A Primer on the Statistics of Longest Increasing Subsequences and Quantum States

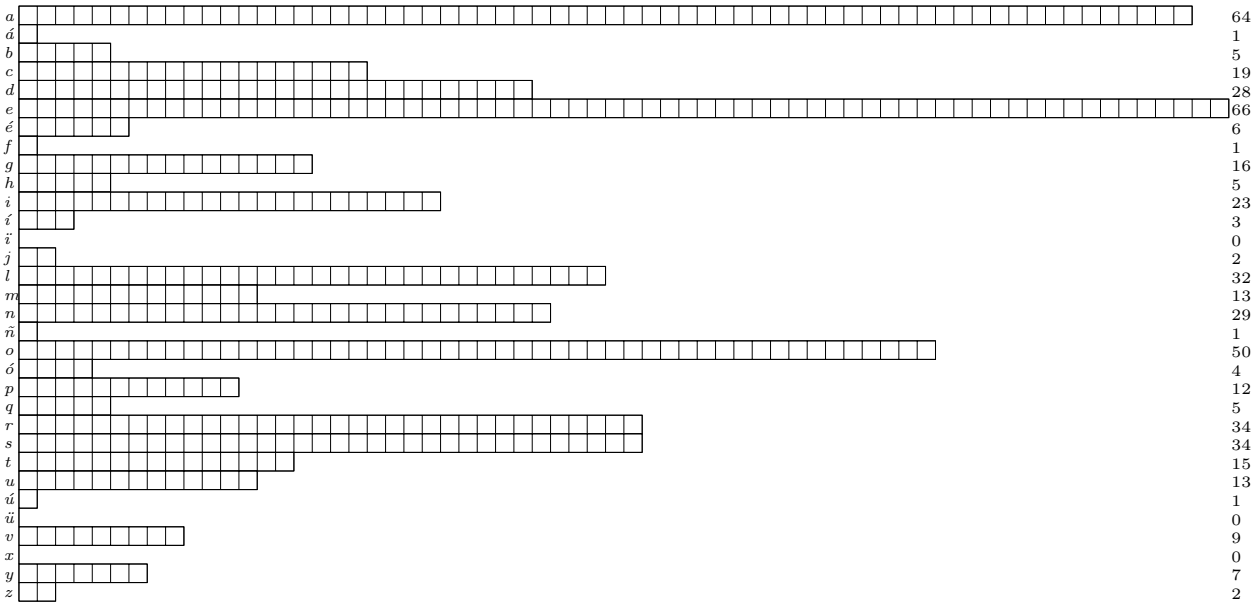
Ryan O'Donnell* John Wright†

Abstract

We give an introduction to the statistics of quantum states, with a focus on recent results giving tight bounds for the problems of learning and testing identity of mixed states. Along the way, we survey the sometimes surprising connections between this area and topics as diverse as classical distribution testing, longest increasing subsequences and the RSK algorithm, and representation theory of the symmetric and general linear groups.

1 Spanish cryptograms

Suppose you encounter a cryptogram (substitution cipher) written in Spanish. To decipher it, you'll probably want to know the frequency of letters in Spanish text. So you download *Don Quixote* [Cer15] and pick out a sample of 500 letters, drawn randomly with replacement; say, z, v, s, r, $\tilde{n}$, ..., q. The resulting histogram of 32 rows might look like this:



What can you infer from this sample? It's reasonable for you to estimate that the true frequency p_a of the letter a in Spanish is approximately $\widehat{p}_a = \frac{64}{500} = 12.8\%$. Similarly, you might estimate $\widehat{p}_{\acute{a}} = \frac{1}{500} = 0.2\%$, $\widehat{p}_b = \frac{5}{500} = 1\%$, $\widehat{p}_c = \frac{19}{500} = 3.8\%$, $\widehat{p}_d = \frac{28}{500} = 5.6\%$, $\widehat{p}_e = \frac{66}{500} = 13.2\%$, etc.[1]

*Computer Science Department, Carnegie Mellon University. Supported by NSF grant CCF-1618679. odonnell@cs.cmu.edu

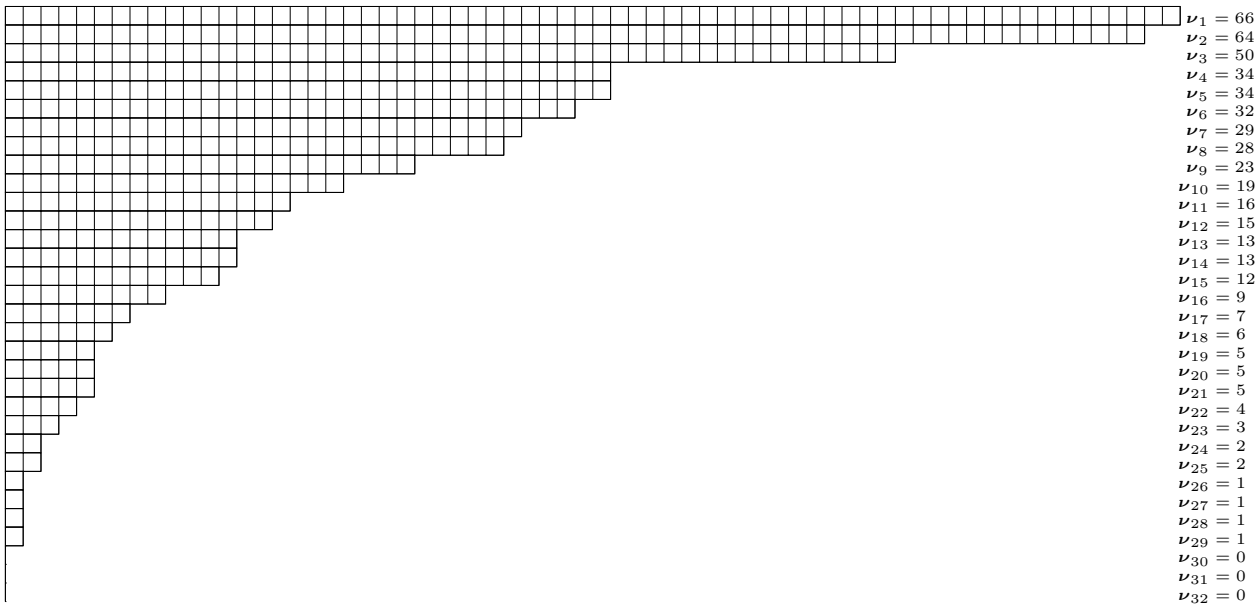
†Center for Theoretical Physics, Massachusetts Institute of Technology. Supported by NSF grant CCF-6931885. jswright@mit.edu

[1] Gaines's cryptanalysis book [Gai14] reports $p_a = 12.7\%$, $p_b = 1.4\%$, $p_c = 3.9\%$, $p_d = 5.6\%$, $p_e = 13.2\%$, ...

Of course, the finite sample size $n = 500$ means there will be some statistical error; for example, with $\widehat{p}_a \approx \widehat{p}_e \approx 13\%$, the true frequency of a and e might plausibly be anywhere between 10% and 16%. So on the basis of this sample, you would be unwise to confidently declare that e is the most probable letter in Spanish. On the other hand, it *would* be reasonable for you to conclude that *the most frequent letter has frequency* $\approx \frac{66}{500} = 13.2\%$.

This question — *What is the frequency of the most frequent letter?* — is an example of a *letter-permutation-invariant* statistic. That is, it doesn't depend on the *names* of the letters: it would be the same if you applied any of the 32! possible permutations to these names (as is done in a cryptogram). Other letter-permutation-invariant statistics include: the entropy of the letter frequencies; the total probability of the top-10 most frequent letters; the number of letters with frequency at least 1%; and so forth. In any long Spanish cryptogram, these statistics would be approximately the same. Indeed, knowing them would give you a good way to *test* whether a new cryptogram is in Spanish or some other language.

As in the Don Quixote example, suppose we form a random "word" $\boldsymbol{w} \in \{a, \ldots, z\}^n$ by sampling n letters independently; say, $\boldsymbol{w}_1 = z$, $\boldsymbol{w}_2 = v$, $\boldsymbol{w}_3 = s$, $\ldots$, $\boldsymbol{w}_n = q$. On the basis of this, we might wish to estimate some letter-permutation-invariant statistic (e.g., entropy, frequency of the most frequent letter, etc.). It's important to note that there are *two* symmetries at play. The first symmetry is the *position-permutation-invariance* of the sample; i.e., the action of the symmetric group S_n. Since the n draws are independent, it doesn't matter that z was the 1st, 107th, and 251st letter, or that v was the 48th, 133rd, 338th, and 350th; it only matters that z occurred 3 times, v occurred 4 times, etc. This is why we immediately simplified to the histogram in our example. The second symmetry is the *letter-permutation-invariance*; i.e., the action of the symmetric group S_d, where $d = 32$ is the number of letters. This symmetry says that the *names* of the letter outcomes don't matter; in other words, the statistic only depends on the (multi)set of probabilities $\{p_a, p_b, \ldots, p_z\}$. Given this, we can simplify our histogram further by eliminating the letter labels, and then *sorting* the rows. This produces a *sorted histogram* like the following:



In this sorted histogram $\boldsymbol{\nu} = \text{SortedHistogram}(\boldsymbol{w})$, the first row has length $\boldsymbol{\nu}_1 = 66$, indicating that the most frequent letter in the sample had frequency 66; the second row has length $\boldsymbol{\nu}_2 = 64$, indicating that the 2nd most frequent letter had frequency 64; etc. By virtue of the two symmetries in our problem — invariance to permuting the $n = 500$ positions, and invariance to permuting the $d = 32$ letter names — the sorted histogram $\boldsymbol{\nu}$ encodes *all* the information we need to estimate

any letter-permutation-invariant statistic, such as entropy, or the probability of the most probable letter. Indeed, if we define $\widehat{\boldsymbol{p}}_i = \boldsymbol{\nu}_i/n$, it would be reasonable to estimate these two quantities by $\sum_{i=1}^{d} \widehat{\boldsymbol{p}}_i \log(1/\widehat{\boldsymbol{p}}_i)$ and $\widehat{\boldsymbol{p}}_1$, respectively.[2]

2 Quantum contraptions

We will now introduce the "quantum" version of the "classical" statistics problem described in the previous section. Suppose you wander into a quantum computing laboratory and find a contraption with a button on the side. Every time you press the button, 5 qubits pop out of the contraption. If a 5-qubit system is in a "pure state", you can represent it as

$$\vec{a}_1|00000\rangle + \vec{a}_2|00001\rangle + \vec{a}_3|00010\rangle + \cdots + \vec{a}_{32}|11111\rangle,$$

where the numbers $\vec{a}_i$'s are complex "amplitudes" satisfying $\sum_i |\vec{a}_i|^2 = 1$. In other words, a 5-qubit pure state can be represented by a unit vector $\vec{a} \in \mathbb{C}^{32}$. (More generally, a system of q qubits has dimension $d = 2^q$, and systems with non-qubit particles may have dimensions that are not powers of 2.)

Actually, the contraption might have some *probabilistic* components inside it; for example, flipping coins, or internal quantum measurement devices. As a consequence, when you press the button, you may get some some kind of *randomly distributed* pure state vector — in other words, a quantum *mixed state*. In principle, the contraption might produce *any* probability distribution over *any* set of unit vectors in $\mathbb{C}^{32}$. However (see Section 8) it is a basic fact of quantum mechanics that we may assume, without loss of generality, that the contraption produces a discrete probability distribution over some basis of 32 *orthonormal* vectors $\vec{a}, \vec{b}, \vec{c}, \ldots \in \mathbb{C}^{32}$. Following quantum notation, let's write these unit vectors as $|1\rangle, |2\rangle, \ldots, |32\rangle \in \mathbb{C}^{32}$, and write $p_1, p_2, \ldots, p_{32}$ for the associated probabilities. In other words, every time you press the button, the contraption spits out $|i\rangle$ with probability p_i ($i = 1 \ldots 32$). Although we've numbered them $1 \ldots 32$, we may still refer to the vectors as *letters*.

Since you've never encountered the contraption before, both the probabilities p_i and the orthonormal vectors $|i\rangle$ are unknown to you. Not only that, you can't just "look at" the output vectors to tell what they are; quantum mechanics only allows you to choose a "measurement" to perform on them (discussed further in Section 8), and this measurement *itself* produces a probabilistic outcome.[3] These difficulties notwithstanding, you may press the button n times, and we'll assume that the resulting outputs are independent and unentangled. For example, if you press the button $n = 6$ times, the contraption might spit out the sequence

$$|7\rangle,\ |12\rangle,\ |4\rangle,\ |20\rangle,\ |7\rangle,\ |31\rangle;$$

this would occur with probability $p_7 \cdot p_{12} \cdot p_4 \cdot p_{20} \cdot p_7 \cdot p_{31}$. At this point, you can perform any measurement you like on the particles. *Quantum tomography* refers to the task of using the samples to estimate the mixed state of the contraption's output. In the general d-dimensional case, this (roughly speaking) means estimating the probabilities $p_1, \ldots, p_d$ and the vectors $|1\rangle, \ldots, |d\rangle$.

[2] This strategy of estimating a statistic of p by computing the statistic for the empirical distribution $\widehat{\boldsymbol{p}}$ is known as the *plug-in estimator*. Though a good baseline estimate, it is often suboptimal; see, for example, [WY16, JVHW17] for optimal entropy estimators which outperform the plug-in estimator.

[3] Although, if a "little birdie" told you the vectors $|1\rangle, \ldots, |d\rangle$, you could "measure in this basis" and thereby exactly "look at" the output vectors. This would reduce you to a classical scenario like that of sampling from unknown Spanish letter frequencies, $p_1, \ldots, p_{32}$.

As in the preceding discussion of Spanish cryptograms, for the moment we'll only concern ourselves with estimating statistics of the (multi)set of probabilities $\{p_1, \ldots, p_d\}$. Most such statistics have a natural physical meaning; for example, the largest probability gives a measure of how "pure" the contraption's output is, and the entropy $\sum_{i=1}^{d} p_i \log(1/p_i)$ is called the *von Neumann entropy* of the mixed quantum state. In this case, we again have two symmetries at play. First, we have the same position-permutation-invariance as before; i.e., the action of the symmetric group S_n. This is because the n button presses are assumed to produce independent and unentangled outcomes. Second, since we only care about statistics depending on $\{p_1, \ldots, p_d\}$ and we don't care about the identity of the orthonormal basis $|1\rangle, \ldots, |d\rangle$ of $\mathbb{C}^d$, we have the symmetry of the *unitary group* $U(d)$ acting as "rotations/reflections" on bases.

When estimating properties of the set $\{p_a, \ldots, p_z\}$ of Spanish letter frequencies, we "factored out" the S_n and S_d symmetries when we reduced our sample to its sorted histogram of n boxes and d rows. As it turns out (see Section 11) there is a similar way to "factor out" the S_n and $U(d)$ symmetries when trying to estimate properties of the probabilities $\{p_1, \ldots, p_d\}$ associated to the quantum contraption. In Section 4, we'll state an *Optimal Measurement Theorem*, which describes a certain quantum "measurement" that may be performed without loss of generality when estimating statistics of $\{p_1, \ldots, p_d\}$. Surprisingly, the possible measurement outcomes will be sorted histograms of n boxes and d rows! The reason for this has to do with the *representation theory* of the groups S_n and $U(d)$, which is intimately connected with sorted histograms — also known as *Young diagrams*.

The later sections of this survey will explain a little representation theory to justify why the Optimal Measurement Theorem is true. Before that, though, we will spend some time analyzing the probability distribution on Young diagrams that arises from the Optimal Measurement Theorem. As we'll see, this distribution is unfortunately not as simple as "draw an n-letter word $\boldsymbol{w}$ from the probability distribution $\{p_1, \ldots, p_d\}$ and form its sorted histogram". Rather, it has to do with an interesting combinatorial property of $\boldsymbol{w}$: the lengths of its *longest increasing subsequences*.

3 Longest increasing subsequences: Robinson, Schensted, Knuth

Let w be a length-n word over the ordered alphabet $\{a, b, c, d\}$; for example, suppose $n = 10$ and

$$w = dbbcdbaabc.$$

We define $\mathrm{LIS}(w)$ to be the *length of the longest increasing subsequence* of w. (Throughout, "increasing" will mean "nondecreasing"; in other words, in alphabetical order.) How can we easily determine this length? For our example $w = d\underline{bb}cd\underline{b}a\underline{abc}$, a little trial and error will convince you that the underlined subsequence $bbbbc$ is maximal, so $\mathrm{LIS}(w) = 5$. For longer words, we'll need to be more systematic.

There is a natural dynamic program for computing $\mathrm{LIS}(w)$ known as *patience sorting* that involves processing the letters of w one-by-one (see [AD99] for a survey on this topic). As we do this, we maintain a growing array L in which

$$L[j] = \text{the "alphabetically smallest" letter that can end a length-}j\text{ increasing subsequence.}$$

For example, after processing $w = dbbcdbaabc$, our array will look like

$$L = \boxed{a}\boxed{a}\boxed{b}\boxed{b}\boxed{c}.$$

This corresponds to the following five increasing subsequences:

$$\begin{aligned} L[1] &= a, \quad \text{because of } dbbcdb\underline{a}abc; \\ L[2] &= a, \quad \text{because of } dbbcdb\underline{aa}bc; \\ L[3] &= b, \quad \text{because of } dbbcdb\underline{aab}c; \\ L[4] &= b, \quad \text{because of } d\underline{bb}cd\underline{b}aa\underline{b}c; \\ L[5] &= c, \quad \text{because of } d\underline{bb}cd\underline{b}aa\underline{bc}, \end{aligned}$$

and it can be checked that there are no subsequences of length six or greater. The overall longest increasing subsequence (of the word processed so far) is simply the length of the array; and, when a new letter is processed, it's not hard to update the entries of the array. To test your understanding, you might confirm that if an 11th letter were to "arrive" at the end of our w, the four possibilities would be:

$$\begin{array}{|c|c|c|c|c|}\hline a&a&b&b&c\\\hline\end{array} + a = \begin{array}{|c|c|c|c|c|}\hline a&a&a&b&c\\\hline\end{array} \qquad \begin{array}{|c|c|c|c|c|}\hline a&a&b&b&c\\\hline\end{array} + b = \begin{array}{|c|c|c|c|c|}\hline a&a&b&b&b\\\hline\end{array}$$

$$\begin{array}{|c|c|c|c|c|}\hline a&a&b&b&c\\\hline\end{array} + c = \begin{array}{|c|c|c|c|c|c|}\hline a&a&a&b&c&c\\\hline\end{array} \qquad \begin{array}{|c|c|c|c|c|}\hline a&a&b&b&c\\\hline\end{array} + d = \begin{array}{|c|c|c|c|c|c|}\hline a&a&b&b&c&d\\\hline\end{array} \tag{1}$$

The algorithm to update the diagram (array) can be thought of as follows:

Insertion: To process a new letter, 'i', find the rightmost position in which it can be placed so as to maintain alphabetical order. If this position is already occupied by some letter, then **bump** that letter out of the diagram. Otherwise, place i at the end of the diagram, in a new box.

This rightmost position, say j, corresponds to the first entry $L[j]$ which is strictly larger than i. The update $L[j] := i$ therefore works because the subsequence ending in $L[j-1]$ can be appended with i to form a subsequence of length j; those letters to the left of $L[j]$ stay the same because they are already less than or equal to i, and those letters to the right stay the same because there is no increasing subsequence of length $j+1$ or greater ending in i. Considering the four examples in (1), we see that inserting a new 'a' causes the 'b' in the third box to be bumped; inserting a new 'b' causes the 'c' in the fifth box to be bumped; inserting a 'c' creates a new box at the end; and inserting a 'd' also creates a new box at the end. The value of $\mathrm{LIS}(w)$ increases precisely when a new box is created.

When a letter is "bumped" during the insertion process, it seems a shame to just throw it in the trash. Following an idea of **R**obinson [Rob38], **S**chensted [Sch61], and **K**nuth [Knu70] ("**RSK**"), let's instead *recursively "insert" the bumped letter into a subsequent row of the diagram.* When this RSK algorithm is applied to the word $w = dbbcdbaabc$, we get the following growing sequence of filled Young diagrams:

$$\begin{array}{|c|}\hline d\\\hline\end{array} \xrightarrow{b} \begin{array}{|c|}\hline b\\\hline d\\\hline\end{array} \xrightarrow{b} \begin{array}{|c|c|}\hline b&b\\\hline d\\\cline{1-1}\end{array} \xrightarrow{c} \begin{array}{|c|c|c|}\hline b&b&c\\\hline d\\\cline{1-1}\end{array}$$

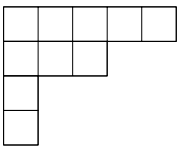
$$\xrightarrow{d} \begin{array}{|c|c|c|c|}\hline b&b&c&d\\\hline d\\\cline{1-1}\end{array} \xrightarrow{b} \begin{array}{|c|c|c|c|}\hline b&b&b&d\\\hline c\\\cline{1-1} d\\\cline{1-1}\end{array} \xrightarrow{a} \begin{array}{|c|c|c|c|}\hline a&b&b&d\\\hline b\\\cline{1-1} c\\\cline{1-1} d\\\cline{1-1}\end{array} \tag{2}$$

$$\xrightarrow{a} \begin{array}{|c|c|c|c|}\hline a&a&b&d\\\hline b&b\\\cline{1-2} c\\\cline{1-1} d\\\cline{1-1}\end{array} \xrightarrow{b} \begin{array}{|c|c|c|c|}\hline a&a&b&b\\\hline b&b&d\\\cline{1-3} c\\\cline{1-1} d\\\cline{1-1}\end{array} \xrightarrow{c} \begin{array}{|c|c|c|c|c|}\hline a&a&b&b&c\\\hline b&b&d\\\cline{1-3} c\\\cline{1-1} d\\\cline{1-1}\end{array}$$

It is not too hard to check that the RSK algorithm, when applied to any word w of length n, produces what is known as a *semistandard Young tableau of size* n: a filled n-box Young diagram in which the rows have increasing entries and the columns have *strictly* increasing entries. Because of the second property, the number of rows will never be more than the number of letters in the alphabet.

Given a semistandard Young tableau (SSYT), its *shape* is the Young diagram (sorted histogram) produced by deleting the entries. We'll write $\lambda = \text{RSKshape}(w)$ for the shape of the SSYT produced by applying the RSK algorithm to word w; thus, e.g.,

$$\text{RSKshape}(dbbcdbaabc) = \tag{3}$$



As we've seen, the top row of the diagram graphically encodes the dynamic program for determining the length of the longest increasing subsequence. Thus if $\lambda = \text{RSKshape}(w)$, then $\lambda_1 = \text{LIS}(w)$. Is there any meaning to the lengths of the subsequent rows of λ? Greene's Theorem [Gre74] implies that there is:

Greene's Theorem: *If* $\lambda = \text{RSKshape}(w)$*, then:*

$$\begin{aligned}
\lambda_1 &\text{ is the length of the longest increasing subsequence in } w;\\
\lambda_1 + \lambda_2 &\text{ is the length of the longest union of 2 increasing subsequences in } w;\\
\lambda_1 + \lambda_2 + \lambda_3 &\text{ is the length of the longest union of 3 increasing subsequences in } w;\\
\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 &\text{ is the length of the longest union of 4 increasing subsequences in } w; \quad \text{etc.}
\end{aligned}$$

For example, in our word $w = dbbcdbaabc$, Greene's Theorem and (3) tell us that w should have 2 disjoint increasing subsequences of total length $5 + 3 = 8$, and indeed here they are, underlined/overlined: $d\overline{bbcdb}\underline{aabc}$. (It's a coincidence that they're both contiguous.)

4 Symmetric properties of probabilities: classical vs. quantum

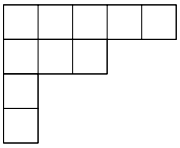
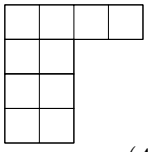
Now let's return to quantum contraptions. Suppose — as we were discussing — that we have a quantum contraption that outputs a d-dimensional mixed state with unknown probabilities $p_1, \ldots, p_{32}$ for an unknown orthonormal basis $|1\rangle, \ldots, |d\rangle$ of $\mathbb{C}^d$. (In our example, d was 32.) And suppose we want to estimate some statistic only depending on the multiset $\{p_1, \ldots, p_d\}$; for example, the maximum p_i (which we recall is one way of quantifying how "pure" the contraption's output is). We press the button n times, obtain n independent unentangled outputs, and now must make some kind of quantum measurement. As mentioned in Section 2, it is possible without loss of generality to "factor out" the S_n and $U(d)$ symmetries, yielding the following (see [CHW07, MW16, OW15, BOW17]):

Optimal Measurement Theorem: *The optimal*[4] *quantum measurement when one only cares about* $\{p_1, \ldots, p_d\}$ *has the following property: It reports an* n*-box,* d*-row Young diagram* $\boldsymbol{\lambda}$*, and the probability distribution of* $\boldsymbol{\lambda}$ *(over both the outcome of the contraption and the measurement's randomness) is exactly the same as that of* $\text{RSKshape}(\boldsymbol{w})$ *for* $\boldsymbol{w} \sim p^{\otimes n}$*, meaning that* $\boldsymbol{w}$ *is a random length-*n *word in which each letter is* $i \in \{1, \ldots, d\}$ *independently with probability* p_i*.*

[4] Vis-a-vis either of these two cases: (i) Discriminating between two classes of multisets, as in Property Testing. (ii) Estimating a statistic with minimal variance (quadratic risk).

This should be compared to the problem of estimating a letter-permutation-invariant statistic of an unknown probability distribution like the frequencies of the $d = 32$ Spanish letters. In that "classical" scenario, an optimal algorithm *also* gets an n-box, d-row random Young diagram $\boldsymbol{\nu}$; however, this $\boldsymbol{\nu}$ is simply distributed as the *sorted histogram* of a random word $\boldsymbol{w}$.

Let's make a closer comparison between the classical and quantum scenarios. In both cases, we want to use n samples to estimate a permutation-invariant property of the probability distribution $p = (p_1, \ldots, p_d)$. In both cases, we can imagine that a random word $\boldsymbol{w} \in \{1, \ldots, d\}^n$ is chosen from the product probability distribution $p^{\otimes n}$. In the classical case, we get to see the Young diagram $\boldsymbol{\nu} = \text{SortedHistogram}(\boldsymbol{w})$; in the quantum case, we get to see the "LIS information" $\boldsymbol{\lambda} = \text{RSKshape}(\boldsymbol{w})$. For example, if $\boldsymbol{w} = dbbcdbaabc$, then

$\boldsymbol{\lambda} = \text{RSKshape}(dbbcdbaabc) =$  $\qquad \boldsymbol{\nu} = \text{SortedHistogram}(dbbcdbaabc) =$  (4)

A first immediate observation is that the quantum case is at least as hard as the classical case. One way to see this is that $\boldsymbol{\nu}$ contains all the information you could ever want, whereas $\boldsymbol{\lambda}$ doesn't; another way is via Footnote 3.

A second observation is that the LIS information $\boldsymbol{\lambda}$ will always be more "top-heavy" than the sorted histogram $\boldsymbol{\nu}$. More precisely, we will always have that $\boldsymbol{\lambda}$ *majorizes* $\boldsymbol{\nu}$, written $\boldsymbol{\lambda} \succ \boldsymbol{\nu}$, meaning that $\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_k \geq \boldsymbol{\nu}_1 + \cdots + \boldsymbol{\nu}_k$ for all $1 \leq k \leq d$, with equality for $k = d$. This follows directly from Greene's Theorem, since one can always find k increasing subsequences in $\boldsymbol{w}$ whose union has length at least $\boldsymbol{\nu}_1 + \cdots + \boldsymbol{\nu}_k$, simply by taking all of the most frequently occurring letter as one subsequence, all of the 2nd-most frequently occurring letter as a 2nd subsequence, ..., all of the kth-most frequently occurring letter as the kth subsequence.

A third observation concerns symmetry with respect to permuting $\{1, \ldots, d\}$. So far we've assumed we're only interested in properties of the multiset $\{p_1, \ldots, p_d\}$, such as the maximum p_i, or the entropy of p. This is why we could reduce to the sorted histogram $\boldsymbol{\nu}$ in the classical case, and why (according to the Optimal Measurement Theorem) we can reduce to the RSK output $\boldsymbol{\lambda}$ in the quantum case. Now it's very clear that the distribution of the sorted histogram $\boldsymbol{\nu}$ is invariant to permuting $p_1, \ldots, p_d$, but it's far from clear that this is true of the RSK output $\boldsymbol{\lambda}$. In fact, it may seem almost definitely false! The very nature of the RSK algorithm, and the phrase "longest increasing subsequence", are both intimately tied up with the *ordering* on the d-letter alphabet. But nevertheless, the following surprising fact is true: The distribution on $\boldsymbol{\lambda}$ (that is, $\text{RSKshape}(\boldsymbol{w})$ for $\boldsymbol{w} \sim p^{\otimes n}$) is unchanged no matter how the probabilities $p_1, \ldots, p_d$ are permuted. The reason for this will be mentioned in Section 5, but for now you might think about the case $d = 2$, wherein $\boldsymbol{\lambda}$ is fully determined by the length of its first row, $\text{LIS}(\boldsymbol{w})$. Thus the fact says that the length of the longest increasing subsequence in a random word with 60% 1's and 40% 2's has the same distribution as in a random word with 40% 1's and 60% 2's...

Because of this symmetry property, we will sometimes assume — without loss of generality — that $p_1 \geq p_2 \geq \cdots \geq p_d$. In this case, we can combine the previous observations to get an interesting inequality. As mentioned, $\boldsymbol{\lambda}$ always majorizes the sorted histogram $\boldsymbol{\nu}$ of $\boldsymbol{w}$. In turn, the sorted histogram always majorizes the *unsorted* histogram of $\boldsymbol{\nu}$, call it $\boldsymbol{\eta}$. Taking expectations of the statement $\boldsymbol{\lambda} \succ \boldsymbol{\eta}$ yields

$$(\mathbf{E}[\boldsymbol{\lambda}_1], \mathbf{E}[\boldsymbol{\lambda}_2], \ldots, \mathbf{E}[\boldsymbol{\lambda}_d]) \succ (p_1 n, p_2 n, \ldots, p_d n), \tag{5}$$

a statement we will use several times later. These inequalities help us understand lower bounds on the λ_i's; we'd like to get some comparable upper bounds so as to really nail down the distribution

on $\boldsymbol{\lambda}$. This issue will be taken up in Section 7, but first we digress to describe a few more properties of the RSK algorithm.

5 The RSK bijection and Schur symmetric polynomials

In fact, we have so far described only *half* of the RSK algorithm. In addition to the semistandard tableau described in Section 3, known as the *insertion tableau*, the full RSK algorithm applied to a word w also maintains a second tableau known as the *recording tableau*. This tableau is updated in parallel with the insertion tableau: when the tth new box is added to the insertion tableau, a new box is added to the recording tableau in the same position, filled with "timestamp" t. In (2) we illustrated how the insertion tableau grows on the example word $w = dbbcdbaabc$; the full insertion/recording tableau output of RSK on $w = dbbcdbaabc$ would be:

$$\mathrm{RSK}(w) = \left(\begin{array}{|c|c|c|c|c|} \hline a & a & b & b & c \\ \hline b & b & d \\ \cline{1-3} c \\ \cline{1-1} d \\ \cline{1-1} \end{array} \;,\; \begin{array}{|c|c|c|c|c|} \hline 1 & 3 & 4 & 5 & 10 \\ \hline 2 & 8 & 9 \\ \cline{1-3} 6 \\ \cline{1-1} 7 \\ \cline{1-1} \end{array} \right). \tag{6}$$

For general w, the usual notation is $\mathrm{RSK}(w) = (P, Q)$, where P is the insertion tableau and Q is the recording tableau. Since the tableaus[5] always grow down-and-to-the-right, Q is always a *standard tableau*. This means that both its rows and columns are *strictly* increasing, and that it contains exactly the numbers 1 through n, where n is the length of word w.

When combined with the insertion tableau, the recording tableau gives all the additional information needed to *reverse* the steps of the RSK algorithm and thereby *invert* the RSK mapping. For example, given just the output tableaus in (6), we could recover $w = dbbcdbaabc$ as follows: First, the recording tableau tells us that the 10th and final box was created in position 5 of the first row. This can only happen if c was the final letter inserted into the first row; hence $w_{10} = c$ and

$$\mathrm{RSK}(w_1 w_2 \cdots w_9) = \left(\begin{array}{|c|c|c|c|} \hline a & a & b & b \\ \hline b & b & d \\ \cline{1-3} c \\ \cline{1-1} d \\ \cline{1-1} \end{array} \;,\; \begin{array}{|c|c|c|c|} \hline 1 & 3 & 4 & 5 \\ \hline 2 & 8 & 9 \\ \cline{1-3} 6 \\ \cline{1-1} 7 \\ \cline{1-1} \end{array} \right).$$

At this step, the recording tableau tells us that the box in position 3 of the second row was the final box created. As a result, d must have been inserted into the second row in the final step, and this could only have happened if it was previously bumped down by b in the first row. In conclusion, $w_9 = b$ and

$$\mathrm{RSK}(w_1 \cdots w_8) = \left(\begin{array}{|c|c|c|c|} \hline a & a & b & d \\ \hline b & b \\ \cline{1-2} c \\ \cline{1-1} d \\ \cline{1-1} \end{array} \;,\; \begin{array}{|c|c|c|c|} \hline 1 & 3 & 4 & 5 \\ \hline 2 & 8 \\ \cline{1-2} 6 \\ \cline{1-1} 7 \\ \cline{1-1} \end{array} \right).$$

Continuing in this manner allows us to recover the entire string w.

Remarkably, this argument shows that *any* pair of tableaus (P, Q) can be inverted into a word w so long as P is semistandard, Q is standard, and P and Q have the same shape. Hence, the RSK algorithm gives a bijection between words and pairs of tableaus, one standard and one semistandard, which we state formally below.

Before doing so, we have yet to touch on the most basic application of the RSK algorithm, which is to *permutations* rather than words. Given a permutation $\pi \in S_n$, if we write it as

[5] Sometimes spelled 'tableaux'.

$\pi = (\pi(1), \dots, \pi(n))$, then we can view it as an n-letter word on the alphabet $\{1, \dots, n\}$ which just happens to have no repetitions. As a result, if $\mathrm{RSK}(\pi) = (P, Q)$, then P has n boxes, contains each integer in $\{1, \dots, n\}$ exactly once, and is semistandard; this implies that it is in fact *standard*, like Q. Conversely, any pair of standard tableaus (P, Q) of the same shape inverts to a permutation π. As a result, the RSK algorithm also gives a bijection between permutations and pairs of standard tableaus. These two bijections are formalized as follows.

Theorem 5.1 (RSK correspondence)**.** *Given an integer n and an n-box Young diagram λ, let $\mathrm{SYT}(\lambda)$ be the set of standard Young tableaus with shape λ. Then the RSK algorithm witnesses the bijection*

$$\pi \in S_n \overset{\mathrm{RSK}}{\longleftrightarrow} (P, Q) \in \bigcup_{n\text{-box } \lambda} \mathrm{SYT}(\lambda) \times \mathrm{SYT}(\lambda). \tag{7}$$

Further, for $d \leq n$, let $\mathrm{SSYT}_d(\lambda)$ be the set of semistandard Young tableaus with shape λ and entries in $\{1, \dots, d\}$. Then the RSK algorithm witnesses the bijection

$$w \in \{1, \dots, d\}^n \overset{\mathrm{RSK}}{\longleftrightarrow} (P, Q) \in \bigcup_{n\text{-box } \lambda} \mathrm{SSYT}_d(\lambda) \times \mathrm{SYT}(\lambda). \tag{8}$$

It's customary to write $\dim \lambda = |\mathrm{SYT}(\lambda)|$ for the number of standard tableaus of shape λ. Taking cardinalities of both sides of (7), we see that

$$n! = \sum_{n\text{-box } \lambda} (\dim \lambda)^2. \tag{9}$$

(The notation $\dim \lambda$ comes from the representation theory of the symmetric group, as we'll see in Section 11. In this context, (9) is also a consequence of the decomposition of the regular representation of S_n into irreducible representations.) A conclusion is that if a permutation $\boldsymbol{\pi} \sim S_n$ is drawn uniformly at random, then $\mathbf{Pr}[\mathrm{RSKshape}(\boldsymbol{\pi}) = \lambda] = (\dim \lambda)^2 / n!$. Incidentally, there is a famous explicit formula for $\dim \lambda$, the *Hook Length formula* [FRT54]:

$$\dim \lambda = \frac{n!}{\prod_{\square \in \lambda} hl(\square)}, \quad \text{where } hl(\square) = \#\{\text{boxes in } \lambda \text{ due east and south of } \square, \text{ including } \square\}. \tag{10}$$

Analogously to (9), suppose we "count" both sides of (8) according to the product measure $p^{\otimes n}$ on words formed by a probability distribution $p = (p_1, \dots, p_d)$ on letters. The conclusion is that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{w} \sim p^{\otimes n}}[\mathrm{RSKshape}(\boldsymbol{w}) = \lambda] = s_\lambda(p) \cdot \dim \lambda, \tag{11}$$

where s_λ denotes the *Schur polynomial* indexed by λ, defined by

$$s_\lambda(x_1, \dots, x_d) = \sum_{T \in \mathrm{SSYT}_d(\lambda)} \prod_{\square \in T} x_{T(\square)}, \quad \text{where } T(\square) \text{ is the entry of tableau } T \text{ in box } \square.$$

It is a surprising and non-obvious fact that the Schur polynomials are in fact *symmetric* in the variables $x_1, \dots, x_n$. (Hint for the proof: it suffices to show that they are invariant under interchanging x_i and x_{i+1}; for this, there's a relatively simple bijection of tableaus…) Indeed, when ranging over n-box diagrams λ, they form a linear *basis* for the set of all d-variable degree-n symmetric polynomials. (We will encounter another, more familiar, such basis later in Section 10: the *power sum*

symmetric polynomials.) Finally, we mention an alternative, more compact formula for the Schur polynomials, which can be proven using some classical combinatorics (see, e.g., [Sta99, Ch. 7]):

$$s_\lambda(x_1, \dots, x_d) = \frac{\det\left(\left(x_i^{\lambda_j + d - j}\right)_{ij}\right)}{\prod_{i<j}(x_i - x_j)}. \tag{12}$$

Swapping any two variables x_s, x_t in the above formula simply creates a negative sign in the numerator and the denominator; thus this formula gives another testament that the Schur polynomials are symmetric.

6 Two majorization theorems for the RSK algorithm

In Section 4 we described a "majorization" result about the RSK algorithm that is an immediate consequence of Greene's Theorem: If w is any word with $\lambda = \text{RSKshape}(w)$ and $\nu = \text{SortedHistogram}(w)$, then $\lambda \succ \nu$, meaning that $\sum_{i=1}^{k} \lambda_i \geq \sum_{i=1}^{k} \nu_i$ for all k. In this section we mention two additonal, newly proven [OW16, OW17] majorization results concerning RSK.

The first result is highly intuitive. Suppose that $p_1 \geq p_2 \geq \cdots \geq p_d$ is a sorted probability distribution on $\{1, \dots, d\}$, and q is another. Further, suppose that $q \succ p$; roughly speaking, this means that a word $\boldsymbol{w}$ drawn randomly from $q^{\otimes n}$ tends to have more letters from "earlier in the alphabet" than if it is drawn from $p^{\otimes n}$. In either case, the sortedness of p and q ensures that the smaller letters of $\boldsymbol{w}$ tend to collect up higher in the Young diagram produced by $\text{RSK}(\boldsymbol{w})$, whereas the larger letters, outnumbered by the smaller letters, will be bumped into the lower rows. As a result, we might expect $\text{RSKshape}(\boldsymbol{w})$ to be more "top-heavy" for $\boldsymbol{w} \sim q^{\otimes n}$ than for $\boldsymbol{w} \sim p^{\otimes n}$. This is exactly what the first majorization theorem says:

Coupling Majorization Theorem [OW16]: *Let p, q be sorted probability distributions on $\{1, \dots, d\}$ with $q \succ p$. Let $\boldsymbol{\lambda} = \text{RSKshape}(\boldsymbol{w})$ for $\boldsymbol{w} \sim p^{\otimes n}$, and let $\boldsymbol{\mu} = \text{RSKshape}(\boldsymbol{z})$ for $\boldsymbol{z} \sim q^{\otimes n}$. Then there is a probabilistic coupling $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ such that $\boldsymbol{\mu} \succ \boldsymbol{\lambda}$ always. (As a consequence, $\mathbf{E}[\boldsymbol{\mu}_1 + \cdots + \boldsymbol{\mu}_k] \geq \mathbf{E}[\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_k]$ for all k.)*

Here, a probabilistic coupling $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ refers to a probability distribution on pairs of Young diagrams such that the first diagram has marginal $\boldsymbol{\lambda}$ and the second diagram has marginal $\boldsymbol{\mu}$. Although this theorem is rather intuitive, a fairly intricate bijective proof was required.

The second majorization theorem we present is concerned with the "lower rows" of the Young diagrams produced by RSK. For $\lambda = \text{RSKshape}(w)$, Greene's Theorem tells us an excellent interpretation for the length of the first row, λ_1: it's equal to $\text{LIS}(w)$. The lengths of the lower rows, though, are a little harder to interpret. Let's say we want to understand the shape of rows k and below when RSK is applied to word w. We'll take the example of $k = 2$ and our favorite word $w = dbbcdbaabc$, whose growing insertion tableau was shown in (2). We want to focus on the Young diagram formed by rows 2 and below, so we sit next to the entrance of row 2 and watch as letters come in (after being bumped from row 1). In the example (2), we see a letter d come in at "time" 2, a letter c at time 6, a letter b at time 7, another letter b at time 8, and a letter d at time 9. Let's annotate the original string w with superscripts, indicating these "times of being inserted into row 2":

$$w = d^2 b^7 b^8 c^6 d^9 baabc.$$

(To emphasize: the first d has superscript 2 because it was bumped at time 2 whereas the second d has superscript 9 because it was bumped at time 9.) In our example it's a coincidence that all

letters that made it to row 2 are from the first half of w. In any case, let's write w^{bump} for *dcbbd*, the letters that entered into row 2 *in the order they entered*; i.e., sorted according to the subscripts above. Then by definition of the RSK algorithm, the insertion tableau of RSK(w) at rows 2 and below will equal the insertion tableau of RSK(w^{bump}):

$$\xrightarrow{d} \begin{array}{|c|}\hline d \\ \hline \end{array} \xrightarrow{c} \begin{array}{|c|}\hline c \\ \hline d \\ \hline \end{array} \xrightarrow{b} \begin{array}{|c|}\hline b \\ \hline c \\ \hline d \\ \hline \end{array} \xrightarrow{b} \begin{array}{|c|c|}\hline b & b \\ \hline c \\ \cline{1-1} d \\ \cline{1-1} \end{array} \xrightarrow{d} \begin{array}{|c|c|c|}\hline b & b & d \\ \hline c \\ \cline{1-1} d \\ \cline{1-1} \end{array} = \text{RSK}(w^{\text{bump}}).$$

On the other hand, let's write w^{orig} for *dbbcd*, the letters that entered into row 2 *in the order they originally appear in* w. We could imagine applying RSK to this subsequence of w, although it's sort of counterfactual:

$$\xrightarrow{d} \begin{array}{|c|}\hline d \\ \hline \end{array} \xrightarrow{b} \begin{array}{|c|}\hline b \\ \hline d \\ \hline \end{array} \xrightarrow{b} \begin{array}{|c|c|}\hline b & b \\ \hline d \\ \cline{1-1} \end{array} \xrightarrow{c} \begin{array}{|c|c|c|}\hline b & b & c \\ \hline d \\ \cline{1-1} \end{array} \xrightarrow{d} \begin{array}{|c|c|c|c|}\hline b & b & c & d \\ \hline d \\ \cline{1-1} \end{array} = \text{RSK}(w^{\text{orig}}).$$

Comparing the shapes of the Young diagrams produced, we have

$$\text{RSKshape}(w^{\text{orig}}) = \begin{array}{|c|c|c|c|}\hline \multicolumn{1}{|c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c|}{} \\ \hline \multicolumn{1}{|c}{} & \multicolumn{1}{|c|}{} \\ \cline{1-2} \multicolumn{1}{|c}{} & \multicolumn{1}{|c|}{} \\ \cline{1-2} \end{array} \succ \begin{array}{|c|c|c|}\hline \multicolumn{1}{|c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c|}{} \\ \hline \multicolumn{1}{|c}{} & \multicolumn{1}{|c|}{} \\ \cline{1-2} \multicolumn{1}{|c}{} & \multicolumn{1}{|c|}{} \\ \cline{1-2} \end{array} = \text{RSKshape}(w^{\text{bump}}).$$

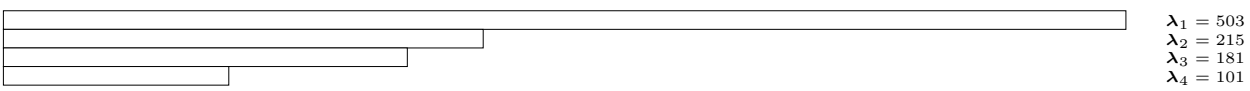
The second majorization theorem we present says that this is a general phenomenon:

Lower-Row Majorization Theorem [OW17]: *Let $w \in \{1, \dots, d\}^n$ be a word and let $1 \le k \le d$. When applying RSK to w, some of its letters enter into the kth row. Let w^{bump} denote the sequence of these letters in the order they enter, and let w^{orig} denote the sequence in the order they originally appear in w. Then* $\text{RSKshape}(w^{\text{orig}}) \succ \text{RSKshape}(w^{\text{bump}})$.

The proof of this theorem required a rather complicated analysis of a *geometric* interpretation of the RSK algorithm known as Viennot's construction [Vie81]. Unlike the Coupling Majorization Theorem, we admit to not having great intuition for the Lower-Row Majorization Theorem. (Indeed, when recalling it, we often forget whether the conclusion should have $\succ$ or $\prec$!) However, it seems to be an invaluable tool for reasoning about the lower rows produced by the RSK algorithm.

7 Probabilistic combinatorics of longest increasing subsequences

Let's return to the problem of understanding the shape of $\boldsymbol{\lambda} = \text{RSKshape}(\boldsymbol{w})$ when $\boldsymbol{w} \sim p^{\otimes n}$.[6] Throughout this section we will assume that $p_1 \ge p_2 \ge \cdots \ge p_d$, which is without loss of generality as we have discussed. For concreteness, consider the following example $\boldsymbol{\lambda}$ when $n = 1000$ and $p = (0.5, 0.2, 0.2, 0.1)$:



Interestingly, it appears that $\boldsymbol{\lambda}_i/n \approx p_i$ for each $1 \le i \le 4$, albeit with somewhat large error for $p_2 = p_3$. If this were typically true in general, then the *normalized Young diagram* $\boldsymbol{\lambda}/n =$

[6] Let us mention the distinct but related topic of understanding the shape of $\boldsymbol{\lambda} = \text{RSKshape}(\boldsymbol{\pi})$ when $\boldsymbol{\pi} \sim S_n$ is a random permutation, in which case $\boldsymbol{\lambda}$ is said to have *Plancherel* distribution. The Plancherel distribution is extremely well-studied; see [IO02] for an excellent overview of this area.

$(\boldsymbol{\lambda}_1/n, , \ldots, \boldsymbol{\lambda}_d/n)$ would provide us with a good estimate $(\widehat{p}_1, \ldots, \widehat{p}_n)$ of the sorted probability distribution $(p_1, \ldots, p_d)$. In turn, this would let us estimate any statistic of the multiset $\{p_1, \ldots, p_d\}$. So how might we show that $\boldsymbol{\lambda}_i \sim p_i n$ for large n?

Let's start with the $i = 1$ case. As described in Section 3, $\boldsymbol{\lambda}_1 = \mathrm{LIS}(\boldsymbol{w})$, so we'd like to show that the longest increasing subsequence in $\boldsymbol{w} \sim p^{\otimes n}$ has length roughly $p_1 n$. The lower bound is simple: indeed, we already determined (see (5)) that $\mathbf{E}[\boldsymbol{\lambda}_1] \geq p_1 n$. This is because $\mathrm{LIS}(\boldsymbol{w})$ is always at least the number of 1's in $\boldsymbol{w}$, a quantity with mean $p_1 n$.

Let's now heuristically reason about an upper bound for $\boldsymbol{\lambda}_1 = \mathrm{LIS}(\boldsymbol{w})$. The longest increasing subsequence in $\boldsymbol{w}$ can always be determined as follows: First, take some partition of the positions $(1, \ldots, n)$ into d contiguous blocks, $B_1, \ldots, B_d$. Next, form an increasing sequence in $\boldsymbol{w}$ by taking all of the letter-1's in block B_1, all of the letter-2's in block B_2, and so forth. Finally, maximize this procedure over all partitions into blocks. Now for any partition, the number of letters i that $\boldsymbol{w}$ has in block B_i will be tightly concentrated around $p_i|B_i|$. Thus the length of the increasing subsequence of $\boldsymbol{w}$ formed from the partition should be not much more than $p_1|B_1| + p_2|B_2| + \cdots + p_d|B_d|$. But $p_i \leq p_1$ for all i, so this is at most $p_1(|B_1| + \cdots + |B_d|) = p_1 n$. Indeed, one can formalize this argument using the Chernoff bound and get that $\mathrm{LIS}(\boldsymbol{w}) \leq p_1 n + O(d\sqrt{n}\log n)$ with high probability. We will later see a noticeably tighter upper bound.

The fact that indeed $\boldsymbol{\lambda}_i \sim p_i n$ for all $i \in [d]$ was first shown by Vershik and Kerov [VK81]. Since then, several works have determined that in the limit as $n \to \infty$, the deviation of the normalized Young diagram $\boldsymbol{\lambda}/n$ from the probability vector p is distributed like a random vector arising from the spectrum of certain *random matrix ensembles*; specifically, it has a partly *Gaussian*, partly *Tracy–Widom* limiting distribution; this was first shown for the case of uniform p_i's by [Ker03, TW01, Joh01] and generalized to the case of nonuniform p_i's by [ITW01, HX13, MÍ2]. Unfortunately, these limiting results don't necessarily give us concrete error bounds on the deviations of $\boldsymbol{\lambda}_i$ from $p_i n$: they heavily rely on considering p fixed and then taking $n \to \infty$. In particular, the error bounds can have an uncontrolled dependence on quantities like d and $\min_{p_i \neq p_j}(p_i - p_j)^{-1}$.

Still, it is very useful to rely on these results for intuition. Most useful has been the following ansatz, which is suggested by these limiting results.

$$\textbf{Ansatz: } \boldsymbol{\lambda}_i \approx p_i n \pm 2\sqrt{p_i d_i n}.$$

Here d_i is the number of occurrences of p_i in $(p_1, \ldots, p_d)$.

One of the main goals in [OW16, OW17] is to prove sharp, explicit bounds on the closeness of the normalized Young diagram $\boldsymbol{\lambda}/n$ to the sorted probability vector p. For instance, in Section 12, we sketch a proof of the ℓ_2-bound

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}}\big[\|\boldsymbol{\lambda}/n - p\|_2^2\big] \leq \frac{d}{n},$$

which is indeed consistent with the ansatz. Going beyond this single global error bound, [OW17] was able to show some *per-row* error bounds, which help in analyzing the Hellinger distance and χ^2-divergence of $\boldsymbol{\lambda}/n$ from p. The easiest to state such bound is the following:

$$p_i n - 2\sqrt{\tau_i n} \leq \mathbf{E}[\boldsymbol{\lambda}_i] \leq p_i n + 2\sqrt{\tau_i n}, \tag{13}$$

where $\tau_i = \min\{1, p_i d\}$. We note that this is suggested by the ansatz, as $p_i d_i \leq \tau_i$ always. In the remainder of this section, we will sketch the proof of the upper bound in (13).

Our starting point is the fact that much stronger asymptotics can be obtained in case the largest probability p_1 is noticeably larger than the second-largest probability p_2. For example, in a long sequence of random English letters, the longest increasing subsequence will almost surely be essentially the same as the number of e's; thus its distribution will be very close to having mean $p_e n$

and standard deviation $\sqrt{p_e(1-p_e)n}$. On the other hand, in Spanish, where $p_a \approx p_e$, the longest increasing subsequence may involve a mix of a's and e's, and its length has a greater chance of deviating noticeably above $p_a n \approx p_e n$.

To make this observation more formal, let $\boldsymbol{w}^{(\infty)} = \boldsymbol{w}_1\boldsymbol{w}_2\boldsymbol{w}_3\cdots$ be an infinite random word with each $\boldsymbol{w}_i \sim p$ independently, and set $\boldsymbol{w}^{(n)} = \boldsymbol{w}_1 \cdots \boldsymbol{w}_n$ to be its length-n prefix. Consider the (indefinite) process of performing RSK on $\boldsymbol{w}^{(\infty)}$, and let $\boldsymbol{\lambda}^{(n)} = \mathrm{RSKshape}(\boldsymbol{w}^{(n)})$ be the "snapshot" of the RSK shape at time n. Then Its, Tracy, and Widom [ITW01] showed that

$$\mathbf{E}[\boldsymbol{\lambda}_1^{(n)}] - p_1 n \quad \xrightarrow{n\to\infty} \quad \sum_{i>1} \frac{p_i}{p_1 - p_i}. \tag{14}$$

The limiting quantity on the right is finite if and only if $p_1 > p_2$ strictly. Supposing that $p_1 - p_2 \geq \delta$, its value is at most $\sum_{i=2}^{d} \frac{p_i}{\delta} \leq 1/\delta$. So at an intuitive level, (14) tells us that in a random length-n word with letter probabilities satisfying $p_1 \geq p_2 + \delta$, the expected length of the longest increasing subsequence is just an additive $1/\delta$ larger than the expected length $p_1 n$ of the all-1's subsequence.

Unfortunately, (14) is merely a limiting statement; it could be true that $\mathbf{E}[\boldsymbol{\lambda}_1^{(n)}] - p_1 n$ only becomes smaller than, say, $2/\delta$ once $n \geq 2^d \cdot 2^{1/\delta}$ — or even worse. Indeed, the proof of (14) in [ITW01] involves asymptotic hacking on the explicit formula (11) (using formulas (10) and (12)) and it heavily relies on d and $\min_{p_i \neq p_j}(p_i - p_j)^{-1}$ being treated as "constant" while $n \to \infty$. However, the combinatorial RSK perspective allows us a nice trick which lets us convert these heavily asymptotic statements to perfectly concrete ones.

The trick is to show that

$$\mathbf{E}[\boldsymbol{\lambda}_1^{(n)}] - p_1 n \text{ is an } \textit{increasing} \text{ function of } n; \qquad \text{i.e.,} \quad \mathbf{E}[\boldsymbol{\lambda}_1^{(n+1)} - \boldsymbol{\lambda}_1^{(n)}] \geq p_1. \tag{15}$$

To show this, let $\boldsymbol{\delta}^{(n+1)} = \boldsymbol{\lambda}_1^{(n+1)} - \boldsymbol{\lambda}_1^{(n)}$. By definition, $\boldsymbol{\delta}^{(n+1)}$ is the 0/1 indicator random variable for the event that, in the infinite RSK process, inserting letter $\boldsymbol{w}_{n+1}$ creates a new box in the first row. Thus $\mathbf{E}[\boldsymbol{\delta}^{(n+1)}]$ is the probability of this event, and we need to show the probability is at least p_1.

To show this, we recall that the RSK output distribution depends only on the multiset $\{p_1, \ldots, p_d\}$, and not on the ordering of the letters; hence, we can "reverse the alphabet" to $1 > 2 > \cdots > d$ without changing the distribution of $\boldsymbol{\lambda}^{(t)}$ for any t. But upon doing this, it becomes evident that the probability that the $(n+1)$th box is in the first row is at least p_1. This is because we get a new box in the first row whenever $\boldsymbol{w}_{n+1} = 1$ (which is now the *last* letter "in alphabetical order").

Thus we have established (15). But now we have an increasing sequence, $\mathbf{E}[\boldsymbol{\lambda}_1^{(n)}] - p_1 n$, and we know its limiting value thanks to (14). This means that the limiting value must be an upper bound for *all* n! That is,

$$\mathbf{E}[\boldsymbol{\lambda}_1^{(n)}] - p_1 n \leq \sum_{i>1} \frac{p_i}{p_1 - p_i}, \quad \text{for all } n. \tag{16}$$

This already gives the upper bound we desire for (13) in the case when $p_1 \geq p_2 + \frac{2}{\sqrt{n}}$. However it can become arbitrarily bad when p_2 gets close to p_1, and it gives nothing at all when $p_1 = p_2$. To get around this, we would like to slightly "shift" some probability mass of p onto p_1 so that: (i) the expected LIS is not changed too much; and, (ii) there is a decent separation between p_1 and p_2. Formally, let $\delta = \frac{1}{\sqrt{n}}$, and construct a sorted probability distribution $q = (q_1, \ldots, q_d)$ with $q_1 = p_1 + \delta$, $q_2 \leq p_2$, and $q \succ p$. (This q can be constructed by simply moving the bottom δ-mass of p onto p_1. We note q cannot be constructed if $p_1 > 1 - \delta$, but in this case our desired

bound is trivially true.) We can now apply the Coupling Majorization Theorem from Section 6 (indeed, just the last statement in it, with $k = 1$.) Using the notation from that theorem, we have

$$\mathbf{E}[\boldsymbol{\lambda}_1] \leq \mathbf{E}[\boldsymbol{\mu}_1] \leq q_1 n + \sum_{i>1} \frac{q_i}{q_1 - q_i} \leq q_1 n + \sqrt{n} = \left(p_1 + \frac{1}{\sqrt{n}}\right) n + \sqrt{n} = p_1 n + 2\sqrt{n}, \tag{17}$$

as stated in (13).

Next, we would like to generalize this to get a similar upper bound on $\mathbf{E}[\boldsymbol{\lambda}_k]$ for any row $1 \leq k \leq d$. For this we use the Lower-Row Majorization Theorem. For $\boldsymbol{w} \sim p^{\otimes n}$, it tells us that

$$\boldsymbol{\lambda}_k = \mathrm{RSKshape}(\boldsymbol{w})_k = \mathrm{RSKshape}(\boldsymbol{w}^{\mathrm{bump}})_1 \leq \mathrm{RSKshape}(\boldsymbol{w}^{\mathrm{orig}})_1 = \mathrm{LIS}(\boldsymbol{w}^{\mathrm{orig}}).$$

Analyzing $\boldsymbol{w}^{\mathrm{orig}}$ directly still seems difficult, because it still requires understanding which letters are bumped to the kth row. However, all the letters bumped into the kth row are at least k. Hence $\boldsymbol{w}^{\mathrm{orig}}$ is a subsequence of $\boldsymbol{w}^{\geq k}$, the subsequence of $\boldsymbol{w}$ formed by removing all letters less than k. Because adding letters cannot decrease the longest increasing subsequence, we have that $\mathrm{LIS}(\boldsymbol{w}^{\mathrm{orig}}) \leq \mathrm{LIS}(\boldsymbol{w}^{\geq k})$. But $\boldsymbol{w}^{\geq k}$ is simple to analyze: it's distributed exactly as $p_{\geq k}^{\otimes \boldsymbol{m}}$, where $\boldsymbol{m} \sim \mathrm{Binomial}(n, p_k + \cdots + p_d)$ and $p_{\geq k}$ is the probability distribution $\frac{1}{p_k + \cdots + p_d}(p_k, \ldots, p_d)$. So we can conclude that

$$\mathbf{E}[\boldsymbol{\lambda}_k] \leq \mathbf{E}[\mathrm{LIS}(\boldsymbol{w}^{\geq k})] \leq \mathop{\mathbf{E}}_{\boldsymbol{m}}[(p_{\geq k})_1 \boldsymbol{m} + 2\sqrt{\boldsymbol{m}}] \leq p_k n + 2\sqrt{(p_k + \cdots + p_d)n},$$

where the second inequality uses (17) and Jensen's inequality. As $p_k + \cdots + p_d \leq \tau_k$, we get the claimed upper bound in (13).

8 Mechanics of quantum mechanics

We have not yet given any justification for the Optimal Measurement Theorem, which concerns a certain quantum measurement that outputs Young diagrams. Now is the time to delve into the mathematics of quantum states and measurements.

In the physical world, a "quantum measurement" is a device that takes in a quantum particle system (of some fixed dimension D) and outputs some classical information. Its output should always be considered a random variable. Even when the input is a deterministic pure state vector $v \in \mathbb{C}^D$, the output will be randomly distributed (in a well-defined way, based on the device itself and the input state v). And on top of this, we will consider measuring quantum contraption outputs, which themselves are randomized.

Speaking of quantum contraptions, we imagined a scenario where, at the push of a button, the contraption outputs a d-dimensional state which is one of the orthonormal vectors $|1\rangle, \ldots, |d\rangle \in \mathbb{C}^d$ with probabilities $p_1, \ldots, p_d$. In an effort to learn about these vectors and probabilities, we have considered pushing the button n times. Suppose $d = 32$, $n = 6$, and the output is the sequence

$$v_1,\ v_2,\ v_3,\ v_4,\ v_5,\ v_6.$$

Here each v_t is one of $|1\rangle, \ldots, |32\rangle \in \mathbb{C}^{32}$ — although we don't yet know these basis vectors. One thing we might do is build some cleverly chosen measuring device M that accepts 32-dimensional inputs and reads out some classical information. We could then apply it to each of $v_1, \ldots, v_6$. A more sophisticated thing to do is build 6 different measuring devices, $M_1, \ldots, M_6$, each taking a 32-dimensional input, and apply M_t to v_t, $t = 1 \ldots 6$. An even more sophisticated strategy might involve adaptivity — we could build and apply different 32-dimensional measuring devices based

on the outcomes of previous measurements. However the *most* sophisticated thing we could do is build a single measurement device $\mathcal{M}$ that takes as input *all* 6 *samples simultaneously.*

If you think of a single $v_t \in \mathbb{C}^{32}$ as the state of 5 qubits, then collectively $v_1, \ldots, v_6$ represent the state of $5 \times 6 = 30$ qubits. This in turn is defined by some $2^{30} = (2^5)^6$-dimensional vector. In general, if we have n "unentangled" d-dimensional systems with pure states $v_1, \ldots, v_n \in \mathbb{C}^d$, then their state is defined by a vector of dimension $D = d^n$. Specifically, it is the vector $v_1 \otimes v_2 \otimes \cdots \otimes v_n \in (\mathbb{C}^d)^{\otimes n}$. This situation is least complicated when each vector v_t is one of d orthonormal possibilities $|1\rangle, \ldots, |d\rangle$, as we have been considering. In that case, $(\mathbb{C}^d)^{\otimes n}$ should be thought of as the vector space spanned by d^n vectors that, by fiat, are orthonormal and are named

$$|i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_n\rangle, \quad i_t \in \{1, \ldots, d\}.$$

For typographical simplicity, we usually write these vectors simply as $|i_1 i_2 \cdots i_n\rangle$, where $i_1 i_2 \cdots i_n$ ranges over all "words" in $\{1, \ldots, d\}^n$. So if, e.g., we have a contraption with 4-dimensional outputs $|1\rangle, |2\rangle, |3\rangle, |4\rangle \in \mathbb{C}^4$, and we press its button twice, the possible outputs are 4^2 orthonormal vectors in $(\mathbb{C}^4)^{\otimes 2}$ named

$$|11\rangle,\ |12\rangle,\ |13\rangle,\ |14\rangle,\ |21\rangle,\ |22\rangle,\ |23\rangle,\ |24\rangle,\ |31\rangle,\ |32\rangle,\ |33\rangle,\ |34\rangle,\ |41\rangle,\ |42\rangle,\ |43\rangle,\ |44\rangle.$$

Let's return to the notion of measurement devices for a D-dimensional particle system. One of the most general kind of measurement devices works as follows. Let f be an ordered orthonormal basis ("frame") $|f_1\rangle, \ldots, |f_D\rangle$ for $\mathbb{C}^D$. Then we can build a measurement device M_f that, on input a pure state $|v\rangle \in \mathbb{C}^D$, produces the following classical read-outs:

$$\text{"}j\text{" with probability } |\langle f_j|v\rangle|^2 = \langle f_j|v_i\rangle\langle v_i|f_j\rangle, \quad j = 1 \ldots D.$$

Here we are using the "bra-ket" notation in which $|f_j\rangle$ and $|v\rangle$ denote column vectors, and $\langle f_j|$ denotes the (complex conjugate-)transposed row vector of $|f_j\rangle$. So $\langle f_j|v\rangle = \langle f_j||v\rangle$ is just the usual inner-product of $|f_j\rangle$ and $|v\rangle$, the number $|\langle f_j|v\rangle|^2 = \langle f_j|v\rangle\,\overline{\langle f_j|v\rangle} = \langle f_j|v_i\rangle\langle v_i|f_j\rangle$ is its squared magnitude, and the fact that these quantities sum to 1 is a consequence of the Pythagorean theorem (and that all the vectors involved have unit length).

We have described what happens when a "pure state" $|v\rangle$ is fed into M_f. What happens if we feed in a *randomly chosen* pure state? Specifically, say we have a "mixed state" $\mathcal{R}$, meaning a probability distribution over some pure states $|v_1\rangle, \ldots, |v_r\rangle$, in which outcome $|v_i\rangle$ occurs with probability q_i. Here the $|v_i\rangle$'s are arbitrary unit vectors in $\mathbb{C}^D$, and r might be more or less than D. If we make a draw from $\mathcal{R}$, feed the result into the measurement device M_f, and observe the outcome, what do we see? We get

$$\text{"}j\text{" with probability } \sum_{i=1}^{r} q_i |\langle f_j|v_i\rangle|^2 = \sum_{i=1}^{r} q_i \langle f_j|v_i\rangle\langle v_i|f_j\rangle = \langle f_j| \left(\sum_{i=1}^{r} q_i |v_i\rangle\langle v_i| \right) |f_j\rangle. \tag{18}$$

Notice that these probabilities only depend on the $D \times D$ *matrix*

$$\sigma = \sum_{i=1}^{r} q_i |v_i\rangle\langle v_i|. \tag{19}$$

This matrix σ is called the *density matrix* for the mixed state $\mathcal{R}$, and we see that two mixed states $\mathcal{R}$ and $\mathcal{R}'$ with the same density matrix produce identical measurement outcomes, and thus *cannot be distinguished by* any *measurement devices* M_f! Accordingly, two such mixed states are

considered physically identical, and they're mathematically represented by the same object, the density matrix σ.

As an example, let $D = 2$ and define the unit vectors

$$|a\rangle = \begin{bmatrix}1\\0\end{bmatrix}, \qquad |b\rangle = \begin{bmatrix}0\\1\end{bmatrix}, \qquad |a'\rangle = \begin{bmatrix}3/5\\4/5\end{bmatrix}, \qquad |b'\rangle = \begin{bmatrix}-4/5\\3/5\end{bmatrix}.$$

Now if we define the mixed state $\mathcal{R}$ = "$|a\rangle$ or $|b\rangle$ with probability $\frac{1}{2}$ each" and the mixed state $\mathcal{R}'$ = "$|a'\rangle$ or $|b'\rangle$ with probability $\frac{1}{2}$ each", they both have the same density matrix, namely

$$\sigma = \frac{1}{2}\begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}3/5\\4/5\end{bmatrix}\begin{bmatrix}3/5 & 4/5\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-4/5\\3/5\end{bmatrix}\begin{bmatrix}-4/5 & 3/5\end{bmatrix} = \begin{bmatrix}1/2 & 0\\0 & 1/2\end{bmatrix} = \frac{1}{2}\mathbb{1}, \tag{20}$$

where $\mathbb{1}$ denotes the identity matrix. In particular, suppose an engineer designs and builds a quantum contraption with 1-qubit ($D = 2$) output given by $\mathcal{R}$. Then a statistician wanders into the lab, presses the contraption's button several times, and estimates its output as $\mathcal{R}'$. At first it might look like the statistician estimated the probabilities $p_1 = p_2 = \frac{1}{2}$ perfectly but the vectors $|a\rangle, |b\rangle$ poorly, since $|a'\rangle$, $|b'\rangle$ look quite different. But in fact the statistician should be given full points for a 100% correct estimate! It *only* makes sense to try to estimate the density matrix σ of an unknown mixed state, and the quality of an estimate matrix $\widehat{\sigma}$ should be measured in terms of some matrix-distance between σ and $\widehat{\sigma}$.

Let's summarize some properties of a D-dimensional density matrix σ, all of which follow from (19). First, σ is *positive-semidefinite*, meaning that it is Hermitian (equal to its complex conjugate transpose $\sigma^\dagger$) and that $\langle g|\sigma|g\rangle \geq 0$ for all vectors $|g\rangle \in \mathbb{C}^D$. Second, σ has *trace* $\mathrm{tr}(\sigma)$ equal to 1, where the trace is the sum of ρ's diagonal entries. An easy way to see this is to use the *linearity* of trace, $\mathrm{tr}(cA + B) = c\,\mathrm{tr}(A) + \mathrm{tr}(B)$, and the *cyclic property* of trace, $\mathrm{tr}(AB) = \mathrm{tr}(BA) = \sum_{i,j=1}^{D} A_{ij}B_{ji}$. Applying these to (19) gives

$$\mathrm{tr}(\sigma) = \mathrm{tr}\left(\sum_{i=1}^{r} q_i |v_i\rangle\langle v_i|\right) = \sum_{i=1}^{r} q_i\,\mathrm{tr}(|v_i\rangle\langle v_i|) = \sum_{i=1}^{r} q_i\,\mathrm{tr}(\langle v_i|v_i\rangle) = \sum_{i=1}^{r} q_i\,\mathrm{tr}\big(\big[\,1\,\big]\big) = \sum_{i=1}^{r} q_i = 1.$$

Since σ is positive-semidefinite, it will always have an orthonormal basis of eigenvectors, call them $|1\rangle, \ldots, |D\rangle$, with associated nonnegative eigenvalues, call them $p_1, \ldots, p_D \geq 0$. Further, the trace of a matrix equals the sum of its eigenvalues.[7] Thus $p_1 + \cdots + p_D = 1$, we can view the p_i's as a probability distribution over the eigenvectors $|i\rangle$, and $\sigma = \sum_{i=1}^{D} p_i |i\rangle\langle i|$. In particular, every positive-semidefinite matrix of trace 1 corresponds to a mixed state over d orthonormal pure state outcomes, justifying a claim made in Section 2.

Please note that for a given density matrix σ, its *spectrum* — i.e., the *multiset* of eigenvalues $\{p_1, \ldots, p_D\}$ — is uniquely determined, but it doesn't have an inherent ordering. Furthermore, corresponding orthonormal eigenvectors $|1\rangle, \ldots, |D\rangle$ are *not* uniquely determined. Taking the example from (20), we see that the 2-dimensional density matrix $\sigma = \frac{1}{2}\mathbb{1}$ has eigenvalues $(\frac{1}{2}, \frac{1}{2})$, but for associated eigenvectors we can choose literally *any* pair of orthonormal vectors in $\mathbb{C}^2$. The D-dimensional analogue of this state, $\sigma = \frac{1}{D}\mathbb{1}$, is called the *maximally mixed state*; it is the unique state with spectrum corresponding to the uniform probability distribution $(\frac{1}{D}, \ldots, \frac{1}{D})$.

Let's make a final observation of relevance for quantum contraptions. Suppose a quantum contraption outputs $|1\rangle, \ldots, |d\rangle$ with probabilities $p_1, \ldots, p_d$, and hence has density matrix $\rho =$

[7] This follows because trace is *unitarily invariant*: $\mathrm{tr}(U\sigma U^\dagger) = \mathrm{tr}(\sigma U^\dagger U) = \mathrm{tr}(\sigma\mathbb{1}) = \mathrm{tr}(\sigma)$ for any unitary U. Choosing U to be a unitary matrix that moves the orthonormal basis $|1\rangle, \ldots, |D\rangle$ to the standard basis of $\mathbb{C}^D$, we get that $U\sigma U^\dagger$ is a diagonal matrix with $p_1, \ldots, p_D$ on the diagonal, and the claim follows.

$\sum_{i=1}^d p_i |i\rangle\langle i|$. If we hit its button n times and view the output collectively, we get $|w\rangle \in (\mathbb{C}^d)^{\otimes n}$ with probability $\mathbf{Pr}_{p^{\otimes n}}[w]$, where w runs over all words $i_1 i_2 \cdots i_n \in \{1, \ldots, d\}^n$. *This* probability distribution on pure states has density matrix

$$\sigma = \sum_{i_1,\ldots,i_n=1}^{d} \left(\prod_{t=1}^{n} p_{i_t}\right) \left(\bigotimes_{t=1}^{n} |i_t\rangle\right) \left(\bigotimes_{t=1}^{n} \langle i_t|\right) = \bigotimes_{t=1}^{n} \left(\sum_{i=1}^{d} p_i |i\rangle\langle i|\right) = \bigotimes_{t=1}^{n} \rho = \rho^{\otimes n}, \tag{21}$$

where $\otimes$ also denotes the matrix Kronecker product. Thus quantum tomography problems can be thought of as *estimating properties of a density matrix ρ given the ability to measure $\rho^{\otimes n}$.*

9 Noncommutative probability

Before thinking about measurements of the bigger state $\rho^{\otimes n}$, let's first discuss measuring a single density matrix $\rho \in \mathbb{C}^{d \times d}$. Measurement can be thought of as a way of generating classical random outcomes from a "base source" of quantum randomness, namely a positive $d \times d$ matrix ρ with trace 1. In this section we'll consistently make an analogy to a similar situation in classical probability: generating classical random outcomes from a "base source" of classical randomness, namely a probability distribution $p \in \mathbb{R}^d$ (which is a vector of positive numbers adding to 1). Indeed, if you restrict attention to *diagonal* density matrices ρ, the two situations become identical.

So far we have seen that, given ρ, you can generate d classical random outcomes with an M_f measurement, where $f = (|f_1\rangle, \ldots, |f_d\rangle)$ is an orthonormal basis of $\mathbb{C}^d$. Let's write the resulting outcome probabilities from (18) in a slightly different way, using the cyclic property of trace (and the fact that the trace of a single number is itself):

$$\text{measuring } \rho \text{ with } M_f \text{ yields outcome "}j\text{" with probability } \langle f_j|\rho|f_j\rangle = \operatorname{tr}(\langle f_j|\rho|f_j\rangle) = \operatorname{tr}(\rho|f_j\rangle\langle f_j|).$$

Writing $E_j = |f_j\rangle\langle f_j|$ (the matrix which projects onto f_j), the above is $\operatorname{tr}(\rho E_j)$, which also equals $\operatorname{tr}(\rho^\dagger E_j)$ because $\rho^\dagger = \rho$. Now the trace of a matrix product $X^\dagger Y$ is the same as the entrywise dot-product between matrices X and Y:

$$\operatorname{tr}(X^\dagger Y) = \sum_{i,j=1}^{d} (X^\dagger)_{ij} Y_{ji} = \sum_{i,j=1}^{d} X^*_{ji} Y_{ji} = \langle X, Y\rangle,$$

where we use the $\langle \cdot, \cdot \rangle$ notation for matrix dot-product. Thus we can further write:

$$\text{measuring } \rho \text{ with } M_f \text{ yields outcome "}j\text{" with probability } \langle \rho, E_j\rangle, \quad E_j = |f_j\rangle\langle f_j|.$$

This can be compared with the simplest way of generating classical outcomes given a classical base source of randomness $p \in \mathbb{R}^d$: namely, simply drawing from p and reporting the outcome. If we do this, the probability of outcome j is $\langle p, e_j\rangle$, where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the jth coordinate.

So far we have only used our base sources of randomness (ρ or p) to generate outcomes from the set $\{1, \ldots, d\}$. In the classical case, we could generate outcomes in some other set Ω as follows: First, draw $\boldsymbol{j}$ from p. Next, add some additional coin flips $\boldsymbol{x}$. Then form a final outcome $\boldsymbol{\omega} \in \Omega$ via some deterministic function h of $\boldsymbol{j}$ and $\boldsymbol{x}$. A similar thing is possible in the quantum case: First, draw $|v\rangle$ from ρ. Next, add some additional qubits initialized to, say, $|0\rangle$, thereby increasing the dimension to D. Next, perform a measurement M_F using some D-dimensional frame F, producing an outcome $\boldsymbol{J} \in \{1, \ldots, D\}$. Lastly, form a final outcome $\boldsymbol{\omega} \in \Omega$ by applying a deterministic

function $h : \{1, \ldots, D\} \to \Omega$ to $\boldsymbol{J}$. This whole process — call it $\mathcal{M}$ — can be viewed as a "generalized measurement" of ρ, with outcomes in Ω. And it turns out that this is the *most general* kind of measurement allowed by the laws of quantum mechanics. As an example, the measurement described in the Optimal Measurement Theorem can be thought of as a general measurement of $\rho^{\otimes n}$ with outcome set Ω equal to the collection of all n-box, d-row Young diagrams.

There is a relatively simple way to mathematically describe any such general measurement $\mathcal{M}$ (which, in quantum lingo, is called a "POVM"). A little calculation shows that, corresponding to any $\mathcal{M}$, there exist positive-semidefinite matrices $E_1, E_2, \ldots, E_{|\Omega|} \in \mathbb{C}^{d \times d}$ satisfying $\sum_{\omega \in \Omega} E_\omega = \mathbb{1}$, such that

$$\text{measuring } \rho \text{ with } \mathcal{M} \text{ yields outcome } \omega \text{ with probability } \langle \rho, E_\omega \rangle.$$

In case $\mathcal{M}$ is of the basic type M_f, the matrices E_ω are just $|f_j\rangle\langle f_j|$, $1 \leq j \leq d$. Again, we can compare these general measurements to the classical case. If we let $e_1, e_2, \ldots, e_{|\Omega|}$ be any nonnegative vectors in $\mathbb{R}^d$ with $\sum_{\omega \in \Omega} e_\omega = (1, 1, \ldots, 1)$, then we can use a base probability distribution $p \in \mathbb{R}^d$ to generate outcome ω with probability $\langle p, e_\omega \rangle$. In both scenarios, we have a useful special case: a two-outcome measurement, or equivalently, a probabilistic *event*. In the classical case, if $e \in \mathbb{R}^d$ satisfies $0 \leq e \leq (1, \ldots, 1)$, we can think of it as an "event" that occurs with probability $\langle p, e \rangle$. Similarly, in the quantum case, if $E \in \mathbb{C}^{d \times d}$ satisfies $0 \preceq E \preceq \mathbb{1}$ (in the positive-semidefinite ordering), we can think of E as an "event" that occurs with probability $\langle \rho, E \rangle$ (arising from the two-outcome measurement with outcomes $\{0, 1\}$ and matrices $E_1 = E$, $E_0 = \mathbb{1} - E$).

We can also describe the quantum analogue of real-valued random variables, called *observables*. If $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we can form a classical real random variable $\boldsymbol{x}$ from a probability distribution $p \in \mathbb{R}^d$ by taking $\boldsymbol{x}$ to have value x_j with probability p_j. The expectation of this random variable is $\langle p, x \rangle$. In the quantum case, suppose we associate the real values $x_1, \ldots, x_d$ to the outcomes of a basic measurement M_f with frame $|f_1\rangle, \ldots, |f_d\rangle$. This yields a real random variable $\boldsymbol{x}$ in which value x_j occurs with probability $\left\langle \rho, |f_j\rangle\langle f_j| \right\rangle$. The expectation of this random variable is

$$\sum_{j=1}^{d} \left\langle \rho, |f_j\rangle\langle f_j| \right\rangle x_j = \left\langle \rho, \sum_{j=1}^{d} x_j |f_j\rangle\langle f_j| \right\rangle = \langle \rho, X \rangle, \quad \text{where } X = \sum_{j=1}^{d} x_j |f_j\rangle\langle f_j|.$$

Here the "observable" X is a $d \times d$ Hermitian matrix, with eigenvalue/vector pairs $x_j, |f_j\rangle$; conversely, to any Hermitian X we can associate a real-valued random variable using its eigenvalue/vector pairs. Notice also that if we *square* all the values x_j, we get the eigenvalue/vectors of the Hermitian matrix X^2. In other words, the expected value of $\boldsymbol{x}^2$ is $\langle \rho, X^2 \rangle$. Given these observations, it's natural to introduce — for any Hermitian ("observable") $X \in \mathbb{C}^{d \times d}$ — the notations

$$\mathbf{E}_\rho[X] = \langle \rho, X \rangle, \qquad \mathbf{Var}_\rho[X] = \mathbf{E}_\rho[X^2] - \mathbf{E}_\rho[X]^2, \qquad \mathbf{stddev}_\rho[X] = \sqrt{\mathbf{Var}_\rho[X]}.$$

Some familiar properties hold: for example, $\mathbf{E}[cX + Y] = c\mathbf{E}[X] + \mathbf{E}[Y]$, and $\mathbf{E}[\mathbb{1}] = 1$, and $\mathbf{Var}_\rho[X] \geq 0$. The main thing to watch out for is that observables need not commute! In fact, $XY = YX$ occurs if and only if the product XY is itself an observable (i.e., Hermitian); thus $\mathbf{E}_\rho[XY] = \mathbf{E}_\rho[YX]$ holds whenever it is "well-defined". As a general substitute for XY, one can sometimes use the always-Hermitian matrix $\frac{1}{2}(XY + YX)$. Incidentally, though it's irrelevant for this survey, you might try proving as an exercise the famous *Heisenberg uncertainty principle* (in Robertson's form [Rob29]): for all observables X, Y,

$$\mathbf{stddev}_\rho[X] \cdot \mathbf{stddev}_\rho[Y] \geq |\mathbf{E}_\rho[\tfrac{i}{2}(XY - YX)]|.$$

10 Testing for the uniform distribution/maximally mixed state

Let's return to the problem of estimating properties of an unknown quantum contraption; in other words, estimating properties of an unknown density matrix $\rho \in \mathbb{C}^{d\times d}$ given the ability to measure n samples, $\sigma = \rho^{\otimes n}$. As in classical statistical testing, we focus on finding tests with good error guarantees while keeping n as small as possible. Recalling Section 2 we may think of ρ as a mixed state, outputting one of $|1\rangle, \dots, |d\rangle$ with probabilities $p_1, \dots, p_d$, where $|1\rangle, \dots, |d\rangle$ is an unknown orthonormal basis of $\mathbb{C}^d$, and the probabilities p_i are also unknown.

To begin, we'll focus on testing whether ρ is the *maximally mixed state*, $\frac{1}{d}\mathbb{1}$, mentioned near the end of Section 8; in other words, testing whether ρ's spectrum, the multiset $\{p_1, \dots, p_d\}$, is $\{\frac{1}{d}, \dots, \frac{1}{d}\}$. This is the quantum analogue of the classical problem of testing whether an unknown probability distribution is the uniform distribution (see, e.g., [GR00, Pan08]).

The basic idea behind testing whether a probability distribution is uniform is to estimate the degree-2 *power sum symmetric polynomial*, $\mathrm{pow}_2(p) = \sum_{i=1}^d p_i^2$.[8] This expression is called the *purity* of ρ in the quantum case, and the *collision probability* of p in the classical case. The latter term refers to the fact that $\mathrm{pow}_2(p) = \mathbf{Pr}_{\boldsymbol{w}\sim p^{\otimes 2}}[\boldsymbol{w}_1 = \boldsymbol{w}_2]$, the probability that two independent draws from p yield the same letter. This quantity is minimized when p is the uniform distribution, with minimal value $\frac{1}{d}$. (Also, it has maximal value 1 when p is "pure"; i.e., $p_i = 1$ for some i.) Furthermore, $\mathrm{pow}_2(p)$ is close to minimal if and only if p is close to uniform: specifically,

$$\mathrm{pow}_2(p) - \tfrac{1}{d} = \delta_p, \text{ where } \delta_p := \|p - \tfrac{1}{d}\mathbb{1}\|_2^2 \text{ is the } \ell_2^2\text{-distance between } p \text{ and the uniform distribution.} \tag{22}$$

Let's work our way up to the quantum case by first studying the classical case. A natural way to estimate $\mathrm{pow}_2(p)$ in the classical case is simply to draw an n-letter word $\boldsymbol{w} \sim p^{\otimes n}$ and compute the random variable

$$\boldsymbol{c}_{(2)} := \operatorname*{avg}_{1\le s\neq t\le n} \{1[\boldsymbol{w}_s = \boldsymbol{w}_t]\}, \quad \text{which has } \mathbf{E}[\boldsymbol{c}_{(2)}] = \operatorname*{avg}_{s\neq t}\{\mathbf{Pr}[\boldsymbol{w}_s = \boldsymbol{w}_t]\} = \mathrm{pow}_2(p).$$

In statistics parlance, $\boldsymbol{c}_{(2)}$ is an *unbiased estimator* of $\mathrm{pow}_2(p)$, and hence $\boldsymbol{c}_{(2)} - \frac{1}{d}$ is an unbiased estimate of δ_p. It's only a small chore to explicitly compute $\mathbf{E}[\boldsymbol{c}_{(2)}^2]$ and hence $\mathbf{Var}[\boldsymbol{c}_{(2)}]$ in terms of $\mathrm{pow}_2(p)$ and $\mathrm{pow}_3(p)$ (the latter being the probability that 3 letters drawn from p are all equal):

$$\begin{aligned}\mathbf{Var}[\boldsymbol{c}_{(2)}] &= \frac{1}{\binom{n}{2}}(\mathrm{pow}_2(p) - \mathrm{pow}_2(p)^2) + \frac{2(n-2)}{\binom{n}{2}}(\mathrm{pow}_3(p) - \mathrm{pow}_2(p)^2) \\ &\le O\left(\frac{\delta_p}{n^2} + \frac{1}{dn^2} + \frac{\delta_p^{3/2}}{n} + \frac{\delta_p}{dn}\right),\end{aligned} \tag{23}$$

where the inequality used (22), some arithmetic, and $\sum_i \gamma_i^3 \le (\sum_i \gamma_i^2)^{3/2}$. If we fix a threshold $\theta \le 1$ and set $n = K\max\{\theta^{-1}d^{-1/2}, \theta^{-1/2}\}$ with K a large constant, then $\mathbf{Var}[\boldsymbol{c}_{(2)}] \le .0001\max\{\delta_p^2, \theta^2\}$. Of course, $\mathbf{Var}[\boldsymbol{c}_{(2)}]$ is also the variance of $\boldsymbol{c}_{(2)} - \frac{1}{d}$, whose mean is δ_p. Summarizing:

Theorem 10.1. *With $n = O(\max\{\theta^{-1}d^{-1/2}, \theta^{-1/2}\})$ samples, the estimator $\boldsymbol{c}_{(2)} - \frac{1}{d}$ has mean equal to δ_p (the ℓ_2^2-distance of p from uniform), and standard deviation at most $.01\max\{\delta_p, \theta\}$; hence by Chebyshev we can use it to decide (with high confidence) whether $\delta_p \le .9\theta$ or $\delta_p \ge \theta$.*

[8] The usual notation for this is $p_k(x)$; however, this clashes with our notation $p_1, \dots, p_d$ for probabilities. The works [OW16, OW17, BOW17] evade this clash by writing $\alpha_1, \dots, \alpha_d$ for probabilities.

In the language of Property Testing, this gives a $(.9\theta, \theta)$-*tolerant testing algorithm* for p being ℓ_2^2-close to the uniform distribution. Noting that $\sqrt{d\delta_p}$ is an upper bound on the total variation distance of p from uniformity (by Cauchy–Schwarz), we can set $\theta = \epsilon^2/d$ and immediately derive an algorithm which tests whether p is the uniform distribution or has total variation distance at least ϵ from it, using $n = O(\sqrt{d}/\epsilon^2)$. Such a result was first obtained by in [CDVV14, VV14], and it was later obtained using the collision tester in [DGPP16].

In a very similar way, we can estimate how close a quantum contraption's density matrix $\rho \in \mathbb{C}^{d\times d}$ is to the maximally mixed state $\frac{1}{d}\mathbb{1}$, in ℓ_2^2-distance (also known as squared *Frobenius*, or *Hilbert–Schmidt*, distance). We remark that this distance is again

$$\langle \rho - \tfrac{1}{d}\mathbb{1}, \rho - \tfrac{1}{d}\mathbb{1}\rangle = \langle \rho, \rho\rangle - \tfrac{2}{d}\operatorname{tr}(\rho) + \tfrac{1}{d} = \operatorname{tr}(\rho^2) - \tfrac{1}{d} = \sum_{i=1}^{d} p_i^2 - \tfrac{1}{d} = \operatorname{pow}_2(p) - \tfrac{1}{d} = \delta_p.$$

Again, we'd like a small-variance unbiased estimator for $\operatorname{pow}_2(p)$, but now it must be an *observable*. Let's warm up by considering the case $n = 2$, writing $\sigma = \rho^{\otimes 2}$. We are looking for an observable Hermitian operator X, acting on $(\mathbb{C}^d)^2$, such that $\mathbf{E}_\sigma[X] = \langle \sigma, X\rangle = \operatorname{pow}_2(p)$. In the classical case with $n = 2$, the random variable $\boldsymbol{c}_{(2)}$ involved drawing $\boldsymbol{w} \sim p^{\otimes 2}$ and checking if $\boldsymbol{w}_1 = \boldsymbol{w}_2$. Another way to view this is checking whether $\boldsymbol{w}_2\boldsymbol{w}_1 = \boldsymbol{w}_1\boldsymbol{w}_2$; i.e., whether swapping the two letters produces the same word. This suggests letting X be the operator on $(\mathbb{C}^d)^{\otimes 2}$ that "swaps the two tensor components": $X(|u\rangle \otimes |v\rangle) = |v\rangle \otimes |u\rangle$. Let's denote X by $\mathcal{P}_{(1\,2)}$, for reasons we'll see in Section 11. Note that this swapping operator is Hermitian, and it can be defined independently of any basis for $\mathbb{C}^d$ (which is good, because a contraption-tester doesn't have any fixed basis in mind). That said, it's advantageous for analysis to consider $\mathcal{P}_{(1\,2)}$ in the natural tensor basis of ρ's eigenvalues $|1\rangle, \ldots, |d\rangle$: it becomes a $d^2 \times d^2$ permutation matrix, and it maps $|i_1 i_2\rangle$ to $|i_2 i_1\rangle$ for any $1 \le i_1, i_2 \le d$. In other words, $\mathcal{P}_{(1\,2)} = \sum_{i_1,i_2} |i_2 i_1\rangle\langle i_1 i_2|$. Recalling (21), we also have $\rho^{\otimes 2} = \sum_{i_1,i_2} p_{i_1} p_{i_2} |i_1 i_2\rangle\langle i_1 i_2|$. Thus in the $d^2 \times d^2$ matrix dot-product $\langle \rho^{\otimes 2}, \mathcal{P}_{(1\,2)}\rangle$, we only get contributions when $i_1 = i_2$. Specifically,

$$\mathbf{E}_{\rho^{\otimes 2}}[\mathcal{P}_{(1\,2)}] = \langle \rho^{\otimes 2}, \mathcal{P}_{(1\,2)}\rangle = \sum_{i=1}^{d} p_i^2 = \operatorname{pow}_2(p). \tag{24}$$

Thus the observable $\mathcal{P}_{(1\,2)}$ is an "unbiased estimator" for the quantum purity. As for its variance, $\mathbf{E}_{\rho^{\otimes 2}}[\mathcal{P}_{(1\,2)}^2] = \mathbf{E}_{\rho^{\otimes 2}}[\mathbb{1}] = 1$, so the variance is $\mathcal{P}_{(1\,2)}$ is $1 - \operatorname{pow}_2(p)^2$, which is not very small. But of course we have only used $n = 2$ so far.

As with the definition of the estimator $\boldsymbol{c}_{(2)}$, we can drive down the variance by taking larger n and averaging over all possible $\binom{n}{2}$ transpositions. So let's define an observable on $(\mathbb{C}^d)^{\otimes n}$ by

$$\mathcal{C}_{(2)} = \operatorname*{avg}_{1\le s\ne t\le n} \{\mathcal{P}_{(s\,t)}\}, \text{ where } \mathcal{P}_{(s\,t)} \text{ acts on } (\mathbb{C}^d)^{\otimes n} \text{ by swapping the } s\text{th and } t\text{th tensor components.}$$

Although each $\mathcal{P}_{(1\,2)}$ here is defined on $(\mathbb{C}^d)^{\otimes n}$ rather than $(\mathbb{C}^d)^{\otimes 2}$, the expectation computation (24) still holds. $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{P}_{(s\,t)}]$ equals the probability that a random word $\boldsymbol{w} \sim p^{\otimes n}$ satisfies $\boldsymbol{w}^{(s\,t)} = \boldsymbol{w}$, where $\boldsymbol{w}^{(s\,t)}$ denotes the word $\boldsymbol{w}$ with its sth and tth letters swapped, and this is indeed $\operatorname{pow}_2(p)$. Thus $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}] = \operatorname{pow}_2(p)$.

As for the computation of $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}^2]$ and hence $\mathbf{Var}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}]$, it's nearly identical to that of $\mathbf{E}[\boldsymbol{c}_{(2)}^2]$. We will do it in some detail in Section 11, but to be brief, here: Upon squaring $\mathcal{C}_{(2)}$, we get three kinds of contributions, arising from the three cycle-types that can arise from the product $(s\,t)(s'\,t')$ of two transpositions: either the identity, a 3-cycle, or the product of disjoint 2-cycles. When we compute expectations, these contributions yield 1, $\operatorname{pow}_3(p)$, and $\operatorname{pow}_2(p)^2$, respectively.

The only difference from the classical case comes when $\{s,t\} = \{s',t'\}$, wherein we have $\mathcal{P}_{(s\,t)}^2 = \mathbb{1}$ in comparison with the classical $1[\boldsymbol{w}_s = \boldsymbol{w}_t]^2 = 1[\boldsymbol{w}_s = \boldsymbol{w}_t]$. In the end, very similarly to (23), we get

$$\mathbf{Var}[\mathcal{C}_{(2)}] = \frac{1}{\binom{n}{2}}(1 - \mathrm{pow}_2(p)^2) + \frac{2(n-2)}{\binom{n}{2}}(\mathrm{pow}_3(p) - \mathrm{pow}_2(p)^2). \tag{25}$$

The fact that this is a bit worse (larger) than in the classical case actually makes the parameters simpler; in the expression following (23) the first two terms get replaced by $O(\frac{1}{n^2})$, and it suffices to bound the second two terms by $O(\frac{\delta_p}{n})$; thus $\mathbf{Var}[\boldsymbol{c}_{(2)}] \le O(\frac{1}{n^2} + \frac{\delta_p}{n})$. Then taking $n = K/\theta$ with K a large constant we get $\mathbf{Var}[\boldsymbol{c}_{(2)} - \frac{1}{d}\mathbb{1}] = \mathbf{Var}[\boldsymbol{c}_{(2)}] \le .0001\max\{\delta_p^2, \theta^2\}$ again. And once again, $\sqrt{d\delta_p}$ is an upper bound on the matrix ℓ_1-distance (or *trace distance*) between ρ and the maximally mixed state, by a matrix form of Cauchy–Schwarz. We can therefore obtain tolerant testers for whether ρ is close to the maximally mixed state:

Theorem 10.2. *Given $n = O(1/\theta)$ samples of $\rho \in \mathbb{C}^{d\times d}$, we can decide (with high confidence) whether $\delta_p \le .9\theta$ or $\delta_p \ge \theta$. As a consequence, given $n = O(d/\epsilon^2)$ samples, we can decide (with high confidence) whether ρ is the maximally mixed state $\frac{1}{d}\mathbb{1}$ or has trace distance at least ϵ from it.*

Theorem 10.2 was first obtained in [OW15], but the viewpoints described in this section are from [BOW17]. The sample complexity $n = O(d/\epsilon^2)$ in the theorem is tight: [OW15] proved that even distinguishing "$\rho = \frac{1}{d}\mathbb{1}$" from "$\rho$ has eigenvalues $\{\frac{1+\epsilon}{d}, \frac{1-\epsilon}{d}, \frac{1+\epsilon}{d}, \frac{1-\epsilon}{d}, \dots, \frac{1+\epsilon}{d}, \frac{1-\epsilon}{d}\}$" requires $\Omega(d/\epsilon^2)$ samples. (Previously, [CHW07] had shown this statement, and therefore an $\Omega(d)$ lower bound, when $\epsilon = 1$.) Elaborating on the techniques used to prove Theorem 10.2, [BOW17] also showed tight results for testing identity of ρ to *any* fixed density matrix, with respect to "infidelity" and other distance measures.

11 Representation theory gives a nice basis for observables

Let's go over the variance computation for the quantum purity estimator $\mathcal{C}_{(2)}$ in a more expansive fashion. We defined $\mathcal{C}_{(2)}$ as the average of $\mathcal{P}_{(s\,t)}$ over all transpositions $(s\,t) \in S_n$, where $\mathcal{P}_{(s\,t)}$ acts on $(\mathbb{C}^d)^{\otimes n}$ by transposing the sth and tth tensor components. More generally, for any permutation $\pi \in S_n$ we could define the operator $\mathcal{P}_\pi$ that acts by permuting the n tensor components according to π. We have $\mathcal{P}_\pi\mathcal{P}_{\pi'} = \mathcal{P}_{\pi\pi'}$; in other words, $\mathcal{P}$ is a *representation*[9] of the symmetric group S_n on the vector space $(\mathbb{C}^d)^{\otimes n}$. We may then define, for any "cycle type" κ of permutations in S_n,

$$\mathcal{C}_\kappa = \operatorname*{avg}_{\pi \text{ of cycle type } \kappa} \{\mathcal{P}_\pi\}.$$

For example, if κ is the cycle type $(4,3)$, then $\mathcal{C}_\kappa$ is the average of all operators $\mathcal{P}_\pi$ where π is the product of a 4-cycle and a (disjoint) 3-cycle. Incidentally, when we speak of cycle types, we generally don't write cycles of length 1; strictly speaking we should, in which case the cycle type $(4,3)$ would be more properly written as $(4,3,1,1,\dots,1)$, with the number of 1's being $n-7$. When all 1's are included and the parts are sorted, a cycle type is nothing more than a *partition* of n; i.e., a Young diagram. We also mention that the cycle types are in correspondence with the *conjugacy classes* of the symmetric group S_n.

Now for a density matrix $\rho \in \mathbb{C}^{d\times d}$ with spectrum $p_1,\dots,p_d$, we saw that the expectation $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{P}_{(s\,t)}]$ equals the probability that a random word $\boldsymbol{w} \sim p^{\otimes n}$ is invariant to transposing the sth

[9] It would be more common to see the notation $\mathcal{P}(\pi)\mathcal{P}(\pi') = \mathcal{P}(\pi\pi')$, but we used subscripts instead to avoid writing things like $\mathcal{P}((s\,t))$.

and tth letters — i.e., that it satisfies $\boldsymbol{w} = \boldsymbol{w}^{(s\,t)}$. This is just $\mathbf{Pr}[\boldsymbol{w}_s = \boldsymbol{w}_t] = \sum_{i=1}^d p_i^2 = \mathrm{pow}_2(p)$. Since this is the same for every transposition $(s\,t)$, we of course also have $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}] = \mathrm{pow}_2(p)$.

More generally, let κ be a cycle type for S_n and define the generalized power sum symmetric polynomial

$$\mathrm{pow}_\kappa(p) = \mathrm{pow}_{\kappa_1}(p) \cdot \mathrm{pow}_{\kappa_2}(p) \cdot \mathrm{pow}_{\kappa_3}(p) \cdots$$

(Note that it doesn't matter whether or not we include the 1-cycles in the cycle type κ, since $\mathrm{pow}_1(p) = 1$ anyway.) Now if $\pi \in S_n$ has cycle type κ, it is not hard to see that[10]

$$\mathbf{E}_{\rho^{\otimes n}}[\mathcal{P}_\pi] = \mathop{\mathbf{Pr}}_{\boldsymbol{w} \sim p^{\otimes n}}[\boldsymbol{w} = \boldsymbol{w}^\pi] = \mathrm{pow}_\kappa(p); \quad \text{hence } \mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_\kappa] = \mathrm{pow}_\kappa(p). \tag{26}$$

Here $\boldsymbol{w}^\pi$ is the word formed from $\boldsymbol{w}$ by permuting its n positions according to π. To illustrate this with an example, let's take $\kappa = (4, 3)$ again. Suppose π is of this cycle type, say $\pi = (1\,2\,3\,4)(5\,6\,7)$. Then $\boldsymbol{w} = \boldsymbol{w}^\pi$ if and only if the first 4 letters of $\boldsymbol{w}$ are the same and also the 5th, 6th, and 7th letters are the same. These two (independent) events occur with probability $\sum_{i=1}^d p_i^4 = \mathrm{pow}_4(p)$ and $\sum_{i=1}^d p_i^3 = \mathrm{pow}_3(p)$, respectively, and hence the probability both occur is indeed $\mathrm{pow}_{(4,3)}(p)$.

To compute the variance of the purity estimator $\mathcal{C}_{(2)}$, we needed to first compute $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}^2]$, where

$$\mathcal{C}_{(2)}^2 = \mathop{\mathrm{avg}}_{(s\,t),(s'\,t')} \{\mathcal{P}_{(s\,t)}\mathcal{P}_{(s'\,t')}\}.$$

The product of two uniformly random transpositions in S_n is either the identity (probability $1/\binom{n}{2}$), a 3-cycle (probability $2(n-2)/\binom{n}{2}$), or of cycle type $(2, 2)$ (probability $\binom{n-2}{2}/\binom{n}{2}$). Hence

$$\mathcal{C}_{(2)}^2 = \tfrac{1}{\binom{n}{2}} \cdot \mathbb{1} + \tfrac{2(n-2)}{\binom{n}{2}} \cdot \mathcal{C}_{(3)} + \tfrac{\binom{n-2}{2}}{\binom{n}{2}} \cdot \mathcal{C}_{(2,2)}. \tag{27}$$

Therefore

$$\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_{(2)}^2] = \frac{1}{\binom{n}{2}} + \frac{2(n-2)}{\binom{n}{2}}\mathrm{pow}_3(p) + \frac{\binom{n-2}{2}}{\binom{n}{2}} \cdot \mathrm{pow}_{(2,2)}(p),$$

from which (25) follows (note that $\mathrm{pow}_{(2,2)}(p) = \mathrm{pow}_2(p)^2$).

Let's look more closely at these "cycle type observables" $\mathcal{C}_\kappa$. One thing to note is that they commute: $\mathcal{C}_\kappa\mathcal{C}_{\kappa'} = \mathcal{C}_{\kappa'}\mathcal{C}_\kappa$ for any two cycle types; in fact, it's not hard to show that $\mathcal{C}_\kappa$ commutes with *every* $\mathcal{P}_\pi$. (Indeed the collection $\{\mathcal{C}_\kappa\}_\kappa$ is a basis for the "center of the group algebra $\mathbb{C}S_n$".) Let's define

$$\mathcal{A} = \{\text{real linear combinations of the observables } \mathcal{C}_\kappa\};$$

the right-hand side of (27) is an example element of $\mathcal{A}$. This $\mathcal{A}$ is not only a (real) vector space of dimension equal to the number of cycle types (conjugacy classes) of S_n, it has a (compatible) commutative multiplication operation. Thus it is a commutative *algebra* over the reals.

It's not a coincidence that the observable $\mathcal{C}_{(2)}$ we used to estimate the quantum purity $\sum_i p_i^2$ is a member of $\mathcal{A}$. Suppose we have a quantum contraption with output $\rho \in \mathbb{C}^{d \times d}$ and we've come up with some observable X on $(\mathbb{C}^d)^{\otimes n}$ with a certain expectation $\mu = \mathbf{E}_{\rho^{\otimes n}}[X]$. (E.g., we might be trying to estimate a statistic μ of ρ's eigenvalues p; or, perhaps we are trying to decide if the multiset p has a certain property, and X's eigenvalues are all 0 or 1 corresponding to "no" and "yes"

[10] In what follows, we allow ourselves the liberty of writing "$\mathbf{E}_{\rho^{\otimes n}}[\mathcal{P}_\pi]$" even though $\mathcal{P}_\pi$ is not usually Hermitian and therefore not an "observable". (It's only Hermitian when $\pi = \pi^{-1}$.) Nevertheless, the expression $\langle \rho^{\otimes n}, X \rangle$ makes sense for any operator X on $(\mathbb{C}^d)^{\otimes n}$, and the final operator $\mathcal{C}_\kappa$ that we care about *is* Hermitian. That's because $\mathcal{C}_\kappa$ is a real linear combination of Hermitian operators of the form $\frac{1}{2}(\mathcal{P}_\pi + \mathcal{P}_{\pi^{-1}})$, since π and π^{-1} always have the same cycle type.

outcomes.) Since we are indifferent to the eigen*vectors* $|1\rangle, \ldots, |d\rangle$ of ρ, we may as well "average X over all unitary transformations of $\mathbb{C}^d$"; i.e., replace it with $\operatorname{avg}_{\boldsymbol{U}}\{\boldsymbol{U}^{\otimes n} X (\boldsymbol{U}^{-1})^{\otimes n}\}$, where $\boldsymbol{U}$ is a uniformly random element of the unitary group $U(d)$. It is not hard to show that this can only decrease the variance of X (which is good), and the resulting X has the property that it is a linear combination of the permutation operators $\mathcal{P}_\pi$. (The latter fact is nontrivial; it is a consequence of the *Schur–Weyl duality* theorem from representation theory.) Furthermore, since the n outputs of the contraption ρ are independent, we may as well "average X over all permutations in S_n"; i.e., replace the new X with $\operatorname{avg}_{\boldsymbol{\pi} \sim S_n}\{\mathcal{P}_{\boldsymbol{\pi}} X \mathcal{P}_{\boldsymbol{\pi}}^{-1}\}$. Again, this can only decrease the variance, and the resulting X is now in $\mathcal{A}$. Thus we have shown that we may as well only consider observables in $\mathcal{A}$. This will be the justification for the Optimal Measurement Theorem, as we will shortly see.

What's convenient about the observables $\mathcal{C}_\kappa$ is that they have a straightforward definition and a nice formula for their expectation: $\mathbf{E}_{\rho^{\otimes n}}[\mathcal{C}_\kappa] = \operatorname{pow}_\kappa(p)$. What's *inconvenient* about the $\mathcal{C}_\kappa$'s is multiplying them; even the simple computation of $\mathcal{C}_{(2)}^2$ in (27) was a little tiresome. One thing that would be nice would be to have a different, "orthogonal" basis $(\Pi_\lambda)_\lambda$ for $\mathcal{A}$, meaning one with the property that

$$\Pi_\lambda \cdot \Pi_{\lambda'} = \begin{cases} \Pi_\lambda & \text{if } \lambda = \lambda', \\ 0 & \text{if } \lambda \neq \lambda'; \end{cases} \qquad \text{i.e., the } \Pi_\lambda\text{'s are orthogonal projections on } (\mathbb{C}^d)^{\otimes n}.$$

(The fact that we chose the letter λ to index the basis is not accidental...) Then linear combinations of these basis elements would be very easy to multiply.

There is another bonus of finding such a nice basis of orthogonal projections: we can build a general quantum measurement ("POVM") from it, taking the "E" matrices to be the orthogonal projections Π_λ. Since every element of the algebra $\mathcal{A}$ is a linear combination of the Π_λ's, we can construct any observable we may have wanted by first performing this measurement — thereby getting some random $\boldsymbol{\lambda}$ — and then deterministically post-processing $\boldsymbol{\lambda}$.

By the end of this section, we will see an orthogonal basis $(\Pi_\lambda)_\lambda$ for $\mathcal{A}$ in which the λ's range over all n-box, d-row Young diagrams, and

$$\mathop{\mathbf{Pr}}_{\rho^{\otimes n}}[\boldsymbol{\lambda} = \lambda] = \langle \rho^{\otimes n}, \Pi_\lambda \rangle = \dim(\lambda) \cdot s_\lambda(p). \tag{28}$$

As described in equation (11) from Section 5, this is precisely the probability distribution on Young diagrams that arises from RSKshape($\boldsymbol{w}$) when $\boldsymbol{w} \sim p^{\otimes n}$. Thus we see the full justification for the Optimal Measurement Theorem.

Let's now look for the desired "nice orthogonal basis" $(\Pi_\lambda)_\lambda$. When $n = 2$, things are very simple: there are only two cycle types in S_2, and $\mathcal{A}$ is just the span of $\mathbb{1}$ (the identity operator on $(\mathbb{C}^d)^{\otimes 2}$) and $\mathcal{P}_{(1\,2)}$ (the swapping operator). The nice basis we're looking for is

$$\Pi_{\text{sym}} = \tfrac{1}{2} \cdot \mathbb{1} + \tfrac{1}{2} \mathcal{P}_{(1\,2)}, \qquad \Pi_{\text{alt}} = \tfrac{1}{2} \cdot \mathbb{1} - \tfrac{1}{2} \mathcal{P}_{(1\,2)}.$$

For $n = 3$, we have three cycle types, and the desired nice basis for $\mathcal{A}$ is

$$\Pi_{\text{sym}} = \operatorname*{avg}_{\pi \in S_3} \{\mathcal{P}_\pi\}, \qquad \Pi_{\text{alt}} = \operatorname*{avg}_{\pi \in S_3} \{\operatorname{sgn}(\pi) \cdot \mathcal{P}_\pi\}, \qquad \Pi_{\text{std}} = \tfrac{1}{3}\big(2 \cdot \mathbb{1} - \mathcal{P}_{(1\,2\,3)} - \mathcal{P}_{(1\,3\,2)}\big).$$

For $n = 4$... well, the pattern isn't easy to spot. You might not be surprised to learn, though, that it has something to do with the representation theory of the symmetric group. Specifically, in the general-n case, the basis for $\mathcal{A}$ we're looking for has one member Π_λ for each n-box Young diagram λ. (This is the correct count of basis members, since it also equals the number of cycle

types in S_n.) These Young diagrams index the *irreducible representations* of S_n, and as such they also index the *(normalized) characters* of S_n, which are certain functions $\widehat{\chi}_\lambda : S_n \to \mathbb{Q}$ with the property that $\widehat{\chi}_\lambda(\pi)$ only depends on the cycle type of π. (Because of this, we'll sometimes write $\widehat{\chi}_\lambda(\kappa)$, where κ is the cycle type of π.) The "orthogonal basis" of $\mathcal{A}$ we are looking for turns out to be

$$\Pi_\lambda = \tfrac{(\dim \lambda)^2}{n!} \cdot \sum_{\pi \in S_n} \widehat{\chi}_\lambda(\pi) \cdot \mathcal{P}_\pi, \tag{29}$$

(recall $\dim \lambda = \#\mathrm{SYT}(\lambda)$). To see that $\Pi_\lambda \in \mathcal{A}$, you can observe that the coefficient on $\mathcal{P}_\pi$ in its definition only depends on π's cycle type. To see that the Π_λ's are "orthogonal" requires some representation theory; basically, you expand the product $\Pi_\lambda \Pi_{\lambda'}$, use the fact that each character is the trace of the associated representation, and then use the fact that different matrix elements of representations are orthogonal.

You might want an explicit formula for the normalized group character $\widehat{\chi}_\lambda(\pi)$ for S_n, but unfortunately you can't expect a very good one — computing symmetric group characters is #P-complete [Hep94]! In the next section we'll see that computing $\widehat{\chi}_\lambda(\pi)$ is efficient when the cycle type of π is considered to be of "constant" size (with 1-cycles ignored). For now, we'll take the following implicit definition of the character values $\widehat{\chi}_\lambda(\kappa)$, sometimes called the *Murnaghan–Nakayama rule*. It says that the characters essentially express how to write the Schur basis of symmetric functions in terms of the power sum basis.

$$\mathrm{pow}_\kappa(x_1, \dots, x_d) = \sum_\lambda \widehat{\chi}_\lambda(\kappa) \dim(\lambda) \cdot s_\lambda(x_1, \dots, x_d). \tag{30}$$

Given this definition, the formula (28) follows immediately by applying $\mathbf{E}_{\rho^{\otimes n}}[\cdot]$ to (29) and using (26). Thus we have now fully explained the justification for the Optimal Measurement Theorem.

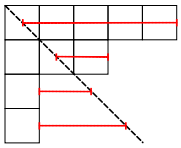
12 Symmetric group characters on small conjugacy classes

It may look like we've made some backward progress in terms of estimating statistics of the spectrum $\{p_1, \dots, p_d\}$ of ρ. Initially we considered the simple observables $\mathcal{C}_\kappa$, which have expectation $\mathrm{pow}_\kappa(p)$. Now we've justified the Optimal Measurement Theorem which tells us we may instead measure using the Π_λ's and thereby obtain a Young diagram $\boldsymbol{\lambda}$ distributed as $\mathrm{RSKshape}(\boldsymbol{w})$ for $\boldsymbol{w} \sim p^{\otimes n}$. But given such a $\boldsymbol{\lambda}$, how would we recover an estimator for, say, $\mathrm{pow}_2(p)$? The answer lies in the combination of formulas (29) and (30): to get an estimator with mean $\mathrm{pow}_\kappa(p)$, we need to output $\widehat{\chi}_{\boldsymbol{\lambda}}(\kappa)$.

Happily, there is a "good" (efficient) formula for computing the (normalized) character $\widehat{\chi}_\lambda(\kappa)$ when the cycle type κ has "constant" size (see [VK81, KO94, OO96, Oko08]). To describe it, it's helpful to introduce some notation. First, let κ be a partition of the integer k (with $k \le n$) and let $\pi \in S_n$ be of cycle type κ. We think of π as fixed and $\lambda = (\lambda_1, \dots, \lambda_d)$ as variable. It's more elegant to work with the following "shifted" parameters $L_1, \dots, L_d$, where $L_i = \lambda_i - (i - 1/2)$. (These expressions have a natural pictorial meaning; L_i is the displacement from the main diagonal of the right edge of the ith row in λ. See the figure below for an example with $\lambda = (5, 3, 1, 1)$.) Next, introduce the notation $\Sigma_\kappa(\lambda) = n(n-1)\cdots(n-k+1) \cdot \widehat{\chi}_\lambda(\pi)$. (The prefactor here is the number of ways of "embedding" an element of S_k into S_n.) Finally, the symmetric group characters are specified by the fact that $\Sigma_\kappa(\lambda)$ is the unique polynomial of the form

$$\mathrm{pow}_\kappa(L) + \Big\{\text{lower-degree power sum polynomials of } L\Big\}$$

such that $\Sigma_\kappa(\lambda) = 0$ whenever λ has fewer than k boxes.



$$\begin{aligned} L_1 &= \lambda_1 - 0.5 = 4.5 \\ L_2 &= \lambda_2 - 1.5 = 1.5 \\ L_3 &= \lambda_3 - 2.5 = -1.5 \\ L_4 &= \lambda_4 - 3.5 = -2.5 \end{aligned}$$

To take the simplest example, suppose $\kappa = (2)$, so $k = 2$; in other words, we are interested in characters' values on transpositions. We are told that

$$\Sigma_{(2)}(\lambda) = \mathrm{pow}_2(L) + a \cdot \mathrm{pow}_1(L) + b$$

for some constants a, b, and that $\Sigma_{(2)}(\lambda) = 0$ whenever λ has fewer than two boxes — i.e., when $\lambda = (1, 0, \ldots, 0)$ or $(0, 0, \ldots, 0)$. The two constraints let us solve for the two unknowns and we find that $a = 0$, $b = -\sum_{i=1}^{d}(i - 1/2)^2$. We can therefore finally conclude,

$$\text{for } \pi \text{ a transposition}, \quad \widehat{\chi}_\lambda(\pi) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{d}\lambda_i^2 - \sum_{i=1}^{d}(2i-1)\lambda_i\right). \tag{31}$$

With this formula in hand, we can show a sample-efficient method for learning the complete spectrum $\{p_1, \ldots, p_d\}$ of an unknown ρ: the *Empirical Young Diagram (EYD)* method, first proposed by [ARS88, KW01]. Without loss of generality, assume $p_1 \geq p_2 \geq \cdots \geq p_d$. The EYD method obtains $\boldsymbol{\lambda}$ from $\rho^{\otimes n}$ as in the Optimal Measurement Theorem and then simply outputs the estimates $\widehat{p}_i = \boldsymbol{\lambda}_i/n$. Following [OW16], let's see that this method has the guarantee

$$\mathbf{E}_{\boldsymbol{\lambda}}\left[\|\widehat{p} - p\|_2^2\right] \leq \frac{2\sum_{i=1}^{d} ip_i}{n} \leq \frac{d}{n}. \tag{32}$$

(The latter inequality is because $p_1 = p_2 = \cdots = p_d = 1/d$ yields the largest value of $\sum_i ip_i$ when subject to $p_1 \geq p_2 \geq \cdots \geq p_d$.) A consequence of (32) is that $n = O(d/\epsilon)$ samples suffice to estimate the sorted spectrum p to ℓ_2^2-accuracy ϵ (with high probability), and hence $n = O(d^2/\epsilon^2)$ samples suffice to estimate it to total variation distance ϵ, by Cauchy–Schwarz (improving on the previous bound of $n = O(d^2/\epsilon^2 \cdot \log(d/\epsilon))$ samples due to [HM02, CM06]).

To obtain (32), we begin with

$$\begin{aligned} \mathbf{E}_{\boldsymbol{\lambda}}\left[\|\widehat{p} - p\|_2^2\right] &= \mathbf{E}\left[\sum_{i=1}^{d}(\boldsymbol{\lambda}_i/n - p_i)^2\right] = \frac{1}{n^2}\mathbf{E}\left[\sum_{i=1}^{d}\boldsymbol{\lambda}_i^2\right] - \frac{2}{n}\sum_{i=1}^{d}p_i\mathbf{E}[\boldsymbol{\lambda}_i] + \mathrm{pow}_2(p) \\ &= \frac{1}{n^2}\mathbf{E}\left[n(n-1)\widehat{\chi}_{\boldsymbol{\lambda}}(2) + \sum_{i=1}^{d}(2i-1)\boldsymbol{\lambda}_i\right] - \frac{2}{n}\sum_{i=1}^{d}p_i\mathbf{E}[\boldsymbol{\lambda}_i] + \mathrm{pow}_2(p) \qquad \text{(by (31))} \\ &= (2 - 1/n)\mathrm{pow}_2(p) + \frac{1}{n^2}\sum_{i=1}^{d}(2i - 1 - 2p_in)\mathbf{E}[\boldsymbol{\lambda}_i], \end{aligned} \tag{33}$$

where in the last line we used that $\mathbf{E}[\widehat{\chi}_{\boldsymbol{\lambda}}(2)] = \mathrm{pow}_2(p)$. We now use the majorization statement (5) from Section 4, namely that $(\mathbf{E}[\boldsymbol{\lambda}_i])_i \succ (p_in)_i$. Since the sequence $(2i - 1 - 2p_in)_i$ is increasing in i, and the sequence $(\mathbf{E}[\boldsymbol{\lambda}_i])_i$ is decreasing in i, a basic rearrangement inequality tells us that the inner product $\sum_{i=1}^{d}(2i - 1 - 2p_in)\mathbf{E}[\boldsymbol{\lambda}_i]$ only increases if we replace $(\mathbf{E}[\boldsymbol{\lambda}_i])_i$ with a sequence that it majorizes. Thus we get the following bound, implying (32):

$$(33) \leq (2 - 1/n)\mathrm{pow}_2(p) + \frac{1}{n^2}\sum_{i=1}^{d}(2i - 1 - 2p_in)p_in + \mathrm{pow}_2(p) = \frac{2\sum_{i=1}^{d} ip_i}{n} - \left(\frac{1 + \mathrm{pow}_2(p)}{n}\right).$$

Elaborations of this method for bounding the sample-complexity of learning ρ's spectrum appear in [OW16, OW17]; results include learning just the k largest eigenvalues with sample complexity depending only on k, and learning with respect to Hellinger, KL-, and χ^2-divergence error.

13 Further results and reading

In the previous section we described how $n = O(d^2/\epsilon^2)$ samples suffice to estimate suffice to estimate the spectrum of ρ to total variation distance ϵ. In fact, it has been shown [OW16] that this many samples also suffice to estimate all of ρ itself to trace distance ϵ, and further that $n = \Omega(d^2/\epsilon^2)$ is necessary [HHJ$^+$16]. (Similar results for learning ρ with respect to infidelity and other measures appear in [HHJ$^+$16, OW17].) Describing the estimation algorithm — due to Keyl [Key06] — and its analysis [OW16] would take us too far afield, but suffice it to say it involves further analysis of Schur polynomials, representation theory of the unitary group, the Harish-Chandra–Itzykson–Zuber formula, Gelfand–Tsetlin patterns, the theory of random matrices, and other interesting topics. For further reading on the topics discussed in this survey, recommendations include Canonne's survey on (classical) distribution testing [Can15], Romik's book on longest increasing subsequences [Rom14], Fulton's book on Young tableaus [Ful97], and Méliot's book on representation theory of the symmetric group [Mél17]. For applications of Schur-Weyl duality to quantum computing in addition to state estimation, we recommend the theses of Harrow [Har05] and Christandl [Chr06]. The Π_λ measurement from above is known as *weak Schur sampling* and can be implemented efficiently on a quantum computer [MW16]. Some applications require a generalization known as the *strong Schur transform*, which can also be computed efficiently on a quantum computer [BCH05, Har05].

Acknowledgments

The authors would like to thank Lane Hemaspaandra for inviting this article to appear in SIGACT News, Costin Bădescu for allowing us to describe some of [BOW17] here, and Costin Bădescu, Clément Canonne, Persi Diaconis, Ilias Diakonikolas, Valentin Féray, Oded Goldreich, Christian Houdré, Christian Ikenmeyer, Greg Kuperberg, Tongyang Li, Māris Ozols, Igor Pak, and Ronald de Wolf for helpful comments.

References

[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

[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. 12

[BCH05] Dave Bacon, Isaac Chuang, and Aram Harrow. The quantum Schur transform: I. efficient qudit circuits. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005. 13

[BOW17] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. Technical report, arXiv, 2017. 4, 8, 10, 13

[Can15] Clément Canonne. A survey on distribution testing: Your data is big. But is it blue? Technical Report 63, Electronic Colloquium on Computational Complexity, 2015. http://www.cs.columbia.edu/~ccanonne/files/misc/2015-survey-distributions.pdf. 13

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. Society for Industrial and Applied Mathematics, 2014. 10

[Cer15] Miguel de Cervantes Saavedra. *El ingenioso hidalgo don Quijote de la Mancha*. Published by Francisco de Robles, 1605, 1615. http://www.gutenberg.org/ebooks/2000. 1

[Chr06] Matthias Christandl. *The Structure of Bipartite Quantum States*. PhD thesis, University of Cambridge, 2006. 13

[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. 4, 10

[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. 12

[DGPP16] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. Technical report, arXiv, 2016. 10

[FRT54] James Sutherland Frame, Gilbert de Beauregard Robinson, and Robert Thrall. The hook graphs of the symmetric groups. *Canadian Journal of Mathematics*, 6(316):316–324, 1954. 5

[Ful97] William Fulton. *Young tableaux: with applications to representation theory and geometry*. Cambridge University Press, 1997. 13

[Gai14] Helen Gaines. *Cryptanalysis: a study of ciphers and their solution*. Dover, 2014. 1

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000. 10

[Gre74] Curtis Greene. An extension of Schensted's theorem. *Advances in Mathematics*, 14:254–265, 1974. 3

[Har05] Aram Harrow. *Applications of coherent classical communication and the Schur transform to quantum information theory*. PhD thesis, Massachusetts Institute of Technology, 2005. 13

[Hep94] Charles Hepler. *On the complexity of computing characters of finite groups*. PhD thesis, University of Calgary, 1994. 11

[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*, pages 913–925, 2016. 13

[HM02] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, 2002. 12

[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. 7

[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. 6

[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. 7, 7, 7

[Joh01] Kurt Johansson. Discrete orthogonal polynomial ensembles and the Plancherel measure. *Annals of Mathematics*, 153(1):259–296, 2001. 7

[JVHW17] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017. 2

[Ker03] Sergei Kerov. *Asymptotic representation theory of the symmetric group and its applications in analysis*. American Mathematical Society, 2003. 7

[Key06] Michael Keyl. Quantum state estimation and large deviations. *Reviews in Mathematical Physics*, 18(01):19–60, 2006. 13

[Knu70] Donald Knuth. Permutations, matrices, and generalized Young tableaux. *Pacific Journal of Mathematics*, 34(3):709–727, 1970. 3

[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. 12

[KW01] Michael Keyl and Reinhard Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5):052311, 2001. 12

[MÍ12] Pierre-Loïc Méliot. Fluctuations of central measures on partitions. In *Proceedings of the 27th Annual International Conference on Formal Power Series and Algebraic Combinatorics*, 2012. 7

[Mél17] Pierre-Loïc Méliot. *Representation theory of symmetric groups*. CRC Press, 2017. 13

[MW16] Ashley Montanaro and Ronald de Wolf. *A Survey of Quantum Property Testing*. Number 7 in Graduate Surveys. Theory of Computing Library, 2016. 4, 13

[Oko08] Andreĭ Okounkov. Characters of symmetric groups #1, 2008. MSRI Introductory Workshop on Combinatorial Representation Theory. http://jessica2.msri.org/attachments/12622/12622.pdf. 12

[OO96] Andreĭ Okounkov and Grigori Olshanski. Shifted Schur functions. Technical Report q-alg/9605042, arXiv, 1996. 12

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 529–538, 2015. 4, 10

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 899–912, 2016. 6, (document), 7, 8, 12, 12, 13

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography II. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 962–974, 2017. 6, (document), 7, 8, 12, 13

[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. 10

[Rob29] Howard Robertson. The uncertainty principle. *Physical Review*, 34(1):163, 1929. 9

[Rob38] Gilbert de Beauregard Robinson. On the representations of the symmetric group. *American Journal of Mathematics*, 60(3):745–760, 1938. 3

[Rom14] Dan Romik. *The surprising mathematics of longest increasing subsequences*. Cambridge University Press, 2014. 13

[Sch61] Craige Schensted. Longest increasing and decreasing subsequences. *Canadian Journal of Mathematics*, 13(2):179–191, 1961. 3

[Sta99] Richard Stanley. *Enumerative combinatorics Volume 2*. Cambridge University Press, Cambridge, 1999. 5

[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. 7

[Vie81] Gérard Viennot. Équidistribution des permutations ayant une forme donnée selon les avances et coavances. *Journal of Combinatorial Theory, Series A*, 31(1):43–55, 1981. 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. 7, 12

[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. 10

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016. 2