90] Shaiy Pilpel. Descending subsequences of random permutations. *Journal of Combinatorial Theory, Series A*, 53(1):96–116, 1990. 3.5

[PP14] Igor Pak and Greta Panova. On the complexity of computing Kronecker coefficients. *Computational Complexity*, pages 1–36, 2014. 3.8.1

[Rob38] Gilbert de Beauregard Robinson. On the representations of the symmetric group. *American Journal of Mathematics*, 60(3):745–760, 1938. 1.5, 3.2

[Rom14] Dan Romik. *The surprising mathematics of longest increasing subsequences*. Cambridge University Press, 2014. 2.2, 3.1, 3.5, 3.7.2, 4.2.1

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009. (9)

[RS92] Ronitt Rubinfeld and Madhu Sudan. Self-testing polynomial functions efficiently and over rational domains. In *Proceedings of the 3rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 23–32, 1992. 1.3.2

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996. (document), 1.3.2

[RSW04] Victor Reiner, Dennis Stanton, and Dennis White. The cyclic sieving phenomenon. *Journal of Combinatorial Theory. Series A*, 108(1):17–50, 2004. 6.4

[RZ12] Rodolfo Ríos-Zertuche. *Near-involutions, the pillowcase distribution, and quadratic differentials*. PhD thesis, Princeton University, 2012. 6.4.6, 6.4, 6.4

[Sag01] Bruce E Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*. Springer, 2001. 2.2.12, 2.3, 2.3.1, 2.3.1, 9.1.6, 9.1

[Sch61] Craige Schensted. Longest increasing and decreasing subsequences. *Canadian Journal of Mathematics*, 13(2):179–191, 1961. 1.5, 3.1, 3.2

[Sch63] Marcel-Paul Schützenberger. Quelques remarques sur une construction de Schensted. *Mathematica Scandinavica*, 12:117–128, 1963. 9.1

[Sei59] Abraham Seidenberg. A simple proof of a theorem of Erdös and Szekeres. *Journal of the London Mathematical Society*, 1(3):352–352, 1959. 3.5

[Śni06] Piotr Śniady. Asymptotics of characters of symmetric groups, genus expansion and free probability. *Discrete mathematics*, 306(7):624–665, 2006. 3.8.1

[Sra15] Suvrit Sra. On inequalities for normalized Schur functions. *European Journal of Combinatorics*, 2015. 5.2, 5.3.1

[SS06] Qi-Man Shao and Zhong-Gen Su. The Berry-Esseen bound for character ratios. *Proceedings of the American Mathematical Society*, 134(7):2153–2159, 2006. (7)

[SS10] Michael Saks and Comandur Seshadhri. Estimating the longest increasing sequence in polylogarithmic time. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 458–467, 2010. 6

[Sta99] Richard P Stanley. *Enumerative combinatorics Volume 2*. Cambridge University Press, Cambridge, 1999. 2.2.13, 2.4.1, 2.4.1, 2.4.1, 2.4.1

[Sta11] Richard P Stanley. *Enumerative combinatorics Volume 1*. Cambridge University Press, Cambridge, 2011. 3.7.2

[Ste95] John Steele. Variations on the monotone subsequence theme of Erdös and Szekeres. In *Discrete probability and algorithms*. Springer, 1995. 3.5

[Ste11] Benjamin Steinberg. *Representation theory of finite groups: an introductory approach*. Springer Science and Business Media, 2011. 2.1, 2.1.1

[SW07] Xiaoming Sun and David Woodruff. The communication and streaming complexity of computing the longest common and increasing subsequences. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 336–345, 2007. 6

[TW01] Craig Tracy and Harold Widom. On the distributions of the lengths of the longest monotone subsequences in random words. *Probability Theory and Related Fields*, 119(3):350–380, 2001. 3.7.1, 3.7.1

[TW09] Craig Tracy and Harold Widom. The distributions of random matrix theory and their applications. In *New Trends in Mathematical Physics*, pages 753–765. Springer, 2009. 3.5.6

[Ula61] Stanislaw Ulam. Monte Carlo calculations in problems of mathematical physics. *Modern Mathematics for the Engineers*, pages 261–281, 1961. 3.5

[Val08] Paul Valiant. *Testing symmetric properties of distributions*. PhD thesis, Massachusetts Institute of Technology, 2008. 1.5, (9)

[vEB77] Peter van Emde Boas. Preserving order in a forest in less than logarithmic time and linear space. *Information processing letters*, 6(3):80–82, 1977. 6

[VK77] Anatoly Vershik and Sergei Kerov. Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux. *Soviet Mathematics Doklady*, 18:118–121, 1977. 3.5, 3.6

[VK81] Anatoly Vershik and Sergei Kerov. Asymptotic theory of characters of the symmetric group. *Functional analysis and its applications*, 15(4):246–255, 1981. 3.7, 3.8.1, 3.8.1

[VK85] Anatoly Vershik and Sergei Kerov. Asymptotic of the largest and the typical dimensions of irreducible representations of a symmetric group. *Functional Analysis and its Applications*, 19(1):21–31, 1985. 3.5, 4.2.1

[vL13] Mark van Leeuwen, 2013. `http://mathoverflow.net/a/140739/658`. 1

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 685–694, 2011. 1.3.1, 1.3.2, 10.2

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pages 403–412, 2011. 1.3.1

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, 2014. 1.3.2, 10.1

[Wal14] Michael Walter. *Multipartite quantum states and their marginals*. PhD thesis, ETH Zurich, 2014. 2.4

[Was81] Antony John Wassermann. *Automorphic actions of compact groups on operator algebras*. PhD thesis, University of Pennsylvania, 1981. 3.8.1, 3.8.1

[Xu08] Hua Xu. *Aspects of Random Matrix Theory: Concentration and Subsequence Problems*. PhD thesis, Georgia Institute of Technology, 2008. (9)