GENERATING FUNCTIONS AND SUBSETS OF $\mathbb{N}$ WITH BOUNDED SPACING

ROHAN BULUSWAR

Abstract. This paper seeks to utilize the theory of symbolic combinatorics to analyze a specific problem involving choosing integers with bounded distance from each other. First, it introduces constructions on combinatorial classes and their associated actions on ordinary generating functions. Second, it uses these ideas to construct a class corresponding to the problem, find its generating function, and unpack it to calculate its coefficients. Finally, it provides a partial argument about the rate of growth of the expected value of the random variable defined by the 'bounded spacing' problem.

Contents

1. Introduction

Suppose that from the set $A = \{1, 2, ..., n\}$, we choose at random from a uniform distribution a subset $B \subset A$ such that $|B| = k$, where $n, k \in \mathbb{N}$ and $k \leq n$. (Note: from here we use $[1, n]$ in place of $\{1, 2, .., n\}$.)We can then sort B and represent it as $\{b_1, b_2, ..., b_k\}$ where $b_1 < b_2 < ... < b_k$. From here we can define $C = \{b_2 - b_1, b_3 - b_2, ..., b_k - b_{k-1}\}$. This new set yields a random variable $d = \max(C)$. In other words, we are interested in the maximal distance between any two consecutive elements of our subset. This construction may raise a number of problems related to the probability distribution of d, but the first and foremost is to find a way to calculate $P(d = i)$ for some constant i and fixed values of n and k. The solution to this problem is the main focus of this paper and will be used to introduce two key concepts in combinatorics: symbolic constructions and generating functions.

We can begin by slightly altering our goal, for reasons that will become clearer later: rather than seeking to calculate $P(d = i)$, we will try to understand $P(d \leq i)$. There is ultimately little difference since $P(d = i) = P(d \leq i) - P(d \leq i - 1)$. We will denote by $S(n, k, i)$ the number of subsets of $[1, n]$ with size k and maximal distance less than or equal to i. Since we are choosing the random subset uniformly, understanding $P(d \leq i)$ is in essence equivalent to understanding $S(n, k, i)$. Specifically, we find

$$P(d \leq i) = \frac{1}{\binom{n}{k}} S(n, k, i).$$

The first step is to find a new way to represent the problem. Specifically, instead of choosing k elements of $[1, n]$, we can think of a sequence of n chosen and unchosen numbers. That is, if a represents an unchosen number and b represents a chosen number, we have a string of k bs and $n - k$ as where d is one more than the longest stretch of consecutive as between two bs. For example, $baab$, with a maximum of two as in a row, corresponds to picking $\{1, 4\}$ from $[1, 4]$, which have a distance of three. In order to gain some value from this new interpretation of our problem, we need to state it in a more formal framework. To do this, we can turn to symbolic combinatorics.

2. Symbolic Constructions in Combinatorics

The symbolic approach to combinatorics and generating functions, laid out by Phillipe Flajolet and Robert Sedgewick in their book "Analytic Combinatorics," is they key to solving this problem. In this section, I seek to explain the relevant definitions and lemmas, and fill in some of the details of proofs left out of the book. Moreover, we will make the first of two important steps towards solving our central problem.

We can begin by giving a definition of a combinatorial class:

Definition 2.1. A (combinatorial) class $\mathcal{W}$ is a countable set and a size function denoted by $|\cdot|$ with the following two properties: $\forall \omega \in \mathcal{W}$, $|\omega|$ is a nonnegative integer, and for all $n \in \mathbb{N}_0$, there exist finitely many $\omega \in \mathcal{W}$ such that $|\omega| = n$.

Now we can introduce some useful notation. Typically, we denote the number of elements in $\mathcal{W}$ with size n by W_n. Moreover, we denote the set of elements in $\mathcal{W}$ with size n by $\mathcal{W}_n$. That is, $|\mathcal{W}_n| = W_n$. Per the second condition in the above definition, W_n must be a nonnegative integer. Additionally, when we have an object

ω that may belong to multiple classes, we specify $|\omega|_{\mathcal{W}}$ to be the size of ω in $\mathcal{W}$. From here, we can introduce one of the most important ideas in combinatorics:

Definition 2.2. The ordinary generating function (OGF) of a combinatorial class $\mathcal{W}$ is given by

$$W(z) = \sum_{n=0}^{\infty} W_n z^n = \sum_{\omega \in \mathcal{W}} z^{|w|}$$

It is important to note that to begin with, we are treating these generating functions as algebraic objects, rather than analytic ones. This distinction means that we do not yet need to worry about whether these sums converge for all (or even any) real or complex inputs. Instead, the z is just a placeholder variable, and we will think of generating functions as formal power series over $\mathbb{Z}$. Specifically, Benjamin Sambale provides the following definition [2]:

Definition 2.3. A formal power series over K is an infinite sequence $\alpha = (a_0, a_1, ...)$ where $a_i \in K$. The set of all power series over K forms a vector space and is denoted by $K[[X]]$, with addition and scalar multiplication performed componentwise as usual. That is, if $\alpha = (a_0, a_1, ...)$, $\beta = (b_0, b_1, ...)$, and $\lambda \in K$, then

- $\alpha + \beta = (a_0 + b_0, a_1 + b_1, a_2 + b_2, ...)$
- $\lambda\alpha = (\lambda a_0, \lambda a_1, \lambda a_2 ...)$

In Sambale's definition [2], K is a field. However, he notes that for all of the results we will need, an integral domain will suffice. Thus, we can continue using $\mathbb{Z}$. Next, we want to define multiplication of two power series. To do that, we can introduce some notation: $1 = (1, 0, 0, ...)$, $X = (0, 1, 0, 0, ...)$, $X^2 = (0, 0, 1, 0, 0, ...)$, etc. Then we take the standard convolution formula for polynomials:

Definition 2.4. If $\alpha = (a_0, a_1, ...)$ and $\beta = (b_0, b_1, ...)$, then we say

$$\alpha \cdot \beta := \sum_{n=0}^{\infty} (\sum_{i=0}^{n} a_i b_{n-i}) X^n$$

Now that we have an elementary understanding of how to work with generating functions as formal power series, we can return to combinatorial classes. The main idea of this part of the paper is that we can build up combinatorial classes from their components, and in parallel use old generating functions to compute new ones.

Proposition 2.5. *Given disjoint combinatorial classes $\mathcal{A}$ and $\mathcal{B}$, let $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ be their union. If we define the size function on C by $|\omega|_{\mathcal{C}} = |\omega|_{\mathcal{A}}$ if $\omega \in \mathcal{A}$ and $|\omega|_{\mathcal{C}} = |\omega|_{\mathcal{B}}$ if $\omega \in \mathcal{B}$, then $\mathcal{C}$ is a valid combinatorial class with generating function $C(z) = A(z) + B(z)$.*

Proof. The proof follows almost directly from the definitions, and is thus omitted. □

Now that we have seen that addition of generating functions corresponds to the union of disjoint classes, we might wonder if there is a set theoretic operation that corresponds to multiplication of generating functions. As it happens, there is.

Proposition 2.6. *Given combinatorial classes $\mathcal{A}$ and $\mathcal{B}$ (not necessarily disjoint), let $\mathcal{C} = \mathcal{A} \times \mathcal{B}$ be their cartesian product. That is, $\mathcal{C} = \{(\alpha, \beta) | \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$. For all $\gamma = (\alpha, \beta) \in \mathcal{C}$, let $|\gamma|_{\mathcal{C}} = |\alpha|_{\mathcal{A}} + |\beta|_{\mathcal{B}}$. Then $\mathcal{C}$ defines a valid class with generating function $C(z) = A(z) \cdot B(z)$.*

Proof. Suppose we have some $\gamma = (\alpha, \beta)$ such that $|\gamma| = n$. Then it could be that $|\alpha| = 0$ and $|\beta| = n$, or $|\alpha| = 1$ and $|\beta| = n - 1$, etc. In other words, we find $\mathcal{C}_n = (\mathcal{A}_0 \times \mathcal{B}_n) \cup (\mathcal{A}_1 \times \mathcal{B}_{n-1}) \cup ... \cup (\mathcal{A}_n \times \mathcal{B}_0)$, such that these unions are all disjoint. Hence, we have the following:

$$|\mathcal{C}_n| = |(\mathcal{A}_0 \times \mathcal{B}_n) \cup (\mathcal{A}_1 \times \mathcal{B}_{n-1}) \cup (\mathcal{A}_2 \times \mathcal{B}_{n-2}) \cup ... \cup (\mathcal{A}_n \times \mathcal{B}_0)|$$

$$|\mathcal{C}_n| = \sum_{i=0}^{n} |\mathcal{A}_i \times \mathcal{B}_{n-i}|$$

$$C_n = \sum_{i=0}^{n} A_i B_{n-i}$$

$$C(z) = \sum_{n=0}^{\infty} (\sum_{i=0}^{n} A_i B_{n-i}) z^n = A(z) \cdot B(z)$$

Finally, we must verify that $\mathcal{C} = \mathcal{A} \times \mathcal{B}$ defines a valid combinatorial class. First, consider $\gamma = (\alpha, \beta)$ where $\alpha \in \mathcal{A}$, $\beta \in \mathcal{B}$, and $\gamma \in \mathcal{C}$. Then because $\mathcal{A}$ and $\mathcal{B}$ are combinatorial classes, $|\alpha|$ and $|\beta|$ are nonnegative integers, so $|\alpha| + |\beta| = |\gamma|$ is a nonnegative integer, and the first condition is satisfied. Next, we must show that C_n is finite. Once again, for all i, A_i and B_{n-i} are just nonnegative integers, so $A_i B_{n-i}$ is a nonnegative integer. Then, C_n is just a finite sum of nonnegative integers, and is thus itself a nonnegative integer. Thus, $\mathcal{C}$ is a valid combinatorial class. □

The final important basic combinatorial construction is sequences.

Definition 2.7. Given a combinatorial class $\mathcal{A}$, we define $\mathcal{B} = \mathrm{SEQ}(\mathcal{A})$ as the set of all n-tuples of $\mathcal{A}$, where n ranges over all nonnegative integers. Formally, we say that

$$\mathcal{B} = \epsilon + \mathcal{A} + \mathcal{A} \times \mathcal{A} + \mathcal{A} \times \mathcal{A} \times \mathcal{A} + ...$$

Here, ϵ represents the empty sequence: a class with a single element of size zero, and generating function 1, or when written as a formal power series, (1,0,0,...). Because sequences are simply defined as the disjoint union of cartesian products, they inherit the definition of the size function given in Proposition 2.5 and Proposition 2.6.

Specifically, if $\alpha_i \in \mathcal{A}$ for $1 \leq i \leq k$ and $\beta \in \mathcal{B}$, then we have

$$|\beta| = |(\alpha_1, \alpha_2, ..., \alpha_k)| = |\alpha_1| + |\alpha_2| + ... + |\alpha_k|$$

Once again, we must determine whether $\mathcal{B} = \mathrm{SEQ}(\mathcal{A})$ is in fact a valid combinatorial class given this definition of its and its size function.

Proposition 2.8. *If $\mathcal{A}$ is a class, then $\mathcal{B}$ = SEQ($\mathcal{A}$) defines a valid class if and only if $A_0 = 0$ (i.e. $\mathcal{A}$ contains no elements of size zero).*

Proof. First, we can check that $|\beta| = |(\alpha_1, ..., \alpha_k)| \in \mathbb{N}_0$ for all $\beta \in \mathcal{B}$. This property is immediate, since $|\beta| = \sum |\alpha_i|$ is simply a finite sum of nonnegative integers, and $\mathbb{N}_0$ is in fact closed under addition. The second condition, that there exist finitely many $\beta \in \mathcal{B}$ with $|\beta| = n$ for any $n \in \mathbb{N}_0$, is slightly trickier.

We will show that it is this condition that is satisfied exactly whenever $A_0 = 0$, i.e. $\mathcal{A}$ contains no elements of size zero. First, suppose that there exists $\alpha \in \mathcal{A}$ with $|\alpha| = 0$. Then $0 = |(\alpha)|, |(\alpha, \alpha)|, |(\alpha, \alpha, \alpha)|, ...$ Thus, we have found infinitely many elements of $\mathcal{B}$ with size 0, and the construction is not valid.

Second, suppose that in fact $A_0 = 0$. Then we want to show that for any nonnegative n, B_n is finite. Let i be the smallest positive integer such that $A_i \neq 0$. In other words, i is the smallest size of an element in A. Because $A_0 = 0$, $i > 0$. Because i is the smallest size of an element in A, any l-tuple of elements in A has size at least li. Therefore, because i is positive, the sequence li grows without bound and must eventually be greater than n. In other words, there must exist some m such that if $l > m$, then all elements of $\mathcal{A}^l$ have size greater than n, and $A^l_n = 0$. As we will see later, in the terminology of formal power series, we have proven that the sequence of generating functions of $\mathcal{A}^i$ is a null sequence. Recall that $\mathcal{B} = \epsilon + \mathcal{A} + \mathcal{A}^2 + \mathcal{A}^3 + ...$ Hence, $B_n = \epsilon_n + A_n + A^2_n + ...$ Because the cartesian product always produces a valid combinatorial class, and due to the second key property of classes, we know that A^i_n is finite for all $i \geq 0$. However, as we just proved above, this infinite sum only has finitely many non-zero terms (at most until A^m_n), and is thus itself finite. Hence, $\mathcal{B}$ is in fact a valid combinatorial class. □

Now we arrive at the question of whether or not the generating function of $\mathcal{B} = \mathrm{SEQ}(\mathcal{A})$, $B(z)$, is easily expressed in terms of $A(z)$. Because the sequence construction involves an infinite union of classes, it is clear that we will need to grapple with infinite sums of power series. However, in order to have a rigorous discussion of the convergence of such series, we must introduce a norm. We can begin with a brief intermediary definition upon which we construct a norm (both are provided by Sambale [2]).

Definition 2.9. Given a formal power series $A(z) = (\alpha_0, \alpha_1, \alpha_2, ...)$, we say $\inf(A(z)) := \min(\{n \in \mathbb{N}_0 : \alpha_n \neq 0\})$. By convention, we say $\inf(0) = \infty$.

In other words, we are checking how many consecutive zeroes appear at the beginning of the sequence representation of $A(z)$. In the context of combinatorial classes, $\inf(A(z))$ is just the size of the smallest element of $\mathcal{A}$. This alternative characterization of the infimum of a generating function is worth noting as it will be useful later. Next,

Definition 2.10. Given a formal power series $A(z)$, we define $||A(z)|| := 2^{-\inf(A(z))}$. Again, by convention, we say that $||0|| = 0$.

Now we have a norm and know that we are interested in infinite sums. As every calculus student studying infinite series knows, the first test of convergence is checking that the limit of the terms is zero. Therefore, we introduce the following definition:

Definition 2.11. A sequence of formal power series $F_0(z), F_1(z), F_2(z), ...$, is called a null sequence if $\lim_{i\to\infty} ||F_i(z)|| = 0$.

Given an infinite sum of formal power series, we might expect that if it is to converge, it is necessary for the absolute values of the terms to approach zero. This is indeed the case. However, somewhat unexpectedly, the converse is also true. That is, it is also sufficient to check that the limit of the absolute values of the terms is zero in order to prove convergence of the series. Specifically, we have the following statement, a proof of which is given by Sambale in his Lemma 2.10 [2].

Lemma 2.12. *Given a sequence of formal power series $F_0(z), F_1(z), F_2(z), ...$, if $\lim_{i\to\infty} ||F_i(z)|| = 0$, then the associated infinite series*

$$\sum_{i=0}^{\infty} F_i(z)$$

is convergent.

Finally, we have all the tools necessary to approach the problem of the generating functions of sequence constructions.

Proposition 2.13. *If $\mathcal{A}$ is a class and $\mathcal{B} = SEQ(\mathcal{A})$, then $B(z) = \frac{1}{1-A(z)}$.*

Proof. From the statements of Proposition 2.5, Proposition 2.6, and Definition 2.7, we get immediately $B(z) = 1 + A(z) + A(z)^2 + A(z)^3 + ...$ This is just a geometric series of $A(z)$, so all we have to show is that the familiar formula

$$\frac{1}{1-A(z)} = 1 + A(z) + A(z)^2 + A(z)^3 + ...$$

still holds.

However, it is important to remember here that by equality, we mean equality as a formal power series (i.e. the same sequence of coefficients), and not in terms of values when plugging in z. This equality is not a trivial fact and must be carefully proven.

The first step is to check that the sum $1 + A(z) + A(z)^2 + A(z)^3 + ...$ is itself a formal power series, i.e. converges. We will want to make use of Lemma 2.12, so we just need to show that the sequence $(A(z)^i)$ is a null sequence. That is, we need to prove

$$\lim_{i\to\infty} ||A(z)^i|| = 0.$$

Because the norm of a formal power series is just a real number, we can use the normal definition of limits of sequences. Specifically, we need to show that for any $\varepsilon > 0$, there exists some natural number N such that for all $i > N$, $|A(z)^i| < \varepsilon$. We can begin by fixing $\varepsilon > 0$. Then because ε is positive, it is clear that there must exist some $n \in \mathbb{N}$ such that $2^{-n} < \varepsilon$. Then by Definition 2.10, we just need to find N such that for all $i > N$, $\inf(A(z)^i) \geq n$. Recalling our alternative characterization of the infimum of a generating function, we need to find some N such that for all $i > N$, the size of the smallest element of $\mathcal{A}^i$ is at least n.

Thankfully, we essentially exactly proved that we can do this in the proof of Proposition 2.8. However, we can revisit it here. Let d be the size of the smallest element in $\mathcal{A}$. By assumption, $d > 0$. Then it follows that id is the size of the smallest element of $\mathcal{A}^i$. Our goal, then, is to find some N such that $di \geq n$ whenever $i > N$. Now it should be clear that we can accomplish this by setting $N = \lceil \frac{n}{d} \rceil$. Hence, the sequence of powers of $A(z)$ is in fact a null sequence, and their sum is therefore well defined.

Moreover, multiplication by a constant does not change the properties of convergence, so we can define a formal power series

$$F(z) = (1 - A(z)) \sum_{i=0}^{\infty} A(z)^i.$$

As we might expect, sums distribute over limits, so we can say that

$$F(z) = \sum_{i=0}^{\infty} A(z)^i - A(z) \sum_{i=0}^{\infty} A(z)^i,$$

which can be rewritten as

$$F(z) = \sum_{i=0}^{\infty} A(z)^i - \sum_{i=1}^{\infty} A(z)^i.$$

Moreover, removing finitely many terms does not affect convergence, so we can say that

$$F(z) = A(z)^0 + \sum_{i=1}^{\infty} A(z)^i - \sum_{i=1}^{\infty} A(z)^i$$
$$F(z) = 1 + 0 = 1$$

Therefore, we have

$$1 = (1 - A(z)) \sum_{i=0}^{\infty} A(z)^i,$$

Lemma 2.5 from [2] states that a formal power series has a unique multiplicative inverse iff its first coefficient is nonzero. Because the first coefficient of $A(z)$ is assumed to be zero (otherwise the sequence construction would be invalid), $1-A(z)$ must have a first coefficient of 1 and hence a unique inverse. Therefore, it is legal to multiply both sides of the equation by $\frac{1}{1-A(z)}$ to get the desired result:

$$\frac{1}{1 - A(z)} = \sum_{i=0}^{\infty} A(z)^i$$

This is exactly what we wanted to show, and the proof is complete. □

While the sequence construction was the last fundamental operation necessary to proceed with tackling the problem set out in the introduction, we will find ourselves in need of one more. That is the finite sequence, which we can define as follows.

Definition 2.14. Given a combinatorial class $\mathcal{A}$, we say that

$$\mathrm{SEQ}_{<k}(\mathcal{A}) = \sum_{i=0}^{k-1} \mathcal{A}^i$$

The last step in building up the theory we need is to determine the generating function of finite sequences.

Proposition 2.15. *Suppose $\mathcal{A}$ is a class and let $\mathcal{B} = SEQ_{<k}(\mathcal{A})$. Then*

$$B(z) = \frac{1 - A(z)^k}{1 - A(z)}.$$

Proof. From the elementary properties of generating functions laid out in Proposition 2.5 and Proposition 2.6, we see immediately that

$$\begin{aligned}
B(z) &= 1 + A(z) + ... + A(z)^{k-1} \\
B(z)(1 - A(z)) &= (1 - A(z))(1 + A(z) + ... + A(z)^{k-1}) \\
B(z)(1 - A(z)) &= \sum_{i=0}^{k-1} A(z)^i - A(z) \sum_{i=0}^{k-1} A(z)^i \\
B(z)(1 - A(z)) &= \sum_{i=0}^{k-1} A(z)^i - \sum_{i=1}^{k} A(z)^i \\
B(z)(1 - A(z)) &= 1 + \sum_{i=1}^{k-1} A(z)^i - \sum_{i=1}^{k-1} A(z)^i - A(z)^k
\end{aligned}$$

$$B(z) = \frac{1 - A(z)^k}{1 - A(z)}$$

□

Now that we have a dictionary about operations on classes and what they do to their respective generating functions, we can build up to the class that solves our problem through constructions on simple ones, and thus access its generating function. From there, we can unpack it to determine its coefficients, and thus solve the central problem.

3. The Generating Function for Bounded Spacings

3.1. **Constructing the Generating Function.** The purpose of this section is to explain in detail how to apply the theory of symbolic combinatorics to solve the problem detailed in the introduction. While Flajolet and Sedgewick do include the final solution to this problem in [1], this section will once again seek to fill in the details that they left to the reader. To begin, we can briefly review the problem statement itself and the secondary characterization that will allow us to define a combinatorial class.

We begin by fixing natural numbers n and k where $k \leq n$. Next, we can define a random variable d in the following manner: begin by choosing a subset of size k from $\{1, 2, ..., n\}$ at random, with a uniform distribution. Then if we sort this subset, d is defined to be the maximum distance between any two consecutive elements. We want to understand the probability distribution of d for a given n and k. One way to do this, since d is a discrete random variable, is to find an expression for $P(d = i)$ where i is some constant. As it happens, it will be easier to first calculate $P(d \leq i)$. Since we are choosing from a uniform distribution, this is equivalent to finding the number of ways to pick k elements from $[1, n]$ with a maximal distance of at most i. We denote this number by $S(n, k, i)$, which we will ultimately find an explicit formula for.

Next, we can review the second way of viewing the problem discussed in the introduction. Specifically, picking k integers from $[1, n]$ is the same as classifying each integer from 1 to n as either 'chosen' or 'unchosen', and ensuring that we only have k that are 'chosen'. If we write an unchosen number as a and a chosen number as b, what we have is just a word or string of k 'b's and $n - k$ 'a's. The second condition of the problem is that any two consecutive chosen integers have a difference of no more than i. In order to satisfy this requirement, we just have to ensure that any string of consecutive 'a's between two 'b's is strictly less than i. Because the distance between any two chosen numbers is one more than the amount of unchosen numbers between them, the inequality here must be strict to correspond to the unstrict inequality in the definition of $S(n, k, i)$. For example, zero 'a's corresponds to two consecutive integers, which have a difference of one.

To begin formally defining the class that solves our problem, we can start with the two basic classes $\mathcal{A}$ and $\mathcal{B}$. Each has just one element, with size one: a and b, which we will use to build our words. We assign our letters size one so that the size of our words is just n. Hence, $A(z) = B(z) = z$. Now we are left with the problem of how to use $\mathcal{A}$ and $\mathcal{B}$ to construct $\mathcal{W}(k,i)$ where $W(k,i)_n = S(n,k,i)$.

First, we need some sequence of unchosen numbers (although this may be zero). This sequence is not limited by i because it is not *between* two chosen numbers, just before the first one. Therefore, we start with $SEQ(\mathcal{A})$. Next, we have a chosen number, which corresponds simply to $\mathcal{B}$. Once again, we have some sequence of unchosen numbers. This time, however, it must be less than i in length. This corresponds to our finite sequence construction, $SEQ_{<i}(\mathcal{A})$. Because the elements from these classes are placed one after another, we are taking one from each, i.e. tuples. Therefore, the appropriate way to combine them is the cartesian product. Our progress so far, then, is

$$SEQ(\mathcal{A}) \times \mathcal{B} \times SEQ_{<i}(\mathcal{A})$$

The last two pieces are repeated once for each chosen number, with the exception of the last one. Again, because the sequence of 'a's following the *last* 'b' is not between two of them, there is no limit on its length. Therefore, we have that

$$\mathcal{W}(k,i) = SEQ(\mathcal{A}) \times (\mathcal{B} \times SEQ_{<i}(\mathcal{A}))^{k-1} \times \mathcal{B} \times SEQ(\mathcal{A})$$

Referring back to Proposition 2.6, Proposition 2.13, and Proposition 2.15, we can now quickly access the generating function of $\mathcal{W}(k,i)$:

$$W(z) = (\frac{1}{1-z})(z^{k-1})(\frac{1-z^i}{1-z})^{k-1}(z)(\frac{1}{1-z})$$

$$W(z) = \frac{z^k(1-z^i)^{k-1}}{(1-z)^{k+1}}$$

Note: although the OGF of $\mathcal{W}(k,i)$ would usually be denoted by $W(k,i)(z)$, it is shortened here to $W(z)$ to avoid awkward notation, and it is understood that it takes k and i as parameters.

3.2. **Analyzing the Generating Function.** Now that we have gone to all this work of building up the theory of combinatorial classes and constructions in order to access the generating function of $\mathcal{W}(k,i)$, one might reasonably wonder how we are any closer to solving the original problem. It may seem that the notion of generating functions, while interesting to consider in relation to constructions, do not provide any new information. This could not be further from the truth. Rather, the generating function is one of the most (if not the most) important and useful ideas in combinatorics. Generating functions are interesting both in their algebraic sense as formal power series, as this paper has focused on, and for their analytic properties. Flajolet and Sedgewick in [1] explore in more depth the use of complex analysis to better understand the power of generating functions.

In our case, the answer is that we must now find another way to express the coefficients of $W(z)$ when written as a Taylor series. First, let's introduce a brief piece of notation. Given a formal power series $W(z) = \omega_0 + \omega_1 z + \omega_2 z^2 + ...$, we say that $[z^n](W(z)) = \omega_n$. This is read as "the n-th coefficient of $W(z)$ is equal to ω_n." When possible to avoid ambiguity, for the sake of ease of reading, it may instead be written as just $[z^n]W(z)$. In other words, it is just a way to express the n-th coordinate of the expression of $W(z)$ as a formal power series. For example, $[z^3](z - 2z^3) = -2$, $[z^2](z - 2z^3) = 0$, and $[z](z - 2z^3) = 1$.

The first important observation to make is that $[z^a]F(z) = [z^{a+b}](z^b F(z))$. This is because if we have that

$$F(z) = f_0 + f_1 z + f_2 z^2 + ... + f_a z^a + ...,$$

then

$$z^b F(z) = 0 + 0z + ... + f_0 z^b + ... + f_a z^{a+b} + ...,$$

which gives us that $[z^a]F(z) = f_a = [z^{a+b}](z^b F(z))$. This property allows us to make the first step towards unpacking $[z^n]W(z)$:

$$[z^n]W(z) = [z^n]\left(\frac{z^k(1-z^i)^{k-1}}{(1-z)^{k+1}}\right) = [z^{n-k}]\left(\frac{(1-z^i)^{k-1}}{(1-z)^{k+1}}\right) = [z^{n-k}]\left(\frac{W(z)}{z^k}\right)$$

Now we can focus on understanding the coefficients of $\frac{(1-z^i)^{k-1}}{(1-z)^{k+1}}$. Because we know how to handle multiplication of formal power series, we can first examine $(1-z^i)^{k-1}$ and $(\frac{1}{1-z})^{k+1}$ separately. First, for $(1-z^i)^{k-1}$. We can notice that in the expansion of this binomial, we are only multiplying terms of z^0 and z^i. Therefore, the only terms with nonzero coefficients will be those with exponents that are divisible by i. In order to find those exponents, we can simply apply the binomial formula. (No justification of this theorem is given here as it is generally quite well-known. However, any interested readers can find a proof of the binomial theorem in the language of formal power series in Theorem 4.1 of [2].) Specifically,

$$[z^{ji}](1-z^i)^{k-1} = (-1)^j \binom{k-1}{j},$$

where we say that $\binom{x}{y} = 0$ if $x < y$, and all other coefficients are zero. In other terms,

$$(1-z^i)^{k-1} = \sum_{a=0}^{\infty} (-1)^a \binom{k-1}{a} z^{ai}$$

Now we move on to the second piece of the puzzle: $(\frac{1}{1-z})^{k+1}$. By substituting $A(z) = z$ from the proof of Proposition 2.13, we get the familiar geometric series

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + ... = \sum_{n=0}^{\infty} z^n$$

Recalling the definition of power series multiplication given in Definition 2.4, we can see that

$$\left(\frac{1}{1-z}\right)^2 = \sum_{n=0}^{\infty} \left(\sum_{i=0}^{n} 1 \cdot 1\right) z^n = \sum_{n=0}^{\infty} (n+1) z^n$$

While the pattern is not yet obvious, we can continue to apply the definition of multiplication and see where it leads. In particular,

$$\left(\frac{1}{1-z}\right)^3 = \left(\frac{1}{1-z}\right)^2 \left(\frac{1}{1-z}\right) = \sum_{n=0}^{\infty} \left(\sum_{i=0}^{n} (i+1) \cdot 1\right) z^n = \sum_{n=0}^{\infty} \left(\sum_{i=1}^{n+1} i\right) z^n.$$

By applying the well known formula that

$$\sum_{i=1}^{n+1} = \frac{(n+2)(n+1)}{2} = \binom{n+2}{2},$$

we can simplify this expression to

$$\left(\frac{1}{1-z}\right)^3 = \sum_{n=0}^{\infty} \binom{n+2}{2} z^n.$$

Now we can recognize that

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n = \sum_{n=0}^{\infty} \binom{n}{0} z^n$$

$$\left(\frac{1}{1-z}\right)^2 = \sum_{n=0}^{\infty} (n+1) z^n = \sum_{n=0}^{\infty} \binom{n+1}{1} z^n$$

$$\left(\frac{1}{1-z}\right)^3 = \sum_{n=0}^{\infty} \binom{n+2}{2} z^n$$

Therefore, it is reasonable to guess that

$$\left(\frac{1}{1-z}\right)^a = \sum_{n=0}^{\infty} \binom{n+a-1}{a-1} z^n.$$

But how could we prove this? Given that the problem involves combinations and sums, induction would be a reasonable approach, and would likely succeed. In this case, however, we can actually use our theory of combinatorial constructions to arrive at a justification that is likely more insightful, and perhaps quicker.

To begin, we might try to think of a class that would have $\frac{1}{1-z}$ as its generating function, and then take its cartesian product with itself a times, i.e. choose a-tuples. We know by now that $\frac{1}{1-z} = 1 + z + z^2 + ...$, so we are interested in a set of objects where there is one of each nonnegative size (0,1, etc.). The most natural choice is just those numbers themselves, i.e. the nonnegative integers. Then we are trying to find the number of a-tuples of nonnegative integers that have size n. Since we add the sizes of elements of tuples, we are just asking how many ways we can choose a nonnegative integers that add to n. This is a well-known problem in combinatorics with a straightforward argument, often referred to as 'Stars and Bars'. The result of the 'Stars and Bars' problem is exactly the quantity we found/guessed earlier: $\binom{n+a-1}{a-1}$. A proof is not discussed here, but one can be found on page 38 of [3]. Replacing a with $k+1$, we get that

$$\left(\frac{1}{1-z}\right)^{k+1} = \sum_{n=0}^{\infty} \binom{n+k}{k} z^n$$

Now that we have the series expansion of each term in our product, we can once again make use of Definition 2.4 to access the coefficients of $\frac{W(z)}{z^k}$, and therefore those of $W(z)$ itself. First, the following lines provide a brief recap of our argument up to this point

$$S(n,k,i) = [z^n](W(z)) = [z^n](\frac{z^k(1-z^i)^{k-1}}{(1-z)^{k+1}})$$
$$[z^n](\frac{z^k(1-z^i)^{k-1}}{(1-z)^{k+1}}) = [z^{n-k}](\frac{(1-z^i)^{k-1}}{(1-z)^{k+1}})$$

As mentioned earlier, we now apply Definition 2.4 to this most recent expression

$$[z^a]\left(\frac{(1-z^i)^{k-1}}{(1-z)^{k+1}}\right) = \sum_{n=0}^{a}([z^{a-n}](\frac{1}{1-z})^{k+1})\cdot([z^a](1-z^i)^{k-1})$$

While this may seem like a formidable expression, it's just taking a convolution of two sequences that we already know. First, to simplify, we can recall that $[z^a](1-z^i)^{k-1} = 0$ unless a is a multiple of i. These terms will contribute nothing to the sum, so we can restrict to just multiples of i. Hence, we have

$$[z^a]\left(\frac{W(z)}{z^k}\right) = \sum_{j=0}([z^{a-ij}](\frac{1}{1-z})^{k+1})\cdot([z^{ij}](1-z^i)^{k-1})$$

Although we could at this point give an expression for how many terms there will be in this sum, that will not be a straightforward question about the final solution. Therefore, the upper bound of the sum is left blank with the understanding that it is only sensible for a finite number of terms. In this case, that it is until $ij > a$. As we will soon see, this will be clearly reflected in the expressions themselves, as long as we again accept that $\binom{x}{y} := 0$ whenever $x < y$. Now, at last, we can substitute the formulas previously found for these coefficients. This gives us that

$$[z^a]\frac{W(z)}{z^k} = \sum_{j=0}\binom{a-ij+k}{k}(-1)^j\binom{k-1}{j}$$

Finally, substituting $n-k$ for a, we get that

$$S(n,k,i) = [z^n]W(z) = \sum_{j=0}(-1)^j \binom{k-1}{j}\binom{n-k-ij+k}{k}$$
$$S(n,k,i) = \sum_{j=0}(-1)^j \binom{k-1}{j}\binom{n-ij}{k}$$

Here, the sum continues until $k-1<j$ or $n-ij<k$. Because of these simultaneous conditions, it is more difficult to determine a general formula in n,k, and i for the number of summands. Nonetheless, this expression is the answer we have been looking for. At last, we have that

$$P(d \leq i) = \frac{1}{\binom{n}{k}} \sum_{j=0}(-1)^j \binom{k-1}{j}\binom{n-ij}{k},$$

and

$$P(d=i) = P(d \leq i+1) - P(d \leq i)$$
$$P(d=i) = \frac{1}{\binom{n}{k}}\left(\sum_{j=0}(-1)^j \binom{k-1}{j}\binom{n-(i+1)j}{k} - \sum_{j=0}(-1)^j \binom{k-1}{j}\binom{n-ij}{k}\right)$$

While this last expression is not particularly nice, it is important to remember that the two sums are both finite, potentially with a different number of terms. Therefore, we cannot, in fact, simply combine them into one. Nonetheless, having accessed the probability distribution of d opens up a wide variety of new properties and ideas about it to analysis.

4. The Rate of Growth of the Expected Value of d

The purpose of this section is to make a conjecture related to the expected value of d and give some justification for it, although we unfortunately cannot provide a complete proof. *To the best of my knowledge*, this proposition has not already been discussed in any mathematical literature, i.e. may be original. That is one reason why this paper unfortunately cannot render any absolute verdict on its truth.

Recall that d as a random variable has parameters of n and k. To start, let's fix some k. Because d is a random variable, we can consider its expected value, $\mathbb{E}(d)$. In this case, we will want to consider how this expected value varies with n, so we will specify the expected value of d given a certain n as $\mathbb{E}(d_n)$. A quick visual analysis of $\mathbb{E}(d_n)$ vs. n will reveal an interesting property - it is very nearly linear in growth no matter which k is chosen (although the slope does depend on k). In particular, it invites the following statement:

Conjecture 4.1. *For any fixed k,*

$$\lim_{n\to\infty} \mathbb{E}(d_{n+1}) - \mathbb{E}(d_n) = c_k,$$

where c_k is some constant (a positive real number) depending only on k.

There are three main parts to this section. First, we will briefly present graphs of $\mathbb{E}(d_n)$ vs. n for a few different values of k to motivate the conjecture, and also show c_k vs k, using estimated values of c_k. Secondly, we will present a proof of the conjecture for the simplest case: $k = 2$. Finally, we will prove that $\mathbb{E}(d_{n+1}) - \mathbb{E}(d_n)$ does not grow without bound or blow up to infinity, but at least can be bounded by constants only depending on k.

4.1. **Numerical Evidence.** Now that we have a precise formula for $P(d = i)$, it is possible to write a program in Python to precisely compute $\mathbb{E}(d)$ for any n and k. In that way, we can also easily construct some graphs that show the exact value of $\mathbb{E}(d)$ as a function of n. Each of the four graphs that follow display forty data points, with a gap of 10 between consecutive values of n.

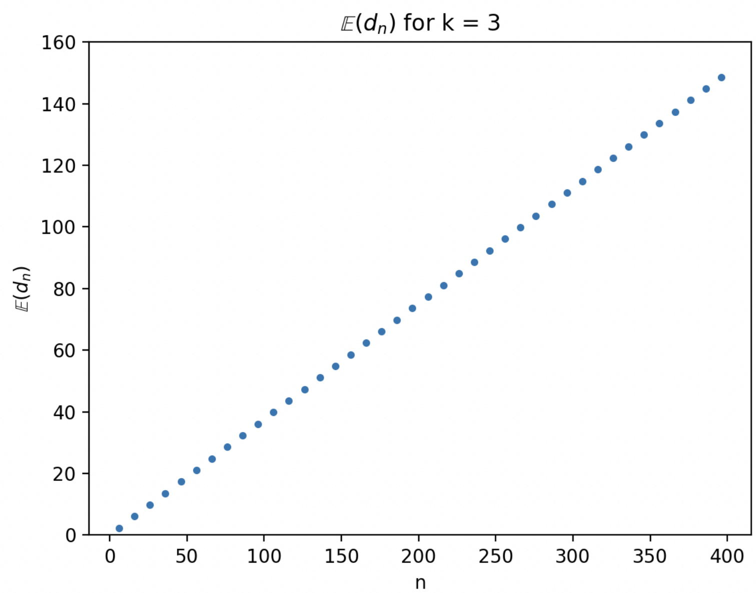

FIGURE 1. $\mathbb{E}(d_n)$ vs. n for $k = 3$

There are a few key takeaways from looking at Figure 1, Figure 2, Figure 3, Figure 4 together. First, and most importantly, is that they all appear strikingly linear. As k and n grow large, these computations take increasingly longer to complete, so these graphs of course give a very small snapshot of the possibilities. k, after all, can be any natural number. Nonetheless, they provide motivation that Conjecture 4.1 may be an interesting question to consider in the first place. These plots are not, however, exactly linear - that seems to only be true for $k = 2$, as we will discuss in Section 4.2. Secondly, the slopes are always positive. This is

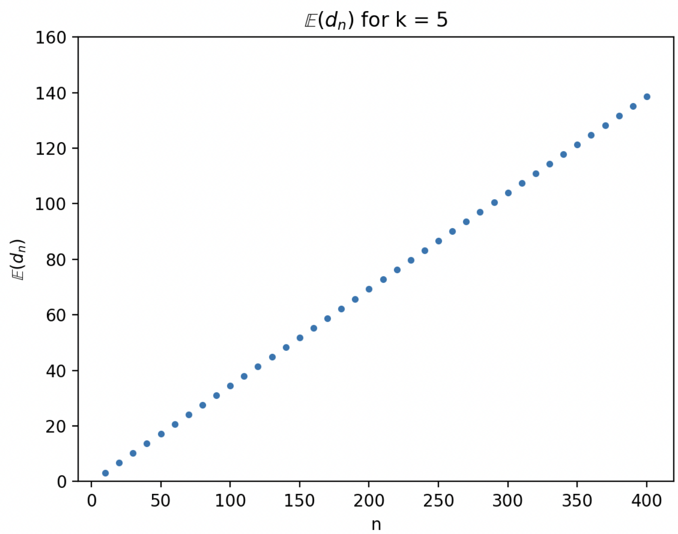

FIGURE 2. $\mathbb{E}(d_n)$ vs. n for $k = 5$

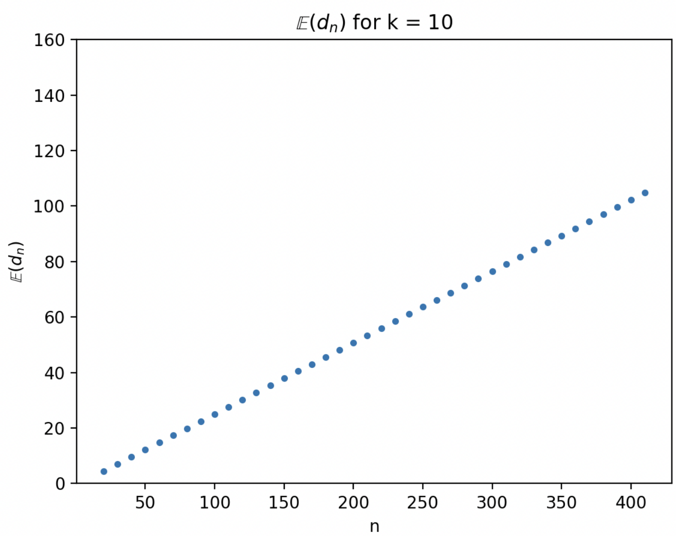

FIGURE 3. $\mathbb{E}(d_n)$ vs. n for $k = 10$

reasonable to expect, since as the size of the original set (n) from which we are picking the same size subset (k) increases, we would expect that the choices tend to become farther apart, thus increasing $\mathbb{E}(d)$. Finally, and more subtly, the slopes themselves seem to tend to decrease as k increases. That is, the effect of changing n on $\mathbb{E}(d)$ is less pronounced for larger values of k. This idea seems to be confirmed

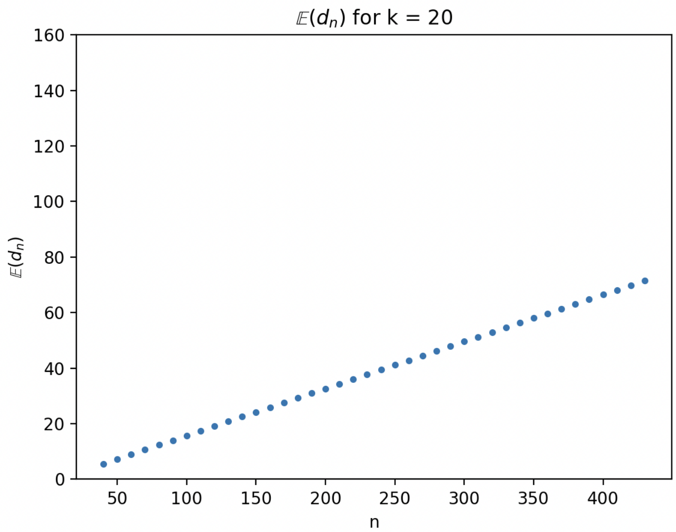

FIGURE 4. $\mathbb{E}(d_n)$ vs. n for $k = 20$

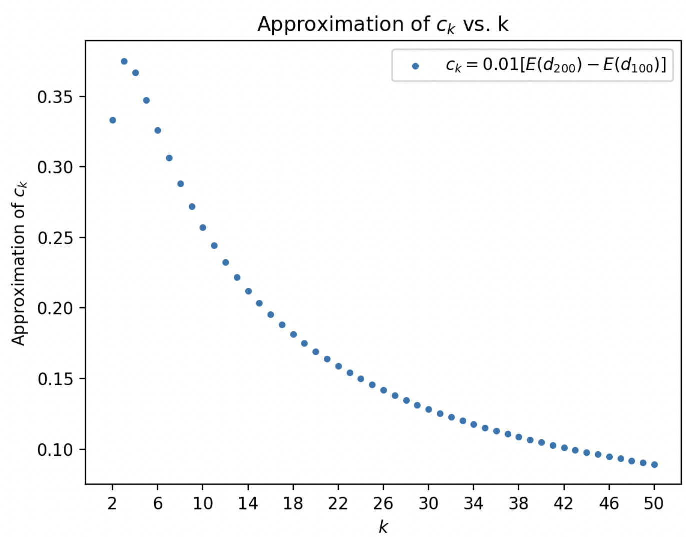

FIGURE 5. c_k vs. k

by Figure 5, which displays an approximation of c_k for values of k ranging from 2 to 29. (In each case, the approximation of c_k is given by $\frac{1}{100}(\mathbb{E}(d_{200}) - \mathbb{E}(d_{100}))$.) With the exception of $k = 2$, the graph is every decreasing, although it remains above the y-axis. As it happens, this approximation of c_k is itself very nearly the reciprocal of a linear equation. In particular, Figure 6 compares $\frac{1}{100}(\mathbb{E}(d_{200}) - \mathbb{E}(d_{100}))$ to

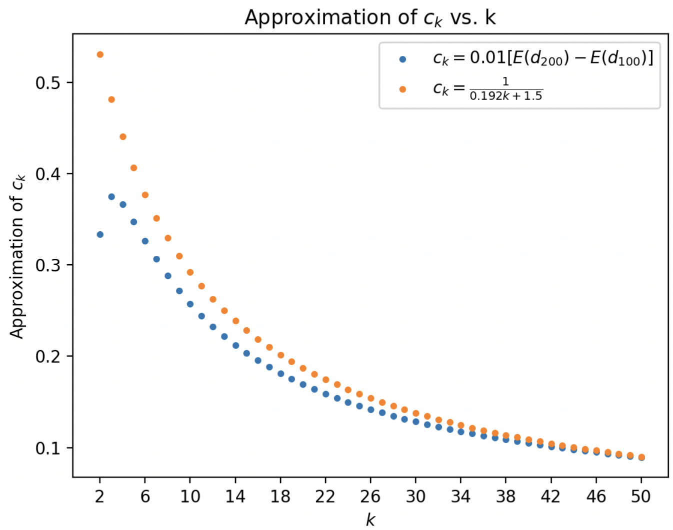

FIGURE 6. A rational function approximating c_k

$\frac{1}{0.192k+1.5}$ for a wider range of values of k, and the two scatterplots are remarkably similar. However, whether or not the two sequences are truly asymptotically equal is regrettably beyond the scope of this paper.

4.2. **The k=2 case.** Now, let's consider the case of $k = 2$. We won't need to use the formula obtained at the end of Section 3 here, since we are taking the maximum of only one difference, since there is only one pair of consecutive numbers chosen. Since d is a discrete random variable ranging here from 1 to $n-1$ (the farthest two numbers we can choose are clearly 1 and n), we can see that

$$\mathbb{E}(d_n) = \sum_{a=1}^{n-1} P(d = a) * a$$

The probability of d being equal to a is just the number of ways that can happen divided by the total number of cases, $\binom{n}{2}$. As for the number of cases when $d = a$, we can observe that the pairs can be $(1, a+1), (2, a+2), ..., (a-n, n)$, so there $a-n$ cases. Hence, we find that

$$\begin{aligned}
\mathbb{E}(d_n) &= \sum_{a=1}^{n-1} \frac{1}{\binom{n}{2}}(n-a) * a \\
&= \frac{2}{n(n-1)} \sum_{a=0}^{n-1} a(n-a) \\
&= \frac{2}{n(n-1)}(n \sum_{a=0}^{n-1} a - \sum_{a=0}^{n-1} a^2)
\end{aligned}$$

Thankfully, we have the following well-known formulas:

$$\sum_{a=1}^{n} a = \frac{n(n+1)}{2}$$

$$\sum_{a=1}^{n} a^2 = \frac{n(n+1)(2n+1)}{6}$$

Substituting, we get that

$$\begin{aligned}\mathbb{E}(d_n) &= \frac{2}{n(n-1)}[n\frac{n(n-1)}{2} - \frac{(n-1)(n)(2n-1)}{6}] \\ &= n - \frac{2n-1}{3} = \frac{n+1}{3}\end{aligned}$$

Therefore, we can see that

$$\mathbb{E}(d_{n+1}) - \mathbb{E}(d_n) = \frac{n+2}{3} - \frac{n+1}{3} = \frac{1}{3}$$

$$\lim_{n\to\infty} \mathbb{E}(d_{n+1}) - \mathbb{E}(d_n) = \lim_{n\to\infty} \frac{1}{3} = \frac{1}{3}$$

Therefore, we have verified that the conjecture holds for $k = 2$, and in particular that $c_2 = \frac{1}{3}$.

4.3. **Bounding the Change in $\mathbb{E}(d)$.** Once again, we begin with the definition of expected value:

$$\mathbb{E}(d_n) = \sum_{a=1} P(d = a) * a$$

To make any more progress, we have to determine what the upper bound of this sum should be, i.e. the greatest possible value of d given n and k. We claim that it is $n-k+1$. First, we can see that it is always attainable by picking $1, 2, 3, ..., k-1$, and n. This is a way of choosing k elements from $[1, n]$, and the greatest distance between consecutive elements is just $n-(k-1) = n-k+1$. Next, we have to show that $n-k+2$ or anything greater is impossible. To prove this, we can refer back to our secondary interpretation of the problem: a sequence of k chosen numbers and $n-k$ unchosen numbers. If $d \geq n-k+2$, then by definition there exist a pair of chosen numbers with at least $n-k+1$ consecutive unchosen numbers between them. This is clearly impossible, since we know that there are only $n-k$ unchosen numbers in the entire sequence. Now, we can be more specific:

$$\begin{aligned}\mathbb{E}(d_n) &= \sum_{i=1}^{n-k+1} P(d = i) * i \\ &= \sum_{i=1}^{n-k+1} [P(d \leq i) - P(d \leq i-1)] * i\end{aligned}$$

$$= \sum_{i=1}^{n-k+1} P(d \le i) * i - \sum_{i=1}^{n-k+1} P(d \le i-1) * i$$
$$= \sum_{i=1}^{n-k+1} P(d \le i) * i - \sum_{i=0}^{n-k} P(d \le i) * (i+1)$$
$$= [P(d \le n-k+1)(n-k+1)] + [\sum_{i=1}^{n-k} P(d \le i) * i] - [\sum_{i=1}^{n-k} P(d \le i) * (i+1)] - [P(d \le 0)(1)]$$

Since $n-k+1$ is the maximum value of d, $P(d \le n-k+1) = 1$. Moreover, because we must pick distinct integers, it is impossible for $d = 0$, so $P(d \le 0) = 0$. Therefore, we have

$$\mathbb{E}(d_n) = n-k+1 + \sum_{i=1}^{n-k} P(d \le i)(i-(i+1))$$
$$= n-k+1 - \sum_{i=1}^{n-k} P(d \le i)$$

Again, because $P(d \le 0) = 0$, we can extend the sum on the right side to

$$\mathbb{E}(d_n) = n-k+1 - \sum_{i=0}^{n-k} P(d \le i)$$
$$= \sum_{i=0}^{n-k} 1 - P(d \le i)$$

Now, we can use this expression to begin to understand $\mathbb{E}(d_{n+1}) - \mathbb{E}(d_n)$. Now that we are dealing with multiple different probability distributions, we will use $P(d_n \le i)$ to denote the probability of d being less than or equal to i given that we are choosing from $[1, n]$. In other words, the above expression would more properly be written as

$$\mathbb{E}(d_n) = \sum_{i=0}^{n-k} 1 - P(d_n \le i)$$

Therefore, we find that

$$\mathbb{E}(d_{n+1}) - \mathbb{E}(d_n) = \sum_{i=0}^{n-k+1} 1 - P(d_{n+1} \le i) - \sum_{i=0}^{n-k} 1 - P(d_n \le i)$$

Recall that because $n-k+1$ is the maximum value of d give n and k, $P(d_n \le n-k+1) = 1$ and $1 - P(d_n \le n-k+1) = 0$. Hence, we can increase the upper bound of the second sum by one without affecting its value. Also, for ease of reading, $\mathbb{E}(d_{n+1}) - \mathbb{E}(d_n)$ will be written as $\Delta\mathbb{E}(d_n)$ In other words,

$$\Delta\mathbb{E}(d_n) = \sum_{i=0}^{n-k+1} 1 - P(d_{n+1} \le i) - \sum_{i=0}^{n-k+1} 1 - P(d_n \le i)$$

$$= \sum_{i=0}^{n-k+1} 1 - P(d_{n+1} \leq i) - (1 - P(d_n \leq i))$$
$$= \sum_{i=0}^{n-k+1} P(d_n \leq i) - P(d_{n+1} \leq i)$$

Again, because we are picking distinct integers, d cannot be zero. Hence, $P(d_n \leq 0) - P(d_{n+1} \leq 0) = 0 - 0 = 0$. Therefore, we can slightly simplify our sum to get

$$\Delta\mathbb{E}(d_n) = \sum_{i=1}^{n-k+1} P(d_n \leq i) - P(d_{n+1} \leq i)$$
$$\Delta\mathbb{E}(d_n) = \sum_{i=1}^{n-k+1} \frac{1}{\binom{n}{k}} S(n,k,i) - \frac{1}{\binom{n+1}{k}} S(n+1,k,i)$$

Now, for the sake of brevity, let's denote $S(n,k,i)$ by a_i. As for $S(n+1,k,i)$, recall that it is just the number of ways to pick k elements of $[1, n+1]$ with $d \leq i$. All of those combinations either include or exclude $n+1$ in particular. The number that exclude $n+1$ is clearly just a_i, since it is the same as just picking from $[1, n]$ in the first place. Then let's denote by b_i the number of combinations which include $n+1$. This gives us that

$$\Delta\mathbb{E}(d_n) = \sum_{i=1}^{n-k+1} \frac{1}{\binom{n}{k}} a_i - \frac{1}{\binom{n+1}{k}} (a_i + b_i)$$
$$\Delta\mathbb{E}(d_n) = \sum_{i=1}^{n-k+1} (\frac{1}{\binom{n}{k}} a_i - \frac{1}{\binom{n+1}{k}} a_i) - \sum_{i=1}^{n-k+1} \frac{1}{\binom{n+1}{k}} b_i$$

Next, it will be helpful to have the following combinatorial identity:

$$\frac{1}{\binom{n}{k}} - \frac{1}{\binom{n+1}{k}} = \frac{k!(n-k!)}{n!} - \frac{k!(n-k+1)!}{(n+1)!}$$
$$= \frac{k!(n-k)!}{n!}(1 - \frac{n-k+1}{n+1})$$
$$= \frac{1}{\binom{n}{k}}(\frac{k}{n+1})$$

Substituting into our previous equation for $\Delta\mathbb{E}(d_n)$ yields

$$\Delta\mathbb{E}(d_n) = \left[(\frac{k}{n+1}) \sum_{i=1}^{n-k+1} \frac{a_i}{\binom{n}{k}}\right] - \left[\frac{1}{\binom{n+1}{k}} \sum_{i=1}^{n-k+1} b_i\right]$$
$$\Delta\mathbb{E}(d_n) = \left[(\frac{k}{n+1}) \sum_{i=1}^{n-k+1} P(d_n \leq i)\right] - \left[\frac{1}{\binom{n+1}{k}} \sum_{i=1}^{n-k+1} b_i\right]$$

Now that we have split $\Delta\mathbb{E}(d_n)$ into two separate sums, we can bound it by simply finding bounds on each individual sum. Let's call the first sum A. Then because A is a sum over $n-k+1$ probabilities, and each probability is necessarily at most 1, $A \leq n-k+1$. This is obviously a very crude estimate, since most of these probabilities are well below 1. Nonetheless, it will be enough for the purposes of bounding $\Delta\mathbb{E}(d_n)$ between constants. It is similarly immediate that $A \geq 0$, but we can also very quickly deduce that $A \geq 1$, since $n-k+1$ is the largest possible value of d, so $P(d_n \leq n-k+1) = 1$. In other words, we are concluding that

$$\frac{k}{n+1} \leq (\frac{k}{n+1}) \sum_{i=1}^{n-k+1} P(d_n \leq i) \leq \frac{k}{n+1}(n-k+1)$$

Secondly, we can find very similar bounds on the second sum. Recall that b_i is defined as the number of subsets of k elements of $[1, n+1]$ that include $n+1$ and have $d \leq i$. Again, it is clear by the definition of b_i that $b_i \geq 0$. On the other hand, because one element of the subset is already prescribed, we are actually only selecting the remaining $k-1$ elements from $[1, n]$. Therefore, $b_i \leq \binom{n}{k-1}$, so

$$0 \leq \frac{1}{\binom{n+1}{k}} \sum_{i=1}^{n-k+1} b_i \leq (n-k+1)\frac{\binom{n}{k-1}}{\binom{n+1}{k}}$$

$$0 \leq \frac{1}{\binom{n+1}{k}} \sum_{i=1}^{n-k+1} b_i \leq (n-k+1)\frac{n!}{(k-1)!(n-k+1)!} * \frac{(n-k+1)! * k!}{(n+1)!}$$

$$0 \leq \frac{1}{\binom{n+1}{k}} \sum_{i=1}^{n-k+1} b_i \leq (n-k+1)\frac{k}{n+1} = k(\frac{n-k+1}{n+1})$$

Now we have upper and lower bounds on each of the two terms, we can find bounds on their difference. First, to get a lower bound, we take a lower bound on the first (added) term and an upper bound on the second (subtracted). That is,

$$\frac{k}{n+1} - k(\frac{n-k+1}{n+1}) \leq \Delta\mathbb{E}(d_n)$$

$$\frac{k}{n+1}(1-n+k-1) \leq \Delta\mathbb{E}(d_n)$$

$$-k \leq -k(\frac{n-k}{n+1}) \leq \Delta\mathbb{E}(d_n)$$

On the other hand, to find the upper bound, we can simply take the upper bound of the first term and a lower bound on the second:

$$\Delta\mathbb{E}(d_n) \leq k(\frac{n-k+1}{n+1}) \leq k$$

Therefore, at last, we have shown that

$$-k \leq \Delta\mathbb{E}(d_n) \leq k$$

It is important to note that these are, again, rather crude bounds. The data shown in Section 4.1 suggested that for all k, $c_k < 1$, which is much smaller than the upper bound here of k, which can be arbitrarily large. Secondly, these calculations fail to prove that $\Delta\mathbb{E}(d_n) > 0$, despite it being supported both by intuition and the numerical evidence shown in 4.1. Nevertheless, it is a (hopefully) useful and encouraging step towards a proof of Conjecture 4.1. We have shown that $\Delta\mathbb{E}(d_n)$, a quantity which we are hoping to show converges as n tends to infinity, is at least bounded by constants (while k is a parameter of the problem, it is a constant in this context since it does not depend on n).

Acknowledgments

First, I would like to thank my mentor, Mark Olson, for first introducing me to the subject of analytic and symbolic combinatorics and for his guidance in writing this paper. Second, I would like to thank Peter May for running the REU program and ensuring that it is open to anyone who is interested in mathematics research, even those such as myself with only one year of college-level coursework. Finally, I would like to thank my parents for their support of my interest in math, before, during, and after the program.

5. Bibliography

References

[1] Philippe Flajolet and Robert Sedgewick. Analytic Combinatorics. Cambridge University Press. 2009.

[2] Benjamin Sambale. An invitation to formal power series. https://arxiv.org/pdf/2205.00879.pdf

[3] William Feller. An Introduction to Probability Theory and Its Applications, Volume I. John Wiley & Sons, Inc. 1968.