UNIVERSITY OF CALIFORNIA,
IRVINE

Couplings, Component Counting Processes, and Probabilistic Number Theory

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Joseph Paul Squillace

Dissertation Committee:
Professor Michael Cranston, Chair
Associate Professor Nathan Kaplan
Professor Roman Vershynin

2020

DEDICATION

To my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisors Michael Cranston and Nathan Kaplan. The mathematics department has provided me the opportunity to teach and study a variety of topics. I would also like to thank all of the educators and students that I have met along my mathematical journey.

VITA

Joseph Paul Squillace

EDUCATION

Doctor of Philosophy in Mathematics **2020**
University of California, Irvine *Irvine, CA*

Master of Arts in Mathematics **2014**
San Francisco State University *San Francisco, CA*

Bachelor of Arts in Mathematics **2011**
University of California, Berkeley *Berkeley, CA*

Certificate of Transfer Completion **2009**
College of the Desert *Palm Desert, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2016–2020**
University of California, Irvine *Irvine, CA*

Graduate Research Assistant **2012–2013**
San Francisco State University *San Francisco, CA*

TEACHING EXPERIENCE

Teaching Assistant **2015–2020**
University of California, Irvine *Irvine, CA*

Teaching Assistant **2012–2013**
San Francisco State University *San Francisco, CA*

REFEREED JOURNAL PUBLICATIONS

Squillace, J. *Estimating the fractal dimension of sets determined by nonergodic parameters.* Discrete Contin. Dyn. Syst. **37**, no. 11, 5843–5859, 2017.

AWARDS

Outstanding Teaching Award **2017-2018**
University of California, Irvine

Graduate Distinguished Achievement Award **2014**
San Francisco State University

ABSTRACT OF THE DISSERTATION

Couplings, Component Counting Processes, and Probabilistic Number Theory

By

Joseph Paul Squillace

Doctor of Philosophy in Mathematics

University of California, Irvine, 2020

Professor Michael Cranston, Chair

Our results are concerned with couplings, component counts of combinatorial objects, and probabilistic number theory. In the theory of couplings, we are concerned with the general problem of proving the existence of joint distributions $p(\cdot,\cdot)$ of two discrete random variables M and N subject to infinitely many constraints of the form $p(M=i,N=j)=0$. The constraints placed on the joint distributions will require, for many elements j in the range of N, $p(M=i,N=j)=0$ for infinitely many values of i in the range of M, where the corresponding values of i depend on j. To prove the existence of such joint distributions, we apply a theorem proved by Volker Strassen on the existence of joint distributions with prespecified marginal distributions. In the case in which N is uniformly distributed in a combinatorial structure with C_i components of size i, we seek to measure the amount of dependence in the process $(C_i)_{i\leq n}$ by coupling N with a variable M such that M has Z_i components of size i and the Z_i's are independent, with $\sum_{i\leq n}(C_i-Z_i)^+\leq 1$.

In the combinatorial example of noncrossing partitions, we provide two derivations of the probability distribution of the component counts of a uniformly distributed noncrossing partition. Upon applying a bijection between the set of noncrossing partitions and Dyck paths consisting of up-steps and down-steps, our results specify the joint and marginal distributions of the block counts of the number of consecutive up-steps in a uniformly random

chosen Dyck path.

In number theory, we give an analogue of the Erdös-Kac Theorem by providing a family of integer-valued random variables on $\{1, \ldots, n\}$ whose number of distinct prime factors is roughly $\log\log n + \chi \cdot \sqrt{\log\log n}$ for large values of n, where χ is a standard normal variable. Our final result involves couplings of a Zeta-distributed variable. Given $s > 1$ and $n \in \mathbb{N}$, consider a Zeta(s)-distributed integer-valued random variable with prime factorization $Z(s) = \prod_p p^{\alpha_p(s)}$ and the truncation $Z_n(s) \coloneqq \prod_{p \leq n} p^{\alpha_p(s)}$. The prime powers $\alpha_p(s)$ are independent with $\alpha_p(s) \sim \text{Geometric}\left(\frac{1}{p^s}\right)$, and we also consider a random variable $M(n) = \prod_{p \leq n} p^{Z_p}$, where the Z_p's are independent with $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$. We apply the concept of pivot mass and a theorem proved by Strassen in order to prove the existence of couplings of a Zeta-distributed random variable $Z(s)$ and $M(n)$ in which we can make probabilistic divisibility statements of the form "$Z_n(s)$ divides $M(n) P(n)$" for some random prime $P(n) \leq n$. In particular, we will prove that for each $n \in \mathbb{N}$ and an integer $k \geq 4$, there exists an $\varepsilon(k) > 0$ such that when $s \in (1, 1 + \varepsilon(k))$ we can couple $Z(s)$ and $M(n)$ such that if $Z(s) \leq k$, then $Z_n(s)$ always divides $M(n) P(n)$ for some random prime $P(n) \leq n$.

Chapter 1

Introduction

Our work is based on probabilistic number theory, component counts of combinatorial objects, and the interplay between probability and combinatorics. Chapters 2-5 are self-contained, while Chapters 6 and 7 cite theorems from the previous chapters. In this chapter we summarize each of our main results and mention some connections between the chapters.

In Chapter 2 we provide a generalization of the Erdös-Kac Theorem (also known as the fundamental theorem of probabilistic number theory) by providing a family of integer-valued random variables on $\{1, \ldots, n\}$ whose number of distinct prime factors is roughly $\log\log n + \chi \cdot \sqrt{\log\log n}$ for large values of n, where χ is a standard normal variable. As an application of our work, we show that an integer-valued random variable with the Harmonic(n)-distribution has roughly $\log\log n + \chi \cdot \sqrt{\log\log n}$ distinct prime factors when n is large.

In Chapter 3 we take a detour from number theory to combinatorics by studying noncrossing partitions. In particular, we apply results on the number of noncrossing partitions with a specified block type to derive of the joint and marginal distributions of the block counts in a uniformly chosen noncrossing partition on $\{1, \ldots, n\}$. We then apply a bijection between the set of noncrossing partitions and the set of Dyck paths (consisting of up-steps and down-

steps) to specify the joint and marginal probability distributions of the blocks of consecutive up-steps in a uniformly random chosen Dyck path consisting of n up-steps and n down-steps.

In Chapter 4 we discuss a conjecture posed by Richard Arratia and develop the concept of pivot mass which is applied in Chapters 6 and 7. Arratia's Conjecture concerns a uniformly distributed random variable $N(n) \in \{1, \ldots, n\}$ with prime factorization $\prod_{p \leq n} p^{C_p(n)}$ and a random integer $M(n) = \prod_{p \leq n} p^{Z_p}$ consisting of independent variables Z_p that are geometrically distributed with[1] $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$ and the conjecture asks whether it is possible to construct a probability space with marginals corresponding to $N(n)$ and $M(n)$ such that $N(n)$ always[2] divides $M(n)P(n)$ for some random prime $P(n) \leq n$. The variables $C_p, p \leq n$, are dependent, so the conjecture is an attempt to measure the amount of the dependence in the process $(C_p)_{p \leq n}$. This problem has not been resolved; however, we apply the pivot mass concept in order to provide a way to test the conjecture on a computer. That is, for any given value of n, we can use computer code to determine whether or not $N(n)$ divides $M(n)P(n)$ for some prime $P(n) \leq n$. The significance of this fact is that any such coupling of $N(n)$ and $M(n)$ would correspond to an $\infty \times n$ matrix with marginals $M(n)$ and $N(n)$ such that the entry in row $M(n) = i$ and column $N(n) = j$ is 0 if j does not divide $iP(n)$ for any prime $P(n) \leq n$.

In Chapter 5 we discuss the concept of stochastic domination of random variables and its connection to couplings. We prove that the variable $C_p(n)$ is stochastically dominated by Z_p, for each prime $p \leq n$, by constructing a coupling in which $C_p(n)$ is pointwise no larger than Z_p.

In Chapter 6, we prove the analogue of Arratia's Conjecture in many combinatorial settings. This was the first successful application of the pivot mass concept in regards to providing couplings.

[1] I.e., $\mathbb{P}(Z_p = k) = \frac{1-\frac{1}{p}}{p^k}$ for nonnegative integers k.

[2] In the desired coupling, any realization $M = i, N = j$ must have j dividing pi for some prime $p \leq n$.

In Chapter 7, we consider a Zeta(s)-distributed random variable $Z(s) = \prod_p p^{\alpha_p(s)}$ and its truncation $Z_n(s) := \prod_{p \leq n} p^{a_p(s)}$. Given $n \in \mathbb{N}$ and $k \geq 4$, we prove that there exists an $\varepsilon(k) > 0$ such that when $s \in (1, 1 + \varepsilon(k))$ we can couple $Z(s)$ and $M(n)$ such that if $Z(s) \leq k$, then $Z_n(s)$ divides $M(n)P(n)$ for some random prime $P(n) \leq n$.

We conclude this chapter by specifying some connections between the following chapters. Chapters 3, 6, and 7 concern component counts in a combinatorial structure. Chapters 4, 5, 6, and 7 involve couplings. Chapters 2, 4, and 7 discuss number theory. The uniform distribution appears in Chapters 3, 4, 5, and 6. The variable $M(n)$ appears in Chapters 4, 5, and 7.

Chapter 2

A Generalization of the Erdös-Kac Central Limit Theorem

2.1 Introduction

In 1917, Hardy and Ramanujan (p. 270 of [6]) proved that the number of distinct prime factors of a natural number n, denoted $\omega(n)$, is about $\log\log n$. Informally speaking, the Erdös-Kac Theorem generalizes the Hardy-Ramanujan Theorem by showing that $\omega(n)$ is about $\log\log n + \chi \cdot \sqrt{\log\log n}$, where $\chi \sim \mathcal{N}(0,1)$. More precisely, we have the following theorem (p. 738 of [4]).

Theorem 1. *(Erdös-Kac) Let P_n denote the uniform distribution[1] on $\{1,2,\ldots,n\}$. As $n \to \infty$,*

$$P_n\left(m \leq n : \omega(m) - \log\log n \leq x\left(\log\log n\right)^{1/2}\right) \to \mathbb{P}(\chi \leq x).$$

Our goal is to find random variables, other than a uniformly distributed variable, on $[n] :=$

[1] I.e., if U is uniformly distributed on $\{1,2,\ldots,n\}$, then for any subset $A \subseteq \{1,2,\ldots,n\}$ we have $P_n(A) = \mathbb{P}(U \in A)$.

$\{1, 2, \ldots, n\}$ which also have roughly $\log\log n + \chi \cdot \sqrt{\log\log n}$ distinct prime factors.

2.1.1 The Main Theorem

Let us define a probability distribution $\mathbb{P}_n$ on $[n]$ given by

$$\mathbb{P}_n(i) = \mathbf{1}_{\{i \in [n]\}} \left(\frac{1}{n} + \varepsilon_{i,n} \right), \tag{2.1}$$

where for a given set A, $\mathbf{1}_A$ denotes the indicator function of A, and for all k-tuples $(p_1, \ldots, p_k)$ of distinct primes,

$$\lim_{n \to \infty} \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{l p_1 p_2 \cdots p_k, n} = 0. \tag{2.2}$$

Due to the axioms of probability, we necessarily have

$$\sum_{i=1}^{n} \varepsilon_{i,n} = 0 \tag{2.3}$$

and

$$\varepsilon_{i,n} \in \left[-\frac{1}{n}, 1 - \frac{1}{n} \right]. \tag{2.4}$$

The motivation for defining $\mathbb{P}_n$ in terms of the uniform distribution was motivated by Durrett's proof (Theorem 3.4.16 in [3]) of the Erdös-Kac Theorem. Upon replacing the uniform distribution P_n with the new distribution $\mathbb{P}_n$ in Durrett's proof, we arrive at some constraints that the terms $\varepsilon_{i,n}, i \leq n$, must satisfy in order to conclude that a random variable with the $\mathbb{P}_n$ distribution has about $\log\log n + \chi \cdot \sqrt{\log\log n}$ distinct prime factors. Our main result is the following theorem, where $\lfloor \cdot \rfloor$ denotes the floor function.

Theorem 2. *Let $\chi \sim \mathcal{N}(0, 1)$. Suppose the following statements are true.*

1. *There exists a constant C such that*

$$\sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{lp_1 \cdots p_k, n} \leq \frac{C}{n} \tag{2.5}$$

for all n and, for each k, all k-tuples $(p_1, \ldots, p_k)$ consisting of distinct primes of size at most $n^{1/\log\log n}$.

2. *There exists a constant D such that for all n and for all prime p with $p > n^{1/\log\log n}$ we have*

$$\sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n} \leq \frac{D}{p}. \tag{2.6}$$

Let $\mathbb{P}_n^*$ denote a probability distribution obtained by imposing the restrictions (2.5) *and* (2.6) *on $\mathbb{P}_n$. As $n \to \infty$,*

$$\mathbb{P}_n^* \left(m \leq n : \omega(m) - \log\log n \leq x (\log\log n)^{1/2} \right) \to \mathbb{P}(\chi \leq x).$$

Remark. If $\varepsilon_{i,n} = 0$ for all $i \leq n$, then $\mathbb{P}_n^* = P_n$ and we obtain Theorem 1. In Example 5 of §2.4, we determine the values of $\varepsilon_{i,n}, i \leq n$, for the Harmonic distribution on $[n]$, and then we specify values of C and D for the Harmonic distribution. As a result, Theorem 2 implies that a integer-valued random variable on $[n]$ with the Harmonic distribution has about $\log\log n + \chi \cdot \sqrt{\log\log n}$ distinct prime factors.

2.1.2 Outline

In §2.2 we mention a heuristic of Kac for the number of distinct prime factors of a uniformly distributed variable on $\mathbb{N}$ and then we give an analogue of the heuristic for any random variable whose probability distribution is given by $\mathbb{P}_n$. The latter was used to arrive at the

parameters $\log\log n$ and $\sqrt{\log\log n}$ appearing in the statement of Theorem 2. The proof of Theorem 2 applies the method of moments and is motivated by Durrett's proof of the Erdös-Kac Theorem. In §2.3 we prove Theorem 2. In §2.4 we apply Theorem 2 to show that the number of distinct prime factors of a positive integer integer in $[n]$ with the harmonic distribution tends to the normal distribution $\mathcal{N}(\log\log n, \log\log n)$ as $n \to \infty$.

2.2 Kac's Heuristic and its Analogue for $\mathbb{P}_n$

Kac's heuristic[2] for the uniform distribution (pp. 134-135 of [3]) suggests the statement of Theorem 1 and is based on the fact that given a random integer $n \in \mathbb{N}$, the events $\{p \text{ divides } n\}$ and $\{q \text{ divides } n\}$ are independent when p and q are distinct primes.

Now we consider the probability distribution $\mathbb{P}_n$ and its behavior in the limit as $n \to \infty$. Given a prime p, let A_p denote the set of positive integers divisible by p. Then

$$\begin{aligned}
\mathbb{P}_\infty(A_p) &:= \lim_{n\to\infty} \mathbb{P}_n(A_p) \\
&= \lim_{n\to\infty} \sum_{l=1}^{\lfloor \frac{n}{p} \rfloor} \mathbb{P}_n(lp) \\
&\overset{(2.1)}{=} \lim_{n\to\infty} \sum_{l=1}^{\lfloor \frac{n}{p} \rfloor} \mathbf{1}_{\{lp\le n\}} \left(\frac{1}{n} + \varepsilon_{lp,n}\right) \\
&= \lim_{n\to\infty} \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n} + \lim_{n\to\infty} \sum_{l=1}^{\lfloor \frac{n}{p} \rfloor} \varepsilon_{lp,n} \\
&\overset{(2.2)}{=} \frac{1}{p}.
\end{aligned}$$

[2]The heuristic is based on the connection between independence and a Gaussian law of errors.

If $q \neq p$ is another prime, then, similarly,

$$\begin{aligned}\mathbb{P}_\infty(A_p \cap A_q) &= \lim_{n\to\infty}\left(\frac{\left\lfloor \frac{n}{pq}\right\rfloor}{n} + \sum_{l=1}^{\left\lfloor \frac{n}{pq}\right\rfloor} \varepsilon_{lpq,n}\right) \\ &\overset{(2.2)}{=} \frac{1}{pq} \\ &= \mathbb{P}_\infty(A_p)\,\mathbb{P}_\infty(A_q).\end{aligned}$$

Therefore, the events A_p and A_q are independent. In general, (2.2) ensures that, for any positive integer k, the events $A_{p_1}, \ldots, A_{p_l}$ are independent for $1 < l \leq k$; i.e., $\mathbb{P}_\infty(A_{p_1} \cap \cdots \cap A_{p_l}) = \mathbb{P}_\infty(A_{p_1}) \cdots \mathbb{P}_\infty(A_{p_l})$ for $1 < l \leq k$.

Let $\delta_p(n) = \mathbf{1}_{p|n}$ so that

$$\omega(n) = \sum_{p\leq n} \delta_p(n)$$

is the number of distinct prime factors of n. The indicator variables δ_p behave like Bernoulli variables X_p that are independent and identically distributed with

$$\begin{aligned}\mathbb{P}(X_p = 1) &= \frac{1}{p}, \\ \mathbb{P}(X_p = 0) &= 1 - 1/p.\end{aligned}$$

The mean of $\sum_{p\leq n} X_p$ is

$$\begin{aligned}\mathbb{E}\left(\sum_{p\leq n} X_p\right) &= \sum_{p\leq n} \mathbb{P}(X_p = 1) \\ &= \sum_{p\leq n} \frac{1}{p},\end{aligned}$$

and the variance of $\sum_{p\leq n} X_p$ is

$$\begin{aligned}
\operatorname{Var}\left(\sum_{p\leq n} X_p\right) &= \sum_{p\leq n} \operatorname{Var}\left(X_p\right) \\
&= \sum_{p\leq n} \left(\mathbb{E}\left(X_p^2\right) - \left(\mathbb{E}\left(X_p\right)\right)^2\right) \\
&= \sum_{p\leq n} \mathbb{E}\left(X_p^2\right) - \sum_{p\leq n} \left(\left(\mathbb{E}\left(X_p\right)\right)^2\right) \\
&= \sum_{p\leq n} \mathbb{E}\left(X_p\right) - \sum_{p\leq n} \left(\left(\frac{1}{p}\right)^2\right) \\
&= \sum_{p\leq n} \frac{1}{p} - \sum_{p\leq n} \frac{1}{p^2}.
\end{aligned}$$

Note that $\sum_{p\leq n} \frac{1}{p} = \log\log n + O\left(1\right)$ by Mertens' Second Theorem (p. 22 of [13]); the series $\sum_{p\leq n} \frac{1}{p^2}$ converges (by comparison with $\sum_{1\leq n} \frac{1}{n^2}$). Therefore, the mean and variance are

$$\begin{aligned}
\sum_{p\leq n} \frac{1}{p} &= \log\log n + O\left(1\right), \\
\sum_{p\leq n} \frac{1}{p} - \sum_{p\leq n} \frac{1}{p^2} &= \log\log n + O\left(1\right).
\end{aligned}$$

This concludes the analogue of Kac's heuristic for $\mathbb{P}_n$ and justifies the parameters of the normal distribution appearing in Theorem 2.

2.3 Proving Theorem 2

Lemma 3. *As $n \to \infty$,*

$$\left(\sum_{n^{1/\log\log n} < p \leq n} \left(\frac{1}{p} + \sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n}\right)\right) / \left(\log\log n\right)^{1/2} \to 0.$$

Proof of Lemma 3. We have

$$-\frac{\left\lfloor \frac{n}{p} \right\rfloor}{n} \overset{(2.4)}{\leq} \sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n} \overset{(2.6)}{\leq} \frac{D}{p}$$

for all n and all primes p with $p > n^{1/\log\log n}$. Therefore,

$$\frac{1}{p} + \sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n} \in \left[0, \frac{D+1}{p}\right] \tag{2.7}$$

for all n. Thus, we have

$$\left(\sum_{n^{1/\log\log n} < p \leq n} \left(\frac{1}{p} + \sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n} \right) \right) / \left(\log\log n\right)^{1/2} \to 0$$

due to (2.7) along with the fact that Durrett (p. 135 of [3]) shows

$$\left(\sum_{n^{1/\log\log n} < p \leq n} \frac{1}{p} \right) / \left(\log\log n\right)^{1/2} \to 0.$$

This proves Lemma 3. □

The following lemma is proved by Durrett (p. 156 of [3]).

Lemma 4. *If $\varepsilon > 0$, then $n^{1/\log\log n} \leq n^{\varepsilon}$ for large n and hence*

$$\frac{n^{r/\log\log n}}{n} \to 0 \tag{2.8}$$

for all $r < \infty$.

Proof of Theorem 2. Let $g_n(m) \coloneqq \sum_{p \leq n^{1/\log\log n}} \delta_p(m)$ and let $\mathbb{E}_n$ denote expectation with

respect to $\mathbb{P}_n^*$. Then

$$\begin{aligned}
\mathbb{E}_n\left(\sum_{n^{1/\log\log n}<p\leq n}\delta_p\right) &= \sum_{n^{1/\log\log n}<p\leq n}\mathbb{P}_n^*\left(m:\delta_p\left(m\right)=1\right)\\
&= \sum_{n^{1/\log\log n}<p\leq n}\mathbb{P}_n^*\left(m:m=p,2\cdot p,\ldots,\left\lfloor\frac{n}{p}\right\rfloor\cdot p\right)\\
&= \sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\mathbb{P}_n^*\left(m:m=lp\right)\\
&\overset{(2.1)}{=} \sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\left(\frac{1}{n}+\varepsilon_{lp,n}\right)\\
&= \sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\frac{1}{n}+\sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\varepsilon_{lp,n}\\
&= \sum_{n^{1/\log\log n}<p\leq n}\frac{\left\lfloor\frac{n}{p}\right\rfloor}{n}+\sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\varepsilon_{lp,n}\\
&\leq \sum_{n^{1/\log\log n}<p\leq n}\frac{1}{p}+\sum_{n^{1/\log\log n}<p\leq n}\sum_{l=1}^{\left\lfloor\frac{n}{p}\right\rfloor}\varepsilon_{lp,n},
\end{aligned}$$

so by Lemma 3 it suffices to prove the theorem for g_n; i.e., we can replace $\omega\left(m\right)$ with $g_n\left(m\right)$ in the statement of Theorem 2 without affecting the limiting distribution. Let

$$\begin{aligned}
S_n &\coloneqq \sum_{p\leq n^{1/\log\log n}}X_p,\\
b_n &\coloneqq \mathbb{E}\left(S_n\right),\\
a_n^2 &\coloneqq \operatorname{Var}\left(S_n\right).
\end{aligned}$$

By Lemma 3, b_n and a_n^2 are both $\log\log n+o\left(\left(\log\log n\right)^{1/2}\right)$, so it suffices to show

$$\mathbb{P}_n^*\left(m:g_n\left(m\right)-b_n\leq xa_n\right)\to\mathbb{P}\left(\chi\leq x\right).$$

An application of Theorem 3.4.5 of [3] shows $(S_n - b_n)/a_n \to \chi$, and since $|X_p| \leq 1$, it follows from Durrett's second proof of Theorem 3.4.5 [3] that $\mathbb{E}\left((S_n - b_n)/a_n\right)^r \to \mathbb{E}(\chi^r)$ for all r. Using the notation from that proof (and replacing i_j by p_j) we have

$$\mathbb{E}(S_n^r) = \sum_{k=1}^{r} \sum_{r_i} \frac{r!}{r_1! \cdots r_k!} \frac{1}{k!} \sum_{p_j} \mathbb{E}\left(X_{p_1}^{r_1} \cdots X_{p_k}^{r_k}\right),$$

where the sum $\sum_{r_i}$ extends over all k-tuples of positive integers for which $r_1 + \cdots + r_k = r$, and $\sum_{p_j}$ extends over all k-tuples of distinct primes in $[n]$. Since $X_p \in \{0, 1\}$, the summand in $\sum_{p_j} \mathbb{E}\left(X_{p_1}^{r_1} \cdots X_{p_k}^{r_k}\right)$ is

$$\mathbb{E}(X_{p_1} \cdots X_{p_k}) = \frac{1}{p_1 \cdots p_k}$$

by independence of the X_p's. Moreover,

$$\begin{aligned}
\mathbb{E}_n(\delta_{p_1} \cdots \delta_{p_k}) &= \mathbb{P}_n^*(m : \delta_{p_1}(m) = \delta_{p_2}(m) = \cdots = \delta_{p_k}(m) = 1) \\
&= \mathbb{P}_n^*\left(m : m = p_1 \cdots p_k, 2 \cdot p_1 \ldots p_k, \ldots, \left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor \cdot p_1 \cdots p_k\right) \\
&= \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \mathbb{P}_n^*(m : m = l p_1 \cdots p_k) \\
&\overset{(2.1)}{=} \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \left(\frac{1}{n} + \varepsilon_{l p_1 \cdots p_k, n}\right) \\
&= \frac{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor}{n} + \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{l p_1 \cdots p_k, n}.
\end{aligned}$$

The two moments differ by at most

$$\max\left\{ \frac{1}{p_1 \cdots p_k} - \frac{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor}{n} - \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{l p_1 \cdots p_k, n}, \frac{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor}{n} + \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{l p_1 \cdots p_k, n} - \frac{1}{p_1 \cdots p_k} \right\}.$$

Further,

$$\frac{1}{p_1\cdots p_k}-\frac{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\le\frac{1}{p_1\cdots p_k}-\frac{\frac{n}{p_1\cdots p_k}-1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}$$
$$=\frac{1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n},$$

and

$$\frac{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}{n}+\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}-\frac{1}{p_1\cdots p_k}\le\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}.$$

Thus, the maximum becomes

$$\max\left\{\frac{1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n},\ \sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\right\}.$$

The maximum equals $\frac{1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}$ if and only if $\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\le\frac{1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}$, and the latter is equivalent to $\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\le\frac{1/2}{n}$. On the other hand, if the maximum equals $\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}$, then by (2.5) we have $\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\le\frac{C}{n}$. Therefore, the two rth moments differ by

$$\begin{aligned}
\left|\mathbb{E}\left(S_n^r\right)-\mathbb{E}_n\left(g_n^r\right)\right| &\le \sum_{k=1}^{r}\sum_{r_i}\frac{r!}{r_1!\cdots r_k!}\frac{1}{k!}\sum_{p_j}\left(\max\left\{\frac{1}{n}-\sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n},\ \sum_{l=1}^{\left\lfloor\frac{n}{p_1\cdots p_k}\right\rfloor}\varepsilon_{lp_1\cdots p_k,n}\right\}\right)\\
&\le \sum_{k=1}^{r}\sum_{r_i}\frac{r!}{r_1!\cdots r_k!}\frac{1}{k!}\sum_{p_j}\frac{\max\left\{C,\frac{1}{2}\right\}}{n}\\
&\le \frac{\max\left\{C,\frac{1}{2}\right\}}{n}\left(\sum_{p\le n^{1/\log\log n}}1\right)^r\\
&\le \max\left\{C,\frac{1}{2}\right\}\cdot\frac{n^{r/\log\log n}}{n}\\
&\overset{(2.8)}{\to} 0.
\end{aligned}$$

□

2.4 Applying Theorem 2 to the Harmonic(n) Distribution

Example 5. Given $n \in \mathbb{N}$, let Q_n denote the Harmonic(n) distribution so that

$$Q_n(i) = \mathbf{1}_{\{i \in [n]\}} \frac{1}{i \sum_{i=1}^{n} \frac{1}{i}}.$$

In order to apply Theorem 2 to Q_n, we will prove that the values $\varepsilon_{i,n}$ corresponding to Q_n satisfy (2.2), (2.5), and (2.6). If $1 \leq i \leq n$, (2.1) implies $\varepsilon_{i,n} = \frac{1}{i \sum_{i=1}^{n} \frac{1}{i}} - \frac{1}{n}$. The left hand side of (2.6) becomes

$$\begin{aligned}
\sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \varepsilon_{lp,n} &= \sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \left(\frac{1}{lp \sum_{i=1}^{n} \frac{1}{i}} - \frac{1}{n} \right) \\
&= \frac{\sum_{l=1}^{\left\lfloor \frac{n}{p} \right\rfloor} \frac{1}{l}}{p \sum_{l=1}^{n} \frac{1}{l}} - \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n} \\
&\leq \frac{1}{p} - \left(\frac{\frac{n}{p} - 1}{n} \right) \\
&\leq \frac{1}{p},
\end{aligned}$$

so we can take $D = 1$ in (2.6). Further, the left hand side of (2.5) becomes

$$\begin{aligned}
\sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \varepsilon_{lp_1 \cdots p_k, n} &= \sum_{l=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \left(\frac{1}{lp_1 \cdots p_k \sum_{i=1}^{n} \frac{1}{i}} - \frac{1}{n} \right) \\
&= \frac{\sum_{i=1}^{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor} \frac{1}{i}}{p_1 \cdots p_k \sum_{i=1}^{n} \frac{1}{i}} - \frac{\left\lfloor \frac{n}{p_1 \cdots p_k} \right\rfloor}{n} \\
&\leq \frac{1}{p_1 \cdots p_k} - \left(\frac{\frac{n}{p_1 \cdots p_k} - 1}{n} \right) \\
&= \frac{1}{n},
\end{aligned}$$

so we can take $C = 1$ in (2.5). Moreover,

$$\begin{aligned}\lim_{n\to\infty} \sum_{l=1}^{\left\lfloor \frac{n}{p_1\cdots p_k} \right\rfloor} \varepsilon_{lp_1\cdots p_k,n} &= \lim_{n\to\infty} \left(\frac{\sum_{l=1}^{\left\lfloor \frac{n}{p_1\cdots p_k} \right\rfloor}}{p_1\cdots p_k \sum_{i=1}^{n} \frac{1}{i}} - \frac{\left\lfloor \frac{n}{p_1\cdots p_k} \right\rfloor}{n} \right) \\ &= \frac{1}{p_1\cdots p_k} - \frac{1}{p_1\cdots p_k} \\ &= 0,\end{aligned}$$

so (2.2) also holds. This proves the analogue of the Erdös-Kac Theorem in the case of an integer-valued random variable with the Harmonic (n) distribution by Theorem 2. $\square$

Chapter 3

The Joint and Marginal Block Counts in a Uniformly Distributed Noncrossing Partition

3.1 Introduction

Noncrossing partitions have played a role in free probability, and it is known [11] that the size of the set of noncrossing partitions on $[n] \coloneqq \{1, 2, \ldots, n\}$ is given by $\frac{1}{n+1}\binom{2n}{n}$, the nth Catalan number. Consider a regular n-gon labeled by $1, 2, \ldots, n$ in a cyclic order. Given a set partition of $[n]$, draw a line between any two vertices belonging to the same block.

Definition. A **noncrossing partition** on $[n]$ is a set partition on $[n]$ in which no two blocks cross each other. I.e., if a and b belong to one block and x and y to another then they are not arranged in the order $axby$. Denote the set of all noncrossing partitions on $[n]$ by NC_n.

Definition. Given a uniformly distributed random variable $\tau(n) \sim \mathrm{Unif}(\mathrm{NC}_n)$, denote by $C_i(n)$ the number of blocks of $\tau(n)$ of size i. The process $(C_i(n))_{i \leq n} \coloneqq (C_1(n), \ldots, C_n(n))$

is the **block type** of $\tau(n)$. Denote the set of all block types in NC_n by T_n.

[1]We will provide two formulas for the probability mass function (PMF) $\mathbb{P}(C_i(n) = \cdot)$, $1 \leq i \leq n$, using two separate approaches. In §3.2 we use a formula for the cardinality of T_n given by Simion to obtain the joint PMF of $(C_i(n))_{i \leq n}$, and then we sum over $n-1$ of the components to obtain marginal PMFs. In §3.3 we provide another formula for the marginal PMFs $\mathbb{P}(C_i(n) = \cdot)$ by mirroring a proof given by Goncharov in the analogue problem in which we replace NC_n with the set S_n of permutations on $[n]$ and denote by $C_i(n)$ the number of cycles of length i in the uniformly random permutation $\tau(n) \sim \mathrm{Unif}(S_n)$. The significance of our second formula is that it has only one index of summation. Moreover, any combinatorial structure on $[n]$ that is bijective to NC_n necessarily has the same joint and marginal PMFs as those defined on NC_n. As an application, in §3.4 we use a bijection between NC_n and the set of Dyck paths (sequences of n up-steps and n down-steps) with $2n$ steps to derive the joint PMF and marginal PMFs of the block counts of consecutive up-steps in a uniformly random chosen Dyck path. For example, the block counts of the consecutive up-steps for the Dyck path $UUUDDUDUUDDD$ are $3, 1$, and 2.

3.2 The Joint and Marginal Distributions of the Block Counts $(C_i(n))_{i \leq n}$ and $C_i(n)$

The number of elements in T_n is equal to the number of integer partitions of n, and the latter is denoted by $p(n)$. This is due to the fact that there are $p(n)$ distinct cycle types among elements in S_n, and each cycle type corresponds to an element in NC_n in the following way. Given the cycle type $(m_1, m_2, \ldots, m_n)$ for a permutation, we can identify a noncrossing

[1]In our computations, we often replace $C_i(n)$ with C_i.

partition with type $(m_1, m_2, \ldots, m_n)$ by forming m_1 blocks

$$\{1\}, \{2\}, \ldots, \{m_1\}$$

of size 1 and m_2 blocks

$$\{m_1+1, m_1+2\}, \{m_1+3, m_1+4\}, \ldots, \{m_1+2m_2-1, m_1+2m_2\}$$

of size 2, and so on. Note that each noncrossing partition of type $(m_1, \ldots, m_n) \in T_n$ must satisfy $1 \leq \sum_{i \leq n} m_i \leq n$ and $\sum_{i \leq n} i m_i = n$. Further, if the number of blocks, $\sum m_i$, is at least 2, then the number of noncrossing partitions in NC_n with type $(m_1, \ldots, m_n) \in T_n$ is (p. 371 of [10])

$$\#\mathrm{NC}_n \left(\text{type}\left((m_i)_{i \leq n}\right)\right) = \frac{n\,(n-1)\cdots\left(n - \sum_{i \leq n} m_i + 2\right)}{m_1! m_2! \cdots m_n!}. \tag{3.1}$$

Note that if $\sum_{i \leq n} m_i = 1$, then we must have $m_i = 0$ for $i < n$ and $m_n = 1$ due to the fact that $\sum_{i \leq n} i m_i = n$. Moreover, there is only one noncrossing partition of the type $(0, 0, \ldots, 1)$, namely $\{\{1, 2, \ldots, n\}\}$. Therefore, the joint PMF the random vector $(C_i(n))_{i \leq n}$ is given by

$$\mathbb{P}\left(C_i = m_i \text{ for all } i \leq n\right) \overset{(3.1)}{=} \frac{\mathbf{1}_{\{\sum m_i = 1\}} + \mathbf{1}_{\{\sum m_i \geq 2\}} \frac{n(n-1)\cdots\left(n - \sum_{i \leq n} m_i + 2\right)}{m_1! m_2! \cdots m_n!}}{\#\mathrm{NC}_n} \tag{3.2}$$

$$= (n+1)\, \frac{\mathbf{1}_{\{\sum m_i = 1\}} + \mathbf{1}_{\{\sum m_i \geq 2\}} \frac{n(n-1)\cdots\left(n - \sum_{i \leq n} m_i + 2\right)}{m_1! m_2! \cdots m_n!}}{\binom{2n}{n}}, \tag{3.3}$$

where $\mathbf{1}_{\{.\}}$ denotes an indicator random variable. The marginal distribution of $C_i(n)$ may be obtained by summing over the marginals $C_j(n)$ for all $j \neq i$; i.e.,

$$\mathbb{P}\left(C_i = k\right) = \sum_{m_j \geq 0, j \neq i} \mathbb{P}\left(C_i = k, \text{ and } C_j = m_j \text{ for all } j \neq i\right) \tag{3.4}$$

$$= \frac{\mathbf{1}_{\{\sum m_i=1,k=1,i=n\}}}{\#\mathrm{NC}_n} + \mathbf{1}_{\{\sum m_i\geq 2\}} \sum_{m_j\geq 0, j\neq i} \mathbb{P}\left(C_i = k, \text{ and } C_j = m_j \text{ for all } j\neq i\right) \tag{3.5}$$

$$\overset{(3.3)}{=} \frac{n+1}{\binom{2n}{n}} \left(\mathbf{1}_{\{\sum m_i=1\}} + \mathbf{1}_{\{\sum m_i\geq 2\}} \sum_{m_j\geq 0, j\neq i} \frac{n\left(n-1\right)\cdots\left(n-\sum_{i\leq n} m_i+2\right)}{m_1!\cdots k!\cdots m_n!} \right). \tag{3.6}$$

3.3 A Simpler Formula for $\mathbb{P}\left(C_i\left(n\right)=k\right)$

The proof of the following theorem was obtained by slightly modifying a proof given by Goncharov (see pages 12-13 of [2]) in which he gives the PMF of the number of cycles of length i in a uniformly random permutation in S_n. Unlike equation (3.6), the following formula for the PMF of $C_i\left(n\right)$ has only one index of summation and the formula does not require us to list all elements of T_n.

Theorem 6. *The marginal distribution of block counts is given by*

$$\mathbb{P}\left(C_i\left(n\right)=k\right) = \sum_{l=0}^{\lfloor \frac{n}{i} \rfloor - k} (-1)^l \binom{k+l}{l} \frac{\#NC_{n-(k+l)i}}{\#NC_n} \binom{n}{\left(k+l\right)i}. \tag{3.7}$$

Proof. Consider the set I_i of all possible blocks of size i, formed with elements chosen from $[n]$, so that $\#I_i = \binom{n}{i}$. For each $\alpha \in I_i$, consider the "property" G_α of having block α. I.e., $G_\alpha \coloneqq \{\rho \in \mathrm{NC}_n : \alpha \text{ is a block of } \rho\}$. Then $\#G_\alpha = \#\mathrm{NC}_{n-i}$ since the elements of $[n]$ not in α must be partitioned among themselves.

In order to apply the inclusion-exclusion formula, we need to calculate the term s_r, which is the sum of the probabilities $\mathbb{P}\left(\bigcap_{i=1}^r G_{\alpha_i}\right)$ of the $r-$fold intersection $\bigcap_{i=1}^r G_{\alpha_i}$ of properties G_{α_i}, summing over all sets of r distinct properties $G_{\alpha_i}, 1\leq i\leq r$. There are two cases to consider. If the r properties are indexed by r blocks having no elements in common, then the intersection specifies how ri elements are placed into blocks, and there are $\#\mathrm{NC}_{n-ri}\mathbf{1}_{\{ri\leq n\}}$ partitions in the intersection. There are $\binom{n}{ri}$ such intersections. For the other case, some two distinct properties $G_\alpha, G_{\alpha'}$ have some element ρ in common, so no non-crossing partition can have both of these properties, and the $r-$fold

intersection is empty. Thus,

$$s_r = \frac{\#\mathrm{NC}_{n-ri}\mathbf{1}_{\{ri\leq n\}}}{\#\mathrm{NC}_n}\binom{n}{ri}.$$

Finally, the inclusion-exclusion series for the number of noncrossing partitions having exactly k properties is (see p. 106 of [5]) given by

$$\sum_{l\geq 0}(-1)^l\binom{k+l}{l}s_{k+l} = \sum_{l=0}^{\lfloor n/i\rfloor-k}(-1)^l\binom{k+l}{l}\frac{\#\mathrm{NC}_{n-(k+l)i}\mathbf{1}_{\{ri\leq n\}}}{\#\mathrm{NC}_n}\binom{n}{(k+l)\,i}.$$

□

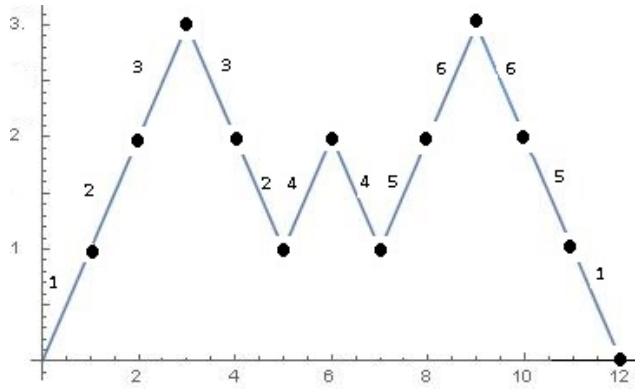
3.4 The Joint and Marginal PMFs for Number of Up-steps in a Uniformly Random Dyck Path

Definition. A **Dyck path** of length $2n$ is a diagonal lattice path from $(0,0)$ to $(2n,0)$ consisting of n up-steps (along the vector $\langle 1,1\rangle$) and n down-steps (along the vector $\langle 1,-1\rangle$), such that the path never goes below the x-axis. Let us denote the set of all Dyck paths of $2n$ steps by D_n.

A bijection between NC_n and D_n is given in [17]. Given a noncrossing partition ρ with k blocks, $1 \leq k \leq n$, arrange the blocks of ρ in increasing order of their maximal elements. Given $1 \leq k \leq n$, let m_i denote these k maximal elements and define $m_0 = 0$. Let $h_i, 1 \leq i \leq k$, be the corresponding block sizes. Then the pairs $(m_i - m_{i-1}, h_i)_{i=1}^k$ determines a Dyck path of length $2n$ as follows. For each i, we first put $m_i - m_{i-1}$ up-steps and then we put h_i down-steps to follow these up steps immediately. As an example, consider the noncrossing partition $\rho \in \mathrm{NC}_6$ given by $\rho = \{\{2,3\},\{4\},\{1,5,6\}\}$ so that $m_1 = 3, m_2 = 4, m_3 = 6, h_1 = 2, h_2 = 1$, and $h_3 = 3$. Then we have $(m_i - m_{i-1}, h_i)_{i=1}^k = ((3,2),(1,1),(2,3))$, which corresponds to the Dyck path $UUUDDUDUUDDD \in D_6$. For the inverse mapping, we label the up steps of the Dyck paths by enumerating them from left to right (so that the kth up-step is labeled k). Next assign to each

down step the same label of its matching up-step. The numbers assigned on a maximal sequence of down steps from a block of the desired noncrossing partition. As a follow-up to our example above, consider the Dyck path $UUUDDUDUUDDD$.

Figure 3.1: The Dyck path $UUUDDUDUUDDD \in D_6$.



The maximal sequence of continuous down-steps is given by $\{2,3\},\{4\},\{1,5,6\}$, and the corresponding noncrossing partition is ρ.

Let $\rho \sim \text{Unif}(D_n)$. We call a sequence of i consecutive U's form a block of size i. Consider the component counting process $(U_i(n))_{i\leq n}$ of ρ, where $U_i(n)$ is the number of blocks of size i. Due to the bijection specified above and equations (3.3) and (3.7), we have the following theorem (recall that T_n denotes the set of all of block types among elements of NC_n).

Theorem 7. *Let $\rho(n)$ be a uniformly distributed random variable in the set of all Dyck paths D_n, and let $U_i(n)$ denote the number of blocks of i consecutive up-steps in $\rho(n)$. Given $(m_i)_{i\leq n} \in T_n$ and $1 \leq k \leq n$, we have*

$$\mathbb{P}\left((U_i(n))_{i\leq n} = (m_i)_{i\leq n}\right) = (n+1)\,\frac{\mathbf{1}_{\{\sum m_i=1\}} + \mathbf{1}_{\{\sum m_i\geq 2\}}\frac{n(n-1)\cdots\left(n-\sum_{i\leq n} m_i+2\right)}{m_1!m_2!\cdots m_n!}}{\binom{2n}{n}}$$

and

$$\mathbb{P}(U_i(n)=k) = \sum_{l=0}^{\lfloor \frac{n}{i}\rfloor - k} (-1)^l \binom{k+l}{l} \frac{\#NC_{n-(k+l)i}}{\#NC_n} \binom{n}{(k+l)\,i}.$$

This concludes our discussion on the component counts of a uniformly random Dyck path.

Chapter 4

On the Prime-Power Process of a Uniformly Distributed Random Integer

4.1 Introduction

Given $n \in \mathbb{N}$, let $N(n)$ denote a uniformly chosen integer in $[n] \coloneqq \{1, 2, \ldots, n\}$ with the prime factorization $N(n) = \prod_{p \leq n} p^{C_p(n)}$. Note that the prime-power process $(C_p(n))_{p \leq n}$ is dependent since the values of $C_p(n)$, $p \leq n$, must be chosen in conjunction such that $N(n) \leq n$. Let us also consider a random variable $M(n) = \prod_{p \leq n} p^{Z_p}$, where $(Z_p)_{p \leq n}$ is a sequence of independent geometric random variables such that $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$ – i.e.,

$$\mathbb{P}(Z_p = k) = \frac{1 - \frac{1}{p}}{p^k} \tag{4.1}$$

for each nonnegative integer k. It is known (p. 3 of [1]) that as $n \to \infty$, the process $(C_p(n))_{p \leq n}$ converges to $(Z_p)_{p \leq n}$ in distribution. Our goal is to resolve the following conjecture, which appears on page 17 of [1].

Conjecture. (Arratia) For all $n \geq 1$, it is possible to construct $N(n)$ uniformly distributed from

1 to n, $M(n)$ and a prime $P(n)$ such that we always[1] have

$$N(n) \,|\, M(n) P(n).$$

Definition. Let X and Y be random variables defined on probability spaces[2] $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ and $(\Omega_Y, \mathcal{F}_Y, \mathbb{P}_Y)$.[3] A **coupling** of X and Y is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in which there exists random variables X' and Y' such that X' has the same distribution as X and Y' has the same distribution as Y.

For each of the random variables X considered in this chapter, X and X' will share the same range. Thus, for each coupling of X and Y, the definition implies that there exists a joint probability mass function (PMF) $p(x,y) \coloneqq \mathbb{P}(X' = x, Y' = y)$ whose marginal distributions satisfy $\sum_{y:p(x,y)>0} p(x,y) = \mathbb{P}_X(x)$ and $\sum_{x:p(x,y)>0} p(x,y) = \mathbb{P}_Y(y)$. Equivalently[4], $\mathbb{P}_{X'}(x) = \mathbb{P}_X(x)$ and $\mathbb{P}_{Y'}(y) = \mathbb{P}_Y(y)$ for all x in the range of X and all y in the range of Y.

The conjecture is true if and only if it is possible to couple $M(n)$ and $N(n)$ such that

$$\sum_{p \le n} (C_p(n) - Z_p)^+ \le 1,$$

where $(\cdot)^+$ denotes the positive part.

[1] I.e., for any realization of the vector $(M(n), N(n))$ there exists a prime $P(n) \le n$ such that $N(n)$ divides $M(n) P(n)$.

[2] In our discrete setting, we can consider probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ of the following form: (i) Ω is a nonempty set that is finite or countably infinite; (ii) $\mathcal{F}$ is the power set of Ω; and (iii) the probability measure $\mathbb{P}$ is defined as $\mathbb{P}(E) = \sum_{e \in E} p(e)$ for all $E \in F$, where p is a probability mass function – i.e., $p : \Omega \to [0,1]$ with $\sum_{s \in \Omega} p(s) = 1$.

[3] The probability measure $\mathbb{P}_X$ is defined by $\mathbb{P}_X(i) = \mathbb{P}(X = i)$.

[4] When describing a particular coupling of X and Y, we often write X and Y instead of X' and Y', respectively.

Figure 4.1: An $\infty \times n$ joint PMF for $(M(n), N(n))$.

$N(n)$	1	2	$\cdots$	j	$\cdots$	n
$M(n)$						
$2^0 3^0 \cdots p^0$						
$2^1 3^0 \cdots p^0$						
$\vdots$						
i				$\mathbb{P}(M(n) = i, N(n) = j)$		
$\vdots$						

4.2 The Joint PMF for Arratia's Conjecture

If we are to successively construct a joint PMF $\mathbb{P}(M(n) = \cdot, N(n) = \cdot)$ satisfying the conjecture, the joint PMF must equal 0 when $M(n)$ and $N(n)$ satisfy $\sum_{p \leq n} (C_p(n) - Z_p)^+ > 1$. Moreover, we know the sums along any row and any column since these marginal distributions are known.

4.2.1 Row Sums and Column Sums in the Joint PMF

If there exists a joint PMF satisfying the conjecture, then there are restrictions on the entries due to the fact that we have formulas for the marginal PMFs:

- For each $1 \leq j \leq n$, the sum along the jth column is

$$\mathbb{P}(N(n) = j) = \frac{1}{n}.$$

- For each $i = \prod_{p \leq n} p^{a_p} \in \mathbb{N}$, applying independence of the variables $Z_p, p \leq n$, we see that the

sum along row i is

$$\begin{aligned}
\mathbb{P}(M(n)=i) &= \mathbb{P}(Z_p = a_p \text{ for all } p \le n) \\
&= \prod_{p\le n} \mathbb{P}(Z_p = a_p) \\
&\overset{(4.1)}{=} \prod_{p\le n} \frac{\left(1-\frac{1}{p}\right)}{p^{a_p}} \\
&= \frac{\prod_{p\le n}\left(1-\frac{1}{p}\right)}{i}.
\end{aligned}$$

Figure 4.2: Row and column sums in the joint PMF.

$M(n)$ \ $N(n)$	1	2	$\cdots$	j	$\cdots$	n	Row sum
$2^0 3^0 \cdots p^0$							$\frac{\prod_{p\le n}\left(1-\frac{1}{p}\right)}{1}$
$2^1 3^0 \cdots p^0$							$\frac{\prod_{p\le n}\left(1-\frac{1}{p}\right)}{2}$
$\vdots$							
i				$\mathbb{P}(M(n)=i, N(n)=j)$			$\frac{\prod_{p\le n}\left(1-\frac{1}{p}\right)}{i}$
$\vdots$							
Column sum	$\frac{1}{n}$	$\frac{1}{n}$		$\frac{1}{n}$		$\frac{1}{n}$	

4.2.2 Pivots

In order to prove the conjecture, we must have $\mathbb{P}(M(n)=i, N(n)=j) = 0$ when $N(n)$ does not divide $M(n)P(n)$ for any prime $P(n) \le n$. The latter happens if and only if $\sum_{p\le n}(C_p(n) - Z_p)^+ \ge 2$. This motivates the following definition.

Definition. Given a column label $j = \prod_{p\le n} p^{a_p}$ and a row label $i = \prod_{p\le n} p^{b_p}$, we call the pair (i,j) a **pivot** if $\sum_{p\le n}(a_p - b_p)^+ \ge 2$.

Example 8. Consider $n = 10$. The prime factorizations of $1, 2, \ldots, 10$ consists of powers of $2, 3, 5,$ and 7. Figure 4.3 shows some of the pivots in the first 10 columns.

Figure 4.3: Some pivots in the case $n = 10$. Column labels and row labels are replaced by their corresponding sequence of prime powers $(C_p(n))_{p \le n}$ and $(Z_p)_{p \le n}$, respectively.

M \ N	0, 0, 0, 0	1, 0, 0, 0	0, 1, 0, 0	2, 0, 0, 0	0, 0, 1, 0	1, 1, 0, 0	0, 0, 0, 1	3, 0, 0, 0	0, 2, 0, 0	1, 0, 1, 0
0, 0, 0, 0				0		0		0	0	0
1, 0, 0, 0								0	0	
0, 1, 0, 0				0				0		0
2, 0, 0, 0									0	
0, 0, 1, 0				0		0		0	0	
1, 1, 0, 0								0		
0, 0, 0, 1				0		0		0	0	0
3, 0, 0, 0									0	
0, 2, 0, 0				0				0		0
1, 1, 0, 0								0		
2, 1, 0, 0										
1, 0, 0, 1								0	0	
⋮										

Given columns $j = \prod_{p \le n} p^{c_p(n)}$ and $k = \prod_{p \le n} p^{c'_p(n)}$, we seek a way to compare the corresponding sets of row labels for which column j or column k must be 0. Based on the definition of a pivot, we compare the probability measures of the sets

$$\left\{ \prod_{p \le n} p^{a_p} \in \mathbb{N} : \sum_{p \le n} \left(c_p(n) - a_p \right)^+ > 1 \right\},$$

$$\left\{ \prod_{p \le n} p^{a_p} \in \mathbb{N} : \sum_{p \le n} \left(c'_p(n) - a_p \right)^+ > 1 \right\}.$$

Definition. The **pivot mass** in column j, denoted $\mathcal{PM}(j)$, is

$$\mathcal{PM}(j) \coloneqq \sum_{i:(i,j) \text{ is a pivot}} \mathbb{P}(M(n) = i).$$

The pivot mass is a number between 0 and 1 that is independent of n (see Theorem 10) and tells us the proportion of the column mass that is contained in the pivot positions when one uses the independence[5] coupling of $N(n)$ and $M(n)$ given by

$$\mathbb{P}(M(n) = i, N(n) = j) = \mathbb{P}(M(n) = i)\,\mathbb{P}(N(n,) = j).$$

We observe that $\mathcal{PM}(j) = 0$ if and only if $j = 1$ or j is prime (which also follows by Theorem 10), and $\mathcal{PM}(j) < 1$ since, for example, (j, j) is not a pivot for any $j \in [n]$.

We can extend this definition to any number of columns.

Definition. Let $L \subseteq \{1, 2, \ldots, n\}$. The **pivot mass** of L is

$$\mathcal{PM}(L) \coloneqq \sum_{\substack{i:(i,j) \text{ is a pivot} \\ \forall j \in L}} \mathbb{P}(M(n) = i).$$

As we will see in §4.3, the concept of pivot mass can be used to address the existence of couplings with known marginals and specified constraints.

4.2.3 A Formula for the Pivot Mass $\mathcal{PM}(j)$

Before we provide a formula for $\mathcal{PM}(j)$, for any $1 \leq j \leq n$, let us motivate its derivation (solution 2 below) with an example.

Example 9. Given $n = 6$, let us compute $\mathcal{PM}(6)$. In both of the solutions below, we make use of

[5]Note that the independence coupling is not a coupling that resolves Arratia's Conjecture due to the fact that the PMF corresponding to the independence coupling is never 0.

geometric series – for any $k > 1$,

$$\sum_{a \geq 0} \left(1 - \frac{1}{k}\right) \frac{1}{k^a} = 1. \tag{4.2}$$

Solution 1. (Direct computation) Let us compute the pivot mass directly. Since $n = 6$, each row label i is of the form $i = 2^{a_2}3^{a_3}5^{a_5}$. We have a pivot located at $(i, 6)$ if and only if i is not divisible by both 2 and 3, so all of the pivot mass lies in column 6 lies within rows of the form $i = 5^{a_5}$, where $a_5 \geq 0$. Therefore, we have

$$\begin{aligned} \mathcal{PM}(6) &= \prod_{p \leq 6} \left(1 - \frac{1}{p}\right) \cdot \sum_{a_5 \geq 0} \frac{1}{5^{a_5}} \\ &\overset{(4.2)}{=} \frac{1}{2}\frac{2}{3}\frac{4}{5} \cdot \frac{5}{4} \\ &= \frac{1}{3}. \end{aligned}$$

However, it is not always convenient to compute the pivot mass based on its definition, especially when there are more than two distinct primes or higher prime powers.

Solution 2. Instead of computing this directly, let us compute the complement of the pivot mass within column 6 and then subtract this quantity from 1. In order for $(i, 6)$ to not be a pivot, i must be divisible by either 2 or 3. The expression $\frac{1}{2}\frac{2}{3}\frac{4}{5}\sum_{a_2,a_3 \geq 0} \frac{1}{2 \cdot 2^{a_2}3^{a_3}}$ counts the mass in column 6 along row labels that are divisible by 2, and $\frac{1}{2}\frac{2}{3}\frac{4}{5}\sum_{a_2,a_3 \geq 0} \frac{1}{3 \cdot 2^{a_2}3^{a_3}}$ counts the mass in column 6 along row labels that are divisible by 3. However, if we add these two quantities, then we are counting the mass in rows i which are divisible by both 2 and 3 twice, so we subtract a sum along rows which are divisible by 6. That is,

$$\begin{aligned} \mathcal{PM}(6) &= 1 - \left(\frac{1}{2}\frac{2}{3}\frac{4}{5} \sum_{a_2,a_3,a_5 \geq 0} \left(\frac{1}{2 \cdot 2^{a_2}3^{a_3}5^{a_5}} + \frac{1}{3 \cdot 2^{a_2}3^{a_3}5^{a_5}} - \frac{1}{2 \cdot 3 \cdot 2^{a_2} \cdot 3^{a_3} \cdot 5^{a_5}}\right)\right) \\ &\overset{(4.2)}{=} 1 - \left(\frac{1}{2} + \frac{1}{3} - \frac{1}{6}\right) \\ &= \frac{1}{3}. \end{aligned}$$

Let $\Omega(j)$ denote the sum of the distinct prime factors of j and let $\omega(j)$ denote the number of the distinct prime factors of j. We have the following theorem.

Theorem 10. *Given a positive integer j we have*

$$\mathcal{PM}(j) = 1 - \frac{1 + \Omega(j) - \omega(j)}{j}.$$

Proof. Fix a natural number $n \geq j$. Instead of computing the pivot mass in column j directly, we compute the complement and subtract from 1. To not have a pivot at (i, j), we need i to have at most one less prime factor or prime power than j. So for each prime factor p_l dividing j, with multiplicity $a_l \geq 1$, i can be any multiple of $\frac{j}{p_i}$. The quantity $\left(\prod_{p \leq n}\left(1 - \frac{1}{p}\right)\right)\sum_{c_2,\ldots,c_p \geq 0}\sum_{p_l | j}\frac{p_l}{N \cdot 2^{c_2}\cdots p^{c_p}}$ counts each of the rows which are divisible by all such values of i, but rows which are divisible by N are counted $\omega(j)$ times while they should only be counted once each. Therefore, we subtract $\frac{\omega(j)-1}{j \cdot 2^{c_2}\cdots p^{c_p}}$ from the sum over the prime divisors of j to obtain

$$\begin{aligned}\mathcal{PM}(j) &= 1 - \left(\prod_{p \leq n}\left(1 - \frac{1}{p}\right)\right)\sum_{c_2,\ldots,c_p \geq 0}\left(\sum_{p_l | j}\frac{p_l}{j \cdot 2^{c_2}\cdots p^{c_p}} - \frac{\omega(j) - 1}{j \cdot 2^{c_2}\cdots p^{c_p}}\right) \\ &\overset{(4.2)}{=} 1 - \frac{1 + \Omega(j) - \omega(j)}{j}.\end{aligned}$$

□

4.3 Stating Arratia's Conjecture in terms of $\mathcal{PM}$

Let S and T be complete separable metric spaces. Denote by p_S the projection of $S \times T$ onto S. Let W be a nonempty closed subset of $S \times T$ and $\varepsilon \geq 0$. The following[6] result is Theorem 11 of [12].

Theorem 11. *(Strassen) There is a probability measure λ in $S \times T$ with marginals μ and ν such*

[6] We thank Anthony Quas [9] for suggesting the use of Hall's Marriage Theorem. Strassen's Theorem is a variant of the marriage theorem.

that $\lambda(W) \geq 1-\varepsilon$*, if and only if for all closed sets* $L \subseteq T$

$$\nu(L) \leq \mu(p_s(W \cap (S \times L))) + \varepsilon. \tag{4.3}$$

Let us define

$$S = (\mathbb{Z}_{\geq 0})^n,$$
$$T = \left\{ (a_i)_{i \leq n} \in (\mathbb{Z}_{\geq 0})^n : \sum_{i \leq n} i a_i = n \right\},$$

corresponding to the set of row labels and the set of column labels respectively. Our goal is to apply Theorem 11 with $\varepsilon = 0$ and

$$W = P^c,$$
$$L = L(n),$$
$$\mu_{(n,x)}(i) = \mathbb{P}(M(n) = i), i \in S,$$
$$\nu_n(j) = \mathbb{P}(N(n) = j), j \in T,$$
$$\lambda = p,$$

where $P = \{(i,j) \in S \times T : (i,j) \text{ is a pivot}\}$, $L(n)$ denotes an arbitrary subset of T, and p is our desired joint PMF, with marginals corresponding to $M(n)$ and $N(n)$, such that $(i,j) \in P$ implies $p(i,j) = 0$. In terms of pivot mass, we can express Strassen's inequality (4.3) as

$$\mathcal{PM}(L) \leq 1 - \frac{\#L}{n} \tag{4.4}$$

for each subset $L \subseteq [n]$; however, this conjecture has yet to be resolved. In an attempt to search for a counter example or to stimulate more interest in the problem, we can apply computer code provided by Carl Woll [16] that checks (4.4) for any subset L. The code provided by Woll can be simplified due to the fact that there is an underlying poset structure among the column labels; e.g., $\mathcal{PM}(4,8) = \mathcal{PM}(4)$, so many of the computations in the current code can be ignored. The final

section of this chapter specifies a poset structure among the columns.

4.4 Future Work: Applying a Post Structure on the Column Labels to Simplify the Computer Code

Given column labels a and b, we write $a \prec b$ if for each row label i, (i, b) is a pivot whenever (i, a) is a pivot. With respect to this poset structure, not all column labels are related; e.g., $4 \not\prec 9$ and $9 \not\prec 4$.

Note that if $a \prec b$, then $\mathcal{PM}(a, b) = \mathcal{PM}(a)$. As a result, when checking (4.4) and given $a, b \in L$ with $a \prec b$, we can replace $\mathcal{PM}(L)$ with $\mathcal{PM}(L - \{b\})$, where $L - \{b\}$ denotes a difference of sets. Our future work will consists of using the poset structure to simplify Woll's code. After that, we hope to either (i) find a counterexample to Arratia's Conjecture or (ii) stimulate more interest in the problem by providing more evidence that the conjecture is true.

Chapter 5

Establishing Stochastic Domination via Couplings

5.1 Introduction

Given $n \in \mathbb{N}$ and a uniformly distributed random variable $N(n)$ taking values in $[n] := \{1, 2, \ldots, n\}$, denote its prime factorization by $N(n) = \prod_{p \leq n} p^{C_p(n)}$. It is known (p.3 of [1]) that the process $(C_p(n))_{p \leq n}$ converges in distribution to a process $(Z_p)_p = (Z_2, Z_3, Z_5, \ldots)$ consisting of independent components Z_p with $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$. We would like to prove, in some sense, that $C_p(n)$ is dominated by Z_p for each prime $p \leq n$.

Definition. Let X and Y be random variables (defined on the same probability space) with cumulative distribution functions F_X and F_Y. We say that X is **pointwise dominated** by Y if $X \leq Y$ – i.e., $X(s) \leq Y(s)$ for each outcome s. We say that X is **stochastically dominated** by Y, denoted $X \leq^D Y$, if $F_X(a) \geq F_Y(a)$ for each $a \in \mathbb{R}$.

The main result is the following theorem.

Theorem 12. *Given $n \in \mathbb{N}$ and any prime $p \leq n$, we have $C_p(n) \leq^D Z_p$.*

Our first proof of Theorem 12 will follow once we prove $\mathbb{P}(Z_p \leq k) = \inf_{n>1} \mathbb{P}(C_p(n) \leq k)$ for all $k \geq 0$. Our second proof consists of the main contribution to this chapter and is based on the fact that stochastic domination can be established by proving the existence of couplings (see §4.1 for the definition of a coupling) with specified constraints.

5.2 Proving Stochastic Domination by the Definition

Given $n \in \mathbb{N}$, any prime $p \leq n$ and a nonnegative integer k, we prove $\mathbb{P}(C_p(n) \leq k) \geq \mathbb{P}(Z_p \leq k)$.

Lemma 13. *For each prime $p \leq n$, we have $\mathbb{P}(Z_p \leq k) = \inf_{n>1} \mathbb{P}(C_p(n) \leq k)$.*

Proof of Lemma 13. Note that $C_p(n) > k$ if and only if p^{k+1} divides $N(n)$. The number of multiples of p^{k+1} that are in $[n]$ is $\left\lfloor \frac{n}{p^{k+1}} \right\rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Therefore, $\mathbb{P}(C_p(n) \leq k) = 1 - \mathbb{P}(C_p(n) > k) = 1 - \frac{\left\lfloor \frac{n}{p^{k+1}} \right\rfloor}{n}$. Direct computation gives

$$\begin{aligned}
\inf_{n>1} \mathbb{P}(C_p(n) \leq k) &= \inf_{n>1} \left(1 - \frac{\left\lfloor \frac{n}{p^{k+1}} \right\rfloor}{n}\right) \\
&= 1 - \sup_{n>1} \left(\frac{\left\lfloor \frac{n}{p^{k+1}} \right\rfloor}{n}\right) \\
&= 1 - \frac{1}{p^{k+1}} \\
&= \mathbb{P}(Z_p \leq k),
\end{aligned}$$

where the latest equation follows from the fact that $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$; i.e., $\mathbb{P}(Z_p = k) = \frac{1-\frac{1}{p}}{p^k}$. This proves the lemma. □

Proof of Theorem 12. Let k be a nonnegative integer. By Lemma 13 we have $\mathbb{P}(C_p(n) \leq k) \geq \mathbb{P}(Z_p \leq k)$ for each n and each prime $p \leq n$. Thus, $C_p(n) \leq^D Z_p$ for each n and each prime $p \leq n$ by definition of stochastic domination. □

5.3 Proving Theorem 12 by Coupling $C_p(n)$ and Z_p with $C_p(n) \leq Z_p$ Pointwise

Theorem 3.1 of [14] states that $X \leq^D Y$ if and only if there exists a coupling of X and Y with $X \leq Y$ pointwise. We use this theorem to provide another proof of Theorem 12 by constructing a coupling of $C_p(n)$ and Z_p with $C_p(n) \leq Z_p$ pointwise.

Proof of Theorem 12. Applying Theorem 3.1 of [14], our goal is to construct a joint probability mass function (PMF) $p_{i,j} := \mathbb{P}(Z_p = i, C_p(n) = j)$ such that $p_{i,j} = 0$ when $i < j$. Consider a prime $p \leq n$. Since $\prod_{p \leq n} p^{C_p(n)} \leq n$, we have $p^{C_p(n)} \leq n$ for all $p \leq n$. Thus, $C_p(n)$ has range $\{0, 1, \ldots, \lfloor \log_p n \rfloor\}$. Any coupling of Z_p and $C_p(n)$ has row and column sums determined by the distributions of Z_p and $C_p(n)$. As previously mentioned, $\mathbb{P}(Z_p = i) = \frac{1-\frac{1}{p}}{p^i}$. Note that $C_p(n) = j$ if and only if p^j divides $N(n)$ and p^{j+1} does not. There are $\left\lfloor \frac{n}{p^j} \right\rfloor$ many integer multiples of p^j in $[n]$ and $\left\lfloor \frac{n}{p^{j+1}} \right\rfloor$ many integer multiples of p^{j+1} in $[n]$. Since $N(n)$ is uniformly distributed, we have $\mathbb{P}(C_p = j) = \frac{\left\lfloor \frac{n}{p^j} \right\rfloor - \left\lfloor \frac{n}{p^{j+1}} \right\rfloor}{n}$.

Figure 5.1: The row sums and column sums of any coupling of Z_p and $C_p(n)$.

Z_p \ $C_p(n)$	0	1	2	$\cdots$	$\lfloor \log_p n \rfloor$	Row sum
0	$p_{0,0}$	$p_{0,1}$	$p_{0,2}$		$p_{0,\lfloor \log_p n \rfloor}$	$\frac{p-1}{p}$
1	$p_{1,0}$	$p_{1,1}$	$p_{1,2}$		$p_{1,\lfloor \log_p n \rfloor}$	$\frac{p-1}{p^2}$
2	$p_{2,0}$	$p_{2,1}$	$p_{2,2}$		$p_{2,\lfloor \log_p n \rfloor}$	$\frac{p-1}{p^3}$
$\vdots$				$\cdots$		$\vdots$
$\lfloor \log_p n \rfloor$	$p_{\lfloor \log_p n \rfloor,0}$	$p_{\lfloor \log_p n \rfloor,1}$	$p_{\lfloor \log_p n \rfloor,2}$		$p_{\lfloor \log_p n \rfloor,\lfloor \log_p n \rfloor}$	$\frac{p-1}{p^{\lfloor \log_p n \rfloor+1}}$
$\vdots$						$\vdots$
Column Sum	$1 - \frac{\lfloor \frac{n}{p} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p} \rfloor - \lfloor \frac{n}{p^2} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p^2} \rfloor - \lfloor \frac{n}{p^3} \rfloor}{n}$	$\cdots$	$\frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor - \left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor+1}} \right\rfloor}{n}$	

In order to ensure $C_p(n) \leq Z_p$ pointwise, we impose the condition $p_{i,j} = 0$ if $i < j$. As a consequence,

we have $\frac{\lfloor \log_p n \rfloor (\lfloor \log_p n \rfloor + 1)}{2}$ many necessary zeros located at each $p_{i,j}$ above the diagonal entries $p_{i,i}$.

Figure 5.2: If $i < j$, then $p_{i,j} = 0$.

Z_p \ $C_p(n)$	0	1	2	3	$\cdots$	$\lfloor \log_p n \rfloor$	Row sum
0	$p_{0,0}$	0	0	0		0	$\frac{p-1}{p}$
1	$p_{1,0}$	$p_{1,1}$	0	0		0	$\frac{p-1}{p^2}$
2	$p_{2,0}$	$p_{2,1}$	$p_{2,2}$	0		0	$\frac{p-1}{p^3}$
$\vdots$						0	$\vdots$
$\lfloor \log_p n \rfloor$	$p_{\lfloor \log_p n \rfloor, 0}$	$p_{\lfloor \log_p n \rfloor, 1}$	$p_{\lfloor \log_p n \rfloor, 2}$	$p_{\lfloor \log_p n \rfloor, 3}$		$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor}$	$\frac{p-1}{p^{\lfloor \log_p n \rfloor + 1}}$
$\vdots$							$\vdots$
Column Sum	$1 - \frac{\lfloor \frac{n}{p} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p} \rfloor - \lfloor \frac{n}{p^2} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p^2} \rfloor - \lfloor \frac{n}{p^3} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p^3} \rfloor - \lfloor \frac{n}{p^4} \rfloor}{n}$	$\cdots$	$\frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor - \left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor + 1}} \right\rfloor}{n}$	

In what follows, we construct a coupling by placing at most two nonzero entries in each row and in each column except the final column. The resulting joint PMF will be of the following form.

Figure 5.3: The pattern of zeros and nonzeros in the coupling.

Z_p \ $C_p(n)$	0	1	2	3	$\cdots$	$\lfloor \log_p n \rfloor - 1$	$\lfloor \log_p n \rfloor$	Row sum
0	$p_{0,0}$	0	0	0		0	0	$\frac{p-1}{p}$
1	$p_{1,0}$	$p_{1,1}$	0	0		0	0	$\frac{p-1}{p^2}$
2	0	$p_{2,1}$	$p_{2,2}$	0		0	0	$\frac{p-1}{p^3}$
3	0	0	$p_{3,2}$	$p_{3,3}$	$\cdots$	0	0	$\frac{p-1}{p^4}$
$\vdots$	$\vdots$				$\ddots$		$\vdots$	$\vdots$
$\lfloor \log_p n \rfloor$	0	0	0	0		$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor - 1}$	$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor}$	$\frac{p-1}{p^{\lfloor \log_p n \rfloor + 1}}$
$\vdots$	0	0	0	0		0	0	$p_{\lfloor \log_p n \rfloor + 1, \lfloor \log_p n \rfloor}$
	0	0	0	0		0	0	$p_{\lfloor \log_p n \rfloor + 2, \lfloor \log_p n \rfloor}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
Column Sum	$1 - \frac{\lfloor \frac{n}{p} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p} \rfloor - \lfloor \frac{n}{p^2} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p^2} \rfloor - \lfloor \frac{n}{p^3} \rfloor}{n}$	$\frac{\lfloor \frac{n}{p^3} \rfloor - \lfloor \frac{n}{p^4} \rfloor}{n}$	$\cdots$	$\frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor - 1}} \right\rfloor - \left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor}{n}$	$\frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor - \left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor + 1}} \right\rfloor}{n}$	

Since $p_{0,0}$ is the only nonzero entry in the 0th row, we necessarily have

$$p_{0,0} = 1 - \frac{1}{p}. \tag{5.1}$$

Let us define $p_{1,0}$ such that $p_{0,0} + p_{1,0} = \text{column 0 sum} = 1 - \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n}$. Then

$$p_{1,0} = \left(1 - \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n}\right) - p_{0,0} \overset{(5.1)}{=} \frac{1}{p} - \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n}. \tag{5.2}$$

Since $p_{1,0}$ and $p_{1,1}$ are the only nonzero entries in row labeled 1, we have $p_{1,0} + p_{1,1} = \frac{p-1}{p^2}$; so

$$p_{1,1} = \left(\frac{1}{p} - \frac{1}{p^2}\right) - p_{1,0} \overset{(5.2)}{=} \frac{\left\lfloor \frac{n}{p} \right\rfloor}{n} - \frac{1}{p^2}. \tag{5.3}$$

Define $p_{2,1}$ such that $p_{1,1} + p_{2,1} = \text{column 1 sum} = \frac{\left\lfloor \frac{n}{p} \right\rfloor - \left\lfloor \frac{n}{p^2} \right\rfloor}{n}$. Then

$$p_{2,1} = \frac{\left\lfloor \frac{n}{p} \right\rfloor - \left\lfloor \frac{n}{p^2} \right\rfloor}{n} - p_{1,1} \overset{(5.3)}{=} \frac{1}{p^2} - \frac{\left\lfloor \frac{n}{p^2} \right\rfloor}{n}. \tag{5.4}$$

Similarly, we have

$$p_{2,2} = \frac{p-1}{p^3} - p_{2,1} \overset{(5.4)}{=} \left(\frac{1}{p^2} - \frac{1}{p^3}\right) - \left(\frac{1}{p^2} - \frac{\left\lfloor \frac{n}{p^2} \right\rfloor}{n}\right) = \frac{\left\lfloor \frac{n}{p^2} \right\rfloor}{n} - \frac{1}{p^3}.$$

If we continue in this fashion, each row and column will have at most two nonzero entries except the last column (since we have yet to assign values to the entries $p_{i,\lfloor \log_p n \rfloor}$ for $i > \lfloor \log_p n \rfloor$). By construction,

$$p_{i,i} = \frac{\left\lfloor \frac{n}{p^i} \right\rfloor}{n} - \frac{1}{p^{i+1}}$$

for $i = 0, 1, \ldots, \lfloor \log_p n \rfloor$ and

$$p_{i+1,i} = \frac{1}{p^{i+1}} - \frac{\left\lfloor \frac{n}{p^{i+1}} \right\rfloor}{n}$$

for $i = 0, 1, \ldots, \lfloor \log_p n \rfloor - 1$. By design, each column has the desired sum except the final column,

and rows $0, 1, 2, \ldots, \lfloor \log_p n \rfloor - 1$ have the desired row sum.

Before we deal with the final column, let us show that the terms we have constructed are nonnegative: The case $i = 0$ has been resolved by equation (5.1) so consider $1 \leq i \leq \lfloor \log_p n \rfloor$. Applying the division algorithm, let $n = qp^i + r$, where $0 \leq r < p^i, q \in \mathbb{N}$. Then

$$\begin{aligned} p_{i,i} &= \frac{\left\lfloor \frac{n}{p^i} \right\rfloor}{n} - \frac{1}{p^{i+1}} \\ &= \frac{q}{n} - \frac{1}{p^{i+1}}, \end{aligned}$$

and the latest expression is positive due to the fact that $q \geq 1$ and $p^{i+1} \geq n$. Moreover,

$$\begin{aligned} p_{i+1,i} &= \frac{1}{p^{i+1}} - \frac{\left\lfloor \frac{n}{p^{i+1}} \right\rfloor}{n} \\ &\geq \frac{1}{p^{i+1}} - \frac{\frac{n}{p^{i+1}}}{n} \\ &= 0. \end{aligned}$$

At the moment, we have only assigned one entry in the final column, namely, $p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor}$. By construction, have

$$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor} = \frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor}{n} - \frac{1}{p^{\lfloor \log_p n \rfloor + 1}}.$$

Moreover, since in rows $i > \lfloor \log_p n \rfloor$ all entries are zero except in the final column, we necessarily have

$$p_{i, \lfloor \log_p n \rfloor} = \text{row } i \text{ sum} = \frac{p-1}{p^{i+1}}$$

when $i > \lfloor \log_p n \rfloor$. Summing over the nonzero entries in the final column, we have

$$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor} + \sum_{i > \lfloor \log_p n \rfloor} \frac{p-1}{p^{i+1}} = \text{column } \lfloor \log_p n \rfloor \text{ sum}$$

$$= \frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor - \left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor + 1}} \right\rfloor}{n}$$
$$= \frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor}{n},$$

where the last equality is due to the fact that $\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor + 1}} \right\rfloor = 0$. Thus, we also have

$$p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor} = \frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor}{n} - \sum_{i > \lfloor \log_p n \rfloor} \frac{p-1}{p^{i+1}}$$
$$= \frac{\left\lfloor \frac{n}{p^{\lfloor \log_p n \rfloor}} \right\rfloor}{n} - \frac{1}{p^{\lfloor \log_p n \rfloor + 1}},$$

which is consistent with the value of $p_{\lfloor \log_p n \rfloor, \lfloor \log_p n \rfloor}$ determined by our construction. This completes the construction of the coupling of $C_p(n)$ and Z_p with $C_p(n) \leq Z_p$ pointwise.

□

Chapter 6

On the Dependence of the Component Counting Process of a Uniform Random Variable

6.1 Introduction

Our results regard the component counting process of a discrete uniform random variable in a combinatorial structure, and these results are provided by establishing the existence of couplings of random variables (see §4.1 for the definition of a coupling). In particular, given $n \in \mathbb{N}$ we provide couplings, with some constraints, of a uniformly distributed random variable N, necessarily consisting of a dependent component process, with another random variable M having the following properties:

1. M has infinite range.
2. M and N have components of sizes between 1 and n.
3. The components of M are independent and nonnegative.

Recall Arratia's Conjecture (which is stated in §4.1 and concerns a uniformly distributed variable $N(n) = \prod_{p \leq n} p^{C_p} \in [n] \coloneqq \{1, \ldots, n\}$ and a random variable $M(n) = \prod_{p \leq n} p^{Z_p}$, where the Z_p's are independent for $p \leq n$ and $Z_p \sim \text{Geometric}\left(\frac{1}{p}\right)$), and also recall that the conjecture is true if and only if there exists a coupling of M and N with $\sum_{p \leq n} (C_p(n) - Z_p)^+ \leq 1$. We impose an analogous constraint on the couplings provided in this chapter, but now we will point out some differences between these couplings and the coupling conjectured by Arratia. First, we drop the requirement that $N(n) \in [n]$; rather, from now on we let $N(n)$ denote a uniform variable in a combinatorial structure over $[n]$ (these structures are defined in §6.1.1). Instead of a prime power process $(C_p(n))_{p \leq n}$, we consider a component counting process $(C_i(n))_{1 \leq i \leq n}$ (here i is any positive integer less than or equal n) of $N(n)$ which satisfies $\sum_{i \leq n} iC_i(n) = n$ – the latter equation is not always true for the prime power process $(C_p(n))_{p \leq n}$ of a uniformly distributed variable over $[n]$. In Arratia's conjecture, there is a natural candidate for $M(n)$ since the prime power process $(C_p(n))_{p \leq n}$ converges in distribution to the process $(Z_p)_{p \leq n}$ described above. However, in each of the examples considered in this chapter, we take advantage of the fact that in either an assembly, multiset, or selection, there exists infinitely many processes $(Z_i(n, x))_{i \leq n}$, indexed by some positive real parameter x, consisting of independent variables $Z_i(n, x), 1 \leq i \leq n$, which furnish natural candidates for a random variable $M(n, x)$ to be compared with a uniform random variable $N(n)$ (see equation (6.1) below).

The combinatorial structures listed in §6.1.1 provide the frameworks in which we obtain our couplings. These are the combinatorial structures for which there exists values of x for which there is a candidate for $M(n, x)$. Theorem 15, the main result of this chapter, is stated in §6.1.2. In §6.2, we describe how our constraints force a significant proportion of the entries of a prospective joint probability mass function (PMF) of our variables to be 0. In §6.3, we introduce the notion of pivot mass, which depends on the constraints placed on the desired joint distribution. A formula for the pivot mass is stated in Theorem 18. Some properties of the pivot mass are proved in §6.3 and §6.4. In our combinatorial structures, we there exists values of x which make the pivot mass arbitrarily small. In §6.5, we apply results on the pivot mass and a theorem proved by Strassen to prove Theorem 15, thereby proving the existence of our couplings.

6.1.1 Three Major Combinatorial Structures

All couplings constructed in this chapter involve a discrete uniform random variable in any one of the following three combinatorial classes. An **assembly** A_n is an example of a combinatorial structure in which the set $[n]$ is partitioned into blocks and for each block of size i one of m_i possible structures is chosen. An example of an assembly is the collection of set partitions of $[n]$, in which case $m_i = 1$ for $i \leq n$ (since the order of the elements in a particular block is irrelevant – i.e., once i numbers $n_1, \ldots, n_i \in [n]$ are chosen and placed in a box of size i, there is a unique block consisting of these i elements). Moreover, for set partitions of $[n]$, we have $\#A_n = B_n$, the nth Bell number. Another example of an assembly is the set S_n of permutations of $[n]$, in which case $m_i = (i-1)!$ (since there are $(i-1)!$ distinct cycles of length i among i chosen numbers $n_1, \ldots, n_i \in [n]$) for $i \leq n$. Further, for permutations of $[n]$, we have $\#A_n = n!$. A **multiset** A_n is a pair $([n], m)$, where $m : A \to \mathbb{N}$ is a function that gives the multiplicity $m(a)$ of each element $a \in [n]$. Equivalently (see Meta-example 2.2 of §2.2 of [2]), the integer n is partitioned into parts, and for each part of size i, one of the m_i objects of weight i is chosen. In the example of integer partitions of a positive integer n, we have $m_i = 1$ (for each part of size i, we have only $m_i = 1$ choice for the size of i) for $1 \leq i \leq n$. When A_n is the set of integer partitions of n, we have $\#A_n = p(n)$, where p is the integer partition function. **Selections** are similar to multisets, but now we require all parts to be distinct. An example of a selection is the set of all integer partitions of a positive integer n with distinct parts. In the case of integer partitions with distinct parts, we have $\#A_n = q(n)$, where q is the integer partition function with distinct parts. To simplify the notation, let us define $k_n \coloneqq \#A_n$ for each of these structures.

These three structures are characterized by the following generating relations between k_n and m_i. Assemblies are characterized by

$$\sum_{n \geq 0} \frac{(k_n) z^n}{n!} = \exp\left(\sum_{i \geq 1} \frac{m_i z^i}{i!} \right),$$

multisets are characterized by

$$\sum_{n\geq 0}(k_n)\,z^n=\prod_{i\geq 1}\left(1-z^i\right)^{-m_i},$$

and selections are characterized by

$$\sum_{n\geq 0}(k_n)\,z^n=\prod_{i\geq 1}\left(1+z^i\right)^{m_i}$$

(§2.2 of [2]). Revisiting the example of an assembly in which A_n denotes the set of all set partitions of $[n]$ (so that $m_i=1$ for $1\leq i\leq n$), it is known (e.g., pp. 20-23 of [15]) that the nth Bell number B_n satisfies the generating equation $\sum_{n\geq 0}\frac{B_n}{n!}z^n=\exp\left(e^z-1\right)$, and the right hand side may be expressed as $\exp\left(\sum_{i\geq 1}\frac{z^i}{i!}\right)$.

6.1.2 Couplings of Random Variables

In each of the assembly, multiset, and selection settings, our methods of arriving at our desired couplings are similar. We start by considering $N(n)\sim \text{Unif}(A_n)$. Given $i\leq n$, if we denote by $C_i(n)$ the number components of $N(n)$ of size i, then $0\leq C_i(n)\leq n$ and $\sum_{i\leq n}iC_i(n)=n$. In particular, the variables $C_i(n), i\leq n$, are dependent and their distributions are determined by the uniform variable $N(n)$. The process $(C_i(n))_{i\leq n}=(C_1(n),\ldots,C_n(n))$ is called the **component counting process** of $N(n)$.

Example 14. In the example $A_n=S_n$, the term $C_i(n)$ is the number of cycles of $N(n)$ of length i, and $(C_1(n),\ldots,C_n(n))$ is often referred to as the *cycle type* of $N(n)$. In the example for which A_n is the collection of set partitions of $[n]$, $C_i(n)$ is the number of blocks of $N(n)$ of size i. In the example for which A_n is the set of integer partitions of n, $C_i(n)$ is the number of i's in the integer partition $N(n)$ of n.

In each of these combinatorial settings, there exists an infinite family $\left((Z_i(n,x))_{i\leq n}\right)_x$, parametrized by positive values of x (specifically, $x>0$ for assemblies, $x\in(0,1)$ for multisets, and $x\in(0,\infty)$

for selections) of infinite sequences $(Z_i(n,x))_{i\leq n}$ of nonnegative integer-valued independent random variables $Z_i(n,x)$ for which

$$\mathcal{L}(C_1(n),\ldots,C_n(n)) = \mathcal{L}\left(Z_1(n,x),\ldots,Z_n(n,x)\,\middle|\,\sum_{i\leq n} iZ_i(n,x) = n\right) \tag{6.1}$$

(§2.3 of [2]). Equation (6.1) states that the probability that the vector $(C_1(n),\ldots,C_n(n))$ belongs to some region $\Gamma \in \mathbb{R}^n$ (where Γ is an element of the n-fold direct product $\prod_{i\leq n}\mathcal{B}(\mathbb{R})$ of the Borel $\sigma-$algebra on $\mathbb{R}$) is the same as the conditional probability that the vector $(Z_1(n,x),\ldots,Z_n(n,x))$ belongs to Γ if we condition on the event $\left\{\sum_{i\leq n} iZ_i(n,x) = n\right\}$. For a fixed x, we consider another random variable $M(n,x)$ whose component counting process[1] is given by $(Z_i(n,x))_{i\leq n}$, so the distribution of $M(n,x)$ is determined by the independent process $(Z_i(n,x))_{i\leq n}$. The main result of this chapter is the following theorem.[2]

Theorem 15. *Let $n \in \mathbb{N}$ and suppose A_n denotes an assembly, multiset, or a selection among elements of $[n]$. Given $N(n) \sim Unif(A_n)$ with component counting process $(C_i(n))_{i\leq n}$, there exists a positive real number $x(n)$ for which, when $x > x(n)$, there exists a process $(Z_i(n,x))_{i\leq n}$ of nonnegative independent random variables satisfying (6.1) such that we can couple $M(n,x)$ and $N(n)$ with*

$$\sum_{i\leq n}(C_i(n) - Z_i(n,x))^+ \leq 1. \tag{6.2}$$

6.2 The Joint Mass Distribution of $(M(n,x), N(n))$

For some fixed value of x, if we are to successively construct a joint PMF $p(\cdot,\cdot)$ with marginal distributions corresponding to $M(n,x)$ and $N(n)$ for which inequality (6.2) holds, we must ensure that $p(\cdot,\cdot) \coloneqq \mathbb{P}(M(n,x) = \cdot, N(n) = \cdot) = 0$ whenever the sum $\sum_{i\leq n}(C_i(n) - Z_i(n,x))^+$ is at least

[1] For fixed x, since the variables $Z_i(n,x), i \leq n$, are independent, it is not always true that $\sum_{i\leq n} iZ_i(n,x) = n$. Therefore, the variable $M(n,x)$ does not always correspond to an element of A_n.

[2] To simplify the notation, we will often replace $Z_i(n,x)$ with Z_i, replace $C_i(n)$ with C_i, replace $N(n)$ with N, and replace $M(n,x)$ with M.

2. We can index the joint distribution by using the range of $N(n)$ and the range of $M(n,x)$ for the column labels and row labels, respectively. In particular, we can label the columns with the range of $(C_i(n))_{i\leq n}$ in lexicographic order. Since we have infinitely many row labels, for each $m \in \mathbb{Z}_{\geq 0}$, we apply the lexicographic ordering on all elements $(m_1,\ldots,m_n) \in (\mathbb{Z}_{\geq 0})^n$ with $\sum_{i\leq n} m_i = m$, starting with $m = 0$ (we start with $m = 0$ since the $Z_i(n,x)$'s are non-negative). With respect to this ordering, we will often enumerate the columns by $1,2,\ldots,k_n$ and the rows by $1,2,\ldots$.

The following example shows that it is possible for several elements of A_n to have the same component process (hence the same column label). Note that in the setting of Arratia's conjecture, it is impossible for two columns to have the same label – the uniqueness of prime factorization in $\mathbb{N}$ ensures that each $(C_p(n))_{p\leq n}$ uniquely determines $N(n)$.

Example 16. Fix $n = 3$ and consider the assembly $A_3 = S_3$ of permutations of $\{1,2,3\}$. The elements of S_3 are $1, (1\,2), (1\,3), (2\,3), (1\,2\,3), (1\,3\,2)$, and their respective component counts are $(3,0,0), (1,1,0), (1,1,0), (1,1,0), (0,0,1), (0,0,1)$. For any $n \in \mathbb{N}$, Cauchy proved that there are $\frac{n!}{(i_1!\cdots i_n!1^{i_1}\cdots n^{i_n})}$ permutations in S_n with cycle type $(i_1,\ldots,i_n)$, so this gives the number of elements in S_n with the component counting process $(C_i(n))_{i\leq n} = (i_1,\ldots,i_n)$.

For our purposes, when we have multiple columns with the same component counting process, we enumerate these columns in any order. The reason that we do not combine these into one column with larger probability mass is due to the fact we are coupling $M(n,x)$ and $N(n)$ instead of coupling the two processes $(C_i(n))_{i\leq n}$ and $(Z_i(n,x))_{i\leq n}$ – i.e., two columns with the same label correspond to different values of $N(n)$. For the interested reader, equations $(2.2), (2.3)$, and (2.4) in §2.2 of [2] give the number of columns with a given column label $(a_1,\ldots,a_n)$ for each of our combinatorial structures.

In each of these three settings, there are additional constraints on any joint PMF of $M(n,x)$ and $N(n)$ since the marginal distributions are known:

- The sum along column $N(n) = j, 1 \leq j \leq k_n$, is

$$\mathbb{P}(N(n) = j) = 1/k_n.$$

- The sum along the$m$th row, $m \in \mathbb{N}$, labeled $(Z_i(n,x))_{i \leq n} = (m_i)_{i \leq n}$ is

$$\mathbb{P}\left((Z_i(n,x))_{i \leq n} = (m_i)_{i \leq n}\right) = \prod_{i \leq n} \mathbb{P}(Z_i(n,x) = m_i),$$

where the latest equation is due to the independence of the process $(Z_i(n,x))_{i \leq n}$.

6.3 Pivot Mass

Given columns j and k, with corresponding component counts $(c_i(n))_{i \leq n}$ and $(c'_i(n))_{i \leq n}$, we seek a way to compare the corresponding sets of row labels in which column j or k must be 0. Based on the definition of a pivot, we compare the probability measures of the sets

$$\left\{\prod_{i \leq n} p_i^{a_i} \in \mathbb{N} : \sum_{i \leq n} (c_i(n) - a_i)^+ > 1\right\},$$
$$\left\{\prod_{i \leq n} p_i^{a_i} \in \mathbb{N} : \sum_{i \leq n} (c'_i(n) - a_i)^+ > 1\right\}.$$

I.e., (6.2) is true for all of our desired couplings, so we measure the probability that $M(n,x)$ takes on a value i for which column j or k has a required 0 in row i. This motivates the following definition.

Definition. We call the pair (i,j), corresponding to the ith row label $(Z_i(n,x))_{i \leq n}$ and the jth column label $(C_i(n))_{i \leq n}$, a **pivot** if $\sum_{i \leq n} (C_i(n) - Z_i(n,x))^+ > 1$. Denote the set of all pivots by P. The **pivot mass** in column $N(n) = j$ is defined as

$$\mathcal{PM}(j) = \mathcal{PM}_{(n,x)}(j) \coloneqq \sum_{i:(i,j) \in P} \mathbb{P}(M(n,x) = i).$$

Given a subset $L(n)$ of column labels of $[n]$, the pivot mass in $L(n)$ is defined as

$$\mathcal{PM}(L(n)) = \mathcal{PM}_{(n,x)}(L(n)) \coloneqq \sum_{\substack{i:(i,j)\in P \\ \forall j\in L(n)}} \mathbb{P}(M(n,x)=i).$$

If we define $I_j \coloneqq \{\text{row labels } i : (i,j) \in P\}$, $I_{L(n)} \coloneqq \{\text{row labels } i : (i,j) \in P \text{ for all } j \in L(n)\}$ and let $\mathbb{P}_M$ denote the distribution function[3] of $M(n,x)$, then

$$\mathcal{PM}(j) = \mathbb{P}_M(I_j),$$
$$\mathcal{PM}(L(n)) = \mathbb{P}_M\left(I_{L(n)}\right).$$

Theorem 18 gives a formula for $\mathcal{PM}(j)$. Fortunately, due to the role of the parameter x, it is not necessary to derive a formula for $\mathcal{PM}(L(n))$ in order to prove Theorem 15. The fact that $\mathcal{PM}(L(n)) \leq \mathcal{PM}(j)$ for any $j \in L(n)$ will be sufficient.

Figure 6.1: If (i_0, j_0) is a pivot, then our desired joint distribution table should have a 0 in the (i_0, j_0) entry.

$N(n)$	1	$\cdots$	j	$\cdots$	j_0	$\cdots$	k_n	Row sum
$M(n,x)$								
1								$\mathbb{P}(M(n,x)=1)$
2								$\mathbb{P}(M(n,x)=2)$
$\vdots$								$\vdots$
i_0					0			$\mathbb{P}(M(n,x)=i_0)$
$\vdots$								$\vdots$
i			$\mathbb{P}(M(n,x)=i, N(n)=j)$					$\mathbb{P}(M(n,x)=i)$
$\vdots$								$\vdots$
Column sum	$\frac{1}{k_n}$	$\cdots$	$\frac{1}{k_n}$	$\cdots$	$\frac{1}{k_n}$	$\cdots$	$\frac{1}{k_n}$	

[3] That is, $\mathbb{P}_M(A) = \mathbb{P}(M(n,x) \in A)$.

Example 17. Revisiting the example $A_3 = S_3$, let us illustrate some key features of a desired joint mass distribution of $(M(3,x), N(3))$. For convenience, in Figure 6.2 and in the proof of the following theorem, we simplify the notation by writing $C_i(n) = C_i$, $Z_i(n,x) = Z_i$, $M(n,x) = M$ and $N(n) = N$.

Figure 6.2: A desired coupling of $M(3,x)$ and $N(3)$ should have a zero at any location $\left((Z_i(3,x))_{i\leq 3}, (C_i(3))_{i\leq 3}\right)$ satisfying $\sum_{i\leq 3}(C_i(3) - Z_i(3,x))^+ > 1$.

M \ N	$(0,0,1)$	$(0,0,1)$	$(1,1,0)$	$(1,1,0)$	$(1,1,0)$	$(3,0,0)$	Row sum
$(0,0,0)$			0	0	0	0	$\mathbb{P}(M=0)$
$(0,0,1)$			0	0	0	0	$\mathbb{P}(M=1)$
$(0,1,0)$						0	$\mathbb{P}(M=2)$
$(1,0,0)$						0	$\mathbb{P}(M=3)$
$(0,0,2)$			0	0	0	0	$\mathbb{P}(M=4)$
$(0,1,1)$						0	$\mathbb{P}(M=5)$
$(0,2,0)$						0	$\mathbb{P}(M=6)$
$(1,0,1)$						0	$\mathbb{P}(M=7)$
$(1,1,0)$						0	$\mathbb{P}(M=8)$
$(2,0,0)$							$\mathbb{P}(M=9)$
$\vdots$							$\vdots$
Column sum	1/6	1/6	1/6	1/6	1/6	1/6	
$\mathcal{PM}(N)$	0	0	$\mathbb{P}(Z_1 = Z_2 = 0)$	$\mathbb{P}(Z_1 = Z_2 = 0)$	$\mathbb{P}(Z_1 = Z_2 = 0)$	$\mathbb{P}(Z_1 \leq 1)$	

Each column with a pivot contains infinitely many pivots. E.g., in Figure 6.2, column $(3,0,0)$ has a pivot in any row of the form (a,b,c) with $a \in \{0,1\}, b,c \geq 0$. Columns labeled $(1,1,0)$ have a pivot in any row of the form $(0,0,l)$ for any $l \in \mathbb{Z}_{\geq 0}$. Moreover, note that the independence of the process $(Z_i(3,x))_{i\leq 3}$ allows us to distribute $\mathbb{P}$ through the parentheses in the row sums and pivot mass expressions. For example, the sum of the probability mass along row $M = 0$ is

$$\begin{aligned}\mathbb{P}(M(3,x) = 0) &= \mathbb{P}(Z_i(3,x) = 0 \text{ for all } i \leq 3) \\ &= \mathbb{P}(Z_1(3,x) = 0)\,\mathbb{P}(Z_2(3,x) = 0)\,\mathbb{P}(Z_3(3,x) = 0),\end{aligned}$$

and the pivot mass in column with label $(1,1,0)$ is

$$\mathcal{PM}((1,1,0)) = \mathbb{P}(Z_1(3,x) = Z_2(3,x) = 0)$$

$$= \mathbb{P}\left(Z_1\left(3,x\right)=0\right)\mathbb{P}\left(Z_2\left(3,x\right)=0\right).$$

The actual value of the row sum and pivot masses depends on the choice of the process $\left(Z_i\left(3,x\right)\right)_{i\leq 3}$. In §6.4, we mention several choices for such processes $\left(Z_i\left(3,x\right)\right)_{i\leq 3}$ which will satisfy equation (6.1).

The following theorem plays a key role in the proof of Theorem 15. Moreover, the notion of pivot mass introduced in this section may be generalized; in a particular setting, one should define pivot mass based on the constraints required of their desired coupling. It is both a combinatorial and probabilistic object since it is a sum of probability masses indexed by the counting constraint (6.2).

Theorem 18. *Consider a fixed column label $N\left(n\right)\in A_n$ and denote its component counting process by $\left(C_i\left(n\right)\right)_{i\leq n}$. Its pivot mass is*

$$\begin{aligned}\mathcal{PM}\left(N\left(n\right)\right)=&1-\sum_{j\leq n}\left(\mathbf{1}_{\{C_j>0\}}\left(1-\mathbb{P}\left(Z_j\leq C_j-2\right)\right)\prod_{\substack{i\neq j,\\ i\leq n}}\left(1-\mathbb{P}\left(Z_i\leq C_i-1\right)\right)\right)\\ &+\left(\sum_{i\leq n}\mathbf{1}_{\{C_i>0\}}-1\right)\prod_{i\leq n}\left(1-\mathbb{P}\left(Z_i\leq C_i-1\right)\right).\end{aligned}$$

Proof. Given $1\leq j\leq n$, let $\overrightarrow{e_j}$ denote the row vector of length n whose jth entry is 1 and whose other entries are 0. Given two vectors $\left(a_i\right)_{i\leq n},\left(b_i\right)_{i\leq n}$ in $\mathbb{R}^n$, we write $\left(a_i\right)_{i\leq n}\leq\left(b_i\right)_{i\leq n}$ if $a_i\leq b_i$ for each $i\leq n$. Since $\sum_{k=1}^{\infty}\mathbb{P}\left(M\left(n,x\right)=k\right)=1$, we have

$$\mathcal{PM}\left(N\right)=1-\sum_{k:(k,N)\notin P}\mathbb{P}\left(M=k\right). \tag{6.3}$$

We have the event equality

$$\left\{(M,N)\notin P\right\}=\left\{\exists j\leq n:\left(Z_i\right)_{i\leq n}\geq\left(C_i\right)_{i\leq n}-\overrightarrow{e_j}\cdot\mathbf{1}_{\{C_j>0\}}\right\}$$

since the pair (M,N) is a not pivot if and only if $Z_i\geq C_i$ for all i except possibly one value j with $Z_j=C_j-1$. Since each $Z_i,1\leq i\leq n$ is nonnegative, we can only have $Z_j=C_j-1$ when $C_j>0$. Note that if $Z_i\geq C_i$ for all i, then any j satisfies $\left(Z_i\right)_{i\leq n}\geq\left(C_i\right)_{i\leq n}-\overrightarrow{e_j}\cdot\mathbf{1}_{\{C_j>0\}}$. On

the other hand, if there exists a value j for which $Z_j = C_j - 1$ and $Z_i \geq C_i$ for all $i \neq j$, then $(Z_i)_{i\leq n} \geq (C_i)_{i\leq n} - \overrightarrow{e_j} \cdot \mathbf{1}_{\{C_j>0\}}$. Therefore, the right hand side of equation (6.3) is

$$1 - \sum_{k:(k,N)\notin P} \mathbb{P}(M = k) = 1 - \mathbb{P}\left(\exists j \leq n : (Z_i)_{i\leq n} \geq (C_i)_{i\leq n} - \overrightarrow{e_j} \cdot \mathbf{1}_{\{C_j>0\}}\right). \tag{6.4}$$

We rewrite the probability $\mathbb{P}\left(\exists j \leq n : (Z_i)_{i\leq n} \geq (C_i)_{i\leq n} - \overrightarrow{e_j} \cdot \mathbf{1}_{\{C_j>0\}}\right)$ by applying an inclusion-exclusion argument. Corresponding to any $j \leq n$ with $C_j > 0, Z_j \geq C_j - 1$, and $Z_i \geq C_i$ for $i \neq j$, we add the term $\mathbb{P}(Z_j \geq C_j - 1, \text{ and } Z_i \geq C_i \text{ for all } i \neq j)$. As a result, we have added those elements with $Z_i \geq C_i$ for all i a total of $\sum_{i=1}^n \mathbf{1}_{\{C_i>0\}}$ many times. Therefore, we compensate by subtracting the term $\left(\sum_{i=1}^n \mathbf{1}_{\{C_i>0\}} - 1\right)\mathbb{P}\left((Z_i)_{i\leq n} \geq (C_i)_{i\leq n}\right)$. Further, applying independence of the process $(Z_i)_{i\leq n}$, we have

$$\begin{aligned}\mathbb{P}(Z_j \geq C_j - 1 \text{ and } Z_i \geq C_i \text{ for all } i \neq j) &= \mathbb{P}(Z_j \geq C_j - 1)\,\mathbb{P}(Z_i \geq C_i \text{ for all } i \neq j) \\ &= \mathbb{P}(Z_j \geq C_j - 1) \prod_{\substack{i\neq j,\\ i\leq n}} \mathbb{P}(Z_i \geq C_i)\end{aligned}$$

and $\mathbb{P}\left((Z_i)_{i\leq n} \geq (C_i)_{i\leq n}\right) = \prod_{i\leq n} \mathbb{P}(Z_i \geq C_i)$. Thus, the right hand side of equation (6.4) becomes

$$1 - \sum_{j\leq n} \left(\mathbf{1}_{\{C_j>0\}} \mathbb{P}(Z_j \geq C_j - 1) \prod_{\substack{i\neq j,\\ i\leq n}} \mathbb{P}(Z_i \geq C_i) \right) + \left(\sum_{i\leq n} \mathbf{1}_{\{C_i>0\}} - 1 \right) \prod_{i\leq n} \mathbb{P}(Z_i \geq C_i). \tag{6.5}$$

Using the fact that $\mathbb{P}(Z_i \geq a) = 1 - \mathbb{P}((Z_i \leq a - 1))$, expression (6.5) becomes

$$\begin{aligned}1 - \sum_{j\leq n} &\left(\mathbf{1}_{\{C_j>0\}} (1 - \mathbb{P}(Z_j \leq C_j - 2)) \prod_{\substack{i\neq j,\\ i\leq n}} (1 - \mathbb{P}(Z_i \leq C_i - 1)) \right) \\ &+ \left(\sum_{i\leq n} \mathbf{1}_{\{C_i>0\}} - 1 \right) \prod_{i\leq n} (1 - \mathbb{P}(Z_i \leq C_i - 1)).\end{aligned}$$

□

The following result shows that only columns with label $(C_i(n))_{i\leq n} = \overrightarrow{e_n}$ have zero pivot mass. In this chapter, we will only apply the ($\Leftarrow$) part of the statement.[4]

Theorem 19. *For any nonempty collection $L(n)$ of column labels, $\mathcal{PM}(L(n)) = 0$ if and only if a column with label $(C_i(n))_{i\leq n} = \overrightarrow{e_n}$ belongs to $L(n)$.*

Proof. ($\Leftarrow$) Given any row label $(Z_i(n,x))_{i\leq n}$, the vector $(C_i(n))_{i\leq n} = \overrightarrow{e_n}$ satisfies

$$\begin{aligned}\sum_{i\leq n}(C_i(n) - Z_i(n,x))^+ &= (C_n(n) - Z_n(n,x))^+ \\ &= (1 - Z_n(n,x))^+ \\ &\leq 1.\end{aligned}$$

Thus, $\mathcal{PM}(\overrightarrow{e_n}) = 0$. Therefore, given $\overrightarrow{e_n} \in L(n)$, we have

$$\mathcal{PM}(L(n)) \leq \mathcal{PM}(\overrightarrow{e_n}) = 0.$$

($\Rightarrow$) Now suppose $\overrightarrow{e_n} \notin L(n)$. Recall that any column label $(C_i(n))_{i\leq n}$ satisfies $\sum_{i\leq n} iC_i(n) = n$. Since $\overrightarrow{e_n}$ is the only column label with $\sum_{i\leq n} C_i(n) = 1$, this gives us one of two cases for each column label in $L(n)$. Either (a) there exists some j with $C_j(n) \geq 2$ or (b) there exists distinct j,k with $C_j(n) \geq 1, C_k(n) \geq 1$. In case (a), using any row label $(Z_i(n,x))_{i\leq n}$ with $Z_j(n,x) = 0$, we have

$$\sum_{i\leq n}(C_i(n) - Z_i(n,x))^+ \geq C_j(n) - Z_j(n,x) \geq 2.$$

In case (b), we can take any $(Z_i(n,x))_{i\leq n}$ with $Z_j(n,x) = Z_k(n,x) = 0$ to ensure that

$$\sum_{i\leq n}(C_i(n) - Z_i(n,x))^+ \geq (C_j(n) - Z_j(n,x)) + (C_k(n) - Z_k(n,x)) \geq 2.$$

[4]Note that ($\Rightarrow$) implies that each column label other than $\overrightarrow{e_n}$ has pivots. Applying equations (2.2)–(2.4) in §2.2 of [2] (which give the number of columns with label $\overrightarrow{e_n}$ in each of these combinatorial settings) we can always determine the number of columns that contain pivots.

Since we have just showed that each column label other than $\overrightarrow{e_n}$ has a pivot, we use the fact that each of these columns has a pivot in the first row (labeled $(Z_i(n,x))_{i\leq n} = (0,0,\ldots,0)$). Note that $\mathbb{P}(M(n,x)=i) > 0$ for all distributions in this chapter (see §6.4), so we have

$$\mathcal{PM}(L(n)) \geq \mathbb{P}(M(n,x)=1) > 0.$$

□

6.4 Pivot Mass can be made Arbitrarily Small for Assemblies, Multisets, and Selections

The following condition on $\mathcal{PM}$ will be verified for our three combinatorial structures:

$$\forall n \in \mathbb{N}\ \forall \varepsilon > 0\ \exists x(n) : x > x(n) \implies \text{equation (6.1) holds and } \mathcal{PM}_{(n,x)}(\cdot) < \varepsilon. \tag{6.6}$$

6.4.1 Assemblies

In the assembly setting, we can take[5] $Z_i(n,x) \sim \text{Po}\left(\frac{m_i x^i}{i!}\right)$ for any $x > 0$ to obtain equation (6.1) (§2.3 of [2]). Recall that the cumulative distribution function (CDF) of a random variable $Z \sim \text{Po}(\lambda)$ is given by $\mathbb{P}(Z \leq k) = \frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$ for $k \in \mathbb{Z}_{\geq 0}$, where $\Gamma(a,b)$ is the **upper incomplete gamma function** – i.e., $\Gamma(a,b) = \int_b^\infty t^{a-1}e^{-t}dt$.

Lemma 20. *For a fixed $a > 0$, we have $\lim_{b\to\infty} \Gamma(a,b) = 0$.*

[5] Although the distribution of $Z_i(n,x)$ does not depend on n, the choice the process $(Z_i(n,x))_{i\leq n}$ satisfying (6.1) does depend on n. I.e., if $(Z_i(n,x))_{i\leq n}$ and $(Z_i(n+1,x))_{i\leq n+1}$ equal $(C_i(n))_{i\leq n}$ and $(C_i(n+1))_{i\leq n+1}$, respectively, conditional on the events $\left\{\sum_{i\leq n} iZ_i = n\right\}$ and $\left\{\sum_{i\leq n+1} iZ_i = n+1\right\}$, respectively, then we need not have $(Z_i(n,x))_{i\leq n} = (Z_i(n+1,x))_{i\leq n}$.

Proof. Since $\Gamma(a,0) = \Gamma(a)$ is convergent for $a > 0$, we have

$$\begin{aligned}\Gamma(a,b) &= \Gamma(a) - \int_0^b t^{a-1}e^{-t}dt \\ &\to \Gamma(a) - \Gamma(a) \text{ as } b \to \infty \\ &= 0.\end{aligned}$$

□

We can apply Lemma 20 and take $x \to \infty$ to obtain

$$\frac{\Gamma\left(C_i, \frac{m_i x^i}{i!}\right)}{(C_i - 1)!} \to 0 \text{ when } C_i > 0 \tag{6.7}$$

and

$$\frac{\Gamma\left(C_j - 1, \frac{m_j x^j}{j!}\right)}{(C_j - 2)!} \to 0 \text{ when } C_j > 1. \tag{6.8}$$

Therefore, Theorem 18 implies that $\mathcal{PM}(N(n))$ equals

$$\begin{aligned}&1 - \sum_{j \le n} \left(\mathbf{1}_{\{C_j > 0\}} \left(1 - \mathbf{1}_{\{C_j > 1\}} \frac{\Gamma\left(C_j - 1, \frac{m_j x^j}{j!}\right)}{(C_j - 2)!}\right) \prod_{\substack{i \neq j, \\ i \le n}} \left(1 - \mathbf{1}_{\{C_i > 0\}} \frac{\Gamma\left(C_i, \frac{m_i x^i}{i!}\right)}{(C_i - 1)!}\right)\right) \\ &+ \left(\sum_{i \le n} \mathbf{1}_{\{C_i > 0\}} - 1\right) \prod_{i \le n} \left(1 - \mathbf{1}_{\{C_i > 0\}} \frac{\Gamma\left(C_i, \frac{m_i x^i}{i!}\right)}{(C_i - 1)!}\right).\end{aligned}$$

If we let $x \to \infty$, we can apply (6.7) and (6.8) to deduce that

$$\begin{aligned}\mathcal{PM}(N(n)) \to 1 - &\sum_{j \le n} \left(\mathbf{1}_{\{C_j > 0\}} (1 - 0) \prod_{\substack{i \neq j, \\ i \le n}} (1 - 0)\right) \\ &+ \left(\sum_{i \le n} \mathbf{1}_{\{C_i > 0\}} - 1\right) \prod_{i \le n} (1 - 0)\end{aligned}$$

$$= 1 - \sum_{j \leq n} \mathbf{1}_{\{C_j > 0\}} + \left(\sum_{i \leq n} \mathbf{1}_{\{C_i > 0\}} - 1 \right)$$
$$= 0.$$

This verifies condition (6.6) for assemblies.

6.4.2 Multisets

In the multiset setting, we can take $Z_i(n, x) \sim \mathrm{NB}\left(m_i, x^i\right)$, for any $x \in (0,1)$, to obtain equation (6.1) (§2.3 of [2]). Recall that the CDF of $Z \sim \mathrm{NB}(r,p)$ is given by $\mathbb{P}(Z \leq k) = 1 - I_p(k+1, r)$, where I_p is the **regularized incomplete beta function**. That is, $I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$, where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$, defined for $\mathrm{Re}(a) > 0$ and $\mathrm{Re}(b) > 0$, is the **beta function** and $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,dt$ is the **incomplete beta function**.

Lemma 21. *Given $a > 0$, $\lim_{x \to 1} I_x(a,b) = 1$.*

Proof. We have

$$\begin{aligned}
\lim_{x \to 1} I_x(a,b) &= \lim_{x \to 1} \frac{B(x;a,b)}{B(a,b)} \\
&= \lim_{x \to 1} \frac{\int_0^x t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt} \\
&= \frac{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt} \\
&= 1.
\end{aligned}$$

□

Applying Theorem 18, $\mathcal{PM}(N(n))$ equals

$$1 - \sum_{j \leq n} \left(\mathbf{1}_{\{C_j > 0\}} \left(1 - \mathbf{1}_{\{C_j > 1\}} \left(1 - I_{x^j}(C_j - 1, m_j)\right)\right) \prod_{\substack{i \neq j, \\ i \leq n}} \left(1 - \mathbf{1}_{\{C_i > 0\}} \left(1 - I_{x^i}(C_i, m_i)\right)\right) \right)$$

$$+\left(\sum_{i\leq n}\mathbf{1}_{\{C_i>0\}}-1\right)\prod_{i\leq n}\left(1-\mathbf{1}_{\{C_i>0\}}\left(1-I_{x^i}\left(C_i,m_i\right)\right)\right).$$

Taking $x\to 1$ and applying Lemma 21, we have

$$\begin{aligned}\mathcal{PM}\left(N\left(n\right)\right)&\to 1-\sum_{j\leq n}\left(\mathbf{1}_{\{C_j>0\}}\left(1-\mathbf{1}_{\{C_j>1\}}\left(1-1\right)\right)\prod_{\substack{i\neq j,\\ i\leq n}}\left(1-\mathbf{1}_{\{C_i>0\}}\left(1-1\right)\right)\right)\\&\quad+\left(\sum_{i\leq n}\mathbf{1}_{\{C_i>0\}}-1\right)\prod_{i\leq n}\left(1-\mathbf{1}_{\{C_i>0\}}\left(1-1\right)\right)\\&=1-\sum_{j\leq n}\mathbf{1}_{\{C_j>0\}}+\left(\sum_{i\leq n}\mathbf{1}_{\{C_i>0\}}-1\right)\\&=0,\end{aligned}$$

which verifies condition (6.6) for multisets.

6.4.3 Selections

In the selection setting, we can take $Z_i\left(n,x\right)\sim\mathrm{Bin}\left(m_i,\frac{x^i}{1+x^i}\right)$ with $0<x<\infty$ in order to obtain equation (6.1) (§2.3 of [2]). In our case, we are taking $p=\frac{x^i}{1+x^i}$, so $p\to 1$ if and only if $x\to\infty$. Recall that the CDF of $Z\sim\mathrm{Bin}\left(n,p\right)$ is given by $\mathbb{P}\left(Z\leq k\right)=I_{1-p}\left(n-k,1+k\right)$. Applying Theorem 18, we can express $\mathcal{PM}\left(N\left(n\right)\right)$ as

$$1-\sum_{j\leq n}\left(\mathbf{1}_{\{C_j>0\}}\left(1-\mathbf{1}_{\{C_{j>1}\}}I_{1-p}\left(m_j-C_j+2,C_j-1\right)\right)\prod_{\substack{i\neq j,\\ i\leq n}}\left(1-\mathbf{1}_{\{C_i>0\}}I_{1-p}\left(m_i-C_i+1,C_i\right)\right)\right)$$
$$+\left(\sum_{i\leq n}\mathbf{1}_{\{C_i>0\}}-1\right)\prod_{i\leq n}\left(1-\mathbf{1}_{\{C_i>0\}}I_{1-p}\left(m_i-C_i+1,C_i\right)\right).$$

Lemma 22. *We have* $\lim_{p\to 1}I_{1-p}\left(n-k,1+k\right)=0$.

Proof.

$$\begin{aligned}\lim_{p\to 1} I_{1-p}(n-k,1+k) &= \lim_{p\to 1}\frac{B(1-p;n-k,1+k)}{B(n-k,1+k)}\\ &= \lim_{p\to 1}\frac{\int_0^{1-p} t^{n-k-1}(1-t)^k}{\int_0^1 t^{n-k-1}(1-t)^k}\\ &= 0.\end{aligned}$$

□

Applying Lemma 22, we see that

$$I_{1-p} = I_{1-\frac{x^i}{1+x^i}} \to 0 \tag{6.9}$$

if $x \to \infty$. Thus, we apply Theorem 18 and Lemma 22 while taking $x \to \infty$ to obtain

$$\mathcal{PM}(N(n)) \overset{(6.9)}{\to} 1 - \sum_{j\le n} \mathbf{1}_{\{C_j>0\}} + \left(\sum_{i\le n} \mathbf{1}_{\{C_i>0\}} - 1\right) = 0,$$

which verifies condition (6.6) for selections.

6.5 Using Pivot Mass to Provide Couplings

We apply Strassen's Theorem (Theorem 11) along with pivot mass to obtain desired couplings.

Proof of Theorem 15. Let us define

$$\begin{aligned}S &= (\mathbb{Z}_{\ge 0})^n,\\ T &= \left\{(a_i)_{i\le n} \in (\mathbb{Z}_{\ge 0})^n : \sum_{i\le n} i a_i = n\right\},\end{aligned}$$

corresponding to the set of row labels and the set of column labels, respectively, and endow both S

and T with the metric d on $\mathbb{Z}^n$ defined as

$$d\left(\left(x_i\right)_{i\leq n},\left(y_i\right)_{i\leq n}\right)\coloneqq\max_{i\leq n}\left|x_i-y_i\right|.$$

Since S is finite and T is countably infinite, both S and T are separable. In both S and T we have

$$\left(x_i\right)_{i\leq n}\neq\left(y_i\right)_{i\leq n}\implies d\left(\left(x_i\right)_{i\leq n},\left(y_i\right)_{i\leq n}\right)\geq 1 \tag{6.10}$$

since our n-tuples are integer-valued. Therefore, every Cauchy sequence in S (or in T) converges in S (or in T). Thus, S and T are complete. Our goal is to apply Theorem 11 with $\varepsilon=0$ and

$$\begin{aligned}W&=P^c,\\ L&=L\left(n\right),\\ \mu_{(n,x)}\left(i\right)&=\mathbb{P}\left(M\left(n,x\right)=i\right),i\in S,\\ \nu_n\left(j\right)&=\mathbb{P}\left(N\left(n\right)=j\right),j\in T,\\ \lambda&=p,\end{aligned}$$

where $P=\{(i,j)\in S\times T:(i,j)\text{ is a pivot}\}$, $L\left(n\right)$ denotes an arbitrary subset of T, and p is our desired joint PMF, with marginals corresponding to $M\left(n,x\right)$ and $N\left(n\right)$, such that $(i,j)\in P$ implies $p\left(i,j\right)=0$. Let us endow W with the metric obtained by restricting the metric

$$d_{S\times T}\left(\left(\left(s_i\right)_{i\leq n},\left(t_i\right)_{i\leq n}\right),\left(\left(s_i'\right)_{i\leq n},\left(t_i'\right)_{i\leq n}\right)\right)\coloneqq\max\left(d\left(\left(s_i\right)_{i\leq n},\left(s_i'\right)_{i\leq n}\right),d\left(\left(t_i\right)_{i\leq n},\left(t_i'\right)_{i\leq n}\right)\right) \tag{6.11}$$

on $S\times T$ to W. To show that W is closed, we first show that S and T are closed. The set T is closed since it is finite. Suppose that

$$\left(\left(s_i\left(k\right)\right)_{i\leq n}\right)_{k\in\mathbb{N}}$$

is a sequence of n-tuples $(s_i(k))_{i\leq n} \in S$ with

$$\lim_{k\to\infty} (s_i(k))_{i\leq n} = l_1$$

for some n-tuple $l_1 \in \mathbb{Z}^n$. To show that S is closed, it suffices to show that $l_1 \in S$. For all $\varepsilon' \in (0,1)$ there exists a constant $K \in \mathbb{N}$ such that

$$k > K \implies d\left((s_i(k))_{i\leq n}, l_1\right) < \varepsilon'.$$

Since $\varepsilon' < 1$, (6.10) implies

$$l_1 = (s_i(K+1))_{i\leq n},$$

so $l_1 \in S$. Therefore, S is closed. Now to show that W is closed in $S \times T$, suppose that

$$\left(\left(s_i(k)_{i\leq n}\right), (t_i(k))_{i\leq n}\right)_{k\in\mathbb{N}}$$

is a sequence of pairs

$$\left(\left(s_i(k)_{i\leq n}\right), (t_i(k))_{i\leq n}\right) \in W$$

of n-tuples $s_i(k)_{i\leq n} \in S, (t_i(k))_{i\leq n} \in T$ with

$$\lim_{k\to\infty} \left(\left(s_i(k)_{i\leq n}\right), (t_i(k))_{i\leq n}\right) = (l_1, l_2).$$

for some n-tuples $l_1, l_2 \in \mathbb{Z}^n$. Since S and T are closed, we have $l_1 \in S$ and $l_2 \in T$. For all $\varepsilon' \in (0,1)$ there exists a constant $K \in \mathbb{N}$ such that

$$k > K \implies d_{S\times T}\left(\left(s_i(k)_{i\leq n}\right), (l_1, l_2)\right) < \varepsilon'.$$

Therefore,

$$k > K \overset{(6.11)}{\Longrightarrow} d\left(\left(s_i\left(k\right)\right)_{i\leq n}, l_1\right), d\left(\left(t_i\left(k\right)\right)_{i\leq n}, l_2\right) < \varepsilon'.$$

Since $\varepsilon' < 1$, we have

$$k > K \overset{(11)}{\Longrightarrow} d\left(\left(s_i\left(k\right)\right)_{i\leq n}, l_1\right) = d\left(\left(t_i\left(k\right)\right)_{i\leq n}, l_2\right) = 0.$$

Therefore, applying (6.10) twice, we obtain

$$\left(l_1, l_2\right) = \left(\left(s_i\left(K+1\right)\right)_{i\leq n}, \left(t_i\left(K+1\right)\right)_{i\leq n}\right) \in W,$$

so W is a closed subset of $S \times T$. Further, $W \neq \emptyset$ since given any column label j, the pair (j, j) belongs to W. Note that the set $L\left(n\right)$ is a closed subset of the column labels since $L\left(n\right)$ is a finite. Moreover,

$$v_n\left(L\left(n\right)\right) = \mathbb{P}\left(N\left(n\right) \in L\left(n\right)\right) = \frac{\#L\left(n\right)}{k_n},$$

and

$$\begin{aligned}
\mu_{(n,x)}\left(p_s\left(W \cap \left(S \times L\right)\right)\right) &= \mathbb{P}\left(M \in p_S\left(W \cap \left(S \times L\left(n\right)\right)\right)\right) \\
&= \mathbb{P}\left(M \in p_S\left(P^c \cap \left(S \times L\left(n\right)\right)\right)\right) \\
&= \mathbb{P}\left(\exists j \in L\left(n\right) : \left(M, j\right) \notin P\right) \\
&= 1 - \mathcal{PM}_{(n,x)}\left(L\left(n\right)\right).
\end{aligned}$$

Therefore, (4.3) from Theorem 11 is equivalent to

$$\frac{\#L\left(n\right)}{k_n} \leq 1 - \mathcal{PM}_{(n,x)}\left(L\left(n\right)\right).$$

The latest inequality is equivalent to

$$\mathcal{PM}_{(n,x)}\left(L\left(n\right)\right) \leq 1 - \frac{\#L\left(n\right)}{k_n}. \tag{6.12}$$

By (6.6), the left hand side can be made arbitrarily small, so (6.12) holds when $1 - \frac{\#L(n)}{k_n} > 0$. When $1 - \frac{\#L(n)}{k_n} = 0$, we must have $L\left(n\right) = A_n$, so that $\mathcal{PM}_{(n,x)}\left(L\left(n\right)\right) = 0$ by Theorem 19. Therefore, by the conclusion of Theorem 11, there exists a joint probability measure p, with marginals $\mathbb{P}\left(M\left(n,x\right) = \cdot\right)$ and $\mathbb{P}\left(N\left(n\right) = \cdot\right)$, such that $p\left(W\right) = 1$. I.e, the probability of having no pivot in this joint distribution is 1. Hence, the proof of Theorem 15 is complete. □

Chapter 7

On the Prime Powers of a Zeta-Distributed Random Variable

7.1 Introduction

Given $s > 1$, the Zeta(s) distribution is a probability distribution whose probability mass function (PMF) involves the Riemann Zeta function ζ, given by $\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$. A connection between ζ and the primes, proved by Euler (see Section 17.2 of [7]), is given by

$$\zeta(s) = \prod_{p} \frac{1}{1 - p^{-s}}. \tag{7.1}$$

If $Z(s)$ is a Zeta(s)-distributed random variable, its PMF is given by

$$\mathbb{P}(Z(s) = j) = \frac{1}{j^s \zeta(s)},\ j \in \mathbb{N}. \tag{7.2}$$

A Zeta(s)-distributed random variable $Z(s)$ has the property that

$$\mathbb{P}(Z(s) = m) = \frac{1}{m^s} \mathbb{P}(Z(s) = 1).$$

Thus, if the values of the range of $Z(s)$ correspond to a ranking, this model can be used when the mth ranked object occurs $1/m^s$ times as often as the 1st ranked item. The Zeta distribution is related to the Zipf distribution [8] which was a model used in linguistics by George Zipf. The Zipf model states that the frequency of any word is inversely proportional to its rank in the frequency table.

7.1.1 Couplings of Random Variables

Our results regard a Zeta(s)-distributed integer-valued random variable $Z(s)$, and in particular, its prime power process $(\alpha_i(s))_{1\leq i}$ when factoring $Z(s)$ uniquely into a product of primes $\prod_{1\leq i} p_i^{\alpha_i(s)}$, where p_i denotes the ith prime. These results are proved by establishing the existence of couplings of random variables (see §4.1 for the definition of a coupling). In particular, we provide couplings, with some constraints, of a Zeta(s)-distributed variable $\prod_{i\geq 1} p_i^{\alpha_i(s)}$, necessarily consisting of a independent prime powers $\alpha_i(s) \sim \text{Geometric}\left(\frac{1}{p_i^s}\right)$, $i \geq 1$, with another random variable $M(n) = \prod_{i\leq\pi(n)} p_i^{Z_i}$, where $\pi(n)$ denotes the number of primes $\leq n$ and the Z_i's, $1 \leq i \leq \pi(n)$, are independent with $Z_i \sim \text{Geometric}\left(\frac{1}{p_i}\right)$.[1] To see that the components $(\alpha_i(s))_{1\leq i}$ are independent with $\alpha_i(s) \sim \text{Geometric}\left(\frac{1}{p_i^s}\right)$, note that for any integer $\prod_{1\leq i} p_i^{a_i}$ we have

$$\begin{aligned}
\mathbb{P}(\alpha_i = a_i \text{ for all } i \geq 1) &= \mathbb{P}\left(Z(s) = \prod_{1\leq i} p_i^{a_i}\right) \\
&\overset{(7.2)}{=} \frac{1}{\left(\prod_{1\leq i} p_i^{a_i}\right)^s \zeta(s)} \\
&\overset{(7.1)}{=} \prod_{1\leq i} \left(\frac{1-\frac{1}{p_i^s}}{p_i^{a_i s}}\right) \\
&= \prod_{1\leq i} \mathbb{P}(\alpha_i(s) = a_i).
\end{aligned}$$

[1] Given $0 < q < 1$, a random variable X has the Geometric(q)-distribution if $\mathbb{P}(X = j) = \frac{1-\frac{1}{p}}{p^j}$ for any natural number j.

7.1.2 Arratia's Conjecture and our Main Result

The constraints we impose on our couplings are motivated by Arratia's Conjecture (stated in §4.1). Using the notation from that conjecture, the conjecture is true if and only if there exists a coupling of $M(n)$ and $N(n)$ such that we always have $\sum_{p \leq n} (C_p(n) - Z_p)^+ \leq 1$. For our purposes, we seek couplings of $Z(s)$ and $M(n)$ such that

$$\sum_{i \leq \pi(n)} (\alpha_i(s) - Z_i)^+ \leq 1, \tag{7.3}$$

where $(\cdot)^+$ denotes the positive part. That is, we seek a joint PMF p such that $p(M(n) = \cdot, Z(s) = \cdot) = 0$ when

$$\sum_{i \leq \pi(n)} (\alpha_i(s) - Z_i)^+ > 1.$$

Consider $Z_n(s) \coloneqq \prod_{i \leq \pi(n)} p_i^{\alpha_i(s)}$ – i.e., the restriction of $Z(s)$ to the first $\pi(n)$ prime factors. Then

$$(7.3) \iff Z_n(s) \text{ always divides } M(n) P(n) \text{ for some random prime } P(n) \leq n. \tag{7.4}$$

Our main result is the following theorem.

Theorem 23. *Given $n \in \mathbb{N}$ and $k \geq 4$, there exists an $\varepsilon(k) > 0$ such that when $s \in (1, 1 + \varepsilon(k))$, we can couple $Z(s)$ and $M(n)$ such that we always have*

$$Z(s) \leq k \implies Z_n(s) \text{ divides } M(n) P(n) \text{ for some random prime } P(n) \leq n. \tag{7.5}$$

By (7.4), we can restate (7.5) as

$$Z(s) \leq k \implies \sum_{i \leq \pi(n)} (\alpha_i(s) - Z_i)^+ \leq 1.$$

Our motivation for coupling $Z(s)$ and $M(n)$ with the constraints described above is due to Arratia's Conjecture and the fact that

$$\alpha_i(s) \to Z_i \text{ in distribution as } s \to 1, \text{ and } C_i(n) \to Z_i \text{ in distribution as } n \to \infty$$

(see p. 3 of [1] for a proof of the fact that $C_i(n)$ converges to Z_i in distribution as $n \to \infty$). In §7.2, we describe how our constraints force a significant proportion of the entries of a prospective joint mass distribution $p(\cdot,\cdot)$ of our variables to be 0. The significance of the parameter k is that we only require $p(M(n), Z(s)) = 0$ when both $Z(s) \leq k$ and $Z_n(s)$ does not divide $M(n)P(n)$ for any prime $P(n) \leq n$. In §7.3, we introduce the notion of pivot mass, which depends on the constraints placed on the desired joint distribution. Some properties of the pivot mass are proved in §7.3 and §7.4. In §7.5, we apply results on the pivot mass and a theorem proved by Strassen (see Theorem 11) to prove Theorem 23.

7.2 Properties of the Couplings corresponding to Theorem 23

If we are to successively construct a joint PMF $p(\cdot,\cdot)$ with marginal distributions corresponding to $M(n)$ and $Z(s)$ such that (7.5) holds, we must ensure that $p(M(n) = \cdot, Z(s) = \cdot) = 0$ when both $Z(s) \leq k$ and $\sum_{i \leq \pi(n)} (\alpha_i(s) - Z_i)^+ > 1$. Since any coupling provided by Theorem 23 consists of discrete marginal distributions with ranges $S \coloneqq \left\{\prod_{i \leq \pi(n)} p_i^{a_i} : a_i \geq 0\right\}$ and $\mathbb{N}$, respectively, any of our desired couplings corresponds to a $\infty \times \infty$ matrix with rows labeled by the elements of S, in increasing order, and with columns labeled $1, 2, 3, \ldots$. Further, the row and column sums of the matrix are known since we have formulas for the distributions of the marginals. I.e., the sum along column $Z(s) = j$ is, by (7.2), $\mathbb{P}(Z(s) = j) = \frac{1}{j^s \zeta(s)}$. Applying independence of the Z_i's, the sum

along row $i = \prod_{i \leq \pi(n)} p_i^{b_i}$ is

$$\begin{aligned}
\mathbb{P}\left(M(n) = \prod_{i \leq \pi(n)} p_i^{b_i}\right) &= \mathbb{P}\left(Z_i = b_i \text{ for all } i \leq \pi(n)\right) \\
&= \prod_{i \leq \pi(n)} \mathbb{P}\left(Z_i = b_i\right) \\
&= \prod_{i \leq \pi(n)} \left(\frac{1 - \frac{1}{p_i}}{p_i^{b_i}}\right) \\
&= \frac{\prod_{i \leq \pi(n)} \left(1 - \frac{1}{p_i}\right)}{i}.
\end{aligned}$$

Figure 7.1: In Theorem 23, if the pair (i_0, j_0) violates (7.5), then our desired joint PMF p should satisfy $p(i_0, j_0) = 0$.

$M(n)$ \ $Z(s)$	1	2	$\cdots$	j	$\cdots$	j_0	$\cdots$	Row sum
1								$\prod_{i \leq n} \left(1 - \frac{1}{p_i}\right)$
2								$\frac{1}{2} \prod_{i \leq n} \left(1 - \frac{1}{p_i}\right)$
$\vdots$								$\vdots$
i_0						0		$\frac{1}{i_0} \prod_{i \leq n} \left(1 - \frac{1}{p_i}\right)$
$\vdots$								$\vdots$
i				$p(i,j)$				$\frac{1}{i} \prod_{i \leq n} \left(1 - \frac{1}{p_i}\right)$
$\vdots$								$\vdots$
Column sum	$\frac{1}{\zeta(s)}$	$\frac{1}{2^s \zeta(s)}$	$\cdots$	$\frac{1}{j^s \zeta(s)}$	$\cdots$	$\frac{1}{j_0^s \zeta(s)}$	$\cdots$	

Example 24. Suppose $n = 4$ and $k = 30$. We display some features of the joint distributions guaranteed by Theorem 23 corresponding to these values of n and k. Each row label is of the form $i = 2^a 3^b$. Column 1 and prime labeled columns have no required 0's. To see this, note that if $Z(s)$

is 1 or prime, then either each α_i is 0 or there exists an i such that $a_i = 1$ and $a_j = 0$ for all $j \neq i$; in either case, $\sum_{i \leq \pi(n)} (a_i - Z_i)^+ \leq 1$ is true for any $Z_1, Z_2 \geq 0$. For example, column 8 satisfies $8 = Z(s) < k$, so we must have a 0 in any row i satisfying $a \leq 1$. Required 0's are also given for some other columns in Figure 7.2. Although 25 is composite, it satisfies $Z_4(s) = 1$, so it never violates (7.5). Column 2^5 satisfies $2^5 = Z(s) > k$, so it satisfies (7.5). In general, any column label $j > k$ never has a required 0.

Consider two columns $Z(s), Z'(s) \leq k$. If $Z_n(s)$ divides $Z'_n(s)$, column $Z(s)$ has a required 0 in row i only if column $Z'(s)$ has a required 0 in row i. The reason for this is that if $Z_n(s)$ does not divide $iP(n)$ for any prime $P(n) \leq n$, then $Z'_n(s)$ also will not divide $iP(n)$ for any prime $P(n) \leq n$. This is why, for example, in each row for which column 8 has a required 0, column 16 also has a required 0.

Figure 7.2: Required zeros in some particular columns for a coupling corresponding to Theorem 23 with $n = 4$ and $k = 30$.

$Z(s)$	2	2^3	$2^2 \cdot 3$	2^4	5^2	3^3	2^5	Row sum
$M(4)$								
1		0	0	0		0		$\frac{1}{3}$
2		0	0	0		0		$\frac{1}{6}$
3		0	0	0		0		$\frac{1}{9}$
2^2				0		0		$\frac{1}{12}$
$2 \cdot 3$		0		0		0		$\frac{1}{18}$
2^3				0		0		$\frac{1}{24}$
3^2		0	0	0				$\frac{1}{27}$
$\vdots$								$\vdots$
Column sum	$\frac{1}{2^s\zeta(s)}$	$\frac{1}{8^s\zeta(s)}$	$\frac{1}{12^s\zeta(s)}$	$\frac{1}{16^s\zeta(s)}$	$\frac{1}{25^s\zeta(s)}$	$\frac{1}{27^s\zeta(s)}$	$\frac{1}{32^s\zeta(s)}$	

Note that we can increase the amount required zeros in the table by increasing the value of k.

7.3 Pivot Mass

Given columns j and l with corresponding prime factorizations $\prod_{1\leq i} p_i^{a_i(s)}$ and $\prod_{1\leq i} p_i^{a_i'(s)}$, we seek a way to compare the corresponding sets of row labels in which column j or l must be 0. Given $j,l\leq k$, based on (7.5) we compare the probability of the random variable $M(n)$ belonging to the sets

$$\left\{\prod_{i\leq\pi(n)} p_i^{b_i}\in\mathbb{N}:\ \text{and}\ \sum_{i\leq n}\left(a_i(s)-b_i\right)^+>1\right\},$$
$$\left\{\prod_{i\leq\pi(n)} p_i^{b_i}\in\mathbb{N}:\ \text{and}\ \sum_{i\leq n}\left(a_i'(s)-b_i\right)^+>1\right\}.$$

I.e., (7.5) is true for all of our desired couplings, so we measure the probability that $M(n)$ takes on a value i for which column j or l has a required 0 in row i. This motivates the following definition.

Definition. Given $n,k\in\mathbb{N}$, we call the pair $(i,Z(s))$, corresponding to the ith row label $\prod_{i\leq\pi(n)} p_i^{Z_i}$ and the column label $Z(s)=\prod_{1\leq i} p_i^{\alpha_i(s)}$, a (n,k)**-pivot** if

$$Z(s)\leq k \text{ and } \sum_{i\leq n}\left(\alpha_i(s)-Z_i\right)^+>1.$$

Denote the set of all (n,k)-pivots by $P(n,k)$. The (n,k)**-pivot mass** in column $Z(s)=j$ is defined as

$$\mathcal{PM}_{(n,k)}(j):=\sum_{i:(i,j)\in P(n,k)}\mathbb{P}\left(M(n)=i\right).$$

Equivalently,

$$\mathcal{PM}_{(n,k)}(j)=\left(\prod_{i\leq\pi(n)}\left(1-\frac{1}{p_i}\right)\right)\sum_{i:(i,j)\in P(n,k)}\frac{1}{i}.$$

Given a subset L of column labels of $\mathbb{N}$, the pivot mass in L is defined as

$$\mathcal{PM}_{(n,k)}(L) \coloneqq \sum_{\substack{i:(i,j)\in P(n,k)\\ \forall j\in L}} \mathbb{P}(M(n)=i).$$

If we define

$$I_j \coloneqq \{\text{row labels } i : (i,j) \in P(n,k)\},$$
$$I_L \coloneqq \{\text{row labels } i : (i,j) \in P(n,k) \text{ for all } j \in L\}$$

and let $\mathbb{P}_M$ denote the distribution function[2] of $M(n)$, then

$$\mathcal{PM}_{(n,k)}(j) = \mathbb{P}_M(I_j)$$

and

$$\mathcal{PM}_{(n,k)}(L) = \mathbb{P}_M(I_L).$$

A consequence of the definition is that, if $Z(s) \leq k$, then $\mathcal{PM}_{(n,k)}(Z(s)) = \mathcal{PM}_{(n,k)}(Z_n(s))$ (this can also be seen by equation (7.6) below).

Example 25. Let us compute the pivot mass of some columns in Example 24. If j is 1 or prime, then $\mathcal{PM}_{(4,30)}(j) = 0$ since we previously showed that column j has no pivots if j is 1 or prime. Column 8 has a pivot in row $i = 2^a 3^b$ if and only if $a \leq 1$ (since 8 divides $iP(4)$ if and only if $a = 2$ and $P(4) = 2$ or if $a \geq 3$). Therefore,

$$\begin{aligned}\mathcal{PM}_{(4,30)}(8) &= \left(\prod_{p\leq 4}\left(1-\frac{1}{p}\right)\right)\sum_{a\leq 1, b\geq 0}\frac{1}{2^a 3^b}\\ &= \frac{1}{2}\frac{2}{3}\frac{3}{2}\frac{3}{2}\\ &= \frac{3}{4}.\end{aligned}$$

[2]The function $\mathbb{P}_{M(n)}$ is defined by $\mathbb{P}_{M(n)}(A) = \mathbb{P}(M(n) \in A)$.

Column 12 has a pivot in row $i = 2^a 3^b$ if and only if $a = 0$ or $a \leq 1$ and $b = 0$. Therefore,

$$\begin{aligned}\mathcal{PM}_{(4,30)}(12) &= \left(\prod_{p\leq 4}\left(1-\frac{1}{p}\right)\right)\left(\sum_{a=0,b\geq 0}\frac{1}{2^a 3^b}+\sum_{a\leq 1,b=0}\frac{1}{2^a 3^b}-\sum_{a=0,b=0}\frac{1}{2^a 3^b}\right)\\ &= \frac{1}{2}\frac{2}{3}\left(\frac{3}{2}+\frac{3}{2}-1\right)\\ &= \frac{2}{3}.\end{aligned}$$

In addition, $\mathcal{PM}_{(4,30)}\left(2^5\right) = 0$ since $2^5 > k$. Theorem 26 gives a formula for $\mathcal{PM}_{(n,k)}(j)$ for any $j \in \mathbb{N}$, where $\Omega(j)$ denotes the sum of the distinct prime factors of j, $\omega(j)$ denotes the number of distinct prime factors of j, and $\mathbf{1}_{\{\cdot\}}$ denotes an indicator random variable. Fortunately, due to the role of the parameter s appearing in the Zeta(s)-distribution, it is not necessary to derive a formula for $\mathcal{PM}(L)$ in order to prove Theorem 23. The fact that $\mathcal{PM}_{(n,k)}(L) \leq \mathcal{PM}_{(n,k)}(j)$ for any $j \in L$ will be sufficient.

Theorem 26. *Consider a fixed column label $Z(s) = \prod_{1\leq i} p_i^{\alpha_i(s)}$. Its (n,k)-pivot mass is*

$$\mathcal{PM}_{(n,k)}(Z(s)) = \mathbf{1}_{\{Z(s)\leq k\}}\left(1-\frac{1+\Omega(Z_n(s))-\omega(Z_n(s))}{Z_n(s)}\right). \tag{7.6}$$

Proof. If $Z(s) > k$, then both sides of (7.6) are 0, so we assume $Z(s) \leq k$. Instead of computing the (n,k)-pivot mass in column $Z(s)$ directly, we compute the complement and subtract from 1. In order for $(i, Z(s))$ to not be an (n,k)-pivot we need i to have at most one less prime factor or prime power than $Z_n(s)$. So for each prime factor p dividing $Z_n(s)$, with multiplicity $a_p \geq 1$, i can be any multiple of $\frac{Z_n(s)}{p}$. The quantity

$$\left(\prod_{p\leq n}\left(1-\frac{1}{p}\right)\right)\sum_{a_2,\ldots,a_p\geq 0}\sum_{p|Z_n(s)}\frac{p}{Z_n(s)\cdot 2^{a_2}\cdots p^{a_p}}$$

counts each of the rows which are divisible by all such values of i, but rows which are divisible by $Z_n(s)$ are counted $\omega(Z_n(s))$ times while they should only be counted once each. Therefore, we

subtract $\frac{\omega(Z_n(s))-1}{Z_n(s)\cdot 2^{a_2}\cdots p^{a_p}}$ from the sum over the distinct prime divisors of $Z_n(s)$ to obtain

$$\begin{aligned}\mathcal{PM}_{(n,k)}(Z(s)) &= 1-\left(\prod_{p\leq n}\left(1-\frac{1}{p}\right)\right)\sum_{a_2,\ldots,a_p\geq 0}\left(\sum_{p|Z_n(s)}\frac{p}{Z_n(s)\,2^{a_2}\cdots p^{a_p}}-\frac{\omega(Z_n(s))-1}{Z_n(s)\,2^{a_2}\cdots p^{a_p}}\right)\\ &= 1-\left(\prod_{p\leq n}\left(1-\frac{1}{p}\right)\right)\sum_{a_2,\ldots,a_p\geq 0}\left(\frac{1+\Omega(Z_n(s))-\omega(Z_n(s))}{Z_n(s)\,2^{a_2}\cdots p^{a_p}}\right)\\ &= 1-\frac{1+\Omega(Z_n(s))-\omega(Z_n(s))}{Z_n(s)}.\end{aligned}$$

□

7.4 An Upper-Bound for $\mathcal{PM}_{(n,k)}(\cdot)$

The following result will be useful in proving Theorem 23.

Theorem 27. *Given $n\in\mathbb{N}$, $k\geq 4$, and any subset $L\subseteq\mathbb{N}$, there exists an $\varepsilon(k)>0$ such that when $s\in(1,1+\varepsilon(k))$ we have*

$$\mathcal{PM}_{(n,k)}(L)\leq 1-\sum_{j\in L}\frac{1}{j^s\zeta(s)}. \tag{7.7}$$

Proof. By definition of $\mathcal{PM}_{(n,k)}$, if there exists a $j\in L$ with $j>k$, then $\mathcal{PM}(L)=0$. Therefore, if suffices to consider sets $L\subseteq\{1,2,\ldots,k\}$. Given any $j\in L$, we have

$$\begin{aligned}\mathcal{PM}_{(n,k)}(L) &\leq \mathcal{PM}_{(n,k)}(j)\\ &\overset{(7.6)}{=} \left(1-\frac{1+\sum_{p\leq n,p|j}(p-1)}{j}\right)\\ &\leq 1-\frac{2}{k}.\end{aligned}$$

Now, we have

$$1-\frac{2}{k}\leq 1-\sum_{j\in L}\frac{1}{j^s\zeta(s)}$$

$$\iff \sum_{j\in L} \frac{1}{j^s \zeta(s)} \leq \frac{2}{k}.$$

Further, since $L \subseteq \{1, 2, \ldots, k\}$, we have $\sum_{j\in L} \frac{1}{j^s} < \sum_{m=1}^{k} \frac{1}{m} =: H_k$ and

$$\sum_{m=1}^{k} \frac{1}{m\zeta(s)} \leq \frac{2}{k}$$
$$\iff \frac{kH_k}{2} \leq \zeta(s)$$

The latest inequality is true for all s sufficiently close to 1, where s depends only on k. □

7.5 Using Pivot Mass and Strassen's Theorem to Provide Couplings

We will apply Strassen's Theorem (Theorem 11) to prove Theorem 23.

Proof of Theorem 23. Let us define

$$S = \left\{ \prod_{i\leq\pi(n)} p_i^{a_i} : a_i \geq 0 \right\},$$
$$T = \mathbb{N},$$

corresponding to the set of row labels and the set of column labels respectively. Let us endow S with the metric d_S on $\mathbb{N}$ defined as

$$d_S(i_1, i_2) \coloneqq |i_1 - i_2|,$$

and endow T with the metric d_T on $\mathbb{N}$ defined a

$$d_T(j_1, j_2) = |j_1 - j_2|.$$

Since S and T are countably infinite, both S and T are separable. In S and T we have

$$i_1 \neq i_2 \implies d_S(i_1, i_2) \geq 1 \tag{7.8}$$

and

$$j_1 \neq j_2 \implies d_T(j_1, j_2) \geq 1.$$

Therefore, every Cauchy sequence in S converges in S and every Cauchy sequence in T converges in T. Thus, S and T are complete. Our goal is to apply Theorem 11 with $\varepsilon = 0$ and

$$\begin{aligned} W &= \{(i,j) \in S \times T : (7.5) \text{ is not violated}\}, \\ \mu_{(n)}(i) &= \mathbb{P}(M(n) = i), i \in S, \\ \nu_s(j) &= \mathbb{P}(Z(s) = j), j \in T, \\ \lambda &= p, \\ & L \text{ is an arbitrary subset of } \mathbb{N}, \end{aligned}$$

where p is our desired joint PMF, with marginals corresponding to $M(n)$ and $Z(s)$, such that if $(i,j) \in P_{(n,k)}$ (by definition, we have $W^c = P_{(n,k)}$), then $p(i,j) = 0$. Let us endow W with the metric obtained by restricting the metric

$$d_{S \times T}((i_1, j_1), (i_2, j_2)) \coloneqq \max(d_S(i_1, i_2), d_T(j_1, j_2))$$

on $S \times T$ to W. To show that W is closed, we first show that S and T are closed. Suppose that $(i_l)_{l \in \mathbb{N}}$ is a sequence in S with $\lim_{l \to \infty} i_l = i$. To show that S is closed, it suffices to show that $i \in S$. For all $\varepsilon' \in (0,1)$ there exists a constant $N \in \mathbb{N}$ such that

$$l > M \implies d_S(i_l, i) < \varepsilon'.$$

Since $\varepsilon' < 1$, (7.8) implies $i = i_{M+1}$, so $i \in S$. Therefore, S is closed. Similarly, T is closed. Now

to show that W is closed in $S \times T$, suppose that $(i_l, j_l)_{l \in \mathbb{N}}$ is a sequence of pairs $(i_l, j_l) \in W$ with

$$\lim_{k \to \infty} ((i_l, j_l)) = (i, j)$$

Since S and T are closed, we have $i \in S$ and $j \in T$. By a similar argument above, there exists an M such that $i = i_{M+1}, j = j_{M+1}$, so that $(i, j) = (i_{M+1}, j_{M+1}) \in W$. Therefore, W is closed. Further, $W \neq \emptyset$ since, e.g., $(1, 1)) \in W$. Note that the set L is a closed subset of the column labels since L is a finite set. Moreover,

$$v_n(L) = \mathbb{P}(Z(s) \in L) = \sum_{j \in L} \frac{1}{j^s \zeta(s)},$$

and

$$\begin{aligned}
\mu_{(n)}\left(p_s\left(P_{(n,k)} \cap (S \times L)\right)\right) &= \mathbb{P}\left(M(n) \in p_S\left(P_{(n,k)} \cap (S \times L)\right)\right) \\
&= \mathbb{P}\left(M(n) \in p_S\left(P^c \cap (S \times L)\right)\right) \\
&= \mathbb{P}\left(\exists j \in L : (M(n), j) \notin P\right) \\
&= 1 - \mathcal{PM}_{(n,k)}(L).
\end{aligned}$$

Therefore, (4.3) of Theorem 11 is equivalent to

$$\sum_{j \in L} \frac{1}{j^s \zeta(s)} \leq 1 - \mathcal{PM}_{(n,k)}(L).$$

By (7.7) the latest inequality can be made true by choosing s sufficiently close to 1. Therefore, by the conclusion of Theorem 11, there exists a joint PMF p, with marginals $\mathbb{P}(M(n) = \cdot)$ and $\mathbb{P}(Z(s) = \cdot)$, such that $p(W) = 1$. I.e, the probability of having no pivot in this joint distribution is 1. Hence, the proof of Theorem 23 is complete. □

Bibliography

[1] Arratia, R. *On the amount of dependence in the prime factorization of a uniform random integer*. In *Contemporary Combinatorics*, **10**, 29–91, Bolyai Society Mathematical Studies., János Bolyai Mathematical Society, Budapest, 2002.

[2] Arratia, R., Barbour, A.D. and Tavaré, S. *Logarithmic Combinatorial Structures: A Probabilistic Approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2003.

[3] Durrett, R. *Probability: Theory and Examples, 5th Edition*. Cambridge University Press, 2019.

[4] Erdös, P. and Kac, M. *The Gaussian law of errors in the theory of additive number theoretic functions*. American Journal of Mathematics, **62** (1/4), 738–742, 1940.

[5] Feller, W. *An Introduction to Probability Theory and its Applications, Vol.1, 3rd Edition*. Wiley, New York, 1968.

[6] Hardy, G. H. and Ramanujan, S. *The normal number of prime factors of a number n*. Quarterly Journal of Mathematics, no. 48, 76–92, 1917.

[7] Hardy, G. and Wright, E. *An introduction to the theory of numbers, 5th Edition*. The Clarendon Press, Oxford University Press, New York, 1979.

[8] Powers, D. *Applications and explanations of Zipf's law*. Association for Computational Linguistics, 151–160, 1998.

[9] Quas, A. Reference request for couplings with conditions. URL https://mathoverflow.net/q/284525, 2017.

[10] Simion, R. *Noncrossing partitions.* Discrete Math. **217** (1-3), 367–409, 2000.

[11] Stanley, R. *Enumerative Combinatorics, Vol.2.* Cambridge Studies in Advanced Mathematics Book 62, Cambridge University Press, 1999.

[12] Strassen, V. *The existence of probability spaces with given marginals.* Ann. Math. Stat., Volume **36**, no. 2, 423-439, 1965.

[13] Tenenbaum, G. and Mendes-France, M. *The Prime Numbers and their Distribution.* American Mathematical Society, Providence, 2000

[14] Thorisson, H. *Coupling, Stationarity, and Regeneration. Probability and Its Applications.* Springer, New York, 2000.

[15] Wilf, H. *Generatingfunctionology, 2nd Edition.* Academic Press, Cambridge, 1994.

[16] Woll, C. Summation over integers satisfying some conditions. URL https://mathematica.stackexchange.com/a/162932/26842, 2017.

[17] Zhao, H. and Zhong, Z. *Two statistics linking Dyck paths and non-crossing partitions.* Electronic Journal of Combinatorics, **18**, no. 1, 2011.