Common Intervals in Permutations

Sylvie Corteel* Guy Louchard † Robin Pemantle‡

January 28, 2004

Abstract

An *interval* of a permutation is a consecutive substring consisting of consecutive symbols. For example, 4536 is an interval in the permutation 71453682. These arise in genetic applications. For the applications, it makes sense to generalise so as to allow gaps of bounded size $\delta - 1$, both in the locations and the symbols. For example, 4527 has gaps bounded by 1 (since 3 and 6 are missing) and is therefore a δ-interval of ****4*5*27**** for $\delta = 2$.

After analysing the distribution of the number of intervals of a uniform random permutation, we study the number of 2-intervals. This is exponentially large, but tightly clustered around its mean. Perhaps surprisingly, the quenched and annealed means are the same. Our analysis is via a multivariate generating function enumerating pairs of potential 2-intervals by size and intersection size.

1 Introduction

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. We are interested in counting the common intervals of a pair of permutations. To be precise, if G_A and G_B are two permutations of $[n]$, we are interested in counting the pairs of intervals (I, J) for which $G_A(I) = G_B(J)$. It is equivalent to count intervals I for which $G_B^{-1} G_A(I)$ is also an interval. Accordingly, we define

Definition 1.1 *The interval $I := [i, i + k - 1] \subseteq [n]$ is called an* interval of the permutation *G if $G^{-1}(I)$ is an interval, that is, if there is a j such that*

$$G[j, j + k - 1] = [i, i + k - 1] \, .$$

The proper intervals are those whose lengths are at least 2 and at most $n - 1$.

Here and throughout, we use vector notation for permutations rather than cycle notation, so that $(\sigma_1, \ldots, \sigma_n)$ denotes the permutation $i \mapsto \sigma_i$ rather than the permutation consisting of a single n-cycle.

Example: Let G be the permutation $(3, 1, 2, 4, 5)$. Then the proper intervals of G are $[1, 2]$, $[4, 5]$, $[1, 3]$ and $[1, 4]$.

*CNRS PRiSM, Université de Versailles Saint-Quentin, 45 Avenue des Etats-Unis, 78035 Versailles France, email: syl@prism.uvsq.fr

†Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email:louchard@ulb.ac.be

‡Department of Mathematics, University of Pennsylvania, 209 S. 33rd Street, Philadelphia, PA 19104 USA, email:pemantle@math.upenn.edu

When G is a random variable, uniformly distributed over all permutations of $[n]$, let X_k denote the number of of intervals of length k of G and let $X = \sum_k X_k$ denote the number of intervals of G. Uno et al [18] compute $\mathbb{E}X_k$; the following easy proposition is proved at the end of this section.

Proposition 1.2 *As $n \to \infty$, the distribution of X converges to a Poisson with mean 2.*

The number of intervals, or runs of a permutation, was studied in the forties by Kaplansky [12] and Wolfowitz [19, 20] from a statistical point of view. See also [14]. Recently several algorithms were designed to efficiently enumerate all common intervals of permutations [10, 18] and their time complexity is $O(n + K)$ where n is the size of the permutation and K the number of intervals. These algorithms were designed because common intervals have several applications. They relate to the consecutive arrangement problem [7]. Genetic algorithms for sequencing problems are based on common intervals [13, 15]. In bioinformatics [4, 5, 9, 10, 11], genomes of prokaryotes can be modelled as a permutation of genes. A common interval is then a set of orthologous genes that appear consecutively, possibly in different orders, in two genomes. Therefore common intervals can be used to detect groups of genes that are functionally associated [10, 11]. As the annotation of genomes is not perfect, the notion of consecutivity in intervals needs to be relaxed. A notion of *gene teams* was defined in [6], where a gene team is a maximal set of orthologous genes, possibly occurring in different orders in the two species, but separated in each case by gaps that do not exceed a fixed threshold, δ. To study these, we consider a generalisation of intervals, namely δ-intervals (the previous case corresponds to $\delta = 1$).

Definition 1.3 *The set $I \subseteq [n]$ is called a* δ-interval of $[n]$ of length k *if I is a set of integers $\{i_1, \ldots, i_k\}$ with $1 \le i_{r+1} - i_r \le \delta$ for each $1 \le r \le k-1$. We call I a* δ-interval of length k of G *if both I and $G^{-1}(I)$ are δ-intervals. Proper δ-intervals are again those of cardinality at least 2 and at most $n-1$.*

Example: $G = (3, 1, 2, 4, 5)$ possesses the $2-$intervals:

$$\{1,2\},\{1,3\},\{1,2,3\},\{2,3\},\{1,2,3,4\},\{1,3,4\},\{2,3,4\},$$
$$\{2,4,5\},\{2,4\},\{2,3,5\},\{2,3,4,5\},\{1,3,4,5\},\{1,2,4,5\},\{1,2,3,5\}$$

In [6] a polynomial time enumeration algorithm for gene teams is presented. Our notion of δ-intervals removes the maximality constraint, whence the number of these may grow exponentially and it is natural to enumerate asymptotically rather than enumerating exactly.

The main purpose of this note is to investigate the asymptotic properties of $X_k^{(\delta)}$, where this denotes the number of $\delta-$intervals of length k of a uniformly chosen random permutation of $[n]$, and of the total number $X^{(\delta)} := \sum_k X_k^{(\delta)}$ of δ-intervals of a random permutation. We are interested in all $\delta > 1$ but in the present manuscript we examine only the case $\delta = 2$. To reduce the number of superscripts, we let Y and Y_k denote $X^{(2)}$ and $X_k^{(2)}$ respectively.

The number $X^{(\delta)}$ of δ-intervals when $\delta > 1$ behaves very differently from X. Whereas X is $\mathcal{O}(1)$ as $n \to \infty$, with all the contributions coming from short intervals, there will typically be many δ-intervals. In fact a thumbnail computation produces numbers α_k in the unit interval ($\alpha_2 \approx 0.57939$) such that for $k \sim \alpha n$ and $\alpha > \alpha_k$, the random variable X_k will be typically exponentially large: the number of δ-intervals of $[n]$ of size k grows exponentially, the probability of G^{-1} of one of these also being a δ-interval decays exponentially, and the growth overcomes the decay when $\alpha > \alpha_k$.

Seeing that $X^{(\delta)}$ grows exponentially in n, it is natural to look at the rescaled quantity $n^{-1} \log X^{(\delta)}$. In the next section we compute the *annealed* mean, $n^{-1} \log \mathbb{E}X^{(\delta)}$. The term "annealed" means that we first take an expectation over the (uniform) measure on permutations. The more interesting quantities are the *quenched* quantities, which refer to the typical, rather than the mean behaviour of $X^{(\delta)}$. Often one has a so-called lottery effect, meaning that the mean of a quantity X comes primarily from an exponentially small number of values that are exponentially larger than the median value, and that consequently, $\mathbb{E} \log X < \log \mathbb{E} X$. For example, when there is a Gaussian limit law, $n^{-1/2}(\log X - n\mu) \to \mathcal{N}(0, \sigma^2)$, then one will typically have a lottery effect. Perhaps surprisingly in light of the discussion in section 4, there is no lottery effect. Our main result, Theorem 4.1 below, is that for $\delta = 2$, we have $\mathbb{E}(X^{(\delta)})^2 = \mathcal{O}(\mathbb{E}X^{(\delta)})^2$. This shows that as $n \to \infty$, the sequence $\overline{X} := X^{(\delta)}/\mathbb{E}X^{(\delta)}$ is tight. The mean of $X^{(\delta)}$ is computed in the next section, with the remaining sections devoted to the computation of the second moment. We start in Section 2 with arguments in the case $\delta = 1$.

2 Intervals

Recall that X_k denotes the number of of intervals of length k of G and that $X = \sum_k X_k$ denotes the number of intervals of G. Uno et al [18] computed

$$\mathbb{E}(X_2) = \frac{2(n-1)}{n} \; ; \; \mathbb{E}(X_3) = \frac{6(n-2)}{n(n-1)} \; ; \; \mathbb{E}(X_k) \le \frac{24}{n^2} \text{ for } k \ge 4$$

Although this was not explicitly stated in [18], it is not hard to show Proposition 1.2.

Proof of Proposition 1.2: Letting $X' := \sum_{k=3}^{n-1} X_k$, we see that

$$\mathbb{E}X' \le \frac{6}{n} + (n-4)\frac{24}{n^2} \le \frac{30}{n}$$

so $X' \to 0$ in probability as $n \to \infty$. Thus it suffices to show that X_2 converges to a Poisson of mean 2. Kaplansky proves this in [12]. A modern approach is to use the Poisson approximation machinery first developed by Chen and Stein, and put in an explicit and usable form in [1]. Given $k \in [n-1]$, let A_k be the event that $G^{-1}\{k, k+1\}$ is an interval. Let $B_k = \{k-1, k, k+1\} \cap [n-1]$. Write p_k for $\mathbb{P}(A_k)$ and $p_{k,l}$ for $\mathbb{P}(A_k \cap A_l)$. Theorem 1 of [2] shows that the total variation distance between X and a Poisson with mean $\sum_k p_k$ is bounded by the sum of three quantities, namely

$$\begin{aligned}
b_1 &:= \sum_k \sum_{j \in B_k} p_j p_k \\
b_2 &:= \sum_k \sum_{k \ne j \in B_k} p_{k,l} \\
b_3 &:= \sum_k \mathbb{E} \left| \mathbb{E} \left(\mathbf{1}_{A_k} - p_k | \sigma(A_j : j \notin B_k) \right) \right| .
\end{aligned}$$

Bounding b_1 and b_2 by $\mathcal{O}(1/n)$ is straightforward. The same bound on b_3 follows from the identity

$$\begin{aligned}
\left| \mathbb{E} \left(\mathbf{1}_{A_k} - p_k | \sigma \right) \right| &= 2 \sup_{H \in \sigma} [\mathbb{P}(H \cap A_k) - p_k \mathbb{P}(H)] \\
&= 2 p_k ||\mu_{A_k} - \mu||_{TV} ,
\end{aligned}$$

together with a coupling argument obtaining μ_{A_k} from μ by switching the values of G^{-1} on k and $G(j)$, and on $k+1$ and $G(j+1)$ for a uniformly chosen j. ■

More generally, one may consider the distribution of X_k.

$$\mathbb{E}(X_k) = \frac{k!(n-k+1)}{n(n-1)\dots(n-k+2)} = \frac{k(n-k+1)}{\binom{n}{k-1}}. \tag{1}$$

Set

$$X := \sum_{k=2}^{n-1} X_k.$$

We see that, as $n \to \infty$, the dominant terms of $\mathbb{E}(X)$ are given by $k = \mathcal{O}(1)$ and $k = n - \mathcal{O}(1)$. Indeed, by Stirling, and setting $k = \alpha n + 1$, we easily derive

$$\mathbb{E}\left(\sum_{k=\varepsilon n}^{n/2} X_k\right) \quad \sim \quad \int_{\varepsilon}^{1/2} \left[\alpha^{\alpha n}(1-\alpha)^{(1-\alpha)n} \alpha n(1-\alpha)n\sqrt{2\pi\alpha(1-\alpha)n}\right] n d\alpha \downarrow 0,$$

exponentially , $n \to \infty$, for fixed ε.

We obtain

$$\mathbb{E}(X) \sim 2 + \frac{8}{n} + \frac{36}{n^2} + \frac{228}{n^3} + \dots$$

As an example of X_k behaviour, let us now turn to X_3 and compute $E_3^2 := \mathbb{E}\left(X_3^2\right)$.

Combinatorial analysis leads to asymptotic series expressions for the moments of X_k and for the probabilities $\mathbb{P}(X_k = j)$. For example,

$$\mathbb{E}X_3 \sim \frac{6}{n} - \frac{6}{n^2} + \cdots \qquad \mathbb{E}X_3^2 \sim \frac{6}{n} + \frac{38}{n^2} + \cdots$$

and

$$\mathbb{P}(X_3 = 0) \sim 1 - \frac{6}{n} + \frac{28}{n^2} + \cdots \qquad \mathbb{P}(X_3 = 1) \sim \frac{6}{n} - \frac{50}{n^2} + \cdots \qquad \mathbb{P}(X_3 = 2) \sim \frac{22}{n^2} + \cdots$$

Details are given in [8]. We turn now to 2-intervals.

3 The mean number of 2-intervals

Let $N(k, n)$ denote the number of 2-intervals that are subsets of $[n]$ and have cardinality k. We will take advantage of the uniformity of G^1. For each of the $N(k, n)$ 2-intervals of cardinality k, its inverse image under G is uniformly distributed on k-subsets of $[n]$. Therefore, the probability is exactly $N(k, n)/\binom{n}{k}$ for any given 2-interval of cardinality k, that its inverse image under G is again a 2-interval. Thus

$$\mathbb{E}Y_k = \frac{N(k, n)^2}{\binom{n}{k}}. \tag{2}$$

To evaluate $N(k, n)$, note that there may be anywhere from 0 to $m_k := \min\{k-1, n-k\}$ "holes" in a 2-interval, where a hole is an element not in the 2-interval but between its endpoints. These may be enumerated by the following procedure. Pick a starting position r with $1 \le r \le n-k-i+1$ and let r be the least element of the 2-interval. Choose any sequence with i occurrences of the word "skip" and $k-1-i$ occurrences of the word "no-skip". If the first word in the sequence is "no-skip" then $r+1$ is the next element of the 2-interval; if the first word is "skip" then $r+2$ is the next element. Continue in this manner until the sequence is used up. This method of enumeration makes it clear that

$$\mathbb{E}Y_k = \frac{N(k,n)^2}{\binom{n}{k}} = \frac{\left(\sum_{i=0}^{m_k}(n+1-k-i)\binom{k-1}{i}\right)^2}{\binom{n}{k}}. \tag{3}$$

The sum may be evaluated exactly for $k/n < 1/2$ and asymptotically for $k/n \geq 1/2$ leading to

$$N(k,n) \sim \begin{cases} \left(n + \frac{3}{2} - \frac{3}{2}k\right) 2^{k-1} & : \text{ if } \frac{k-2n/3}{n^{1/2}} \to -\infty, \\ 2^{k-1}\frac{\sqrt{k}}{2}\Psi_x & : \text{ if } \frac{k-2n/3}{n^{1/2}} \to x \in (-\infty, \infty), \\ \left[\frac{(2k-1-n)^2}{(3k-1-2n)^2}\right]\binom{k-1}{n-k} & : \text{ if } \frac{k-2n/3}{n^{1/2}} \to +\infty, \end{cases} \tag{4}$$

where Ψ_x is the expectation of the positive part of $Z + x$ and Z is a standard normal.

Via equation (3), these asymptotics for $N(k,n)$ lead directly to asymptotics for $\mathbb{E}Y_k$. To obtain asymptotics for $\mathbb{E}Y$, we then sum over k, using a saddle point approximation. The only significant terms are near $k = \alpha_* n$, where α_* will be determined shortly but is evidently greater than $2/3$. We therefore use (4) with $n \to \infty$ and $\alpha := k/n$ to obtain

$$\begin{aligned} \frac{N(k,n)^2}{\binom{n}{k}} &\sim \left[\frac{2k-n}{k}\binom{k}{n-k}\left(\frac{2k-n}{3k-2n}\right)^2\right]^2 \binom{n}{k}^{-1} \\ &\sim A(\alpha)n^{-1/2}\exp\left[nF_1(\alpha)\right] \end{aligned} \tag{5}$$

where

$$\begin{aligned} A(\alpha) &:= \frac{(2\alpha-1)^5}{(3\alpha-2)^4}\sqrt{\frac{1}{2\pi\alpha(1-\alpha)}}; \\ F_1(\alpha) &:= 3\alpha\log\alpha - (1-\alpha)\log(1-\alpha) - 2(2\alpha-1)\log(2\alpha-1). \end{aligned}$$

The maximum of F_1 occurs in $[2/3, 1]$ where $F_1'(\alpha)$ vanishes, which occurs when $\alpha = \alpha_* \approx .7840013296$ solves

$$17\alpha^4 - 33\alpha^3 + 24\alpha^2 - 8\alpha + 1 = 0.$$

A saddle point approximation now gives us

$$\begin{aligned} \mathbb{E}Y = \sum_{k=2}^{n-1}\mathbb{E}Y_k &\sim \sum_{k=3n/4}^{n} A(k/n)\exp\left[nF_1(k/n)\right]n^{-1/2} \\ &\sim A(\alpha_*)\exp(nF_1(\alpha_*))\sqrt{\frac{2\pi}{-F_1''(\alpha_*)}} \\ &= \exp(nF_1(\alpha_*))\sqrt{\frac{(2\alpha_*-1)^{11}(3\alpha_*-2)}{(3-2\alpha_*)^8}} \\ &\approx 2.4253\exp(0.40122467n). \end{aligned}$$

4 Counting pairs of 2-intervals

We will now use the second moment method to show that $Y/\mathbb{E}Y$ is tight as $n \to \infty$ by establishing:

Theorem 4.1 $\mathbb{E}Y^2 = \mathcal{O}(\mathbb{E}Y)^2$.

Just as the mean of Y may be obtained from a saddle point analysis of $\mathbb{E}Y_k$ near $k/n = \alpha_*$, we expect the second moment of Y to be dominated by terms $\mathbb{E}Y_k^2$ with k near some α_{**}. Because we have seen from numerical data that the quenched and annealed behaviour are the same, we expect to find, and do find, that $\alpha_{**} = \alpha_*$.

Again we will take advantage of symmetry. This time, if I and I' are 2-intervals, we will need to know the cardinality of their intersection before we can say the probability that $G^{-1}(I)$ and $G^{-1}(I')$ are both 2-intervals. We therefore define $N(k, k', n, \kappa)$ to be the number of pairs of 2-intervals (I, I') of $[n]$ with $|I| = k, |I'| = k'$ and $|I \cap I'| = \kappa$. For the computation of $\mathbb{E}Y_k^2$ we will want to specialise to the case $k = k'$, so we denote $N(k, n, \kappa) := N(k, k, n, \kappa)$. Our computations will now be analogous to the computation F_1 and its argmax, α_*. Specifically, letting

$$\alpha := \frac{k}{n}\,, \qquad \beta := \frac{k'}{n}\,, \qquad \rho := \frac{\kappa}{n}\,,$$

we will find a rate function $\mathsf{rate}(\alpha, \beta, \rho)$ such that

$$N(k, k', n, \kappa) \sim A_0(\alpha, \beta, \rho) n^{-3/2} \exp(n\mathsf{rate}(\alpha, \beta, \rho)) \tag{6}$$

for all parameter values in a range containing the dominant contributions to the second moment of Y. To obtain the analogue of (5) for second moments, we need the rate function for total number of pairs of subsets A and B of $[n]$ with $|A| = k, |B| = k'$ and $|A \cap B| = \kappa$. The total number is exactly

$$\binom{n}{\kappa, k-\kappa, k'-\kappa, n-k-k'+\kappa} = \frac{n!}{\kappa!(k-\kappa)!(k'-\kappa)!(n-k-k'+\kappa)!}\,,$$

whence for any pair of sets of respective cardinalities k and k' whose intersection has cardinality $\kappa \le k$, the probability that their union is a specific pair of sets is the reciprocal, P, of this. Let ent denote the exponential rate $\lim n^{-1} \log P$ for this probability:

$$\mathsf{ent} := \rho \log \rho + (\alpha - \rho) \log(\alpha - \rho) + (\beta - \rho) \log(\beta - \rho) + (1 - \alpha - \beta + \rho) \log(1 - \alpha - \beta + \rho)\,. \tag{7}$$

We may now evaluate the rate function

$$F_2(\alpha, \beta, \rho) := \lim_{n\to\infty, k/n\to\alpha, k'/n\to\beta, \kappa/n\to\rho} N(k, k', n, \kappa)^2 P(k, k', n, \kappa)$$

for the expected number of pairs of 2-intervals of respective sizes k and k' with overlap κ:

$$F_2 := 2 \cdot \mathsf{rate} + \mathsf{ent}\,. \tag{8}$$

Since $\mathbb{E}Y^2 = \sum_{k,k'} \mathbb{E}Y_k Y_{k'} = \sum_{k,k',\kappa} N(k, k', n, \kappa)^2 P(k, k', n, \kappa)$ is the sum of polynomially many summands, it follows that the exponential order of $\mathbb{E}Y^2$ is the same as the order of the largest summand, namely

$$n^{-1} \log \mathbb{E}Y^2 \to \sup_{\alpha,\beta,\rho} F_2(\alpha, \beta, \rho) := \lambda^{**}\,. \tag{9}$$

Furthermore, from the inequality $\mathbb{E}Y_kY_{k'} \le \frac{1}{2}(\mathbb{E}Y_k^2 + \mathbb{E}Y_{k'}^2)$, we see that the maximum of $\mathbb{E}Y_kY_{k'}$ (for fixed n) can only occur when $k' = k$, and therefore

$$\lambda^{**} = \sup_{\alpha,\rho} F_2(\alpha, \alpha, \rho)\,. \tag{10}$$

Next, consider how $N(k, k', n, \kappa)$ varies with κ for fixed (k, k', n). In other words, enumerate pairs of 2-intervals of fixed sizes k and k' according to the size of their intersection, κ. Observe that $\sum_\kappa N(k, k', n, \kappa)$ counts all pairs of 2-intervals of sizes k and k', so that

$$\sum_\kappa N(k, k', n, \kappa) = N(k, n)N(\kappa, n)\,.$$

Since the number of summands is linear in n, we have at the exponential level that

$$\sup_\rho \mathsf{rate}(\alpha, \beta, \rho) = u(\alpha)u(\beta) \tag{11}$$

where $u(\alpha) = \lim n^{-1}N(k, n)$ as $n \to \infty$ with $k/n \to \alpha$. We will show:

Lemma 4.2 *For α and β in a neighbourhood of α_{**}, this supremum occurs at $\rho = \alpha \cdot \beta$.*

Observe that $-\mathsf{ent}(\alpha, \beta, \cdot)$ has a maximum at $\alpha \cdot \beta$ as well, where it is given by the expression

$$\mathsf{ent}(\alpha, \beta, \alpha \cdot \beta) = h(\alpha) + h(\beta)\,. \tag{12}$$

Here, $h(x) = x \log(x) + (1 - x) \log(1 - x)$ is the usual entropy function. Both ent and ent are smooth, so we see that the function $F_2 = 2 \cdot \mathsf{rate} + \mathsf{ent}$, viewed as a function of the third argument, has a critical point at $\rho = \alpha \cdot \beta$ as well. The next lemma, requiring a four-variable generating function and considerable computation, is:

Lemma 4.3 *The critical point $(\alpha, \beta, \alpha \cdot \beta)$ is in fact a maximum for F_2.*

Corollary 4.4

$$\lambda_{**} = 2F_1(\alpha_*)\,.$$

Proof. It follows from (11) and (12) that

$$\begin{aligned}
\sup_\rho F_2(\alpha, \beta, \rho) &= F_2(\alpha, \beta, \alpha \cdot \beta) \\
&= 2 \cdot \mathsf{rate}(\alpha, \beta, \alpha \cdot \beta) + \mathsf{ent}(\alpha, \beta, \alpha \cdot \beta) \\
&= 2u(\alpha) + 2u(\beta) + h(\alpha) + h(\beta) \\
&= F_1(\alpha) + F_1(\beta)\,.
\end{aligned}$$

Taking the maximum over α and β then gives the result. ■

Theorem 4.1 now follows directly from Corollary 4.4 and:

Lemma 4.5 *1. There is a neighbourhood of $(\alpha_*, \alpha_*, \rho)$ for which*

$$N(k, k', n, \kappa)^2 P(k, k', n, \kappa) \sim A(\alpha, \beta, \rho)n^{-3/2} \exp(nF_2(\alpha, \beta, \rho))$$

uniformly as $n \to \infty$ with $(k/n, k'/n, \kappa/n) \to (\alpha, \beta, \rho)$.

2. The Hessian of F_2 here is non-degenerate, whence

$$\sum_{k,k',\kappa} N(k,k',n,\kappa)^2 P(k,k',n,\kappa) \sim \frac{A(\alpha_*,\alpha_*,\alpha_*^2)}{\sqrt{(2\pi)^3\mathcal{H}}} \exp(2nF_1(\alpha_*))\,.$$

We will prove Lemmas 4.2, 4.3 and 4.5 in the next section by means of the aforementioned four-variable generating function.

5 The generating function

Recall that $N(k,k',\kappa,n)$ denotes the number of pairs of 2-intervals (I,I') of $[n]$ with $|I| = k, |I'| = k'$ and $|I \cap I'| = \kappa$. Define the generating function

$$F(u_1,u_2,t,z) := \sum_{k,k',\kappa,n} N(k,k',\kappa,n) u_1^k u_2^{k'} t^\kappa z^n$$

so that $N(k,\kappa,n) = [u_1^k u_2^k t^\kappa z^n]F$.

Given a pair of subsets of $[n]$, the positions 1 through n may be divided into three parts:

1. An initial sequence of positions before the first common position;
2. A common position followed by zero or more segments of the form: **a sequence of positions not common to either set, in such a way that no two positions in a row are absent from either set, followed by a common position**;
3. A final sequence of positions after the last common position.

We give generating functions for the three parts separately. To enumerate the second of the three parts, note that each segment between common positions can be one of six possible classes of configuration. We list these here, along with the factor contributed by such a step to the generating function.

- *a* Empty. $f_a = 1$.
- *b* A single position, which can belong to either set or neither, but not both. $f_b = z(1 + u_1 + u_2)$.
- *c* A positive number of pairs $(j, j+1)$ where $j \in I \setminus I'$ and $j+1 \in I' \setminus I$. $f_c = z^2u_1u_2/(1 - z^2u_1u_2)$.
- *c'* A positive number of pairs $(j, j+1)$ where $j \in I' \setminus I$ and $j+1 \in I \setminus I'$. $f_{c'} = f_c$.
- *d* the same as (c) but there is a single position in $I \setminus I'$ at the end. $f_d = zu_1f_c$.
- *d'* the same as (c') but there is a single position in $I' \setminus I$ at the end. $f_{d'} = zu_2f_{c'}$.

The generating function for an arbitrary sequence of these is

$$f = \frac{zu_1u_2t}{1 - zu_1u_2t(f_a + f_b + (2 + z(u_1+u_2)f_c)}\,. \tag{13}$$

To enumerate the first (or last) of the three parts, we first write down the generating function

$$F_6(u,z) := \frac{uz}{1-(1+z)uz}$$

for that part of a 2-interval between its first and last point. By symmetry, parts 1 and 3 have the same generating function, which is equal to $1/(1-z)$ times the generating function g for the segment to the right of the last common position but to the left of the last position in $I \cup I'$. We may write g as the sum of several cases.

e Empty. $g_e = 1$.

f A position in neither set, followed by a non-empty string of positions, each of which is in neither set or I, with no two in a row not in I. $g_f = zF_6(u_1, z)$.

f' A position in neither set, followed by a non-empty string of positions, each of which is in neither set or I', with no two in a row not in I'. $g_{f'} = zF_6(u_2, z)$.

g A string of pairs as in case (c) above, followed by a nonempty sequence as in case (f). $g_g = F_6(u_1, z)/(1 - z^2 u_1 u_2)$.

g' A string of pairs as in case (c') above, followed by a nonempty sequence as in case (f'). $g_{g'} = F_6(u_2, z)/(1 - z^2 u_1 u_2)$.

h The same as (g) except with a position in $I' \setminus I$ in the beginning. $g_h = zu_2 F_6(u_1, z)/(1 - z^2 u_1 u_2)$.

h' The same as (g') except with a position in $I \setminus I'$ in the beginning. $g_{h'} = zu_1 F_6(u_2, z)/(1 - z^2 u_1 u_2)$.

Summing, we see that the factor from the first and last parts is

$$g = \frac{1}{1-z}\left(1 + zF_6(u_1, z) + zF_6(u_2, z) + \frac{F_6(u_1, z) + F_6(u_1, z)}{1 - z^2 u_1 u_2} + \frac{u_2 z F_6(u_1, z) + u_1 z F_6(u_2, z)}{1 - z^2 u_1 u_2}\right),$$

and finally,

$$F(u_1, u_2, t, z) = fg^2 . \tag{14}$$

6 Proofs of lemmas

There is only room for sketches here. Details are given in [8].

Let V be the variety where the rational function F blows up, namely the union of the varieties where the denominators of f and g vanish. It simplifies the computation to change variables to $\tau := ztu_1u_2$, $u = zu_1$ and $v = zu_2$. The $[k, k', n, \kappa]$ coefficient of F now becomes the $[k - \kappa, k' - \kappa, n - k - k' + \kappa, \kappa]$ coefficient of the function $\widetilde{F}(x, y, z, \tau)$. Thus

$$\frac{1}{n}\log[n\alpha, n\beta, n, n\rho]F = \frac{1}{n}\log[n\mu, n\nu, n\delta, n\rho]\widetilde{F}$$

when $\alpha = \mu + \rho$, $\beta = \nu + \rho, \rho = \alpha + \beta + \delta - 1$ and $\widetilde{F}$ is F under the change of variables. In the new variables,

$$\widetilde{f} = \frac{\tau(1 - xy)}{1 - xy - \tau(1 + x + y + xy + zxy - z)} .$$

We let h denote the denominator of this expression. Then V is the union of the variety V_h where h vanishes, together with the varieties where the functions $g_1 := 1 - z$, $g_2 := 1 - xy$, $g_3 := 1 - (1 + z)x$ and $g_4 := 1 - (1 + z)y$ vanish.

Let Dom denote the domain of convergence for the power series for

$$F = \sum_{\mathbf{r}:=(k,k',n,\kappa)} a_{\mathbf{r}} \mathbf{x}^{\mathbf{r}} .$$

Let $\log \mathcal{D}$ denote the logarithmic domain, which is always convex:

$$\log \mathcal{D} := \{\log |x_1|, \ldots, \log |x_4|) : \mathbf{x} \in \mathsf{Dom}\} .$$

Asymptotics of $a_{\mathbf{r}}$ in the direction $\mathbf{r}/|\mathbf{r}| \to \mathbf{s}$ are "controlled" by corresponding points $\mathbf{x}(\mathbf{s}) \in V$ as described in the following paragraph (see [8] for details).

First, it is shown in [16, Theorem 6.3] that for each vector $\mathbf{s} \in (\mathbb{R}^+)^4$, there is a point $\mathbf{x}(\mathbf{s}) \in V$ with positive real coordinates such that $\mathbf{x} \in \partial \mathsf{Dom}$ and such that the normal to the support hyperplane to $\log \mathcal{D}$ at $\log \mathbf{x}$ is parallel to $\mathbf{s}$. It then follows from [16, Theorem 3.5], provided a certain function has a non-degenerate Hessian, that if $\mathbf{x}(\mathbf{s})$ lies in some smooth component of V for all $\mathbf{s}$ in some neighbourhood $\mathcal{N}$, then there is a function A such that the coefficients of

$$F = \sum_{\mathbf{r}:=(k,k',n,\kappa)} a_{\mathbf{r}} \mathbf{x}^{\mathbf{r}}$$

are uniformly asymptotically approximated on $\mathcal{N}$ by

$$a_{\mathbf{r}} \sim A(\frac{r_1}{r_3}, \frac{r_2}{r_3}, \frac{r_4}{r_3}) n^{-3/2} \exp(-\mathbf{r} \cdot \mathbf{x}(\mathbf{r}))\,.$$

The key computational step necessary to make use of these results is to verify that the point $\mathbf{x}(\mathbf{r})$ is in the part V_h of the pole set V that corresponds to the vanishing of h. Combinatorially, the factor f in F enumerates pairs of subsets for which the first and last common positions are 1 and n respectively. The other divisors correspond to the cases where one or both intervals begin at $\Theta(n)$ or end at $n - \Theta(n)$. We now give an argument that the pairs of sets beginning at 1 and ending at n count "nearly all" pairs, meaning that the rate $n^{-1} \log N$ is the same.

Let $\widehat{N}(k, k', \kappa, n)$ of pairs of 2-intervals in which both the subsets contains 1 and n. Suppose, for a given (k, k', κ, n) that

$$\widehat{N}(k, k', \lambda n, \kappa) \geq \widehat{N}(k, k', n, \kappa) \tag{15}$$

for some fixed $\lambda < 1$ and arbitrarily large n (this can happen, for example when $\mu = \alpha - \rho$ is small). Then

$$\begin{aligned} \widehat{N}(k, k', \lambda n, \kappa)^2 P(k, k', \lambda n, \kappa) &\geq \widehat{N}(k, k', n, \kappa)^2 P(k, k', \lambda n, \kappa) \\ &> \widehat{N}(k, k', n, \kappa)^2 P(k, k', n, \kappa) \end{aligned}$$

by an exponential factor. But $\sup_{\alpha,\beta,\rho} F_2(\alpha, \beta, \rho) > 0$, so if this supremum is achieved at $\alpha = k/n, \beta = k'/n, \rho = \kappa/n$, then the inequality would be reversed. Thus, in a neighbourhood of where the supremum of F_2 is achieved, it is not possible for (15) to hold. Consequently, in this neighbourhood the coefficients of $\tilde{f}$ are, on the exponential scale, as large as those of $\widetilde{F}$, whence the minimal point occurs on V_h.

Proof of Lemma 4.2: Choose a pair of 2-intervals of sizes k and k' uniformly from pairs where both contain 1 and n. Define the random variable $Z_{k,k',n}$ to be n^{-1} times the cardinality of the intersection of these two sets. If $Z_{k,k',n} \to \gamma$ in probability as $n \to \infty$ with $k/n \to \alpha$ and $k'/n \to \beta$, then the supremum of $\mathsf{rate}(\alpha, \beta, \cdot)$ occurs at γ. But we see from the construction of the generating function that the indicator functions of the two sets form two independent Markov chains, both constrained to go from 1 to n in a given number of steps. This implies asymptotically independent statistics, meaning that the limit in probability of $Z_{k,k',n}$ is kk'/n. ∎

Proof of Lemma 4.3: We argue here only what we need, namely the lemma for $\beta = \alpha$ and α in a neighbourhood of α_*. We have seen that rate and $-\mathsf{ent}$ both have maxima at $(\alpha_*, \alpha_*, \alpha_*^2)$, so this is a matter of checking the second derivative of $2 \cdot \mathsf{rate} + \mathsf{ent}$ for negative definiteness and then checking that this local maximum is in fact a global maximum for F_2. To do this, we evaluate the point $\mathbf{x}$ controlling asymptotics of $F_2(\alpha, \alpha, \rho)$.

Using Maple, we solve for $\mathbf{x} \in V_h$ and $(x_1 \partial h/\partial x_1, x_2 \partial h/\partial x_2, x_3 \partial h/\partial x_3, x_4 \partial h/\partial x_4)$ parallel to $\mathbf{s}$ for which $\alpha = s_1/s_3, \beta = s_2/s_3, \rho = s_4/s_3)$. We get

$$\begin{aligned} u_0 = v_0 &= \frac{2\alpha - 1 - R}{2(\alpha + \delta - 1)}\,; \\ z_0 &= -\frac{(2\alpha - 1 - R)\delta}{4\delta^2 - 6\delta + 2\delta R + 4\delta\alpha - 6\alpha + 3 + 4\alpha^2 - R}\,; \\ \tau_0 &= -\frac{4\delta^2 - 6\delta + 2\delta R + 4\delta\alpha - 6\alpha + 3 + 4\alpha^2 - R}{2\delta\alpha - \delta - \delta R - 4\alpha + 4\alpha^2 + q + 1 - 2\alpha R}\,; \end{aligned}$$

here, $\delta = 1 - 2\alpha + \rho$ and

$$R := \sqrt{8\alpha^2 - 12\alpha + 5 + 8\delta\alpha - 8\delta + 4\delta^2}\,.$$

We take two derivatives and evaluate at $\rho = \alpha^2$, obtaining a rational function of α. Doubling and adding the second derivative of `ent`, we may verify negativity of the second derivative in ρ for all $\alpha > 2/3$. This establishes a local maximum at $\rho = \alpha^2$. We have already seen that the global maximum of $F_2(\alpha, \beta, \rho)$ occurs at $\beta = \alpha$. It is straightforward to check numerically that there is no maximum in a set bounded away from the diagonal $\beta = \alpha$ exceeding the value of $F_2(\alpha, \alpha, \alpha^2)$ at $\alpha = \alpha_*$. ∎

Proof of Lemma 4.5: For this we need to compute the solution $\mathbf{x}(\mathbf{s})$ in all four variables, rather than just when $\beta = \alpha$. The result is rather messy and may be found in [8]. In a neighbourhood of $(\alpha_*, \alpha_*, \alpha_*^2)$ we may use rigorous numerical estimates to verify that the Hessian is non-degenerate. This establishes the second assertion of the lemma. The first follows from [16, Theorem 3.5]. ∎

References

[1] R. Arratia, L. Goldstein and L. Gordon. Two moments suffice for Poisson approximation: the Chen-Stein method. *Annals of Probability*, 17:9–25, 1989.

[2] R. Arratia, L. Goldstein and L. Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5:403–424, 1990.

[3] Y. Baryshnikov and R. Pemantle. *Manuscript in preparation.*

[4] A. Bergeron and J. Stoye, On the Similarity of Sets of Permutations and Its Applications to Genome Comparison, COCOON 2003, *Lecture Notes in Computer Science*, 2697: 68-79, (2003).

[5] A. Bergeron, S. Heber and J. Stoye, Common intervals and sorting by reversals: a marriage of necessity. *Proceedings of ECCB 2002*: 54-63, (2002).

[6] A. Bergeron, S. Corteel and M. Raffinot: The Algorithmic of Gene Teams. WABI 2002, *Lecture Notes in Computer Science*, 2452: 464-476, (2002).

[7] K. S. Booth and G. S. Lueker, Testing for the Consecutive Ones Property, Interval Graphs, and Graph Planarity Using PQ-Tree Algorithms. *J. Comput. Syst. Sci.* 13(3): 335-379 (1976)

[8] S. Corteel, G. Louchard and R. Pemantle. Common intervals in permutations (unabridged version). `http://www.math.upenn.edu/~pemantle/louchard/version040122.ps`

[9] G. Didier, Common Intervals of Two Sequences. WABI 2003, *Lecture Notes in Computer Science*, 2812: 17-24 (2003).

[10] S. Heber and J. Stoye, Algorithms for Finding Gene Clusters. WABI 2001, *Lecture Notes in Computer Science*, 2149: 252-263, (2001).

[11] S. Heber and J. Stoye, Finding All Common Intervals of k Permutations. CPM 2001, *Lecture Notes in Computer Science*, 2089: 207-218, (2001).

[12] I. Kaplansky. The asymptotic distributions of runs of consecutive elements. *Annals of Mathematical Statistics*, 16:200–203, 1945.

[13] S. Kobayashi, I. Ono and M. Yamamura, An Efficient Genetic Algorithm for Job Shop Scheduling Problems. *ICGA 1995*: 506-511, (1995).

[14] V.K. Kolchin, A.S. Sevastyanov, and P.C. Chistiakov. *Random Allocations*. Wiley, 1978.

[15] H. Mühlenbein, M. Gorges-Schleuter, and O. Krämer. Evolution algorithms in combinatorial optimization.*Parallel Comput.*, 7:65-85, (1988).

[16] R. Pemantle and M. Wilson Asymptotics of multivariate sequences, part I: smooth points of the singular variety. *J. Comb. Theory, Series A,* 97:129–161, 2001.

[17] R. Pemantle and M. Wilson Asymptotics of multivariate sequences, part II: multiple points of the singular variety. *Combinatorics, Probability and Computing, to appear.*

[18] T. Uno and M. Yagiura, Fast Algorithms to Enumerate All Common Intervals of Two Permutations. *Algorithmica* 26(2): 290-309 (2000).

[19] J. Wolfowitz. Additive partition functions and a class of statistical hypotheses. *Annals of Mathematical Statistics*, 13:247–279, 1942.

[20] J. Wolfowitz. Note on runs of consecutive elements. *Annals of Mathematical Statistics*, 15:97–98, 1944.