Random Matrices and Their Applications
MSRI Publications
Volume **40**, 2001

Random Words, Toeplitz Determinants, and Integrable Systems I

ALEXANDER R. ITS, CRAIG A. TRACY, AND HAROLD WIDOM

Abstract. It is proved that the limiting distribution of the length of the longest weakly increasing subsequence in an inhomogeneous random word is related to the distribution function for the eigenvalues of a certain *direct sum* of Gaussian unitary ensembles subject to an overall constraint that the eigenvalues lie in a hyperplane.

1. Introduction

A class of problems — important for their applications to computer science and computational biology as well as for their inherent mathematical interest — is the statistical analysis of a string of random symbols. The symbols, called *letters*, are assumed to belong to an alphabet $\mathcal{A}$ of fixed size k. The set of all such strings (or *words*) of length N, $\mathcal{W}(\mathcal{A}, N)$, forms the sample space in the statistical analysis of these strings. A natural measure on $\mathcal{W}$ is to assign each letter equal probability, namely $1/k$, and to define the probability measure on words by the product measure. Thus each letter in a word occurs independently and with equal probability. We call such random word models *homogeneous*.

Of course, for some applications, each letter in the alphabet does not occur with the same frequency and it is therefore natural to assign to each letter i a probability p_i. If we again use the product measure for the words (letters in a word occur independently), then the resulting random word models are called *inhomogeneous*.

Fixing an ordering of the alphabet $\mathcal{A}$, a *weakly increasing subsequence* of a word

$$w = \alpha_1 \alpha_2 \dots \alpha_N \in \mathcal{W}$$

is a subsequence $\alpha_{i_1} \alpha_{i_2} \dots \alpha_{i_m}$ such that $i_1 < i_2 < \cdots < i_m$ and $\alpha_{i_1} \le \alpha_{i_2} \le \cdots \le \alpha_{i_m}$. The positive integer m is called the *length* of this weakly increasing subsequence. For each word $w \in \mathcal{W}$ we define $l_N(w)$ to equal the *length of the longest weakly increasing subsequence* in w. We now define the fundamental object of this paper:

$$F_N(n) := \mathrm{Prob}\,(l_N(w) \leq n)$$

where Prob is the inhomogeneous measure on random words. Of course, Prob depends upon N and the probabilities p_i.

Our results are of two types. To state our first results, we order the p_i so that

$$p_1 \geq p_2 \geq \cdots \geq p_k$$

and decompose out alphabet $\mathcal{A}$ into subsets $\mathcal{A}_1$, $\mathcal{A}_2$, ...such that $p_i = p_j$ if and only if i and j belong to the same $\mathcal{A}_\alpha$. Setting $k_\alpha = |\mathcal{A}_\alpha|$, we show that the limiting distribution function as $N \to \infty$ for the appropriately centered and normalized random variable l_N is related to the distribution function for the eigenvalues ξ_i in the *direct sum* of mutually independent $k_\alpha \times k_\alpha$ Gaussian unitary ensembles (GUE), conditional on the eigenvalues ξ_i satisfying $\sum \sqrt{p_i}\,\xi_i = 0$. (See [13], for example, for the notion of a GUE and other concepts in random matrix theory.) In the case when one letter occurs with greater probability than the others, this result implies that the limiting distribution of $(l_N - Np_1)/\sqrt{N}$ is Gaussian with variance equal to $p_1(1-p_1)$. In the case when all the probabilities p_i are distinct, we compute the next correction in the asymptotic expansion of the mean of l_N and find that

$$\mathrm{E}(l_N) = Np_1 + \sum_{j>1} \frac{p_j}{p_1 - p_j} + O(N^{-1/2}), \quad N \to \infty.$$

This last formula agrees quite well with finite N simulations. We expect this asymptotic formula remains valid when one letter occurs with greater probability than the others.

These results generalize work on the homogeneous model by Johansson [11] and by Tracy and Widom [19]. Since all the probabilities p_i are equal in the homogeneous model, the underlying random matrix model is $k \times k$ traceless GUE. That is, the direct sum reduces to just one term. In [19] the integrable system underlying the finite N homogeneous model was shown to be related to Painlevé V. In the isomonodromy formulation of Painlevé V [9], the associated 2×2 matrix linear ODE has two simple poles in the finite complex plane and one Poincaré index 1 irregular singular point at infinity. In Part II we will show that the finite N inhomogeneous model is represented by the isomonodromy deformations of the 2×2 matrix linear ODE which has $m+1$ simple poles in the finite complex plane and, again, one Poincaré index 1 irregular singular point at infinity. The number m is the total number of the subsets $\mathcal{A}_\alpha$, and the poles are located at zero point and at the points $-p_{i_\alpha}$ ($i_\alpha = \max \mathcal{A}_\alpha$). The integers k_α appear as the formal monodromy exponents at the respective points $-p_{i_\alpha}$. We will also analyse the monodromy meaning of the asymptotic results obtained in this part.

The results presented here are part of the recent flurry of activity centering around connections between combinatorial probability of the Robinson–Schensted–Knuth type on the one hand and random matrices and integrable systems on the other. From the point of view of probability theory, the quite surprising feature of these developments is that the methods came from Toeplitz determinants, integrable differential equations of the Painlevé type and the closely related Riemann–Hilbert techniques. The first to discover this connection at the level of distribution functions was Baik, Deift and Johansson [1] who showed that the limiting distribution of the length of the longest increasing subsequence in a random permutation is equal to the limiting distribution function of the appropriately centered and normalized largest eigenvalue in the GUE [17]. This result has been followed by a number of developments relating random permutations, random words and more generally random Young tableaux to the distribution functions of random matrix theory [2; 3; 4; 6; 8; 10; 12; 14; 18].

After the completion of this paper, Stanley [16] showed that the measure (2–1) also underlies the analysis of certain (generalized) riffle shuffles of Bayer and Diaconis [5]. Stanley relates this measure to quasisymmetric functions and does not require that p have finite support. (Many of our results generalize to the case when p does not have finite support, but we do not consider this here.)

2. Random Words

2.1. Probability Measure on Words and Partitions. The Robinson–Schensted–Knuth (RSK) algorithm is a bijection between two-line arrays w_A (or generalized permutation matrices) and ordered pairs (P, Q) of semistandard Young tableaux (SSYT). For a detailed account, see [15, Chapter 7], for example; we will use without further reference various results from symmetric function theory, which can be found in the same reference.

When the two-line arrays have the special form

$$w_A = \begin{pmatrix} 1 & 2 & \cdots & N \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \end{pmatrix},$$

with $\alpha_i \in \mathcal{A} = \{1, 2, \dots, k\}$, we identify each w_A with a word $w = \alpha_1\alpha_2\cdots\alpha_N$ of length N composed of letters from the alphabet $\mathcal{A}$; furthermore, in this case the insertion tableaux P have shape $\lambda \vdash N$, $l(\lambda) \le k$, with entries coming from $\mathcal{A}$ and the recording tableaux Q are standard Young tableau (SYT) of the same shape λ. As usual, f^λ denotes the number of SYT of shape λ and $d_\lambda(k)$ the number of SSYT of shape λ whose entries come from $\mathcal{A}$.

We define a probability measure, Prob, on $\mathcal{W}(\mathcal{A}, N)$, the set of all words w of length N formed from the alphabet $\mathcal{A}$, by the two requirements:

1. For each word w consisting of a single letter $i \in \mathcal{A}$, $\text{Prob}(w = i) = p_i$, $0 < p_i < 1$, with $\sum p_i = 1$.

2. For each $w = \alpha_1\alpha_2\dots\alpha_N \in \mathcal{W}$ and any $i_j \in \mathcal{A}$, $j = 1, 2, \dots, N$,

$$\operatorname{Prob}(\alpha_1\alpha_2\dots\alpha_N = i_1 i_2 \dots i_N) = \prod_{j=1}^{N} \operatorname{Prob}(\alpha_j = i_j) \qquad \text{(independence)}.$$

Of course, Prob depends both on N and the probabilities $\{p_i\}$.

Under the RSK correspondence, the probability measure Prob induces a probability measure on partitions $\lambda \vdash N$, which we will again denote by Prob. This induced measure is expressed in terms of f^λ and the Schur function. To see this we first recall that a tableau T has *type* $\alpha = (\alpha_1, \alpha_2, \dots)$, denoted $\alpha = \operatorname{type}(T)$, if T has $\alpha_i = \alpha_i(T)$ parts equal to i. We write

$$x^T = x_1^{\alpha_1(T)} x_2^{\alpha_2(T)} \dots$$

The combinatorial definition of the Schur function of shape λ in the variables $x = (x_1, x_2, \dots)$ is the formal power series

$$s_\lambda(x) = \sum_T x^T$$

summed over all SSYT of shape λ. The $p = \{p_1, \dots, p_k\}$ specialization of $s_\lambda(x)$ is $s_\lambda(p) = s_\lambda(p_1, p_2, \dots, p_k, 0, 0, \dots)$.

For each word $w \leftrightarrow (P, Q)$, the N entries of P consist of the N letters of w since P is formed by successive row bumping the letters from w. Because of the independence assumption,

$$p^P = p_1^{\alpha_1(P)} p_2^{\alpha_2(P)} \dots p_k^{\alpha_k(P)}$$

gives the weight assigned to word w. From the combinatorial definition of the Schur function, we observe that its p specialization is summing the weights of words w that under RSK have shape $\lambda \vdash N$. The recording tableau Q keeps track of the *order* of the letters in the word. The weights of any words with the same number of letters of each type are equal (independence), so we need merely count the number of such Q, i.e. f^λ, and multiply this by the weight of any given such word to arrive at the induced measure on partitions,

$$\operatorname{Prob}(\{\lambda\}) = s_\lambda(p)\, f^\lambda, \tag{2–1}$$

which satisfies the normalization $\sum_{\lambda\vdash N} \operatorname{Prob}(\lambda) = 1$. For the homogeneous case $p_i = 1/k$, the measure reduces to

$$\operatorname{Prob}(\lambda) = s_\lambda(1/k, 1/k, \dots, 1/k)\, f^\lambda = \frac{d_\lambda(k)\, f^\lambda}{k^N}, \quad \lambda \vdash N.$$

The Poissonization of this homogeneous measure is called the Charlier ensemble in [11].

If $l_N(w)$ equals the length of the longest *weakly* increasing subsequence in the word $w \in \mathcal{W}(\mathcal{A}, N)$, then by the RSK correspondence $w \leftrightarrow (P, Q)$, the number of boxes in the first row of P, λ_1, equals $l_N(w)$. Hence,

$$\text{Prob}\,(l_N(w) \le n) = \sum_{\substack{\lambda \vdash N \\ \lambda_1 \le n}} s_\lambda(p)\, f^\lambda. \tag{2–2}$$

2.2. Toeplitz Determinant Representation. Gessel's theorem [7] — more precisely, its dual version (see [19, § II], whose notation we follow) — is the formal power series identity

$$\sum_{\substack{\lambda \vdash N \\ \lambda_1 \le n}} s_\lambda(x) s_\lambda(y) = \det\,(T_n(\varphi))$$

where $T_n(\varphi)$ is the $n \times n$ Toeplitz matrix whose i, j entry is φ_{i-j}, where φ_i is the i^{th} Fourier coefficient of

$$\varphi(z) = \prod_{n=1}^{\infty}(1 + y_n z^{-1}) \prod_{n=1}^{\infty}(1 + x_n z), \quad z = e^{i\theta}.$$

If we define the (exponential) generating function

$$G_I(n; \{p_i\}, t) = \sum_{N=0}^{\infty} \text{Prob}\,(l_N(w) \le n)\, \frac{t^N}{N!},$$

then an immediate consequence of Gessel's identity with p specialization of the x variables and exponential specialization of the y variables and the RSK correspondence is

$$G_I(n; \{p_i\}, t) = \det\,(T_n(f_I)) \tag{2–3}$$

where

$$f_I(z) = e^{t/z} \prod_{j=1}^{k}(1 + p_j z). \tag{2–4}$$

3. Limiting Distribution

We start with the probability distribution (2–1) on the set of partitions $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_k\} \vdash N$. For f^λ we use the formula

$$f^\lambda = \frac{N!\,\Delta(h)}{h_1!\,h_2! \dots h_k!}$$

where

$$h_i = \lambda_j + k - i$$

and

$$\Delta(h) = \Delta(h_1, h_2, \dots, h_k) = \prod_{1 \le i < j \le k}(h_i - h_j). \tag{3–1}$$

Equivalently,

$$f^\lambda = \frac{\Delta(h)}{\prod_{i=1}^{k-1}\prod_{j=i}^{k-1}(\lambda_i + k - j)} \binom{N}{\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_k}.$$

The (classical) definition of the Schur function is

$$s_\lambda(p) = \frac{\det\bigl(p_i^{h_j}\bigr)}{\Delta(p)} = \frac{1}{\Delta(p)} \sum_{\sigma\in S_k} (-1)^\sigma p_1^{h_{\sigma(1)}} p_2^{h_{\sigma(2)}} \dots p_k^{h_{\sigma(k)}}. \tag{3–2}$$

This holds when all the p_i are distinct but in general the two determinants require modification, which we now describe. We order the p_i so that

$$p_1 \ge p_2 \ge \cdots \ge p_k \tag{3–3}$$

and decompose our alphabet $\mathcal{A} = \{1, 2, \dots, k\}$ into subsets $\mathcal{A}_1, \mathcal{A}_2, \dots$ such that $p_i = p_j$ if and only if i and j belong to the same $\mathcal{A}_\alpha$. Set $i_\alpha = \max \mathcal{A}_\alpha$. Think of the p_i as indeterminates and for all indices i differentiate the determinant $i_\alpha - i$ times with respect to p_i if $i \in \mathcal{A}_\alpha$. Then replace the p_i by their given values. (That this is correct follows from l'Hôpital's rule.) If we set $k_\alpha = |\mathcal{A}_\alpha|$ and write p_α for p_{i_α} then we see that $\Delta(p)$ becomes

$$\Delta'(p) = \prod_\alpha (1!\,2!\cdots(k_\alpha - 1)!) \prod_{\alpha<\beta} (p_\alpha - p_\beta)^{k_\alpha\, k_\beta} \tag{3–4}$$

and (after performing row operations) that the ith row of $\det\bigl(p_i^{h_j}\bigr)$ becomes $\bigl(h_j^{i_\alpha - i}\, p_i^{h_j - i_\alpha + i}\bigr)$. Equivalently, the partial product $\prod_{i\in\mathcal{A}_\alpha} p_i^{h_{\sigma(i)}}$ from the summand in (3–2) gets multiplied by

$$\prod_{i\in\mathcal{A}_\alpha} \Bigl(h_{\sigma(i)}^{i_\alpha - i}\, p_i^{-i_\alpha + i}\Bigr) = \Bigl(\prod_{i\in\mathcal{A}_\alpha} h_{\sigma(i)}^{i_\alpha - i}\Bigr) p_\alpha^{-k_\alpha(k_\alpha - 1)/2}. \tag{3–5}$$

In the case of distinct p_i we write our formula as

$$\begin{aligned}\mathrm{Prob}(\lambda) &= s_\lambda(p_1, \dots, p_k)\, f^\lambda \\ &= \frac{\Delta(h)}{\Delta(p)} \frac{1}{\prod_{i=1}^{k-1}\prod_{j=i}^{k-1}(\lambda_i + k - j)} \\ &\quad \times \sum_{\sigma\in S_k} (-1)^\sigma p_1^{k-\sigma(1)} \dots p_k^{k-\sigma(k)}\, p_1^{\lambda_{\sigma(1)}} \dots p_k^{\lambda_{\sigma(k)}} \binom{N}{\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_k}.\end{aligned}$$

Let $M_q(\lambda)$ denote the multinomial distribution associated with a sequence $q = \{q_1, \dots, q_k\}$,

$$M_q(\lambda) = q_1^{\lambda_1} \cdots q_k^{\lambda_k} \binom{N}{\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_k}.$$

If p_σ denotes the sequence $\{p_{\sigma^{-1}(1)}, \ldots, p_{\sigma^{-1}(k)}\}$, we may write

$$\mathrm{Prob}(\lambda) = \frac{\Delta(h)}{\Delta(p)} \frac{1}{\prod_{i=1}^{k-1}\prod_{j=i}^{k-1}(\lambda_i + k - j)} \sum_{\sigma\in S_k} (-1)^\sigma p_1^{k-\sigma(1)} \cdots p_k^{k-\sigma(k)} M_{p_\sigma}(\lambda). \tag{3–6}$$

This is the formula for distinct p_i. In the general case we must replace $\Delta(p)$ by $\Delta'(p)$ and each partial product $\prod_{i\in\mathcal{A}_\alpha} p_i^{k-\sigma(i)}$ appearing in the sum on the right must be multiplied by the factor (3–5).

The multinomial distribution $M_q(\lambda)$ has the property that the total measure of any region where $|\lambda_i - Nq_i| > \varepsilon N$ for some i and some $\varepsilon > 0$ tends exponentially to zero as $N \to \infty$. All the other terms appearing in (3–6) or its modification are uniformly bounded by a power of N. Since $\lambda_{i+1} \le \lambda_i$ for all i it follows that the contribution of the terms involving $M_q(\lambda)$ in (3–6) will tend exponentially to zero unless $q_{i+1} \le q_i$ for all i. Since $q_i = p_{\sigma^{-1}(i)}$ this shows that the contribution to (3–6) of the summand corresponding to σ is exponentially small unless σ leaves each of the sets $\mathcal{A}_\alpha$ invariant. It follows that if we denote the set of such permutations by S_k' then we may restrict the sum in (3–6) to the $\sigma \in S_k'$ without affecting the limit. Observe that when $\sigma \in S_k'$ all the $M_{p_\sigma}(\lambda)$ appearing in (3–6) equal $M_p(\lambda)$.

Write

$$\lambda_i = Np_i + \sqrt{Np_i}\,\xi_i.$$

In terms of the ξ_i the multinomial distribution $M_p(\lambda)$ converges to

$$(2\pi)^{-(k-1)/2}\, e^{-\sum \xi_i^2/2}\, \delta\Big(\sum \sqrt{q_i}\,\xi_i\Big). \tag{3–7}$$

(See Section 3.1.) Here $\delta(\sum \sqrt{q_i}\,\xi_i)$ denotes Lebesgue measure on the hyperplane $\sum \sqrt{q_i}\,\xi_i = 0$.

We now consider the contribution of the other terms in (3–6) as modified. Again, they are uniformly bounded by a power of N and the total measure of any region where $|\lambda_i - Np_i| > \varepsilon N$ for some i and some $\varepsilon > 0$ tends exponentially to zero as $N \to \infty$. Thus in determining the asymptotics of the other terms we may assume that $\lambda_i \sim Np_i$ for all i.

The constant $\Delta'(p)$ is given by (3–4). As for $\Delta(h)$, observe that the factor

$$h_i - h_j = \lambda_i - \lambda_j - i + j$$

in the product in (3–1) is asymptotically equal to $N\,(p_i - p_j)$ when i and j do not belong to the same $\mathcal{A}_\alpha$ and to $\sqrt{N\,p_\alpha}\,(\xi_i - \xi_j)$ if $i, j \in \mathcal{A}_\alpha$. It follows that

$$\Delta(h) \sim N^{k\,(k-1)/2 - \sum_\alpha k_\alpha\,(k_\alpha-1)/4} \prod p_\alpha^{k_\alpha(k_\alpha-1)/4} \prod_{\alpha<\beta} (p_\alpha - p_\beta)^{k_\alpha\,k_\beta} \prod_\alpha \Delta_\alpha(\xi),$$

where $\Delta_\alpha(\xi)$ is the Vandermonde determinant of those ξ_i with $i \in \mathcal{A}_\alpha$.

The next factor in (3–6), the reciprocal of the double product, is asymptotically

$$N^{-k\,(k-1)/2}\prod_{i=1}^{k-1}p_i^{i-k}.$$

As for the sum in (3–6) as modified, observe that since each σ now belongs to S_k' each product appearing there is equal to $\prod p_i^{k-i}$. Each such product is to be multiplied by

$$\prod_{\alpha}\left(\left(\prod_{i\in\mathcal{A}_\alpha}h_{\sigma(i)}^{i_\alpha-i}\right)p_\alpha^{-k_\alpha(k_\alpha-1)/2}\right).$$

(See (3–5).) Hence the sum itself is equal to

$$\prod_i p_i^{k-i}\prod_\alpha p_\alpha^{-k_\alpha(k_\alpha-1)/2}\sum_{\sigma\in S_k'}(-1)^\sigma\prod_\alpha\prod_{i\in\mathcal{A}_\alpha}h_{\sigma(i)}^{i_\alpha-i}.$$

Since each $\sigma\in S_k'$ is uniquely expressible as a product of $\sigma_\alpha\in S(\mathcal{A}_\alpha)$ (where $S(\mathcal{A}_\alpha)$ is the group of permutations of $\mathcal{A}_\alpha$) we have

$$\begin{aligned}\sum_{\sigma\in S_k'}(-1)^\sigma\prod_\alpha\prod_{i\in\mathcal{A}_\alpha}h_{\sigma(i)}^{i_\alpha-i}&=\prod_\alpha\sum_{\sigma_\alpha\in S(\mathcal{A}_\alpha)}(-1)^{\sigma_\alpha}\prod_{i\in\mathcal{A}_\alpha}h_{\sigma_\alpha(i)}^{i_\alpha-i}=\prod_\alpha\Delta_\alpha(h)\\&\sim N^{\sum k_\alpha(k_\alpha-1)/4}\prod_\alpha\big(p_\alpha^{k_\alpha(k_\alpha-1)/4}\,\Delta_\alpha(\xi)\big).\end{aligned}$$

Putting all this together shows that the limiting distribution is

$$(2\pi)^{-(k-1)/2}\prod_\alpha\big(1!\,2!\cdots(k_\alpha-1)!\big)^{-1}\prod_\alpha\Delta_\alpha(\xi)^2\,e^{-\sum\xi_i^2/2}\,\delta\Big(\sum\sqrt{p_i}\,\xi_i\Big).\tag{3–8}$$

This has a random matrix interpretation. It is the distribution function for the eigenvalues in the direct sum of mutually independent $k_\alpha\times k_\alpha$ Gaussian unitary ensembles, conditional on the eigenvalues ξ_i satisfying $\sum\sqrt{p_i}\,\xi_i=0$.

It remains to determine the support of the limiting distribution. In terms of the ξ_i the inequalities $\lambda_{i+1}\le\lambda_i$ are equivalent to

$$\xi_{i+1}\le\frac{N(p_i-p_{i+1})}{\sqrt{N}p_i}+\sqrt{\frac{p_i}{p_{i+1}}}\,\xi_i.$$

In the limit $N\to\infty$ this becomes no restriction if $p_{i+1}<p_i$ but becomes $\xi_{i+1}\le\xi_i$ if $p_{i+1}=p_i$. Otherwise said, the support of the limiting distribution is restricted to those $\{\xi_i\}$ for which $\xi_{i+1}\le\xi_i$ whenever i and $i+1$ belong to the same $\mathcal{A}_\alpha$. (In the random matrix interpretation it means that the eigenvalues within each GUE are ordered.) We denote this set of ξ_i by Ξ.

It now follows from (2–2) and (3–8) (also recall the ordering (3–3)) that

$$\begin{aligned}\lim_{N\to\infty}\text{Prob}\left(\frac{l_N-Np_1}{\sqrt{Np_1}}\le s\right)&=(2\pi)^{-(k-1)/2}\prod_\alpha(1!\,2!\cdots(k_\alpha-1)!)^{-1}\\&\times\int\cdots\int_{\substack{\xi_i\in\Xi\\\xi_1\le s}}\prod_\alpha\Delta_\alpha(\xi)^2\,e^{-\sum\xi_i^2/2}\,\delta(\textstyle\sum\sqrt{p_i}\,\xi_i)\,d\xi_1\cdots d\xi_k.\end{aligned}\tag{3–9}$$

When the probabilities are not all equal this may be reduced to a k_1-dimensional integral as follows. Let i denote the indices in $\mathcal{A}_1$ and j the other indices. We have to integrate

$$\prod_\alpha \Delta_\alpha(\xi)^2\, e^{-\frac{1}{2}\sum \xi_i^2 - \frac{1}{2}\sum \xi_j^2}\, \delta\Big(\sum \sqrt{p_i}\xi_i + \sum \sqrt{p_j}\xi_j\Big)$$

over the subset of Ξ where $\xi_1 \le s$. Since $\xi_1 = \max \xi_i$ and since the integrand is symmetric in the ξ_i and the ξ_j within their groups we may (by changing the normalization constant) integrate over all $\xi_i \le s$ and all ξ_j. We first fix the ξ_i and integrate over the ξ_j. These have to satisfy

$$\sum \sqrt{p_j}\xi_j = -\sum \sqrt{p_i}\xi_i = -\sqrt{p_1}\sum \xi_i.$$

If we write

$$\xi_j = \eta_j + x\sqrt{p_j}, \tag{3–10}$$

where $\{\eta_j\}$ is orthogonal to $\{\sqrt{p_j}\}$, then

$$x = \frac{\sum \sqrt{p_j}\xi_j}{\sum p_j} = -\frac{\sqrt{p_1}}{1-k_1 p_1}\sum \xi_i.$$

(Recall that $\mathcal{A}_1$ has k_1 indices.) For each $\alpha > 1$ we have $\Delta_\alpha(\xi) = \Delta_\alpha(\eta)$ since the p_j within groups are equal and

$$\sum \xi_j^2 = \sum \eta_j^2 + x^2 \sum p_j = \sum \eta_j^2 + \frac{p_1}{1-k_1 p_1}\Big(\sum \xi_i\Big)^2.$$

So the distribution function is equal to a constant times

$$\int_{-\infty}^{s} \cdots \int_{-\infty}^{s} \Delta(\xi)^2\, e^{-\frac{1}{2}\left(\sum \xi_i^2 + \frac{p_1}{1-k_1 p_1}(\sum \xi_i)^2\right)}\, d\xi_1 \cdots d\xi_{k_1} \int \prod_{\alpha>1} \Delta_\alpha(\eta)^2\, e^{-\frac{1}{2}\sum \eta_j^2}\, d\eta,$$

where the η integration is over the orthogonal complement of $\{\sqrt{p_j}\}$. The η integral is just another constant. Therefore the distribution function equals

$$\frac{1}{c_{k_1,\,p_1}} \int_{-\infty}^{s} \cdots \int_{-\infty}^{s} \Delta(\xi)^2\, e^{-\frac{1}{2}\left(\sum \xi_i^2 + \frac{p_1}{1-k_1 p_1}(\sum \xi_i)^2\right)}\, d\xi_1 \cdots d\xi_{k_1},$$

where $c_{k_1,\,p_1}$ is the integral over all of $\mathbb{R}^{k_1}$.

To evaluate this we make the substitution (3–10), but with j replaced by i and each p_j replaced by $1/\sqrt{k}$. The integral becomes

$$\int \prod_j \Delta(\eta)^2\, e^{-\frac{1}{2}\sum \eta_j^2}\, d\eta \int e^{-\frac{x^2}{2}\left(\frac{1}{k_1} + \frac{p_1}{1-k_1 p_1}\right)}\, dx,$$

taken over $x \in \mathbb{R}$ and η in hyperplane $\sum \eta_i = 0$ with Lebesgue measure. The x integral equals $\sqrt{2\pi k_1(1-k_1 p_1)}$ while the first integral equals

$$(2\pi)^{(k_1-1)/2}\, 1!\, 2! \cdots k_1!.$$

(For the last, observe that the right side of (3–9) must equal 1 when $s=\infty$.) Hence

$$c_{k_1,\,p_1}=(2\pi)^{k_1/2}\,1!\,2!\cdots k_1!\,\sqrt{k_1(1-k_1p_1)}.$$

3.1. Distinct Probabilities: the Next Approximation. If all the p_i are different then $P(\lambda):=\mathrm{Prob}(\lambda)$ equals

$$\frac{\Delta(h)}{\Delta(p)}\,\frac{1}{\prod_{i=1}^{k-1}\prod_{j=i}^{k-1}(\lambda_i+k-j)}\,\prod_{i=1}^{k}p_i^{k-i}\,M_p(\lambda) \tag{3–11}$$

plus an exponentially small correction. We recall that

$$\lambda_j=Np_j+\sqrt{Np_j}\,\xi_j$$

and compute the Fourier transform of the measure P with respect to the ξ variables. Beginning with M_p, we have

$$\begin{aligned}\widehat{M}_p(x)&=\int e^{i\sum x_j\xi_j}\,dM_p(\lambda)=e^{-i\sum\sqrt{Np_j}\,x_j}\int e^{i\sum x_j\lambda_j/\sqrt{Np_j}}\,dM_p(\lambda)\\&=e^{-i\sum\sqrt{Np_j}\,x_j}\,\Big(\sum p_j\,e^{ix_j/\sqrt{Np_j}}\Big)^N,\end{aligned}$$

since M_p is the multinomial distribution. An easy computation gives

$$\widehat{M}_p(x)=\left(1+\frac{i}{\sqrt{N}}Q(x)+O\Big(\frac{1}{N}\Big)\right)e^{-\frac12\sum x_j^2+\frac12(\sum\sqrt{p_j}x_j)^2},$$

where $Q(x)$ is a homogeneous polynomial of degree three. (In particular the limit of M_p is the inverse Fourier transform of the exponential in the above formula, which equals (3–7).)

As for the other nonconstant factors in (3–11), we have

$$\begin{aligned}\prod_{i=1}^{k-1}\prod_{j=i}^{k-1}(\lambda_i+k-j)&=\prod_{i=1}^{k-1}\big(Np_i+\sqrt{Np_i}\xi_i+O(1)\big)^{k-i}\\&=N^{k(k-1)/2}\prod_{i=1}^{k-1}p_i^{k-i}\left(1+\frac{1}{\sqrt{N}}\sum_{i=1}^{k-1}(k-i)\,\frac{\xi_i}{\sqrt{p_i}}+O\Big(\frac{1}{N}\Big)\right)\end{aligned}$$

and

$$\begin{aligned}\Delta(h)&=\prod_{i<j}\big(N(p_i-p_j)+\sqrt{N}(\sqrt{p_i}\xi_i-\sqrt{p_j}\xi_j)+O(1)\big)\\&=N^{k(k-1)/2}\Delta(p)\left(1+\frac{1}{\sqrt{N}}\sum_{i<j}\frac{\sqrt{p_i}\xi_i-\sqrt{p_j}\xi_j}{p_i-p_j}+O\Big(\frac{1}{N}\Big)\right).\end{aligned}$$

Thus the factors in (3–11) aside from M_p contribute

$$1+\frac{1}{\sqrt{N}}\left(\sum_{i<j}\frac{\sqrt{p}_i\xi_i-\sqrt{p}_j\xi_j}{p_i-p_j}-\sum_{i<j}\frac{\xi_i}{\sqrt{p}_i}\right)+O\Big(\frac{1}{N}\Big)$$
$$=1+\frac{1}{\sqrt{N}}\left(\sum_{i<j}\sqrt{\frac{p_j}{p_i}}\,\frac{\sqrt{p}_j\xi_i-\sqrt{p}_i\xi_j}{p_i-p_j}\right)+O\Big(\frac{1}{N}\Big).$$

Using the fact that multiplication by ξ_j corresponds, after taking Fourier transforms, to $-i\partial_{x_j}$ and combining this with the preceding we deduce that $\widehat{P}(x)$, the Fourier transform of $P(\lambda)$ with respect to the ξ variables, equals

$$\left(1+\frac{i}{\sqrt{N}}\sum_{i<j}\sqrt{\frac{p_j}{p_i}}\,\frac{\sqrt{p}_jx_i-\sqrt{p}_ix_j}{p_i-p_j}+\frac{i}{\sqrt{N}}\,Q(x)+O\Big(\frac{1}{N}\Big)\right)e^{-\frac12\sum x_j^2+\frac12(\sum\sqrt{p}_jx_j)^2}$$

plus a correction which is exponentially small in N.

The Mean. We have

$$\mathrm{E}(\xi_1)=\int\xi_1\,dP(\lambda)=-i\,\partial_{x_1}\widehat{P}(x)\big|_{x=0}.$$

From the preceding discussion we see that this equals

$$\frac{1}{\sqrt{Np_1}}\sum_{j>1}\frac{p_j}{p_1-p_j}+O\Big(\frac{1}{N}\Big).$$

Hence

$$\mathrm{E}(l_N)=\mathrm{E}(\lambda_1)=Np_1+\sum_{j>1}\frac{p_j}{p_1-p_j}+O\Big(\frac{1}{\sqrt{N}}\Big),\quad N\to\infty. \tag{3–12}$$

This last formula is, in fact, an accurate approximation for $\mathrm{E}(l_N)$ (for distinct p_i) for moderate values of N. Table 1 summarizes various simulations of l_N and compares the means of these simulated values with the asymptotic formula. We remark that even though the proof assumed distinct p_i, we expect the asymptotic formula to remain valid for $p_1>p_2\ge\cdots\ge p_k$. (See the last set of simulations in Table 1.)

The Variance. We write our approximation as $P=P_0+N^{-1/2}P_1+O(N^{-1})$ with corresponding expected values $\mathrm{E}=\mathrm{E}_0+N^{-1/2}\mathrm{E}_1+O(N^{-1})$. (In fact P_1 is a distribution, not a measure, but the meaning is clear.) Then the variance of λ_1 is equal to

$$Np_1\big(\mathrm{E}(\xi_1^2)-\mathrm{E}(\xi_1)^2\big)$$
$$=Np_1\left(\mathrm{E}_0(\xi_1^2)-\mathrm{E}_0(\xi_1)^2+\frac{1}{\sqrt{N}}\mathrm{E}_1(\xi_1^2)-\frac{2}{\sqrt{N}}\mathrm{E}_0(\xi_1)\,\mathrm{E}_1(\xi_1)+O\big(\frac{1}{N}\big)\right).$$

Of course $\mathrm{E}_0(\xi_1)=0$, but also

$$\mathrm{E}_1(\xi_1^2)=-\partial^2_{x_1,\,x_1}\,\widehat{P}_1(x)\big|_{x=0}=0.$$

k	Probabilities of $\{1,\dots,k\}$	N	N_S	Mean	$\mathrm{E}(l_N)$
2	$\{\frac{5}{7},\frac{2}{7}\}$	50	20 000	36.37	36.38
		100	20 000	72.12	72.10
		500	20 000	357.73	357.81
2	$\{\frac{6}{11},\frac{5}{11}\}$	50	20 000	30.54	32.27
		100	20 000	58.52	59.55
		200	20 000	113.71	114.09
		400	20 000	223.16	223.18
3	$\{\frac{1}{2},\frac{5}{14},\frac{1}{7}\}$	50	10 000	27.53	27.90
		100	10 000	52.79	52.90
		500	10 000	252.80	252.90
		1000	10 000	502.78	502.90
3	$\{\frac{3}{8},\frac{1}{3},\frac{7}{24}\}$	50	10 000	23.96	30.25
		100	10 000	44.33	49.00
		500	10 000	197.65	199.00
		1000	2 000	386.08	386.50
3	$\{\frac{3}{8},\frac{5}{16},\frac{5}{16}\}$	50	10 000	23.92	28.75
		100	10 000	44.16	47.50
		200	10 000	83.15	85.00
		400	10 000	159.30	160.00
		800	10 000	310.08	310.00

Table 1. Simulations of the length of the longest weakly increasing subsequence in inhomogeneous random words of length N for two- and three-letter alphabets. N_S is the sample size. The last column gives the asymptotic expected value (3–12).

Since $\mathrm{E}_0(\xi_1^2) - \mathrm{E}_0(\xi_1)^2 = 1 - p_1$, we find that the variance of λ_1 equals

$$Np_1(1-p_1) + O(1)$$

and so its standard deviation equals $\sqrt{Np_1(1-p_1)} + O(N^{-1/2})$.

Acknowledgments

This work was begun during the MSRI Semester Random Matrix Models and Their Applications. We wish to thank D. Eisenbud and H. Rossi for their support during this semester. This work was supported in part by the National Science Foundation through grants DMS–9801608, DMS–9802122 and DMS–9732687. The last two authors thank Y. Chen for his kind hospitality at Imperial College where part of this work was done as well as the EPSRC for the award of a Visiting Fellowship, GR/M16580, that made this visit possible.

References

[1] J. Baik, P. Deift, and K. Johansson, "On the distribution of the length of the longest increasing subsequence of random permutations", *J. Amer. Math. Soc.* **12** (1999), 1119–1178.

[2] J. Baik, P. Deift, and K. Johansson, "On the distribution of the length of the second row of a Young diagram under Plancherel measure", *Geom. Funct. Anal.* **10** (2000), 702–731.

[3] J. Baik and E. M. Rains, "The asymptotics of monotone subsequences of involutions", preprint (arXiv: math.CO/9905084).

[4] J. Baik and E. M. Rains, "Algebraic aspects of increasing subsequences", preprint (arXiv: math.CO/9905083).

[5] D. Bayer and P. Diaconis, "Trailing the dovetail shuffle to its lair", *Ann. Applied Prob.* **2** (1992), 294–313.

[6] A. Borodin, A. Okounkov and G. Olshanski, "Asymptotics of Plancherel measures for symmetric groups", *J. Amer. Math. Soc.* **13** (2000), 481–515.

[7] I. M. Gessel, "Symmetric functions and P-recursiveness", *J. Comb. Theory, Ser. A,* **53** (1990), 257–285.

[8] C. Grinstead, unpublished notes on random words, $k = 2$.

[9] M. Jimbo and T. Miwa, "Monodromy preserving deformation of linear ordinary differential equations with rational coefficients, II", *Physica D* **2** (1981), 407–448.

[10] K. Johansson, "Shape fluctuations and random matrices", *Commun. Math. Phys.* **209** (2000), 437–476.

[11] K. Johansson, "Discrete orthogonal polynomial ensembles and the Plancherel measure", preprint (arXiv: math.CO/9906120).

[12] G. Kuperberg, "Random words, quantum statistics, central limits, random matrices", preprint (arXiv: math.PR/9909104).

[13] M. L. Mehta, *Random matrices*, 2nd ed., Academic Press, San Diego, 1991.

[14] A. Okounkov, "Random matrices and random permutations", preprint (arXiv: math.CO/9903176).

[15] R. P. Stanley, *Enumerative combinatorics*, Vol. 2, Cambridge University Press, Cambridge, 1999.

[16] R. P. Stanley, "Generalized riffle shuffles and quasisymmetric functions", preprint (arXiv: math.CO/9912025).

[17] C. A. Tracy and H. Widom, "Level-spacing distributions and the Airy kernel", *Commun. Math. Phys.* **159** (1994), 151–174.

[18] C. A. Tracy and H. Widom, "Random unitary matrices, permutations and Painlevé", *Commun. Math. Phys.* **207** (1999), 665–685.

[19] C. A. Tracy and H. Widom, "On the distributions of the lengths of the longest monotone subsequences in random words", preprint (arXiv: math.CO/9904042).

Alexander R. Its
Department of Mathematics
Indiana University–Purdue University Indianapolis
Indianapolis, IN 46202
United States
itsa@math.iupui.edu

Craig A. Tracy
Department of Mathematics
Institute of Theoretical Dynamics
University of California
Davis, CA 95616
United States
tracy@itd.ucdavis.edu

Harold Widom
Department of Mathematics
University of California
Santa Cruz, CA 95064
United States
widom@math.ucsc.edu