On a variance related to the Ewens sampling formula

Eugenijus Manstavičius, Žydrūnas Žilinskas

Faculty of Mathematics and Informatics, Vilnius University
Naugarduko str. 24, LT-03225 Vilnius, Lithuania
eugenijus.manstavicius@mif.vu.lt

Received: 12, March 2011 / **Revised:** 14 November 2011 / **Published online:** 7 December 2011

Abstract. A one-parameter multivariate distribution, called the Ewens sampling formula, was introduced in 1972 to model the mutation phenomenon in genetics. The case discussed in this note goes back to Lynch's theorem in the random binary search tree theory. We examine an additive statistics, being a sum of dependent random variables, and find an upper bound of its variance in terms of the sum of variances of summands. The asymptotically best constant in this estimate is established as the dimension increases. The approach is based on approximation of the extremal eigenvalues of appropriate integral operators and matrices.

Keywords: random permutation, cycle structure, integral operator, matrix eigenvalue, Jacobi polynomial.

1 Introduction and result

Let $\mathbf{S}_n$ denote the symmetric group of permutations σ acting on $n \geq 1$ letters. Each $\sigma \in \mathbf{S}_n$ has a unique representation (up to the order) by the product of independent cycles κ_i:

$$\sigma = \kappa_1 \cdots \kappa_w, \tag{1}$$

where $w = w(\sigma)$ denotes the number of cycles. Denote by $k_j(\sigma) \geq 0$ the number of cycles in (1) of length j for $1 \leq j \leq n$ and $\bar{k}(\sigma) := (k_1(\sigma), \ldots, k_n(\sigma))$. The latter is called the *cycle vector* of permutation σ. Let $\nu_{n,\theta}$ be the *Ewens probability measure* on $\mathbf{S}_n$ defined by

$$\begin{aligned}\nu_{n,\theta}\bigl(\{\sigma\}\bigr) &= \theta^{w(\sigma)}/\bigl(\theta(\theta+1)\cdots(\theta+n-1)\bigr)\\ &=: \theta^{w(\sigma)}/\bigl(\theta^{(n)}\bigr), \quad \sigma \in \mathbf{S}_n,\end{aligned}$$

where $\theta > 0$ is a parameter. If we set $\ell(\bar{s}) = 1s_1 + \cdots + ns_n$ for a vector $\bar{s} = (s_1, \ldots, s_n) \in \mathbf{Z}_+^n$, then $\ell(\bar{k}(\sigma)) \equiv n$ and, as it is shown in [1],

$$\begin{aligned}\nu_{n,\theta}\bigl(\bar{k}(\sigma) = \bar{s}\bigr) &= \mathbf{1}\bigl\{\ell(\bar{s}) = n\bigr\}\frac{n!}{\theta^{(n)}}\prod_{j=1}^{n}\left(\frac{\theta}{j}\right)^{s_j}\frac{1}{s_j!}\\ &= P\bigl(\bar{\xi}_\theta = \bar{s} \bigm| \ell(\bar{\xi}_\theta) = n\bigr),\end{aligned} \tag{2}$$

where $\bar{\xi}_\theta = (\xi_{1,\theta}, \ldots, \xi_{n,\theta})$ and $\xi_{j,\theta}$, $1 \le j \le n$, are mutually independent Poisson r.vs given in some probability space $\{\Omega, \mathcal{F}, P\}$ with parameter $\mathbf{E}\xi_{j,\theta} = \theta/j$.

The probability in (2), assigned to the vector $\bar{s} \in \mathbf{Z}_+^n$, is called the *Ewens sampling formula*. It has been introduced by W.J. Ewens [2] to describe the sampling distribution of a sample of n genes from a large population. In this case, the *allelic partition* $\bar{s} = (s_1, \ldots, s_n)$, contains all the information available in a sample, that is, s_j denotes the number of alleles represented j times in it, $j = 1, \ldots, n$. In the so-called neutral alleles model of population genetics, the parameter θ is interpreted as the mutation rate (see [3]). For a comprehensive account of recent applications in combinatorics and statistics, we refer to [1] and [4]. Now, we just mention that the case $\theta = 2$ explored in the present paper has some connections to the random binary search tree theory (see [5]).

Apart from $w(\sigma)$, other statistics, called *completely additive functions*,

$$h(\sigma) := h_{\bar{a}}(\sigma) := a_1 k_1(\sigma) + \cdots + a_n k_n(\sigma), \tag{3}$$

where $\bar{a} := (a_1, \ldots, a_n) \in \mathbf{R}^n$ is a non-zero vector, appear in applications rather often. For instance, $h(\sigma)$ with $a_j = \log j$, $j \le n$, is a good approximation for the logarithm of the group-theoretical order of $\sigma \in \mathbf{S}_n$ (see [1]). The case with $a_j = \{xj\}$, where $\{u\}$ stands for the fractional part of $u \in \mathbf{R}$, is met in the theory of random permutation matrices (see [6]).

By (2), under $\nu_{n,\theta}$, the function $h(\sigma)$ is a sum of dependent r.vs. This fact raises some obstacles, seen already in the analysis of power moments carried out by the first author [7, 8] and [9] in the case $\theta = 1$. To overcome the difficulties arising from the dependency, proving limit theorems for $h(\sigma)$, one needs specified approaches (see, for instance, [10, 11] or [12] and the references therein). We now draw the reader's attention to the variance.

Denote by $A_{n,\theta}(\bar{a}) := \mathbf{E}_{n,\theta} h(\sigma)$ and $D_{n,\theta}(\bar{a}) := \mathbf{V}\mathrm{ar}_{n,\theta} h(\sigma)$ the mean value and the variance of function $h(\sigma)$ under the probability measure $\nu_{n,\theta}$. Set

$$\tau_{n,\theta} := \sup_{\bar{a} \ne \bar{0}} \biggl(D_{n,\theta}(\bar{a}) \Big/ \sum_{j \le n} \mathbf{V}\mathrm{ar}_{n,\theta}\bigl(a_j k_j(\sigma)\bigr) \biggr).$$

The problem is to estimate its discrepancy from 1 which is an indicator of the dependence among the summands. The first author [7] has succeeded to explore the case $\theta = 1$.

Theorem M. *We have*

$$\tau_{n,1} = \frac{3}{2} + \mathrm{O}\biggl(\frac{1}{n}\biggr)$$

as $n \to \infty$.

A sketchy *proof* of this theorem is given in [7]. It is based on the spectral analysis of some integral operators. Nevertheless, the same approach to $\tau_{n,\theta}$ for other $\theta > 0$ leads to different operators in each case. Therefore, having the aim to expose our method in full detail and to give an instance of another appearing operator, we now chose the case $\theta = 2$.

Theorem. *We have*

$$\tau_{n,2} = \frac{4}{3} + \mathrm{O}\left(\frac{1}{n}\right)$$

as $n \to \infty$.

The idea of our proof goes back to the number theoretical papers by J. Kubilius [13] and [14] and is explained in the next section. To our knowledge, apart from [7], it has not been applied in combinatorics.

The lower bound of $\tau_{n,2}$ is found in Section 3 and the upper estimate is obtained in the last section.

2 An idea

First of all, we express the variance as a quadratic form. It appears that the eigenvalues of appropriate integral operators well approximate the eigenvalues of the involved matrices as their order increases. Further, the eigenfunctions of operators are used to find the vectors $\bar{a}$ giving the extremal values of the variances.

To find an expression of the variance $D_{n,\theta}(\bar{a})$, we apply the following G.A. Watterson's result.

Lemma 1. *Set* $x_{(r)} := x(x-1)\cdots(x-r+1)$ *if* $r \in \mathbf{Z}_+$ *and, for arbitrary* $l \in \mathbf{N}$ *and* $r_1, \dots, r_l \in \mathbf{Z}_+$, *define* $m := 1r_1 + \cdots + lr_l$. *Then*

$$\begin{aligned}
&\mathbf{E}_{n,\theta}\bigl(k_1(\sigma)_{(r_1)} \cdots k_l(\sigma)_{(r_l)}\bigr) \\
&\quad = \mathbf{1}\{m \le n\} \binom{\theta + n - m - 1}{n - m} \binom{\theta + n - 1}{n}^{-1} \prod_{j=1}^{l} \left(\frac{\theta}{j}\right)^{r_j}.
\end{aligned}$$

Proof. See [15] or [1, p. 96]. □

Lemma 2. *We have* $D_{n,2}(\bar{a}) = 2B(\bar{a}) - 4\Delta(\bar{a})$, *where*

$$B(\bar{a}) = \sum_{j \le n} \frac{a_j^2}{j} \left(1 - \frac{j}{n+1}\right),$$

and

$$\Delta(\bar{a}) = \sum_{\substack{i,j \le n \\ i+j > n}} \frac{a_i a_j}{ij} \left(1 - \frac{i}{n+1}\right)\left(1 - \frac{j}{n+1}\right) + \frac{1}{(n+1)^2} \sum_{i+j \le n} a_i a_j.$$

Proof. Since $x_{(0)} = 1$, applying Lemma 1 for $l = m = j \cdot 1$ and $r_i = 0$ if $1 \le i \le j-1$, we have

$$A_{n,2}(\bar{a}) = 2 \sum_{j \le n} \frac{a_j}{j} \left(1 - \frac{j}{n+1}\right).$$

Similarly,

$$\begin{aligned}\mathbf{E}_{n,2}h^2(\sigma) &= \sum_{i,j\le n} a_i a_j \mathbf{E}_{n,2}\big(k_i(\sigma)k_j(\sigma)\big)\\ &= 4\sum_{i+j\le n}\frac{a_ia_j}{ij}\left(1-\frac{i+j}{n+1}\right)+2\sum_{j\le n}\frac{a_j^2}{j}\left(1-\frac{j}{n+1}\right).\end{aligned}$$

In the second step, we used Lemma 1 separately in the cases $i\neq j$ and $i=j$. In the latter, dealing with $\mathbf{E}_{n,2}k_j(\sigma)^2$, we also applied the relation $x^2=x_{(2)}+x_{(1)}$. Inserting these expressions into the equality $D_{n,2}(\bar a)=\mathbf{E}_{n,2}h^2(\sigma)-(A_{n,2}(\bar a))^2$, we obtain the desired formula.

The lemma is proved. □

Corollary 1. *For all $n\ge 1$, we have*

$$D_{n,2}(\bar a)\le 4B(\bar a) \tag{4}$$

and

$$\left|\sum_{j\le n}\mathbf{V}\mathrm{ar}_{n,2}\big(a_jk_j(\sigma)\big)-2B(\bar a)\right|\le\frac{6}{n}B(\bar a). \tag{5}$$

Proof. If a_j, $j\le n$, are of one sign, then, omitting $\Delta(\bar a)\ge 0$ in the expression of variance obtained in Lemma 2, we have $D_{n,2}(\bar a)\le 2B(\bar a)$ for all $n\ge 1$ and $\bar a\in\mathbf{R}^n$.

Further, splitting $a_j=a_j^+-a_j^-$, where a_j^+ and a_j^- are respectively the positive and negative parts of a_j, we define $\bar a'=(a_1^+,\dots,a_n^+)$ and $\bar a''=(a_1^-,\dots,a_n^-)$. Then, by virtue of $(x+y)^2\le 2x^2+2y^2$, $x,y\in\mathbf{R}$,

$$D_{n,2}(\bar a)=D_{n,2}\big(\bar a'-\bar a''\big)\le 2D_{n,2}\big(\bar a'\big)+2D_{n,2}\big(\bar a''\big).$$

Now, applying the just proved inequality twice, we obtain

$$D_{n,2}(\bar a)\le 4B\big(\bar a'\big)+4B\big(\bar a''\big)=4B(\bar a).$$

To prove (5), from Lemma 1, we have

$$\begin{aligned}&\sum_{j\le n}\mathbf{V}\mathrm{ar}_{n,2}\big(a_jk_j(\sigma)\big)-2B(\bar a)\\ &\quad=4\sum_{j\le n/2}\frac{a_j^2}{j^2}\left(1-\frac{2j}{n+1}\right)-4\sum_{j\le n}\frac{a_j^2}{j^2}\left(1-\frac{j}{n+1}\right)^2\\ &\quad=-\frac{4}{(n+1)^2}\sum_{j\le n/2}a_j^2-4\sum_{n/2\le j\le n}\frac{a_j^2}{j^2}\left(1-\frac{j}{n+1}\right)^2\\ &\quad=:-4\big(\Sigma'+\Sigma''\big).\end{aligned}$$

Now,

$$\Sigma' \le \frac{1}{n} \sum_{j \le n/2} \frac{a_j^2}{j}\left(1 - \frac{j}{n+1}\right)$$

and

$$\Sigma'' \le \frac{3}{2n} \sum_{n/2 < j \le n} \frac{a_j^2}{j}\left(1 - \frac{j}{n+1}\right).$$

Inserting the last inequalities into the previous expression, we complete the proof of (5).

The corollary is proved. □

By virtue of (5), instead of $\tau_{n,2}$, we may examine the ratio $\Delta(\bar{a})/B(\bar{a})$. Our idea lays in choosing the vectors $\bar{a}$ so that these quadratic forms could be approximated by appropriate Riemann integrals. The natural choice is

$$a_j := a_{j,n} := \frac{j}{n+1}\, g\left(\frac{j}{n+1}\right), \quad 1 \le j \le n,$$

where $g : [0,1] \to \mathbf{R}$ is a continuous function. For convenience, in the sums below, we formally add one more summand corresponding to $j = n+1$ though it equals zero at all places of appearance. Then

$$B(\bar{a}) \approx \int_0^1 x(1-x) g^2(x)\, \mathrm{d}x.$$

Setting

$$\gamma_j(\bar{a}) := \left(1 - \frac{j}{n+1}\right) \sum_{n-j+1 < i \le n+1} \frac{a_i}{i}\left(1 - \frac{i}{n+1}\right) + \frac{j}{(n+1)^2} \sum_{i \le n-j+1} a_i$$

for $1 \le j \le n+1$, we have a more convenient expression

$$\Delta(\bar{a}) = \sum_{j \le n+1} \frac{a_j}{j} \gamma_j(\bar{a}).$$

This leads to

$$\Delta(\bar{a}) \approx \int_0^1 g(x)\left[(1-x)\int_{1-x}^1 (1-u) g(u)\, \mathrm{d}u + x \int_0^{1-x} u g(u)\, \mathrm{d}u\right] \mathrm{d}x.$$

Now, assume that $g(x)$, $0 \le x \le 1$, is a solution to the equation

$$(1-x)\int_{1-x}^1 (1-u) g(u)\, \mathrm{d}u + x \int_0^{1-x} u g(u)\, \mathrm{d}u = \lambda x (1-x) g(x), \tag{6}$$

with some $\lambda \in \mathbf{R}$. Then, for the additive function $h(\sigma) = h_{\bar{a}}(\sigma)$ defined in (3) via such $\bar{a}$,

$$D_{n,2}(\bar{a}) \approx 2(1-2\lambda)B(\bar{a}).$$

If our intuition is true, among all these λ we may look for the values giving the extremes of ratio $D_{n,2}/B(\bar{a})$. Following this idea, we have to examine the operator

$$g(x) \mapsto \frac{1}{x}\int_{1-x}^{1}(1-u)g(u)\,\mathrm{d}u + \frac{1}{1-x}\int_{0}^{1-x} ug(u)\,\mathrm{d}u \tag{7}$$

defined on the space of continuous functions $g : [0,1] \to \mathbf{R}$ and its eigenvalues λ. Using the substitutions $g(u) = p(2u-1)$ and $y = 2x-1$ with a continuous function $p : [-1,1] \to \mathbf{R}$, from (6), we arrive at the equation

$$(1-y)\int_{-y}^{1}(1-t)p(t)\,\mathrm{d}t + (1+y)\int_{-1}^{-y}(1+t)p(t)\,\mathrm{d}t = 2\lambda\bigl(1-y^2\bigr)p(y).$$

The solutions to it are twice differentiable, therefore they also satisfy

$$\lambda\bigl(1-y^2\bigr)p''(y) - 4\lambda yp'(y) + p(-y) - 2\lambda p(y) = 0.$$

If $\lambda \neq 0$, for even and uneven functions $p(y)$, respectively, this leads to the differential equations

$$\bigl(1-y^2\bigr)p''(y) - 4yp'(y) + \bigl(\pm\lambda^{-1} - 2\bigr)p(y) = 0. \tag{8}$$

The latter are well known in the theory of Jacobi polynomials $P_r^{(1,1)}(t)$, $r \geq 0$, which are defined (see in [16, Sect. V.2] or [17, Sect. II.7]) by

$$\bigl(1-t^2\bigr)P_r^{(1,1)}(t) = \frac{(-1)^r}{2^r r!}\frac{\mathrm{d}^r}{\mathrm{d}t^r}\bigl(1-t^2\bigr)^{r+1}$$

or by

$$P_r^{(1,1)}(t) = \frac{1}{2^r}\sum_{k=0}^{r}\binom{r+1}{r-k}\binom{r+1}{k}(t-1)^k(t+1)^{r-k}.$$

We will use their properties listed in the next lemma.

Lemma 3. *Let δ_{mr} be the Kronecker symbol and $0 \leq m \leq r$. Then*

(i) $$\int_{-1}^{1}\bigl(1-t^2\bigr)P_m^{(1,1)}(t)P_r^{(1,1)}(t)\,\mathrm{d}t = \frac{8(r+1)\delta_{mr}}{(2r+3)(r+2)};$$

(ii) $P_r^{(1,1)}(t)$ *satisfies the differential equation*

$$\bigl(1-t^2\bigr)p''(t) - 4tp'(t) + r(r+3)p(t) = 0;$$

(iii) $$P_r^{(1,1)}(-t) = (-1)^r P_r^{(1,1)}(t).$$

Proof. See [16, Sect. V.2] and [17, Sect. II.7]. □

Corollary 2. *The operator defined by* (7) *has the eigenfunctions*

$$g_r(x) := P_r^{(1,1)}(2x-1), \quad 0 \le x \le 1,$$

corresponding to the eigenvalues

$$\lambda_r := \frac{(-1)^r}{(r+1)(r+2)},$$

where $r = 0, 1, \dots$.

Proof. Applying $2 + r(r+3) = (r+1)(r+2)$, the properties (ii) and (iii) given in the lemma, we see that equation (8) is satisfied by $p(y) = P_r^{(1,1)}(y)$ if $\lambda = \lambda_r$. Recalling the former substitutions, we complete the proof. □

3 The lower bound

Let us keep our previous notation. As we have noted, the numbers λ_r and the vectors $\bar{a}^r := (a_{r1}, \dots, a_{rn})$, where $a_{rj} := (j/(n+1))g_r(j/(n+1))$, $1 \le j \le n$ and $r \ge 0$, are worth to be exploited. The technical calculations are presented in a few lemmata, the most of them are based on the next well known Koksma inequality.

Lemma 4. *If* $f : [0,1] \to \mathbf{R}$ *is continuously differentiable and* $N \in \mathbf{N}$*, then*

$$\left| \frac{1}{N} \sum_{j \le N} f\left(\frac{j}{N}\right) - \int_0^1 f(x)\,\mathrm{d}x \right| \le \frac{1}{N} \int_0^1 \left|f'(x)\right| \mathrm{d}x.$$

Proof. See, for instance, [18, Sect. 2.5]. □

Afterwards, all remainder term estimates will be dependent on r only.

Lemma 5. *For each* $r \ge 0$*, we have*

$$B(\bar{a}^r) = \frac{(r+1)}{(2r+3)(r+2)} + \mathrm{O}\left(\frac{1}{n}\right)$$

and

$$\Delta(\bar{a}^r) = \frac{(-1)^r}{(2r+3)(r+2)^2} + \mathrm{O}\left(\frac{1}{n}\right)$$

as $n \to \infty$.

Proof. By the Koksma inequality and relation (i) of Lemma 3, we have

$$\begin{aligned} B(\bar a^r) &= \int_0^1 x(1-x)g_r^2(x)\,\mathrm{d}x + \mathrm{O}\biggl(\frac{1}{n}\biggr) \\ &= \frac{1}{8}\int_{-1}^1 \bigl(1-t^2\bigr)\bigl(P_r^{(1,1)}(t)\bigr)^2\,\mathrm{d}t + \mathrm{O}\biggl(\frac{1}{n}\biggr) \\ &= \frac{r+1}{(2r+3)(r+2)} + \mathrm{O}\biggl(\frac{1}{n}\biggr). \end{aligned}$$

Calculating $\gamma_j(\bar a^r)$, we introduce the temporary notation $K := n-j+1$,

$$f_1(x) := \biggl(1-\frac{jx+K}{n+1}\biggr)g_r\biggl(\frac{jx+K}{n+1}\biggr), \qquad f_2(x) := \frac{Kx}{n+1}g_r\biggl(\frac{Kx}{n+1}\biggr)$$

and arrive at

$$\gamma_j(\bar a^r) := \frac{K}{(n+1)^2}\sum_{1\le i\le j} f_1\biggl(\frac{i}{j}\biggr) + \frac{j}{(n+1)^2}\sum_{1\le i\le K} f_2\biggl(\frac{i}{K}\biggr) + \mathrm{O}\biggl(\frac{Kj}{(n+1)^3}\biggr).$$

Now, applying Lemma 4 for $f(x) = f_l(x)$, $l=1,2$, and exploiting also equation (6), we obtain

$$\begin{aligned} \gamma_j(\bar a^r) &= \biggl(1-\frac{j}{n+1}\biggr)\int_{1-j/(n+1)}^1 (1-t)g_r(t)\,\mathrm{d}t \\ &\quad + \frac{j}{n+1}\int_0^{1-j/(n+1)} tg_r(t)\,\mathrm{d}t + \mathrm{O}\biggl(\frac{j}{n^2}\biggl(1-\frac{j}{n+1}\biggr)\biggr) \\ &= \lambda_r\frac{j}{n+1}\biggl(1-\frac{j}{n+1}\biggr)g_r\biggl(\frac{j}{n+1}\biggr) + \mathrm{O}\biggl(\frac{j}{n^2}\biggl(1-\frac{j}{n+1}\biggr)\biggr) \end{aligned} \tag{9}$$

for $1\le j\le n+1$. Hence again by (i) of Lemma 3,

$$\begin{aligned} \Delta(\bar a^r) &= \frac{1}{n+1}\sum_{j\le n+1} g_r\biggl(\frac{j}{n+1}\biggr)\gamma_j(\bar a^r) \\ &= \lambda_r\int_0^1 x(1-x)g_r^2(x)\,\mathrm{d}x + \mathrm{O}\biggl(\frac{1}{n}\biggr) \\ &= \frac{\lambda_r(r+1)}{(2r+3)(r+2)} + \mathrm{O}\biggl(\frac{1}{n}\biggr) = \frac{(-1)^r}{(2r+3)(r+2)^2} + \mathrm{O}\biggl(\frac{1}{n}\biggr). \end{aligned}$$

The lemma is proved. □

Corollary 3. *We have*

$$\tau_{n,2} \geq \frac{4}{3} + \mathrm{O}\left(\frac{1}{n}\right) \tag{10}$$

as $n \to \infty$.

Proof. It suffices to apply (5) and Lemma 5 for $r = 1$. Indeed,

$$\begin{aligned}\tau_{n,2} &\geq D_{n,\theta}\big(\bar{a}^1\big) \Big/ \sum_{j\leq n} \mathbf{V}\mathrm{ar}_{n,\theta}\big(a_{j1}k_j(\sigma)\big) \\ &= \big(2B\big(\bar{a}^1\big) - 4\Delta\big(\bar{a}^1\big)\big) / \big(2B\big(\bar{a}^1\big) + \mathrm{O}\big(n^{-1}\big)\big) \\ &= \left(\frac{16}{45} + \mathrm{O}\big(n^{-1}\big)\right) \Big/ \left(\frac{4}{15} + \mathrm{O}\big(n^{-1}\big)\right) = \frac{4}{3} + \mathrm{O}\big(n^{-1}\big).\end{aligned}$$

The corollary is proved. □

4 The upper bound

Let $\|\cdot\|$ denote the Euclidean norm in $\mathbf{R}^n$. Dealing with the quadratic forms $B(\bar{a})$ and $\Delta(\bar{a})$, it is convenient to apply the substitution

$$a_j = j^{1/2}\left(1 - \frac{j}{n+1}\right)^{-1/2} x_j, \quad 1 \leq j \leq n,$$

converting $B(\bar{a})$ to the sum of squares, that is, to $\|\bar{x}\|^2$. Then $\Delta(\bar{a})$ becomes the quadratic form $\mathcal{Q}(\bar{x}) =: \bar{x}Q\bar{x}'$, where $\bar{x}'$ is the vector-column and Q is a symmetric matrix. If q_{ij}, $1 \leq i, j \leq n$, are the entries of the latter, then

$$q_{ij} = (ij)^{-1/2}\left(1 - \frac{i}{n+1}\right)^{1/2}\left(1 - \frac{j}{n+1}\right)^{1/2}$$

if $i + j > n$ and

$$q_{ij} = \frac{(ij)^{1/2}}{(n+1)^2}\left(1 - \frac{i}{n+1}\right)^{-1/2}\left(1 - \frac{j}{n+1}\right)^{-1/2}$$

if $i + j \leq n$.

Now, by virtue of (5),

$$\begin{aligned}\tau_{n,2} &= 1 - 2\inf_{\bar{x}\neq\bar{0}} \frac{\mathcal{Q}(\bar{x})}{\|\bar{x}\|^2} + \mathrm{O}\left(\frac{1}{n}\right) \\ &= 1 - 2\inf_{\bar{x}\neq\bar{0}} \frac{1}{\|\bar{x}\|^2}\sum_{j=1}^{n} \mu_j x_j^2 + \mathrm{O}\left(\frac{1}{n}\right) \\ &\leq 1 - 2\min_{1\leq j\leq n} \mu_j + \mathrm{O}\left(\frac{1}{n}\right),\end{aligned} \tag{11}$$

where μ_j, $1 \leq j \leq n$, denote the eigenvalues of the matrix Q. It remains to find their minimal value.

Lemma 6. *Let $\bar v \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}$ be arbitrary. If M is a symmetric real $n \times n$ matrix, then there exists its eigenvalue μ such that*

$$|\alpha - \mu| \|\bar v\| \le \|\bar v M - \alpha \bar v\|.$$

Proof. This is Lemma 5.6 in [19]. □

Lemma 7. *Let $r \ge 0$ be fixed and λ_r be defined in Corollary 2. There exists an eigenvalue μ_r (relabelled if necessary) of Q such that*

$$\mu_r = \lambda_r + \mathrm{O}\bigl(n^{-1}\bigr)$$

provided that n is sufficiently large.

Proof. We apply the previous lemma with $\alpha = \lambda_r$ and $\bar v = \bar v^r := (v_{r1}, \dots, v_{rn})$, where

$$v_{ri} = \frac{i^{1/2}}{n+1}\left(1 - \frac{i}{n+1}\right)^{1/2} g_r\left(\frac{i}{n+1}\right), \quad i \le n.$$

If $\bar y^r = (y_{r1}, \dots, y_{rn}) = \bar v^r Q$, then, recalling the previous notation and using (9), we have

$$\begin{aligned} y_{rj} &= \sum_{i \le n} q_{ij} v_{ri} = j^{-1/2}\left(1 - \frac{j}{n+1}\right)^{-1/2} \gamma_j\bigl(\bar a^r\bigr) \\ &= \lambda_r v_{rj} + \mathrm{O}\left(\frac{j^{1/2}}{n^2}\left(1 - \frac{j}{n+1}\right)^{1/2}\right). \end{aligned}$$

Hence

$$\bigl\|\bar v^r Q - \lambda_r \bar v^r\bigr\| = \mathrm{O}\bigl(n^{-1}\bigr).$$

By Lemma 5, $\|\bar v^r\|^2 = B(\bar a^r) \ge c_r > 0$ for sufficiently large n.

The claim now follows from Lemma 6. □

The just proved lemma gives some numeration of the first eigenvalues μ_r of the matrix Q. There is no repetition of them if n is sufficiently large. Actually, we can even chose some unbounded sufficiently slowly increasing sequence of natural numbers r_n such that

$$\max_{0 \le r \le r_n} |\mu_r - \lambda_r| \le r_n^{-2} \tag{12}$$

as $n \to \infty$. Extend the numeration to list the remaining eigenvalues. The latter μ_r, $r_n < r \le n-1$, maybe, are written with repetitions or repeating those with small indexes. However, we will prove that μ_1 is the minimal among all of μ_r, where $0 \le r \le n-1$.

Lemma 8. *We have*

$$\begin{aligned} \sum_{0 \le r \le n-1} \mu_r^2 &= \sum_{r \ge 0} \lambda_r^2 + \mathrm{O}\left(\frac{\log n}{n}\right) \\ &= \frac{\pi^2}{3} - 3 + \mathrm{O}\bigl(n^{-1} \log n\bigr). \end{aligned} \tag{13}$$

and

$$\min_{0\le r\le n-1}\mu_r=\mu_1 \tag{14}$$

if n *is sufficiently large.*

Proof. Observe that

$$\sum_{r\ge 0}\lambda_r^2=\sum_{r\ge 0}\biggl(\frac{1}{(r+1)^2}+\frac{1}{(r+2)^2}-2\biggl(\frac{1}{r+1}-\frac{1}{r+2}\biggr)\biggr)=\frac{\pi^2}{3}-3. \tag{15}$$

On the other hand, using the well known property of the matrix eigenvalues, we arrive at

$$\begin{aligned}\sum_{0\le r\le n-1}\mu_r^2&=\sum_{i,j\le n}q_{ij}^2\\&=\sum_{\substack{i,j\le n\\ i+j>n}}\frac{1}{ij}\biggl(1-\frac{i}{n+1}\biggr)\biggl(1-\frac{j}{n+1}\biggr)\\&\quad+\frac{1}{(n+1)^4}\sum_{i+j\le n}ij\biggl(1-\frac{i}{n+1}\biggr)^{-1}\biggl(1-\frac{j}{n+1}\biggr)^{-1}\\&=:\Sigma_1+\Sigma_2.\end{aligned} \tag{16}$$

Now

$$\begin{aligned}\Sigma_1&=\biggl(\sum_{n/2<j\le n}\frac{1}{j}\biggl(1-\frac{j}{n+1}\biggr)\biggr)^2\\&\quad+2\sum_{j\le n/2}\frac{1}{j}\biggl(1-\frac{j}{n+1}\biggr)\sum_{n-j<i\le n}\frac{1}{i}\biggl(1-\frac{i}{n+1}\biggr)\\&=:\Sigma_{11}+2\Sigma_{12}.\end{aligned}$$

Further, approximating the sums by appropriate integrals, from Lemma 4 we obtain

$$\Sigma_{11}=\biggl(\int_{1/2}^{1}\frac{1-x}{x}\,\mathrm{d}x+\mathrm{O}\biggl(\frac{1}{n}\biggr)\biggr)^2=\biggl(\log 2-\frac{1}{2}\biggr)^2+\mathrm{O}\bigl(n^{-1}\bigr)$$

and

$$\begin{aligned}\Sigma_{12}&=\sum_{j\le n/2}\frac{1}{j}\biggl(1-\frac{j}{n+1}\biggr)\biggl(\int_{1-j/n}^{1}\frac{1-u}{u}\,\mathrm{d}u+\mathrm{O}\biggl(\frac{1}{n}\biggr)\biggr)\\&=\int_0^{1/2}\frac{1-x}{x}\bigl(-\log(1-x)-x\bigr)\,\mathrm{d}x+\mathrm{O}\biggl(\frac{\log n}{n}\biggr)\\&=:I+\mathrm{O}\bigl(n^{-1}\log n\bigr).\end{aligned}$$

If $\mathrm{Li}_2(x)$ denotes the dilogarithm function, then

$$\begin{aligned} I &= \mathrm{Li}_2\left(\frac{1}{2}\right) + \int_0^{1/2} \big(\log(1-x) - (1-x)\big)\,\mathrm{d}x \\ &= \mathrm{Li}_2\left(\frac{1}{2}\right) + \frac{\log 2}{2} - \frac{7}{8} = \frac{\pi^2}{12} - \frac{\log^2 2}{2} + \frac{\log 2}{2} - \frac{7}{8} \end{aligned} \tag{17}$$

(see [20, Sect. 27.7.3, p. 1004]).

Similarly,

$$\begin{aligned} \Sigma_2 &= \left(\frac{1}{n+1}\sum_{j\le n/2}\frac{j}{n+1-j}\right)^2 \\ &\quad + \frac{2}{(n+1)^2}\sum_{n/2<j\le n}\frac{j}{n+1-j}\sum_{i\le n-j}\frac{i}{n+1-i} \\ &=: \Sigma_{21} + 2\Sigma_{22}. \end{aligned}$$

Now again

$$\Sigma_{21} = \left(\int_0^{1/2}\frac{x}{1-x}\,\mathrm{d}x + \mathrm{O}\left(\frac{1}{n}\right)\right)^2 = \left(\log 2 - \frac{1}{2}\right)^2 + \mathrm{O}\big(n^{-1}\big)$$

and

$$\begin{aligned} \Sigma_{22} &= \frac{1}{n+1}\sum_{n/2<j\le n}\frac{j}{n+1-j}\left(\int_0^{1-j/n}\frac{v}{1-v}\,\mathrm{d}v + \mathrm{O}\left(\frac{1}{n}\right)\right) \\ &= \int_{1/2}^{1}\frac{u}{1-u}\big(-\log u - (1-u)\big)\,\mathrm{d}u + \mathrm{O}\left(\frac{\log n}{n}\right) \\ &= I + \mathrm{O}\big(n^{-1}\log n\big). \end{aligned}$$

Inserting the obtained values of Σ_{ij}, $i,j \in \{1,2\}$, into formulas for Σ_i, $i \in \{1,2\}$, and the latter into (16), we have

$$\sum_{0\le r\le n-1}\mu_r^2 = 2\left(\log 2 - \frac{1}{2}\right)^2 + 4I + \mathrm{O}\big(n^{-1}\log n\big).$$

This and (17) yields

$$\sum_{0\le r\le n-1}\mu_r^2 = \frac{\pi^2}{3} - 3 + \mathrm{O}\big(n^{-1}\log n\big).$$

Now, equality (13) follows from the earlier found sum (15).

As we have noted after Lemma 7, the possible minimum, except μ_1, could be among μ_r, if $r_n < r \le n-1$. However, using the inequality (12), we have the estimate

$$\sum_{r_n<r\le n-1} \mu_r^2 = \sum_{r>r_n} \lambda_r^2 + \mathrm{O}\biggl(\sum_{0\le r\le r_n} |\lambda_r - \mu_r|\biggr) + \mathrm{O}\biggl(\frac{\log n}{n}\biggr)$$
$$\le \int_{r_n}^{\infty} \frac{\mathrm{d}u}{(u+1)^2(u+2)^2} + \mathrm{o}(1) = \mathrm{o}(1),$$

showing that $\max_{r_n<r\le n-1} |\mu_r| = \mathrm{o}(1)$ as $n \to \infty$. Hence by (12), the minimal eigenvalue is μ_1 provided that n is sufficiently large.

The lemma is proved. □

Proof of Theorem. By virtue of the upper estimate (11) and (14) we see that $\tau_{n,2} \le 1 - 2\mu_1 + \mathrm{O}(n^{-1})$. Lemma 7 now yields $\tau_{n,2} \le 4/3 + \mathrm{O}(n^{-1})$. Recalling the lower estimate (10) we complete the proof.

The theorem is proved. □

Acknowledgement

The authors thank the referees for the benevolent remarks which have helped to improve the exposition of the paper.

References

1. R. Arratia, A.D. Barbour, S. Tavaré, *Logarithmic Combinatorial Structures: A Probabilistic Approach*, EMS Monographs in Mathematics, EMS Publishing House, Zürich, 2003.
2. W.J. Ewens, The sampling theory of selectively neutral alleles, *Theor. Popul. Biol.*, **3**, pp. 87–112, 1972.
3. J.F.C. Kingman, *Mathematics of Genetic Diversity*, SIAM, Philadelphia, PA, 1980.
4. N.S. Johnson, S. Kotz, N. Balakrishnan, *Discrete Multivariate Distributions*, Wiley, New York, 1997.
5. W.C. Lynch, More combinatorial properties of certain trees, *Comput. J.*, **7**, pp. 299–302, 1965.
6. K. Wieand, Eigenvalue distributions of random permutation matrices, *Ann. Probab.*, **28**, pp. 1563–1587, 2000.
7. E. Manstavičius, Conditional probabilities in combinatorics. The cost of dependence, in: *Prague Stochastics 2006: 7th Prague Symposium on Asymptotic Statistics and 15th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, J. Huškov'a, M. Janžura (Eds.), Prague, Charles University, Matfyzpress, 2006, pp. 523–532.

8. E. Manstavičius, Summability of additive functions on random permutations, in: *Proceedings of the 4th International Conference "Analytic and Probabilistic Methods in Number Theory"*, A. Laurinčikas, E. Manstavičius (Eds.), TEV, Vilnius, 2007, pp. 99–108.

9. E. Manstavičius, Moments of additive functions defined on the symmetric group, *Acta Appl. Math.*, **97**, pp. 119–127, 2007.

10. G.J. Babu, E. Manstavičius, Limit processes with independent increments for the Ewens sampling formula, *Ann. Inst. Statist. Math.*, **54**(3), pp. 607–620, 2002.

11. E. Manstavičius, An analytic method in probabilistic combinatorics, *Osaka J. Math.*, **46**, pp. 273–290, 2009.

12. E. Manstavičius, A limit theorem for additive functions defined on the symmetric group, *Lith. Math. J.*, **51**, pp. 211–237, 2011.

13. J. Kubilius, On the estimate of the second central moment for arbitrary additive arithmetic functions, *Liet. Mat. Rink.*, **23**, pp. 110–117, 1983 (in Russian).

14. J. Kubilius, Improved estimate of the second central moment for additive arithmetic functions, *Liet. Mat. Rink.*, **25**, pp. 104–110, 1985 (in Russian).

15. G.A. Watterson, The sampling theory of selectively neutral alleles, *Adv. Appl. Probab.*, **6**, pp. 463–488, 1974.

16. T.S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach Scientific Publishers, New York, 1978.

17. Y.L. Geronimus, *Theory of Orthogonal Polynomials*, Gos. Izd. Techn.-Theor. Liter., Moscow, 1950 (in Russian).

18. L. Kuipers, H. Niederreiter, *Uniform Distribution of Sequences*, Wiley & Sons, New York, 1974.

19. P.D.T.A. Elliott, *Arithmetic Functions and Integer Products*, Springer-Verlag New York Inc., New York, 1985.

20. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, M. Abramowitz, I.A. Stegun (Eds.), USA Department of Commerce, National Bureau of Standards, 1972.