Effective bounds of the variance of statistics on multisets of necklaces

Arvydas Karbonskis*, Eugenijus Manstavičius*

Institute of Mathematics, Vilnius University
Naugarduko 24, LT-03225 Vilnius
E-mail: arvydas.karbonskis@mif.stud.vu.lt

Received October 28, 2020; published online February 18, 2021

Abstract. The variance of a linear statistics on multisets of necklaces is explored. The upper and lower bounds with optimal constants are obtained.

Keywords: Turán–Kubilius inequality; polynomials over finite field; additive function

AMS Subject Classification: 60C05

1 Introduction and results

Let \((\mathcal{P}, \| \cdot \|)\) be an initial set of weighted objects and

\[
\pi(j) := |\{p \in \mathcal{P} : \|p\| = j\}| < \infty
\]

for every \(j = 1, 2, \ldots\). Examine the set \(\mathcal{G}\) with the extended weight function \(\| \cdot \|\) of multisets comprised of \(p \in \mathcal{P}\). Namely, \(a \in \mathcal{G}\) if \(a = \{p_1, \ldots, p_r\}\) and \(\|a\| = \|p_1\| + \cdots + \|p_r\|\) including the empty multiset \(\emptyset\) of weight 0. Then

\[
m(n) := |\mathcal{G}_n| := |\{a \in \mathcal{G} : \|a\| = n\}| = \sum_{\ell(\bar{k}) = n} \prod_{j=1}^{n} \left( \pi(j) + k_j - 1 \right),
\]

where \(\ell(\bar{k}) = k_1 + \cdots + nk_n\) if \(\bar{k} = (k_1, \ldots, k_n) \in \mathbb{N}_0^n\) and \(n \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}\).

In the present paper, we deal with the multisets for which \(m(n) = q^n\), where \(q \geq 2\) is an arbitrary natural number. If \(q\) is a prime power, then \(\mathcal{G}\) may be interpreted as \(\mathbb{F}_q^*[t]\), the set of monic polynomials over a finite field \(\mathbb{F}_q\). Then \(\mathcal{P}\) is the subset of irreducible polynomials. For an arbitrary such \(q\), there exist combinatorial constructions,
called multisets of necklaces satisfying \( m(n) = q^n \) (see, [1, Example 2.12, p. 43]). For multisets, we have the following relations

\[
\pi(n) = \frac{1}{n} \sum_{d|n} q^{n/d} \mu(d), \quad q^n = \sum_{d|n} d\pi(d),
\]

where in the summations, \( d \) runs over natural divisors of \( n \) and \( \mu(d) \) stands for the Möbius function. The equalities are equivalent to the formal power series relation

\[
\sum_{n=0}^{\infty} q^n x^n = \frac{1}{1 - qx} = \prod_{j=1}^{\infty} (1 - x^j)^{-\pi(j)}.
\]

Take an \( a \in \mathcal{G}_n \) uniformly at random, that is, sample it with probability \( \nu_n(\{a\}) = q^{-n}, \ n \in \mathbb{N} \) and \( \nu_0(\{\emptyset\}) = 1 \). If \( k_j(a) \geq 0 \) is the number of elements \( p_i \) in \( a \in \mathcal{G}_n \) of weight \( j \), then \( \bar{k}(a) = (k_1(a), \ldots, k_n(a)) \) is the structure vector of \( a \in \mathcal{G}_n \) satisfying \( \ell(\bar{k}(a)) = n \). Its distribution is

\[
\nu_n(\bar{k}(a) = s) = \mathbf{1}\{\ell(s) = n\} q^{-n} \prod_{j=1}^{n} \left( \frac{1}{2} \left( \pi(j) + s_j - 1 \right) \right),
\]

where \( s = (s_1, \ldots, s_n) \in \mathbb{N}_0^n \) and \( \mathbf{1}\{\cdot\} \) stands for the indicator function.

We are interested in the distribution with respect to \( \nu_n \) of the linear statistics

\[
h(\bar{c}) := h(\bar{c}, a) = c_1 k_1(a) + \cdots + c_n k_n(a), \quad \bar{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n.
\]

The number of components in \( a \) is such a function, namely, it equals \( k_1(a) + \cdots + k_n(a) \). We refer to [1] for more sophisticated examples.

The present paper is devoted to the variance of \( h(\bar{c}) \) which is a sum of dependent random variables (r.v.s) as the relation \( h(\bar{j}, a) = \ell(\bar{k}(a)) = n \) for each \( a \in \mathcal{G}_n \) shows. Estimating it, we propose an approach to overcome technical obstacles stemming from dependence.

In the sequel, the expectations and variances with respect to \( \nu_n \) will be denoted by \( \mathbf{E}_n \) and \( \mathbf{V}_n \) while, when the probability space \( (\Omega, \mathcal{F}, P) \) is not specified, we will respectively use the notation \( \mathbf{E} \) and \( \mathbf{V} \). The summation indexes \( i, j, l, k, m, m_1 \) and \( m_2 \) will be natural numbers.

**Theorem 1** If \( \bar{c} \in \mathbb{R}^n \) and \( n \in \mathbb{N}_0 \), then

\[
\mathbf{V}_n h(\bar{c}) = \sum_{1 \leq j, k \leq n} c_j^2 \pi(j) q^{-j k} - \sum_{i, l, j, k \leq n \atop i + l, j + k > n} c_i c_j \pi(i) \pi(j) q^{-i - j}.
\]

The sketch of the proof is given at the beginning of Section 2.

It is known [1] that, for a fixed \( j \), the r.v. \( k_j(a) \) converges in distribution to the r.v. \( \gamma_j \) distributed according the negative binomial law \( NB(\pi(j), q^{-j}) \). If \( \{\gamma_1, \gamma_2, \ldots\} \) are mutually independent, define the statistics

\[
Y_n = c_1 \gamma_1 + \cdots + n \gamma_n.
\]

We shall see that the first sum on the right-hand side in (4) is close to \( \mathbf{V} Y_n \); therefore, estimating \( \mathbf{V}_n h(\bar{c}) \), we use the following quadratic forms:

\[
B_n(\bar{c}) := \sum_{1 \leq j, k \leq n} c_j^2 \pi(j) q^{-j k}, \quad R_n(\bar{c}) = \sum_{m \leq n} m q^{-2m} \left( \sum_{j|m} c_j \pi(j) \right)^2.
\]
Theorem 2 If \( n \geq 2 \), then
\[
V_n h(\bar{c}) \leq B_n(\bar{c}) + \frac{1}{2} R_n(\bar{c}).
\]
The inequality becomes an equality for
\[
c_j = c_j^* := \frac{3}{\pi(j)} \sum_{d | j} dq^d \mu \left( \frac{j}{d} \right) - (2n + 1)j, \quad 1 \leq j \leq n.
\]

Corollary 1 If \( n \geq 2 \) and \( \bar{c} \neq 0 \), then
\[
V_n h(\bar{c}) < \frac{3}{2} B_n(\bar{c}) < \left( \frac{3}{2} - \frac{q - 1}{q} nq^{-n} \right) V_n.
\]

The proofs of the last two theorems presented in Section 2 are built upon the ideas and auxiliary results obtained in [4], [2] and [5].

2 Proofs

We firstly recall known facts about random multisets which can be found in [3] and [1, Section 2.3]. Let \( \gamma(x) = (\gamma_1(x), \gamma_2(x), \ldots) \) be the infinite dimensional vector of independent r.v.s having the negative binomial distributions \( NB(\pi(j), x^j) \), namely,
\[
P(\gamma_j = m) = \left( \frac{\pi(j)}{m} + m - 1 \right) (1 - x^j)^{\pi(j)} x^{jm}, \quad m = 0, 1, \ldots
\]
where \( 0 < x \leq q^{-1} \). Then \( \gamma_j^{(q^{-1})} = \gamma_j \) which has been introduced in Introduction. For convenience, we extend \( k(a) \) to \( k(a) := (k_1(a), \ldots, k_n(a), 0, \ldots) \) and use infinite dimensional vectors. Set \( \theta(x) = 1 \gamma_1(x) + \cdots + n \gamma_n(x) + (n + 1) \gamma_{n+1}(x) + \cdots \) The latter r.v. is well defined if \( 0 < x < q^{-1} \), since the condition of the Borel–Cantelli lemma is satisfied:
\[
\sum_{j=1}^{\infty} P(\gamma_j^{(x)} \neq 0) = \sum_{j=1}^{\infty} (1 - (1 - x^j)^{\pi(j)}) < \infty.
\]
Lemma 1 If \( \bar{s} = (s_1, \ldots, s_j, s_{j+1}, \ldots) \in \mathbb{N}_0^\infty \) and \( 0 < x < q^{-1} \), then
\[
\nu_n(\bar{k}(a) = \bar{s}) = P(\bar{\gamma}^{(x)} = \bar{s} | \theta^{(x)} = n).
\]

Proof. Actually, this is Lemma 2.2 in [3] stated there for \( \mathbb{F}_q[t] \). The details remain the same in the more general case. \( \square \)

Lemma 2 For a functional \( \Psi : \mathbb{N}_0^\infty \to \mathbb{R} \) such that \( \mathbb{E}|\Psi(\bar{\gamma}^{(x)})| < \infty \), we have
\[
\mathbb{E}\Psi(\bar{\gamma}^{(x)}) = (1 - qx) \left( \Psi(\bar{0}) + \sum_{n=1}^\infty \mathbb{E}_n \Psi(\bar{k}(a)) q^n x^n \right), \quad 0 < x < q^{-1}.
\]

Proof. Apply Lemma 1 in the double averaging as follows:
\[
\mathbb{E}\Psi(\bar{\gamma}^{(x)}) = \sum_{n=0}^\infty \mathbb{E}\left(\Psi(\bar{\gamma}^{(x)}) | \theta^{(x)} = n\right) P(\theta^{(x)} = n)
\]
\[
= \sum_{n=0}^\infty \mathbb{E}_n \Psi(\bar{k}(a)) P(\theta^{(x)} = n). \quad \square
\]

Proof of Theorem 1. It is straightforward. Applying the last lemma for the relevant \( \Psi \), one can easily find the needed mixed moments of \( k_j(a), 1 \leq j \leq n \), and further, the variance of the linear combination \( h(a) \). \( \square \)

To prove Theorems 2 and 3, we will apply the following lemmas concerning particular matrices and quadratic forms.

Lemma 3 Let \( U = ((u_{ij})), i, j \leq n, \) be the symmetric matrix with the entries
\[
u_{ij} = 1\{i + j > n\}(ij)^{-1/2}.
\]

The spectrum of \( U \) is the set \( \{1, -1/2, 1/3, \ldots, (-1)^{n-1}/n\} \). The eigenvectors corresponding to the first three eigenvalues are proportional to \( \bar{e}_r = (e_{r1}, \ldots, e_{rn}) \), where \( r = 1, 2, 3 \) and, for \( j \leq n \),
\[
e_{1j} = \sqrt{j}, \quad e_{2j} = (3j - 2n - 1)\sqrt{j}, \quad e_{3j} = (10j^2 - 6(2n + 1)j + 3n^2 + 3n + 2)\sqrt{j}.
\]

Proof. This is the byproduct of works [4] and [2]. \( \square \)

Afterwards, let \( \bar{e}_r, 1 \leq r \leq n \), be the orthogonal basis of \( \mathbb{R}^n \) comprised of the eigenvectors of \( U \) and \( \bar{x}' \) means the transposed vector \( \bar{x} \).

Lemma 4 If \( b_m \in \mathbb{R} \) and \( 1 \leq m \leq n \) and \( n \geq 2 \), then
\[
-\frac{1}{2} \sum_{1 \leq m \leq n} mb_m^2 \leq \sum_{1 \leq m_1, m_2 \leq n} b_{m_1} b_{m_2} \leq \sum_{1 \leq m \leq n} mb_m^2.
\]

If \( n \geq 3 \) and
\[
\sum_{m \leq n} mb_m = 0,
\]

http://www.journals.vu.lt/LMR
then
\[ \sum_{1 \leq m_1, m_2 \leq n \atop m_1 + m_2 > n} b_{m_1} b_{m_2} \leq \frac{1}{3} \sum_{1 \leq m \leq n} m b_m^2. \] (11)

Moreover, each bound in (9) and (11) are achieved, respectively, for \( b_m = e_{rm}/\sqrt{m} \), where \( r = 2, 1, 3 \) and \( e_{rm} \) have been defined in Lemma 3.

Proof. Inequalities (9) are seen from Lemma 3 after the substitution \( b_m = x_m/\sqrt{m} \), \( m \leq n \), since the extreme eigenvalues are 1 and \(-1/2\).

After the same substitution, we further examine the quadratic form with the matrix \( U \). Condition (10) reckons the subspace of vectors \( \bar{x} = (x_1, \ldots, x_n) \) satisfying \( x_1 + \cdots + x_j \sqrt{j} + \cdots + x_n \sqrt{n} = \bar{x} \cdot \bar{e}_1' = 0 \). This subspace is spanned over the first eigenvector. In other words, under (10), only the form values obtained in the subspace \( L \subset \mathbb{R}^n \) spanned over the vectors \( \bar{e}_2, \ldots, \bar{e}_n \) count. Hence
\[ \max_{\bar{x} \in L} \|\bar{x}\|^{-2} \bar{x}U \bar{x}' \leq \max_{2 \leq r \leq n} (1-r^{-1})^{r-1} / r = 1/3. \]

Returning to \( b_m \), from this we obtain inequality (11). \( \Box \)

Proof of Theorem 2. After grouping the summands, expression (4) can be rewritten as follows:
\[ V_n h(\bar{c}) = B_n(\bar{c}) - \sum_{m_1, m_2 \leq n \atop m_1 + m_2 > n} \left( q^{-m_1} \sum_{i|m_1} c_i \pi(i) \right) \left( q^{-m_2} \sum_{j|m_2} c_j \pi(j) \right). \]

Now evidently estimate (5) follows from Lemma 4 with
\[ b_m = q^{-m} \sum_{j|m} c_j \pi(j), \quad m \leq n. \]

Moreover, it becomes an equality if we take \( c_j = c_j^* \) satisfying
\[ q^{-m} \sum_{j|m} c_j^* \pi(j) = 3m - 2n - 1, \]
which by the Möbius inversion formula and (1) may be rewritten as (6). \( \Box \)

To prove the first assertion of Corollary 1, it suffices to estimate the inner sum in \( R_n(\bar{c}) \), namely,
\[ \left( \sum_{j|m} c_j \pi(j) \right)^2 \leq \sum_{j|m} c_j^2 \pi(j)/j \sum_{j|m} j \pi(j) = \sum_{j|m} c_j^2 \pi(j)/j \cdot q^m. \]

Further, using the expression of \( VY_n \), we just estimate the remainder:
\[ VY_n - B_n(\bar{c}) = \sum_{j \leq n} c_j^2 \pi(j) \sum_{k>n/j} q^{-jk} \geq nq^{-n} \sum_{j \leq n} c_j^2 \pi(j)/j \cdot q^j \frac{q^j - 1}{q^j} \cdot q \]
\[ \geq nq^{-n-1} (q - 1) VY_n. \]
Plugging both estimates into (5), we obtain the first inequality in Corollary 1 with \( \leq \) instead of \(<\). In fact, we obtained the strict inequality since Cauchy’s inequality applied in the last step is strict if \( \bar{c} \) is not proportional to \( \bar{j} \), and in this exceptional case, \( V h(\bar{c}) = 0 \).

**Proof of Theorem 3.** Observe that \( V_n h(\bar{c}) = V_n (h(\bar{c}) - t n) = V_n h(\bar{c} - t \bar{j}) \) for every \( t \in \mathbb{R} \). Hence the right-hand inequality follows from (5) applied for the shifted statistics.

To get the lower bound of variance, we combine (4) and (11). We start with

\[
V_n h(\bar{c} - t \bar{j}) = B_n (\bar{c} - t \bar{j}) - \sum_{m_1, m_2 \leq n} \tilde{b}_{m_1} \tilde{b}_{m_2},
\]

where

\[
\tilde{b}_m = q^{-m} \sum_{j \mid m} (c_j - t c_j) \pi(j)
\]

and \( m \leq n \). By the definition of \( t_c \) the latter sequence satisfies condition (10). Hence by (11),

\[
\sum_{m_1, m_2 \leq n} \tilde{b}_{m_1} \tilde{b}_{m_2} \leq \frac{1}{3} \sum_{m \leq n} m b_m^2 = \frac{1}{3} R_n (\bar{c} - t \bar{j}).
\]

This and (4) imply the lower bound. Moreover, the latter is sharp since Lemma 4 assures this by a choice of a particular sequence \( b_m, m \leq n \). \( \square \)

**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] Ž. Baronėnas, E. Manstavičius, P. Šapokaitė. A sharp inequality for the variance with respect to the ewens sampling formula. 2019, arXiv:2003.05975v1.

[3] J.C. Hansen. Factorization in \( f_q[x] \) and brownian motion. *Combin. Probab. Comput.*, **2**:285–299, 1993.

[4] J. Klimavičius, E. Manstavičius. The Turán–Kubilius inequality on permutations. *Annales Univ. Sci. Budapest., Sect. Comp.*, **48**:45–51, 2018.

[5] E. Manstavičius. Sharp bounds for the variance of linear statistics on random permutations. *Random Struct. Alg.*, 2020. https://doi.org/10.1002/rsa.20951.