Electron. J. Probab. **27** (2022), article no. 105, 1–25.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP815

The distribution of the number of distinct values in a finite exchangeable sequence

Theodore Zhu*

Abstract

Let K_n denote the number of distinct values among the first n terms of an infinite exchangeable sequence of random variables $(X_1, X_2, \ldots)$. We prove for $n = 3$ that the extreme points of the convex set of all possible laws of K_3 are those derived from i.i.d. sampling from discrete uniform distributions and the limit case with $\mathbb{P}(K_3 = 3) = 1$. We also consider the problem in higher dimensions and variants of the problem for finite exchangeable sequences and exchangeable random partitions.

Keywords: occupancy problem; exchangeable sequences; exchangeable random partitions; Ewens-Pitman two-parameter family.
MSC2020 subject classifications: 60G09; 60C05.
Submitted to EJP on March 17, 2021, final version accepted on June 28, 2022.
Supersedes arXiv:2103.07518.

1 Introduction

For an infinite sequence of real-valued random variables $(X_1, X_2, \ldots)$, let

$$K_n = K_n(X_1, \ldots, X_n) := \#\{X_i : 1 \le i \le n\},$$

the number of distinct values appearing in the first n terms. This article focuses on the case in which the sequence $(X_1, X_2, \ldots)$ is *exchangeable*, meaning that its distribution is invariant under finite permutations of the indices. It is a well-known and celebrated result of de Finetti that any infinite exchangeable sequence is a mixture of i.i.d. sequences. We explore ideas related to the following central question:

> Given a probability distribution $(a_1, \ldots, a_n)$ on $[n] := \{1, \ldots, n\}$, is there an infinite exchangeable sequence of random variables $(X_1, X_2, \ldots)$ such that $\mathbb{P}(K_n = k) = a_k$ for $1 \le k \le n$?

*University of California, Berkeley. E-mail: tdz@math.berkeley.edu

The functional K_n has been studied extensively in the context of the *occupancy problem* as well as other closely related formulations including the birthday problem, the coupon collector's problem, and random partition structures [5, 8, 13]. Much of the literature pertains to the asymptotic behavior of K_n in the classical version in which the X_i are i.i.d. discrete uniform random variables, as well as the general i.i.d. case. See [6] for a recent survey with many references. Asymptotics of K_n have also been studied for a random walk $(X_1, X_2, \ldots)$ with stationary increments [15], [3, Section 7.3].

Let us first consider the problem for small values of n. For $n = 1$, the random variable K_1 is just the constant 1. Next, it is easy to see that any probability distribution on $\{1, 2\}$ can be achieved as the law of K_2 for some exchangeable sequence; indeed, for any $a \in [0, 1]$, i.i.d. sampling from a distribution with a single atom having weight $\sqrt{a}$ yields $\mathbb{P}(K_2 = 1) = a$. However, the problem is not trivial for $n = 3$, as evident by the following bound due to Jim Pitman (personal communication.) The proof is presented in Section 3.

Proposition 1.1. *For K_3 the number of distinct values in the first 3 terms of an infinite exchangeable sequence of random variables $(X_1, X_2, \ldots)$,*

$$\mathbb{P}(K_3 = 2) \leq \frac{3}{4}.$$

Here we present the main result of this article. Let $\boldsymbol{v}_{n,m}$ denote the law of $K_{n,m} := K_n(X_{m,1}, \ldots, X_{m,n})$ where $X_{m,i}$ are i.i.d. with uniform distribution on m elements, i.e.

$$\boldsymbol{v}_{n,m} = \big(\mathbb{P}(K_{n,m} = k) : 1 \leq k \leq n\big)$$

and let $\boldsymbol{v}_{n,\infty} = (0, \ldots, 0, 1)$, corresponding to the limit case $m = \infty$ since

$$\mathbb{P}(K_{n,m} = n) = \frac{m(m-1)\cdots(m-n+1)}{m^n} \longrightarrow 1 \quad \text{as } m \to \infty.$$

Let

$$V_n := \{\boldsymbol{v}_{n,m} : m = 1, 2, \ldots, \infty\}$$

and let H_n denote the convex hull of V_n.

Theorem 1.2. *For $n = 3$,*

(i) The set of extreme points of H_n is V_n.

(ii) The set of possible laws of K_n for an infinite exchangeable sequence $(X_1, X_2, \ldots)$ is H_n.

It is natural to conjecture that the assertions in Theorem 1.2 hold true for larger values of n. Yuri Yakubovich [17] proved that (i) holds for all $n \geq 3$. However, Yakubovich exhibits a counterexample to (ii) for $n = 7$. The results in [17] are further discussed in Section 4. It remains a conjecture that (ii) holds for $n = 4, 5$ and fails for all $n \geq 6$, and more generally it remains an open problem to characterize the the set of possible laws of K_n for $n \geq 4$.

The rest of this article is organized as follows. Section 2 establishes notation and the fundamentals of our approach. Section 3 covers some properties of the law of K_3 leading to a proof of Theorem 1.2, and Section 4 extends some of these results to higher dimensions. Section 5 considers a variant of the main problem for finite exchangeable sequences by appealing to the framework of exchangeable random partitions, and Section 6 explores a remarkable symmetry for K_3 in the Ewens-Pitman two-parameter partition model.

2 Preliminaries

For an i.i.d. sequence $(X_1, X_2, \ldots)$, there is an associated *ranked discrete distribution* $(p_1, p_2, \ldots)$ with $p_1 \geq p_2 \geq \ldots \geq 0$ and $\sum_{i=1}^{\infty} p_i \leq 1$ where the p_i are the weights of the atoms for the law of X_i in decreasing order, and $1 - \sum_{i=1}^{\infty} p_i$ is the weight of the continuous component.

Consider the set

$$\nabla_\infty := \Big\{(p_1, p_2, \ldots) : p_1 \geq p_2 \geq \ldots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1\Big\},$$

sometimes referred to as the infinite dimensional *Kingman simplex* as in [12]. The uniform distribution on m elements corresponds to

$$\boldsymbol{u}_m := \Big(\underbrace{\frac{1}{m}, \ldots, \frac{1}{m}}_{m \text{ times}}, 0, 0, \ldots\Big) \in \nabla_\infty.$$

and any non-atomic law corresponds to $\boldsymbol{u}_\infty := (0, 0, \ldots) \in \nabla_\infty$. With Theorem 1.2 in mind, note that

$$\big\{\boldsymbol{u}_m : m = 1, 2, \ldots, \infty\big\}$$

is precisely the set of extreme points of ∇_∞ [1, Theorem 4.1]. Any $(p_1, p_2, \ldots) \in \nabla_\infty$ has a unique representation as a convex combination of $\boldsymbol{u}_m$, $m = 1, 2, \ldots, \infty$ given by

$$(p_1, p_2, \ldots) = p_* \boldsymbol{u}_\infty + \sum_{i=1}^{\infty} (p_i - p_{i+1}) \boldsymbol{u}_i, \qquad p_* = 1 - \sum_{i=1}^{\infty} p_i.$$

This is a discrete version of Khintchine's representation theorem for unimodal distributions [9].

It is easy to see that the law of K_n for an i.i.d sequence depends only on the ranked frequencies of the atoms. Let

$$q_{n,i}(p_1, p_2, \ldots) := \mathbb{P}\big(K_n = i\big)$$

where $K_n = K_n(X_1, \ldots, X_n)$ for i.i.d. X_i with ranked frequencies $(p_1, p_2, \ldots)$. Then for $n = 3$, it is easy to see that

$$\begin{aligned}
q_{3,1}(p_1, p_2, \ldots) &= \sum_{i=1}^{\infty} p_i^3 \\
q_{3,2}(p_1, p_2, \ldots) &= \sum_{i=1}^{\infty} 3p_i^2(1 - p_i) \\
q_{3,3}(p_1, p_2, \ldots) &= 1 - \sum_{i=1}^{\infty} \big[3p_i^2 - 2p_i^3\big].
\end{aligned}$$

For the general exchangeable case, de Finetti's theorem guarantees that the law of K_n for an exchangeable sequence of random variables $(X_1, X_2, \ldots)$ is a mixture of laws of K_n for i.i.d. sequences. In other words, the set of laws of K_n derived from exchangeable sequences is the convex hull of those derived from i.i.d. sequences. This property allows us to focus on the i.i.d. case and simplify our treatment to ranked discrete distributions.

Note that there is an equivalent reformulation of the problem in the setting of exchangeable random partitions; see e.g. [13] for relevant background on the subject. For an exchangeable random partition $\Pi = (\Pi_n)$ of $\mathbb{N}$, let K_n denote the number of

clusters in the restriction Π_n of Π to $[n]$. Through *Kingman's representation theorem* [10] for exchangeable random partitions of $\mathbb{N}$ in terms of random ranked discrete distributions, the possible laws of K_n in this setting are identical to the possible laws of K_n as defined originally in this paper as the number of distinct values in the first n terms of an exchangeable sequence $(X_1, X_2, \ldots)$. In Sections 5–7, we explore some related problems in the framework of exchangeable random partitions.

Notations and conventions. If a ranked discrete distribution $(p_1, p_2, \ldots)$ has finitely many atoms, i.e. there exists m such that $p_i = 0$ for all $i > m$, we call it a *finite* distribution and abbreviate it as $(p_1, \ldots, p_m)$ when convenient. Since all of the functionals that we work with on ∇_∞ are symmetric functions of the arguments, we understand an equivalence between an unordered discrete distribution $(p_1, p_2, \ldots)$ and its ranked version. Unless otherwise stated, it is implicit in the appearance of $(p_1, p_2, \ldots)$ or $(p_1, \ldots, p_m)$ that the conditions $p_i \geq 0$ and $\sum p_i \leq 1$ hold.

3 Laws of K_3

To simplify notation in this section, let

$$q_i := q_{3,i} = \mathbb{P}(K_3 = i)$$

where q_i may be treated as a functional on ∇_∞.

Lemma 3.1. *For $(p_1, \ldots, p_m)$ with $m \geq 3$ and $p_1 \leq \ldots \leq p_m$,*

$$q_2(p_1 + p_2, p_3 \ldots, p_m) \geq q_2(p_1, p_2, p_3, \ldots, p_m).$$

Proof. Let $a = p_1$ and $b = p_2$. We have

$$q_2(a, b, p_3, \ldots, p_m) = 3a^2(1-a) + 3b^2(1-b) + \sum_{i=3}^{m} 3p_i^2(1-p_i)$$

and

$$q_2(a+b, p_3, \ldots, p_m) = 3(a+b)^2(1-a-b) + \sum_{i=3}^{m} 3p_i^2(1-p_i).$$

Then

$$\begin{aligned} q_2(a+b, p_3, \ldots, p_m) - q_2(a, b, p_3, \ldots, p_m) &= 3(a+b)^2(1-a-b) - 3a^2(1-a) - 3b^2(1-b) \\ &= 6ab(1-a-b) - 3a^2b - 3ab^2 \\ &= 3ab(2 - 3(a+b)) \\ &\geq 0 \end{aligned}$$

since a and b are the two smallest values among $\{a, b, p_3, \ldots, p_m\}$ so $a + b \leq \frac{2}{m} \leq \frac{2}{3}$ for $m \geq 3$. □

This shows that for any $(p_1, \ldots, p_m)$ with $m \geq 3$, merging the two smallest values among $\{p_1, \ldots, p_m\}$ does not decrease q_2.

Proof of Proposition 1.1. By de Finetti's theorem, it suffices to prove the inequality for i.i.d. sequences. Since

$$q_2(p_1, p_2, \ldots) = \sum_{i=1}^{\infty} 3p_i^2(1-p_i) = \lim_{m\to\infty} \sum_{i=1}^{m} 3p_i^2(1-p_i) = \lim_{m\to\infty} q_2(p_1, \ldots, p_m),$$

it is enough to establish the inequality $q_2(p_1,\dots,p_m) \le \frac{3}{4}$ for finite discrete distributions $(p_1,\dots,p_m)$. If $m=2$, then $q_2(p_1,p_2) = 3p_1^2(1-p_1)+3p_2^2(1-p_2)$ which attains its maximum value of $\frac{3}{4}$ subject to $p_1,p_2 \ge 0$ and $p_1+p_2 \le 1$ at $p_1=p_2=\frac{1}{2}$. For $m \ge 3$, by Lemma 3.1 repeatedly merging the two smallest values until no more than two nonzero values remain gives $q_2(p_1,\dots,p_m) \le q_2(\frac{1}{2},\frac{1}{2}) = \frac{3}{4}$. □

Consider the law of K_3 for an i.i.d. sequence $(X_1,X_2,\dots)$ where each X_i has the uniform distribution $\boldsymbol{u}_N := (\frac{1}{N},\dots,\frac{1}{N})$. A probability distribution (q_1,q_2,q_3) of K_3 (on $\{1,2,3\}$) can be represented by any pair of its coordinates; here we shall work with $(q_1,q_3) := \big(\mathbb{P}(K_3=1),\mathbb{P}(K_3=3)\big)$. Then

$$q_1(\boldsymbol{u}_N) := \mathbb{P}(K_3(\boldsymbol{u}_N)=1) = \frac{1}{N^2}$$
$$q_3(\boldsymbol{u}_N) := \mathbb{P}(K_3(\boldsymbol{u}_N)=3) = \frac{(N-1)(N-2)}{N^2}.$$

The set of points $\{\boldsymbol{v}_N : N \in \mathbb{N}\} = \{(1,0),(\frac{1}{4},0),(\frac{1}{9},\frac{2}{9}),(\frac{1}{16},\frac{6}{16}),(\frac{1}{25},\frac{12}{25}),(\frac{1}{36},\frac{20}{36}),\dots\}$ where

$$\boldsymbol{v}_N := (q_1(\boldsymbol{u}_N), q_3(\boldsymbol{u}_N)) = \Big(\frac{1}{N^2}, \frac{(N-1)(N-2)}{N^2}\Big) \tag{3.1}$$

are shown in Figures 1 and 2, with line segments connecting consecutive points.

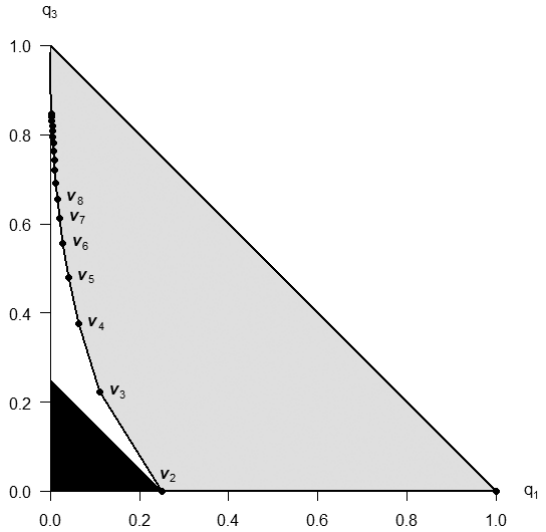

Figure 1: Probability distributions of K_3 represented as points $(q_1,q_3) = \big(\mathbb{P}(K_3 = 1),\mathbb{P}(K_3=3)\big)$ with q_1 horizontal and q_3 vertical. Shaded in black is the restricted region specified by Proposition 1.1. The gray region is the closed convex hull of $\{\boldsymbol{v}_N : N \in \mathbb{N}\}$ where $\boldsymbol{v}_N$ corresponds to the distribution of K_3 for i.i.d. sampling from a discrete uniform distribution on N elements, as defined in (3.1).

The slope of the line connecting $\boldsymbol{v}_N = \big(\frac{1}{N^2},\frac{(N-1)(N-2)}{N^2}\big)$ and $\boldsymbol{v}_{N+1} = \big(\frac{1}{(N+1)^2},\frac{N(N-1)}{(N+1)^2}\big)$ is

$$\frac{\frac{N(N-1)}{(N+1)^2} - \frac{(N-1)(N-2)}{N^2}}{\frac{1}{(N+1)^2} - \frac{1}{N^2}} = -\frac{(N-1)(3N+2)}{2N+1}. \tag{3.2}$$

This is increasing in N which proves Theorem 1.2(i). The equation of the Nth line is given by

$$q_3 - \frac{(N-1)(N-2)}{N^2} = -\frac{(N-1)(3N+2)}{2N+1}\Big(q_1 - \frac{1}{N^2}\Big)$$

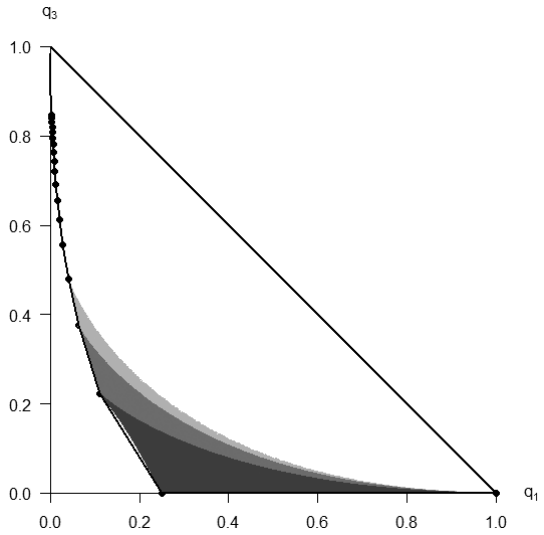

Figure 2: The shaded regions (nested) correspond to the images of $\{(p_1,\dots,p_m) : p_i \geq 0, \sum p_i = 1\}$ under the map $(p_1,\dots,p_m) \mapsto \big(q_1(p_1,\dots,p_m), q_3(p_1,\dots,p_m)\big)$, i.e. distributions of K_3 for i.i.d. sampling from a discrete distribution with at most m atoms for $m = 3$ (dark), $m = 4$ (dark and medium), and $m = 5$ (dark, medium, and light). The existence of the gap between the left boundary of the dark region and the line segment connecting $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ is a consequence of Lemma 3.6. The midpoint of $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$, for example, does not correspond to i.i.d. sampling from any any discrete distribution; however, it does correspond to the the exchangeable sequence with law given by i.i.d. sampling from $\boldsymbol{u}_2$ with probability $\frac{1}{2}$ and i.i.d. sampling from $\boldsymbol{u}_3$ with probability $\frac{1}{2}$.

or after rearranging,

$$q_3 + \frac{(N-1)(3N+2)}{2N+1} q_1 = \frac{2N-2}{2N+1}. \tag{3.3}$$

For $\boldsymbol{p} = (p_1,\dots,p_m)$, define according to the left-hand side of (3.3) the functional

$$L_N(\boldsymbol{p}) := q_3(\boldsymbol{p}) + \frac{(N-1)(3N+2)}{2N+1} q_1(\boldsymbol{p})$$

which may be reexpressed as

$$\begin{aligned}
L_N(\boldsymbol{p}) &= 1 - \big(1 - L_N(\boldsymbol{p})\big) \\
&= 1 - \left(1 - q_3(\boldsymbol{p}) - q_1(\boldsymbol{p}) - \left[\frac{(N-1)(3N+2)}{2N+1} - 1\right] q_1(\boldsymbol{p})\right) \\
&= 1 - q_2(\boldsymbol{p}) + \frac{3(N^2-N-1)}{2N+1} q_1(\boldsymbol{p}) \\
&= 1 - \sum_{i=1}^m 3p_i^2(1-p_i) + \frac{3(N^2-N-1)}{2N+1} \sum_{i=1}^m p_i^3 \\
&= 1 - 3\sum_{i=1}^m p_i^2 + \frac{3N(N+1)}{2N+1} \sum_{i=1}^m p_i^3.
\end{aligned}$$

Note that L_N is nonlinear as a function of discrete distributions $\boldsymbol{p}$. Define

$$f(N) := \frac{3N(N+1)}{2N+1}$$

so

$$L_N(\boldsymbol{p}) = 1 - 3\sum_{i=1}^{m} p_i^2 + f(N)\sum_{i=1}^{m} p_i^3.$$

To better understand the sequence of values $f(N)$, note that f is increasing and

$$N < \frac{2N+2}{2N+1}(N) = \frac{2}{3}\cdot\underbrace{\frac{3N(N+1)}{2N+1}}_{f(N)} = \frac{2N}{2N+1}(N+1) < N+1.$$

The first few values are $f(1) = 2$, $f(2) = \frac{18}{5}$, $f(3) = \frac{36}{7}$, $f(4) = \frac{60}{9}$.

Lemma 3.2. *For $N \geq 1$ and any $\boldsymbol{p} = (p_1, \ldots, p_m)$ with $p_1 \geq \ldots \geq p_m \geq 0$ and $\sum p_i \leq 1$,*

$$L_N(\boldsymbol{p}) \geq \frac{2N-2}{2N+1}.$$

Geometrically, Lemma 3.2 asserts that for any $\boldsymbol{p} = (p_1, \ldots, p_m)$, the point $\big(q_1(\boldsymbol{p}), q_3(\boldsymbol{p})\big)$ lies on or above each of the lines connecting $\boldsymbol{v}_N$ and $\boldsymbol{v}_{N+1}$ for $N \in \mathbb{N}$. It will be shown in the proof that for $N \geq 2$, $L_N(\boldsymbol{p}) = \frac{2N-2}{2N+1}$ if and only if $\boldsymbol{p} = \boldsymbol{u}_N$ or $\boldsymbol{p} = \boldsymbol{u}_{N+1}$; as for $N = 1$, $L_1(\boldsymbol{p}) = q_3(\boldsymbol{p}) = 0$ is attained if and only if $\boldsymbol{p} = (p_1, p_2)$ with $p_1 + p_2 = 1$.

The strategy for proving Lemma 3.2 is to show that L_N is minimized at precisely $\boldsymbol{v}_N$ and $\boldsymbol{v}_{N+1}$ by reducing the domain of minimization in stages, first to $(p_1, \ldots, p_m)$ with $\sum p_i = 1$, then to the uniform distributions, and finally to $\boldsymbol{u}_N$ and $\boldsymbol{u}_{N+1}$. The key to the proof is the following *merging* lemma, which generalizes Lemma 3.1.

Lemma 3.3. *For $N \geq 1$ and $(p_1, \ldots, p_m)$ with $m \geq 2$,*

$$L_N(p_1 + p_2, p_3, \ldots, p_m) - L_N(p_1, p_2, p_3, \ldots, p_m) = 3p_1p_2\big[(p_1+p_2)f(N) - 2\big]$$

which is positive, negative, or zero according to the sign of $p_1 + p_2 - \frac{2}{f(N)}$.

Proof. Let $a = p_1$ and $b = p_2$. We have

$$L_N(a, b, p_3, \ldots, p_m) = 1 - 3a^2 - 3b^2 - 3\sum_{i=3}^{m} p_i^2 + f(N)(a^3 + b^3) + f(N)\sum_{i=3}^{m} p_i^3$$

and

$$L_N(a+b, p_3, \ldots, p_m) = 1 - 3(a+b)^2 - 3\sum_{i=3}^{m} p_i^2 + f(N)(a+b)^3 + f(N)\sum_{i=3}^{m} p_i^3.$$

Then

$$\begin{aligned} L_N(a+b, p_3, \ldots, p_m) - L_N(a, b, p_3, \ldots, p_m) &= -6ab + f(N)(3a^2b + 3ab^2) \\ &= 3ab\big[(a+b)f(N) - 2\big]. \end{aligned}$$

□

The proof of Lemma 3.2 is organized according to the following lemmas.

Lemma 3.4. *Let $\mathcal{P}$ denote the set of all finite ranked discrete distributions, and let $\mathcal{P}^1$ denote the set of finite ranked discrete distributions $(p_1, \ldots, p_m)$ with $\sum p_i = 1$. Then for any $N \geq 1$, we have the equality of sets*

$$\operatorname*{arg\,min}_{\boldsymbol{p}\in\mathcal{P}} L_N(\boldsymbol{p}) = \operatorname*{arg\,min}_{\boldsymbol{p}\in\mathcal{P}^1} L_N(\boldsymbol{p})$$

Proof. Let $\boldsymbol{p}_0 = (p_1, \ldots, p_m) \in \mathcal{P}$ such that $\sum_{i=1}^m p_i < 1$. Let ε satisfy $0 < \varepsilon < \min\{\frac{3}{f(N)}, 1 - \sum_{i=1}^m p_i\}$. Then

$$\begin{aligned} L_N(\varepsilon, p_1, \ldots, p_m) &= -3\varepsilon^2 + f(N)\varepsilon^3 + L_N(p_1, \ldots, p_m) \\ &= \varepsilon^2(f(N)\varepsilon - 3) + L_N(p_1, \ldots, p_m) \\ &< L_N(p_1, \ldots, p_m). \end{aligned}$$

This shows that if $\boldsymbol{p}_0 \notin \mathcal{P}^1$, then $\boldsymbol{p}_0 \notin \arg\min_{\boldsymbol{p}\in\mathcal{P}} L_N(\boldsymbol{p})$. □

Lemma 3.5. *Let $\mathcal{P}^1$ denote the set of finite ranked discrete distributions $(p_1, \ldots, p_m)$ with $\sum p_i = 1$, and let $\mathcal{U} := \{\boldsymbol{u}_m : m \in \mathbb{N}\}$. Then for $N \geq 2$, we have the equality of sets*

$$\arg\min_{\boldsymbol{p}\in\mathcal{P}^1} L_N(\boldsymbol{p}) = \arg\min_{\boldsymbol{p}\in\mathcal{U}} L_N(\boldsymbol{p})$$

Proof. Let $\boldsymbol{p}_0 = (p_1, \ldots, p_m)$, not necessarily ranked, such that $\sum_{i=1}^m p_i = 1$. Suppose $\boldsymbol{p}_0$ has a pair of distinct nonzero values, say $a = p_1$ and $b = p_2$ with $a, b > 0$ and $a \neq b$. Consider the three cases as designated in Lemma 3.3, noting that $\frac{2}{f(N)} < 1$ for $N \geq 2$.

(i) If $a + b < \frac{2}{f(N)}$, then $L_N(a + b, p_3, \ldots, p_m) < L_N(a, b, p_3, \ldots, p_m)$ by Lemma 3.3.

(ii) If $a + b > \frac{2}{f(N)}$, then

$$\begin{aligned} &L_N(\tfrac{a+b}{2}, \tfrac{a+b}{2}, p_3, \ldots, p_m) - L_N(a, b, p_3, \ldots, p_m) \\ &\quad = \big(L_N(a + b, p_3, \ldots, p_m) - L_N(a, b, p_3, \ldots, p_m)\big) \\ &\qquad - \big(L_N(a + b, p_3, \ldots, p_m) - L_N(\tfrac{a+b}{2}, \tfrac{a+b}{2}, p_3, \ldots, p_m)\big) \\ &\quad = 3ab\big((a + b)f(N) - 2\big) - 3(\tfrac{a+b}{2})^2\big((a + b)f(N) - 2\big) \\ &\quad = 3\big(ab - (\tfrac{a+b}{2})^2\big)\big((a + b)f(N) - 2\big) \end{aligned}$$

which is negative since $ab - (\frac{a+b}{2})^2 < 0$ and $(a + b)f(N) - 2 > 0$.

(iii) If $a + b = \frac{2}{f(N)} < 1$, then there must exist a third nonzero value, say $p_3 = c > 0$. If $c = \frac{2}{f(N)}$, then $a \neq c$ and $a + c > \frac{2}{f(N)}$ so $L_N(\frac{a+c}{2}, \frac{a+c}{2}, b, p_4, \ldots, p_m) < L_N(a, b, c, p_4, \ldots, p_m)$ by case (ii). If $c \neq \frac{2}{f(N)}$, then merging a and b, which does not change L_N, followed by averaging $a + b$ and c gives $L_N(\frac{a+b+c}{2}, \frac{a+b+c}{2}, p_4, \ldots, p_m) < L_N(a, b, c, p_4, \ldots, p_m)$ by case (ii) again.

Since permuting values in any discrete distribution does not change L_N, the analysis above holds for all ranked discrete distributions and thus shows that among $\boldsymbol{p} \in \mathcal{P}^1$, L_N cannot be minimized at any $\boldsymbol{p}$ with a pair of distinct nonzero values, i.e. any non-uniform distribution. □

Remark. As mentioned previously, for $N = 1$,

$$\arg\min_{\boldsymbol{p}\in\mathcal{P}} L_1(\boldsymbol{p}) = \{(p_1, p_2) : p_1 \geq p_2 \geq 0, p_1 + p_2 = 1\}$$

which differs from the general case $N \geq 2$. The reason the proof of Lemma 3.5 fails for $N = 1$ is that $f(1) = 2$, so $\frac{2}{f(1)} = 1$ and case (iii) of the proof breaks down.

Lemma 3.6. *Let $\mathcal{U} := \{\boldsymbol{u}_m : m \in \mathbb{N}\}$. Then for $N \geq 1$,*

$$\arg\min_{\boldsymbol{p}\in\mathcal{U}} L_N(\boldsymbol{p}) = \{\boldsymbol{u}_N, \boldsymbol{u}_{N+1}\}$$

Proof. The claim is obvious based on Figure 1, which shows that the slopes between $\boldsymbol{v}_N$ and $\boldsymbol{v}_{N+1}$ for $N \in \mathbb{N}$ are decreasing in N. Indeed, the slope of the Nth line segment is computed in (3.2) as

$$-\frac{(N-1)(3N+2)}{2N+1} = -\frac{3N^2-N-2}{2N+1} = 2 - \frac{3N(N+1)}{2N+1} = 2 - f(N)$$

which is decreasing in N. □

Proof of Lemma 3.2. The claim holds trivially for $N = 1$. For $N \geq 2$, applying Lemmas 3.4, 3.5, and 3.6 yields

$$\arg\min_{\boldsymbol{p}\in\mathcal{P}} L_N(\boldsymbol{p}) = \arg\min_{\boldsymbol{p}\in\mathcal{P}^1} L_N(\boldsymbol{p}) = \arg\min_{\boldsymbol{p}\in\mathcal{U}} L_N(\boldsymbol{p}) = \{\boldsymbol{u}_N, \boldsymbol{u}_{N+1}\}$$

and therefore for any $\boldsymbol{p} = (p_1, \ldots, p_m)$ with $p_i \geq 0$ and $\sum p_i \leq 1$,

$$L_N(\boldsymbol{p}) \geq L_N(\boldsymbol{u}_N) = L_N(\boldsymbol{u}_{N+1}) = \frac{2N-2}{2N+1}.$$

Proof of Theorem 1.2. Part (i) was proven earlier by the slope computation (3.2) and illustrated in Figure 1. For part (ii), Lemma 3.2 asserts that $(q_1(\boldsymbol{p}), q_2(\boldsymbol{p}), q_3(\boldsymbol{p})) \in \mathrm{conv}(V_3)$ for any finite ranked discrete distribution $\boldsymbol{p}$. Extension to infinite discrete distributions $(p_1, p_2, \ldots)$ follows because $\lim_{m\to\infty} q_i(p_1, \ldots, p_m) = q_i(p_1, p_2, \ldots)$, and then extension to exchangeable sequences holds by convexity. □

4 Higher dimensions

This section aims to extend some of the results in the previous section to K_n for larger n. Here $q_{n,i} := \mathbb{P}(K_n = i)$. We begin by generalizing Lemma 3.1 and Proposition 1.1.

Lemma 4.1. *For $n \geq 3$ and $(p_1, \ldots, p_m)$ with $m \geq 3$, $\sum_{i=1}^m p_i = 1$, $p_1 \leq \ldots \leq p_m$,*

$$q_{n,2}(p_1 + p_2, p_3, \ldots, p_m) \geq q_{n,2}(p_1, p_2, p_3, \ldots, p_m).$$

The proof requires the following inequality:

Lemma 4.2. *For $a, b > 0$ and $n \geq 2$,*

$$4\tbinom{n-1}{n} ab(a+b)^{n-2} \leq (a+b)^n - a^n - b^n \leq nab(a+b)^{n-2}$$

Proof. We have

$$(a+b)^n - a^n - b^n = \sum_{k=1}^{n-1} \binom{n}{k} a^k b^{n-k} = ab \sum_{k=0}^{n-2} \binom{n}{k+1} a^k b^{n-2-k}. \tag{4.1}$$

Observe that

$$\binom{n}{k+1} = \frac{n(n-1)(n-2)!}{(k+1)k!(n-k-1)(n-k-2)!} = \frac{n(n-1)}{(k+1)(n-k-1)} \binom{n-2}{k};$$

the denominator $(k+1)(n-k-1)$ is no greater than $(n/2)^2$, and is minimized at $k = 0$ and $k = n-2$, so

$$\binom{n}{k+1} \geq \frac{n(n-1)}{(n/2)^2} \binom{n-2}{k} = 4\frac{n-1}{n} \binom{n-2}{k} \tag{4.2}$$

and

$$\binom{n}{k+1} \leq n \binom{n-2}{k}. \tag{4.3}$$

The result follows by substituting inequalities (4.2) and (4.3) into (4.1) and appealing to the binomial theorem. □

Proof of Lemma 4.1. Let $a = p_1$ and $b = p_2$. We can compute

$$q_{n,2}(a, b, p_3, \dots, p_m) = \mathbb{P}\big(K_n(a, b, p_3, \dots, p_m) = 2\big)$$

by conditioning on the appearance of the first two values:

$$\begin{aligned} q_{n,2}(a, b, p_3, \dots, p_m) = &\sum_{k=1}^{n-1} \binom{n}{k} a^k b^{n-k} + \sum_{k=1}^{n-1} \binom{n}{k} a^k \sum_{i=3}^{m} p_i^{n-k} \\ &+ \sum_{k=1}^{n-1} \binom{n}{k} b^k \sum_{i=3}^{m} p_i^{n-k} + \sum_{3 \le i < j \le m} \sum_{k=1}^{n-1} \binom{n}{k} p_i^k p_j^{n-k}. \end{aligned}$$

Note that the first term, which is an expression for the probability that the first two values both appear and are the only ones to appear in the first n observations, is also equal to $(a+b)^n - a^n - b^n$. Similarly,

$$q_{n,2}(a+b, p_3, \dots, p_m) = \sum_{k=1}^{n-1} \binom{n}{k} (a+b)^k \sum_{i=3}^{m} p_i^{n-k} + \sum_{3 \le i < j \le m} \sum_{k=1}^{n-1} \binom{n}{k} p_i^k p_j^{n-k}.$$

For $m \ge 3$, the difference after appropriate cancellations and applying Lemma 4.2 is

$$\begin{aligned} &q_{n,2}(a+b, p_3, \dots, p_m) - q_{n,2}(a, b, p_3, \dots, p_m) \\ &\quad = \sum_{k=1}^{n-1} \binom{n}{k} \big[(a+b)^k - a^k - b^k\big] \sum_{i=3}^{m} p_i^{n-k} - \sum_{k=1}^{n-1} \binom{n}{k} a^k b^{n-k} \\ &\quad = \underbrace{\sum_{k=1}^{n-2} \binom{n}{k} \big[(a+b)^k - a^k - b^k\big] \sum_{i=3}^{m} p_i^{n-k}}_{\ge 0} + n \underbrace{\big[(a+b)^{n-1} - a^{n-1} - b^{n-1}\big]}_{\ge 4\binom{n-2}{n-1} ab(a+b)^{n-3} \ge 2ab(a+b)^{n-3}} \sum_{i=3}^{m} p_i \\ &\qquad\qquad - \underbrace{\big[(a+b)^n - a^n - b^n\big]}_{\le nab(a+b)^{n-2}} \\ &\quad \ge nab(a+b)^{n-3} \Big[2 \sum_{i=3}^{m} p_i - (a+b)\Big]. \end{aligned}$$

Since $\sum_{i=1}^m p_i = 1$ and $a \le b \le p_3 \le \ldots \le p_m$, it follows that $\sum_{i=3}^m p_i \ge \frac{m-2}{m}$ and $a+b \le \frac{2}{m}$, so

$$2 \sum_{i=3}^{m} p_i - (a+b) \ge 2\Big(\frac{m-2}{m}\Big) - \frac{2}{m} = \frac{2(m-3)}{m} \ge 0$$

and therefore merging the two smallest values among $\{p_1, \dots, p_m\}$ does not decrease $q_{n,2}$ provided that there are at least 3 nonzero values. □

Lemma 4.3. *For any $(p_1, \dots, p_m)$ and $n \ge 3$,*

$$q_{n,2}(p_1, \dots, p_m, p_*) \ge q_{n,2}(p_1, \dots, p_m)$$

where $p_* := 1 - \sum_{i=1}^m p_i$.

Proof. We have

$$q_{n,2}(p_1, \dots, p_m) = \sum_{1 \le i < j \le m} \sum_{k=1}^{n-1} \binom{n}{k} p_i^k p_j^{n-k} + \sum_{i=1}^{m} n p_i^{n-1} p_*$$

and

$$q_{n,2}(p_1,\dots,p_m,p_*) = \sum_{1\le i<j\le m}\sum_{k=1}^{n-1}\binom{n}{k}p_i^k p_j^{n-k} + \sum_{i=1}^{m}\sum_{k=1}^{n-1}\binom{n}{k}p_i^k p_*^{n-k},$$

so

$$q_{n,2}(p_1,\dots,p_m,p_*) - q_{n,2}(p_1,\dots,p_m) = \sum_{i=1}^{m}\sum_{k=1}^{n-2} p_i^k p_*^{n-k} \ge 0.$$

Theorem 4.4. *For any exchangeable sequence of random variables $(X_1, X_2, \dots)$ and any $n \ge 3$,*

$$\mathbb{P}(K_n = 2) \le 1 - 2^{-(n-1)}.$$

Proof. As in the proof of Proposition 1.1, it suffices to show that $q_{n,2}(p_1,\dots,p_m) \le 1-2^{-(n-1)}$ for any $(p_1,\dots,p_m)$. If $m=2$ and $p_1+p_2=1$, then

$$q_{n,2}(p_1,p_2) = 1 - p_1^n - p_2^n$$

which attains its maximum of $1-2^{-(n-1)}$ at $p_1=p_2=\frac{1}{2}$. For $m\ge 3$, by Lemmas 4.1 and 4.3 we have

$$q_{n,2}(p_1,\dots,p_m) \le q_{n,2}(p_1,\dots,p_m,p_*) \le q_{n,2}\left(\tfrac{1}{2},\tfrac{1}{2}\right) = 1-2^{-(n-1)}.$$

The difficulty in extending the proof of Theorem 1.2(ii) to the problem in higher dimensions is that there is no simple generalization of Lemma 3.3. Lemma 3.3 is essential because it asserts that whether merging two values in a discrete distribution increases, decreases, or preserves the functionals L_N is determined by only the sum of the two value to be merged. The corresponding functionals for the higher dimensional problem are more complicated and do not have the same convenient property.

Recently, Yakubovich [17] resolved the previously standing conjecture regarding the assertions in Theorem 1.2 for $n \ge 3$.

Recall some notation from Section 1: for $n\ge 3$ and $m = 1,2,\dots,\infty$, denote by $\boldsymbol{v}_{n,m}$ the law of $K_n(X_{m,1},\dots,X_{m,n})$ where $X_{m,i}$ are i.i.d. with uniform distribution on m elements, i.e.

$$\boldsymbol{v}_{n,m} = \big(\mathbb{P}(K_{n,m}=k) : 1\le k\le n\big)$$

and $\boldsymbol{v}_{n,\infty} = (0,\dots,0,1)$. By a standard combinatorial argument, we have the formula

$$\boldsymbol{v}_{n,m}(k) = \frac{S(n,k)\binom{m}{k}k!}{m^n} \qquad (1\le k\le n)$$

where the $S(n,k)$ are Stirling numbers of the second kind. Let

$$V_n := \{\boldsymbol{v}_{n,m} : m = 1,2,\dots,\infty\}$$

and let H_n denote the convex hull of V_n. Yakubovich proved the following:

Proposition 4.5. *For $n\ge 3$, the set of extreme points of H_n is V_n.*

This is a consequence of the following two lemmas, in which orthogonal vectors to bounding hyperplanes are found, revealing the geometry of H_n.

Lemma 4.6. *Let $n\ge 3$ be odd. Let $\boldsymbol{\gamma}_{n,1} := \delta_n = (0,\dots,0,1)\in\mathbb{R}^n$, and for $r\ge 2$ define $\boldsymbol{\gamma}_{n,r}\in\mathbb{R}^n$ by*

$$\boldsymbol{\gamma}_{n,r}(k) := \frac{(-1)^{k-1}\binom{n-1}{k-1}}{S(n,k)\binom{n+r-3}{k-1}(k-1)!} \qquad \text{for } k=1,\dots,n.$$

Then for $r\ge 1$ we have $\langle\boldsymbol{\gamma}_{n,r},\boldsymbol{v}_{n,m}\rangle = 0$ for $m = r, r+1,\dots,r+n-2$ and $\langle\boldsymbol{\gamma}_{n,r},\boldsymbol{v}_{n,m}\rangle > 0$ for $m<r$ or $m>r+n-2$.

Proof. The assertion for $r = 1$ is obvious because the probability of observing n distinct values in a n-sample from a uniform distribution on m elements is 0 for $m = 1, \dots, n-1$ and positive for $m \geq n$. For $r \geq 2$, observe that

$$\langle \boldsymbol{\gamma}_{n,r}, \boldsymbol{v}_{n,m} \rangle = \sum_{k=1}^{n} \frac{(-1)^{k-1}\binom{n-1}{k-1}}{S(n,k)\binom{n+r-3}{k-1}(k-1)!} \frac{S(n,k)\binom{m}{k}k!}{m^n} = \sum_{k=1}^{n} \frac{(-1)^{k-1}\binom{m-1}{k-1}\binom{n-k+r-2}{r-2}}{\binom{n+r-3}{r-2}m^{n-1}} \tag{4.4}$$

because $k\binom{m}{k} = m\binom{m-1}{k-1}$ and $\binom{n-1}{k-1}\binom{n+r-3}{k-1}^{-1} = \binom{n-k+r-2}{r-2}\binom{n+r-3}{r-2}^{-1}$. Note that the denominator in the summand does not depend on k. It can be shown using generating functions that the numerator evaluates to 0 for $r \leq m \leq r+n-2$ and is otherwise positive for all odd n (proof omitted). □

Lemma 4.7. *Let $n \geq 4$ be even. Let $\boldsymbol{\gamma}''_{n,2} := \delta_n = (0, \dots, 0, 1) \in \mathbb{R}^n$. For $r \geq 2$ define $\boldsymbol{\gamma}'_{n,r} \in \mathbb{R}^n$ by*

$$\boldsymbol{\gamma}'_{n,r}(k) := \frac{(-1)^{k-1}\binom{n-2}{k-1}}{S(n,k)\binom{n+r-4}{k-1}(k-1)!} \qquad \text{for } k = 1, \dots, n-1,$$
$$\boldsymbol{\gamma}'_{n,r}(n) := 0$$

and for $r \geq 3$ define $\boldsymbol{\gamma}''_{n,r} \in \mathbb{R}^n$ by

$$\boldsymbol{\gamma}''_{n,r}(1) := 0,$$
$$\boldsymbol{\gamma}''_{n,r}(k) := \frac{(-1)^{k}\binom{n-2}{k-2}}{S(n,k)\binom{n+r-5}{k-2}(k-2)!} \qquad \text{for } k = 2, \dots, n.$$

Then for $r \geq 2$ we have $\langle \boldsymbol{\gamma}'_{n,r}, \boldsymbol{v}_{n,m} \rangle = 0$ for $m = \infty, r, r+1, \dots, r+n-3$ and $\langle \boldsymbol{\gamma}'_{n,r}, \boldsymbol{v}_{n,m} \rangle > 0$ for $m < r$ or $r+n-3 < m < \infty$, and we have $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle = 0$ for $m = 1, r, r+1, \dots, r+n-3$ and $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle > 0$ for $1 < m < r$ or $m > r+n-3$.

Proof. First, for $r \geq 2$ we have $\langle \boldsymbol{\gamma}'_{n,r}, \boldsymbol{v}_{n,\infty} \rangle = 0$ since $\boldsymbol{v}_{n,\infty} = \delta_n$, and for $1 \leq m < \infty$ we have

$$\langle \boldsymbol{\gamma}'_{n,r}, \boldsymbol{v}_{n,m} \rangle = \sum_{k=1}^{n-1} \frac{(-1)^{k-1}\binom{n-2}{k-1}}{S(n,k)\binom{n+r-4}{k-1}(k-1)!} \frac{S(n,k)\binom{m}{k}k!}{m^n} = \sum_{k=1}^{n-1} \frac{(-1)^{k-1}\binom{m-1}{k-1}\binom{n-k+r-3}{r-2}}{\binom{n+r-4}{r-2}m^{n-1}} \tag{4.5}$$

which is up to a factor of m the same as (4.4) with n replaced by $n-1$. Therefore it evaluates to 0 for $r \leq m \leq r+n-3$ and is positive for other integer values of m for all even n.

Next, the assertion about $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle$ for $r = 2$ holds by the same reasoning as in the case for n odd and $r = 1$. For $r \geq 3$ we have $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,\infty} \rangle > 0$ and for $1 \leq m < \infty$ we have

$$\begin{aligned}
\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle &= \sum_{k=2}^{n} \frac{(-1)^{k}\binom{n-2}{k-2}}{S(n,k)\binom{n+r-5}{k-2}(k-2)!} \frac{S(n,k)\binom{m}{k}k!}{m^n} \\
&= \sum_{k=2}^{n} \frac{(-1)^{k}(m-1)\binom{n-k+r-3}{r-3}\binom{m-2}{k-2}}{\binom{n+r-5}{r-3}m^{n-1}} \\
&= \sum_{k=1}^{n-1} \frac{(-1)^{k-1}(m-1)\binom{n-k+r-4}{r-3}\binom{m-2}{k-1}}{\binom{n+r-5}{r-3}m^{n-1}}
\end{aligned}$$

because $k(k-1)\binom{m}{k} = m(m-1)\binom{m-2}{k-2}$ and $\binom{n-2}{k-2}\binom{n+r-5}{k-2}^{-1} = \binom{n-k+r-3}{r-3}\binom{n+r-5}{r-3}^{-1}$. We see that $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle = 0$ for $m = 1$. For $m > 1$ by shifting the variables accordingly,

specifically $r-1 \mapsto r$ and $m-1 \mapsto m$, we obtain (4.5) up to a positive factor and thus $\langle \boldsymbol{\gamma}''_{n,r}, \boldsymbol{v}_{n,m} \rangle = 0$ for $r \le m \le r+n-3$, and is positive for other integer values of $m > 1$ and $m = \infty$ by definition of $\boldsymbol{\gamma}''_{n,r}$. □

The set $H_n \in \mathbb{R}^n$ is a $(n-1)$-dimensional *apeirotope*, or a generalized polytope which has infinitely many facets, lying in the $(n-1)$-dimensional affine subspace $\{(x_1, \ldots, x_n) : x_1 + \ldots + x_n = 1\}$ intersected with the positive orthant in $\mathbb{R}^n$. Lemmas 4.6 and 4.7 show that the geometry of H_n depends on the parity of n. Specifically,

- for odd $n \ge 3$, the facets of H_n are $(n-2)$-dimensional polytopes given by the vertices $\boldsymbol{v}_{n,1}, \boldsymbol{v}_{n,2}, \ldots, \boldsymbol{v}_{n,n-2}, \boldsymbol{v}_{n,\infty}$ and the vertices $\boldsymbol{v}_{n,r}, \boldsymbol{v}_{n,r+1}, \ldots, \boldsymbol{v}_{n,r+n-2}$ for $r = 1, 2, \ldots$;
- for even $n \ge 4$, the facets of H_n are $(n-2)$-dimensional polytopes given by the vertices $\boldsymbol{v}_{n,1}, \boldsymbol{v}_{n,2}, \ldots, \boldsymbol{v}_{n,n-2}, \boldsymbol{v}_{n,\infty}$, the vertices $\boldsymbol{v}_{n,1}, \boldsymbol{v}_{n,r}, \ldots, \boldsymbol{v}_{n,r+n-3}$ for $r = 2, 3, \ldots$, and the vertices $\boldsymbol{v}_{n,\infty}, \boldsymbol{v}_{n,r}, \ldots, \boldsymbol{v}_{n,r+n-3}$ for $r = 2, 3, \ldots$.

For some intuition regarding the structural difference between the two cases, see Figure 1 ($n = 3$) and Figure 3 ($n = 4$).

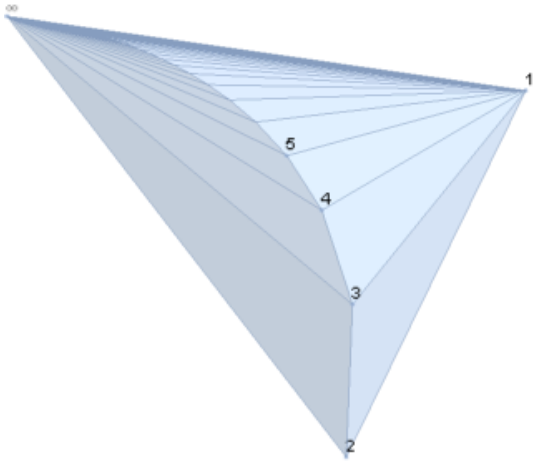

Figure 3: The set $H_4 \in \mathbb{R}^4$ projected onto the first, second, and fourth coordinates. Each point $v_{4,m}$ is labeled by m, for $m = 1, 2, 3, 4, 5, \infty$.

Yakubovich [17] also found the following counterexample to assertion (ii) in Theorem 1.2 for $n = 7$. Consider the distribution of K_7 induced by i.i.d. sampling from the discrete distribution $\boldsymbol{p}^{(t)} := (\frac{1}{4+t}, \frac{1}{4+t}, \frac{1}{4+t}, \frac{1}{4+t}, \frac{t}{4+t}) = \frac{4-4t}{4+t}\boldsymbol{u}_4 + \frac{5t}{4+t}\boldsymbol{u}_5$ for some $t > 0$. The corresponding distribution of K_7 can be computed according to

$$\boldsymbol{v}_7^{(t)}(k) := \mathbb{P}(K_7(\boldsymbol{p}^{(t)}) = k) = \sum \binom{7}{n_1, \ldots, n_5} \prod_{i=1}^{5} (\boldsymbol{p}^{(t)}(i))^{n_i}$$

where the sum is taken over quintuples $(n_1, \ldots, n_5)$ with $\sum n_i = 7$ and $\#\{i : n_i > 0\} = k$. In particular, it can be verified numerically that for $t \in (0, 0.13)$, $\langle \boldsymbol{\gamma}_{7,2}, \boldsymbol{v}_7^{(t)} \rangle < 0$ and hence by Lemma 4.6, $\boldsymbol{v}_7^{(t)} \notin V_7$. This result is proved in [17] using a calculus argument, by showing that function $t \mapsto \langle \boldsymbol{\gamma}_{7,2}, \boldsymbol{v}_7^{(t)} \rangle$, which takes the value 0 at $t = 0$ by Lemma 4.6, has a negative one-sided derivative at $t = 0$, implying that $\langle \boldsymbol{\gamma}_{7,2}, \boldsymbol{v}_7^{(t)} \rangle < 0$ for small positive values of t.

A similar counterexample works for $n = 6$, with $\boldsymbol{p}^{(t)} := (\frac{1}{5+t}, \frac{1}{5+t}, \frac{1}{5+t}, \frac{1}{5+t}, \frac{1}{5+t}, \frac{t}{5+t})$ and the hyperplane inequality with $\boldsymbol{\gamma}'_{6,3}$ as in Lemma 4.7. It appears empirically that similar modifications can be made to produce counterexamples for larger values of n.

5 Finite exchangeable sequences

In this section, we consider the distribution of K_n for a *finite exchangeable sequence* $(X_1, \dots, X_m)$ with $m \geq n$. Note the deviation from the original problem: the first m terms of an infinite exchangeable sequence always form a finite exchangeable sequence, but a finite exchangeable sequence need not have an embedding into an infinite one, nor one with more terms. See [2] for some nice geometric pictures of this property; [7] for an extension of de Finetti's theorem to finite exchangeable sequences in which such a sequence can be identified as a "mixture" of i.i.d. random variables, but allowing for a signed mixing measure; and [11] for conditions for the existence of an embedding of a finite exchangeable sequence in a longer one. The presence of negative signs in the mixture confirms that the laws of K_n in this setting are not simply derived from the i.i.d. case by convexity.

The set of possible laws of K_n for finite exchangeable sequences $(X_1, \dots, X_m)$ form decreasing nested subsets for $m \geq n$, all of which contain that for infinite exchangeable sequences. To analyze this problem, we shift to the framework of *exchangeable random partitions*, for which we provide some background below.

A *partition* of $[m] := \{1, \dots, m\}$ is an unordered collection of disjoint non-empty subsets $\{A_i\}$ of $[m]$ with $\bigcup_i A_i = [m]$. The A_i are called the *clusters* of the partition. The *restriction* of a partition $\{A_i\}$ of $[m]$ to $[n]$ where $n < m$ is the partition of $[n]$ whose clusters are the nonempty members of $\{A_i \cap [n]\}$.

Any infinite sequence of random variables $(X_1, X_2, \dots)$ induces a random partition of $\mathbb{N}$ according to the relation $i \sim j$ if and only if $X_i = X_j$. More precisely, a random partition Π of $\mathbb{N}$ is a sequence (Π_m) where for each m, Π_m is a random partition of $[m]$, and for $n < m$, the restriction of Π_m to $[n]$ is Π_n. For the random partition Π of $\mathbb{N}$ induced by a sequence $(X_1, X_2, \dots)$, the clusters of Π_m are the indices associated to each distinct value among $\{X_1, \dots, X_m\}$. For example, if

$$(X_1(\omega), X_2(\omega), \dots) = (7, 6, 7, 8, 8, 7 \dots),$$

then

$$\Pi_1(\omega) = \{\{1\}\}, \qquad \Pi_2(\omega) = \{\{1\}, \{2\}\}, \qquad \Pi_3(\omega) = \{\{1, 3\}, \{2\}\},$$

$$\Pi_4(\omega) = \{\{1, 3\}, \{2\}, \{4\}\}, \quad \Pi_5(\omega) = \{\{1, 3\}, \{2\}, \{4, 5\}\}, \quad \Pi_6(\omega) = \{\{1, 3, 6\}, \{2\}, \{4, 5\}\}.$$

Observe that K_n as previously defined for a sequence $(X_1, X_2, \dots)$ counts the number of clusters of Π_n for the associated partition Π. When $(X_1, X_2, \dots)$ is exchangeable, it induces an *exchangeable random partition* Π of $\mathbb{N}$, meaning that for each m, the distribution of Π_m is invariant under any deterministic permutation of $[m]$. In this scenario, associated to Π is a function p defined for all finite sequences of positive integers such that for any m and any partition $\{A_1, \dots, A_k\}$ of $[m]$,

$$\mathbb{P}(\Pi_m = \{A_1, \dots, A_k\}) = p(|A_1|, \dots, |A_k|).$$

Here p is called the *exchangeable partition probability function (EPPF)* associated to Π. A consequence of exchangeability is that the EPPF is a symmetric function of its arguments. The probability mass function for K_n can therefore be expressed in terms of the EPPF as

$$\mathbb{P}(K_n = k) = \sum_{\substack{n_1 + \dots + n_k = n \\ n_1 \geq \dots \geq n_k \geq 1}} C(n_1, \dots, n_k) p(n_1, \dots, n_k) \tag{5.1}$$

where

$$C(n_1, \dots, n_k) := \frac{n!}{\prod_{j=1}^{n} (j!)^{s_j} s_j!}, \qquad s_j = s_j(n_1, \dots, n_k) := \#\{i : n_i = j\} \tag{5.2}$$

counts the number of partitions of $[n]$ whose cluster sizes in descending order are given by $n_1, \ldots, n_k$. Furthermore, the EPPF p satisfies the following *consistency* relation:

$$p(n_1, \ldots, n_k) = p(n_1, \ldots, n_k, 1) + \sum_{i=1}^{k} p(n_1, \ldots, \ n_i + 1 \ , \ldots, n_k). \tag{5.3}$$

Reposed in this alternate framework, the goal of this section is to understand the possible distributions of $K_n = K_n(\Pi_m)$ for an exchangeable random partition Π_m of $[m]$ for $m \geq n$, meaning the number of clusters of the restriction $\Pi_{m\downarrow n}$ of Π_m to $[n]$. A consequence of the exchangeability of Π_m is that $\Pi_{m\downarrow n}$ is an exchangeable random partition of $[n]$, whose EPPF is determined recursively by the EPPF for Π_m and the consistency relations (5.3). Note that for $m = n$, $K_n(\Pi_n)$ can have any general probability distribution on $[n]$: for example, given such a probability distribution $(a_1, \ldots, a_n)$, define an EPPF according to

$$p(n-k+1, \underbrace{1, \ldots, 1}_{k-1 \text{ singletons}}) = \frac{a_k}{\binom{n}{k-1}}, \qquad k = 1, \ldots, n$$

where the rest of the values are either 0 or specified by symmetry. By construction, p corresponds to an exchangeable random partition of $[n]$ such that $\mathbb{P}(K_n = k) = a_n$ for $1 \leq k \leq n$. However, for $m > n$, the consistency relations (5.3) must be satisfied, so it is not immediately clear given n and $m > n$ what restrictions there are on the distribution of K_n, if any. The next proposition shows that there are indeed nontrivial restrictions on the law of K_n in this setting.

Proposition 5.1. *Let $n \geq 3$, and let Π_{n+1} be an exchangeable random partition of $[n+1]$. Then we have the sharp bound*

$$\mathbb{P}(K_n(\Pi_{n+1}) = n-1) \leq \frac{\max\{4, n-1\}}{n+1}$$

Proof. We have

$$\mathbb{P}(K_n = n-1) = \binom{n}{2} p(2, 1^{n-2}) = \binom{n}{2} \left[p(3, 1^{n-2}) + (n-2) p(2, 2, 1^{n-3}) + p(2, 1^{n-1}) \right] \tag{5.4}$$

where 1^m is shorthand for m clusters of size 1. We consider the appearance of each of the three terms $p(3, 1^{n-2})$, $p(2, 2, 1^{n-3})$, and $p(2, 1^{n-1})$ in the expansion (5.3) of $p(n_1, \ldots, n_k)$ for $(n_1, \ldots, n_k)$ with $\sum_{i=1}^{k} n_i = n$ and $n_1 \geq \ldots \geq n_k \geq 1$.

- $p(3, 1^{n-2})$ appears in the expansion of only $p(2, 1^{n-2})$ with coefficient 1 and $p(3, 1^{n-3})$ with coefficient 1. $p(3, 1^{n-3})$ appears in the expansion of $\mathbb{P}(K_n = n-2)$ according to (5.1) with coefficient $C(3, 1^{n-3}) = \binom{n}{3}$.
- $p(2, 2, 1^{n-3})$ appears in the expansion of only $p(2, 1^{n-2})$ with coefficient $n-2$ and $p(2, 2, 1^{n-4})$ with coefficient 1. $p(2, 2, 1^{n-4})$ appears in the expansion of $\mathbb{P}(K_n = n-2)$ according to (5.1) with coefficient $C(2, 2, 1^{n-4}) = 3\binom{n}{4}$.
- $p(2, 1^{n-1})$ appears in the expansion of only $p(2, 1^{n-2})$ with coefficient 1 and $p(1^n)$ with coefficient n. $p(1^n)$ appears in the expansion of $\mathbb{P}(K_n = n)$ with coefficient $C(1^n) = 1$.

Hence the problem reduces to maximizing (5.4) subject to the linear constraints

$$\left[\binom{n}{2} + \binom{n}{3} \right] p(3, 1^{n-2}) + \left[\binom{n}{2} (n-2) + 3\binom{n}{4} \right] p(2, 2, 1^{n-3}) + \left[\binom{n}{2} + n \right] p(2, 1^{n-1}) \leq 1.$$

The maximum value of (5.4) is evidently equal to

$$\max\left\{\frac{\binom{n}{2}}{\binom{n}{2}+\binom{n}{3}}, \frac{\binom{n}{2}(n-2)}{\binom{n}{2}(n-2)+3\binom{n}{4}}, \frac{\binom{n}{2}}{\binom{n}{2}+n}\right\},$$

with the first expression corresponding to Π_{n+1} having 1 cluster of size 3 and $n-2$ clusters of size 1 with probability 1; the second expression corresponding to Π_{n+1} having 2 clusters of size 2 and $n-3$ clusters of size 1 with probability 1; and the third expression corresponding to Π_{n+1} having 1 cluster of size 2 and $n-1$ clusters of size 1 with probability 1. Simplifying each of the three expressions yields

$$\max\left\{\frac{3}{n+1}, \frac{4}{n+1}, \frac{n-1}{n+1}\right\} = \frac{\max\{4, n-1\}}{n+1}.$$

It follows from Proposition 5.1 that for $n = 3$, there are no restrictions on the distribution of $K_3(\Pi_4)$ on $\{1,2,3\}$. The same claim cannot be made for $n \geq 4$, as $\mathbb{P}(K_4(\Pi_5) = 3) \leq \frac{4}{5}$ and $\mathbb{P}(K_n(\Pi_{n+1}) = n-1) \leq \frac{n-1}{n+1}$ for $n \geq 5$.

The remainder of the section will focus on $K_3(\Pi_n)$ for $n \geq 3$. Intuitively, as $n \to \infty$, the set of probability distributions of $K_3(\Pi_n)$ should tend to the corresponding set for $K_3(\Pi)$ for exchangeable random partitions Π of $\mathbb{N}$, which was explicitly characterized in Section 2. Fix $n \geq 3$, and as before, consider the parameterization $q_1 = \mathbb{P}(K_3(\Pi_n) = 1)$ and $q_3 = \mathbb{P}(K_3(\Pi_n) = 3)$. By repeated application of (5.3), q_1 and q_3 may be written in terms of the EPPF as

$$q_1 = p(3) = \sum_{\substack{1\leq k\leq n\\ n_1+\ldots+n_k=n\\ n_1\geq\ldots\geq n_k\geq 1}} A(n_1,\ldots,n_k)p(n_1,\ldots,n_k)$$

and

$$q_3 = p(1,1,1) = \sum_{\substack{1\leq k\leq n\\ n_1+\ldots+n_k=n\\ n_1\geq\ldots\geq n_k\geq 1}} B(n_1,\ldots,n_k)p(n_1,\ldots,n_k)$$

for uniquely defined nonnegative integer coefficients $A(n_1,\ldots,n_k)$ and $B(n_1,\ldots,n_k)$. The problem is to describe the set of points (q_1, q_3) arising in this manner subject to

$$\sum_{\substack{1\leq k\leq n\\ n_1+\ldots+n_k=n\\ n_1\geq\ldots\geq n_k\geq 1}} C(n_1,\ldots,n_k)p(n_1,\ldots,n_k) = 1$$

where $C(n_1,\ldots,n_k)$ is as defined in (5.2). Observe that, in vector notation,

$$\begin{aligned}(q_1,q_3) &= \Big(\sum A(n_1,\ldots,n_k)p(n_1,\ldots,n_k), \sum B(n_1,\ldots,n_k)p(n_1,\ldots,n_k)\Big)\\ &= \sum C(n_1,\ldots,n_k)p(n_1,\ldots,n_k)\Big(\tfrac{A(n_1,\ldots,n_k)}{C(n_1,\ldots,n_k)}, \tfrac{B(n_1,\ldots,n_k)}{C(n_1,\ldots,n_k)}\Big)\end{aligned}$$

This shows that any (q_1, q_3) is a convex combination of points of the form $\big(\frac{A(\boldsymbol{n})}{C(\boldsymbol{n})}, \frac{B(\boldsymbol{n})}{C(\boldsymbol{n})}\big)$, and thus the set of probability distributions of $K_3(\Pi_n)$ over all exchangeable random partitions Π_n of $[n]$, expressed in the parameterization (q_1, q_3), is the convex hull of the finite set of points

$$S_n := \Big\{\Big(\tfrac{A(n_1,\ldots,n_k)}{C(n_1,\ldots,n_k)}, \tfrac{B(n_1,\ldots,n_k)}{C(n_1,\ldots,n_k)}\Big) : 1 \leq k \leq n, n_1+\ldots+n_k = 1, n_1 \geq \ldots \geq n_k \geq 1\Big\}.$$

Listed below is the sequence (s_n) for the number of extreme points of the convex hull of S_n for $3 \leq n \leq 35$, computed using SciPy [16].

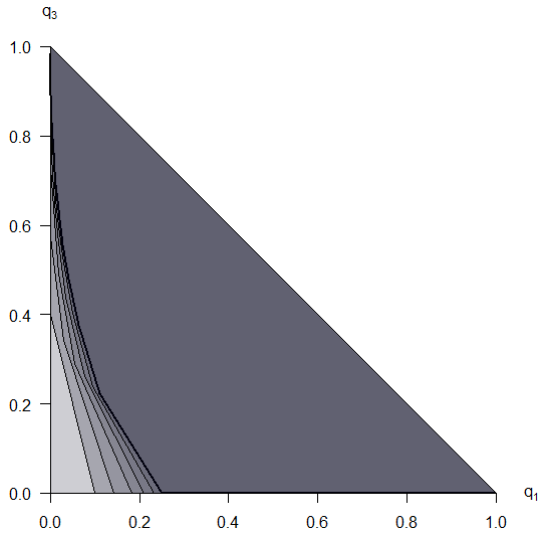

Figure 4: The nested regions are the possible probability distributions of $K_3(\Pi_n)$ for Π_n an exchangeable random partition of $[n]$ for $n = 4, 5, 7, 12, 19, 41$, which tend to the region corresponding to K_3 for infinite exchangeable sequences, as described in Theorem 1.2 and shown in Figure 1.

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
s_n	3	3	4	4	5	5	6	6	7	6	8	7	8	8	9	8	10

n	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
s_n	9	10	10	11	9	12	11	11	11	13	11	13	12	13	13	14

6 The two-parameter family

It was shown in [14] that any pair of real parameters (α, θ) satisfying either of the conditions

(i) $0 \le \alpha < 1$ and $\theta > -\alpha$; or

(ii) $\alpha < 0$ and $\theta = -m\alpha$ for some $m \in \mathbb{N}$

corresponds to an exchangeable random partition $\Pi_{\alpha,\theta} = (\Pi_n)$ of $\mathbb{N}$ according to the following sequential construction known as the Chinese restaurant process: for each $n \in \mathbb{N}$, conditionally given $\Pi_n = \{C_1, \ldots, C_k\}$, Π_{n+1} is formed by having $n+1$

$$\text{attach to cluster } C_i \text{ with probability } \frac{|C_i| - \alpha}{n+\theta}, \quad 1 \le i \le k\ ;$$
$$\text{form a new cluster with probability } \frac{\theta + k\alpha}{n+\theta}.$$

The corresponding EPPF is given by

$$p_{\alpha,\theta}(n_1, \ldots, n_k) = \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)\prod_{j=1}^{k}(1-\alpha)_{n_j-1}}{(\theta)_n}$$

where $n = n_1 + \ldots + n_k$ and

$$(x)_m := x(x+1)\cdots(x+m-1) = \frac{\Gamma(x+m)}{\Gamma(x)}.$$

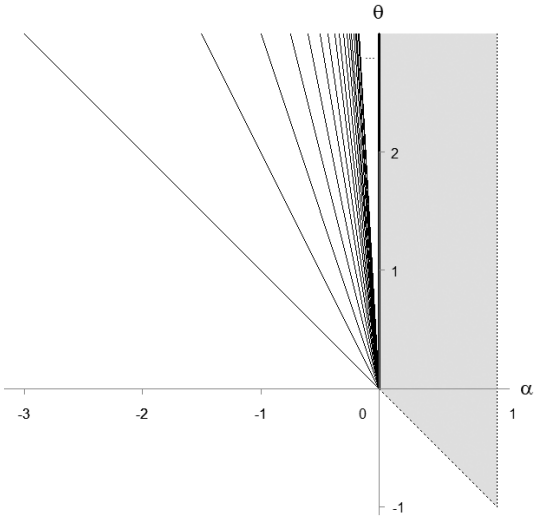

Figure 5: The (α, θ) parameter space.

Let $\mathbb{P}_{\alpha,\theta}$ denote the law of $\Pi_{\alpha,\theta}$. The distribution of K_3 for $\Pi_{\alpha,\theta}$ is given by

$$q_1(\alpha,\theta) = \frac{(1-\alpha)(2-\alpha)}{(1+\theta)(2+\theta)} \tag{6.1}$$

$$q_2(\alpha,\theta) = \frac{3(1-\alpha)(\theta+\alpha)}{(1+\theta)(2+\theta)}$$

$$q_3(\alpha,\theta) = \frac{(\theta+\alpha)(\theta+2\alpha)}{(1+\theta)(2+\theta)} \tag{6.2}$$

where

$$q_i(\alpha,\theta) := \mathbb{P}_{\alpha,\theta}(K_3 = i).$$

For $m > 0$, let

$$A_m := \left\{(m+m\theta,\theta) : -\tfrac{m}{m+1} < \theta < \tfrac{1-m}{m}\right\} \subseteq \{(\alpha,\theta) : 0 \le \alpha < 1, \theta > -\alpha\}$$

and let $A_0 := \{(0,\theta) : \theta > 0\}$, the parameter subspace corresponding to the well-known one-parameter Ewens sampling formula [4]. The line segments and one ray $\{A_m\}_{m\ge 0}$ with inverse slope m in the (α,θ) plane, each of which would pass through the point $(\alpha,\theta) = (0,-1)$ if extended, partition the parameter subspace $\{(\alpha,\theta) : 0 \le \alpha < 1, \theta > -\alpha\}$. Hence the distribution of K_3 can be reparameterized in m and θ as

$$q_1^{(m)}(\theta) = \frac{(1-m-m\theta)(2-m-m\theta)}{(1+\theta)(2+\theta)} \tag{6.3}$$

$$q_2^{(m)}(\theta) = \frac{3(1-m-m\theta)[m+(m+1)\theta]}{(1+\theta)(2+\theta)}$$

$$q_3^{(m)}(\theta) = \frac{[m+(m+1)\theta][2m+(2m+1)\theta]}{(1+\theta)(2+\theta)} \tag{6.4}$$

It can be checked by calculus that for each fixed $m > 0$,

- the function $q_1^{(m)}(\theta)$ is strictly decreasing for $\theta \in (-\frac{m}{m+1}, \frac{1-m}{m})$ with $\lim_{\theta\to-\frac{m}{m+1}} q_1^{(m)}(\theta) = 1$ and $\lim_{\theta\to\frac{1-m}{m}} q_1^{(m)}(\theta) = 0$.

- the function $q_3^{(m)}(\theta)$ is strictly increasing for $\theta \in (-\frac{m}{m+1}, \frac{1-m}{m})$ with $\lim_{\theta\to-\frac{m}{m+1}} q_3^{(m)}(\theta) = 0$ and $\lim_{\theta\to\frac{1-m}{m}} q_3^{(m)}(\theta) = 1$.
- the function $q_2^{(m)}(\theta)$ is strictly increasing on $\big(-\frac{m}{m+1}, \tau(m)\big]$ and strictly decreasing on $\big[\tau(m), \frac{1-m}{m}\big)$, with a unique maximum value of $9 - 6\big(\sqrt{(m+1)(m+2)} - m\big)$ at $\theta = \tau(m) := \frac{-m^2-3m+\sqrt{(m+1)(m+2)}}{1+3m+m^2}$, which is also the unique value of θ in the domain at which $q_1^{(m)}(\theta) = q_3^{(m)}(\theta)$.

The properties above also hold for $m = 0$ after slight modification by replacing each instance of $\frac{1-m}{m}$ with $\lim_{m\to0^+} \frac{1-m}{m} = \infty$, and this remark also applies to subsequent discussion.

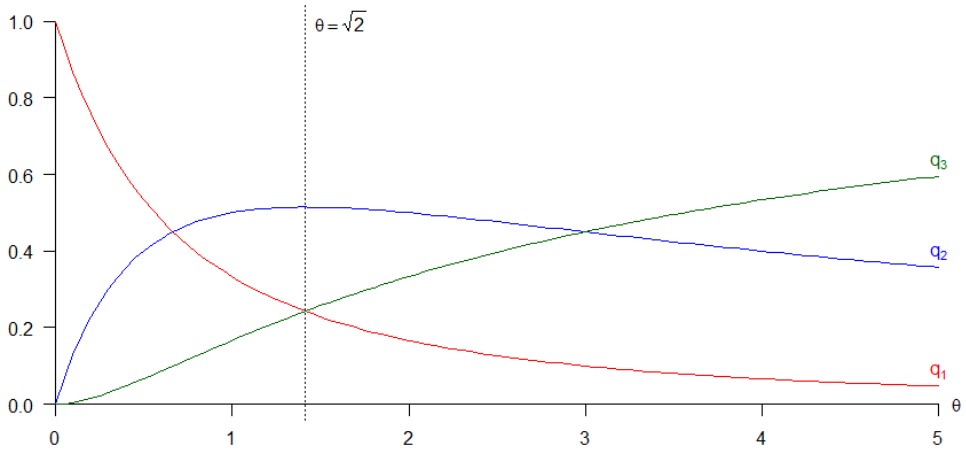

Figure 6: Graphs of $q_i^{(m)}(\theta)$ for $m = 0$ and $\theta \in [0, 5]$. Observe that q_1 and q_3 intersect at the same value of θ as where q_2 attains its maximum value. The corresponding graphs for every $m > 0$ also share this property.

Duality. The last observation implies that for $m \geq 0$ and any real number p such that $0 < p < 9 - 6(\sqrt{(m+1)(m+2)} - m)$, there are exactly two values $\theta_\pm^{(m)}(p)$ with

$$-\frac{m}{m+1} < \theta_-^{(m)}(p) < \tau(m) < \theta_+^{(m)}(p) < \frac{1-m}{m}. \tag{6.5}$$

satisfying

$$q_2^{(m)}(\theta_-^{(m)}(p)) = q_2^{(m)}(\theta_+^{(m)}(p)).$$

For $p = 9 - 6(\sqrt{(m+1)(m+2)} - 2)$, define $\theta_-^{(m)}(p) = \theta_+^{(m)}(p) = \varphi(m)$. As $\theta_\pm^{(m)}(p)$ are defined as the solutions to the equation

$$\frac{3(1-m-m\theta)[m+(m+1)\theta]}{(1+\theta)(2+\theta)} = p$$

or equivalently the quadratic equation

$$p(1+\theta)(2+\theta) - 3(1-m-m\theta)[m+(m+1)\theta] = 0, \tag{6.6}$$

we have the polynomial identity

$$(\theta - \theta_+^{(m)}(p))(\theta - \theta_-^{(m)}(p)) = \theta^2 + \frac{3p-3+6m^2}{p+3m+3m^2}\theta + \frac{2p-3m+3m^2}{p+3m+3m^2}$$

after rearranging (6.6). It follows that

$$\theta_+^{(m)}\theta_-^{(m)} = \frac{2p - 3m + 3m^2}{p + 3m + 3m^2}. \tag{6.7}$$

For $-\frac{m}{m+1} < \theta < \frac{1-m}{m}$, define the *$m$-dual* $\theta_*^{(m)}$ of θ according to (6.5). Rearranging (6.7) and simplifying gives the explicit formula

$$\theta_*^{(m)} = \frac{2 - m(3+m)(1+\theta)}{\theta + m(3+m)(1+\theta)}. \tag{6.8}$$

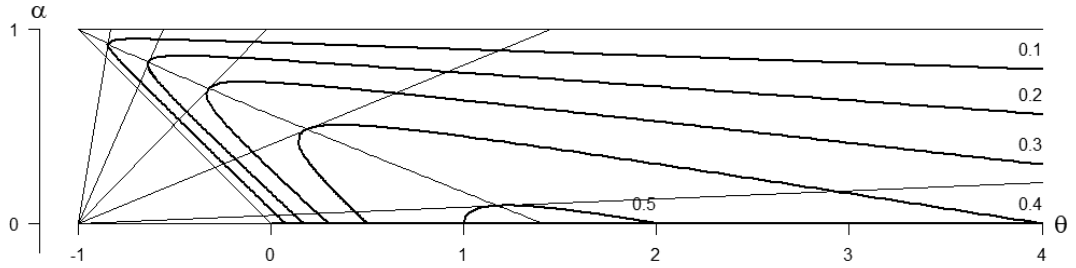

Figure 7: Contour plot of $q_2(\alpha, \theta)$. The level curves for $q_2(\alpha, \theta) \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ are shown, along with their tangent lines where they meet the curve $q_1(\alpha, \theta) = q_3(\alpha, \theta)$. Observe that each tangent line passes through the point $(\alpha, \theta) = (0, -1)$. Note that here α is plotted on the vertical axis, for convenience of display.

Theorem 6.1. *For $m \geq 0$ and $-\frac{m}{m+1} < \theta < \frac{1-m}{m}$, we have*

$$q_1^{(m)}(\theta_*^{(m)}) = q_3^{(m)}(\theta) \qquad \textit{and} \qquad q_3^{(m)}(\theta_*^{(m)}) = q_1^{(m)}(\theta).$$

Proof. It suffices to verify the first of the two identities since (6.8) is constructed as an involution. Let $D(m, \theta)$ be the denominator in (6.8). Substituting and simplifying yields

$$\begin{aligned}
1 + \theta_*^{(m)} &= \frac{2 + \theta}{D(m, \theta)} \ ; \\
2 + \theta_*^{(m)} &= \frac{(1+\theta)(1+m)(2+m)}{D(m, \theta)} \ ; \\
1 - m - m\theta_*^{(m)} &= \frac{(1+m)[m + (m+1)\theta]}{D(m, \theta)} \ ; \\
2 - m - m\theta_*^{(m)} &= \frac{(2+m)[2m + (2m+1)\theta]}{D(m, \theta)}.
\end{aligned}$$

Hence we have

$$\begin{aligned}
q_1^{(m)}(\theta_*^{(m)}) &= \frac{(1 - m - m\theta_*^{(m)})(2 - m - m\theta_*^{(m)})}{(1 + \theta_*^{(m)})(2 + \theta_*^{(m)})} \\
&= \frac{[m + (m+1)\theta][2m + (2m+1)\theta]}{(1+\theta)(2+\theta)} \\
&= q_3^{(m)}(\theta)
\end{aligned}$$

as desired. □

Symmetry. A consequence of Theorem 6.1 is a surprising symmetry in the set of laws of K_3 arising from the two-parameter model. To make this observation explicit, for any $m \geq 0$ we solve for $q_3 = q_3^{(m)}$ in terms of $q_1 = q_1^{(m)}$ as defined in (6.3) and (6.4) to obtain the formula

$$q_3 = \varphi_m(q_1) := 1 + \frac{3}{4}m + \frac{5}{4}q_1 - \frac{3}{4}\sqrt{m^2 + 6q_1m + q_1(8+q_1)}. \tag{6.9}$$

Rearranging to eliminate the radical yields the relation

$$(4+3m)(q_1+q_3) + 5q_1q_3 - 2(q_1^2+q_3^2) - 2 - 3m = 0$$

which verifies the symmetry. For $m = 0$ the identity reduces to

$$h(q_1,q_3) := 4(q_1+q_3) + 5q_1q_3 - 2(q_1^2+q_3^2) - 2 = 0. \tag{6.10}$$

This appears to be an exclusive property of the case $n = 3$, as no similar symmetry appears to manifest for larger n.

Theorem 6.2. *The mapping* $(\alpha,\theta) \mapsto (q_1,q_3)$ *defined by* (6.1) *and* (6.2) *is a bijection between the regions*

$$\{(\alpha,\theta) : 0 \leq \alpha < 1,\ \theta > -\alpha\} \qquad \textit{and} \qquad \{(q_1,q_3) : h(q_1,q_3) \geq 0,\ q_1+q_3 < 1\}$$

where $h(q_1,q_3)$ *is defined as in* (6.10).

Proof. Consider $\varphi(m,q_1) := \varphi_m(q_1)$ as in (6.9). To show the desired bijection, it suffices to show that for every fixed $0 < q_1 < 1$ that (i) $\varphi(m,q_1)$ is increasing in m, and (ii) $\lim_{m\to\infty}\varphi(m,q_1) = 1-q_1$.

(i)

$$\frac{\partial}{\partial m}\varphi(m,q_1) = \frac{3}{4}\Big(1 - \frac{2m+6q_1}{2\sqrt{m^2+6q_1m+q_1(8+q_1)}}\Big) > \frac{3}{4}\Big(1 - \frac{2m+6q_1}{2\sqrt{m^2+6q_1m+9q_1^2}}\Big) = 0$$

(ii)

$$\begin{aligned}
\lim_{m\to\infty}\varphi(m,q_1) &= \lim_{m\to\infty} 1 + \frac{5}{4}q_1 + \frac{3}{4}\left(\frac{m^2-(m^2+6q_1m+q_1(8+q_1))}{m+\sqrt{m^2+6q_1m+q_1(8+q_1)}}\right)\\
&= \lim_{m\to\infty} 1 + \frac{5}{4}q_1 + \frac{3}{4}\left(\frac{-6q_1 - \frac{q_1(8+q_1)}{m}}{1+\sqrt{1+\frac{6q_1}{m}+\frac{q_1(8+q_1)}{m^2}}}\right)\\
&= 1-q_1
\end{aligned}$$

□

Explicit inverse. Define the ratios

$$r(\alpha,\theta) := \frac{q_1(\alpha,\theta)}{q_2(\alpha,\theta)} = \frac{2-\alpha}{3(\theta+\alpha)}, \qquad s(\alpha,\theta) := \frac{q_2(\alpha,\theta)}{q_3(\alpha,\theta)} = \frac{3(1-\alpha)}{(\theta+2\alpha)}$$

These ratios uniquely define the law of K_3 for the corresponding (α,θ). The map $(\theta,\alpha) \mapsto (r,s)$ can be explicitly inverted as

$$\alpha(r,s) = \frac{9r-2s}{9r-s+3rs}, \qquad \theta(r,s) = \frac{3-9r+4s}{9r-s+3rs}$$

Expressed in terms of q_1 and q_3, this gives the inversion formulas

$$\begin{aligned}
\alpha(q_1,q_3) &= \frac{4q_1+4q_3+5q_1q_3-2q_1^2-2q_3^2-2}{5q_1+2q_3+4q_1q_3-4q_1^2-q_3^2-1},\\
\theta(q_1,q_3) &= -\frac{8q_1+5q_3+4q_1q_3-4q_1^2-q_3^2-4}{5q_1+2q_3+4q_1q_3-4q_1^2-q_3^2-1}.
\end{aligned}$$

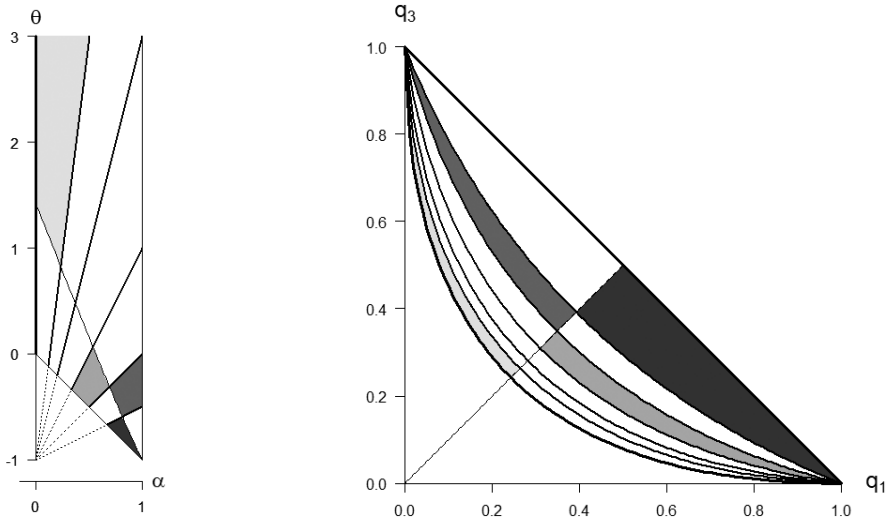

Figure 8: The bijection of Theorem 6.2. The regions colored in different shades of gray reveal the geometry of the bijection.

Note that the numerator in the formula for $\alpha(q_1, q_3)$ is equal to $h(q_1, q_3)$ as defined in (6.10). It is easy to verify that these formulas give an algebraic inverse. Observe that the denominator which is the same in both formulas is nonvanishing on the region $\{(q_1, q_3) : h(q_1, q_3) \geq 0,\ q_1 + q_3 < 1\}$, since

$$2(5q_1 + 2q_3 + 4q_1q_3 - 4q_1^2 - q_3^2 - 1) = h(q_1, q_3) + 6q_1 - 6q_1^2 + 3q_1q_3 > 0.$$

Corollary 6.3. *For any parameters (α, θ) with $0 \leq \alpha < 1$ and $\theta > -\alpha$, there exists a unique pair (α_*, θ_*) with $0 \leq \alpha_* < 1$ and $\theta_* > -\alpha_*$ such that*

$$q_{2\pm1}(\alpha, \theta) = q_{2\mp1}(\alpha_*, \theta_*).$$

Explicit formulas for α_* and θ_* in terms of α and θ can be computed as

$$\alpha^* = \frac{(2 - 3\alpha)(1 + \theta) - \alpha^2}{(\theta + 3\alpha)(1 + \theta) + \alpha^2}$$
$$\theta^* = \frac{\alpha(2 + \theta)}{(\theta + 3\alpha)(1 + \theta) + \alpha^2}.$$

Exceptional parameters. $\alpha < 0$, $\theta = -m\alpha$ for some $m \in \mathbb{N}$

It is well-known that in this case, the exchangeable random partition (Π_n) of $\mathbb{N}$ generated according to the Chinese restaurant construction is distributed as if by sampling from a symmetric Dirichlet distribution with m parameters equal to $-\alpha$ [13]. Hence for fixed $m \in \mathbb{N}$, as $\alpha \downarrow -\infty$ the exchangeable random partition of $\mathbb{N}$ corresponding to the parameter pair $(\alpha, \theta) = (\alpha, -m\alpha)$ converges in distribution to that obtained by sampling from the discrete uniform distribution on m elements. For K_3, the (α, θ) to (q_1, q_3) correspondence can be seen in Figure 9.

7 Complements

In this section, we point out an interesting convexity property for the the law of K_3. With notation as in Section 3, for $\boldsymbol{p} \in \nabla_\infty$, let

$$\boldsymbol{Q}(\boldsymbol{p}) := \big(q_1(\boldsymbol{p}), q_3(\boldsymbol{p})\big)$$

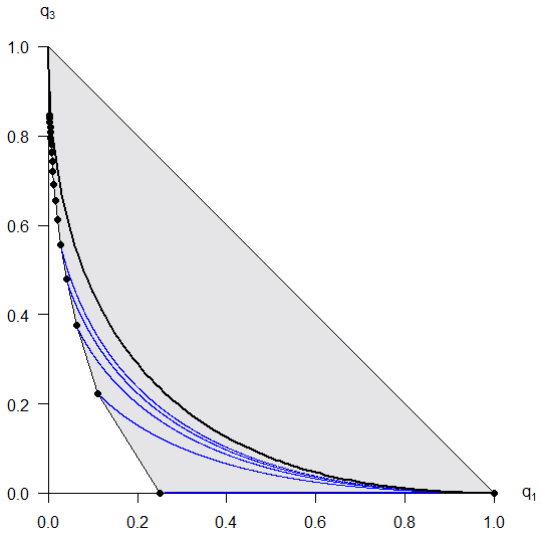

Figure 9: The blue curves correspond to the images of $(\alpha, \theta) = (\alpha, -m\alpha)$ for $\alpha \in (-\infty, 0)$ and fixed m under the $(\alpha, \theta) \mapsto (q_1, q_3)$ map, for $m = 2, 3, 4, 5, 6$. The curve defined by (6.10) is included in black.

be the mapping from a ranked discrete distribution to its corresponding law of K_3 obtained by i.i.d. sampling. In Section 3 we established that the range of $\boldsymbol{Q}$ is a (strict) subset of the closed convex hull of the set of points $\{\boldsymbol{Q}(\boldsymbol{u}_N) : N \in \mathbb{N}\}$. Note that the range of $\boldsymbol{Q}$ includes only distributions of K_3 which arise from i.i.d. sampling. Here are some preliminary efforts to better understand the geometry of this mapping.

Proposition 7.1. *For any $0 \le \lambda \le 1$ and $N \ge 1$,*

$$\boldsymbol{Q}(\lambda \boldsymbol{u}_N + (1-\lambda)\boldsymbol{u}_{2N}) = \lambda^2 \boldsymbol{Q}(\boldsymbol{u}_N) + (1-\lambda^2)\boldsymbol{Q}(\boldsymbol{u}_{2N})$$

Proof. We have

$$\lambda \boldsymbol{u}_N + (1-\lambda)\boldsymbol{u}_{2N} = \big(\underbrace{\tfrac{1+\lambda}{2N}, \dots, \tfrac{1+\lambda}{2N}}_{N \text{ times}}, \underbrace{\tfrac{1-\lambda}{2N}, \dots, \tfrac{1-\lambda}{2N}}_{N \text{ times}}\big).$$

Hence

$$q_1(\lambda \boldsymbol{u}_N + (1-\lambda)\boldsymbol{u}_{2N}) = N\Big(\frac{1+\lambda}{2N}\Big)^3 + N\Big(\frac{1-\lambda}{2N}\Big)^3 = \frac{1+3\lambda^2}{4N^2}$$

and

$$\begin{aligned} q_3(\lambda \boldsymbol{u}_N + (1-\lambda)\boldsymbol{u}_{2N}) &= \binom{N}{3}\Big(\frac{1+\lambda}{2N}\Big)^3 + \binom{N}{2} N \Big(\frac{1+\lambda}{2N}\Big)^2 \Big(\frac{1-\lambda}{2N}\Big) \\ &\quad + N\binom{N}{2}\Big(\frac{1+\lambda}{2N}\Big)\Big(\frac{1-\lambda}{2N}\Big)^2 + \binom{N}{3}\Big(\frac{1-\lambda}{2N}\Big)^3 \\ &= \binom{N}{3}\frac{1+3\lambda^2}{4N^3} + N\binom{N}{2}\frac{1-\lambda^2}{4N^3} \\ &= \frac{N-1}{3}\Big(\frac{2N-1-3\lambda^2}{4N^2}\Big). \end{aligned}$$

On the other side,

$$\lambda^2 q_1(\boldsymbol{u}_N) + (1-\lambda^2) q_1(\boldsymbol{u}_{2N}) = \frac{\lambda^2}{N^2} + \frac{1-\lambda^2}{4N^2} = \frac{1+3\lambda^2}{4N^2}$$

and

$$\begin{aligned}\lambda^2 q_3(\boldsymbol{u}_N) + (1-\lambda^2) q_3(\boldsymbol{u}_{2N}) &= \lambda^2 \binom{N}{3} \frac{1}{N^3} + (1-\lambda^2) \binom{2N}{3} \frac{1}{8N^3} \\ &= \frac{N(N-1)(N-2)}{6} \cdot \frac{\lambda^2}{N^3} + \frac{2N(2N-1)(2N-2)}{6} \cdot \frac{1-\lambda^2}{8N^3} \\ &= \frac{N-1}{3}\Big(\frac{2N-1-3\lambda^2}{4N^2}\Big). \end{aligned}$$

□

References

[1] S. Dharmadhikari and K. Joag-Dev, *Unimodality, convexity, and applications*, Probability and Mathematical Statistics, Academic Press, Inc., Boston, MA, 1988. MR954608

[2] P. Diaconis, *Finite forms of de Finetti's theorem on exchangeability*, Synthese **36** (1977), no. 2, 271–281, Foundations of probability and statistics, II. MR517222

[3] R. Durrett, *Probability: theory and examples*, fourth ed., Cambridge Series in Statistical and Probabilistic Mathematics, vol. 31, Cambridge University Press, Cambridge, 2010. MR2722836

[4] W. J. Ewens, *The sampling theory of selectively neutral alleles*, Theoret. Population Biol. **3** (1972). MR325177

[5] W. Feller, *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons, Inc., New York-London-Sydney, 1968. MR0228020

[6] A. Gnedin, B. Hansen, and J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv. **4** (2007), 146–171. MR2318403

[7] S. Janson, T. Konstantopoulos, and L. Yuan, *On a representation theorem for finitely exchangeable random vectors*, J. Math. Anal. Appl. **442** (2016), no. 2, 703–714. MR3504021

[8] S. Karlin, *Central limit theorems for certain infinite urn schemes*, J. Math. Mech. **17** (1967), 373–401. MR0216548

[9] A. Ya. Khintchine, *On unimodal distributions*, Izvestiya Nauchno-Issledovatel'skogo Instituta Matematiki i Mekhaniki **2** (1938), no. 2, 1–7.

[10] J. F. C. Kingman, *The representation of partition structures*, J. London Math. Soc. (2) **18** (1978), no. 2, 374–380. MR509954

[11] T. Konstantopoulos and L. Yuan, *On the extendibility of finitely exchangeable probability measures*, Trans. Amer. Math. Soc. **371** (2019), no. 10, 7067–7092. MR3939570

[12] L. A. Petrov, *A two-parameter family of infinite-dimensional diffusions on the Kingman simplex*, Funktsional. Anal. i Prilozhen. **43** (2009), no. 4, 45–66. MR2596654

[13] J. Pitman, *Combinatorial stochastic processes*, Lecture Notes in Mathematics, vol. 1875, Springer-Verlag, Berlin, 2006, Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. MR2245368

[14] J. Pitman, *Exchangeable and partially exchangeable random partitions*, Probab. Theory Related Fields **102** (1995), no. 2, 145–158. MR1337249

[15] F. Spitzer, *Principles of random walk*, second ed., Springer-Verlag, New York-Heidelberg, 1976, Graduate Texts in Mathematics, Vol. 34. MR0388547

[16] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, F. Yu, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods **17** (2020), 261–272.

[17] Y. Yakubovich, *On the distribution of the number of distinct values in a finite sample*, Unpublished (2021).

Acknowledgments. Many thanks to my advisor Jim Pitman for suggesting this problem and providing invaluable guidance, to Yuri Yakubovich for his significant contributions, and to the referees for their insightful feedback and excellent suggestions.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System: `https://vtex.lt/services/ejms-peer-review/`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `https://imstat.org/shop/donation/`