Discrete Mathematics and Theoretical Computer Science DMTCS vol. **8**, 2006, 189–216

Common Intervals in Permutations

Sylvie Corteel[1] and Guy Louchard[2] and Robin Pemantle[3]

[1] *CNRS PRiSM, Université de Versailles Saint-Quentin, 45 Avenue des Etats-Unis, 78035 Versailles France, email: syl@prism.uvsq.fr*
[2] *Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email:louchard@ulb.ac.be*
[3] *Department of Mathematics, University of Pennsylvania, 209 S. 33rd Street, Philadelphia, PA 19104 USA, email:pemantle@math.upenn.edu*

received January 24, 2005, *revised April 13, 2006, June 26, 2006*, *accepted June 26, 2006*.

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 generalize 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 389415627 for $\delta = 2$.

After analyzing 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.

Keywords: intervals in random permutations, gene teams, annealed mean

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 $[1, 4]$.

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. We will show in section 2 that 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 [11] and Wolfowitz [18, 19] from a statistical point of view. See also [13]. Recently several algorithms were designed to efficiently enumerate all common intervals of permutations [9, 17] 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 [12, 14]. In bioinformatics [4, 5, 8, 9, 10], genomes of prokaryotes can be modeled 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 [9, 10]. 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 generalization of intervals, namely δ-intervals (the previous case corresponds to $\delta = 1$).

Definition 1.2 *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 paper 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 for $\delta = 2$, 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 Y_k will be typically exponentially large: the number of 2-intervals of $[n]$ of size k grows exponentially, the probability of G^{-1} of one of these also being a 2-interval decays exponentially, and the growth overcomes the decay when $\alpha > \alpha_k$.

Seeing that Y grows exponentially in n, it is natural to look at the rescaled quantity $n^{-1} \log Y$. We compute the *annealed* mean, $n^{-1} \log \mathbb{E} Y$. 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 behavior of Y. Often one has a so-called lottery effect, meaning that the mean of a quantity Y comes primarily from an exponentially small number of values that are exponentially larger than the median value, and that consequently, $\mathbb{E}\log Y < \log \mathbb{E}Y$. For example, when there is a Gaussian limit law, $n^{-1/2}(\log Y - n\mu) \to \mathcal{N}(0, \sigma^2)$, then one will typically have a lottery effect. Perhaps surprisingly in light of the discussion in section 7, there is no lottery effect. Our main result, Theorem 4.1 below, is that for $\delta = 2$, we have $\mathbb{E}Y^2 = \mathcal{O}(\mathbb{E}Y)^2$. This shows that as $n \to \infty$, the sequence $\overline{Y} := Y/\mathbb{E}Y$ is tight on $(0, \infty)$, meaning that when Y is rescaled by its mean, the probability of $Y/\mathbb{E}Y < \epsilon$ is uniformly bounded by some $g(\epsilon)$ going to zero as $\epsilon \to 0$.

The paper is organized as follows. In Section 2 we study the case $\delta = 1$. We recall previous results and show that the distribution of X converges to a Poisson with mean 2. In Section 3 we compute the mean value of 2-intervals. We use basic counting arguments and then apply a saddle point argument. Section 4 states the main result which is that the quenched and annealed means are the same, that is $\mathbb{E}Y^2 = O(\mathbb{E}Y)^2$. Then we outline in Section 4 all the steps of the proof. The proof itself is presented in Sections 5, 6 and 7. In Section 5 we present a 4-variable generating function for pairs of 2-intervals of size k and k' which overlap on κ positions. Then in Section 6 we compute a rate function thanks to saddle point analysis which gives us the asymptotics of the number of pairs of 2-intervals of size k and k' which overlap on κ positions. This gives the exponential part of $\mathbb{E}Y^2$. The full computation of the asymptotics of the non-exponential part of $\mathbb{E}Y^2$ seems daunting. We will show in section 7 that $\mathbb{E}Y^2 - (\mathbb{E}Y)^2 \neq o(\mathbb{E}Y)^2$, and that a Gaussian limit is not possible (Theorem 7.1). We are left to conjecture that $\overline{Y}$ converges in distribution to an unidentified positive real random variable, whose properties are discussed in section 7.

A preliminary version of this paper was presented at the Third International Colloquium of Mathematics and Computer Science held at the Vienna University of Technology (Sep.2004).

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 [17] computed

$$\begin{aligned} \mathbb{E}(X_2) &= \frac{2(n-1)}{n}\,; \\ \mathbb{E}(X_3) &= \frac{6(n-2)}{n(n-1)}\,; \\ \mathbb{E}(X_k) &\leq \frac{24}{n^2} \ \text{ for } k \geq 4 \end{aligned}$$

Although this was not explicitly stated in [17], it is not hard to show that in fact

Proposition 2.1 *As $n \to \infty$, the distribution of X converges to a Poisson with mean 2.*

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

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

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 [11]. We give here an independent argument via a more modern approach, using the Poisson approximation machinery first developed by Chen and Stein, and put in an explicit and usable form in [1].

Proof: Recall that G is a random permutation of $[n]$. 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] concludes convergence to a Poisson with mean $\sum_k p_k$ provided the following three quantities go to zero. In fact the distance in total variation to the Poisson for any fixed n is bounded above by $2(b_1(n) + b_2(n) + b_3(n))$, so the argument will show that the total variation distance is $\mathcal{O}(1/n)$.

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

Each $p_k = (n-1)/\binom{n}{2} = 2/n$, since there are $\binom{n}{2}$ pairs of positions, and $n-1$ pairs of adjacent positions. Similarly there are $\binom{n-2}{2}$ pairs of pairs of adjacent positions and $\binom{n}{4}$ quadruples of positions. It follows that for $|k-j| \geq 2$, $p_{k,j} = (1/3)\binom{n-2}{2}/\binom{n}{4} = 4/(n^2-n)$. Immediately we see that $b_1 \leq n[3(2/(n-2))^2] = \mathcal{O}(1/n)$ and that $b_2 \leq 3n4/(n^2-n) = \mathcal{O}(1/n)$.

It remains only to show that $b_3 = \mathcal{O}(1/n)$. Let σ denote $\sigma(A_j : j \notin B_k)$. A simple identity is

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

where $|| \cdot ||_{TV}$ is the total variation distance, μ is the unconditional probability measure on σ and μ_{A_k} is the μ conditioned on A_k. Now there is an easy way to generate a random permutation conditional on A_k: pick G_0 uniformly at random, pick a pair of positions $\{j, j+1\}$ independently uniformly at random, switch the values of G^{-1} on k and $G(j)$, and switch the values G^{-1} on $k+1$ and $G(j+1)$. With probability $1 - \mathcal{O}(1/n)$, this does not change whether A_j occurs for any $j \neq k$. It follows that $||\mu_{A_k} - \mu||_{TV} = \mathcal{O}(1/n)$, which completes the proof. $\square$

More generally, consider X_k. There are $T_{1,k} = n-k+1$ possible k-tuples and each of them has a probability

$$\Pi_{1,k} = k! \frac{1}{n(n-1)\dots(n-k+2)}$$

of being made of k consecutive integers (again irrespective of their order, we denote this property by C_k). Hence

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

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

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

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 $E_3^2 := \mathbb{E}\left(X_3^2\right)$. First we have $T_{1,3} = n-2$ triplets contributing by $\mathbb{E}(X_3)$ to E_3^2. Next we have $T_{2,3} = 2(n-3)$ couple of triplets with two common values and the probability of one of these couples contributing by 1 to E_3^2 is given by

$$\Pi_{2,3} = \frac{1}{n(n-1)(n-2)(n-3)}[2!T_{2,3}].$$

Next we have $T_{3,3} = 2(n-4)$ couples of triplets with one common value and the probability of one of these couples contributing by 1 to E_3^2 is given by

$$\Pi_{3,3} = \frac{1}{n(n-1)(n-2)(n-3)(n-4)}[(2!)^2 T_{3,3}].$$

Finally, there are $T_{4,3} = (n-4)(n-5)$ couples of disjoint triplets and the probability of one of these couples contributing by 1 to E_3^2 is given by

$$\Pi_{4,3} = (3!)^2 \frac{T_{4,3}}{\prod_0^5 (n-i)} = (3!)^2 \frac{1}{n(n-1)(n-2)(n-3)}.$$

Note again that

$$\sum_1^4 T_{i,3} = (n-2) + 2(n-3) + 2(n-4) + (n-4)(n-5) = (n-2)^2$$

as it should. This leads to

$$E_3^2 = \mathbb{E}\left(X_3^2\right) \sim \frac{6}{n} + \frac{38}{n^2} + \dots \tag{3}$$

What is the asymptotic distribution of X_3? Let H_i denote the event: the triplet $[\sigma_i, \sigma_{i+1}, \sigma_{i+2}]$ is made of three consecutive integers (again irrespective of their order). We obtain, by inclusion-exclusion,

$$\begin{aligned}
\Pr[X_3 = 0] &= \Pr\left[\prod_{i=1}^{n-2}[I - H_i]\right] \\
&= 1 - T_{1,3}\Pi_{1,3} + T_{2,3}\Pi_{2,3}/2 + T_{3,3}\Pi_{3,3}/2 + T_{4,3}\Pi_{4,3}/2 + \dots \\
&= 1 - (n-2)\frac{3!}{n(n-1)} + (n-3)\frac{1}{n(n-1)(n-2)(n-3)}[4 + 4(n-4)]
\end{aligned}$$

$$
\begin{aligned}
&+ \quad (n-4)\mathcal{O}\left(\frac{1}{n^4}\right) + \frac{(n-4)(n-5)}{2}(3!)^2\frac{1}{n(n-1)(n-2)(n-3)} + \dots \\
\Pr[X_3 = 1] &= \Pr\left[\sum_{i=1}^{n-2} H_i \prod_{j\neq i}[I - H_j]\right] \\
&= T_{1,3}\Pi_{1,3} - T_{2,3}\Pi_{2,3} - T_{3,3}\Pi_{3,3} - T_{4,3}\Pi_{4,3} + \dots \\
&= (n-2)\frac{3!}{n(n-1)} - 2(n-3)\frac{1}{n(n-1)(n-2)(n-3)}[4 + 4(n-4)] \\
&- \quad 2(n-4)\mathcal{O}\left(\frac{1}{n^4}\right) - (n-4)(n-5)(3!)^2\frac{1}{n(n-1)(n-2)(n-3)} + \dots \\
\Pr[X_3 = 2] &= \Pr\left[\sum_i \sum_{j>i} H_i H_j \prod_{k\neq(i,j)} [I - H_k]\right] \\
&= T_{2,3}\Pi_{2,3}/2 + T_{3,3}\Pi_{3,3}/2 + T_{4,3}\Pi_{4,3}/2 + \dots \\
&= (n-3)\frac{1}{n(n-1)(n-2)(n-3)}[4 + 4(n-4)] + (n-4)\mathcal{O}\left(\frac{1}{n^4}\right) \\
&+ \quad \frac{(n-4)(n-5)}{2}(3!)^2\frac{1}{n(n-1)(n-2)(n-3)} + \dots
\end{aligned}
$$

This leads to

$$
\begin{aligned}
\Pr[X_3 = 0] &\sim 1 - \frac{6}{n} + \frac{28}{n^2} + \dots \\
\Pr[X_3 = 1] &\sim \frac{6}{n} - \frac{50}{n^2} + \dots \\
\Pr[X_3 = 2] &\sim \frac{22}{n^2} + \dots . \qquad (4)
\end{aligned}
$$

Note that this also leads to $\mathbb{E}(X_3) \sim \frac{6}{n} - \frac{6}{n^2} + \dots$ and $\mathbb{E}(X_3^2) \sim \frac{6}{n} + \frac{38}{n^2} + \dots$ which is of course comptatible with (1) and (3).

A simulation of X_3 based on $M = 3000$ trials with $n = 40$ is given in Figure 1 (asymptotic =line, observed =circle). The fit is reasonable.

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. 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

$$\mathbb{E}Y_k = \frac{N(k, n)^2}{\binom{n}{k}} . \qquad (5)$$

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. Let $T_{i,k}$ denote the

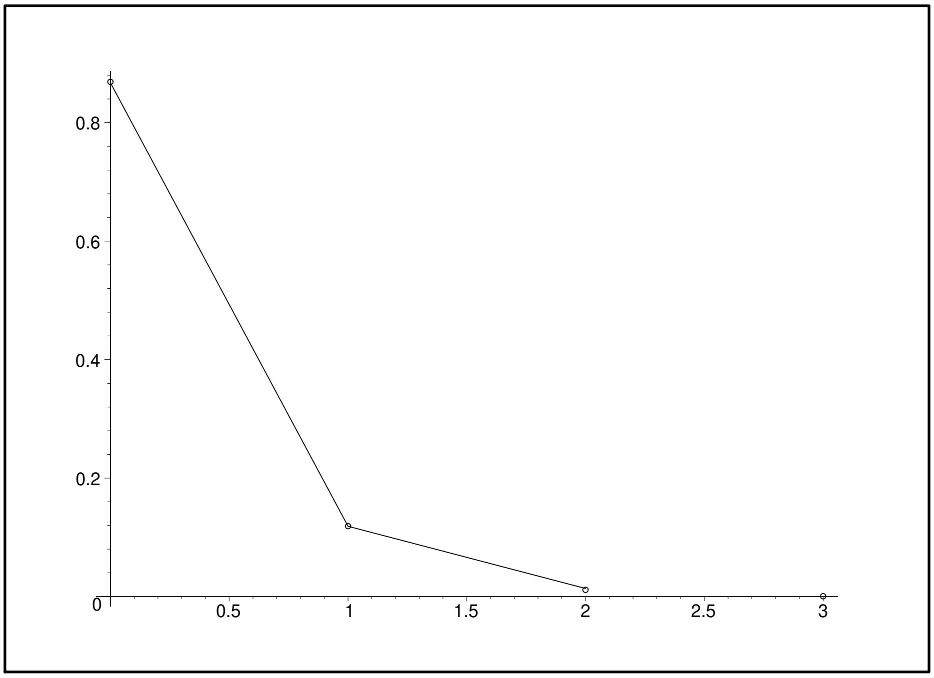

Fig. 1: Observed and limiting (4) X_3 distribution

number of 2-intervals of cardinality k with i holes. 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

$$T_{i,k} := (n-k+1-i) \cdot \binom{k-1}{i}$$

$0 \le i \le m_k$, and

$$\mathbb{E}Y_k = \frac{\left(\sum_{i=0}^{m_k}(n+1-k-i)\binom{k-1}{i}\right)^2}{\binom{n}{k}} . \tag{6}$$

When $k/n < 1/2$ then

$$N(k,n) = \sum_{i=0}^{k-1}(n+1-k-i)\binom{k-1}{i} = \left(n + \frac{3}{2} - \frac{3}{2}k\right) 2^{k-1} . \tag{7}$$

When $k/n > 1/2$ the sum is better approximated than evaluated exactly. We find that

$$N(k,n) = \sum_{i=0}^{n-k}(n+1-k-i)\binom{k-1}{i} ,$$

which has its maximum term near the endpoint $i = n-k$ when $k \ge \frac{2}{3}n$, and near $k/2$ when $k \le \frac{2}{3}n$. More categorical asymptotics than we need are available by using a normal approximation near $k = 2n/3$. For $n/2 < k < 2n/3)$ with $(2n/3-k)/\sqrt{n} \to \infty$, we have

$$\begin{aligned} N(k,n) &= \sum_{i=0}^{n-k}(n+1-k-i)\binom{k-1}{i} \\ &\sim \sum_{i=0}^{k-1}(n+1-k-i)\binom{k-1}{i} \\ &= \left(n+\frac{3}{2}-\frac{3}{2}k\right) 2^{k-1} . \end{aligned} \tag{8}$$

which is asymptotically the same as when $k < n/2$. On the other hand, for $(k-2n/3)/\sqrt{n} \to \infty$, we have

$$\begin{aligned} N(k,n) &= \sum_{i=0}^{n-k}(n+1-k-i)\binom{k-1}{i} \\ &\sim \binom{k-1}{n-k}\left(\frac{2k-1-n}{3k-1-2n}\right)^2 . \end{aligned} \tag{9}$$

Finally, for $x := k^{-1/2}(k - 2n/3) = \mathcal{O}(1)$, the normal approximation yields directly

$$N(k,n) = 2^{k-1}\binom{\sqrt{k}}{2}\Psi_x\,, \tag{10}$$

where Ψ_x is the expected positive part of $Z + x$ with Z a standard normal.

Via equation (6), 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 (9) with $n \to \infty$ and $\alpha := k/n$ to obtain

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

where

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

When $\alpha < 2/3$, we obtain, by a similar analysis,

$$F_1(\alpha) = \alpha\log(\alpha) + (1-\alpha)\log(1-\alpha) + 2\log(2)\alpha.$$

A plot of $F_1(\alpha)$ is given in Figure 2. When $\alpha \geq 2/3$, the function $F_1'(\alpha)$ vanishes when α solves

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

The unique root in $[2/3, 1]$ is $\alpha_* \approx .7840013296\ldots$. A saddle point approximation now gives us

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

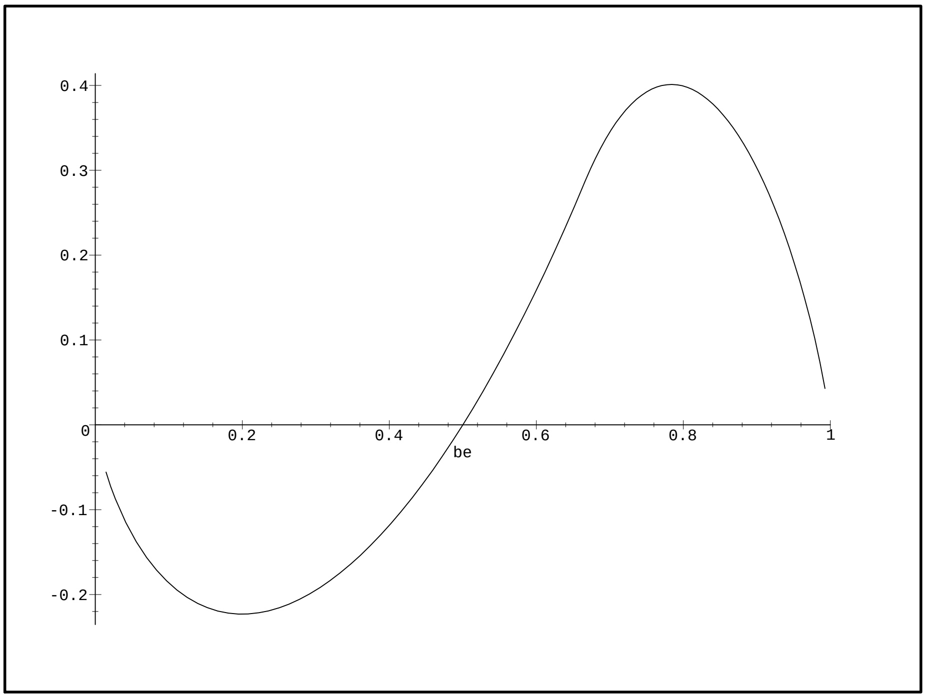

Fig. 2: $F_1(\alpha)$

4 The second moment of the number of 2-intervals

In the next three sections, we show:

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

This section outlines, in three subsections, the argument for this. The subsequent section derives the crucial generating function and the following one uses the generating function to verify certain key computations.

4.1 Counting pairs of 2-intervals

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 behavior 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 determine 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 specialize 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

$$\begin{aligned} \alpha &:= \frac{k}{n} \\ \beta &:= \frac{k'}{n} \\ \rho &:= \frac{\kappa}{n} \end{aligned}$$

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

$$N(k, k', n, \kappa) \sim A_0(\alpha, \beta, \rho) n^{-3/2} \exp(n \mathsf{rate}(\alpha, \beta, \rho)) \tag{12}$$

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 will 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 given by

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

It follows that 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

$$P(k, k', n, \kappa) := \binom{n}{\kappa, k-\kappa, k'-\kappa, n-k-k'+\kappa}^{-1}$$

and that the rate function for this probability, $\lim n^{-1} \log P$, which we denote ent for entropy, is given by

$$\mathsf{ent} := \rho \log \rho + (\alpha - \rho) \log(\alpha - \rho) + (\beta - \rho) \log(\beta - \rho) + (1 - \alpha - \beta + \rho) \log(1 - \alpha - \beta + \rho) \,. \quad (13)$$

Observe that the function ent satisfies the identity

$$\mathsf{ent}(\alpha, \beta, \alpha \cdot \beta) = h(\alpha) + h(\beta) \quad (14)$$

where $h(x) = x \log(x) + (1 - x) \log(1 - x)$ is the usual entropy function, and in particular that $\rho = \alpha \cdot \beta$ is a maximum for the function $-\mathsf{ent}$ for fixed α, β.

The rate function for the expected number of pairs of 2-intervals of respective sizes k and k' with overlap κ is given by

$$F_2 := 2 \cdot \mathsf{rate} + \mathsf{ent} \,, \quad (15)$$

which we now analyze.

4.2 The exponential order of the second moment

Since $\mathbb{E}Y^2 = \sum_{k,l} \mathbb{E}Y_k Y_l = \sum_{k,l,\kappa} N(k, l, n, \kappa)^2 P(k, l, n, \kappa)$ 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

$$n^{-1} \log \mathbb{E}Y^2 \to \sup_{\alpha,\beta,\rho} F_2(\alpha, \beta, \rho) := \lambda^{**} \,. \quad (16)$$

In order to compute λ^{**} we must find the location, $(\alpha_{**}, \beta_{**}, \rho_{**})$, of the maximum of F_2. Without computing, we may narrow the search considerably. First, from the inequality $\mathbb{E}Y_k Y_l \le \frac{1}{2}(\mathbb{E}Y_k^2 + \mathbb{E}Y_l^2)$, we see that the maximum of $\mathbb{E}Y_k Y_l$ (for fixed n) can only occur when $l = k$, and therefore

$$\lambda^{**} = F_2(\alpha, \alpha, \rho) \quad (17)$$

for some α, ρ.

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

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

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

$$\sup_\rho \mathsf{rate}(\alpha, \beta, \rho) = u(\alpha) + u(\beta) \quad (18)$$

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

Later, we will show, for α and β in a range containing α_{**}, that this supremum occurs at $\rho = \alpha \cdot \beta$. Assume this for now. We have previously remarked that $-\mathsf{ent}(\alpha, \beta, \cdot)$ has a maximum at $\alpha \cdot \beta$ as well. Both functions are smooth, so the function $F_2 = 2 \cdot \mathsf{rate} + \mathsf{ent}$ has a critical point there as well. With some work, using a four-variable generating function, we will show that this critical point is a maximum. It will then follow from (18) and (14) 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 would give

$$\lambda^{**} = 2F_1(\alpha_*)\,. \tag{19}$$

4.3 End of the outline

More precise information about $\mathbb{E}Y^2$ is obtained by a saddle point summation. In particular, from the form of the generating function, it will follow that there is a $n^{-3/2}$ correction term:

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

in a neighborhood of the maximum. The product $N(k,k',n,\kappa)^2P(k,k',n,\kappa)$ is then asymptotically

$$A_1(\alpha,\beta,\rho)n^{-3/2}\exp(nF_2(\alpha,\beta,\rho))$$

and the sum will be asymptotically $C\exp(nF_2(\alpha_*,\alpha_*,\alpha_*^2))$ where

$$C=\frac{A_1(\alpha_*,\alpha_*,\alpha_*^2)}{\sqrt{(2\pi)^3\mathcal{H}}}$$

and $\mathcal{H}$ is the Hessian of the rate function F_2 in appropriate coordinates. In particular the polynomial correction term is canceled by the summation, and the demonstration that $\mathbb{E}Y^2=\mathcal{O}(\mathbb{E}Y)^2$ is completed.

To summarize, the foregoing outline proves Theorem 4.1 once we have verified several assertions:

(m1) that the supremum of $\mathsf{rate}(\alpha,\beta,\cdot)$ occurs at $\alpha\cdot\beta$;

(m2) that the critical point of F_2 at $\rho=\alpha\cdot\beta$ is a maximum;

(m3) that exact asymptotics for $\mathbb{E}Y_kY_l$ have an $n^{-3/2}$ correction term to the exponential and that the discrete saddle point summation exactly cancels this out.

We remark that this argument does not show that $\mathbb{E}Y^2-(\mathbb{E}Y)^2=\Theta(\mathbb{E}Y)^2$. In principle we could compute $A(\alpha_*,\alpha_*,\alpha_*^2)$ and $\mathcal{H}$ and compare to the constant computed in (11), but this computation seems daunting. Instead, a separate argument ruling out the possibility that $\mathbb{E}(Y^2)-(\mathbb{E}Y)^2=o(\mathbb{E}Y)^2$ is given in the last section.

Before turning to the generating function, we establish some helpful numerical bounds.

Lemma 4.2 *If $\alpha < 12/17$ or if $\alpha > 0.847$ then*

$$F_2(\alpha,\alpha,\rho) < 2F_1(\alpha_*)\,.$$

Proof: Suppose first that $\alpha > 0.847$. A crude upper bound for $F_2(\alpha,\alpha,\rho)$ is $\lim_{n\to\infty} N(\lfloor \alpha n \rfloor, n)^2$. This latter quantity is equal to 2η where $\eta := \alpha\log(\alpha) - (1-\alpha)\log(1-\alpha) - (2\alpha-1)\log(2\alpha-1)$. We may verify that when $\alpha > 0.847$ then $\eta < 0.4001$, which is known to be less than $F_1(\alpha_*)$. Verification: $d\eta/d\alpha$ is the log of an algebraic expression and vanishes exactly once on the interval $[1,2,1]$ namely at $\alpha = (5+\sqrt{5})/10 \approx 0.7236$; in particular, η is decreasing on $[0.73, 1]$ and $\eta(0.847) < 0.4001 < F_1(\alpha_*)$. For the case when $\alpha < 12/17$, the most straightforward method is to divide the α-δ plane into small squares. In each we may easily compute separate upper and lower bounds for $2 \cdot \mathtt{rate}$ and $\mathtt{ent}$, since these functions will have extremes at a vertex of the grid square. With the grid squares sufficiently small, one may then check that the sum of the bounds does not exceed $2F_1(\alpha_*)$. For details, see

`http://www.math.upenn.edu/~pemantle/CLP04-grid1.mw`

□

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

As a warm-up, let us compute the generating function for $N(k,n)$ and recover the asymptotic formulae (7) – (9). Recall from building the generic indicator function that a 2-interval consists of any number of unused locations, followed by a single used location, followed by any sequence of pairs (unused , used) and single used locations, followed by any number of unused locations. Each pair (unused , used) contributes a factor of uz^2 while each single used location gives a factor of uz. The generating function for an arbitrary sequence of pairs and singles is thus $uz/(1-(uz+uz^2)) = uz/(1-(1+z)uz)$. Taking into account the initial and final strings of unused positions gives

$$F_3(u,z) := \sum N(k,n) u^k z^n = \frac{uz}{(1-z)^2(1-(1+z)uz)}\,. \tag{20}$$

To recover the asymptotics in the different regimes, write

$$F_4(z) = [u^k]F_3(u,z) = \frac{(1+z)^{k-1} z^k}{(1-z)^2}$$

and

$$N(k,n) = [z^n]F_4(z) = \frac{1}{2\pi i}\int \frac{(1+z)^{k-1} z^k}{(1-z)^2 z^{n+1}}\,dz\,.$$

Writing $F_5 := \alpha \log(1+z) - (1-\alpha)\log z$ and setting $(k-1)/n := \alpha$ we then have

$$N(k,n) = \frac{1}{2\pi i}\int \frac{\exp(nF_5(z))}{(1-z)^2}\,dz\,.$$

There is a double pole at $z = 1$ and a saddle point at $F_5'(z) = 0$, that is, at $z_* := (1-\alpha)/(2\alpha-1)$. when $\alpha < 2/3$ the double pole is the dominant singularity and leads by residues to

$$N(k,n) \sim -n2^{\alpha n}F_5'(1) = n2^{\alpha n}(1-3\alpha/2)$$

which is (7). When $\alpha > 2/3$ the saddle point is dominant and leads to

$$N(k,n) \sim \frac{1}{(1-z_*)^2\sqrt{-2\pi n F_5''(z_*)}}\exp(nF_5(z_*))$$

which agrees with (9).

Now to derive F, we follow a similar route. A canonical way to build a pair of sets and keep track of the intersection is as follows.

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.

The crucial part of the generating function is the one that grows exponentially with κ, namely the second of the three parts. Nevertheless, in order to compute a valid leading order asymptotic, we must keep track of all three parts.

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. $f_{d'} = zu_2f_{c'}$.

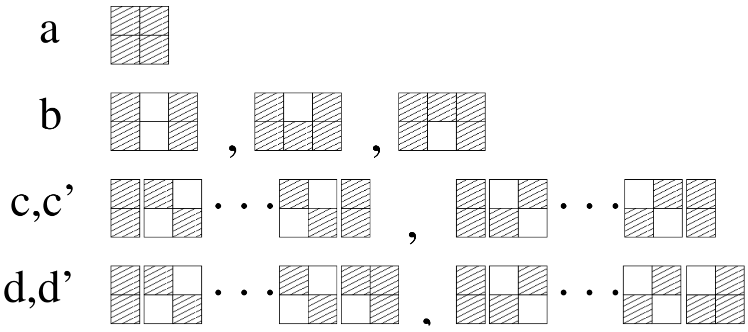

Fig. 3: shaded squares in each row correspond to positions present in that set

The generating function for an arbitrary sequence of these is

$$\begin{aligned} f &= \frac{zu_1u_2t}{1 - zu_1u_2t(f_a + f_b + (2 + z(u_1 + u_2)f_c)} \qquad (21) \\ &= \frac{zu_1u_2t(1 - z^2u_1u_2)}{(1 - z^2u_1u_2)(1 - zu_1u_2t - z^2u_1u_2t(1 + u_1 + u_2)) - zu_1u_2t(2 + z(u_1 + u_2))z^2u_1u_2}. \end{aligned}$$

For the first and last of the three parts, we first recall from (20) 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). $g_g = F_6(u_1, z)/(1 - z^2u_1u_2)$.
- g′ A string of pairs as in case (c′) above, followed by a nonempty sequence as in case (f′). $g_{g'} = F_6(u_2, z)/(1 - z^2u_1u_2)$.
- h The same as (g) except with a position in $I' \setminus I$ in the beginning. $g_h = zu_2F_6(u_1, z)/(1 - z^2u_1u_2)$.

h′ The same as (g′) except with a position in $I \setminus I'$ in the beginning. $g_{h'} = zu_1F_6(u_2, z)/(1-z^2u_1u_2)$.

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

$$g = \frac{1}{1-z}\left(1 + zF_6(u_1, z) + zF_6(u_2, z) + \frac{F_6(u_1, z) + F_6(u_2, z)}{1 - z^2u_1u_2} + \frac{u_2zF_6(u_1, z) + u_1zF_6(u_2, z)}{1 - z^2u_1u_2}\right).$$

Finally,

$$F(u_1, u_2, t, z) = fg^2 . \tag{22}$$

6 The rate function

Let $\mathcal{L}$ denote the logarithmic gradient, that is, $\mathcal{L}Q$ is the vector whose j^{th} coordinate is $x_j\partial Q/\partial x_j$.

Let $h(x, y, z, \tau) = 1 - xy - \tau(1 + x + y + z + xy - xyz)$ and let V_o be the set of smooth points of the variety where h vanishes. Let (μ, ν, δ) denote $(\alpha - \rho, \beta - \rho, 1 - \alpha - \beta + \rho)$. In the next subsection we will prove:

Theorem 6.1 *Let* $\mathbf{s}_0 := (\alpha_{**}, \beta_{**}, \rho_{**})$ *be the argmax for* F_2 *as in (16). There is a neighborhood* $\mathcal{N}$ *of* $\mathbf{s}_0$ *in* $\mathbb{RP}^{d-1}$ *and a continuous map* $\mathbf{x} : \mathcal{N} \to V \cap (\mathbb{R}^+)^4$ *such for every* $\mathbf{s} \in \mathcal{N}$ *there is a point* $\mathbf{x}(\mathbf{s})$ *satisfying:*

1. $h(\mathbf{x}) = 0$*;*
2. $\mathcal{L}h(\mathbf{x}) = \mathbf{s}$*;*
3. *if* $\alpha = s_1 + s_4$*,* $\beta = s_2 + s_4$ *and* $\rho = s_4$ *with* $\mathbf{s}$ *normalized so that* $\sum_j s_j = 1$*, then*

$$\mathsf{rate}(\alpha, \beta, \rho) = -\mu \log x_1 - \nu \log x_2 - \delta \log x_3 - \rho \log x_4 \tag{23}$$

and

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

Readers not interested where this comes from may skip now to section 6.1.

The general approach to extracting asymptotics from the generating function is taken from [3]. Let F be a rational generating function in d variables, written as the quotient of polynomials P/Q with $Q(\mathbf{0}) \neq 0$. The coefficients of $F = \sum_{\mathbf{r}} a_{\mathbf{r}}\mathbf{z}^{\mathbf{r}}$ may be evaluated via Cauchy's integral formula

$$a_{\mathbf{r}} = \frac{1}{(2\pi i)^d}\int_T \mathbf{z}^{-\mathbf{r}}F(\mathbf{z})\,\frac{d\mathbf{z}}{\mathbf{z}} .$$

The cycle of integration, T, is, initially, the product of small circles around the origin in each coordinate. But, letting Dom denote the domain of holomorphy of F (that is $(\mathbb{C}^*)^d \setminus V$ where V is the variety on which Q vanishes), we may replace T by anything in the homology class $[T] \in H_d(\mathsf{Dom})$.

It is shown in [3] that $H_d(\mathsf{Dom})$ is generated by cycles of the following sort. Fix a vector $\mathbf{s} \in (\mathbb{R}^d)^+$, projecting to a direction also denoted $\mathbf{s}$ in $\mathbb{RP}^{d-1}$. positive orthant of $\mathbb{RP}^{d-1}$, for which one wishes to compute asymptotics of $a_{\mathbf{r}}$ as $\mathbf{r} \to \mathbf{s}$. Let $\{S_\beta\}$ be a Whitney decomposition of V into strata (e.g., if V is smooth, then one has simply $\{V\}$). For generic values of $\mathbf{s}$, the function $h_{\mathbf{s}} := -\sum_{j=1}^d s_j \log|z_j|$ will be a Morse function on all strata of V and there will be a finite set of critical points $\{\mathbf{x}_{\beta,j}\}$ of the restriction to stratum S_β of $h_{\mathbf{s}}$. Then $H_d(\mathsf{Dom})$ is generated by cycles C such that

- C is the product of a cycle C_1 in some S_β with an arbitrarily small cycle C_2 orthogonal to S_β;
- C_1 passes through some $\mathbf{x}_{\beta,j}$ and $h_\mathbf{s}$ is maximized on C_1 at $\mathbf{x}_{\beta,j}$.

The integral of $\omega := \mathbf{z}^{-\mathbf{r}} F d\mathbf{z}/\mathbf{z}$ over C_2 will be computable as a residue, ω_1. The integral of ω_1 over C_1 will be a saddle point integral, in the sense that it will be the integral of $P_1(\mathbf{z})\mathbf{z}^{-\mathbf{r}}\, d\mathbf{z}/\mathbf{z}$ where for $\mathbf{r} \to \mathbf{s}$, the dominant term in the integrand, $\mathbf{z}^{-\mathbf{r}}$, has maximum modulus and stationary phase at $\mathbf{x}_{\beta,j}$. It is easy to evaluate $\int_{C_1} \omega_1$ as a stationary phase integral. Asymptotics of $a_\mathbf{r}$ are then obtained by summing over critical points $\mathbf{x}_{\beta,j}$ in the support of $[T]$ (when $[T]$ is written as a linear combination of these, which have positive coefficient). Among these, only those with the highest value of $h_\mathbf{s}$ need be considered.

Carrying out this programme will require several steps:

i Find the critical points (routine but messy exercise in computer algebra)

ii Determine which of these are in the support of $[T]$ (nontrivial topological problem with tidy answer in terms of local tangent cone)

iii Compute the functions $\mathtt{rate}$ and A_0 (straightforward)

iv Optimize F_2 in α, β and ρ (cajole the computer into performing the right simplifications)

v Sum in a neighborhood of the optimum (fairly routine discrete saddle point computation)

The first three of these are carried out in the next subsection, and the last two in the subsequent one.

6.1 Finding the dominating point

We now specialize to the generating function $F = fg^2$. It will turn out to simplify the computations if we 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,

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

The other divisors in V are factors in the denominator of g, namely, $g_1 := 1 - z$, $g_2 := 1 - xy$, $g_3 := 1 - (1+z)x$ and $g_4 := 1 - (1+z)y$. The variety $V_{\widetilde{f}}$ where $\widetilde{f}$ vanishes is not smooth, since $\nabla \widetilde{f}$ vanishes on the curve $\gamma := (-1, -1, 1/t, t)$. Thus $V_{\widetilde{f}}$ has the strata $V_o := V_{\widetilde{f}} \setminus \gamma$ and γ. The other divisors of V are smooth.

Define the function $\mathcal{D} : V \to \mathbb{CP}^3$ by $\mathcal{D}(\mathbf{x}) = \mathcal{L}\mathbf{x}$ for smooth points of V and otherwise by letting $\mathcal{D}(\mathbf{x})$ be the closure of the limit points of $\mathcal{L}(\mathbf{y})$ as $\mathbf{y} \to \mathbf{x}$. The point $\mathbf{x}(\mathbf{s})$ "controls" the asymptotics in the direction $\mathbf{s}$, as captured by the following result taken from [15].

Proposition 6.2 *For any* $\mathbf{s}$ *in the positive orthant of* $\mathbb{RP}^3$*, there is a point* $\mathbf{x}(\mathbf{s})$ *with the following properties.*

1. $\mathbf{x} \in V$
2. $\mathbf{s} \in \mathcal{D}(\mathbf{x})$
3. *if* $\mathbf{x}$ *is a smooth point of* V*, and if* $\alpha = s_1 + s_4$*,* $\beta = s_2 + s_4$ *and* $\rho = s_4$ *with* $\mathbf{s}$ *normalized so that* $\sum_j s_j = 1$*, then*

$$\mathsf{rate}(\alpha, \beta, \rho) = -\mu \log x_1 - \nu \log x_2 - \delta \log x_3 - \rho \log x_4$$

Remark 6.3 *The only difference between this proposition and Theorem 6.1 is that we would like* $\mathbf{x} \in V_o$ *rather than just* $\mathbf{x} \in V$*.*

Proof: Let $\log \mathcal{D}$ be the logarithmic domain of convergence of $\widetilde{F}$, that is, the closure of the set of $\mathbf{x} \in \mathbb{R}^4$ such that $\sum a_{\mathbf{r}} \exp(\mathbf{r} \cdot \mathbf{x})$ is finite. Since $\widetilde{F}$ has nonnegative coefficients, we may use the argument of [15, Theorem 6.3] to see that there will be a *minimal* point $\mathbf{x}(\mathbf{s})$ for which conclusions (1) and (2) hold. Here a minimal point means a point $\mathbf{x}$ which is the only intersection of V with the polydisk $\{\mathbf{y} : |y_j| \leq x_j \forall j\}$ in $\mathbb{C}^4$. The minimal point in question will be the exponential $(e^{u_1}, e^{u_2}, e^{u_3}, e^{u_4})$ of the contact point $\mathbf{u}$ for the support hyperplane to $\log \mathcal{D}$ in direction $\mathbf{s}$. In other words, choosing $\mathbf{u}$ to maximize $\mathbf{s} \cdot \mathbf{u}$ on $\log \mathcal{D}$ will yield a point $\mathbf{x} = \exp(\mathbf{u})$ which is minimal in direction $\mathbf{s}$. The third conclusion follows from [15, Theorem 3.5]. □

Finding the asymptotics for $n^{-1} \log[n\mu, n\nu, n\delta, n\rho]\widetilde{F}$ is now a matter of locating the minimal point.

Lemma 6.4 *Let* $\mathbf{s}_0$ *be the maximizing direction* $(\alpha_{**}, \beta_{**}, \rho_{**})$ *for* F*. There is a neighborhood* $\mathcal{N}$ *of* $\mathbf{s}_0$ *in* $\mathbb{RP}^3$ *such that for any* $\mathbf{s} \in \mathcal{N}$*, the critical point* $\mathbf{x} = (x_0, y_0, z_0, \tau_0)$ *with maximum value of* $h_{\mathbf{s}}$ *among those in the support of* $[T]$ *is in* V_o *and the coordinates are positive and real.*

Proof: Exponentiating $\log \mathcal{D}$, we obtain a set E in the positive orthant of $\mathbb{R}^4$ that may be described as follows. First we compute the intersection with the plane $\tau = 0$, or in other words, the positive minimal points of the divisors of g. Recall that these divisors are $\{z = 1\}, \{x(1+z) = 1\}, \{y(1+z) = 1\}$ and $\{xy = 1\}$. Let E' denote the connected component of the complement of the coordinate hyperplanes and these four divisors that contains the origin in its closure. It is not hard to describe E': it is the region below the graph of the function $z = \min\{1, 1/x - 1, 1/y - 1\}$. A lower boundary is the square $\{z = 0, 0 \leq x \leq 1, 0 \leq y \leq 1\}$, an upper boundary is the square $\{z = 1, 0 \leq x \leq 1/2, 0 \leq y \leq 1/2\}$ and there are sloping curved ruled surfaces for the remaining upper boundary defined by $S_1 := \{z = 1/x - 1, 0 \leq y \leq x \leq 1\}$ and $S_2 := \{z = 1/y - 1, 0 \leq x \leq y \leq 1\}$. The divisor $\{xy = 1\}$ intersects E' only at the point $(1, 1, 0)$ and plays no role in bounding $\log \mathcal{D}$.

In $\mathbb{R}^4$, the set E will be a subset of the cylinder $E'' := E' \times [0, \infty)$. In particular, it will be bounded "below" (lowest τ) by $E' \times \{0\}$ and "above" (highest τ) by $t = \psi(x, y, z) := (1 - xy)/(1 + x + y + z + xy - xyz)$. There will be side boundaries at the graph of ψ restricted to the boundaries of E'.

Now we rule out finding the minimal point at any place other than on the "upper" boundary, $V_f = V_o \cup \gamma$. As long as $\mathbf{s}$ is strictly positive, the support hyperplane to $\log \mathcal{D}$ normal to $\mathbf{s}$ must contact $\log \mathcal{D}$ either at a smooth point whose normal is strictly positive or at a non-smooth point whose normals together have positive values in each coordinate. Exponentiating, we see that the minimal point must be on the closure of the "upper" surface, namely the graph of ψ on E'. We must now rule out the following places for the minimal point to occur:

1. the graph E_1 of ψ on S_1;
2. the graph E_2 of ψ on S_2;
3. the graph E_3 of ψ on the upper square, $[0,1/2]\times[0,1/2]\times\{1\}$.

To rule out E_1, we compute $\mathcal{L}(\mathbf{y})$ as $\mathbf{y}$ in the graph of ψ in the interior of E' converges to the graph on E_1. Recalling that the direction corresponding to such a point on $V\cap E_1$ is given by $\mathbf{v}:=\mathcal{L}\widetilde{f}$ there, we compute the ratio of $(v_1+v_4)/(v_1+v_2+v_3+v_4)$:

$$-\frac{-2\,x^2-4\,x^2y-x^2y^2-1+x^3y+xy+2\,x^3y^2}{3\,x^2+6\,x^2y+x^2y^2-2\,x^3y+x^4y^2+1-2\,x^3y^2}\,. \tag{24}$$

Routine calculus shows this to attain a maximum value of $12/17$ on $V_o\cap S_1$, at $(1/2,1/2,1,1/4)$. The same is true of the ratio $(v_2+v_4)/(v_1+v_2+v_3+v_4)$. Thus points on the graph of ψ on E_1 control asymptotics of $N(k,k',n,\kappa)$ only when $\alpha,\beta\le 12/17$. We see from Lemma 4.2 that $\mathbb{E}X_kX_{k'}$ is always exponentially less than $(\mathbb{E}X)^2$ when $\alpha,\beta\le 12/17$. This rules out E_1, and an analogous argument rules out E_2.

To rule out E_3, we argue combinatorially. Without the factor of $1/(1-z)^2$, we have a generating function $\widehat{F}:=(1-z)^2\widetilde{F})$ that counts the number $\widehat{N}(k,k',\kappa,n)$ of pairs of 2-intervals in which the union of the two intervals contains 1 and n. Suppose, for a given (k,k',κ,n) that

$$\widehat{N}(k,k',\lambda n,\kappa)\ge\widehat{N}(k,k',n,\kappa) \tag{25}$$

for some $\lambda\le 1$ (this can happen, for example when $\mu=\alpha-\rho$ is small). Then

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

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 neighborhood of where the supremum is achieved, it is not possible for (25) to hold. Consequently, in this neighborhood the coefficients of $\widehat{F}$ are, on the exponential scale, as large as those of $\widetilde{F}$, whence the minimal point cannot occur on the divisor $1-z$.

Finally, we must rule out $\mathbf{x}(\mathbf{s})\in\gamma$. If $\mathbf{x}\in\gamma$, and (μ,ν,δ,ρ) are the coordinates of $\mathbf{s}$ normalized to sum to 1, with $\alpha=\mu+\rho$ and $\beta=\nu+\rho$, then

$$\mathsf{rate}(\alpha,\beta,\rho)=\mu\log x_1-\nu\log x_2-\delta\log x_3-\rho\log x_4\,.$$

The curve γ is parametrized by $(1,1,1/t,t)$, so $\mathsf{rate}=(\rho-\delta)\log x_3$. Recalling that the minimal point must lie in the region $x_3:=z\le 1$, and that $\rho>\delta$ for $\alpha,\beta\ge 2/3$, we see that we have $\mathsf{rate}\le 0$ anywhere controlled by a point on γ with $\alpha,\beta\ge 2/3$. It follows that for $\alpha,\beta\ge 12/17>2/3$, the minimal point cannot lie on γ. By Lemma 4.2, directions with $\alpha<12/17$ are not in a neighborhood of $(\alpha_{**},\alpha_{**},\rho_**)$, which completes the proof of the lemma. □

Proof of Theorem 6.1: By Proposition 6.2, the rate function rate is controlled by a point $\mathbf{x}(\alpha,\beta,\rho)$ with properties stated in the proposition. The foregoing lemma then shows that $\mathbf{x}\in V_o\cap(\mathbb{R}^+)^4$, which proves Theorem 6.1. □

Let us interpret Theorem 6.1 and its proof. Values of (α, β, ρ) whose corresponding minimal points lie on V_o correspond to values of (k, k', n, κ) for which almost all the pairs of 2-intervals counted by $N(k, k', n, \kappa)$ nearly span the interval $[n]$. In other words, when the minimal point is on V_o, the exponential rate for coefficients of F is the same as that for f, which enumerates pairs of 2-intervals both of which have 1 and n for their first and last element respectively[(i)]. We have interpreted f as generating a sequence of blocks between common values; it is well known that the statistics of these blocks are asymptotically independent. In particular, the intersection of two independent random 2-intervals, chosen uniformly from pairs of sizes (k, k'), will be almost asymptotically of size $N(k/n)N(k'/n)$ in probability. In other words, when the conclusion of Theorem 6.1 holds, the assertion (1) also holds.

6.2 Optimization and algebraic simplification

Everything now rests on verifying (2), namely that the critical point for $F_2(\alpha, \beta, \cdot)$ at $\alpha \cdot \beta$ is actually a maximum. This is mostly one long computer algebra computation. For replicability, we outline the trickier steps. Verification of the maximum can be (and later will have to be) restricted to the diagonal $\alpha = \beta$. But in order also to verify (3) we will need a summation over all α and β, so we begin with all the variables.

To find the critical point $\mathbf{x}(\alpha, \beta, \rho) = (u_0, v_0, z_0, \tau_0) \in V_o$ for $\widetilde{f}$, we tell Maple to compute the Groebner basis for the ideal generated by

```
[h ,
(alpha+beta-1+delta)*vv[1] - (1-alpha-delta)*vv[4] ,
vv[2]*(1-alpha-delta) - vv[1]*(1-beta-delta) ,
(1-alpha-delta)*vv[3] - delta * vv[1]
];
```

where

```
h := 1 - u*v - (1 + u*v + z - z*u*v + u + v)*t;
vv := simplify([u*diff(h,u) , v*diff(h,v) , z*diff(h,z) , t*diff(h,t)]);
```

this is done using the `Basis` command of the `Groebner` package in Maple 10 (earlier versions use variants such as `gbasis` and name the package differently).

In order to get Maple to halt we had to first use the term ordering `tdeg[z,t,alpha,beta,delta,u,v]` and then compute a basis in the order `plex[z,t,u,alpha,beta,delta,v]` for the Groebner basis coming from the previous computation. The resulting Groebner basis has 27 generators, that first one of which (it will be the the last rather than the first in versions of Maple before Maple 10) is an elimination polynomial for v, that is, it contains v, α, β and δ but not z, u or t. Factoring out $(v + 1)$ and the constant term $(1 - \alpha - \delta)$, we solve for v to obtain

```
v0 := 1/2*(alpha^2+4*beta*alpha-4*alpha+2*alpha*delta-beta^2
-2*beta*delta+1-(5-16*alpha-8*delta-16*beta+18*beta^2
-16*beta^2*delta-32*alpha^2*beta+40*beta*alpha+20*beta*delta
+18*alpha^2+20*alpha*delta+4*delta^2+alpha^4+4*alpha^3*delta
-8*beta^3+8*alpha*delta^2*beta+8*beta^3*alpha+4*beta^3*delta
+14*beta^2*alpha^2-8*alpha^3+4*alpha^2*delta^2
```

(i) As an aside: when the minimal point is on E_1, the first 2-interval tends to span only $(1 - \Theta(1))n$ of the interval $[n]$; when it is on E_2, the second 2-interval is similarly short; when it is on E_3, the union of the two 2-intervals is short.

```
+12*beta*delta*alpha^2-16*alpha^2*delta+beta^4
-32*alpha*beta*delta+4*delta^2*beta^2-8*delta^2*beta
-8*alpha*delta^2-32*beta^2*alpha+12*beta^2*alpha*delta
+8*beta*alpha^3)^(1/2))/(-delta+2*beta^2+2*beta*delta+1-3*beta)
```

Luckily this can be simplified. First, we find the elimination polynomial for u in terms of v, α, β, δ, which is the 3^{rd} basis element, divided by $(1+v)^2$:

```
upoly := -u*delta*v-u*beta*v+u*v+alpha+delta-u-1+2*alpha*u;
```

Solving for u and plugging in the value v_0 above yields

$$u_0 = 2\,\frac{(2\,\beta-1)\,(-1+\alpha+\delta)}{2\,\alpha\,\delta-2\,\beta\,\delta-1+4\,\beta+\alpha^2-4\,\beta\,\alpha-\beta^2-Q}\,.$$

where

$$Q = \sqrt{(\alpha^2+4\,\alpha\,\delta-6\,\alpha+6\,\beta\,\alpha-6\,\beta+\beta^2+4\,\beta\,\delta-8\,\delta+4\,\delta^2+5)\,(-1+\alpha+\beta)^2}\,. \qquad (26)$$

We then recover a simplified version of v_0 by symmetry, switching α and β:

$$v_0 = 2\,\frac{(2\,\alpha-1)\,(-1+\delta+\beta)}{2\,\beta\,\delta-2\,\alpha\,\delta-1+4\,\alpha+\beta^2-4\,\beta\,\alpha-\alpha^2-Q}\,.$$

Continuing with elimination polynomials, we find that

$$z_0 = -\frac{\left(\alpha^2+4\,\beta\,\alpha-4\,\alpha+2\,\alpha\,\delta+3\,\beta^2+2\,\beta\,\delta+3-Q-2\,\delta-6\,\beta\right)\delta}{-2\,\delta+5\,\beta^2+6\,\beta\,\delta+1-4\,\beta-\alpha^2+2\,\alpha\,\delta+Q-4\,\alpha\,\beta\,\delta+2\,\alpha^2\beta-2\,\beta^3-4\,\beta^2\delta-2\,\beta\,Q}$$

and

$$\begin{aligned}\tau_0 \quad = \quad & -\left((2\,\beta-1)\left(-\beta^2+2\,\beta-2\,\beta\,\delta-1+\alpha^2+2\,\delta-2\,\alpha\,\delta-Q\right)\right)\\ & \div\left(7\,\alpha+5\,\delta+9\,\beta-9\,\beta^2+5\,\beta^2\delta+5\,\alpha^2\beta-14\,\beta\,\alpha-10\,\beta\,\delta-5\,\alpha^2-8\,\alpha\,\delta-2\,\delta^2+3\,\beta^3\right.\\ & \left.+Q+\alpha^3-3+3\,\alpha^2\delta+8\,\alpha\,\beta\,\delta+2\,\delta^2\beta+2\,\alpha\,\delta^2+7\,\beta^2\alpha-\beta\,Q-\alpha\,Q-\delta\,Q\right)\,.\end{aligned}$$

When we set $\beta=\alpha$, things become a little simpler. We get

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

here,

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

This is now manageable, meaning we can take two derivatives in δ. We get a mess, several pages long, but when we evaluate it at $\delta = (1-\alpha)^2$ (i.e., $\rho = \alpha^2$), the radical becomes a polynomial and we get -8 times the following rational function of α for the second derivative.

$$\begin{aligned} & \left[-2 + 36\,\alpha - 4940\,\alpha^4 + 1472\,\alpha^3 - 672\,\alpha^{12} - 296\,\alpha^2 + 18580\,\alpha^9 \right. \\ - & \left. \quad 9424\,\alpha^{10} + 3232\,\alpha^{11} + 11816\,\alpha^5 - 20728\,\alpha^6 + 27008\,\alpha^7 + 64\,\alpha^{13} - 26146\,\alpha^8\right] / \\ & \Big/ \left[\alpha^2\left(2 - 12\,\alpha + 26\,\alpha^2 - 24\,\alpha^3 + 8\,\alpha^4\right)^2\left(-4\,\alpha + 2 + 2\,\alpha^2\right)^2\left(1 - 2\,\alpha + 2\,\alpha^2\right)(-1+\alpha)^2\right] \end{aligned} \tag{27}$$

We may verify, by sign-change rules for polynomials, that twice this plus the second derivative $1/(\alpha^2(1-\alpha)^2)$ of the entropy function is negative for all $\alpha > 2/3$. This shows $\rho = \alpha^2$ to be a local maximum for all $\alpha > 2/3$. [Our research assistant points out that if the derivatives are taken directly in ρ rather than in δ, then the second derivative comes out to be $-(2\alpha-1)/[\alpha^2(1-\alpha)^2(1-2\alpha+2\alpha^2)]$; doubling and adding the second derivative of the entropy function gives a quantity which is negative for $\alpha > (\sqrt{3}-1)/2 \approx 0.634$.]

There are several crude numeric ways to see it is a global maximum. For example one may compute an upper bound of -10 for the second derivative at $\delta = (1-\alpha)^2$ for $\alpha > 12/17$. One may then compute an upper bound for the third derivative of 1700 on the region $\alpha \le 0.847$ and any δ. Taylor's Theorem then tells us that the region in which the local maximum is global extends at least $3/170$ in the δ direction on either side of the curve $\delta = (1-\alpha)^2$. Away from this curve one may then divide the α-δ plane into grid squares sufficiently small so that when one bounds $2 \cdot$ `rate` and `ent` separately from above on the grid square, the sum of the bounds is less than $2F_1(\alpha_*)$. For details, see

`http://www.math.upenn.edu/~pemantle/CLP04-grid2.mw`

The final piece of the proof of Theorem 4.1 is the verification of (3).

From Theorem 6.1 we know that equation (12) holds with `rate` given by (23). Having maximized the smooth function F_2, we see that for a fixed n,

$$E(k, k', n, \kappa) := N(k, k', n, \kappa)^2 P(k, k', n, \kappa)$$

has a global maximum at

$$\mathbf{r} := (r_1, r_2, r_3, r_4) = (n + \mathcal{O}(1))(\alpha_*, \alpha_*, 1, \alpha_*^2)\,.$$

If $|r_i' - r_i| \le n^{1/2+\epsilon}$ for $1 \le i \le 4$ and any $\epsilon < 1/6$, then

$$E(\mathbf{r}') \sim E(\mathbf{r}) \exp(-n^{-1} Q(\mathbf{r}' - \mathbf{r}))$$

where Q is the quadratic approximation to F_2.

Checking that Q is non-degenerate, we see that we may use a Gaussian approximation to sum the values of $E(\mathbf{r}')$ to yield

$$\sum_{\mathbf{r}'} E(\mathbf{r}') = \sqrt{\frac{n^3}{(2\pi)^3 \mathcal{H}}} E(\mathbf{r}) = \mathcal{O}(\exp(2nF_1(\alpha_*))$$

with $\mathcal{H}$ denoting the determinant of the quadratic form Q. This proves Theorem 4.1. $\square$

7 A lower variance bound and a related model

Let us call a subset of $[n]$ a strong δ-interval if it intersects all of the intervals of size δ of $[n]$, including cyclic ones, e.g., $\{n, 1, \ldots, \delta - 1\}$. For $\delta = 2$, this means a strong interval must be an interval and also must intersect $\{1, 2\}$, $\{1, n\}$ and $\{n-1, n\}$. So, for example, the set $\{1, 3, 4, 6, 7, 8\}$ is always a 2-interval of $[n]$, $n \geq 8$, but is only a strong 2-interval if $n = 8$ or 9. A strong δ-interval of the permutation G is a strong interval such that I and $G^{-1}(I)$ are strong intervals of $[n]$. This is a symmetrized definition, aimed at making some of the analysis easier without fundamentally changing the problem. We may reason heuristically that the least and greatest elements of a typical interval of G will be respectively $\mathcal{O}(1)$ and $n - \mathcal{O}(1)$, and that the strong intervals of G are a positive fraction of all intervals of G, so that in some sense, the number of strong δ-intervals of a random permutation should behave like the number of δ-intervals of a random permutation.

Given a permutation G, let us look at the complement $I^c := [n] \setminus I$ of a 2-interval I of G. To say that I is a strong 2-interval of $[n]$ is exactly to say that I^c is an independent set of the cycle graph C with edges between each j and $(j+1)_{\mathrm{mod} n}$. To say that $G^{-1}(I)$ is a strong 2-interval is equivalent to saying that I^c is an independent set of the cycle graph $G(C)$ with edges between $G(j)$ and $G((j+1)_{\mathrm{mod} n})$. Let $H = H(G)$ denote graph with vertex set $[n]$ whose edges are the union of the two n-cycles C and $G(C)$. Then a strong 2-interval of G is exactly the complement of an independent set of $H(G)$. There is a natural question, analogous to the question of how many 2-intervals there are in a typical permutation, namely,

Problem: Determine the (quenched) behavior of the random number Z of independent sets of the graph $C \cup G(C)$.

This appears difficult, and the simplest heuristics misleading. One might expect that Z is asymptotically log-normal for the following reason. Adding or deleting an edge should change the number of independent sets by a factor of $\Theta(1)$. Changing an edge to a different edge should therefore also change by such a factor. The randomness in G is the result of $\Theta(n)$ random selections, thus there should be a variance of $\Theta(n)$ in $\log Z$. We know this reasoning fails for 2-intervals, so it probably fails for strong 2-intervals as well.

Indeed, one may think that for graphs of high girth and average degree d, the log of the number of independent sets is very well approximated by the following mean field heuristic. If n is the number of vertices, then the number of vertex subsets of size $k = \alpha n$ is

$$\sim C n^{-1/2} \exp\left[n\alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}\right] .$$

Such a set contains $\sim \alpha^2 n^2/2$ pairs, each being an edge of the graph with probability $\sim d/n$. A mean field heuristic would say that the probability of a k-set being independent is roughly $\exp(-\alpha^2 dn/2)$. Multiplying by the number of k-sets and taking the log gives

$$\sim \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha} - \frac{\alpha^2 d}{2} .$$

For fixed d one may optimize in α. For $d = 4$, one finds that the optimal α is roughly $0.26064\ldots$ and that the number of independent sets would then be roughly $\exp(0.43786\ldots n)$. For the random graph H this surely fails, since the number of 2-intervals is exponentially lower than this. It appears that d-regularity is a stronger constraint than average degree d. Perhaps a large family of graphs, such as those of average degree d, tends to be subject to the lottery phenomenon, with the typical log number of independent sets falling below the mean of $\approx \exp(.437n)$, while the more homogeneous family of d-regular graphs exhibits the quenched behavior even in the mean.

Amid all this speculation, let us prove that the variance of Y must be at least of order $(\mathbb{E}Y)^2$, ruling out Gaussian behavior. We give a proof for Z instead of Y because the bookkeeping is simpler, with the proof for Y being entirely analogous.

Theorem 7.1 *There exists a positive number δ such that there is no number c for which $Z \in [c, (1+\delta)c]$ with probability at least $1-\delta$. It follows that $Z/\mathbb{E}Z$, which has been shown to be tight, cannot converge to a constant in probability.*

Lemma 7.2 *Let K be any finite graph with degrees bounded by d and let μ be the probability measure which is uniform on independent sets of K. Then for any non-adjacent vertices x, y of K,*

$$\mu\{I : x, y \in I\} \geq \epsilon_d := 2^{-2d-2}\,.$$

Proof: Let K' be the subgraph of K induced on the vertices of K at distance at least 2 from both x and y. Each independent set I of K' has at most 2^{2d+2} supersets that are subsets of the vertices of K, and at least one of these is an independent set containing both x and y. □

Remark 7.3 *This is a bad bound, but on the other hand it is sharp (let K be the union of two stars). What is the right constant for d-regular graphs?*

Proof of Theorem. Let G be a permutation for which $\{i, i+1\}$ and $\{i', i'+1\}$ are both intervals, and both ascending, i.e., $G^{-1}(i) = j, G^{-1}(i+1) = j+1, G^{-1}(i') = j'$ and $G^{-1}(i'+1) = j'+1$ for some i, j, i', j'. Suppose further that $i' \geq i+3$ and that G has no other intervals of size 2. Let $\mathcal{A}$ denote the set of such permutations, G. Reasoning from Proposition 2.1, we know that $\mathbb{P}(A) = (1/4)(2e^{-2} + o(1))$. For $G \in \mathcal{A}$, define $\phi(G)$ to be the permutation G' such that $G'(j+1) = i'+1$, $G(j'+1) = i+1$, and G' agrees with G except on $j+1$ and $j'+1$. The graph $H(\phi(G))$ is the graph $H(G)$ with two extra edges, namely $\{i, i'+1\}$ and $\{i', i+1\}$. The set of independent sets of $\phi(G)$ is therefore a subset of the set of independent sets of G, namely those not containing both endpoints of either new edge. Taking just one new edge into consideration and using Lemma 7.2, we see that the number of independent sets of $\phi(G)$ is most $1 - 2^{-10}$ times the number of independent sets of G. On the other hand, the measure of the collection $\{\phi(G) : G \in A\}$ is $(1/2)e^{-4} + o(1)$ by reasoning similar to that used in the proof of Proposition 2.1. [The probability of no intervals of size 2 is e^{-2} and the probability of seeing (i, i', j, j') as above is $(1/4) \cdot 2e^{-2}$.] Choosing $\delta < 2^{-10}$ completes the proof. □

8 Conclusion

We could also have dealt with a more general problem. We could have allowed gaps of bounded size $\delta - 1$ in the positions and gaps of bounded size $\gamma - 1$ in the positions in the symbols. We call these δ, γ-intervals. We conjecture that similar results hold for these δ, γ-intervals. Let $X^{(\delta,\gamma)}$ be the number of δ, γ-intervals of G.

For example, the case $\delta = 2$ and $\gamma = 1$ would follow using our previous analysis as,

$$\mathbb{E}X^{(2,1)} = \sum_k \frac{N(n,k)(n-k+1)}{\binom{n}{k}}$$

and
$$\mathbb{E}(X^{(2,1)})^2 = \sum_{k,\ell,\kappa} N(k,\ell,n,\kappa)(n-k-\ell+\kappa+1)P(k,\ell,n,\kappa).$$

Acknowledgements

Thanks to Kate Davidson (undergraduate research assistant at The University of Pennsylvania) for help with the rigorous numerical bounds in Lemma 4.2 and the computation of the maximum in equation (27). We are also grateful to both referees, whose comments led to improvements in the presentation.

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. Convolutions of inverse linear functions via multivariate residues. *Preprint.*

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

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

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

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

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

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

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

[15] 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.

[16] R. Pemantle and M. Wilson Asymptotics of multivariate sequences, part II: multiple points of the singular variety. *Combinatorics, Probability and Computing*, 13:735–761, 2004.

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

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

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