Efficient quantum tomography II

Ryan O'Donnell[*] John Wright[†]

November 23, 2016

Abstract

Following [OW16], we continue our analysis of: (i) "Quantum tomography", i.e., learning a quantum state, i.e., the quantum generalization of learning a discrete probability distribution; (ii) The distribution of Young diagrams output by the RSK algorithm on random words. Regarding (ii), we introduce two powerful new tools:

- A precise upper bound on the expected length of the longest union of k disjoint increasing subsequences in a random length-n word with letter distribution $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d$. Our bound has the correct main term and second-order term, and holds for *all* n, not just in the large-n limit.
- A new majorization property of the RSK algorithm that allows one to analyze the Young diagram formed by the *lower* rows $\lambda_k, \lambda_{k+1}, \dots$ of its output.

These tools allow us to prove several new theorems concerning the distribution of random Young diagrams in the *nonasymptotic* regime, giving concrete error bounds that are optimal, or nearly so, in all parameters. As an one example, we give a fundamentally new proof of the celebrated fact that the expected length of the longest increasing sequence in a random length-n permutation is bounded by $2\sqrt{n}$. This is the $k = 1$, $\alpha_i \equiv \frac{1}{d}$, $d \to \infty$ special case of a much more general result we prove: the expected length of the kth Young diagram row produced by an α-random word is $\alpha_k n \pm 2\sqrt{\alpha_k d n}$.

From our new analyses of random Young diagrams we derive several new results in quantum tomography, including:

- Learning the eigenvalues of an unknown state to ϵ-accuracy in Hellinger-squared, chi-squared, or KL distance, using $n = O(d^2/\epsilon)$ copies.
- Learning the top-k eigenvalues of an unknown state to ϵ-accuracy in Hellinger-squared or chi-squared distance using $n = O(kd/\epsilon)$ copies or in ℓ_2^2 distance using $n = O(k/\epsilon)$ copies.
- Learning the optimal rank-k approximation of an unknown state to ϵ-fidelity (Hellinger-squared distance) using $n = \widetilde{O}(kd/\epsilon)$ copies.

We believe our new techniques will lead to further advances in quantum learning; indeed, they have already subsequently been used for efficient von Neumann entropy estimation.

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

[†] Department of Physics, Massachusetts Institute of Technology. Most of this work performed while affiliated with the Department of Computer Science, Carnegie Mellon University. Partially supported by a Simons Fellowship in Theoretical Computer Science. jswright@mit.edu

1 Introduction

The *Robinson–Schensted–Knuth (RSK) algorithm* is a well-known combinatorial algorithm with diverse applications throughout mathematics, computer science, and physics. Given a word w with n letters from the alphabet $[d]$, it outputs two semistandard Young tableaus $(P, Q) = \mathrm{RSK}(w)$ with common shape given by some Young diagram $\lambda \in \mathbb{N}^d$ $(\lambda_1 \geq \cdots \geq \lambda_d)$. We write $\lambda = \mathrm{shRSK}(w)$, and mention that λ can be defined independently of the RSK algorithm as in Theorem 1.2 below. In the RSK algorithm, the process generating the first row is sometimes called *patience sorting*, and it is equivalent to the basic dynamic program for computing w's longest (weakly) increasing subsequence.

Definition 1.1. Given a word $w \in [d]^n$, a *subsequence* is a sequence of letters $(w_{i_1}, \dots, w_{i_\ell})$ such that $i_1 < \cdots < i_\ell$. The *length* of the subsequence is ℓ. We say that the subsequence is *weakly increasing*, or just *increasing*, if $w_{i_1} \leq \cdots \leq w_{i_\ell}$. We write $\mathrm{LIS}(w)$ for the length of the longest weakly increasing subsequence in w.

Hence $\lambda_1 = \mathrm{LIS}(w)$, a result known as Schensted's Theorem [Sch61]. Further rows of λ are characterized by Greene's Theorem as giving the "higher order LIS statistics" of w.

Theorem 1.2 ([Gre74]). *Suppose $\lambda = \mathrm{shRSK}(w)$. Then for each k, $\lambda_1 + \cdots + \lambda_k$ is equal to the length of the longest union of k disjoint (weakly) increasing subsequences in w.*

For background on the RSK algorithm, see e.g. [Ful97, Rom14] and the references therein.

Many applications involve studying the behavior of the RSK algorithm when its input is drawn from some random distribution. A famous case is the uniform distribution over length-n permutations $\boldsymbol{\pi} \sim S_n$ (in which case $d = n$); here the resulting random Young diagram $\boldsymbol{\lambda} = \mathrm{RSK}(\boldsymbol{\pi})$ is said to have *Plancherel distribution*. Starting with the work of Ulam [Ula61], a line of research has studied the distribution of the longest increasing subsequence of $\boldsymbol{\pi}$; its results are summarized as follows: $\mathbf{E}[\mathrm{LIS}(\boldsymbol{\pi})] \to 2\sqrt{n}$ as $n \to \infty$ [LS77, VK77] (in fact, $\mathbf{E}[\mathrm{LIS}(\boldsymbol{\pi})] \leq 2\sqrt{n}$ for all n [VK85, Pil90]), and the deviations of $\mathrm{LIS}(\boldsymbol{\pi})$ from this value can be characterized by the Tracy–Widom distribution from random matrix theory [BDJ99]. The RSK algorithm has played a central role in many of these developments, and these results have been shown to apply not just to the first row $\boldsymbol{\lambda}_1 = \mathrm{LIS}(\boldsymbol{\pi})$ but also to the entire shape of $\boldsymbol{\lambda}$ [Joh01, BOO00]. In a different stream of research, the Plancherel distribution arises naturally in quantum algorithms which perform Fourier sampling over the symmetric group. Here, its properties have been used to show that any quantum algorithm for graph isomorphism (or, more generally, the hidden subgroup problem on the symmetric group) which uses the "standard approach" must perform highly entangled measurements across many copies of the coset state [HRTS03, MRS08, HMR$^+$10].

In this work, we consider a more general setting, sometimes called the *inhomogeneous random word model*, in which the input to the RSK algorithm is a random word $\boldsymbol{w}$ whose letters are selected independently from some probability distribution.

Definition 1.3. Given a probability distribution $\alpha = (\alpha_1, \dots, \alpha_d)$ on alphabet $[d]$, an *n-letter α-random word* $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_n)$, written as $\boldsymbol{w} \sim \alpha^{\otimes n}$, is a random word in which each letter $\boldsymbol{w}_i$ is independently drawn from $[d]$ according to α. The *Schur–Weyl distribution* $\mathrm{SW}^n(\alpha)$ is the distribution on Young diagrams given by $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$. Although it is not obvious, it is a fact that the distribution $\mathrm{SW}^n(\alpha)$ does not depend on the ordering of α's components. Thus unless otherwise stated, we will assume that α is sorted; i.e., $\alpha_1 \geq \cdots \geq \alpha_d$.

(The *homogeneous* random word model is the special case in which $\alpha_i = \frac{1}{d}$, for each $i \in [d]$. It is easy to see that in this case, $\mathrm{SW}^n(\alpha)$ converges to the Plancherel distribution as $d \to \infty$.) Aside from arising naturally in combinatorics and representation theory, the Schur–Weyl distribution also appears in a large number of problems in quantum learning and data processing, as we will see below.

Much of the prior work on the Schur–Weyl distribution has occurred in the *asymptotic* regime, in which d and α are held constant and $n \to \infty$. An easy exercise in Chernoff bounds shows that $\mathrm{LIS}(\boldsymbol{w})/n \to \alpha_1$ as $n \to \infty$. Generalizing this, a sequence of works [TW01, Joh01, ITW01, HX13, Mél12] have shown that in this regime, $\boldsymbol{\lambda}$ is equal to $(\alpha_1 n, \dots, \alpha_d n)$ plus some lower-order fluctuations distributed as the eigenvalues of certain random matrix ensembles. From these works, we may extract the following ansatz, coarsely describing the limiting behavior of the rows of $\boldsymbol{\lambda}$.

Ansatz: For all $k \in [d]$, $\boldsymbol{\lambda}_k \approx \alpha_k n \pm 2\sqrt{\alpha_k d_k n}$.

Here d_k is the number of times α_k occurs in $(\alpha_1, \dots, \alpha_d)$. We survey this literature below in Section 1.5.

1.1 A nonasymptotic theory of the Schur–Weyl distribution

In this work, motivated by problems in quantum state learning, we study the Schur–Weyl distribution in the *nonasymptotic* regime. Previous efforts in this direction were the works [HM02, CM06] and, more extensively, our previous paper [OW16]. Our goal is to prove worst-case bounds on the shape of $\boldsymbol{\lambda}$ which hold for all n, independent of d and α. When possible, we would like to translate certain features of the Schur–Weyl distribution present in the asymptotic regime — in particular, the ansatz and its consequences — down into the nonasymptotic regime.

Clearly, nonasymptotic results cannot depend on the quantity d_k, which can be sensitive to arbitrarily small changes in α that are undetectable when n is small. (Consider especially when α is uniform versus when α is uniform but with each entry slightly perturbed.) Instead, our results are in terms of the quantity $\min\{1, \alpha_k d\}$, for each $k \in [d]$, which always upper bounds $\alpha_k d_k$.

Our first result tightly bounds the expected row lengths, in line with the ansatz.

Theorem 1.4. *For $k \in [d]$, set $\nu_k = \min\{1, \alpha_k d\}$. Then*

$$\alpha_k n - 2\sqrt{\nu_k n} \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_k \leq \alpha_k n + 2\sqrt{\nu_k n},$$

This improves on a result from [OW16], which showed an upper bound in the $k = 1$ case with error $+2\sqrt{2}\sqrt{n}$ for general α and with error $+2\sqrt{n}$ for α the uniform distribution. Setting $\alpha = (\frac{1}{d}, \dots \frac{1}{d})$ and letting $d \to \infty$, the $k = 1$ case of Theorem 1.4 recovers the above-mentioned celebrated fact that the length of the longest increasing subsequence of a random permutation of n is at most $2\sqrt{n}$ in expectation. Our result gives only the second proof of this statement since it was originally proved independently by Vershik and Kerov in 1985 [VK85] and by Pilpel in 1990 [Pil90]. Next, we bound the mean-squared error of the estimator $\boldsymbol{\lambda}_k/n$ for α_k.

Theorem 1.5. *For $k \in [d]$, set $\nu_k = \min\{1, \alpha_k d\}$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} (\boldsymbol{\lambda}_k - \alpha_k n)^2 \leq O(\nu_k n).$$

Again, this is in line with the ansatz. This theorem can be used to derive tight bounds (up to constant factors) on the convergence of the *normalized Young diagram* $\underline{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_1/n, \dots, \boldsymbol{\lambda}_d/n)$ to α in a variety of distance measures, including Hellinger-squared distance and the KL and chi-squared divergences. Now in fact, using related techniques, in [OW16] we were able to prove convergence bounds for some distance measures with stronger constants:

Theorem 1.6 ([OW16]). $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} \|\underline{\boldsymbol{\lambda}} - \alpha\|_2^2 \leq \frac{d}{n}$ *and* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} \|\underline{\boldsymbol{\lambda}} - \alpha\|_1 \leq \frac{d}{\sqrt{n}}$.

In this work, we extend Theorem 1.6 to other, more challenging distance measures.

Theorem 1.7. *Let $d(\cdot,\cdot)$ be any of $d_{\mathrm{H}^2}(\cdot,\cdot)$, $d_{\mathrm{KL}}(\cdot,\cdot)$, or $d_{\chi^2}(\cdot,\cdot)$. Then* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} d(\underline{\boldsymbol{\lambda}}, \alpha) \leq \frac{d^2}{n}$.

Not only are Theorems 1.6 and 1.7 in line with the ansatz, they even have the correct constant factors, as predicted below by Theorem 1.24 in the asymptotic regime.

Finally, we show similar results for truncated distances, in which only the top k entries of $\boldsymbol{\lambda}$ and the top k entries of α are compared with each other. In [OW16], this was carried out for truncated ℓ_1 distance.

Theorem 1.8 ([OW16]). $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} d_{\mathrm{TV}}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha) \leq \frac{1.92k + .5}{\sqrt{n}}$.

By following the proof of this result, our Theorem 1.4 immediately implies the same bound with 1.5 in place of 1.92. In addition, we prove similar bounds for truncated ℓ_2^2, Hellinger, and chi-squared distances.

Theorem 1.9. $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} d_{\ell_2^2}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha) \leq \frac{46k}{n}$.

Theorem 1.10. *Let $d(\cdot,\cdot)$ be either $d_{\mathrm{H}^2}^{(k)}(\cdot,\cdot)$ or $d_{\chi^2}^{(k)}(\cdot,\cdot)$. Then* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim \mathrm{SW}^n(\alpha)} d(\underline{\boldsymbol{\lambda}}, \alpha) \leq \frac{46kd}{n}$.

These results follow the ansatz, though our techniques are not yet strong enough to achieve optimal constant factors.

1.2 Techniques

Our main techniques include a pair of majorization theorems for the RSK algorithm. Here we refer to the following definition.

Definition 1.11. For $x, y \in \mathbb{R}^d$, we say that x *majorizes* y, denotes $x \succ y$ if $x_{[1]} + \cdots + x_{[k]} \geq y_{[1]} + \cdots + y_{[k]}$ for all $k \in [d]$, with equality for $k = d$. Here the notation $x_{[i]}$ means the ith largest value among the x_j's. In the case of Young diagrams λ and μ, we also use the standard notation $\lambda \trianglerighteq \mu$. *Weak majorization*, denoted with either $\succ_w$ or $\trianglerighteq_w$, is the case when the equality constraint may not necessarily hold.

Several of our results require understanding the behavior of an individual row $\boldsymbol{\lambda}_k$, for $k \in [d]$. However, the RSK algorithm's sequential behavior makes understanding rows after the first quite difficult. So instead, we adopt the strategy of proving bounds only for the first row (which can sometimes be done directly), and then translating them to the kth row via the following new theorem.

Theorem 1.12. *Fix an integer $k \geq 1$ and an ordered alphabet $\mathcal{A}$. Consider the RSK algorithm applied to some string $x \in \mathcal{A}^n$. During the course of the algorithm, some letters of x get bumped from the kth row and inserted into the $(k+1)$th row. Let $x^{(k)}$ denote the string formed by those letters* in the order they are so bumped. *On the other hand, let $\overline{x}$ be the subsequence of x formed by the letters of $x^{(k)}$* in the order they appear in x. *Then* $\mathrm{shRSK}(\overline{x}) \trianglerighteq \mathrm{shRSK}(x^{(k)})$.

Our other key tool is the following result allowing us to bound how much larger $\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_k$ is than its intended value $\alpha_1 n + \cdots + \alpha_k n$ in expectation.

Theorem 1.13. *Let α be a sorted probability distribution on $[d]$ and let $k \in [d]$. Then for all $n \in \mathbb{N}$,*

$$\mathbf{E}\Big[\sum_{i=1}^{k} \boldsymbol{\lambda}_i^{(n)}\Big] - \sum_{i=1}^{k} \alpha_i n \leq \mathrm{Excess}_k(\alpha), \quad \textit{where } \mathrm{Excess}_k(\alpha) = \sum_{i \leq k < j} \frac{\alpha_j}{\alpha_i - \alpha_j}.$$

Furthermore, using the notation $E_k^{(n)}(\alpha)$ for the left-hand side, it holds that $E_k^{(n)}(\alpha) \nearrow \mathrm{Excess}_k(\alpha)$ as $n \to \infty$ provided that all α_i's are distinct.

The fact that $E_1^{(n)}(\alpha) \to \mathrm{Excess}_1(\alpha)$ as $n \to \infty$ when all the α_i's are fixed and distinct was originally proven in by Its, Tracy, and Widom [ITW01]. We extend this to the general k case, and also show that the sequence $E_1^{(n)}(\alpha)$ is increasing in n, so that $\mathrm{Excess}_k(\alpha)$ is an upper bound for all n. So long as α_k and α_{k+1} are sufficiently separated we have found that $\mathrm{Excess}_k(\alpha)$ gives a surprisingly accurate bound on $\mathbf{E}[\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_k] - (\alpha_1 n + \cdots + \alpha_k n)$. When α_k and α_{k+1} are not well-separated, on the other hand, $\mathrm{Excess}_k(\alpha)$ can be arbitrarily large. In this case, we consider a mildly perturbed distribution α' in which α'_k and α'_{k+1} *are* well-separated and then apply Theorem 1.13 to α' instead. Supposing that $\alpha' \succ \alpha$, we may then relate the bounds we get on $\mathrm{SW}^n(\alpha')$ back to $\mathrm{SW}^n(\alpha)$ using Theorem 1.11 from [OW16].

Theorem 1.14 ([OW16])**.** *Let α, $\beta \in \mathbb{R}^d$ be sorted probability distributions with $\beta \succ \alpha$. Then for any $n \in \mathbb{N}$ there is a coupling $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ of $\mathrm{SW}^n(\alpha)$ and $\mathrm{SW}^n(\beta)$ such that $\boldsymbol{\mu} \trianglerighteq \boldsymbol{\lambda}$ always.*

1.3 Quantum state learning

Our main application of these bounds is to problems in the area of quantum state learning. Here, one is given n copies of a mixed state $\rho \in \mathbb{C}^{d \times d}$ and asked to learn some property of ρ. For example, one might attempt to learn the entire $d \times d$ matrix (*quantum tomography*), just its spectrum $\alpha = (\alpha_1, \ldots, \alpha_d)$ (*quantum spectrum estimation*), or some other more specific property such as its von Neumann entropy, its purity, and so forth. These problems play key roles in various quantum computing applications, including current-day verification of experimental quantum devices and hypothesized future quantum protocols such as entanglement verification. We allow ourselves arbitrary entangled measurements, and our goal is to learn while using as few copies n as possible.

The standard approach to designing entangled measurements for quantum state learning [ARS88, KW01] uses a powerful tool from representation theory called *Schur–Weyl duality*, which states that

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

Here the direct sum ranges over all partitions $\lambda \vdash n$ of height at most d, and Sp_λ and V_λ^d are the irreps of the symmetric and general linear groups corresponding to λ. Measuring $\rho^{\otimes n}$ according to the projectors $\{\Pi_\lambda\}_\lambda$ corresponding to the λ-subspaces is called *weak Schur sampling* and is the optimal measurement if one is interested only in learning ρ's spectrum α (or some function of α). The outcome of this measurement is a random $\boldsymbol{\lambda}$ whose distribution depends only on α; in fact:

Fact 1.15. *When performed on $\rho^{\otimes n}$, the measurement outcome $\boldsymbol{\lambda}$ of weak Schur sampling is distributed exactly as the Schur–Weyl distribution $\mathrm{SW}^n(\alpha)$, where α is ρ's spectrum.*

(See, for example, the discussion of this in [OW16].) Following weak Schur sampling, $\rho^{\otimes n}$ collapses to the subspace corresponding to $\boldsymbol{\lambda}$, and if one wishes to learn about more than just ρ's spectrum, one must perform a further measurement within this subspace. An algorithm which does so is said to have performed *strong* Schur sampling. Note that weak Schur sampling refers to a specific measurement, whereas strong Schur sampling refers to a class of measurements.

Fact 1.15, when paired with our results from Section 1.1, immediately suggests the following algorithm for estimating ρ's spectrum: perform weak Schur sampling, receive the outcome $\boldsymbol{\lambda}$, and output $\underline{\boldsymbol{\lambda}}$. This is exactly the *empirical Young diagram (EYD) algorithm* introduced independently by Alicki, Ruckinci, and Sadowski [ARS88] and Keyl and Werner [KW01]. To date, this is the best known spectrum estimation algorithm, and it has recently been proposed for current-day experimental implementation [BAH$^+$16]. Our Theorem 1.7 immediately implies the following.

Theorem 1.16. *The spectrum α can be learned in Hellinger-squared distance, KL divergence, and chi-squared divergence using $n = O(d^2/\epsilon)$ copies.*

Previously, it was known from the works of Hayashi and Matsumoto [HM02] and Christandl and Mitcheson [CM06] that $n = O(d^2/\epsilon) \cdot \log(d/\epsilon^2)$ copies sufficed for KL divergence (and hence for Hellinger-squared). We note that Theorem 1.6 from [OW16] gave learning bounds of $O(d/\epsilon)$ and $O(d^2/\epsilon^2)$ for spectrum learning under ℓ_2^2 and ℓ_1 distance, respectively. Combined with the lower bound from [OW15] showing that the EYD algorithm requires $n = \Omega(d^2/\epsilon^2)$ copies for ℓ_1 learning, we have given optimal bounds for the EYD algorithm in terms of all five distance metrics.

For the more difficult problem of quantum tomography, the optimal number of copies needed to learn ρ in trace distance was recently determined to be $n = \Theta(d^2/\epsilon^2)$ — the upper bound from our previous work [OW16] and the lower bound from the independent work of Haah et al. [HHJ$^+$16]. The optimal complexity of learning ρ in infidelity — i.e., outputting an estimate $\widehat{\boldsymbol{\rho}}$ such that $1 - F(\rho, \widehat{\boldsymbol{\rho}}) \leq \epsilon$ — remains open, however. Essentially the best prior result is by Haah et al. [HHJ$^+$16], who showed that $n = O(d^2/\epsilon) \cdot \log(d/\epsilon)$ copies suffice. For our results, we find it convenient to work with the very closely related *quantum Hellinger-squared distance* $d_{\mathrm{H}^2}(\cdot, \cdot)$. This is known to be the same as infidelity $1 - F(\rho, \widehat{\boldsymbol{\rho}})$ up to a factor of 2 (see Section 2 for details), and hence learning in quantum Hellinger-squared distance is equivalent to learning in infidelity up to a small constant. We show the following theorem.

Theorem 1.17. *A state $\rho \in \mathbb{C}^{d \times d}$ can be learned in quantum Hellinger-squared distance with copy complexity*

$$n = O\left(\min\left\{\frac{d^2}{\epsilon}\log\left(\frac{d}{\epsilon}\right), \frac{d^3}{\epsilon}\right\}\right).$$

The left-hand term in the min gives a new proof of the fidelity bound of Haah et al. [HHJ$^+$16]. The right-hand term in the min is new; previously it was known only how to learn ρ in fidelity using $n = O(h(d)/\epsilon)$ copies for some unspecified function $h(\cdot)$ (see the list of citations in [HHJ$^+$16]). Along with our trace distance bound of $n = O(d^2/\epsilon^2)$ — which implies a fidelity bound of $O(d^2/\epsilon^2)$ — we now have three incomparable upper bounds on the complexity of fidelity tomography, none of which match the best known lower bound of $\Omega(d^2/\epsilon)$ from [HHJ$^+$16]. Settling the complexity of fidelity learning remains an important open problem.

To perform full-state tomography, we analyze *Keyl's algorithm* [Key06]. After performing weak Schur sampling and receiving a random $\boldsymbol{\lambda}$, it performs a subsequent measurement in the $\boldsymbol{\lambda}$-subspace whose measurement outcomes correspond to $d \times d$ unitary matrices. We denote by $\mathrm{K}_{\boldsymbol{\lambda}}(\rho)$ the distribution on unitary matrices observed given $\boldsymbol{\lambda}$ and ρ. The algorithm receives a random $\boldsymbol{V} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)$

from this measurement and then outputs the density matrix $\boldsymbol{V}\mathrm{diag}(\boldsymbol{\lambda}/n)\boldsymbol{V}^{\dagger}$. We will only require one fact about this algorithm from [OW16], and so we defer the full description of Keyl's measurement and algorithm to the papers [Key06, OW16].

1.4 Principal component analysis

Next, we consider natural "principal component analysis" (PCA)-style versions of the above problems. Here, rather than learning the whole state or spectrum, the goal is to learn the "largest" k-dimensional part of the state or spectrum. These problems arise naturally when the state is "fundamentally" low rank, but has been perturbed by a small amount of noise. For spectrum estimation, this involves learning the first k α_i's under the ordering $\alpha_1 \geq \cdots \geq \alpha_d$. Previous work [OW16] used Theorem 1.8 to learn the first k α_i's in trace distance using $n = O(k^2/\epsilon^2)$ copies. Using our Theorems 1.9 and 1.10, we extend this result to other distance measures.

Theorem 1.18. *The first k α_i's can be learned in Hellinger-squared distance or chi-squared divergence using $n = O(kd/\epsilon)$ copies, and in ℓ_2^2 distance using $n = O(k/\epsilon)$ copies.*

For full-state PCA, the natural variant is to output a rank-k matrix $\widehat{\rho}$ which is almost as good as the best rank-k approximation to ρ. For trace distance, the work of [OW16] showed that $n = O(kd/\epsilon^2)$ copies are sufficient to output an estimate with error at most ϵ more than the error of the best rank-k approximation. In this work, we show the following fidelity PCA result.

Theorem 1.19. *There is an algorithm that, for any $\rho \in \mathbb{C}^{d\times d}$ and $k \in [d]$, outputs a random rank-k (or less) hypothesis $\widehat{\boldsymbol{\rho}}$ such that*

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho)] \leq \alpha_{>k} + O\left(\frac{kdL}{n}\right) + O\left(kL\sqrt{\frac{\alpha_{>k}}{n}}\right),$$

where $L = \min\{k, \ln n\}$ and $\alpha_{>k} = \alpha_{k+1} + \cdots + \alpha_d$.

Let us spend time interpreting this result. The Hellinger-squared error of the best rank-k approximation to ρ — the projection of ρ to its top-k eigenspace — is given by $\alpha_{>k}$. When ρ is exactly of rank k, then $\alpha_{>k} = 0$, and this bound tells us that

$$n = O\left(\min\left\{\frac{kd}{\epsilon}\log\left(\frac{d}{\epsilon}\right), \frac{k^2 d}{\epsilon}\right\}\right)$$

copies are sufficient to learn ρ up to error ϵ. The left-hand term in the min was shown previously by Haah et al. [HHJ+16] using different techniques, whereas the right-hand term is new. In the case that ρ is *not* rank-k, let us first make the reasonable assumption that $k \leq d/\ln n$. Then

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho)] \leq \alpha_{>k} + Z_1 + Z_2, \text{ where } Z_1 = O\left(\frac{kd\ln n}{n}\right),\ Z_2 = O\left(\sqrt{\frac{\alpha_{>k}kd\ln n}{n}}\right).$$

Noting that Z_2 is the geometric mean of $\alpha_{>k}$ and Z_2, we get that for any $\delta > 0$,

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho)] \leq (1+\delta)\cdot\alpha_{>k} + O_\delta\left(\frac{kd\ln n}{n}\right).$$

Hence, this tells us that $n = O(kd/\epsilon)\cdot\log(d/\epsilon)$ copies are sufficient to learn ρ to error $(1+\delta)\cdot\alpha_{>k}+\epsilon$ (essentially recovering the exactly rank-k case). Finally, in the unlikely case of $k > d/\ln n$, a similar argument shows that $n = O(kd/\epsilon)\cdot\log^2(d/\epsilon)$ copies are sufficient to learn ρ to error $(1+\delta)\cdot\alpha_{>k}+\epsilon$.

1.5 Asymptotics of the Schur–Weyl distribution

In this section, we survey the known results on the Schur–Weyl distribution in the asymptotic setting. Though we are primarily interested in proving convergence results with explicit error bounds in the nonasymptotic setting, the asymptotic regime is useful for understanding the high-level features of the Schur–Weyl distribution. Indeed, the early quantum computing papers [ARS88, KW01] on this topic operated in this regime.

The earliest theorem in this area is due to Vershik and Kerov [VK81], who showed the following:

Theorem 1.20 ([VK81])**.** *Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution, and let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$. Then for all $k \in [d]$, as $n \to \infty$ we have $\boldsymbol{\lambda}_k/n \to \alpha_k$ in probability.*

This theorem has been reproven in a variety of works, including independently by [ARS88] and [KW01] in the quantum computing literature.

Subsequent work determined the lower-order asymptotics of the Schur–Weyl distribution. As it turns out, the qualitative features of the distribution depend on whether α has any repeated values. The simplest case, when all the α_i's are distinct, was first handled in the work of Alicki, Rudnicki, and Sadowski [ARS88].

Theorem 1.21 ([ARS88])**.** *Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution in which every entry is distinct. Let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, and let $(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_d)$ be centered jointly Gaussian random variables with $\mathbf{Var}[\boldsymbol{g}_i] = \alpha_i(1-\alpha_i)$ and $\mathbf{Cov}[\boldsymbol{g}_i, \boldsymbol{g}_j] = -\alpha_i\alpha_j$, for $i \neq j$.*[1] *Then as $n \to \infty$,*

$$\left(\frac{\boldsymbol{\lambda}_1 - \alpha_1 n}{\sqrt{n}}, \ldots, \frac{\boldsymbol{\lambda}_d - \alpha_d n}{\sqrt{n}}\right) \to (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_d)$$

in distribution.

In other words, in the case of distinct α_i's, the Schur–Weyl distribution acts in the asymptotic regime like the multinomial distribution with parameter α. The intuition is that given an α-random word $\boldsymbol{w}$, the count of 1's is so much greater than the count of any other letter that the longest increasing subsequence is not much longer than the all-1's subsequence. Similarly, the longest pair of disjoint increasing subsequences is not much longer than the all-1's and all-2's subsequences, and so forth. This theorem has been reproven many times, such as by [HX13, Buf12, Mél12, FMN13].

On the other hand, when α is *degenerate*, i.e. the α_i's are *not* all distinct, then $\mathrm{SW}^n(\alpha)$ has a surprisingly non-Gaussian limiting behavior. The first paper along this line of work was by Baik, Deift, and Johansson [BDJ99]; it characterized the Plancherel distribution (a special case of the Schur–Weyl distribution) in terms of the eigenvalues of the *Gaussian unitary ensemble.*

Definition 1.22. The *Gaussian unitary ensemble* GUE_d is the distribution on $d \times d$ Hermitian matrices $\boldsymbol{X}$ in which (i) $\boldsymbol{X}_{i,i} \sim \mathcal{N}(0,1)$ for each $i \in [d]$, and (ii) $\boldsymbol{X}_{i,j} \sim \mathcal{N}(0,1)_{\mathbb{C}}$ and $\boldsymbol{X}_{j,i} = \overline{\boldsymbol{X}_{i,j}}$ for all $i < j \in [d]$. Here $\mathcal{N}(0,1)_{\mathbb{C}}$ refers to the *complex* standard Gaussian, distributed as $\mathcal{N}(0, \frac{1}{2}) + i\mathcal{N}(0, \frac{1}{2})$. The *traceless GUE*, denoted GUE_d^0, is the probability distribution on $d \times d$ Hermitian matrices $\boldsymbol{Y}$ given by

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

where $\boldsymbol{X} \sim \mathrm{GUE}_d$.

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

[1] This is a degenerate Gaussian distribution, supported on $\sum_i \boldsymbol{g}_i = 0$.

Fact 1.23. *Given $\boldsymbol{Y} \sim \mathrm{GUE}_d^0$, then as $d \to \infty$,*

$$\left(\frac{\mathrm{eig}_1(\boldsymbol{Y})}{\sqrt{d}}, \dots, \frac{\mathrm{eig}_d(\boldsymbol{Y})}{\sqrt{d}}\right)$$

converges almost surely to the semicircle law with density $\sqrt{4-x^2}/2\pi$, $-2 \leq x \leq 2$, where $\mathrm{eig}_i(\boldsymbol{Y})$ denotes the ith largest eigenvalue of $\boldsymbol{Y}$.

The traceless GUE was first used to characterize the Schur–Weyl distribution in the special case when α is the uniform distribution, the homogeneous random word case. In this case, Tracy and Widom [TW01] showed such a characterization for just the first row $\boldsymbol{\lambda}_1$, and Johansson [Joh01] extended their result to hold for the entire diagram $\boldsymbol{\lambda}$, as follows (cf. the quantum mechanical proof of this theorem by Kuperberg [Kup02]).

Theorem 1.24 ([Joh01])**.** *Let $\alpha = (\frac{1}{d}, \dots, \frac{1}{d})$ be the uniform distribution. Let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, and let $\boldsymbol{X} \sim \mathrm{GUE}_d^0$. Then as $n \to \infty$,*

$$\left(\frac{\boldsymbol{\lambda}_1 - n/d}{\sqrt{n/d}}, \dots, \frac{\boldsymbol{\lambda}_d - n/d}{\sqrt{n/d}}\right) \to (\mathrm{eig}_1(\boldsymbol{X}), \dots, \mathrm{eig}_d(\boldsymbol{X}))$$

in distribution.

Using Fact 1.23, we expect that for a typical $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda}_1 \approx \frac{n}{d} + 2\sqrt{n}, \qquad \boldsymbol{\lambda}_d \approx \frac{n}{d} - 2\sqrt{n}$$

and that the remaining $\boldsymbol{\lambda}_i$'s interpolate between these two values. (Let us also mention a line of work that has considered the case of uniform α in the nonasymptotic setting. Here, rather than fixing d and letting n tend towards infinity, d is allowed to grow to infinity while n scales as $n = O(d^2)$. In this case, Biane [Bia01] has shown a limiting theorem for the shape of $\boldsymbol{\lambda}$, and Méliot [Mél10] has characterized the fluctions of $\boldsymbol{\lambda}$ around its mean with a certain Gaussian process. The paper of Ivanov and Olshanski [IO02], which proves similar results for the Plancherel distribution, serves as an excellent introduction to this area.)

In the case of general α — the inhomogeneous random word case — it is convenient to group the indices $\{1, \dots, d\}$ according to the degeneracies of α.

Notation 1.25. Suppose there are m distinct values among the α_i's, and write $\alpha^{(k)}$ for the kth largest distinct value. We will block the indices as

$$[1, d] = [1, d^{(1)}] \cup [d^{(1)} + 1, d^{(1)} + d^{(2)}] \cup \cdots \cup [d - d^{(m)} + 1, d],$$

where every α_i in the kth block has the value $\alpha^{(k)}$. Given a partition λ of height d, we will write $\lambda_i^{(k)}$ for the ith index in the kth block, i.e. $\lambda_i^{(k)} = \lambda_{d_{<k}+i}$, where $d_{<k} = d^{(1)} + \cdots + d^{(k-1)}$. (We will only use this notation in this subsection; in particular, for Theorem 1.26.)

In the inhomogeneous case, Its, Tracy, and Widom [ITW01] gave a limiting characterization for the first row $\boldsymbol{\lambda}_1$, and Houdré and Xu [HX13], in a work that first appeared in 2009, extended their result to hold for the entire diagram $\boldsymbol{\lambda}$. Roughly, their characterization shows that within each block, $\boldsymbol{\lambda}$ acts GUE-like, as in Theorem 1.24, but across blocks $\boldsymbol{\lambda}$ acts Gaussian-like, as in Theorem 1.21. We cite here a related theorem of Méliot [Mél12], which cleanly decouples these two limiting effects.

Theorem 1.26 ([Mél12])**.** *Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a sorted probability distribution. Let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, let $(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m)$ be centered jointly Gaussian random variables with covariances $\delta_{k\ell} d^{(k)} - d^{(k)} d^{(\ell)} \sqrt{\alpha^{(k)} \alpha^{(\ell)}}$, and let $\boldsymbol{Y}^{(k)} \sim \mathrm{GUE}^0_{d^{(k)}}$, for each $k \in [m]$. Then as $n \to \infty$,*

$$\left\{ \frac{\boldsymbol{\lambda}_i^{(k)} - \alpha^{(k)} n}{\sqrt{\alpha^{(k)} n}} \right\}_{k \in [m], i \in [d^{(k)}]} \to \left\{ \frac{\boldsymbol{g}_k}{d^{(k)}} + \mathrm{eig}_i(\boldsymbol{Y}^{(k)}) \right\}_{k \in [m], i \in [d^{(k)}]}$$

in distribution.

Note that this theorem recovers Theorem 1.21 in the case when all the α_i's are distinct and Theorem 1.24 in the case when α is the uniform distribution. More generally, if we define $\boldsymbol{\lambda}[k] = \boldsymbol{\lambda}_1^{(k)} + \cdots + \boldsymbol{\lambda}_{d^{(k)}}^{(k)}$, then the random variables $(\boldsymbol{\lambda}[k] - \alpha^{(k)} d^{(k)} n)/\sqrt{\alpha^{(k)} d^{(k)} n}$ converge to Gaussian random variables with covariance $\delta_{k\ell} - \sqrt{\alpha^{(k)} d^{(k)} \alpha^{(\ell)} d^{(\ell)}}$. Hence, within blocks, $\boldsymbol{\lambda}$ experiences GUE fluctuations, whereas across blocks, $\boldsymbol{\lambda}$ experiences Gaussian fluctuations.

Theorem 1.26 predicts qualitatively different limiting behaviors between the cases when two α_i's are exactly equal and when two α_i's are unequal, even if they are close. Hence its convergence rate naturally depends on quantities like

$$\max_{i : \alpha_i \neq \alpha_{i+1}} \left(\frac{1}{\alpha_i - \alpha_{i+1}} \right),$$

and it is therefore not applicable in the nonasymptotic regime. Nevertheless, we have found it useful when reasoning about the Schur–Weyl distribution; in particular, by disregarding the Gaussian term in Theorem 1.26, we have our ansatz.

1.6 Future work

Bavarian, Mehraban, and Wright have used the techniques in this work to study the accuracy of the empirical entropy estimator for learning the von Neumann entropy. In preliminary work [BMW16], they have shown the following result:

Theorem 1.27. *The bias of the empirical entropy estimator satisfies*

$$H(\alpha) - \frac{3d^2}{2n} \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} H(\underline{\boldsymbol{\lambda}}) \leq H(\alpha).$$

Furthermore, the estimator has mean absolute error

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} |H(\underline{\boldsymbol{\lambda}}) - H(\alpha)| \leq \frac{3d^2}{2n} + \sqrt{\frac{2 + \log(d+e)^2}{n}}.$$

Hence, the empirical entropy is ϵ-close to the true von Neumann entropy with high probability when $n = O(d^2/\epsilon + \log(d)^2/\epsilon^2)$.

This gives an expression similar to both the bias and the mean absolute error of the classical empirical entropy estimator [WY16].

1.7 Organization

Section 2 contains the preliminaries, Section 3 contains the proof of Theorem 1.13, Section 4 contains our results on the concentration of the Schur-Weyl distribution, Section 5 contains our tomography results, and Section 6 contains our lower-rows majorization theorem.

Acknowledgments We would like to thank Mohammad Bavarian and Saeed Mehraban for allowing us to mention the entropy estimation result.

2 Preliminaries

Please refer to Section 2 of [OW15] for many of the definitions and notations used in this paper. We will also introduce additional notation in this section, and establish some simple results.

Notation 2.1. Given a sequence $\eta = (\eta_1, \dots, \eta_d)$ we write $\eta_{\le k} = \eta_1 + \cdots + \eta_k$ and we write $\eta_{>k} = \eta_{k+1} + \cdots + \eta_d$.

The following observation concerning Lipschitz constants of the RSK algorithm is very similar to one made in [BL12, Proposition 2.1]:

Proposition 2.2. *Suppose $w, w' \in [d]^n$ differ in exactly one coordinate. Write $\lambda = \mathrm{shRSK}(w)$, $\lambda' = \mathrm{shRSK}(w')$. Then:*

- $\left|\lambda_{\le k} - \lambda'_{\le k}\right| \le 1$ *for every $k \in [d]$.*
- $|\lambda_k - \lambda'_k| \le 2$ *for every $k \in [d]$.*

Proof. It suffices to prove the first statement; then, using the k and $k-1$ cases, we get the second statement via the triangle inequality. Also, by interchanging the roles of w and w', it suffices to prove $\lambda_{\le k} - \lambda'_{\le k} \le 1$. This follows from Greene's Theorem: $\lambda_{\le k}$ is the length of the longest disjoint union U of k increasing subsequences in w. If w' is formed by changing one letter in w', we can simply delete this letter from U (if it appears) and get a disjoint union of k increasing subsequences in w' of length at least $\lambda_{\le k} - 1$. But Greene's Theorem implies this is a lower bound on $\lambda'_{\le k}$. □

Remark 2.3. The bound of 2 in the second statement may be tight; e.g., $\mathrm{shRSK}(232122) = (4,1,1)$, $\mathrm{shRSK}(233122) = (3,3,0)$.

Proposition 2.4. *Let α, α' be probability distributions on $[d]$ and let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, $\boldsymbol{\lambda}' \sim \mathrm{SW}^n(\alpha')$. Then:*

- $\left|\mathbf{E}[\underline{\boldsymbol{\lambda}}_{\le k}] - \mathbf{E}[\underline{\boldsymbol{\lambda}}'_{\le k}]\right| \le d_{\mathrm{TV}}(\alpha, \alpha')$ *for every $k \in [d]$.*
- $\left|\mathbf{E}[\underline{\boldsymbol{\lambda}}_k] - \mathbf{E}[\underline{\boldsymbol{\lambda}}'_k]\right| \le 2d_{\mathrm{TV}}(\alpha, \alpha')$ *for every $k \in [d]$.*

Proof. Again, it suffices to prove the first statement, as the second one easily follows. Write $\epsilon = d_{\mathrm{TV}}(\alpha, \alpha')$. Thus there is a coupling $(\boldsymbol{a}, \boldsymbol{a}')$ such that $\boldsymbol{a} \sim \alpha$, $\boldsymbol{a}' \sim \alpha'$, and $\mathbf{Pr}[\boldsymbol{a} \ne \boldsymbol{a}'] = \epsilon$. Making n independent draws from the coupled distribution and calling the resulting words $(\boldsymbol{w}, \boldsymbol{w}')$, it follows that $\mathbf{E}[\triangle(\boldsymbol{w}, \boldsymbol{w}')] = \epsilon n$, where $\triangle$ denotes Hamming distance. Thus repeated application of Proposition 2.2 yields $\left|\mathbf{E}[\boldsymbol{\lambda}_{\le k}] - \mathbf{E}[\boldsymbol{\lambda'}_{\le k}]\right| \le \epsilon n$, where $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$, $\boldsymbol{\lambda}' = \mathrm{shRSK}(\boldsymbol{w}')$. But now $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$, $\boldsymbol{\lambda}' \sim \mathrm{SW}^n(\alpha')$, so the result follows after dividing through by n. □

The following lemma, while simple, is crucial for our nonasymptotic estimates:

Lemma 2.5. *Let α be a probability distribution on $[d]$ with $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_d$. Fix $k \in [d]$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} [\boldsymbol{\lambda}_{\le k} - \alpha_{\le k} n]$$

is a nondecreasing function of n.

Proof. We begin by "reversing" the alphabet $[d]$, so that $1 > 2 > \cdots > d$; recall that this does not change the distribution of $\boldsymbol{\lambda}$. Further, we will consider all n simultaneously by letting $\boldsymbol{\Lambda}$ be drawn from the Schur–Weyl process associated to $\boldsymbol{w} \sim \alpha^{\otimes\infty}$. Now

$$\mathbf{E}\left[\boldsymbol{\Lambda}^{(n)}_{\leq k}\right] = \sum_{t=1}^{n} \mathbf{Pr}[t^{\text{th}} \text{ letter of } \boldsymbol{w} \text{ creates a box in the first } k \text{ rows}].$$

If $\boldsymbol{w}_t \in [k]$ (i.e., it is among the k largest letters), then it will surely create a box in the first k rows. Since this occurs with probability $\alpha_{\leq k}$ for each t, we conclude that

$$\mathbf{E}\left[\boldsymbol{\Lambda}^{(n)}_{\leq k} - \alpha_{\leq k} n\right] = \sum_{t=1}^{n} \mathbf{Pr}[\boldsymbol{w}_t > k \text{ and it creates a box in the first } k \text{ rows}].$$

This is evidently a nondecreasing function of n. □

2.1 Distance measures

Definition 2.6. Let $\alpha, \beta \in \mathbb{R}^d$ be probability distributions. Then the truncated *Hellinger-squared distance* is given by

$$d^{(k)}_{\mathrm{H}^2}(\alpha,\beta) = d^{(k)}_{\mathrm{H}}(\alpha,\beta)^2 = \sum_{i=1}^{k} (\sqrt{\alpha_i} - \sqrt{\beta_i})^2,$$

and the $k = d$ case gives $d_{\mathrm{H}}(\alpha,\beta) = d^{(d)}_{\mathrm{H}}(\alpha,\beta)$ and $d_{\mathrm{H}^2}(\alpha,\beta) = d^{(d)}_{\mathrm{H}^2}(\alpha,\beta)$. The truncated *chi-squared divergence* is given by

$$d^{(k)}_{\chi^2}(\alpha,\beta) = \sum_{i=1}^{k} \beta_i \left(\frac{\alpha_i}{\beta_i} - 1\right)^2,$$

and the $k = d$ case gives $d_{\chi^2}(\alpha,\beta) = d^{(d)}_{\chi^2}(\alpha,\beta)$. The truncated ℓ_2^2 *distance* is given by

$$d^{(k)}_{\ell_2^2}(\alpha,\beta) = \sum_{i=1}^{k} (\alpha_i - \beta_i)^2,$$

and the $k = d$ case gives $d_{\ell_2^2}(\alpha,\beta) = d^{(d)}_{\ell_2^2}(\alpha,\beta)$. Finally, the *Kullback-Liebler (KL) divergence* is given by

$$d_{\mathrm{KL}}(\alpha,\beta) = \sum_{i=1}^{d} \alpha_i \ln\left(\frac{\alpha_i}{\beta_i}\right).$$

Proposition 2.7. *These distance measures are related as*

$$d_{\mathrm{H}^2}(\alpha,\beta) \leq d_{\chi^2}(\alpha,\beta), \quad d^{(k)}_{\mathrm{H}^2}(\alpha,\beta) \leq d^{(k)}_{\chi^2}(\alpha,\beta), \quad \textit{and } d_{\mathrm{KL}}(\alpha,\beta) \leq d_{\chi^2}(\alpha,\beta).$$

Proof. For the first and second inequalities, the bound follows term-by-term:

$$(\sqrt{\alpha_i} - \sqrt{\beta_i})^2 = \beta_i \left(\sqrt{\frac{\alpha_i}{\beta_i}} - 1\right)^2 \leq \beta_i \left(\sqrt{\frac{\alpha_i}{\beta_i}} - 1\right)^2 \left(\sqrt{\frac{\alpha_i}{\beta_i}} + 1\right)^2 = \beta_i \left(\frac{\alpha_i}{\beta_i} - 1\right)^2.$$

On the other hand, the third inequality is proven considering the whole sum at once:

$$\sum_{i=1}^{d} \alpha_i \ln\left(\frac{\alpha_i}{\beta_i}\right) \leq \sum_{i=1}^{d} \alpha_i \left(\frac{\alpha_i}{\beta_i} - 1\right) = \sum_{i=1}^{d} \frac{\alpha_i^2}{\beta_i} - 1,$$

where the inequality uses $\ln(x) \leq x - 1$ for all $x > 0$, and it can be checked that the right-most quantity is equal to $d_{\chi^2}(\alpha,\beta)$. □

Definition 2.8. Let ρ, σ be density matrices. The *fidelity* is given by $F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$. Related is the *affinity*, given by $A(\rho, \sigma) = \mathrm{tr}(\sqrt{\rho}\sqrt{\sigma})$. Finally, the *quantum Hellinger-squared distance* is given by

$$d_{\mathrm{H}^2}(\rho, \sigma) = d_{\mathrm{H}}(\rho, \sigma)^2 = \mathrm{tr}((\sqrt{\rho} - \sqrt{\sigma})^2) = 2 - 2A(\rho, \sigma).^{2}$$

By definition, $d_{\mathrm{H}}(\rho, \sigma) = \|\sqrt{\rho} - \sqrt{\sigma}\|_F$, and hence it satisfies the triangle inequality.

Proposition 2.9. *These distance measures are related as* $F(\rho, \sigma)^2 \leq A(\rho, \sigma) \leq F(\rho, \sigma)$*. As a result, if* $d_{\mathrm{H}^2}(\rho, \sigma) = \epsilon$*, then* $1 - \epsilon/2 \leq F(\rho, \sigma)$*.*

Proof. The upper bound $A(\rho, \sigma) \leq F(\rho, \sigma)$ is immediate, as $\mathrm{tr}(M) \leq \|M\|_1$ for any matrix M. As for the lower bound, it follows from Equations (28) and (32) from [ANSV08]. □

As a result, fidelity and affinity are essentially equivalent in the "$1 - \epsilon$" regime, and further it suffices to upper bound the Hellinger-squared distance if we want to lower bound the fidelity. For other properties of the affinity, see [LZ04, MM15]. Though we only ever apply the fidelity to density matrices, we will sometimes apply the affinity to arbitrary positive semidefinite matrices, as in Theorem 1.19.

3 Bounding the excess

In this section we will study the quantity

$$E_k^{(n)}(\alpha) = \mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} [\boldsymbol{\lambda}_{\leq k} - \alpha_{\leq k} n],$$

where α is a sorted probability distribution $[d]$, and $k \in [d]$. One way to think about this quantity is as

$$\mathop{\mathbf{E}}_{\boldsymbol{w} \sim \alpha^{\otimes n}} [\boldsymbol{\lambda}_{\leq k} - \boldsymbol{h}_{\leq k}],$$

where $\boldsymbol{\lambda} = \mathrm{shRSK}(\boldsymbol{w})$ and $\boldsymbol{h} = \mathrm{Histogram}(\boldsymbol{w})$, i.e. $\boldsymbol{h}_i$ is the number of i's in $\boldsymbol{w}$. By Greene's Theorem we know that $\boldsymbol{\lambda} \succ \boldsymbol{h}$ always; thus $E_k^{(n)}(\alpha) \geq 0$. We are therefore concerned with upper bounds, trying to quantify how "top-heavy" $\boldsymbol{\lambda}$ is on average (compared to a typical $\boldsymbol{h}$).

As we will see (and as implicitly shown in work of Its, Tracy, and Widom [ITW01]), the distribution of $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$ is very close to that of a certain modification of the multinomial distribution that favors top-heavy Young diagrams.

Definition 3.1. For a sorted probability distribution α with all α_i's distinct, define the function $\phi_{n,\alpha} : \mathbb{R}^n \to \mathbb{R}$ by

$$\phi_{n,\alpha}(h) = 1 + \sum_{1 \leq i < j \leq d} \frac{\alpha_j}{\alpha_i - \alpha_j} \left(\frac{h_i}{\alpha_i n} - \frac{h_j}{\alpha_j n} \right).$$

For $\boldsymbol{h} \sim \mathrm{Mult}(n, \alpha)$ we have $\mathbf{E}[\boldsymbol{h}_\ell] = \alpha_\ell n$; thus $\mathbf{E}[\phi_{n,\alpha}(\boldsymbol{h})] = 1$. We may therefore think of $\phi_{n,\alpha}(h)$ as a *relative density* with respect to the $\mathrm{Mult}(n, \alpha)$ distribution — except for the fact that we don't necessarily have $\phi_{n,\alpha}(h) \geq 0$ always. That will not bother us, though; we will only ever compute expectations relative to this density.

[2] We note that the quantum and classical Hellinger-squared distance are often defined with factors of $\frac{1}{2}$ in front. We have omitted them for simplicity.

Definition 3.2. We define the *modified α-multinomial (signed) distribution* on size-n, d-letter histograms h by $\phi_{n,\alpha}(h)M_{n,\alpha}(h)$, where $M_{n,\alpha}(h)$ is the probability of h under $\mathrm{Mult}(n,\alpha)$. We use the notation

$$\mathop{\mathbf{E}}_{\boldsymbol{h}\sim\mathrm{ModMult}(n,\alpha)}[F(\boldsymbol{h})] = \sum_h \phi_{n,\alpha}(h)M_{n,\alpha}(h)F(h) = \mathop{\mathbf{E}}_{\boldsymbol{h}\sim\mathrm{Mult}(n,\alpha)}[\phi_{n,\alpha}(\boldsymbol{h})F(\boldsymbol{h})].$$

As we will see in the proof of Theorem 3.7 below, for each $\lambda \vdash n$,

$$\text{``}\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[\boldsymbol{\lambda}=\lambda] \approx \mathop{\mathbf{Pr}}_{\boldsymbol{h}\sim\mathrm{ModMult}(n,\alpha)}[\boldsymbol{h}=\lambda]\text{''}. \tag{1}$$

Remark 3.3. The modified α-multinomial distribution is only defined when $\alpha_1 > \alpha_2 > \cdots > \alpha_d$. Note that under this condition, a draw $\boldsymbol{h} \sim \mathrm{Mult}(n,\alpha)$ will have $\boldsymbol{h}_1 \geq \boldsymbol{h}_2 \geq \cdots \geq \boldsymbol{h}_d$ with "very high" probability, and thus be a genuine partition $\boldsymbol{h} \vdash n$. (The "very high" here is only when n is sufficiently large compared to all of the $\frac{1}{\alpha_k-\alpha_{k+1}}$ values, though.)

The approximation (1) is consistent with the ansatz. One can see from the $\left(\frac{\lambda_i}{\alpha_i n} - \frac{\lambda_j}{\alpha_j n}\right)$ part of the formula for $\phi_{n,\alpha}(\lambda)$ that it emphasizes λ's that are "top-heavy". That is, it gives more probability to λ's that exceed their multinomial-expectation at low indices and fall short of their multinomial-expectation at high indices. Furthermore, one can see from the $\frac{\alpha_j}{\alpha_i-\alpha_j}$ part of the formula that this effect becomes more pronounced when two or more α_ℓ's tend toward equality.

The utility of (1) is that we can compute certain expectations under the modified multinomial distribution easily and exactly, since it has a simple formula. Of course, we have to concern ourselves with the approximation in (1); in fact, the error can be quite unpleasant in that it depends on d, and even worse, on the gaps $\alpha_k - \alpha_{k+1}$. Nonetheless, when it comes to using (1) to estimate $E_k^{(n)}(\alpha)$, we will see that the increasing property (Lemma 2.5) will let us evade the approximation error. Toward that end, let us make a definition and some calculations:

Notation 3.4. For any sorted probability distribution α on $[d]$ and any $k \in [d]$ we write

$$\mathrm{Excess}_k(\alpha) = \sum_{i \leq k < j} \frac{\alpha_j}{\alpha_i - \alpha_j}.$$

Remark 3.5. We have $\mathrm{Excess}_k(\alpha) = 0$ if $k = d$, and otherwise $\mathrm{Excess}_k(\alpha)$ is continuous away from $\alpha_k = \alpha_{k+1}$, where it blows up to ∞. We also have the following trivial bound, which is useful if the gap $\alpha_k - \alpha_{k+1}$ is large:

$$\mathrm{Excess}_k(\alpha) \leq k\alpha_{>k}/(\alpha_k - \alpha_{k+1}). \tag{2}$$

Although their proof was a little more elaborate, Its, Tracy, and Widom [ITW01] proved the following result in the special case of $k = 1$:

Proposition 3.6. *If α is a sorted probability distribution on $[d]$ with all α_i's distinct, then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{ModMult}(n,\alpha)}\left[E_k^{(n)}(\alpha)\right] = \mathrm{Excess}_k(\alpha).$$

Proof. By definition we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{ModMult}(n,\alpha)}\left[\boldsymbol{\lambda}_k - \alpha_k n\right] = \mathop{\mathbf{E}}_{\boldsymbol{h}\sim\mathrm{Mult}(n,\alpha)}[\phi_{n,\alpha}(\boldsymbol{h})(\boldsymbol{h}_k - \alpha_k n)].$$

It's convenient to write

$$\phi_{n,\alpha}(h) = 1 + \sum_{1\le i<j\le d} \frac{\alpha_j}{\alpha_i - \alpha_j}\left(\frac{h_i - \alpha_i n}{\alpha_i n} - \frac{h_j - \alpha_j n}{\alpha_j n}\right).$$

Then using the fact that for $\boldsymbol{h} \sim \text{Mult}(n,\alpha)$ we have $\mathbf{E}[\boldsymbol{h}_k] = \alpha_k n$, $\mathbf{Var}[\boldsymbol{h}_k] = \alpha_k(1-\alpha_k)n$, $\mathbf{Cov}[\boldsymbol{h}_k, \boldsymbol{h}_\ell] = -\alpha_k\alpha_\ell n$, we easily obtain:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\text{ModMult}(n,\alpha)}\big[\boldsymbol{\lambda}_k - \alpha_k n\big] = \sum_{j>k} \frac{\alpha_j}{\alpha_k - \alpha_j} - \sum_{i<k} \frac{\alpha_k}{\alpha_i - \alpha_k}.$$

The result follows. □

We now come to the main result of this section:

Theorem 3.7. *Let α be a sorted probability distribution on $[d]$ and let $k \in [d]$. Then for all $n \in \mathbb{N}$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\text{SW}^n(\alpha)}[\boldsymbol{\lambda}_{\le k} - \alpha_{\le k} n] \le \text{Excess}_k(\alpha).$$

Furthermore, using the notation $E_k^{(n)}(\alpha)$ for the left-hand side, it holds that $E_k^{(n)}(\alpha) \nearrow \text{Excess}_k(\alpha)$ as $n \to \infty$ provided that all α_i's are distinct.

Remark 3.8. We expect that $E_k^{(n)}(\alpha) \nearrow \text{Excess}_k(\alpha)$ for all α; however we did not prove this.

Proof. Lemma 2.5 tells us that $E_k^{(n)}(\alpha)$ is nondecreasing in n for all α and k; thus $E_k^{(n)}(\alpha) \nearrow L_k(\alpha)$ for some $L_k(\alpha) \in \mathbb{R} \cup \{\infty\}$. The main claim that will complete the proof is the following (the $k=1$ case of which was proven in [ITW01]):

Claim 3.9. *For fixed α and k,*

$$E_k^{(n)}(\alpha) = \text{Excess}_k(\alpha) \pm O(1/\sqrt{n}) \quad \textit{provided the } \alpha_i\textit{'s are distinct,}$$

where the constant hidden in the $O(\cdot)$ may depend on α in an arbitrary way.

This claim establishes $L_k(\alpha) = \text{Excess}_k(\alpha)$ whenever the α_i's are all distinct. It remains to observe that when the α_i's are not all distinct, $L_k(\alpha) > \text{Excess}_k(\alpha)$ is impossible; this is because $\text{Excess}_k(\alpha) - E_k^{(n)}(\alpha)$ is a continuous function of α for each fixed n (unless $\alpha_k = \alpha_{k+1}$, but in this case $\text{Excess}_k(\alpha) = \infty$ and there is nothing to prove).

We now focus on proving Claim 3.9, following the analysis in [ITW01]. We emphasize that the α_i's are now assumed distinct, and the constants hidden in all subsequent $O(\cdot)$ notation may well depend on the α_i's.

As computed in [ITW01, top of p. 255], for each $\lambda \vdash n$ we have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\text{SW}^n(\alpha)}[\boldsymbol{\lambda} = \lambda] = \left(1 + \frac{1}{\sqrt{n}}\left(\sum_{1\le i<j\le d} \sqrt{\frac{\alpha_j}{\alpha_i}}\,\frac{\sqrt{\alpha_j}\boldsymbol{\xi}_i - \sqrt{\alpha_i}\boldsymbol{\xi}_j}{\alpha_i - \alpha_j}\right) \pm O(\tfrac{1}{n})\right) M_{n,\alpha}(\lambda) \pm e^{-\Omega(n)}, \tag{3}$$

where $\boldsymbol{\xi}_\ell = (\boldsymbol{\lambda}_\ell - \alpha_\ell n)/\sqrt{\alpha_\ell n}$. If we now simply substitute in the definition of $\boldsymbol{\xi}_\ell$ and do some simple arithmetic, we indeed get the following precise form of (1):

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\text{SW}^n(\alpha)}[\boldsymbol{\lambda} = \lambda] = \phi_{n,\alpha}(\lambda) M_{n,\alpha}(\lambda) \pm O\big(\tfrac{M_{n,\alpha}(\lambda)}{n}\big) \pm e^{-\Omega(n)}. \tag{4}$$

Given this, let F be any functional on partitions of n that is subexponentially bounded in n (meaning $|F(\lambda)| \leq e^{o(n)}$ for all $\lambda \vdash n$). Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[F(\boldsymbol{\lambda})] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{ModMult}(n,\alpha)}[\mathbf{1}_{\{\boldsymbol{\lambda}\text{ is sorted}\}}\cdot F(\boldsymbol{\lambda})] \\ \pm O(\tfrac{1}{n})\cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{Mult}(n,\alpha)}[\mathbf{1}_{\{\boldsymbol{\lambda}\text{ is sorted}\}}\cdot |F(\boldsymbol{\lambda})|] \pm e^{-\Omega(n)},$$

where in the final error $e^{\Omega(n)}$ we used the subexponential bound on $|F(\lambda)|$ and also absorbed a factor of $e^{O(\sqrt{n})}$, the number of partitions of n. We can further simplify this: Using $\alpha_1 > \alpha_2 > \cdots > \alpha_d$, an easy Chernoff/union bound gives that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\lambda}\sim\mathrm{Mult}(n,\alpha)}[\boldsymbol{\lambda}\text{ is not sorted}] \leq e^{\Omega(n)} \tag{5}$$

(where certainly the constant in the $\Omega(\cdot)$ depends on all the gaps $\alpha_\ell - \alpha_{\ell+1}$). Thus

$$\left|\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{ModMult}(n,\alpha)}[\mathbf{1}_{\{\boldsymbol{\lambda}\text{ is unsorted}\}}\cdot F(\boldsymbol{\lambda})]\right| = \left|\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{Mult}(n,\alpha)}[\mathbf{1}_{\{\boldsymbol{\lambda}\text{ is unsorted}\}}\cdot \phi_{n,\alpha}(\boldsymbol{\lambda})F(\boldsymbol{\lambda})]\right| \\ \leq \sqrt{\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{Mult}(n,\alpha)}[\mathbf{1}^2_{\{\boldsymbol{\lambda}\text{ is unsorted}\}}]}\sqrt{\phi_{n,\alpha}(\boldsymbol{\lambda})^2F(\boldsymbol{\lambda})^2} \leq e^{-\Omega(n)}, \tag{6}$$

where we used (5), the subexponential bound on F, and $\phi_{n,\alpha}(\boldsymbol{\lambda}) \leq O(1)$. A similar but simpler analysis applies to the first middle term in (6), and we conclude the following attractive form of (1) for subexponentially-bounded F:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[F(\boldsymbol{\lambda})] = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{ModMult}(n,\alpha)}[F(\boldsymbol{\lambda})] \qquad \pm O(\tfrac{1}{n})\cdot \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{Mult}(n,\alpha)}[|F(\boldsymbol{\lambda})|] \pm e^{-\Omega(n)}.$$

Finally, Claim (3.9) now follows from Proposition 3.6, together with the fact that for $\boldsymbol{\lambda} \sim \mathrm{Mult}(n,\alpha)$,

$$\mathbf{E}[|E_k^{(n)}(\boldsymbol{\lambda})|] \leq \sum_{i=1}^k \sqrt{\mathbf{E}[(\boldsymbol{\lambda}_i - \alpha_i n)^2]} = \sum_{i=1}^k \mathbf{stddev}[\boldsymbol{\lambda}_i] = \sum_{i=1}^k \sqrt{n}\sqrt{\alpha_i(1-\alpha_i)} = O(\sqrt{n}). \qquad \square$$

4 Convergence of the Schur–Weyl distribution

In this section, we derive consequences of Theorem 1.13 and Theorem 1.12. To begin, it will help to define two restrictions of a word w.

Notation 4.1. Let $w \in [d]^n$ and let $\lambda = \mathrm{shRSK}(w)$. We use boldface $\boldsymbol{w}$ if $\boldsymbol{w} \sim \alpha^{\otimes n}$, in which case $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$.

- Write $w^{(k..)}$ for the string formed from w by deleting all letters smaller than k, and let $\lambda^{(k..)} = \mathrm{shRSK}(w^{(k..)})$. Then the random variable $\boldsymbol{\lambda}^{(k..)}$ is distributed as $\mathrm{SW}^{\boldsymbol{\ell}}(\underline{\alpha[k{:}]})$, where $\boldsymbol{\ell} \sim \mathrm{Binomial}(n, \alpha_{\geq k})$ and $\underline{\alpha[k{:}]} = (\alpha_i/\alpha_{\geq k})_{i=k}^d$.

- Write $w^{(..k)}$ for the string formed from w by deleting all letters larger than k, and let $\lambda^{(..k)} = \mathrm{shRSK}(w^{(..k)})$. Note that if $(P,Q) = \mathrm{RSK}(w)$, then $\lambda^{(..k)}$ is the shape of the diagram formed by deleting all boxes containing letters larger than k from P, and hence $\lambda_i^{(..k)} \leq \lambda_i$ for all i. Then the random variable $\boldsymbol{\lambda}^{(..k)}$ is distributed as $\mathrm{SW}^{\boldsymbol{m}}(\underline{\alpha[{:}k]})$, where $\boldsymbol{m} \sim \mathrm{Binomial}(n, \alpha_{\leq k})$ and $\underline{\alpha[{:}k]} = (\alpha_i/\alpha_{\leq k})_{i=1}^k$.

We will mainly use the following weaker version of Theorem 6.3.

Theorem 4.2. *Let $\lambda[k{:}]$ denote the Young diagram formed by rows $k, k+1, k+2, \dots$ of λ. Then $\lambda^{(k..)} \trianglerighteq_w \lambda[k{:}]$.*

Proof. This follows by applying Theorem 1.12 to w and noting that the string $\overline{w}$ in that theorem is a substring of $w^{(k..)}$. Hence weak majorization holds trivially. □

4.1 Bounds on the first and last rows

Theorem 4.3. *Let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_1 \leq \alpha_1 n + 2\sqrt{n}.$$

Proof. Write $g = 1/\sqrt{n}$. We assume that $\alpha_1 + 2g \leq 1$, as otherwise the theorem is vacuously true. Let $\beta \in \mathbb{R}^d$ be a sorted probability distribution for which $\beta_1 = \alpha_1 + g$, $\beta_2 \leq \alpha_2$, and $\beta \succ \alpha$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_1 \leq \mathop{\mathbf{E}}_{\boldsymbol{\mu}\sim\mathrm{SW}^n(\beta)} \boldsymbol{\mu}_1 \leq \beta_1 n + \sum_{j>1} \frac{\beta_j}{\beta_1 - \beta_j} \leq \beta_1 n + \frac{1}{g} = \alpha_1 n + ng + \frac{1}{g} = \alpha_1 n + 2\sqrt{n},$$

where the first step is by Theorem 1.14 and the second is by Theorem 1.13. □

Theorem 4.4. *Let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_d \geq \alpha_d n - 2\sqrt{\alpha_d d n}.$$

Proof. Write $g = \sqrt{\alpha_d d/n}$. We assume that $\alpha_d - 2g \geq 0$, as otherwise the theorem is vacuously true. Let $\beta \in \mathbb{R}^d$ be a sorted probability distribution for which $\beta_d = \alpha_d - g$, $\beta_{d-1} \geq \alpha_{d-1}$, and $\beta \succ \alpha$. Then

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} [\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_{d-1}] \leq \mathop{\mathbf{E}}_{\boldsymbol{\mu}\sim\mathrm{SW}^n(\beta)} [\boldsymbol{\mu}_1 + \cdots + \boldsymbol{\mu}_{d-1}] &\leq \beta_1 n + \cdots + \beta_{d-1} n + \sum_{i<d} \frac{\beta_d}{\beta_i - \beta_d} \\ &\leq \beta_1 n + \cdots + \beta_{d-1} n + \frac{d\alpha_d}{g} = \alpha_1 n + \cdots + \alpha_{d-1} n + 2\sqrt{\alpha_d d n},\end{aligned}$$

where the first inequality is by Theorem 1.14, and the second inequality is by Theorem 1.13. As $\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_d = n$, this implies that $\mathbf{E}\,\boldsymbol{\lambda}_d \geq \alpha_d n - 2\sqrt{\alpha_d d n}$. □

We note that Theorem 1.14 can be replaced by Proposition 2.4 at just a constant-factor expense.

4.2 Bounds for all rows

Theorem 4.5. *Let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then*

$$\alpha_k n - 2\sqrt{\alpha_k k n} \leq \mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_k \leq \alpha_k n + 2\sqrt{\alpha_{\geq k} n}.$$

Proof. For the upper bound, we use Theorem 4.2:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_k \leq \mathop{\mathbf{E}}_{\boldsymbol{\ell},\boldsymbol{\lambda}^{(k..)}} \boldsymbol{\lambda}_1^{(k..)} \leq \mathop{\mathbf{E}}_{\boldsymbol{\ell}} \left[\underline{\alpha[k{:}]}_k \boldsymbol{\ell} + 2\sqrt{\boldsymbol{\ell}}\right] \leq \underline{\alpha[k{:}]}_k \,\mathbf{E}\,\boldsymbol{\ell} + 2\sqrt{\mathbf{E}\,\boldsymbol{\ell}} \leq \alpha_k n + 2\sqrt{\alpha_{\geq k} n},$$

where the second step is by Theorem 4.3 and the third is by Jensen's inequality.

For the lower bound, we use the fact that $\boldsymbol{\lambda}_k \geq \boldsymbol{\lambda}_k^{(..k)}$:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \boldsymbol{\lambda}_k \geq \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{\lambda}^{(..k)}} \boldsymbol{\lambda}_k^{(..k)} \geq \mathop{\mathbf{E}}_{\boldsymbol{m}} \left[\underline{\alpha[:k]}_k \boldsymbol{m} - 2\sqrt{\underline{\alpha[:k]}_k k\boldsymbol{m}}\right] \geq \underline{\alpha[:k]}_k \mathbf{E}\,\boldsymbol{m} - 2\sqrt{\underline{\alpha[:k]}_k k\,\mathbf{E}\,\boldsymbol{m}} = \alpha_k n - 2\sqrt{\alpha_k kn},$$

where the second inequality is by Theorem 4.4, and the third is by Jensen's inequality. □

Theorem 1.4 follows from the fact that $\alpha_k k, \alpha_{\geq k} \leq \min\{1, \alpha_k d\}$.

4.3 Chi-squared spectrum estimation

Theorem 4.6. *Let $\alpha \in \mathbb{R}^d$ be a sorted probability distribution. Then for any $k \in [d]$,*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \sum_{i=k}^{d} \boldsymbol{\lambda}_i^2 \leq \sum_{i=k}^{d} (\alpha_i n)^2 + d\alpha_{\geq k} n.$$

Proof. When $k = 1$, this statement is equivalent to Lemma 3.1 from [OW16]. Hence, we may assume $k > 1$. By Theorem 4.2,

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} \sum_{i=k}^{d} \boldsymbol{\lambda}_i^2 \leq \mathop{\mathbf{E}}_{\boldsymbol{\ell},\boldsymbol{\lambda}^{(k..)}} \sum_{i=1}^{d-k+1} (\boldsymbol{\lambda}_i^{(k..)})^2 &\leq \mathop{\mathbf{E}}_{\boldsymbol{\ell}} \left[\sum_{i=k}^{d} (\underline{\alpha[k:]}_i \boldsymbol{\ell})^2 + (d-k+1)\boldsymbol{\ell}\right] \\ &= \sum_{i=k}^{d} (\alpha_i n)^2 + \sum_{i=k}^{d} \alpha_i^2 n \left(\frac{1}{\alpha_{\geq k}} - 1\right) + (d-k+1)\alpha_{\geq k} n \leq \sum_{i=k}^{d} (\alpha_i n)^2 + d\alpha_{\geq k} n.\end{aligned}$$

Here the second inequality used Lemma 3.1 from [OW16], and the third inequality used $\alpha_{\geq k} \geq \alpha_i$ and $k > 1$. □

Theorem 4.7. $\displaystyle\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} d_{\chi^2}(\underline{\boldsymbol{\lambda}}, \alpha) \leq \frac{d^2}{n}$.

Proof. Write the expectation as

$$\mathbf{E}\, d_{\chi^2}(\underline{\boldsymbol{\lambda}}, \alpha) = \frac{1}{n^2} \cdot \mathbf{E} \sum_{i=1}^{k} \frac{\boldsymbol{\lambda}_i^2}{\alpha_i} - 1.$$

To upper bound the expectation, we can apply Theorem 4.6.

$$\begin{aligned}\mathbf{E} \sum_{i=1}^{d} \frac{\boldsymbol{\lambda}_i^2}{\alpha_i} = \sum_{i=1}^{d} \left(\frac{1}{\alpha_i} - \frac{1}{\alpha_{i-1}}\right) \cdot \mathbf{E} \sum_{j=i}^{d} \boldsymbol{\lambda}_j^2 &\leq \sum_{i=1}^{d} \left(\frac{1}{\alpha_i} - \frac{1}{\alpha_{i-1}}\right) \cdot \sum_{j=i}^{d} \left((\alpha_j n)^2 + d\alpha_j n\right) \\ &= \sum_{j=1}^{d} \left((\alpha_j n)^2 + d\alpha_j n\right) \cdot \sum_{i=1}^{j} \left(\frac{1}{\alpha_i} - \frac{1}{\alpha_{i-1}}\right) \\ &= \sum_{j=1}^{d} \left((\alpha_j n)^2 + d\alpha_j n\right) \cdot \frac{1}{\alpha_j} = n^2 + d^2 n.\end{aligned}$$

Dividing through by n^2 and subtracting one completes the proof. □

Combined with Proposition 2.7, Theorem 4.7 implies Theorem 1.7.

4.4 Concentration bounds

In this section, we show that each row $\boldsymbol{\lambda}_i$ concentrates exponentially around its mean. We do so using the method of bounded differences.

Proposition 4.8. *Let $\alpha \in \mathbb{R}^d$ be a probability distribution. Then for any $k \in [d]$,*

$$\mathop{\mathbf{Var}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)}[\boldsymbol{\lambda}_k] \leq 16n.$$

Proof. Let $\boldsymbol{w} \sim \alpha^{\otimes n}$, and consider the martingale $\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(n)}$ defined as

$$\boldsymbol{X}^{(i)} := \mathbf{E}[\boldsymbol{\lambda}_k \mid \boldsymbol{w}_1, \ldots, \boldsymbol{w}_i].$$

Note that $\boldsymbol{X}^{(0)} = \mathbf{E}\,\boldsymbol{\lambda}_k$ and $\boldsymbol{X}^{(n)} = \mathrm{shRSK}(\boldsymbol{w})_k$. Furthermore, by Proposition 2.2, we have that $|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(i-1)}| \leq 2$ always, for all $i \in [n]$. Thus, if we write $\nu_k := \mathbf{E}\,\boldsymbol{\lambda}_k = \boldsymbol{X}^{(0)}$, then by Azuma's inequality

$$\mathbf{Pr}[|\boldsymbol{\lambda}_k - \nu_k| \geq t] \leq 2\exp\left(\frac{-t^2}{8n}\right).$$

We can therefore calculate $\mathbf{Var}[\boldsymbol{\lambda}_k]$ as

$$\mathbf{E}\,(\boldsymbol{\lambda}_k - \nu_k)^2 = \int_{t=0}^{\infty} 2t \cdot \mathbf{Pr}[|\boldsymbol{\lambda}_k - \nu_k| \geq t] \cdot \mathrm{d}t \leq \int_{t=0}^{\infty} 4t \exp\left(\frac{-t^2}{8n}\right) \cdot \mathrm{d}t = -16n \exp\left(\frac{-t^2}{8n}\right)\Bigg|_{t=0}^{\infty} = 16n.$$

□

4.5 Truncated spectrum estimation

Lemma 4.9. *Let $1 \leq i \leq k \leq d$. Then*

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)} (\boldsymbol{\lambda}_i - \alpha_i n)^2 \leq 2 \mathop{\mathbf{E}}_{\boldsymbol{m}, \boldsymbol{\lambda}^{(..k)}} (\boldsymbol{\lambda}_i^{(..k)} - \underline{\alpha[:k]}_i \boldsymbol{m})^2 + 44\alpha_{\geq i} n.$$

Proof. Write $\mathcal{G}$ for the event that $\boldsymbol{\lambda}_i \geq \alpha_i n$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_i - \alpha_i n)^2 = \mathop{\mathbf{E}}_{\boldsymbol{\lambda}}[(\boldsymbol{\lambda}_i - \alpha_i n)^2 \cdot \mathbf{1}[\mathcal{G}]] + \mathop{\mathbf{E}}_{\boldsymbol{\lambda}}[(\boldsymbol{\lambda}_i - \alpha_i n)^2 \cdot \mathbf{1}[\overline{\mathcal{G}}]]. \tag{7}$$

When $\mathcal{G}$ occurs, then $(\boldsymbol{\lambda}_i - \alpha_i n)^2 \leq (\boldsymbol{\lambda}_1^{(i..)} - \alpha_i n)^2$. Hence

$$\begin{aligned}\mathbf{E}[(\boldsymbol{\lambda}_i - \alpha_i n)^2 \cdot \mathbf{1}[\mathcal{G}]] &\leq \mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \alpha_i n)^2 = \mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \underline{\alpha[i:]}_1 \boldsymbol{\ell} + \underline{\alpha[i:]}_1 \boldsymbol{\ell} - \alpha_i n)^2 \\ &\leq 2\,\mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 + 2\,\mathbf{E}(\underline{\alpha[i:]}_1 \boldsymbol{\ell} - \alpha_i n)^2 \leq 2\,\mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 + \frac{2n\alpha_i^2}{\alpha_{\geq i}},\end{aligned}$$

where the second inequality uses $(x+y)^2 \leq 2x^2 + 2y^2$ for all $x, y \in \mathbb{R}$, and the third inequality is because $\boldsymbol{\ell}$ is distributed as Binomial$(n, \alpha_{\geq i})$. Given $\boldsymbol{\ell}$, define $\boldsymbol{\nu} = \mathbf{E}[\boldsymbol{\lambda}_1^{(i..)} \mid \boldsymbol{\ell}]$. Then

$$\begin{aligned}&\mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 = \mathbf{E}(\boldsymbol{\lambda}_1^{(i..)} - \boldsymbol{\nu} + \boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 \\ &= \mathbf{E}[(\boldsymbol{\lambda}_1^{(i..)} - \boldsymbol{\nu})^2 + (\boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 + 2(\boldsymbol{\lambda}_1^{(i..)} - \boldsymbol{\nu})(\boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})] = \mathbf{E}[(\boldsymbol{\lambda}_1^{(i..)} - \boldsymbol{\nu})^2 + (\boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2] \\ &\qquad \leq 16 \mathop{\mathbf{E}}_{\boldsymbol{\ell}} \boldsymbol{\ell} + \mathop{\mathbf{E}}_{\boldsymbol{\ell}}(\boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2 = 16\alpha_{\geq i} n + \mathop{\mathbf{E}}_{\boldsymbol{\ell}}(\boldsymbol{\nu} - \underline{\alpha[i:]}_1 \boldsymbol{\ell})^2,\end{aligned} \tag{8}$$

where the inequality uses Proposition 4.8. Next, we note that because $\boldsymbol{\lambda}^{(i..)}$ is distributed as $\mathrm{SW}^{\boldsymbol{\ell}}(\underline{\alpha[i{:}]})$, $\boldsymbol{\nu}$ is at least $\underline{\alpha[i{:}]}_1\boldsymbol{\ell}$. Hence, to upper-bound $(\boldsymbol{\nu}-\underline{\alpha[i{:}]}_1\boldsymbol{\ell})^2$ we must upper-bound $\boldsymbol{\nu}$, and this can be done by Theorem 4.3: $\boldsymbol{\nu}=\mathbf{E}[\boldsymbol{\lambda}_1^{(i..)}\mid\boldsymbol{\ell}]\le\underline{\alpha[i{:}]}_1\boldsymbol{\ell}+2\sqrt{\boldsymbol{\ell}}$. Thus, (8) can be bounded as

$$16\alpha_{\geq i}n+\mathop{\mathbf{E}}_{\boldsymbol{\ell}}(\boldsymbol{\nu}-\underline{\alpha[i{:}]}_1\boldsymbol{\ell})^2\leq 16\alpha_{\geq i}n+\mathop{\mathbf{E}}_{\boldsymbol{\ell}}4\boldsymbol{\ell}=20\alpha_{\geq i}n.$$

In summary, the term in (7) corresponding to $\mathcal{G}$ is at most $42\alpha_{\geq i}n$.

As for the other term, when $\mathcal{G}$ does not occur, then $(\boldsymbol{\lambda}_i-\alpha_i n)^2\leq(\boldsymbol{\lambda}_i^{(..k)}-\alpha_i n)^2$. Hence

$$\begin{aligned}\mathbf{E}[(\boldsymbol{\lambda}_i-\alpha_i n)^2\cdot\mathbf{1}[\overline{\mathcal{G}}]]&\leq\mathbf{E}(\boldsymbol{\lambda}_i^{(..k)}-\alpha_i n)^2=\mathbf{E}(\boldsymbol{\lambda}_i^{(..k)}-\underline{\alpha[{:}k]}_i\boldsymbol{m}+\underline{\alpha[{:}k]}_i\boldsymbol{m}-\alpha_i n)^2\\&\leq 2\,\mathbf{E}(\boldsymbol{\lambda}_i^{(..k)}-\underline{\alpha[{:}k]}_i\boldsymbol{m})^2+2\,\mathbf{E}(\underline{\alpha[{:}k]}_i\boldsymbol{m}-\alpha_i n)^2\leq 2\,\mathbf{E}(\boldsymbol{\lambda}_i^{(..k)}-\underline{\alpha[{:}k]}_i\boldsymbol{m})^2+\frac{2n\alpha_i^2}{\alpha_{\leq k}},\end{aligned}$$

where the second inequality uses $(x+y)^2\leq 2x^2+2y^2$ for all $x,y\in\mathbb{R}$, and the third inequality is because $\boldsymbol{m}$ is distributed as $\mathrm{Binomial}(n,\alpha_{\leq k})$. As $\alpha_i^2/\alpha_{\leq k}\leq\alpha_{\geq i}$, the proof is complete. □

Theorem 4.10. $\displaystyle\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} d^{(k)}_{\ell_2^2}(\underline{\boldsymbol{\lambda}},\alpha)\leq\frac{46k}{n}.$

Proof. Applying Lemma 4.9 with $j=k$ for all $i\in[k]$,

$$\begin{aligned}n^2\,\mathbf{E}\,d^{(k)}_{\ell_2^2}(\underline{\boldsymbol{\lambda}},\alpha)&=\mathbf{E}\sum_{i=1}^{k}(\boldsymbol{\lambda}_i-\alpha_i n)^2\leq 2\,\mathbf{E}\sum_{i=1}^{k}(\boldsymbol{\lambda}_i^{(..k)}-\underline{\alpha[{:}k]}_i\boldsymbol{m})^2+44kn\\&=2\mathop{\mathbf{E}}_{\boldsymbol{m}}\left[\boldsymbol{m}^2\mathop{\mathbf{E}}_{\boldsymbol{\lambda}^{(..k)}}\|\underline{\boldsymbol{\lambda}}^{(..k)}-\underline{\alpha[{:}k]}\|_2^2\right]+44kn\leq 2\mathop{\mathbf{E}}_{\boldsymbol{m}}\left[\boldsymbol{m}^2\left(\frac{k}{\boldsymbol{m}}\right)\right]+44kn\leq 46kn,\end{aligned}$$

where the second inequality is by Theorem 1.1 of [OW16]. The theorem follows by dividing through by n^2. □

Theorem 4.11. $\displaystyle\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)} d^{(k)}_{\chi^2}(\underline{\boldsymbol{\lambda}},\alpha)\leq\frac{46kd}{n}.$

Proof. Applying Lemma 4.9 with $j=k$ for all $i\in[k]$,

$$\begin{aligned}n^2\,\mathbf{E}\,d^{(k)}_{\chi^2}(\underline{\boldsymbol{\lambda}},\alpha)&=\mathbf{E}\sum_{i=1}^{k}\frac{1}{\alpha_i}(\boldsymbol{\lambda}_i-\alpha_i n)^2\leq 2\,\mathbf{E}\sum_{i=1}^{k}\frac{1}{\alpha_i}(\boldsymbol{\lambda}_i^{(..k)}-\underline{\alpha[{:}k]}_i\boldsymbol{m})^2+\sum_{i=1}^{k}\frac{44\alpha_{\geq i}n}{\alpha_i}\\&\leq 2\mathop{\mathbf{E}}_{\boldsymbol{m}}\left[\frac{\boldsymbol{m}^2}{\alpha_{\leq k}}\mathop{\mathbf{E}}_{\boldsymbol{\lambda}^{(..k)}}d_{\chi^2}(\underline{\boldsymbol{\lambda}}^{(..k)},\underline{\alpha[{:}k]})\right]+44kdn\leq 2\mathop{\mathbf{E}}_{\boldsymbol{m}}\left[\frac{\boldsymbol{m}^2}{\alpha_{\leq k}}\cdot\frac{k^2}{\boldsymbol{m}}\right]+44kdn=2k^2n+44kdn,\end{aligned}$$

where the second inequality is because $\alpha_{\geq i}\leq\alpha_i d$, and the third inequality is by Theorem 4.7. The theorem follows from $k\leq d$ and by dividing through by n^2. □

4.6 Mean squared error

Theorem 4.12. $\displaystyle\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}(\boldsymbol{\lambda}_k-\alpha_k n)^2\leq 42\alpha_k kn+42\alpha_{\geq k}n.$

Proof. Following the proof of Lemma 4.9 for $i = k$, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_k - \alpha_i n)^2 \le \mathbf{E}[(\boldsymbol{\lambda}^{(..k)}_k - \alpha_k n)^2 \cdot \mathbf{1}[\overline{\mathcal{G}}]] + 42\alpha_{\ge k} n, \tag{9}$$

where $\mathcal{G}$ is the event that $\boldsymbol{\lambda}_k \ge \alpha_i n$. Now we borrow a step from the proof of Lemma 5.1 in [OW16]. Because it has support size k, $\underline{\alpha[{:}k]}$ can be expressed as a mixture

$$\underline{\alpha[{:}k]} = p_1 \cdot \mathcal{D}_1 + p_2 \cdot \mathcal{D}_2, \tag{10}$$

of a certain distribution $\mathcal{D}_1$ supported on $[k-1]$ and the uniform distribution $\mathcal{D}_2$ on $[k]$. It can be checked that $p_2 = \underline{\alpha[{:}k]}_k k$. We may therefore think of a draw $\boldsymbol{\lambda}^{(..k)}$ from $\mathrm{SW}^{\boldsymbol{m}}(\underline{\alpha[{:}k]})$ occurring as follows. First, $[\boldsymbol{m}]$ is partitioned into two subsets $\boldsymbol{I}_1, \boldsymbol{I}_2$ by including each $i \in [\boldsymbol{m}]$ into $\boldsymbol{I}_j$ independently with probability p_j. Next we draw strings $\boldsymbol{w}^{(j)} \sim \mathcal{D}_j^{\otimes \boldsymbol{I}_j}$ independently for $j \in [2]$. Finally, we let $\boldsymbol{w}^{(..k)} = (\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}) \in [d]^n$ be the natural composite string and define $\boldsymbol{\lambda}^{(..k)} = \mathrm{shRSK}(\boldsymbol{w}^{(..k)})$. Let us also write $\boldsymbol{\lambda}^{(j)} = \mathrm{shRSK}(\boldsymbol{\lambda}^{(j)})$ for $j \in [2]$. We now claim that

$$\sum_{i=1}^{z} \boldsymbol{\lambda}_i^{(..k)} \le \sum_{i=1}^{z} \boldsymbol{\lambda}_i^{(1)} + \sum_{i=1}^{z} \boldsymbol{\lambda}_i^{(2)} \tag{11}$$

always holds. Indeed, this follows from Greene's Theorem: the left-hand side is $|\boldsymbol{s}|$, where $\boldsymbol{s} \in [d]^n$ is a maximum-length disjoint union of z increasing subsequences in $\boldsymbol{w}$; the projection of $\boldsymbol{s}^{(j)}$ onto coordinates $\boldsymbol{I}_j$ is a disjoint union of z increasing subsequences in $\boldsymbol{w}^{(j)}$ and hence the right-hand side is at least $|\boldsymbol{s}^{(1)}| + |\boldsymbol{s}^{(2)}| = |\boldsymbol{s}|$.

Applying (11) in the $z = k-1$ case, and using the facts that (i) $|\boldsymbol{\lambda}^{(..k)}| = |\boldsymbol{\lambda}^{(1)}| + |\boldsymbol{\lambda}^{(2)}|$, and (ii) $\lambda^{(1)}$ has height at most $k-1$, we see that $\boldsymbol{\lambda}_k^{(2)} \le \boldsymbol{\lambda}_k^{(..k)}$. Hence

$$\mathbf{E}[(\boldsymbol{\lambda}^{(..k)}_k - \alpha_k n)^2 \cdot \mathbf{1}[\overline{\mathcal{G}}]] \le \mathbf{E}(\boldsymbol{\lambda}^{(2)}_k - \alpha_k n)^2 = \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \alpha_k n)^2,$$

where $\boldsymbol{u} \sim \mathrm{Binomial}(\boldsymbol{m}, p_2)$ and $\boldsymbol{\mu} \sim \mathrm{SW}^{\boldsymbol{u}}(\frac{1}{k})$. Hence

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \alpha_k n)^2 &= \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \tfrac{1}{k}\boldsymbol{u} + \tfrac{1}{k}\boldsymbol{u} - \alpha_k n)^2 \\ &\le 2\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \tfrac{1}{k}\boldsymbol{u}) + 2\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}}(\tfrac{1}{k}\boldsymbol{u} - \alpha_k n)^2 \le 2\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \tfrac{1}{k}\boldsymbol{u})^2 + \frac{2n\alpha_k}{k},\end{aligned}$$

where the first inequality used $(x+y)^2 \le 2x^2 + 2y^2$. Given $\boldsymbol{u}$, define $\boldsymbol{\nu} = \mathbf{E}[\boldsymbol{\mu}_k \mid \boldsymbol{u}]$. Then

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \tfrac{1}{k}\boldsymbol{u})^2 &= \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}(\boldsymbol{\mu}_k - \boldsymbol{\nu} + \boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2 \\ &= \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}[(\boldsymbol{\mu}_k - \boldsymbol{\nu})^2 + (\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2 + 2(\boldsymbol{\mu}_k - \boldsymbol{\nu})(\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})] = \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u},\boldsymbol{\mu}}[(\boldsymbol{\mu}_k - \boldsymbol{\nu})^2 + (\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2] \\ &\le 16\mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}}\boldsymbol{u} + \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}}(\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2 = 16\alpha_k k n + \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}}(\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2,\end{aligned} \tag{12}$$

where the inequality uses Proposition 4.8. Next, we note that because $\boldsymbol{\mu}$ is distributed as $\mathrm{SW}^{\boldsymbol{u}}(\frac{1}{k})$, $\boldsymbol{\nu}$ is at most $\frac{1}{k}\boldsymbol{u}$. Hence, to upper-bound $(\boldsymbol{\nu} - \frac{1}{k}\boldsymbol{u})^2$ we must lower-bound $\boldsymbol{\nu}$, and this can be done by Theorem 4.4: $\boldsymbol{\nu} = \mathbf{E}[\boldsymbol{\mu}_k \mid \boldsymbol{u}] \ge \frac{1}{k}\boldsymbol{u} - 2\sqrt{\boldsymbol{u}}$. Thus, (12) can be bounded as

$$16\alpha_k k n + \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}}(\boldsymbol{\nu} - \tfrac{1}{k}\boldsymbol{u})^2 \le 16\alpha_k k n + \mathop{\mathbf{E}}_{\boldsymbol{m},\boldsymbol{u}} 4\boldsymbol{u} = 20\alpha_k k n.$$

In summary, the term in (9) corresponding to $\overline{\mathcal{G}}$ is at most $42\alpha_{\ge k} n$. □

Using the fact that $\alpha_k k, \alpha_{\ge k} \le \min\{1, \alpha_k d\}$, this implies Theorem 1.5.

4.7 An alternate bound on $E_k^{(n)}(\alpha)$

If the gap $\alpha_k - \alpha_{k+1}$ is very tiny (or zero), $\mathrm{Excess}_k(\alpha)$ will not be a good bound on $E_k^{(n)}(\alpha)$. In [OW16] we gave the following bound:

Proposition 4.13. *([OW16, Lemma 5.1].)* $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}] \leq \dfrac{2\sqrt{2}k}{\sqrt{n}}$.

By summing our Theorem 1.4 over all $i \in [k]$, we can replace the constant $2\sqrt{2}$ by 2. We now observe that this bound can also be improved so that it tends to 0 with $\alpha_{>k}$.

Proposition 4.14. $\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}] \leq O\left(\dfrac{k\sqrt{\alpha_{>k}}}{\sqrt{n}}\right)$.

Proof. By Theorem 1.14 we may assume that for some $m \geq 1$ we have $\alpha_k = \alpha_{k+1} = \alpha_{k+2} = \cdots = \alpha_{k+m-1}$ and $\alpha_{k+m+1} = \alpha_{k+m+2} = \cdots = \alpha_d = 0$.

Case 1: $m < k$. In this case, Theorem 4.5 tells us that $\mathbf{E}[\alpha_\ell - \underline{\boldsymbol{\lambda}}_\ell] \leq 2\sqrt{\alpha_\ell \ell/n}$. If $m = 1$ then we are done. Otherwise, $\alpha_k m \leq 2\alpha_{>k}$, and so

$$\mathbf{E}[\alpha_{>k} - \underline{\boldsymbol{\lambda}}_{>k}] \leq m \cdot 2\sqrt{2}\sqrt{\alpha_k k/n} \leq 2\sqrt{2}k\sqrt{\alpha_k m/n} \leq 4k\sqrt{\alpha_{>k}/n},$$

which is equivalent to our desired bound.

Case 2: $m \geq k$. In this case we follow the proof of Proposition 4.13 from [OW16]. Inspecting that proof, we see that in fact the following stronger statement is obtained:

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda}\sim\mathrm{SW}^n(\alpha)}[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}] \leq 2k\sqrt{p_2/n} + 2k\sqrt{p_3/n},$$

where it is easy to check (from [OW16, (25)]) that $p_2 + p_3 = k\alpha_k + \alpha_{>k}$. Now

$$\sqrt{p_2} + \sqrt{p_3} \leq 2\sqrt{p_2 + p_3} \leq 2\sqrt{m\alpha_k + \alpha_{>k}} \leq 2\sqrt{2\alpha_{>k}}$$

where the middle inequality is because we're in Case 2. Combining this with the previous inequality completes the proof. □

5 Tomography with Hellinger/infidelity error

5.1 Setup and notation

In this section we study the number of samples needed in quantum tomography to achieve small quantum Hellinger error (equivalently, infidelity). Throughout this section we consider the Keyl algorithm, in which

$$\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha), \quad \boldsymbol{V} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho), \quad \boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}/n), \quad \text{and the hypothesis is } \widehat{\boldsymbol{\rho}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\dagger,$$

Later, we will also consider "PCA"-style results where the output is required to be of rank k. In that case, $\boldsymbol{\Lambda}$ will be replaced by $\boldsymbol{\Lambda}^{(k)} = \mathrm{diag}(\boldsymbol{\lambda}_1/n, \dots, \boldsymbol{\lambda}_k/n, 0, \dots, 0)$.

Since the Hellinger distance is unitarily invariant, we have $D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho) = D_{\mathrm{H}^2}(\boldsymbol{\Lambda}, \boldsymbol{R})$, where $\boldsymbol{R} = \boldsymbol{V}^\dagger \rho \boldsymbol{V}$. It is also an immediate property of the Keyl distribution that the distribution of $\boldsymbol{R}$

depends only on the spectrum of ρ. Thus we may henceforth assume, without loss of generality, that

$$\rho = A = \operatorname{diag}(\alpha), \quad \text{so } \boldsymbol{V} \sim \mathrm{K}_{\boldsymbol{\lambda}}(A), \quad \boldsymbol{R} = \boldsymbol{V}^{\dagger} A \boldsymbol{V}, \tag{13}$$

and our goal is to bound

$$\mathop{\mathbf{E}}_{\boldsymbol{\lambda},\boldsymbol{V}}[D_{\mathrm{H}^2}(\boldsymbol{\Lambda}, \boldsymbol{R})]. \tag{14}$$

We introduce one more piece of notation. Every outcome $\boldsymbol{V} = V$ is a unitary matrix, for which the matrix $(|V_{ij}|^2)_{ij}$ is doubly-stochastic and hence a convex combination of permutation matrices. We think of V as inducing a random permutation $\boldsymbol{\pi}$ on $[d]$, which we write as

$$\boldsymbol{\pi} \sim V.$$

This arises in expressions like $\boldsymbol{R}_{ii} = (\boldsymbol{V}^{\dagger} A \boldsymbol{V})_{ii}$ and $(\sqrt{\boldsymbol{R}})_{ii} = (\boldsymbol{V}^{\dagger}\sqrt{A}\boldsymbol{V})_{ii}$, which, by explicit computation, are

$$\boldsymbol{R}_{ii} = \sum_{j=1}^{d} |\boldsymbol{V}_{ji}|^2 \alpha_j = \mathop{\mathbf{E}}_{\boldsymbol{\pi}\sim\boldsymbol{V}}[\alpha_{\boldsymbol{\pi}(i)}], \qquad (\sqrt{\boldsymbol{R}})_{ii} = \mathop{\mathbf{E}}_{\boldsymbol{\pi}\sim\boldsymbol{V}}[\sqrt{\alpha_{\boldsymbol{\pi}(i)}}]. \tag{15}$$

In addition to (13), we will henceforth always assume $\boldsymbol{\pi} \sim \boldsymbol{V}$.

5.2 Preliminary tools

We will require the following theorem from [OW16]. It is not explicitly stated therein, but it is derived within its "Proof of Theorem 1.5" (between the lines labeled "(30)" and "(by (8) again)").

Theorem 5.1. *Let ρ be a d-dimensional density matrix ρ with sorted spectrum α, let $j \in [d]$, let $\boldsymbol{\lambda} \sim \mathrm{SW}^n(\alpha)$ and let $\boldsymbol{V} \sim \mathrm{K}_{\boldsymbol{\lambda}}(\rho)$. Then for $\boldsymbol{R} = \boldsymbol{V}^{\dagger}\rho\boldsymbol{V}$, it holds that*

$$\mathbf{E}\left[\sum_{i=1}^{j} \boldsymbol{R}_{ii}\right] \geq 2\sum_{i=1}^{j} \alpha_i - \mathop{\mathbf{E}}_{\boldsymbol{\lambda}'\sim\mathrm{SW}^{n+1}(\alpha)}\left[\sum_{i=1}^{j} \frac{d-i+\boldsymbol{\lambda}'_i}{n+1}\right].$$

We will slightly simplify the bound above:

Corollary 5.2. *In the setting described in Section 5.1, and for any $j \in [d]$,*

$$\mathbf{E}\left[\sum_{i=1}^{j}(\alpha_i - \boldsymbol{R}_{ii})\right] = \mathbf{E}\left[\sum_{i=1}^{j}(\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right] \leq \mathbf{E}[\underline{\boldsymbol{\lambda}}_{\leq j} - \alpha_{\leq j}] + \frac{jd}{n}.$$

Proof. Here we simply used $\mathbf{E}[\boldsymbol{\lambda}'_i] \leq \mathbf{E}[\boldsymbol{\lambda}_i] + 1$, $d - i + 1 \leq d$, $\frac{1}{n+1} \leq \frac{1}{n}$, and then did some rearranging. □

When it comes to analyzing the quantum Hellinger error of the algorithm, we will end up needing to bound expressions like

$$\mathbf{E}\left[\sum_{i=1}^{d}\left(\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}}\right)^2\right] = \mathbf{E}[d_{\mathrm{H}^2}(\alpha \circ \boldsymbol{\pi}, \alpha)].$$

Ultimately, all we will use about the Keyl distribution on $\boldsymbol{V}$ (and hence the distributions of $\boldsymbol{R}$, $\boldsymbol{\pi}$) is that Corollary 5.2 holds. This motivates the following lemma:

Lemma 5.3. *Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d > 0$ and let $\beta_1, \ldots, \beta_d$ be a permutation of $\alpha_1, \ldots, \alpha_d$. Then*

$$\sum_{i=1}^{d} (\sqrt{\alpha_i} - \sqrt{\beta_i})^2 \leq 2 \sum_{j=1}^{d-1} \frac{\alpha_j - \alpha_{j+1}}{\alpha_j} \sum_{i=1}^{j} (\alpha_i - \beta_i).$$

Proof. Let us write LHS and RHS for the left-hand side and right-hand side above. Also, denoting $T_i = \sum_{j=i}^{d-1} 2\frac{\alpha_j - \alpha_{j+1}}{\alpha_j}$, we have RHS $= \sum_{i=1}^{d-1} (\alpha_i - \beta_i) T_i$.

If $\beta_1, \ldots, \beta_d$ is identical to $\alpha_1, \ldots, \alpha_d$ then LHS $=$ RHS $= 0$. Otherwise, suppose that q is the least index such that $\alpha_q \neq \beta_q$. Let $r, s > q$ be such that $\beta_q = \alpha_r$ and $\beta_s = \alpha_q$; then let β' denote the permutation β with its qth and sth entries swapped, so $\beta'_q = \alpha_q$, $\beta'_s = \alpha_r$. Writing LHS$'$ and RHS$'$ for the new values of LHS, RHS, we will show that

$$\text{LHS} - \text{LHS}' \leq \text{RHS} - \text{RHS}'.$$

Repeating this argument until β is transformed into α completes the proof. We have

$$\text{LHS} - \text{LHS}' = (\sqrt{\alpha_q} - \sqrt{\alpha_r})^2 + (\sqrt{\alpha_q} - \sqrt{\alpha_s})^2 - (\sqrt{\alpha_s} - \sqrt{\alpha_r})^2 = 2(\sqrt{\alpha_q} - \sqrt{\alpha_r})(\sqrt{\alpha_q} - \sqrt{\alpha_s}),$$
$$\text{RHS} - \text{RHS}' = (\alpha_q - \alpha_r)T_q + (\alpha_s - \alpha_q)T_s - (\alpha_s - \alpha_r)T_s = (\alpha_q - \alpha_r)(T_q - T_s).$$

Since $s > q$ we have

$$T_q - T_s = \sum_{j=q}^{s-1} 2\frac{\alpha_j - \alpha_{j+1}}{\alpha_j} \geq \frac{2}{\alpha_q} \sum_{j=q}^{s-1} (\alpha_j - \alpha_{j+1}) = \frac{2}{\alpha_q}(\alpha_q - \alpha_s).$$

Thus it remains to show

$$(\sqrt{\alpha_q} - \sqrt{\alpha_r})(\sqrt{\alpha_q} - \sqrt{\alpha_s}) \leq \frac{(\alpha_q - \alpha_r)(\alpha_q - \alpha_s)}{\alpha_q} = (\sqrt{\alpha_q} - \alpha_r/\sqrt{\alpha_q})(\sqrt{\alpha_q} - \alpha_s/\sqrt{\alpha_q}).$$

This indeed holds, because $\alpha_q \geq \alpha_r \implies \sqrt{\alpha_r} \geq \alpha_r/\sqrt{\alpha_q}$, and similarly for s. □

It will be useful to have a bound on the sum of the “multipliers” appearing in Lemma 5.3.

Lemma 5.4. *Suppose $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d > 0$. Let $L = \min\{d, \ln(\alpha_1/\alpha_d)\}$. Then*

$$\sum_{i=1}^{d-1} \frac{\alpha_i - \alpha_{i+1}}{\alpha_i} \leq L.$$

Proof. The bound of d is obvious. Otherwise, the bound involving $\ln(\alpha_1/\alpha_d)$ is equivalent to

$$\frac{\alpha_1}{\alpha_d} \geq \exp\left(\sum_{i=1}^{d-1} \frac{\alpha_i - \alpha_{i+1}}{\alpha_i}\right) = \prod_{i=1}^{d-1} \exp\left(1 - \frac{\alpha_{i+1}}{\alpha_i}\right).$$

But this follows from $\exp(1 - z) \leq 1/z$ for $z \in (0, 1]$, and telescoping. □

Finally, we also have a variant of Lemma 5.3 that can help if some of the α_i's are very small:

Corollary 5.5. *Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k > \zeta \geq \alpha_{k+1} \geq \cdots \geq \alpha_d$, where $k < d$, and let $\beta_1, \dots, \beta_d$ be a permutation of $\alpha_1, \dots, \alpha_d$. Let α'_i be the same as α_i for $i \leq k$, but let $\alpha'_{k+1} = \zeta$.*

$$\sum_{i=1}^{d} (\sqrt{\alpha_i} - \sqrt{\beta_i})^2 \leq 4 \sum_{j=1}^{k} \frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j} \sum_{i=1}^{j} (\alpha_i - \beta_i) + d\zeta + 8kL\zeta,$$

where $L = \min\{k, \ln(\alpha_1/\zeta)\}$.

Proof. Extend the notation α' by defining $\alpha'_i = \max\{\alpha_i, \zeta\}$, and similarly define β'_i. Applying Lemma 5.3 we get

$$\sum_{i=1}^{d} \left(\sqrt{\alpha'_i} - \sqrt{\beta'_i}\right)^2 \leq 2 \sum_{j=1}^{k} \frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j} \sum_{i=1}^{j} (\alpha'_i - \beta'_i).$$

On the left we can use

$$\left(\sqrt{\alpha_i} - \sqrt{\beta_i}\right)^2 \leq 2\left(\sqrt{\alpha'_i} - \sqrt{\beta'_i}\right)^2 + \zeta,$$

which is easy to verify by case analysis. On the right we can use

$$\left| \sum_{i=1}^{j} (\alpha'_i - \beta'_i) - \sum_{i=1}^{j} (\alpha_i - \beta_i) \right| \leq 2k\zeta$$

and then the bound from Lemma 5.4 applied to the sequence $(\alpha'_1, \dots, \alpha'_{k+1})$. □

5.3 Tomography analysis

We begin with with a partial analysis of the general PCA algorithm.

Theorem 5.6. *Let ρ be a d-dimensional density matrix ρ with sorted spectrum α, and let $k \in [d]$. Suppose we perform the Keyl algorithm and produce the rank-k (or less) hypothesis $\widehat{\boldsymbol{\rho}} = \boldsymbol{V}\boldsymbol{\Lambda}^{(k)}\boldsymbol{V}^\dagger$, as described in Section 5.1. Then*

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho)] \leq \alpha_{>k} + 2\,\mathbf{E}\left[d^{(k)}_{\mathrm{H}^2}(\alpha \circ \boldsymbol{\pi}, \alpha)\right] + 2\,\mathbf{E}\left[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}\right] + O\left(\frac{kd}{n}\right).$$

Proof. As described in Section 5.1 — in particular, at (14) — we need to bound $\mathbf{E}[D_{\mathrm{H}^2}(\boldsymbol{\Lambda}^{(k)}, \boldsymbol{R})]$. We have

$$\begin{aligned}
\mathbf{E}\left[D_{\mathrm{H}^2}(\boldsymbol{\Lambda}^{(k)}, \boldsymbol{R})\right] &= \mathbf{E}\left[\mathrm{tr}(\boldsymbol{\Lambda}^{(k)}) + \mathrm{tr}(\boldsymbol{R}) - 2\mathrm{tr}\left(\sqrt{\boldsymbol{\Lambda}^{(k)}}\sqrt{\boldsymbol{R}}\right)\right] \\
&= \mathbf{E}\left[\underline{\boldsymbol{\lambda}}_{>k} + 2\underline{\boldsymbol{\lambda}}_{\leq k} - 2\sum_{i=1}^{k} \sqrt{\underline{\boldsymbol{\lambda}}_i} \cdot (\sqrt{\boldsymbol{R}})_{ii}\right] \\
&= \mathbf{E}[\underline{\boldsymbol{\lambda}}_{>k}] + \mathbf{E}\left[2\underline{\boldsymbol{\lambda}}_{\leq k} - 2\sum_{i=1}^{k} \sqrt{\underline{\boldsymbol{\lambda}}_i}\sqrt{\alpha_{\boldsymbol{\pi}(i)}}\right] && \text{(by (15))} \\
&= \mathbf{E}[\underline{\boldsymbol{\lambda}}_{>k}] + \mathbf{E}\left[\sum_{i=1}^{k} \left(\sqrt{\underline{\boldsymbol{\lambda}}_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}}\right)^2\right] + \mathbf{E}\left[\sum_{i=1}^{k} (\underline{\boldsymbol{\lambda}}_i - \alpha_{\boldsymbol{\pi}(i)})\right]. && (16)
\end{aligned}$$

We bound the three expressions in (16) as follows:

$$\mathbf{E}[\underline{\boldsymbol{\lambda}}_{>k}] \le \alpha_{>k} \qquad \text{(since } \underline{\boldsymbol{\lambda}} \succ \alpha\text{)}$$

$$\begin{aligned}
\mathbf{E}\left[\sum_{i=1}^{k}\left(\sqrt{\underline{\boldsymbol{\lambda}}_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}}\right)^2\right] &\le 2\,\mathbf{E}\left[\sum_{i=1}^{k}\left(\sqrt{\underline{\boldsymbol{\lambda}}_i} - \sqrt{\alpha_i}\right)^2\right] + 2\,\mathbf{E}\left[\sum_{i=1}^{k}\left(\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}}\right)^2\right] \\
&= 2\,\mathbf{E}\left[d_{\mathrm{H}^2}^{(k)}(\underline{\boldsymbol{\lambda}}, \alpha)\right] + 2\,\mathbf{E}\left[d_{\mathrm{H}^2}^{(k)}(\alpha \circ \boldsymbol{\pi}, \alpha)\right] \\
&\le O\left(\frac{kd}{n}\right) + 2\,\mathbf{E}\left[d_{\mathrm{H}^2}^{(k)}(\alpha \circ \boldsymbol{\pi}, \alpha)\right] \qquad \text{(by Theorem 4.11)}
\end{aligned}$$

$$\mathbf{E}\left[\sum_{i=1}^{k}(\underline{\boldsymbol{\lambda}}_i - \alpha_{\boldsymbol{\pi}(i)})\right] = \mathbf{E}\big[\underline{\boldsymbol{\lambda}}_{\le k} - \alpha_{\le k}\big] + \mathbf{E}\left[\sum_{i=1}^{k}(\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right] \le 2\,\mathbf{E}\big[\underline{\boldsymbol{\lambda}}_{\le k} - \alpha_{\le k}\big] + \frac{kd}{n}$$
$$\text{(by Corollary 5.2)}$$

Combining these bounds completes the proof. □

We can now give our analysis for full tomography:

Theorem 5.7. *Suppose ρ has rank r. Then the hypothesis of the Keyl algorithm satisfies*

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}}, \rho)] \le O\left(\frac{rd}{n}\right) \cdot \min\{r, \ln n\}.$$

Proof. When ρ has rank r we know that $\boldsymbol{\lambda}$ will always have at most r nonzero rows; thus $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{(r)}$, and we may use the bound in Theorem 5.6 with $k = r$. In this case, the terms $\alpha_{>r}$ and $2\,\mathbf{E}\big[\underline{\boldsymbol{\lambda}}_{\le r} - \alpha_{\le r}\big]$ vanish, the $O(\frac{rd}{n})$ error is accounted for in the theorem statement, and we will use the simple bound

$$2\,\mathbf{E}\left[d_{\mathrm{H}^2}^{(r)}(\alpha \circ \boldsymbol{\pi}, \alpha)\right] \le 2\,\mathbf{E}\left[\sum_{i=1}^{d}(\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}})^2\right].$$

We now apply Corollary 5.5 with $\zeta = \frac{r}{n}$; note that the "k" in that corollary satisfies $k \le r = \operatorname{rank}\rho$. We obtain

$$2\,\mathbf{E}\left[\sum_{i=1}^{d}(\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}})^2\right] \le 8\sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\,\mathbf{E}\left[\sum_{i=1}^{j}(\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right] + 2d\zeta + 16kL\zeta,$$

where $L \le \min\{k, \ln n\}$. The latter two terms here are accounted for in the theorem statement, so it suffices to bound the first. By Corollary 5.2 we have

$$\sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\,\mathbf{E}\left[\sum_{i=1}^{j}(\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right] \le \sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\,\mathbf{E}[\underline{\boldsymbol{\lambda}}_{\le j} - \alpha_{\le j}] + \sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\frac{jd}{n}.$$

The latter quantity here is at most $\frac{kd}{n}L$ (by Lemma 5.4), which again is accounted for in the theorem statement. As for the former quantity, we use Theorem 3.7 to obtain the bound

$$\sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\,\mathbf{E}[\underline{\boldsymbol{\lambda}}_{\le j} - \alpha_{\le j}] \le \frac{1}{n}\sum_{j=1}^{k}\frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\mathrm{Excess}_j(\alpha). \tag{17}$$

(The quantity $\mathrm{Excess}_j(\alpha)$ may be ∞, but only if $\alpha_j = \alpha_{j+1}$ and hence $\frac{\alpha'_j-\alpha'_{j+1}}{\alpha'_j} = 0$. The reader may check that it is sufficient for us to proceed with the convention $0 \cdot \infty = 0$.) As $\frac{\alpha'_j-\alpha'_{j+1}}{\alpha'_j} = 1 - \frac{\alpha'_{j+1}}{\alpha'_j} \leq 1 - \frac{\alpha_{j+1}}{\alpha_j}$, we can replace α' with α in (17). Then substituting in the definition of $\mathrm{Excess}_k(\alpha)$ yields an upper bound of

$$\frac{1}{n}\sum_{j=1}^{k}\frac{\alpha_j-\alpha_{j+1}}{\alpha_j}\sum_{i\leq j<\ell}\frac{\alpha_\ell}{\alpha_i-\alpha_\ell} = \frac{1}{n}\sum_{\substack{i\leq k\\ \ell>i}}\frac{\alpha_\ell}{\alpha_i-\alpha_\ell}\sum_{j=i}^{\min\{k,\ell-1\}}\frac{\alpha_j-\alpha_{j+1}}{\alpha_j}$$
$$\leq \frac{1}{n}\sum_{\substack{i\leq k\\ \ell>i}}\frac{\alpha_\ell}{\alpha_i-\alpha_\ell}\sum_{j=i}^{\ell-1}\frac{\alpha_j-\alpha_{j+1}}{\alpha_\ell} = \frac{1}{n}\sum_{\substack{i\leq k\\ \ell>i}}1 \leq \frac{kd}{n},$$

which suffices to complete the proof of the theorem. □

Finally, we give our "PCA"-style bound. Theorem 1.19 follows by applying Proposition 4.14.

Theorem 5.8. *For any ρ and $k \in [d]$, if we apply the Keyl algorithm but output the rank-k (or less) hypothesis $\widehat{\boldsymbol{\rho}} = \boldsymbol{V}\boldsymbol{\Lambda}^{(k)}\boldsymbol{V}^\dagger$, we have*

$$\mathbf{E}[D_{\mathrm{H}^2}(\widehat{\boldsymbol{\rho}},\rho)] \leq \alpha_{>k} + O\left(\frac{kdL}{n}\right) + O(L)\cdot\mathbf{E}[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}],$$

where $L = \min\{k, \ln n\}$.

Proof. The proof proceeds similarly to that of Theorem 5.7 but with k in place of r. Now the terms $\alpha_{>k}$ and $\mathbf{E}[\underline{\boldsymbol{\lambda}}_{\leq k} - \alpha_{\leq k}]$ do not vanish but instead go directly into the error bound. It remains to bound

$$2\,\mathbf{E}\left[\sum_{i=1}^{k}(\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}})^2\right].$$

Fix an outcome for $\boldsymbol{\pi} = \pi$ and write the associated permutation $\alpha \circ \pi$ as β. Unfortunately we cannot apply Lemma 5.3 because the subsequence $\beta_1, \ldots, \beta_k$ is not necessarily a permutation of the subsequence $\alpha_1, \ldots, \alpha_k$. What we can do instead is the following. Suppose that some k' of the numbers $\beta_1, \ldots, \beta_k$ do not appear within $\alpha_1, \ldots, \alpha_k$. Place these missing numbers, in decreasing order, at the end of the α-subsequence, forming the new decreasing subsequence $\overline{\alpha}_1, \ldots, \overline{\alpha}_K$, where $K = k + k' \leq 2k$ and $\overline{\alpha}_i = \alpha_i$ for $i \leq k$. Similarly extend the β-subsequence to $\overline{\beta}_1, \ldots, \overline{\beta}_K$ by adding in the "missing" α_i's; the newly added elements can be placed in any order. Note that all of the elements added to the α-subsequence are less than all the elements added to the β-subsequence (because α_k must be between them). Thus we have

$$\sum_{i=1}^{j}(\overline{\alpha}_i - \overline{\beta}_i) \leq \sum_{i=1}^{k}(\alpha_i - \beta_i) \quad \forall\, j > k. \tag{18}$$

We may now apply Corollary 5.5 to $\overline{\alpha}$ and $\overline{\beta}$, with its "d" set to K and with $\zeta = \frac{k}{n}$. We get

$$\sum_{i=1}^{K}\left(\sqrt{\overline{\alpha}_i} - \sqrt{\overline{\beta}_i}\right)^2 \leq 4\sum_{j=1}^{K}\frac{\overline{\alpha}'_j - \overline{\alpha}'_{j+1}}{\overline{\alpha}'_j}\sum_{i=1}^{j}(\overline{\alpha}_i - \overline{\beta}_i) + O\left(\frac{k^2L}{n}\right),$$

where $L = \min\{k, \ln n\}$. We can split the sum over $1 \le j \le K$ into $1 \le j \le k$ and $k < j \le K$. The latter sum can be bounded by $4L \sum_{i=1}^{k} (\alpha_i - \beta_i)$, using Lemma 5.4 and (18). The former sum is what we "would have gotten" had we been able to directly apply Corollary 5.5. That is, we have established

$$\sum_{i=1}^{k} \left(\sqrt{\alpha_i} - \sqrt{\beta_i}\right)^2 \le \sum_{i=1}^{K} \left(\sqrt{\overline{\alpha}_i} - \sqrt{\overline{\beta}_i}\right)^2 \le 4 \sum_{j=1}^{k} \frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j} \sum_{i=1}^{j} (\alpha_i - \beta_j) + O\left(\frac{k^2 L}{n}\right) + O(L) \cdot \sum_{i=1}^{k} (\alpha_i - \beta_i).$$

Now taking this in expectation over $\boldsymbol{\pi}$ yields

$$2\,\mathbf{E}\left[\sum_{i=1}^{k} (\sqrt{\alpha_i} - \sqrt{\alpha_{\boldsymbol{\pi}(i)}})^2\right] \le 8 \sum_{j=1}^{k} \frac{\alpha'_j - \alpha'_{j+1}}{\alpha'_j}\, \mathbf{E}\left[\sum_{i=1}^{j} (\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right] + O\left(\frac{k^2 L}{n}\right) + O(L) \cdot \mathbf{E}\left[\sum_{i=1}^{k} (\alpha_i - \alpha_{\boldsymbol{\pi}(i)})\right].$$

The first term above is handled exactly as in the proof of Theorem 5.7, and Corollary 5.2 takes care of the last term. □

6 The lower-row majorization theorem

The following theorem refers (in (19)) to some terminology "curves". We will actually not define this terminology until inside the proof of theorem; the reader can nevertheless follow the logical flow without knowing the definition.

Theorem 6.1. *Let $b \in \mathcal{A}^m$ be a string of distinct letters and let $A \subseteq \mathcal{A}$ be a set of letters in b deemed "admissible". Let $I_1, \dots, I_c$ be disjoint increasing subsequences (possibly empty) in b of total length L; we assume also that these subsequences are "admissible", meaning they consist only of admissible letters. Finally, assume the following condition:*

$$\text{"one can draw a set of curves through the } I\text{'s such that} \qquad\qquad \\ \text{all inadmissible letters in } b \text{ are southeast of the first curve".} \qquad (19)$$

(As mentioned, the terminology used here will be defined later.)

Let $w \in \mathcal{A}^{m'}$ be a string of distinct letters with the following property: When the RSK algorithm is applied to w, the letters that get bumped into the second row form the string b (in the order that they are bumped).

Then there exists a new set of "admissible" letters $A' \supseteq A$ for w, with $|A'| = |A| + \Delta$ (so $\Delta \in \mathbb{N}$), along with disjoint admissible (with respect to A') increasing subsequences $J_1, \dots, J_c$ in w of total length $L + \Delta$, such that (19) holds for w and the J's with respect A'.

We will prove this theorem, as well as the following lemma, later.

Lemma 6.2. *Let $b \in \mathcal{A}^m$ be a string of distinct letters and let $I_1, \dots, I_c$ be disjoint increasing subsequences in b. Then there are disjoint increasing subsequences $I'_1, \dots, I'_c$ consisting of the same letters as $I_1, \dots, I_c$, just grouped differently, such that it is possible to draw a "set of curves" through $I'_1, \dots, I'_c$.*

Let us now see what Theorem 6.1 and Lemma 6.2 imply:

Theorem 6.3. *Fix an integer $k \ge 1$. Consider the RSK algorithm applied to some string $x \in \mathcal{A}^n$. During the course of the algorithm, some letters of x get bumped from the kth row and inserted into the $(k+1)$th row. Let $x^{(k)}$ denote the string formed by those letters* in the order they are so bumped. *On the other hand, let $\overline{x}$ be the subsequence of x formed by the letters of $x^{(k)}$* in the order they appear in x. *Then* $\mathrm{shRSK}(\overline{x}) \trianglerighteq \mathrm{shRSK}(x^{(k)})$.

Proof. We may assume all the letters in x are distinct, by the usual trick of "standardization"; this does not affect the operation of RSK on x or $\overline{x}$. When RSK is applied to x, let us write more generally $x^{(j)}$ $(1 \leq j \leq k)$ for the sequence of letters bumped from the jth row and inserted into the $(j+1)$th row, in the order they are bumped. We also write $x^{(0)} = x$.

We will show $\mathrm{shRSK}(\overline{x}) \trianglerighteq \mathrm{shRSK}(x^{(k)})$ using Greene's Theorem; it suffices to show that if $I_1^{(k)}, \dots, I_c^{(k)}$ are any disjoint increasing subsequences in $x^{(k)}$ of total length L, there are some c disjoint increasing subsequences $\overline{I}_1, \dots, \overline{I}_c$ of total length at least L in $\overline{x}$. We will find these subsequences by applying Theorem 6.1 k times in succession, with (b, w) equal to $(x^{(j)}, x^{(j-1)})$ for $j = k, k-1, k-2, \dots, 1$.

In the first application, with $b = x^{(k)}$ and $w = x^{(k-1)}$, we will declare *all* letters appearing in b to be "admissible". In particular this means that $I_1^{(k)}, \dots, I_c^{(k)}$ are automatically admissible. After reorganizing these subsequences using Lemma 6.2 (if necessary), we may draw *some* "set of curves" through them. Condition (19) is then vacuously true, as there are no inadmissible letters. Theorem 6.1 thus gives us some Δ_k newly admissible letters, as well as admissible disjoint increasing subsequences $I_1^{(k-1)}, \dots, I_c^{(k-1)}$ in $x^{(k-1)}$ of total length $L + \Delta_k$, such that condition (19) still holds.

We now continue applying Theorem 6.1, $k-1$ more times, until we end up with admissible disjoint increasing subsequences $I_1^{(0)}, \dots, I_c^{(0)}$ in $x^{(0)} = x$ of total length $L + \overline{\Delta}$, where $\overline{\Delta} = \Delta_1 + \cdots + \Delta_k$ is the number of newly admissible letters in x, beyond those letters originally appearing in $x^{(k)}$. Finally, we delete all of these newly admissible letters wherever they appear in $I_1^{(0)}, \dots, I_c^{(0)}$, forming $\overline{I}_1, \dots, \overline{I}_c$; these are then disjoint increasing subsequences of total length at least L. But they also consist only of letters that were originally admissible, i.e. in $x^{(k)}$; hence $\overline{I}_1, \dots, \overline{I}_c$ are subsequences of $\overline{x}$ and we are done. □

We now come to the proof of Theorem 6.1 (which includes the definition of "curves").

Proof of Theorem 6.1. Our proof of the theorem uses Viennot's geometric interpretation of the RSK process [Vie81]; see e.g., [Sag01, Chapter 3.6], [Wer94, Chapter 2]) for descriptions. We may assume that $\mathcal{A} = [D]$ for some $D \in \mathbb{N}$. The word $w = (w_1, \dots, w_n)$ is then identified with its "graph"; i.e., the set of points (i, w_i) in the 2-dimensional plane. We will call these the "white points". (Since w has distinct letters, no two white points are at the same height.) Viennot's construction then produces a set of "shadow lines" through the points; we will call them "jump lines", following the terminology in Wernisch [Wer94]. The points at the northeast corners are called the "skeleton" of w; we will also call these the "black points". They are the graph of the string b (if we use w's indices when indexing b).

Note that an increasing subsequence in b (respectively, w) corresponds to an increasing sequence of black (respectively, white) points; i.e., a sequence in which each successive point is to the northeast of the previous one. We will call such a sequence a *chain*. In Theorem 6.1 we are initially given c disjoint sequences/chains $I_1, \dots, I_c$ in b, of total length L. To aid in the geometrical description, we will imagine that L "beads" are placed on all of these chain points.[3]

We are also initially given a set A of admissible "letters" in b; in the geometric picture, these correspond to "admissible black points". We are initially promised that all the beads are at admissible black points. The end of the theorem statement discusses admissible letters A' in w; these will correspond to "admissible white points". Since $A' \supseteq A$, every white point that is directly west of an admissible black point will be an admissible white point. But in addition, the theorem allows us to designate some Δ additional white points as "newly admissible".

[3] As a technical point, the theorem allows some of these chains to be empty. We will henceforth discard all such "empty" chains, possibly decreasing c. In the end we will also allow ourselves to produce fewer than c chains J_i in w. But this is not a problem, as we can always artificially introduce more empty chains.

Our final goal involves finding c chains of admissible white points, of total length $L + \Delta$. The outline for how we will do this is as follows: First we will add some number Δ of beads to the initial chains, and at the same time designate some Δ white points as newly admissible. Then we will "slide" each bead some amount west and north along its jump line, according to certain rules. At the end of the slidings, all the beads will be on admissible white points, and we will show that they can be organized into c chains. Finally, we must show that the postcondition (19) concerning "curves" is still satisfied, given that it was satisfied initially.

Let's now explain what is meant by "curves" and "sets of curves". A curve will refer to an infinite southwest-to-northeast curve which is the graph of a strictly increasing continuous function that diverges to $\pm\infty$ in the limits to $\pm\infty$. It will typically pass through the beads of a chain. When condition (19) speaks of a "set of curves" passing through chains $I_1, \dots, I_c$, it is meant that for each chain we have have a single curve passing through the beads of that chain, and that the curves do not intersect. With this definition in place, the reader may now like to see the proof of Lemma 6.2 at the end of this section.

Given such a set of curves, we can and will always perturb them slightly so that the only black or white diagram points that the curves pass through are points with beads. As the curves do not intersect, we may order them as $Q_1, Q_2, \dots, Q_c$ from the southeasternmost to the northwesternmost; we can assume that the chains are renumbered correspondingly. The curves thereby divide the plane into *regions*, between the successive curves.

Before entering properly into the main part of the proof of Theorem 6.1 we need one more bit of terminology. Whenever a curve intersects a jump line, we call the intersection point a "crossing". We will further categorize each crossing as either "horizontal" or "vertical", according to whether the point is on a horizontal or vertical jump line segment. (Cf. the "chain curve" crossings in [Wer94, Section 3.1].) In case the crossing is at a jump line corner (as happens when the curves passes through a bead), we classify the crossing as *vertical*.

We now come to the main part of the proof: In the geometric diagram we have chains of beads $I_1, \dots, I_c$ of total length L, as well as a set of curves $Q_1, Q_2, \dots, Q_{c'}$ passing through them, with all inadmissible black points to the southeast of Q_1. Our proof will be algorithmic, with four *phases*.

- **Phase 1:** In which new beads are added, and new white points are deemed admissible.
- **Phase 2:** In which beads are partly slid, to "promote" them to new chains.
- **Phase 3:** In which the new chains, $J_1, \dots, J_c$, are further slid to white points.
- **Phase 4:** In which a new set of curves is drawn through $J_1, \dots, J_c$, to satisfy (19).

We now describe each of the phases in turn.

Phase 1. In this phase we consider the horizontal crossings, if any, of the first (i.e., southeasternmost) curve Q_1. For each horizontal crossing, we consider the white point immediately westward. If that point is currently inadmissible, we declare it to be admissible, and we add a new bead at the crossing point. Certainly this procedure introduces as many beads Δ as it does newly admissible white points. Also, although the new beads are not (yet) at points in the Viennot diagram, we *can* add them to the first chain I_1, in the sense that they fit into the increasing sequence of beads already in I_1. (This is because the curve Q_1 is increasing from southwest to northeast.) We now want to make the following claim:

Key Claim: All white points to the northwest of Q_1 are now admissible.

(Recall that no white point is exactly on Q_1.) To see the claim, consider any white point p to the northwest of Q_1. Consider the jump line segment extending east from p. If that segment crosses Q_1 then it's a horizontal crossing, and Phase 1 makes p admissible. Otherwise, the segment must terminate at a black point q that is on or northwest of Q_1. (The segment cannot be a half-infinite extending eastward to infinity, because of the condition that Q_1 eventually extends infinitely northward.) By the precondition (19), q must be an admissible black point. Thus p is an admissible white point, being at the same height as q.

Phase 2. For this phase it will be notationally convenient to temporarily add a "sentinel curve" Q_{c+1} that is far to the northwest of the last curve Q_c; the only purpose of this curve is to create vertical crossing points on all of the northward half-infinite segments of the jump lines. We now proceed through each bead x in the diagram and potentially "promote" it to a higher-numbered chain. The algorithm is as follows: We imagine traveling west and north from x along its jump line until the first time that a vertical crossing point p is encountered. (Note that such a vertical crossing point must always exist because of the sentinel curve Q_{c+1}.) Let q denote the crossing point on this jump line immediately *preceding* p. (This q will either be the current location of x, or it will be northwest of x.) We now slide bead x to point q. If x indeed moves (i.e., q is not already its current location), we say that x has been "promoted" to a higher curve/chain. As mentioned, we perform this operation for every bead x.

A crucial aspect of this phase is that the beads on a single jump line never "pass" each other, and in particular we never try to place two beads on the same crossing point. This is because whenever we have two consecutive beads on a jump line, there is always a vertical crossing at the higher bead. (Of course, prior to Phase 1 all beads were at vertical crossings, by definition. After Phase 1 we may have some beads at horizontal crossings, but at most one per jump line, and only on the lowest curve Q_1.)

At the end of Phase 2, all beads end up at (distinct) crossing points; in particular, they are all on the curves $Q_1, \dots, Q_c$. (A bead cannot move onto the sentinel curve Q_{c+1}.) Thus the beads may naturally partitioned into at most c chains (increasing sequences), call them $J_1, \dots, J_c$, some of which may be empty. We now have:

Phase 2 Postcondition: For every bead, the first crossing point on its jump line to its northwest is a vertical crossing.

Phase 3. The goal of Phase 3 is to further slide the (nonempty) chains $J_1, \dots, J_c$ northwestward along their jump lines so that they end up at white points, forming white chains. Since we will continue to move beads only northwestward, all the final white resting points will be admissible, by the *Key Claim* above. Another property of Phase 3 will be that each (nonempty) chain J_i will stay strictly inside the region between curves Q_i and Q_{i+1}. Because of this, we will again have the property that beads will never slide past each other or end at the same white point, and the order in which we process the chains does not matter.

So let us fix some (nonempty) chain J_i on curve Q_i and see how its beads can be slid northwestward along their jump lines to white points forming a chain southeast of Q_{i+1}. We begin with the northeasternmost bead on the chain, which (by virtue of Phase 2) is either on a black point or is in the middle of a horizontal jump line segment. In either case, we begin by sliding it west, and we deposit immediately at the first white point encountered. We must check that this white point is still to the southeast of curve Q_{i+1}. This is true because otherwise the first crossing point northwest of the bead's original position would be a horizontal one, in contradiction to the *Phase 2*

Postcondition.

We handle the remaining beads in J_i inductively. Suppose we have successfully deposited the northeasternmost t beads of J_i at white points that form a chain southeast of Q_{i+1}. Let's say the last of these (the southwesternmost of them) is bead x_t, and we now want to successfully slide the next one, x_{t+1}. Let p_t denote the white point into which x_t was slid. Our method will be to slide x_{t+1} west and north along its jump line until the first time it encounters a white point, p_{t+1} that is west of p_t, depositing it there. We need to argue three things: (i) p_{t+1} exists; (ii) p_{t+1} is southeast of the curve Q_{i+1}; (iii) p_{t+1} is south of p_t. If these things are true then we will have deposited x_{t+1} in such a way that it extends the white point chain and is still southeast of Q_{i+1}. This will complete the induction.

To check the properties of p_{t+1}, consider also the last (northwesternmost) white point p'_{t+1} on x_{t+1}'s jump line that is still southeast of Q_{i+1}. At least one must exist because the jump line crosses Q_{i+1} vertically, by the *Phase 2 Postcondition.* This p'_{t+1} must be west of p_t, as otherwise the jump line segment extending north from from it would cross p_t's jump line. It follows that p_{t+1} must exist and be southeast of Q_{i+1}. It remains to check that p_{t+1} is indeed south of p_t. But if this were not true then bead x_{t+1} would have slid north of p_t prior to sliding west of it — impossible, as again it would imply x_{t+1}'s jump line crossing x_t's. This completes the induction, and the analysis of Phase 3.

Phase 4. Here we need to show that we can draw a new set of curves through the final (nonempty) chains $J_1, \dots, J_c$ such that condition (19) is satisfied. This is rather straightforward. As shown in Phase 3, each final chain J_i is confined to a region between the old curves Q_i and Q_{i+1}. Thus there is no difficulty in drawing new nonintersecting curves through these chains, also confined within the regions, that pass through all the beads. Finally, as the "new first curve" is completely to the northwest of the "old first curve" Q_1, the fact that all inadmissible white points are southeast of it follows from the *Key Claim.* □

Proof of Lemma 6.2. Suppose we are given initial chains of beads $I_1, \dots, I_c$. As the beads on each chain are increasing, southwest-to-northeast, there is no difficulty in drawing a curve through each chain. The only catch, for the purposes of getting a "set of curves", is that the curves might intersect. (We can assume by perturbation, though, that two curves never intersect at a bead.) To correct the intersections, consider any two curves Q_i and Q_j that intersect. Redefine these curves as follows: take all the "lower" (southeastern) segments and call that one new curve, and take all the "upper" (northwestern) segments and call that another new curve. All of the beads are still on curves, and thus can still be repartitioned into c chains. The resulting new curves still technically have points of contact but do not essentially cross each other; by a perturbation we can slightly pull apart the points of contact to make sure they become truly nonintersecting.

(An alternative to this proof is the following: For our overall proof of Theorem 6.3 we may as well assume that the initial set of c disjoint sequences in b is of maximum possible total length. In this case, Wernisch's maximum c-chain algorithm [Wer94, Theorem 21] provides a set of increasing, nonintersecting "chain curves", together with a maximizing set of c chains that are in the associated "regions". This makes it easy to draw a nonintersecting set of curves through the chains, as in Phase 4 above.) □

References

[ANSV08] Koenraad Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical*

Physics, 279(1):251–283, 2008. 2.1

[ARS88] Robert Alicki, Sławomir Rudnicki, and Sławomir Sadowski. Symmetry properties of product states for the system of N n-level atoms. *Journal of mathematical physics*, 29(5):1158–1162, 1988. 1.3, 1.3, 1.5, 1.5, 1.21

[BAH$^+$16] Michael Beverland, Gorjan Alagic, Jeongwan Haah, Gretchen Campbell, Ana Maria Rey, and Alexey Gorshhkov. Implementing a quantum algorithm for spectrum estimation with alkaline earth atoms. In *19th Conference on Quantum Information Processing*, 2016. QIP 2016. 1.3

[BDJ99] Jinho Baik, Percy Deift, and Kurt Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *Journal of the American Mathematical Society*, 12(4):1119–1178, 1999. 1, 1.5

[Bia01] Philippe Biane. Approximate factorization and concentration for characters of symmetric groups. *International Mathematics Research Notices*, 2001(4):179–192, 2001. 1.5

[BL12] Nayantara Bhatnagar and Nathan Linial. On the Lipschitz constant of the RSK correspondence. *Journal of Combinatorial Theory, Series A*, 119(1):63–82, 2012. 2

[BMW16] Mohammad Bavarian, Saeed Mehraban, and John Wright. Personal communication, 2016. 1.6

[BOO00] Alexei Borodin, Andrei Okounkov, and Grigori Olshanski. Asymptotics of plancherel measures for symmetric groups. *Journal of the American Mathematical Society*, 13(3):481–515, 2000. 1

[Buf12] Alexey Bufetov. A central limit theorem for extremal characters of the infinite symmetric group. *Functional Analysis and Its Applications*, 46(2):83–93, 2012. 1.5

[CM06] Matthias Christandl and Graeme Mitchison. The spectra of quantum states and the Kronecker coefficients of the symmetric group. *Communications in mathematical physics*, 261(3):789–797, 2006. 1.1, 1.3

[FMN13] Valentin Féray, Pierre-Loïc Méliot, and Ashkan Nikeghbali. Mod-ϕ convergence I: Normality zones and precise deviations. Technical report, arXiv:1304.2934, 2013. 1.5

[Ful97] William Fulton. *Young tableaux: with applications to representation theory and geometry*. Cambridge University Press, 1997. 1

[Gre74] Curtis Greene. An extension of Schensted's theorem. *Advances in Mathematics*, 14:254–265, 1974. 1.2

[HHJ$^+$16] Jeongwan Haah, Aram Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, August 2016. Preprint. 1.3, 1.3, 1.4

[HM02] Masahito Hayashi and Keiji Matsumoto. Quantum universal variable-length source coding. *Physical Review A*, 66(2):022311, 2002. 1.1, 1.3

[HMR$^+$10] Sean Hallgren, Cristopher Moore, Martin Rötteler, Alexander Russell, and Pranab Sen. Limitations of quantum coset states for graph isomorphism. *Journal of the ACM (JACM)*, 57(6):34, 2010. 1

[HRTS03] Sean Hallgren, Alexander Russell, and Amnon Ta-Shma. The hidden subgroup problem and quantum computation using group representations. *SIAM Journal on Computing*, 32(4):916–934, 2003. 1

[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. 1, 1.5, 1.5, 1.5

[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. 1.5

[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. 1, 1.2, 1.5, 3, 3, 3, 3

[Joh01] Kurt Johansson. Discrete orthogonal polynomial ensembles and the Plancherel measure. *Annals of Mathematics*, 153(1):259–296, 2001. 1, 1, 1.5, 1.24

[Key06] Michael Keyl. Quantum state estimation and large deviations. *Reviews in Mathematical Physics*, 18(01):19–60, 2006. 1.3

[Kup02] Greg Kuperberg. Random words, quantum statistics, central limits, random matrices. *Methods and Applications of Analysis*, 9(1):99–118, 2002. 1.5

[KW01] Michael Keyl and Reinhard Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5):052311, 2001. 1.3, 1.3, 1.5, 1.5

[LS77] Benjamin Logan and Larry Shepp. A variational problem for random Young tableaux. *Advances in Mathematics*, 26(2):206–222, 1977. 1

[LZ04] Shunlong Luo and Qiang Zhang. Informational distance on quantum-state space. *Physical Review A*, 69(3):032106, 2004. 2.1

[Mél10] Pierre-Loïc Méliot. Kerov's central limit theorem for Schur-Weyl measures of parameter 1/2. Technical report, arXiv:1009.4034, 2010. 1.5

[Mél12] Pierre-Loïc Méliot. Fluctuations of central measures on partitions. In *24th International Conference on Formal Power Series and Algebraic Combinatorics*, pages 385–396, 2012. 1, 1.5, 1.5, 1.26

[MM15] Paulina Marian and Tudor Marian. Hellinger distance as a measure of Gaussian discord. *Journal of Physics A: Mathematical and Theoretical*, 48(11):115301, 2015. 2.1

[MRS08] Cristopher Moore, Alexander Russell, and Leonard Schulman. The symmetric group defies strong Fourier sampling. *SIAM Journal on Computing*, 37(6):1842–1864, 2008. 1

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, 2015. 1.3, 2

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 2016. (document), 1.1, 1.1, 1.1, 1.6, 1.1, 1.8, 1.2, 1.14, 1.3, 1.3, 1.3, 1.4, 1.4, 4.3, 4.5, 4.6, 4.7, 4.13, 4.7, 5.2

[Pil90] Shaiy Pilpel. Descending subsequences of random permutations. *Journal of Combinatorial Theory, Series A*, 53(1):96–116, 1990. 1, 1.1

[Rom14] Dan Romik. *The surprising mathematics of longest increasing subsequences*. Cambridge University Press, 2014. 1

[Sag01] Bruce E Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*. Springer, 2001. 6

[Sch61] Craige Schensted. Longest increasing and decreasing subsequences. *Canadian Journal of Mathematics*, 13(2):179–191, 1961. 1

[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. 1, 1.5

[Ula61] Stanislaw Ulam. Monte Carlo calculations in problems of mathematical physics. *Modern Mathematics for the Engineers*, pages 261–281, 1961. 1

[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

[VK77] Anatoly Vershik and Sergei Kerov. Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux. *Soviet Mathematics Doklady*, 18:118–121, 1977. 1

[VK81] Anatoly Vershik and Sergei Kerov. Asymptotic theory of characters of the symmetric group. *Functional analysis and its applications*, 15(4):246–255, 1981. 1.5, 1.20

[VK85] Anatoly Vershik and Sergei Kerov. Asymptotic of the largest and the typical dimensions of irreducible representations of a symmetric group. *Functional Analysis and its Applications*, 19(1):21–31, 1985. 1, 1.1

[Wer94] Lorenz Wernisch. *Dominance relation on point sets and aligned rectangles*. PhD thesis, Free University Berlin, 1994. 6, 6

[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. 1.6