Feature Allocations, Probability Functions, and Paintboxes

Tamara Broderick, Jim Pitman, Michael I. Jordan

Abstract

The problem of inferring a clustering of a data set has been the subject of much research in Bayesian analysis, and there currently exists a solid mathematical foundation for Bayesian approaches to clustering. In particular, the class of probability distributions over partitions of a data set has been characterized in a number of ways, including via exchangeable partition probability functions (EPPFs) and the Kingman paintbox. Here, we develop a generalization of the clustering problem, called feature allocation, where we allow each data point to belong to an arbitrary, non-negative integer number of groups, now called features or topics. We define and study an "exchangeable feature probability function" (EFPF)—analogous to the EPPF in the clustering setting—for certain types of feature models. Moreover, we introduce a "feature paintbox" characterization—analogous to the Kingman paintbox for clustering—of the class of exchangeable feature models. We provide a further characterization of the subclass of feature allocations that have EFPF representations.

Keywords: feature, feature allocation, paintbox, EFPF, feature frequency model, Indian buffet process, beta process

1 Introduction

Exchangeability has played a key role in the development of Bayesian analysis in general and Bayesian nonparametric analysis in particular. Exchangeability can be viewed as asserting that the indices used to label the data points are irrelevant for inference, and as such is often a natural modeling assumption. Under such an assumption, one is licensed by de Finetti's theorem [De Finetti, 1931, Hewitt and Savage, 1955] to propose the existence of an underlying parameter that renders the data conditionally independent and identically distributed (iid) and to place a prior distribution on that parameter. Moreover, the theory of infinitely exchangeable sequences has advantages of simplicity over the theory of finite exchangeability, encouraging modelers to take a nonparametric stance in which the underlying "parameter" is infinite dimensional. Finally, the development of algorithms for posterior inference is often greatly simplified by the assumption of exchangeability, most notably in the case of Bayesian nonparametrics, where models based on the Dirichlet process and other combinatorial

priors became useful tools in practice only when it was realized how to exploit exchangeability to develop inference procedures [Escobar, 1994].

The connection of exchangeability to Bayesian nonparametric modeling is well established in the case of models for clustering. The goal of a clustering procedure is to infer a partition of the data points. In the Bayesian setting, one works with random partitions, and, under an exchangeability assumption, the distribution on partitions should be invariant to a relabeling of the data points. The notion of an exchangeable random partition has been formalized by Kingman, Aldous, and others [Kingman, 1978, Aldous, 1985], and has led to the definition of an *exchangeable partition probability function* (EPPF) [Pitman, 1995]. The EPPF is a mathematical function of the cardinalities of the groups in a partition. Exchangeability of the random partition is captured by the requirement that the EPPF be a symmetric function of these cardinalities. Furthermore, the exchangeability of a partition can be related to the exchangeability of a sequence of random variables representing the assignments of data points to clusters, for which a de Finetti mixing measure necessarily exists. This de Finetti measure is known as the *Kingman paintbox* [Kingman, 1978]. The relationships among this circle of ideas are well understood: it is known that there is an equivalence among the class of exchangeable random partitions, the class of random partitions that possess an EPPF, and the class of random partitions generated by a Kingman paintbox; see Pitman [2006] for an overview of these relations. A specific example of these relationships is given by the Chinese restaurant process and the Dirichlet process, but several other examples are known and have proven useful in Bayesian nonparametrics.

Our focus in the current paper is on an alternative to clustering models that we refer to as *feature allocation models*. While in a clustering model each data point is assigned to one and only one class, in a feature allocation model each data point can belong to multiple groups. It is often natural to view the groups as corresponding to traits or features, such that the notion that a data point belongs to multiple groups corresponds to the point exhibiting multiple traits or features. A Bayesian feature allocation model treats the feature assignments for a given data point as random and subject to posterior inference. A nonparametric Bayesian feature allocation model takes the number of features to also be random and subject to inference.

Research on nonparametric Bayesian feature allocation has been based around a single prior distribution, the Indian buffet process of Griffiths and Ghahramani [2006], which is known to have the beta process as its underlying de Finetti measure [Thibaux and Jordan, 2007]. There does not yet exist a general definition of exchangeability for feature allocation models, nor counterparts of the EPPF or the Kingman paintbox.

In this paper we supply these missing constructions. We provide a rigorous treatment of exchangeable feature allocations (in Section 2 and Section 3). In Section 4 we define a notion of *exchangeable feature probability function* (EFPF) that is the analogue for feature allocations of the EPPF for clustering. We then proceed to define a *feature paintbox* in Section 5. Finally, in Section 6 we discuss a class of models that we refer to as *feature frequency models* for which

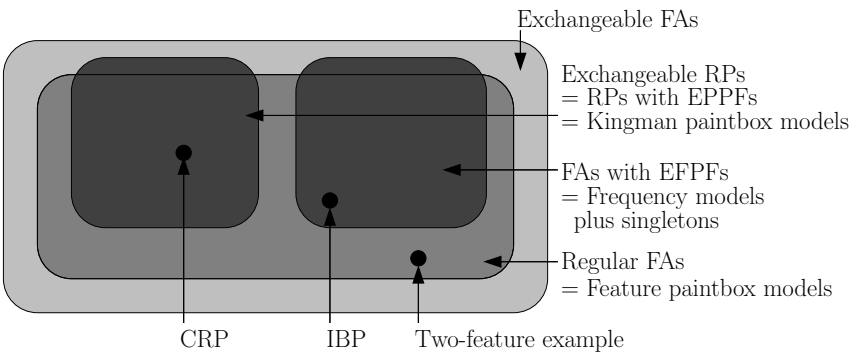

Figure 1: A summary of the relations described in this paper. Rounded rectangles represent classes with the following abbreviations: RP for random partition, FA for random feature allocation, EPPF for exchangeable partition probability function, EFPF for exchangeable feature probability function. The large black dots represent particular models with the following abbreviations: CRP for Chinese restaurant process, IBP for Indian buffet process. The two-feature example refers to Example 9 with the choice $p_{11}p_{00} \neq p_{10}p_{01}$.

the construction of the feature paintbox is particularly straightforward, and we discuss the important role that feature frequency models play in the general theory of feature allocations.

The Venn diagram shown in Figure 1 is a useful guide for understanding our results, and the reader may wish to consult this diagram in working through the paper. As shown in the diagram, random partitions (RPs) are a special case of random feature allocations (FAs), and previous work on random partitions can be placed within our framework. Thus, in the diagram, we have depicted the equivalence already noted of exchangeable RPs, RPs that possess an EPPF, and Kingman paintboxes. We also see that random feature allocations have a somewhat richer structure: the class of FAs with EFPFs is not the same as those having an underlying feature paintbox. But the class of EFPFs is characterized in a different way; we will see that the class of feature allocations with EFPFs is equivalent to the class of FAs obtained from feature frequency models together with singletons of a certain distribution. Indeed, we will find that the class of clusterings with EPPFs is, in this way, analogous to the class of feature allocations with EFPFs when both are considered as subclasses of the general class of feature allocations. The diagram also shows several examples that we use to illustrate and develop our theory.

2 Feature allocations

We consider data sets with N points and let the points be indexed by the integers $[N] := \{1, 2, \ldots, N\}$. We also explicitly allow $N = \infty$, in which case

the index set is $\mathbb{N} = \{1, 2, 3, \ldots\}$. For our discussion of feature allocations and partitioning it is sufficient to focus on the indices rather than the data points; thus, we will be discussing models for collections of subsets of $[N]$ and $\mathbb{N}$.

Our introduction to feature allocations follows Broderick et al. [2012b]. We define a *feature allocation* f_N of $[N]$ to be a multiset of non-empty subsets of $[N]$ called *features*, such that no index n belongs to infinitely many features. We write $f_N = \{A_1, \ldots, A_K\}$, where K is the number of features. An example feature allocation of [6] is $f_6 = \{\{2, 3\}, \{2, 4, 6\}, \{3\}, \{3\}, \{3\}\}$. Similarly, a feature allocation f_∞ of $\mathbb{N}$ is a multiset of non-empty subsets of $\mathbb{N}$ such that no index n belongs to infinitely many features. The total number of features in this case may be infinite, in which case we write $f_\infty = \{A_1, A_2, \ldots\}$. An example feature allocation of $\mathbb{N}$ is $f_\infty = \{\{n : n \text{ is prime}\}, \{n : n \text{ is not divisible by two}\}\}$. Finally, we may have $K = 0$, and $f_\infty = \emptyset$ is a valid feature allocation.

A *partition* is a special case of a feature allocation for which the features are restricted to be mutually exclusive and exhaustive. The features of a partition are often referred to as *blocks* or *clusters*. We note that a partition is always a feature allocation, but the converse statement does not hold in general; neither of the examples given above (f_6 and f_∞) are partitions.

We now turn to the problem of defining exchangeable feature allocations, extending previous work on exchangeable random partitions [Aldous, 1985]. Let $\mathcal{F}_N$ be the space of all feature allocations of $[N]$. A *random feature allocation* F_N of $[N]$ is a random element of $\mathcal{F}_N$. Let $\sigma : \mathbb{N} \to \mathbb{N}$ be a finite permutation. That is, for some finite value N_σ, we have $\sigma(n) = n$ for all $n > N_\sigma$. Further, for any feature $A \subset \mathbb{N}$, denote the permutation applied to the feature as follows: $\sigma(A) := \{\sigma(n) : n \in A\}$. For any feature allocation F_N, denote the permutation applied to the feature allocation as follows: $\sigma(F_N) := \{\sigma(A) : A \in F_N\}$. Finally, let F_N be a random feature allocation of $[N]$. Then we say that a random feature allocation F_N is *exchangeable* if $F_N \stackrel{d}{=} \sigma(F_N)$ for every permutation of $[N]$.

In addition to exchangeability, we also require our distributions on feature allocations to exhibit a notion of coherence across different ranges of the index. Intuitively, we often imagine the indices as denoting time, and it is natural to suppose that the randomness at time n is coherent with the randomness at time $n+1$. More formally, we say that a feature allocation f_M of $[M]$ is the *restriction* of a feature allocation f_N of $[N]$ for $M < N$ if

$$f_M = \{A \cap [M] : A \in f_N, A \cap [M] \neq \emptyset\}.$$

Let $\mathcal{R}_N(f_M)$ be the set of all feature allocations of $[N]$ whose restriction to $[M]$ is f_M.

Let $\mathbb{P}$ denote a probability measure on some probability space supporting (F_n). We say that the sequence of random feature allocations (F_n) is *consistent in distribution* if for all M and N such that $M < N$, we have

$$\mathbb{P}(F_M = f_M) = \sum_{f_N \in \mathcal{R}_N(f_M)} \mathbb{P}(F_N = f_N).$$

We say that the sequence (F_n) is *strongly consistent* if for all M and N such

that $M < N$, we have

$$F_N \overset{a.s.}{\in} \mathcal{R}_N(F_M).$$

Given any (F_n) that is consistent in distribution, the Kolmogorov extension theorem implies that we can construct a sequence of random feature allocations that is strongly consistent and has the same finite dimensional distributions. So henceforth we simply use the term "consistency" to refer to strong consistency.

With this consistency condition, we can define a random feature allocation F_∞ of $\mathbb{N}$ as a consistent sequence of finite feature allocations. Thus F_∞ may be thought of as a random element of the space of such sequences: $F_\infty = (F_n)_{n=1}^\infty$. We say that F_N is a restriction of F_∞ to $[N]$ when it is the Nth element in this sequence. We let $\mathcal{F}_\infty$ denote the space of consistent feature allocation sequences, of which each random feature allocation is a random element. The sigma field associated with this space is generated by the finite-dimensional sigma fields of the restricted random feature allocations F_n.

We say that F_∞ is exchangeable if $F_\infty \overset{d}{=} \sigma(F_\infty)$ for every finite permutation σ. That is, for every permutation σ that changes no indices above N for some $N < \infty$, we require $F_N \overset{d}{=} \sigma(F_N)$, where F_N is the restriction of F_∞ to $[N]$.

3 Labeling features

Now that we have defined consistent, exchangeable random feature allocations, we want to characterize the class of all distributions on these allocations. We begin by considering some alternative representations of the feature allocation that are not merely useful, but indeed key to some of our later results.

A number of authors have made use of matrices as a way of representing feature allocations [Griffiths and Ghahramani, 2006, Thibaux and Jordan, 2007, Doshi et al., 2009]. This representation, while a boon for intuition in some regards, requires care because a matrix presupposes an order on the features, which is not a part of the feature allocation a priori. We cover this distinction in some detail next.

We start by defining an *a priori labeled feature allocation*. Let $\hat{F}_{N,1}$ be the collection of indices in $[N]$ with feature 1, let $\hat{F}_{N,2}$ be the collection of indices in $[N]$ with feature 2, etc. Here, we think of a priori labels as being the ordered, positive natural numbers. This specification is different from (a priori unlabeled) feature allocations as defined above since there is nothing to distinguish the features in a feature allocation other than, potentially, the members of a feature. Consider the following analogy: an a priori labeled feature allocation is to a feature allocation as a classification is to a clustering. Indeed, when each index n belongs to exactly one feature in an a priori feature allocation, feature 1 is just class 1, feature 2 is class 2, and so on.

Another way to think of an a priori labeled feature allocation of $[N]$ is as a matrix of N rows filled with zeros and ones. Each column is associated with a feature. The (n, k) entry in the matrix is one if index n is in feature k and zero otherwise. However, just as—contrary to the classification case—we do

not know the ordering of clusters in a clustering a priori, we do not a priori know the ordering of features in a feature allocation. To make use of a matrix representation for a feature allocation, we will need to introduce or find such an order.

The reasoning above suggests that introducing an order for features in a feature allocation would be useful. The next example illustrates that the probability $\mathbb{P}(F_N = f_N)$ in some sense undercounts features when they contain exactly the same indices: e.g., $A_j = A_k$ for some $j \neq k$. This fact will suggest to us that it is not merely useful, but indeed a key point of our theoretical development, to introduce an ordering on features.

Example 1 (A Bernoulli, two-feature allocation)**.** Given $q_A, q_B \in (0,1)$, draw $Z_{n,A} \overset{iid}{\sim} \text{Bern}(q_A)$ and $Z_{n,B} \overset{iid}{\sim} \text{Bern}(q_B)$, independently, and construct the random feature allocation by collecting those indices with successful draws:

$$F_N := \{\{n : n \leq N, Z_{n,A} = 1\}, \{n : n \leq N, Z_{n,B} = 1\}\}.$$

One caveat here is that if either of the two sets in the multiset F_N is empty, we do not include it in the allocation. Note that calling the features A and B was merely for the purposes of construction, and in defining F_N, we have lost all feature labels. So F_N is a feature allocation, not an a priori labeled feature allocation.

Then the probability of the feature allocation $F_5 = f_5 := \{\{2,3\},\{2,3\}\}$ is

$$q_A^2(1-q_A)^3 q_B^2(1-q_B)^3,$$

but the probability of the feature allocation $F_5 = f_5' := \{\{2,3\},\{2,5\}\}$ is

$$2q_A^2(1-q_A)^3 q_B^2(1-q_B)^3.$$

The difference is that in the latter case the features can be distinguished, and so we must account for the two possible pairings of features to frequencies $\{q_A, q_B\}$.

Now, instead, let $\tilde{F}_N$ be F_N with the features ordered uniformly at random amongst all possible feature orderings. There is just a single possible ordering of f_5, so the probability of $\tilde{F}_5 = \tilde{f}_5 := (\{2,3\},\{2,3\})$ is again

$$q_A^2(1-q_A)^3 q_B^2(1-q_B)^3.$$

However, there are two orderings of f_5', each of which is equally likely. The probability of $\tilde{F}_N = \tilde{f}_5' := (\{2,5\},\{2,3\})$ is

$$q_A^2(1-q_A)^3 q_B^2(1-q_B)^3.$$

The same holds for the other ordering. ■

This example suggests that there are combinatorial factors that must be taken into account when working with the distribution of F_N directly. The example also suggests that we can avoid the need to specify such factors by

instead working with a suitable randomized ordering of the random feature allocation F_N. We achieve this ordering in two steps.

The first step involves ordering the features via a procedure that we refer to as *order-of-appearance labeling*. The basic idea is that we consider data indices $n = 1, 2, 3$, and so on in order. Each time a new data point arrives, we examine the features associated with that data point. Each time we see a new feature, we label it with the lowest available feature label from $k = 1, 2, \ldots$.

In practice, the order-of-appearance scheme requires some auxiliary randomness since each index n may belong to zero, one, or many different features (though the number must be finite). When multiple features first appear for index n, we order them uniformly at random. That simple idea is explained in full detail as follows. Recursively suppose that there are K features among the indices $[N-1]$. Trivially there are zero features when no indices have been seen yet. Moreover, we suppose that we have features with labels 1 through K if $K \geq 1$, and if $K = 0$, we have no features. If features remain without labels, there exists some minimum index n in the data indices such that $n \notin \bigcup_{k=1}^{K} A_k$, where the union is $\emptyset$ if $K = 0$. It is possible that no features contain n. So we further note that there exists some minimum index m such that $m \notin \bigcup_{j=1}^{K} A_j$ but m is contained in some feature of the allocation. By construction, we must have $m \geq N$. Let K_m be the number of features containing m; K_m is finite by definition of a feature allocation. Let (U_k) denote a sequence of iid uniform random variables, independent of the random feature allocation. Assign $U_{K+1}, \ldots, U_{K+K_m}$ to these new features and determine their order of appearance by the order of these random variables. While features remain to be labeled, continue the recursion with N now equal to m and K now equal to $K + K_m$.

Example 2 (Feature labeling schemes)**.** Consider the feature allocation

$$f_6 = \{\{2, 5, 4\}, \{3, 4\}, \{6, 4\}, \{3\}, \{3\}\}. \tag{1}$$

And consider the random variables

$$U_1, U_2, U_3, U_4, U_5 \overset{iid}{\sim} \text{Unif}[0, 1].$$

We see from f_6 that index 1 has no features. Index 2 has exactly one feature, so we assign this feature, $\{2, 5, 4\}$, to have order-of-appearance label 1. While U_1 is associated with this feature, we do not need to break any ties at this point, so it has no effect.

Index 3 is associated with three features. We associate each feature with exactly one of U_2, U_3, and U_4 (the next three available U_k). For instance, pair $\{3, 4\}$ with U_2, $\{3\}$ with U_3, and the other $\{3\}$ with U_4. Suppose it happens that $U_3 < U_2 < U_4$. Then the feature $\{3\}$ paired with U_3 receives label 2 (the next available order-of-appearance label). The feature $\{3, 4\}$ receives label 3. And the feature $\{3\}$ paired with U_4 receives label 4.

Index 4 has three features, but $\{2, 5, 4\}$ and $\{3, 4\}$ are already labeled. So the only remaining feature, $\{6, 4\}$, receives the next available order-of-appearance

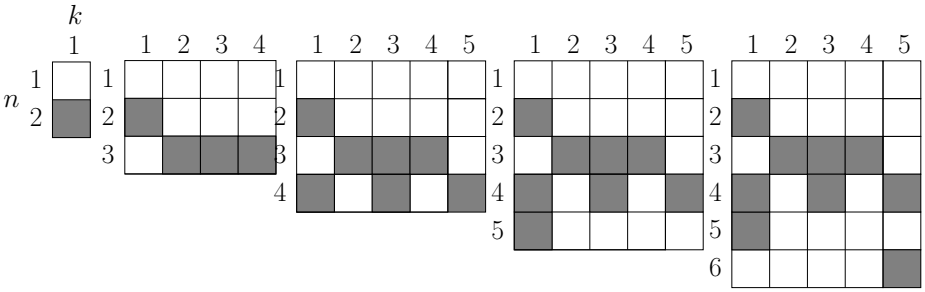

Figure 2: Order-of-appearance binary matrix representations of the sequence of feature allocations on [2], [3], [4], [5], and [6] found by restricting f_6 in Example 2. Rows correspond to indices n, and columns correspond to order-of-appearance feature labels k. A gray square indicates a 1 entry, and a white square indicates a 0 entry. Y_n°, the set of order-of-appearance feature assignments of index n, is easily read off from the matrix as the set of columns with entry in row n equal to 1.

label: 5. U_5 is associated with this feature, but since we do not need to break ties here, it has no effect. Indices 5 and 6 belong to already-labeled features.

So the features can be listed with order-of-appearance indices as

$$A_1 = \{2, 5, 4\}, A_2 = \{3\}, A_3 = \{3, 4\}, A_4 = \{3\}, A_5 = \{6, 4\}. \tag{2}$$

Let Y_n° indicate the set of order-of-appearance feature labels for the features to which index n belongs; i.e., if the features are labeled according to order of appearance as in Eq. (2), then $Y_n^\circ = \{k : n \in A_k\}$. By definition of a feature allocation, Y_n° must have finite cardinality. The order-of-appearance labeling gives $Y_1^\circ = \emptyset, Y_2^\circ = \{1\}, Y_3^\circ = \{2, 3, 4\}, Y_4^\circ = \{1, 3, 5\}, Y_5^\circ = \{1\}, Y_6^\circ = \{5\}$.

Order-of-appearance labeling is well-suited for matrix representations of feature allocations. The rows of the matrix correspond to indices n and the columns correspond to features with order-of-appearance labels k. The matrix representation of the order-of-appearance labeling and resulting feature assignments (Y_n°) for $n \in [6]$ is depicted in Figure 2. ■

Note that when the feature allocation is a partition, there is exactly one feature containing any m, so this scheme reduces to the order-of-appearance scheme for cluster labeling.

Consider an exchangeable feature allocation F_∞. Give order-of-appearance labels to the features of this allocation, and let Y_n° be the set of feature labels for features containing n. So Y_n° is a random finite subset of $\mathbb{N}$. It can be thought of as a simple point process on $\mathbb{N}$; a discussion of measurability of such processes may be found in Kallenberg [2002, p. 178]. Our process is even simpler than a simple point process as it is globally finite rather than merely locally finite.

Note that $(Y_n^\circ)_{n=1}^\infty$ is not necessarily exchangeable. For instance, consider again Example 1. If Y_1° is non-empty, $1 \in Y_1^\circ$ with probability one. If Y_2° is non-empty, with positive probability it may not contain 1. To restore exchangeability

we extend an idea due to Aldous [1985] in the setting of random partitions; in our feature allocation extension, we associate to each feature a draw from a uniform random variable on $[0, 1]$. Drawing these random variables independently we maintain consistency across different values of N. We refer to these random variables as *uniform random feature labels*.

Note that the use of a uniform distribution is for convenience; we simply require that features receive distinct labels with probability one, so any other continuous distribution would suffice. We also note that in a full-fledged model based on random feature allocations these labels often play the role of parameters and are used in defining the likelihood. For further discussion of such constructions, see Broderick et al. [2012b].

Thus, let (ϕ_k) be a sequence of iid uniform random variables, independent of both (U_k) and F_∞. Construct a new feature labeling by taking the feature labeled k in the order-of-appearance labeling and now label it ϕ_k. In this case, let $Y_n^\dagger$ denote the set of feature labels for features to which n belongs. Call this a *uniform random labeling*. $Y_n^\dagger$ can be thought of as a (globally finite) simple point process on $[0, 1]$. Again, we refer the reader to Kallenberg [2002, p. 178] for a discussion of measurability.

Example 3 (Feature labeling schemes (continued))**.** Again consider the feature allocation

$$f_6 = \{\{2, 5, 4\}, \{3, 4\}, \{6, 4\}, \{3\}, \{3\}\}.$$

Now consider the random variables

$$U_1, U_2, U_3, U_4, U_5, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5 \overset{iid}{\sim} \text{Unif}[0, 1].$$

Recall from Example 2 that $U_1, \ldots, U_5$ gave us the order-of-appearance labeling of the features. This labeling allowed us to index the features as in Eq. (2), copied here:

$$A_1 = \{2, 5, 4\}, A_2 = \{3\}, A_3 = \{3, 4\}, A_4 = \{3\}, A_5 = \{6, 4\}.$$

With this order-of-appearance labeling in hand, we can assign a uniform random label to each feature. In particular, we assign the uniform random label ϕ_k to the feature with order-of-appearance label k: $A_1 = \{2, 5, 4\}$ gets label ϕ_1, $A_2 = \{3\}$ gets label ϕ_2, $A_3 = \{3, 4\}$ gets label ϕ_3, $A_4 = \{3\}$ gets label ϕ_4, and $A_5 = \{6, 4\}$ gets label ϕ_5. Let $Y_n^\dagger$ indicate the set of uniform random feature labels for the features to which index n belongs. The uniform random labeling gives

$$Y_1^\dagger = \emptyset, Y_2^\dagger = \{\phi_1\}, Y_3^\dagger = \{\phi_2, \phi_3, \phi_4\}, Y_4^\dagger = \{\phi_1, \phi_3, \phi_5\}, Y_5^\dagger = \{\phi_1\}, Y_6^\dagger = \{\phi_5\}. \tag{3}$$

■

Lemma 4. *Give the features of an exchangeable feature allocation F_∞ uniform random labels, and let $Y_n^\dagger$ be the set of feature labels for features containing n. So $Y_n^\dagger$ is a random finite subset of $[0, 1]$. Then the sequence $(Y_n^\dagger)_{n=1}^\infty$ is exchangeable.*

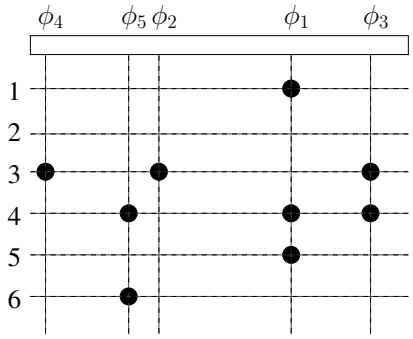

Figure 3: An illustration of the uniform random feature labeling in Example 3. The top rectangle is the unit interval. The uniform random labels are depicted along the interval with vertical dotted lines at their locations. The indices [6] are shown to the left. A black circle shows appears when an index occurs in the feature with a given label. The matrix representations of this feature allocation in Figure 4 can be recovered from this plot.

Proof. Note that $(Y_n^\dagger)_{n=1}^\infty = g((\phi_k)_k, (U_k)_k, F_\infty)$ for some measurable function g. Consider any finite permutation σ that does not change any index n with $n > N$ for some fixed, finite N. Let K represent the (potentially random but finite) number of features in F_N. If we construct order-of-appearance labels using the same $(U_k)_k$ as above and now $\sigma(F_\infty)$ instead of F_∞, the labels will not differ from the original order-of-appearance labels after the first K features. Therefore, there exists some finite permutation τ—which may be a function of $(U_k)_{k=1}^K$, σ, and F_N and hence random—such that $(Y^\dagger_{\sigma(n)})_n = g((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty))$.

Now

$$((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} ((\phi_k)_k, (U_k)_k, \sigma(F_\infty))$$

since the iid sequence $(\phi_k)_k$, the iid sequence $(U_k)_k$, and F_∞ are independent by construction and

$$((\phi_k)_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} ((\phi_k)_k, (U_k)_k, F_\infty)$$

since the feature allocation is exchangeable and the independence used above still holds. So

$$g((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} g((\phi_k)_k, (U_k)_k, F_\infty).$$

It follows that the sequence $(Y_n^\dagger)_n$ is exchangeable. □

We can recover the full feature allocation F_∞ from the sequence $Y_1^\dagger, Y_2^\dagger, \ldots$. In particular, if $\{x_1, x_2, \ldots\}$ are the unique values in $\{Y_1^\dagger, Y_2^\dagger, \ldots\}$, then the features are $\{\{n : x_k \in Y_n^\dagger\} : k = 1, 2, \ldots\}$. The feature allocation can similarly be recovered from the order-of-appearance label collections (Y_n°).

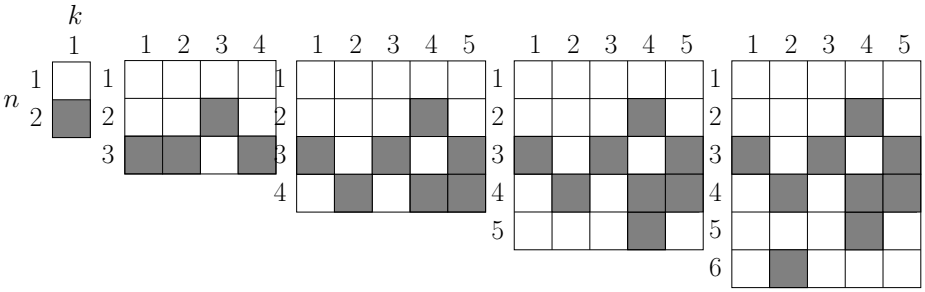

Figure 4: The same consistent sequence of feature allocations in Figure 2 but now with the uniform random order of Example 5 instead of the order of appearance illustrated in Figure 2.

We can also recover a new *random ordered feature allocation* $\tilde{F}_N$ from the sequence $(Y_n^\dagger)$. In particular, $\tilde{F}_N$ is the sequence—rather than the collection—of features $\{n : x_k \in Y_n^\dagger\}$ such that the feature with smallest label ϕ_k occurs first, and so on. This construction achieves our goal of avoiding the combinatorial factors needed to work with the distribution of F_N, while retaining exchangeability and consistency.

Example 5 (Feature labeling schemes (continued))**.** Once more, consider the feature allocation

$$f_6 = \{\{2, 5, 4\}, \{3, 4\}, \{6, 4\}, \{3\}, \{3\}\}.$$

and the uniform random labeling in Eq. (3). If it happens that $\phi_4 < \phi_5 < \phi_2 < \phi_1 < \phi_3$, then the random ordered feature allocation is

$$\tilde{f}_6 = (\{3\}, \{6, 4\}, \{3\}, \{2, 5, 4\}, \{3, 4\}).$$

■

Recall that we were motivated by Example 1 to produce such a random ordering scheme to avoid obfuscating combinatorial factors in the probability of a feature allocation. From another perspective, these factors arise because the random labeling is in some sense more natural than alternative labelings; again, consider random labels as iid parameters for each feature. While order-of-appearance labeling is common due to its pleasant aesthetic representation in matrix form (compare Figures 2 and 4), one must be careful to remember that the order-of-appearance label sets (Y_n°) are not exchangeable. We will use random labeling extensively below since, among other nice properties, it preserves exchangeability of the sets of feature labels associated with the indices.

4 Exchangeable feature probability function

In general, given a probability of a random feature allocation, $\mathbb{P}(F_N = f_N)$, we can find the probability of a random ordered feature allocation $\mathbb{P}(\tilde{F}_N = \tilde{f}_N)$ as

follows. Let H be the number of distinct features of F_N, and let $(\tilde{K}_1, \ldots, \tilde{K}_H)$ be the multiplicities of these distinct features in decreasing order. Then

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = \binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H}^{-1} \mathbb{P}(F_N = f_N), \tag{4}$$

where

$$\binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H} := \frac{K!}{\tilde{K}_1! \cdots \tilde{K}_H!}.$$

For partitions, the effect of this multiplicative factor is the same across all partitions with the same number of clusters; for some number of clusters K, it is just $1/K!$. In the general feature case, the multiplicative factor may be different for different feature configurations with the same number of features.

Example 6 (A Bernoulli, two-feature allocation (continued))**.** Consider F_N constructed as in Example 1. Denote the sizes of the two features by $M_{N,1}$ and $M_{N,2}$. Then

$$\begin{aligned}\mathbb{P}(\tilde{F}_N = \tilde{f}_N) &= \frac{1}{2} q_A^{M_{N,1}} (1 - q_A)^{N - M_{N,1}} q_B^{M_{N,2}} (1 - q_B)^{N - M_{N,2}} \\ &\quad + \frac{1}{2} q_A^{M_{N,2}} (1 - q_A)^{N - M_{N,2}} q_B^{M_{N,1}} (1 - q_B)^{N - M_{N,1}} \\ &= p(N, M_{N,1}, M_{N,2}). \end{aligned} \tag{5}$$

Here, p is some function of the number of indices N and the feature sizes $(M_{N,1}, M_{N,2})$ that we note is symmetric in $(M_{N,1}, M_{N,2})$; i.e., $p(N, M_{N,1}, M_{N,2}) = p(N, M_{N,2}, M_{N,1})$. ■

When the feature allocation probability admits the representation

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = p(N, |A_1|, \ldots, |A_K|) \tag{6}$$

for every ordered feature allocation $\tilde{f}_N = (A_1, \ldots, A_K)$ and some function p that is symmetric in all arguments after the first, we call p the *exchangeable feature probability function* (EFPF). We take care to note that the exchangeable partition probability function (EPPF), which always exists for partitions, is not a special case of the EFPF. Indeed, the EPPF assigns zero probability to any multiset in which an index occurs in more than one feature of the multiset; e.g., $\{\{1\}, \{2\}\}$ is a valid partition and a valid feature allocation of $[2]$, but $\{\{1\}, \{1\}\}$ is a valid feature allocation but not a valid partition of $[2]$. Thus, the EPPF must examine the feature indices of a feature allocation to judge their exclusivity and thereby assign a probability. By contrast, the indices in the multiset provide no such information to the EFPF; only the sizes of the multiset features are relevant in the EFPF case.

Proposition 7. *The class of exchangeable feature allocations with EFPFs is a strict but non-empty subclass of the class of exchangeable feature allocations.*

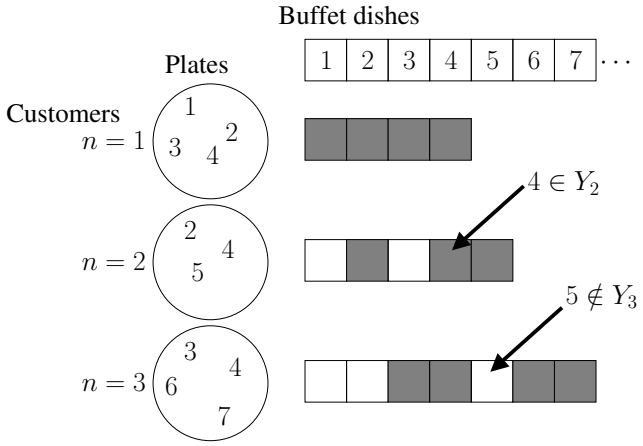

Figure 5: Illustration of an Indian buffet process in the order-of-appearance representation of Figure 2. The buffet (*top*) consists of a vector of dishes, corresponding to features. Each customer—corresponding to a data point—who enters the restaurant first decides whether or not to choose dishes that the other customers have already sampled. The customer then selects a random number of new dishes, not previously sampled by any customer. A gray box in position (n, k) indicates customer n has sampled dish k, and a white box indicates the customer has not sampled the dish. In the example, the second customer has sampled exactly those dishes indexed by 2, 4, and 5: $Y_2^\circ = \{2, 4, 5\}$.

Proof. Example 8 below shows that the class of feature allocations with EFPFs is non-empty, and Example 9 below establishes that there exist simple exchangeable feature allocations without EFPFs. □

Example 8 (Three-parameter Indian buffet process)**.** The Indian buffet process (IBP) [Griffiths and Ghahramani, 2006] is a generative model for a random feature allocation that is specified recursively in a manner akin to the Chinese restaurant process [Aldous, 1985] in the case of partitions. The metaphor involves a set of "customers" that enter a restaurant and sample a set of "dishes." Order the customers by placing them in one-to-one correspondence with the indices $n \in \mathbb{N}$. The dishes in the restaurant correspond to feature labels. Customers in the Indian buffet can sample any non-negative integer number of dishes. The set of dishes chosen by a customer n is just Y_n°, the collection of feature labels for the features to which n belongs, and the procedure described below provides a way to construct Y_n° recursively.

We describe an extended version [Teh and Görür, 2009, Broderick et al., 2012a] of the Indian buffet that includes two extra parameters beyond the single *mass parameter* γ ($\gamma > 0$) originally specified by Griffiths and Ghahramani [2006]; in particular, we include a *concentration parameter* θ ($\theta > 0$) and a *discount parameter* α ($\alpha \in [0, 1)$). We abbreviate this three-parameter IBP

as "3IBP." The single-parameter IBP may be recovered by setting $\theta = 1$ and $\alpha = 0$.

We start with a single customer, who enters the buffet and chooses $K_1^+ \sim$ Poisson(γ) dishes. None of the dishes have been sampled by any other customers since no other customers have yet entered the restaurant. An order-of-appearance labeling gives the dishes labels $1, \ldots, K_1^+$ if $K_1^+ > 0$.

Recursively, the nth customer chooses which dishes to sample in two phases. First, for each dish k that has previously been sampled by any customer in $1, \ldots, n-1$, customer n samples dish k with probability

$$\frac{M_{n-1,k} - \alpha}{\theta + n - 1},$$

for $M_{n,k}$ equal to the number of customers indexed $1, \ldots, n$ who have tried dish k. As each dish represents a feature, sampling a dish represents that the customer index n belongs to that feature. And $M_{n,k}$ is the size of the feature labeled k in the feature allocation of $[n]$.

Next, customer n chooses

$$K_n^+ \sim \text{Poisson}\left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right)$$

new dishes to try. If $K_n^+ > 0$, then the dishes receive unique order-of-appearance labels $K_{n-1}+1, \ldots, K_n$. Here, K_n represents the number of sampled dishes after n customers: $K_n = K_{n-1} + K_n^+$ (with base case $K_0 = 0$).

With this generative model in hand, we can find the probability of a particular feature allocation. We discover its form by enumeration. At each round n, we have a Poisson number of new features, K_n^+, represented. The probability factor associated with these choices is a product of Poisson densities:

$$\prod_{n=1}^{N} \frac{1}{K_n^+!} [C(n, \gamma, \theta, \alpha)]^{K_n^+} \exp\left(-C(n, \gamma, \theta, \alpha)\right),$$

where

$$C(n, \gamma, \theta, \alpha) := \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}.$$

Let R_k be the round on which the kth dish, in order of appearance, is first chosen. Then the denominators for future dish choice probabilities are the factors in the product $(\theta + R_k) \cdot (\theta + R_k + 1) \cdots (\theta + N - 1)$. The numerators for the times when the dish is chosen are the factors in the product $(1-\alpha) \cdot (2-\alpha) \cdots (M_{N,k} - 1 - \alpha)$. The numerators for the times when the dish is not chosen yield $(\theta + R_k - 1 + \alpha) \cdots (\theta + N - 1 - M_{N,k} + \alpha)$. Let $A_{n,k}$ represent the collection of indices in the feature with label k after n customers have entered the restaurant. Then $M_{n,k} = |A_{n,k}|$.

Finally, let $\tilde{K}_1, \ldots, \tilde{K}_H$ be the multiplicities of distinct features formed by this model. We note that there are

$$\left[\prod_{n=1}^{N} K_n^+!\right] / \left[\prod_{h=1}^{H} \tilde{K}_h!\right]$$

rearrangements of the features generated by this process that all yield the same feature allocation. Since they all have the same generating probability, we simply multiply by this factor to find the feature allocation probability.

Multiplying all factors together[1] and taking $f_n = \{A_{N,1}, \ldots, A_{N,K_N}\}$ yields

$$\begin{aligned}
&\mathbb{P}(F_N = f_N) \\
&= \left(\prod_{h=1}^{H} \tilde{K}_h!\right)^{-1} \left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}\right)^{K_N} \exp\left(-\sum_{n=1}^{N} \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right) \\
&\quad \cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(M_{N,k}-\alpha)}{\Gamma(1-\alpha)} \cdot \frac{\Gamma(\theta+N-M_{N,k}+\alpha)}{\Gamma(\theta+N)}\right].
\end{aligned}$$

It follows from Eq. (4) that the probability of a uniform random ordering of the feature allocation is

$$\begin{aligned}
&\mathbb{P}(\tilde{F}_N = \tilde{f}_N) \\
&= \frac{1}{K_N!} \left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}\right)^{K_N} \exp\left(-\sum_{n=1}^{N} \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right) \\
&\quad \cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(M_{N,k}-\alpha)}{\Gamma(1-\alpha)} \cdot \frac{\Gamma(\theta+N-M_{N,k}+\alpha)}{\Gamma(\theta+N)}\right].
\end{aligned} \tag{7}$$

The distribution of $\tilde{F}_N$ has no dependence on the ordering of the indices in $[N]$. Hence, the distribution of F_N depends only on the same quantities—the number of indices and the feature sizes—and the feature multiplicities. So we see that the 3IBP construction yields an exchangeable random feature allocation. Consistency follows from the recursive construction and exchangeability. Therefore, Eq. (7) is seen to be in EFPF form given by Eq. (6). ■

The three-parameter Indian buffet process has an EFPF representation, but the following simple model does not.

Example 9 (A general two-feature allocation)**.** We here describe an exchangeable, consistent random feature allocation whose (randomly ordered) distribution does not depend only on the number of indices N and the sizes of the features of the allocation.

Let $p_{10}, p_{01}, p_{11}, p_{00}$ be fixed frequencies that sum to one. Let Y_n represent the collection of features to which index n belongs. For $n \in \{1, 2\}$, choose Y_n

[1] Readers curious about how the R_k terms disappear may observe that

$$\prod_{k=1}^{K_N} \frac{\Gamma(\theta+R_k)}{\Gamma(\theta+R_k+\alpha-1)} = \prod_{n=1}^{N} \left(\frac{\Gamma(\theta+n)}{\Gamma(\theta+n+\alpha-1)}\right)^{K_N^+}.$$

independently and identically according to:

$$Y_n = \begin{cases} \{1\} & \text{with probability } p_{10} \\ \{2\} & \text{with probability } p_{01} \\ \{1,2\} & \text{with probability } p_{11} \\ \emptyset & \text{with probability } p_{00}. \end{cases}$$

We form a feature allocation from these labels as follows. For each label (1 or 2), collect those indices n with the given label appearing in Y_n to form a feature.

Now consider two possible outcome feature allocations: $f_2 = \{\{2\}, \{2\}\}$, and $f_2' = \{\{1\}, \{2\}\}$. The probability of any ordering $\tilde{f}_2$ of f_2 under this model is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) = p_{10}^0 \, p_{01}^0 \, p_{11}^1 \, p_{00}^1.$$

To see this result, note the distinction between indices $\{1,2\}$ and the feature labels $\{1,2\}$ used in an intermediate step above. Likewise, the probability of any ordering $\tilde{f}_2'$ of f_2' is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}_2') = p_{10}^1 \, p_{01}^1 \, p_{11}^0 \, p_{00}^0.$$

It follows from these two probabilities that we can choose values of $p_{10}, p_{01}, p_{11}, p_{00}$ such that $\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) \neq \mathbb{P}(\tilde{F}_2 = \tilde{f}_2')$. But $\tilde{f}_2$ and $\tilde{f}_2'$ have the same feature counts and N value ($N = 2$). So there can be no such symmetric function p, as in Eq. (5), for this model. ■

5 The Kingman paintbox and feature paintbox

Since the class of exchangeable feature models with EFPFs is a strict subclass of the class of exchangeable feature models, it remains to find a characterization of the latter class. Noting that the sequence of feature collections $Y_n^\dagger$ is an exchangeable sequence when the uniform random labeling of features is used, we might turn to the de Finetti mixing measure of this exchangeable sequence for such a characterization.

Indeed, in the partition case, the Kingman paintbox [Kingman, 1978, Aldous, 1985] provides just such a characterization.

Theorem 10 (Kingman paintbox)**.** *Let $\Pi_\infty := (\Pi_n)_{n=1}^\infty$ be an exchangeable random partition of $\mathbb{N}$, and let $(M_{n,k}^\downarrow, k \geq 1)$ be the decreasing rearrangement of cluster sizes of Π_n with $M_{n,k}^\downarrow = 0$ if Π_n has fewer than k clusters. Then $M_{n,k}^\downarrow/n$ has an almost sure limit $\rho_k^\downarrow$ as $n \to \infty$ for each k. Moreover, the conditional distribution of Π_∞ given $(\rho_k^\downarrow, k \geq 1)$ is as if Π_∞ were generated by random sampling from a random distribution with ranked atoms $(\rho_k^\downarrow, k \geq 1)$.*

When the partition clusters are labeled with uniform random labels rather than by the ranking in the statement of the theorem above, Kingman's paintbox provides the de Finetti mixing measure for the sequence of partition labels

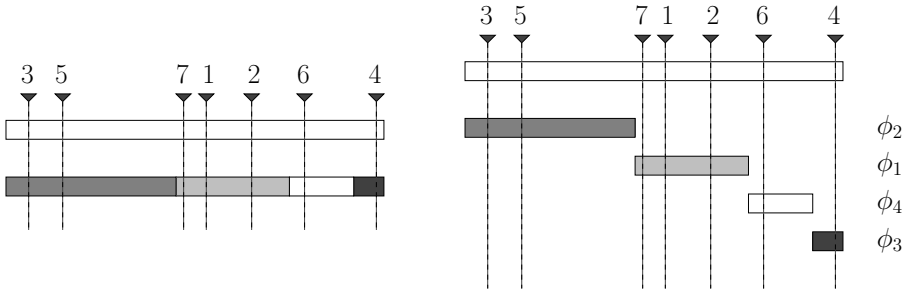

Figure 6: *Left*: An example Kingman paintbox. The upper rectangle represents the unit interval. The lower rectangles represent a partition of the unit interval into four subintervals corresponding to four clusters. The horizontal locations of the seven vertical lines represent seven uniform random draws from the unit interval. The resulting partition of [7] is $\{\{3,5\},\{7,1,2\},\{6\},\{4\}\}$. *Right*: An alternate representation of the same Kingman paintbox, now with each subinterval separated out into its own vertical level. To the right of each cluster subinterval is a uniform random label (with index determined by order of appearance) for the cluster.

of each index n. Two representations of an example Kingman paintbox are illustrated in Figure 6. The Kingman paintbox is so named since we imagine each subinterval of the unit interval as containing paint of a certain color; the colors have a one-to-one mapping with the uniform random cluster labels. A random draw from the unit interval is painted with the color of the Kingman paintbox subinterval into which it falls. While Figure 6 depicts just four subintervals and hence at most four clusters, the Kingman paintbox may in general have a countable number of subintervals and hence clusters. Moreover, these subintervals may themselves be random.

Note that the ranked atoms need not sum to one; in general, $\sum_k \rho_k^{\downarrow} \leq 1$. When random sampling from the Kingman paintbox does not select some atom k with $\rho_k^{\downarrow} > 0$, a new cluster is formed but it is necessarily never selected again for another index. In particular, then, a corollary of the Kingman paintbox theorem is that there are two types of clusters: those with unbounded size as the number of indices N grows to infinity and those with exactly one member as N grows to infinity; the latter are sometimes referred to as *singletons* or collectively as *Kingman dust*. In the feature case, we impose one further regularity condition that essentially rules out dust. Consider any feature allocation F_∞. Recall that we use the notation $Y_n^\dagger$ to indicate the set of features to which index n belongs. We assume that, for each n, with probability one there exists some m with $m \neq n$ such that $Y_m^\dagger = Y_n^\dagger$. Equivalently, with probability one there is no index with a unique feature collection. We call a random feature allocation that obeys this condition a *regular feature allocation*.

We can prove the following theorem for the feature case, analogous to the Kingman paintbox construction for partitions.

Theorem 11 (Feature paintbox)**.** *Let $F_\infty := (F_n)$ be an exchangeable, consistent, regular random feature allocation of $\mathbb{N}$. There exists a random sequence $(C_k)_{k=1}^\infty$ such that C_k is a countable union of subintervals of $[0,1]$ (and may be empty) and such that F_∞ has the same distribution as F'_∞ where F'_∞ is generated as follows. Randomly sample $(U'_n)_n$ iid uniform in $[0,1]$. Let $Y_n := \{k : U'_n \in C_k\}$ represent a collection of feature labels for index n, and let F'_∞ be the induced feature allocation from these label collections.*

Proof. Given F_∞ as in the theorem statement, we can construct $(Y_n^\dagger)_{n=1}^\infty$ as in Lemma 4. Then, according to Lemma 4, $(Y_n^\dagger)_{n=1}^\infty$ is an exchangeable sequence. Note that $Y_n^\dagger$ defines a partition: $n \sim m$ (i.e., n and m belong to the same cluster of the partition) if and only if $Y_n^\dagger = Y_m^\dagger$. This partition is exchangeable since the feature allocation is. Moreover, since we assume there are no singletons in the induced partition (by regularity), the Kingman paintbox theorem implies that the Kingman paintbox atoms sum to one.

By de Finetti's theorem [Aldous, 1985], there exists α such that α is the directing random measure for $(Y_n^\dagger)$. Condition on $\alpha = \mu$. Write $\mu = \sum_{j=1}^\infty q_j \delta_{x_j}$, where the q_j satisfy $q_j \in (0,1]$ and are written in monotone decreasing order: $q_1 \geq q_2 \geq \cdots$. The condition that the atoms of the paintbox sum to one translates to $\sum_{j=1}^\infty q_j = 1$. The (x_j) are the (countable) unique values of $Y_n^\dagger$, ordered to agree with the q_j. The strong law of large numbers yields

$$N^{-1}\#\{n : n \leq N, Y_n^\dagger = x_j\} \to q_j, \quad N \to \infty.$$

Since $\sum_{j=1}^\infty q_j = 1$, we can partition the unit interval into subintervals of length q_j. The jth such subinterval starts at $s_j := \sum_{l=1}^{j-1} q_l$ and ends at $e_j := s_{j+1}$. For $k = 1, 2, \ldots$, define $C_k := \bigcup_{j:\phi_k \in x_j} [s_j, e_j)$. We call the $(C_k)_{k=1}^\infty$ the *feature paintbox*.

Then F_∞ has the same distribution as the following construction. Let $(U'_1, U'_2, \ldots)$ be an iid sequence of uniform random variables. For each n, define $Y_n = \{k : U'_n \in C_k\}$ to be the collection of features, now labeled by positive integers, to which n belongs. Let F'_∞ be the feature allocation induced by the (Y_n). □

A point to note about this feature paintbox construction is that the ordering of the feature paintbox subsets C_k in the proof is given by the order of appearance of features in the original feature allocation F_∞. This ordering stands in contrast to the ordering of atoms by size in the Kingman paintbox. Making use of such a size-ordering would be more difficult in the feature case due to the non-trivial intersections of feature subsets. A particularly important implication is that the conditional distribution of F_∞ given $(C_k)_k$ is not the same as that of F'_∞ given $(C_k)_k$ (cf. Pitman [1995] for similar ordering issues in the partition case).

An example feature paintbox is illustrated in Figure 7. Again, we may think of each feature paintbox subset as containing paint of a certain color (where these colors have a one-to-one mapping with the uniform random labels). Draws from

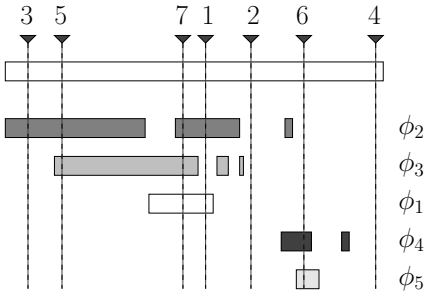

Figure 7: An example feature paintbox. The top rectangle represents the unit interval. Each vertical level below the top rectangle represents a subset of the unit interval corresponding to a feature. To the right of each subset is a uniform random label for the feature. For example, using the notation of Theorem 11, the topmost subset is C_2 corresponding to feature label ϕ_2. The vertical dashed lines represent uniform random draws; i.e., U'_n for index n. The resulting feature allocation of [7] for this realization of the construction is $\{\{3,5,7,1\},\{5,7\},\{7,1\},\{6\},\{6\}\}$. The collection of feature labels for index 7 is $Y_7 = \{\phi_2, \phi_3, \phi_1\}$. The collection of feature labels for index 4 is $Y_4 = \emptyset$.

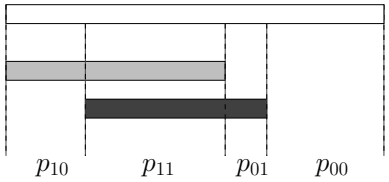

Figure 8: A feature paintbox for the two-feature allocation in Example 9. The top rectangle is the unit interval. The middle rectangle is the feature paintbox subset for feature 1. The lower rectangle is the feature paintbox subset for feature 2.

the unit interval to determine the feature allocation may now be painted with some subset of these colors rather than just a single color.

Next, we revisit earlier examples to find their feature paintbox representations.

Example 12 (A general two-feature allocation (continued))**.** The feature paintbox for the random feature allocation in Example 9 consists of two features. The total measure of the paintbox subset for feature 1 is $p_{10} + p_{11}$. The total measure of the paintbox subset for feature 2 is $p_{01} + p_{11}$. The total measure of the intersection of these two subsets is p_{11}. A depiction of this paintbox appears in Figure 8. ■

Example 13 (Three-parameter Indian buffet process (continued))**.** The 3IBP turns out to be an instance of a general class of exchangeable feature models that we refer to as *feature frequency models*. This class of models not only provides

a straightforward way to construct feature paintbox representations in general, but also plays a key role in our general theory, providing a link between feature paintboxes and EFPFs. In the following section, we define feature frequency models, develop the general construction of paintboxes from feature frequency models, and then return to the construction of the feature paintbox for the 3IBP as an example. We subsequently turn to the general theoretical characterization of feature frequency models. ■

6 Feature frequency models

We now discuss a general class of exchangeable feature models for which it is straightforward to describe the feature paintbox. Let (V_k) be a sequence of (not necessarily independent) random variables with values in $[0, 1]$ such that $\sum_{k=1}^{\infty} V_k < \infty$ almost surely. Let $\phi_k \overset{iid}{\sim} \text{Unif}[0, 1]$ and independent of the (V_k). A *feature frequency model* is built around a random measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$. We may draw a feature allocation given B as follows. For each data point n, independently draw its features like so: for each feature indexed by k, independently make a Bernoulli draw with success probability V_k. If the draw is a success, n belongs to the feature indexed by k (i.e., the feature with label ϕ_k). If the draw is a failure, n does not belong to the feature indexed by k. The feature allocation is induced in the usual way from these labels.

The condition that the frequencies have an almost surely finite sum guarantees, by the Borel-Cantelli lemma, that the number of features exhibited by any index n is almost surely finite, as required in the definition of a feature allocation. We obtain exchangeable feature allocations simply by virtue of the fact that the feature allocations are independently and identically distributed given B. The Bernoulli draws from the feature frequencies guarantee that the feature allocation is regular.

Before constructing the feature paintbox for such a model, we note that V_k is the total length of the paintbox subset for the feature indexed by k. In this sense, it is the frequency of this feature (hence the name "feature frequency model"). And ϕ_k is the uniform random feature label for the feature with frequency V_k. Finally, to achieve the independent Bernoulli draws across k required by the feature allocation specification, we need for the intersection of any two paintbox subsets to have length equal to the product of the two paintbox subset lengths. This desideratum can be achieved with a recursive construction.

First, divide the unit interval into one subset (call it I_1) of length V_1 and another subset (call it I_0) of length $1 - V_1$. Then I_1 is the paintbox subset for the feature indexed by 1. Recursively, suppose we have paintbox subsets for features indexed 1 to $K - 1$. Let e be a binary string of length $K - 1$. Suppose that I_e is the intersection of (a) all paintbox subsets for features indexed by k $(k < K)$ where the kth digit of e is 1 and (b) all paintbox subset complements for features indexed by k $(k < K)$ where the kth digit of e is 0. For every e, we construct $I_{(e,1)}$ to be a subset of I_e with total length equal to V_K times the length of I_e. We construct $I_{(e,0)}$ to be $I_e \backslash I_{(e,1)}$.

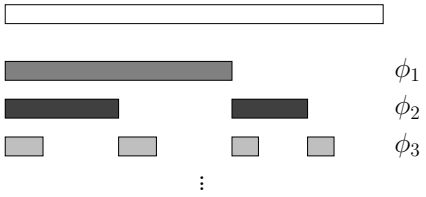

Figure 9: An example feature paintbox for a feature frequency model (Section 6). One such model is the 3IBP (Example 14).

Finally, the paintbox subset for the feature indexed by K is the union of all $I_{e'}$ with e' a binary string of length K such that the final digit of e' is 1. An example of such a paintbox is illustrated in Figure 9.

Example 14 (Three-parameter Indian buffet process (continued))**.** We show that the three-parameter Indian buffet process is an example of a feature frequency model, and thus its feature paintbox can be constructed according to the general recipe that we have just presented.

The underlying random measure for the three-parameter Indian buffet process is known as the *three-parameter beta process* [Teh and Görür, 2009, Broderick et al., 2012a]. This random measure, denoted B, can be constructed explicitly via the following recursion (with $K_0 = 0$ and $n = 1, 2, \ldots$), which extends the results of Thibaux and Jordan [2007]:

$$\begin{aligned}
K_n^+ &\sim \text{Poisson}\left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right), \\
K_n &= K_{n-1} + K_n^+ \\
V_k &\sim \text{Beta}(1-\alpha, \theta+n-1+\alpha), \quad k = K_{n-1}+1, \ldots, K_n \\
\phi_k &\sim \text{Unif}[0,1] \\
B &= \sum_{k=1}^{\infty} V_k \delta_{\phi_k},
\end{aligned}$$

where we recall that the ϕ_k are assumed to be drawn from the uniform distribution for simplicity in this paper, but in general they may be drawn from a continuous distribution that serves as a prior for the parameters defining a likelihood.

Given $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$, the feature allocation is drawn according to the procedure outlined for feature frequency models conditioned on the underlying random measure. Teh and Görür [2009] demonstrate that the distribution of the resulting feature allocation is the same as if it were generated according to a three-parameter Indian buffet process. An alternative proof proceeds as in the two-parameter case covered by Broderick et al. [2012b]. ■

We have seen that the 3IBP can be represented as a feature frequency model. It is straightforward to observe that the two-feature model in Examples 9 and

12 cannot be represented as a feature frequency model unless the intersection of the feature subsets has length p_{11} equal to the product of the feature subset lengths ($p_{10} + p_{11}$ and $p_{01} + p_{11}$); i.e., unless $(p_{10} + p_{11})(p_{01} + p_{11}) = p_{11}$ (cf. Figure 8). Therefore, we have the following result similar to Proposition 7.

Proposition 15. *The class of feature frequency models is a strict but non-empty subclass of the class of exchangeable feature allocations.*

In proving Propositions 15 and 7, we used the 3IBP as an example that belongs to both the class of feature models with EFPFs and the class of feature frequency models. Moreover, in both cases we used two-feature models as an example of exchangeable feature models that do not belong to these subclasses; in particular, we used two-feature models in which the feature combination probabilities $p_{10}, p_{01}, p_{11}, p_{00}$ are not in the necessary proportions. These observations suggest that feature frequency models and EFPFs may be linked. We flesh out the relationship between the two representations in the next few results.

We start with *a priori labeled* features. Recall from Section 3 that an a priori labeled feature allocation is to a feature allocation what a classification is to a clustering; that is, the feature labels are known in advance. The case where we know the feature order in advance is somewhat easier and gives intuition for the type of result we would like in the true feature allocation case. In particular, we prove the results for the case of two a priori labeled features in Theorem 16 and then the case of an unbounded number of a priori labeled features in Theorem 17.

From there, we move on to the (a priori) unlabeled case that is the focus of the paper and prove the equivalence of EFPFs and a slight extension of feature frequency models in Theorem 18.

Theorem 16. *Consider a model with two a priori labeled features: feature 1 and feature 2. If the two features are generated from labeled feature frequencies, the probability of an a priori labeled feature allocation of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2 takes the form $\check{p}(N; M_{N,1}, M_{N,2})$, where we make no symmetry assumptions about $\check{p}$ here and also allow any of $M_{N,1}$ and $M_{N,2}$ to be zero. Conversely, if the probability of any a priori labeled feature allocation can be written as $\check{p}(N; M_{N,1}, M_{N,2})$, then the feature allocation has the same distribution as if it were generated from labeled feature frequencies.*

Proof. Note that throughout this proof we consider the probability of *a particular* labeled feature allocation of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2, as distinct from the probability of all labeled feature allocations of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2. The latter, which is not addressed here, would be the sum over instances of the former. In particular, recalling the matrix representation from Section 3, there are

$$\binom{N}{M_{N,1}}\binom{N}{M_{N,2}}$$

possible $N \times 2$ matrices with $M_{N,1}$ ones in the first column and $M_{N,2}$ ones in the second column.

The reader may feel there is some similarity in this setup to the two-feature allocation of Examples 9 and 12. We note that the quantities $p_{10}, p_{01}, p_{11}, p_{00}$—which retain essentially the same meaning as in Figure 8—may now be random and that their order is pre-specified and non-random.

First, we calculate the probability of a certain labeled feature configuration under this model. Let $M'_{n,10}$ be the number of indices in $[n]$ with feature 1 but not feature 2. Let $M'_{n,01}$ be the number of indices in $[n]$ with feature 2 but not feature 1. Let $M'_{n,00}$ count the indices with neither feature, and let $M'_{n,11}$ count the indices with both features. Then

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \mathbb{E}(p_{10}^{M'_{N,10}} p_{01}^{M'_{N,01}} p_{11}^{M'_{N,11}} p_{00}^{M'_{N,00}}). \tag{8}$$

Denote the total probabilities of features 1 and 2 as, respectively, $q_1 = p_{10} + p_{11}$ and $q_2 = p_{01} + p_{11}$. Suppose that we have a feature frequency model. This assumption implies that

$$p_{10} \overset{a.s.}{=} q_1(1-q_2), \quad p_{01} \overset{a.s.}{=} (1-q_1)q_2, \quad p_{11} \overset{a.s.}{=} q_1 q_2, \quad p_{00} \overset{a.s.}{=} (1-q_1)(1-q_2), \tag{9}$$

where any one of the equalities in Eq. (9) implies the others. It follows that

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \mathbb{E}[q_1^{M_{N,1}}(1-q_1)^{N-M_{N,1}} q_2^{M_{N,2}}(1-q_2)^{N-M_{N,2}}], \tag{10}$$

where $M_{n,1} = M'_{n,10} + M'_{n,11}$ is the total number of indices with feature 1, and likewise $M_{n,2} = M'_{n,01} + M'_{n,11}$ is the total number of indices with feature 2.

So we see that making a feature frequency model assumption yields a feature allocation probability in Eq. (10) that depends only on $N, M_{N,1}, M_{N,2}$. Since we retain the known labeling in this example, the probability is not symmetric in $M_{N,1}$ and $M_{N,2}$.

In the other direction, suppose we know that

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \check{p}(N, M_{N,1}, M_{N,2}) \tag{11}$$

for some function $\check{p}$. Again, we make no symmetry assumptions about $\check{p}$ here, and any of $M_{N,1}$ and $M_{N,2}$ may be zero. Then frequencies $p_{10}, p_{01}, p_{11}, p_{00}$ must exist by the law of large numbers; we note they may be random.

The assumption in Eq. (11) implies that the configurations

$$(M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) = (2,2,0,0)$$
$$(M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) = (0,0,2,2)$$
$$(M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) = (1,1,1,1)$$

have the same probability. That is, by Eq. (8),

$$\mathbb{E}[p_{10}^2 p_{01}^2] = \mathbb{E}[p_{11}^2 p_{00}^2] = \mathbb{E}[p_{10}p_{01}p_{11}p_{00}].$$

It follows that

$$\mathbb{E}[(p_{10}p_{01} - p_{11}p_{00})^2] = \mathbb{E}[p_{10}^2 p_{01}^2 + p_{11}^2 p_{00}^2 - 2p_{10}p_{01}p_{11}p_{00}] = 0.$$

So it must be that $p_{10}p_{01} \overset{a.s.}{=} p_{11}p_{00}$. Recall that this condition is familiar from Example 9.

Adding $p_{10}p_{11}$ to both sides of the almost sure equality and then further adding $p_{11}(p_{01} + p_{11})$ to both sides yields

$$(p_{10} + p_{11})(p_{01} + p_{11}) \overset{a.s.}{=} p_{11}(p_{10} + p_{01} + p_{11} + p_{00}),$$

which reduces to

$$q_1 q_2 \overset{a.s.}{=} p_{11}$$

from the definitions of q_1 and q_2 and from the fact that $p_{10}+p_{01}+p_{11}+p_{00} = 1$.

By Eq. (9) and surrounding text, we see that Eq. (11) implies our model is a feature frequency model. Thus, the equivalence between models with a priori labeled EFPFs and a priori labeled feature frequency models in the case of two features results from simple algebraic manipulations. □

Extending the argument above becomes more tedious when more than two features are involved. In the case of multiple, or even countably many, labeled features, a more elegant proof exists.

Theorem 17. *Consider a model with features a priori labeled* $1, 2, 3, \ldots$. *If the features are generated from labeled feature frequencies, the probability of an a priori labeled feature allocation of* $[N]$ *with* K *or fewer features and* $M_{N,k}$ *occurrences of feature* k *for* $k \in \{1, \ldots, K\}$ *takes the form* $\check{p}(N; M_{N,1}, \ldots, M_{N,K})$, *where we make no symmetry assumptions about* $\check{p}$ *here and note that any of* $M_{N,1}, \ldots, M_{N,K}$ *may be zero. Call* $\check{p}$ *a* labeled EFPF. *Conversely, if the probability of any a priori labeled feature allocation can be written as* $\check{p}(N; M_{N,1}, \ldots, M_{N,K})$, *then the feature allocation has the same distribution as if it were generated from labeled feature frequencies.*

Proof. First, consider the claim that every labeled feature frequency model has a labeled EFPF. This claim is intuitively clear since the independent Bernoulli draws at each atom of the (potentially random) measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$ result in a probability that depends only on the number of occurrences of the corresponding feature and not any interactions between features.

To show this direction formally, we consider a fixed, labeled feature allocation $\hat{f}_N = (A_{N,1}, A_{N,2}, \ldots, A_{N,K})$ with $M_{N,k} := |A_{N,k}|$ and note that

$$\begin{aligned}
&\mathbb{P}(\hat{F}_N = \hat{f}_N) \\
&= \mathbb{E}\left[\mathbb{P}(\hat{F}_N = \hat{f}_N | B)\right] \\
&= \mathbb{E}\left[\left(\prod_{k=1}^{K} V_k^{M_{N,k}}(1 - V_k)^{N - M_{N,k}}\right) \cdot \left(\prod_{k=K+1}^{\infty} (1 - V_k)^N\right)\right].
\end{aligned}$$

It follows that $\mathbb{P}(\hat{F}_N = \hat{f}_N)$ has $\check{p}$ form.

Now consider the other direction. We start with a labeled feature allocation F_∞. In this case, we know that for every labeled feature allocation of $[N]$,

$$\hat{f}_N = (A_{N,1}, \ldots, A_{N,K}),$$

we have that a function $\check{p}$ exists in the form

$$\mathbb{P}(\hat{F}_N = \hat{f}_N) = \check{p}(N, |A_{N,1}|, \ldots, |A_{N,K}|), \tag{12}$$

with no additional symmetry assumptions for $\check{p}$ and where the block sizes $M_{N,k} = |A_{N,k}|$ may be zero.

Let $Z_{n,k}$ be one if n belongs to the kth feature (i.e., $n \in A_{N,k}$) or zero otherwise. Let $b_1, \ldots, b_k$ be values in $\{0, 1\}$. Our goal is to show that conditional on some (as yet unknown) labeled feature frequencies, the probability of feature presence factorizes as independent Bernoulli draws:

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | V_1, \ldots, V_K) = \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1-b_k}. \tag{13}$$

By the assumption on $\check{p}$, the labeled feature sizes $M_{N,1}, \ldots, M_{N,K}$ are sufficient for the distribution of the labeled feature allocation. Let ξ_N be the sigma-field of events invariant under permutations of the first N indices. We note that $M_{N,1}, \ldots, M_{N,K}$ are measurable with respect to ξ_N and start by considering

$$\begin{aligned} &\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \\ &= \prod_{k=1}^{K} \mathbb{P}(Z_{1,k} = b_k | Z_{1,1} = b_1, \ldots, Z_{1,k-1} = b_{k-1}, \xi_N). \end{aligned} \tag{14}$$

Then since the feature sizes are sufficient for the feature allocation distribution, we have

$$\begin{aligned} &\mathbb{P}(Z_{1,k} = b_k | Z_{1,1} = b_1, \ldots, Z_{1,k-1} = b_{k-1}, \xi_N) \\ &= \mathbb{P}(Z_{1,k} = b_k | \xi_N) \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{P}(Z_{n,k} = b_k | \xi_N) \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Z_{n,k} = b_k\} | \xi_N \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Z_{n,k} = b_k\}. \end{aligned}$$

The last line follows since the sum is measurable in ξ_N. By the strong law of large numbers, the final sum converges almost surely as $N \to \infty$ to some potentially random value in $[0, 1]$; call it V_k if $b_k = 1$. By Eq. (14), then, we have

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \xrightarrow{\text{a.s.}} \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1-b_k}. \tag{15}$$

We next observe that the lefthand side of Eq. (15) is a reverse martingale. (ξ_N) is a reversed filtration since $\xi_N \supseteq \xi_{N+1}$ for all N. Moreover, (1) $\mathbb{P}(Z_{1,1} =$

$b_1, \ldots, Z_{1,K} = b_K | \xi_N)$ is measurable with respect to ξ_N; (2) the same quantity is integrable; and (3) by the tower law,

$$\mathbb{E}\left[\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) | \xi_{N+1}\right] = \mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_{N+1}).$$

Since $\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N)$ is a reverse martingale, we have that

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \xrightarrow{\text{a.s.}} \mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_\infty)$$

for $\xi_\infty = \bigcap_{n=1}^{\infty} \xi_n$ by reverse martingale convergence. Together with Eq. (15), this convergence implies that

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_\infty) = \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1 - b_k},$$

and since the V_k are measurable with respect to ξ_∞, the tower law yields Eq. (13), as was to be shown. □

While illustrative, the two previous results do not directly deal with feature allocations as defined earlier in this paper; namely, they do not show any equivalence between EFPFs and feature frequency models in the case where the features are unlabeled (which is exactly the case where EFPFs are defined). We will show in the unlabeled case that every feature frequency model has an EFPF and that every regular feature allocation with an EFPF is an feature frequency model. In fact, we can consider a general—i.e., not necessarily regular—feature allocation and characterize the EFPF representation in this case.

Theorem 18. *Let λ be a non-negative random variable (which may have some arbitrary joint law with the feature frequencies in a feature frequency model). We can obtain an exchangeable feature allocation by generating a feature allocation from a feature frequency model and then, for each index n, including an independent Poisson(λ)-distributed number of features of the form $\{n\}$ in addition to those features previously generated (which may also include index n). A feature allocation of this type has an EFPF. Conversely, every feature allocation with an EFPF has the same distribution as one generated by this construction for some joint distribution of λ and the feature frequencies.*

Proof. Suppose a feature allocation $\tilde{f}$ is generated as described by the construction in Theorem 18 with (potentially random) measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$ giving the frequencies in the feature frequency model component. We wish to show that the feature allocation has an EFPF. We will make use of the fact that an equivalent way to generate the Poisson component of the feature allocation is to draw Poisson $(N\lambda)$ singletons and then assign each uniformly at random to an index in $[N]$.

Consider $\tilde{f}_N = (A_1, A_2, \ldots, A_K)$. Let $S = \{k : |A_k| = 1\}$ represent the feature indices of the singletons of the feature allocation. These features may have been generated either from the feature frequency model or from the Poisson

component. To find the probability of the feature allocation, we consider each possible association of singletons to one of these components. For any such association, let $\tilde{S}$ represent those singletons assigned to the Poisson component; that is, $\tilde{S} \subseteq S$. Let $\tilde{K} = K - |\tilde{S}|$ represent the number of remaining features, which we denote by

$$(\tilde{A}_1, \ldots, \tilde{A}_{\tilde{K}}).$$

Then the probability of this feature allocation satisfies

$$\begin{aligned}
&\mathbb{P}(\tilde{F}_N = \tilde{f}_N) \\
&= \mathbb{E}\left[\mathbb{P}(\tilde{F}_N = \tilde{f}_N | B, \lambda)\right] \\
&= \mathbb{E}\left[\sum_{\tilde{S}:\tilde{S}\subseteq S} N^{-\tilde{S}} \text{Poisson}\left(\tilde{S}|N\lambda\right) \sum_{\substack{(i_1,\ldots,i_{\tilde{K}}) \\ \text{distinct}}}\right. \\
&\left.\frac{1}{K!}\left(V_{i_1}^{|\tilde{A}_1|}(1-V_{i_1})^{N-|\tilde{A}_1|}\cdots V_{i_{\tilde{K}}}^{|\tilde{A}_{\tilde{K}}|}(1-V_{i_{\tilde{K}}})^{N-|\tilde{A}_{\tilde{K}}|} \prod_{\substack{l\in\mathbb{N} \\ l\notin\{i_1,\ldots,i_{\tilde{K}}\}}} (1-V_l)^N\right)\right].
\end{aligned}$$

The final expression depends only on the number of data points N and feature sizes and is symmetric in the feature sizes. So it has EFPF form.

In the other direction, we sidestep the issue of feature ordering by looking at the number of features to which each data index belongs. The advantage of this approach is that this number does not depend on the feature order. The following result is the key to making use of this observation.

Lemma 19. *Let K_n be a sequence of positive integers. For each n, suppose we have (constants)*

$$1 \geq p_{n,1} \geq p_{n,2} \geq \ldots \geq p_{n,K_n} > 0.$$

And, for completeness, suppose $p_{n,k} = 0$ for $k > K_n$. Let $X_{n,k} \sim Bern(p_{n,k})$, independently across n and k and with $k = 1 : K_n$. Define $\#_n := \sum_{k=1}^{K_n} X_{n,k}$. Then the following are equivalent.

1. *$\#_n \xrightarrow{d} \#$ for some finite-valued random variable $\#$ on $\{0, 1, 2, \ldots\}$.*
2. *There exist (constants) $\{p_k\}_{k=1}^{\infty}$ and λ such that $p_k \in [0, 1]$ and $\lambda > 0$ and further such that, $\forall k = 1, 2, \ldots$,*

$$p_{n,k} \to p_k, \quad n \to \infty \tag{16}$$

and

$$\sum_{k=1}^{K_n} p_{n,k} \to \sum_{k=1}^{\infty} p_k + \lambda, \quad n \to \infty. \tag{17}$$

In this case, we further have

$$1 \geq p_1 \geq p_2 \geq \cdots, \tag{18}$$

and

$$\# \stackrel{d}{=} Y + \sum_{k=1}^{\infty} X_k, \tag{19}$$

where $X_k \sim$ *Bern*(p_k)*, independently across* k*, and* $Y \sim$ *Poisson*(λ)*.*

The proof of Lemma 19 appears in Appendix 2; this lemma is essentially a special case of a more general result in Appendix 1.

In this direction of the proof of Theorem 18, we want to show that if we assume that the probability of a feature allocation takes EFPF form, then the allocation has the same distribution as if it were generated according to a feature frequency model with a Poisson-distributed number of singleton features for each n. To see how Lemma 19 may be useful, we let $\hat{\#}$ be the number of features in which index 1 occurs. Recall that in order to use the EFPF, we apply a uniform random ordering to the features of our feature allocation. Examining $\hat{\#}$ is advantageous since it is invariant to the ordering of the features, and we can thereby avoid complicated considerations that may arise related to the feature ordering and consistency of ordering across feature allocations of increasing index sets.

Indeed, recall that once we have chosen a uniform random ordering for the features, the EFPF assumption tells us that any feature allocation with the requisite feature sizes and number of indices has the same probability. Let K_N be the number of features containing indices $[N]$. If $M_{N,k}$ is the size of the kth feature (under the uniform random ordering) after N indices, then there are

$$\binom{N}{M_{N,1}} \cdots \binom{N}{M_{N,K_N}}$$

such configurations. $M_{N,1}/N$ have index 1 in the first feature. For each such allocation, there are equally many configurations of the remaining features. So, for each such allocation, $M_{N,2}/N$ have index 1 in the second feature. And so on. That is, we have that, conditionally on the feature sizes, the number of features with index 1 has the same distribution as a sum of Bernoulli random variables:

$$\sum_{k=1}^{K_N} \tilde{X}_{N,k}, \quad \tilde{X}_{N,k} \stackrel{indep}{\sim} \text{Bern}(M_{N,k}/N). \tag{20}$$

First, we note that the feature sizes are sufficient for the distribution by the EFPF assumption. So we may, in fact, condition on ξ_N, which we define to be the sigma-field of events invariant under permutations of the indices $n = 1, \ldots, N$. That is, $\hat{\#}|\xi_N$ has the same distribution as the sum in Eq. (20).

Second, we note that the sum in Eq. (20) has no dependence on the ordering of the features. In particular, then, let $1 \geq p_{N,1} \geq p_{N,2} \geq \cdots \geq p_{N,K_N}$ be

the sizes of the features divided by N and ordered so as to be monotonically decreasing. Again, note that we are only considering those features including some data index in $[N]$. It follows that

$$\hat{\#}|\xi_N \stackrel{d}{=} \sum_{k=1}^{K_N} \tilde{X}_{N,k}, \quad \tilde{X}_{N,k} \overset{indep}{\sim} \text{Bern}(p_{N,k}). \tag{21}$$

So we see that we have circumvented ordering concerns and can simply use a size ordering in what follows.

At this point, it seems natural to apply Lemma 19 to $\hat{\#}|\xi_N$. To do so, we need to show that $\hat{\#}|\xi_N$ converges in distribution to some random variable with non-negative integer values as $N \to \infty$. To that end, we note that (ξ_N) is a reversed filtration: $\xi_N \supseteq \xi_{N+1}$ for all N. And further $\mathbb{P}(\hat{\#} = j|\xi_N)$ is a reversed martingale since (1) $\mathbb{P}(\hat{\#} = j|\xi_N)$ is measurable with respect to ξ_N; (2) $\mathbb{P}(\hat{\#} = j|\xi_N)$ is integrable; and (3) by the tower law, $\mathbb{E}[\mathbb{P}(\hat{\#} = j|\xi_N)|\xi_{N+1}] = \mathbb{P}(\hat{\#} = j|\xi_{N+1})$. It follows that

$$\mathbb{P}(\hat{\#} = j|\xi_N) \xrightarrow{\text{a.s.}} \mathbb{P}(\hat{\#} = j|\xi_\infty)$$

and hence

$$\hat{\#}|\xi_N \xrightarrow{d} \hat{\#}|\xi_\infty \quad \text{a.s.}$$

for $\xi_\infty = \bigcap_{n=1}^{\infty} \xi_n$ by reverse martingale convergence.

So we may apply Lemma 19 conditional on ξ_∞. By the lemma, we have that, conditional on ξ_∞,

$$\begin{aligned} \hat{\#} &\stackrel{d}{=} Y + \sum_{k=1}^{\infty} X_k \\ Y &\sim \text{Poisson}(\lambda) \\ X_k &\overset{indep}{\sim} \text{Bern}(p_k) \end{aligned}$$

for some $\lambda \geq 0$ and some $1 \geq p_1 \geq p_2 \geq \cdots$. The conditioning on ξ_∞ means that, in general, λ and the frequencies $1 \geq p_1 \geq p_2 \geq \cdots$ may be positive random variables, as was to be shown. □

7 Conclusion

It has been known for some time that the class of exchangeable partitions is the same as the class of partitions generated by the Kingman paintbox, which is in turn the same as the class of partitions with exchangeable partition probability functions (EPPFs). In this paper, we have developed an analogous set of concepts for the feature allocation problem. We defined a feature allocation as an extension of partitions in which indices may belong to multiple groups, now called features. We have developed analogues of the EPPF and the Kingman paintbox, which we refer to as the exchangeable feature partition function

(EFPF) and the feature paintbox, respectively. The feature paintbox allows us to construct a feature allocation via iid draws from an underlying collection of sets in the unit interval. In the special cases of partitions and feature frequency models the construction of these sets is particularly straightforward.

The Venn diagram presented earlier in Figure 1 summarizes our results and also suggests a number of open areas for further investigation. In particular it would be useful to develop a fuller understanding of the regularity condition on feature allocations that allows the connection to the feature paintbox. It would also be of interest to carry the program further by exploring generalizations of the partition and feature allocation framework to other combinatorial representations, such as the setting in which we allow multiplicity within, as well as across, features [Broderick et al., 2011, Zhou et al., 2012].

Acknowledgements

T. Broderick was funded by a National Science Foundation Graduate Research Fellowship. The work of J. Pitman was supported in part by National Science Foundation Award 0806118. Our work has also been supported by the Office of Naval Research under contract/grant number N00014-11-1-0688.

References

D. Aldous. Exchangeability and related topics. *Ecole d'Eté de Probabilités de Saint-Flour XIII1983*, pages 1–198, 1985.

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *Arxiv preprint arXiv:1111.1802*, 2011.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, 7(2):439–476, 2012a.

T. Broderick, M. I. Jordan, and J. Pitman. Clusters and features from combinatorial stochastic processes. *Statistical Science, to appear. Arxiv preprint arXiv:1206.5862*, 2012b.

S. X. Chen and J. S. Liu. Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892, 1997.

B. De Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6.*, 4:251–299, 1931. in Italian.

F. Doshi, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Clearwater Beach, Florida, USA, 2009.

M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, pages 268–277, 1994.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, 2006.

E. Hewitt and L. J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, November 1955.

Oliver Johnson, Ioannis Kontoyiannis, and Mokshay Madiman. Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound poisson measures. *Discrete Applied Mathematics*, 2011.

O. Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374, 1978.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

J. Pitman. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. ISBN 978-3-540-30990-1; 3-540-30990-X. doi: 10.1007/b11601500. URL `http://bibserver.berkeley.edu/csp/april05/bookcsp.pdf`.

Y.W. Teh and D. Görür. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, 2009.

R. Thibaux and M. Jordan. Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, 2007.

YH Wang. On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312, 1993.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands, 2012.

Appendix 1: Intermediate lemmas leading to Lemma 19

To prove Lemma 19, we will make use of a few definitions and lemmas. We start with two definitions. First, suppose we have constants $p_1, p_2, p_3, \ldots$ such that

$$1 \geq p_1 \geq p_2 \geq p_3 \geq \ldots \geq 0$$

and a constant λ such that $0 \leq \lambda < \infty$. Then we say that the random variable $\#$ has the *extended Poisson-binomial distribution* with parameters $(\lambda, p_1, p_2, \ldots)$ if there exist independent random variables $X_0, X_1, X_2, \ldots$ with

$$\begin{aligned} X_0 &\sim \text{Poisson}(\lambda) \\ X_k &\sim \text{Bern}(p_k), \quad k = 1, 2, \ldots \end{aligned}$$

such that

$$\# = X_0 + \sum_{k=1}^{\infty} X_k.$$

The terminology "extended Poisson-binomial distribution" is motivated by the familiar *Poisson-binomial distribution* [Wang, 1993, Chen and Liu, 1997, Johnson et al., 2011], which describes the special case of the above where $\lambda = 0$ and $p_k = 0$ for all $k > K$ for some finite K.

Second, we say that μ is the *spike size-location measure* with parameters $(\lambda, p_1, p_2, \ldots)$ if μ puts mass λ at 0 and mass p_k at p_k for $k = 1, 2, \ldots$. With these definitions in hand, we can state the following lemmas.

Lemma 20. *Let $\#$ have the extended Poisson-binomial distribution with parameters $(\lambda, p_1, p_2, \ldots)$.*

Then

1. *$\#$ is a.s. finite if and only if $\sum_{k=1}^{\infty} p_k < \infty$.*
2. *If $\#$ is a.s. finite, then the parameters $(\lambda, p_1, p_2, \ldots)$ are uniquely determined by the distribution of $\#$.*

In particular, since the parameters $(\lambda, p_1, p_2, \ldots)$ uniquely determine the distribution of $\#$, Lemma 20 tells us that there is a bijection between the distribution of $\#$ and the parameters $(\lambda, p_1, p_2, \ldots)$ when $\#$ is a.s. finite. See Appendix 3 for the proof of Lemma 20.

The next lemma tells us that this correspondence between distributions and parameters is also continuous in a sense.

Lemma 21. *For $n = 1, 2, \ldots$, let $\#_n$ have the extended Poisson-binomial distribution with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$. Let μ_n be the spike size-location measure with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$.*

Then the following two statements are equivalent:

1. *$\#_n$ converges in distribution to a finite-valued limit random variable.*
2. *μ_n converges weakly to some finite measure on $[0, 1]$.*

If the convergence holds, the limiting random variable (call it $\#$) has an extended Poisson-binomial distribution, and the limiting measure (call it μ) is a spike size-location measure. In this case, $\#$ and μ have the same parameters; call the parameters $(\lambda, p_1, p_2, \ldots)$.

This lemma is suggested by, and provides an extension to, previous results on triangular arrays of random variables with row sums converging in distribution; cf., Kallenberg [2002]. See Appendix 4 for the proof of Lemma 21.

Lemma 19 highlights a special case of Lemmas 20 and 21 that we use to prove the equivalence in Theorem 18.

Appendix 2: Proof of Lemma 19

We can rephrase the statement of Lemma 19 in terms of the terminology introduced in Appendix 1. In particular, we are given a sequence of random variables $\#_n$, where $\#_n$ has an extended Poisson-binomial distribution with parameters

$$(0, p_{n,1}, p_{n,2}, \ldots, p_{n,K_n}, 0, 0, \ldots).$$

Then we see that Lemma 19 is essentially a special case of Lemma 21 where λ_n and all but finitely many of the $p_{n,k}$ are equal to zero; this special case is exactly the usual Poisson-binomial distribution.

(1) $\Rightarrow$ (2). We assume that $\#_n$ converges in distribution to some finite-valued random variable $\#$, and we wish to show that the $p_{n,k}$ converge to some limiting p_k as $n \to \infty$ for each k, and likewise that $\sum_{k=1}^{K_n} p_{n,k}$ converges to $\sum_{k=1}^{\infty} p_k + \lambda$ for some non-negative constant λ. The $p_{n,k}$ are just the ordered atom sizes of the spike size-location measures μ_n in Lemma 21. By Lemma 21, the μ_n converge weakly to some spike size-location measure μ.

Denote the parameters of μ by $(\lambda, p_1, p_2, \ldots)$. The convergence of μ_n to μ yields both the desired convergence of the atom sizes (Eq. (16), repeated here)

$$p_{n,k} \to p_k, \quad n \to \infty$$

and the desired convergence of the total mass of μ_n (Eq. (17), repeated here)

$$\sum_{k=1}^{K_n} p_{n,k} \to \sum_{k=1}^{\infty} p_k + \lambda, \quad n \to \infty.$$

(2) $\Rightarrow$ (1). Now we assume that the $p_{n,k}$ converge to some limiting p_k as $n \to \infty$ for each k, and likewise that $\sum_{k=1}^{K_n} p_{n,k}$ converges to $\sum_{k=1}^{\infty} p_k + \lambda$ for some appropriate positive constants $\{p_k\}, \lambda$. We wish to show that $\#_n$ converges in distribution to some finite-valued random variable $\#$.

The assumed convergences guarantee the weak convergence of the spike size-location measures μ_n to some finite measure on $[0, 1]$. Lemma 21 then guarantees that $\#_n$ converges in distribution to some finite-valued random variable $\#$.

Assume (1) and (2). We wish to show that $1 \geq p_1 \geq p_2 \geq \ldots$ (Eq. (18)), but this result follows from the monotonicity of the $p_{n,k}$.

Eq. (19) in the original lemma statement can be rephrased as wanting to show that # has the extended Poisson-binomial distribution with parameters $(\lambda, p_1, p_2, \ldots)$. This follows directly from the final statement in Lemma 21 and our identification of the limiting spike size-location measure μ as having parameters $(\lambda, p_1, p_2, \ldots)$ in a previous part of this proof ("(1) $\Rightarrow$ (2)"). $\square$

Appendix 3: Proof of Lemma 20

Throughout we assume that # has the extended Poisson-binomial distribution with parameters $(\lambda, p_1, p_2, \ldots)$.

(1). We want to show that # is a.s. finite if and only if $\sum_{k=1}^{\infty} p_k < \infty$. Since # is extended Poisson-binomially distributed, we can write $\# = X_0 + \sum_{k=1}^{\infty} X_k$ for independent $X_0 \sim \text{Poisson}(\lambda)$ and $X_k \sim \text{Bern}(p_k)$ for $k = 1, 2, \ldots$. First suppose $\sum_{k=1}^{\infty} p_k < \infty$. Then $\sum_{k=1}^{\infty} X_k$ is a.s. finite by the Borel-Cantelli lemma. Second, suppose $\sum_{k=1}^{\infty} p_k = \infty$. Then $\sum_{k=1}^{\infty} X_k$ is a.s. infinite by the second Borel-Cantelli lemma. Since X_0 is a.s. finite by construction, the result follows.

(2). We want to show that if # is a.s. finite, then the parameters $(\lambda, p_1, p_2, \ldots)$ are uniquely determined by the distribution of #. To that end, let μ be the spike size-location measure with parameters $(\lambda, p_1, p_2, \ldots)$. Note that μ need not be a probability measure but is finite by the assumption that # is a.s. finite together with part (1) of the lemma.

To better understand the distribution of #, we write the probability generating function of #. For s with $0 \leq s \leq 1$, we have

$$\mathbb{E}s^{\#} = e^{-\lambda(1-s)} \prod_{k=1}^{\infty} \left[1 - (1-s)p_k\right],$$

which implies that for s with $0 < s \leq 1$ we have

$$\begin{aligned} -\log \mathbb{E}s^{\#} &= \lambda(1-s) - \sum_{k=1}^{\infty} \log\left[1 - (1-s)p_k\right] \qquad (22) \\ &= \lambda(1-s) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{1}{j}(1-s)^j p_k^j \end{aligned}$$

from the Taylor series expansion of the logarithm

$$= \lambda(1-s) + \sum_{j=1}^{\infty} \frac{1}{j}(1-s)^j \sum_{k=1}^{\infty} p_k^j$$

interchanging the order of summation since the summands are non-negative

$$= (1-s)\mu\{0\} + \sum_{j=1}^{\infty} \frac{1}{j}(1-s)^j \int_{(0,1]} x^{j-1} \mu(dx) \qquad (23)$$

$$= \sum_{j=1}^{\infty} \frac{1}{j}(1-s)^j m_{j-1}, \tag{24}$$

where

$$m_j := \int_{[0,1]} x^j \mu(dx)$$

is the jth moment of the measure μ.

Now the distribution of # uniquely determines the probability generating function of #, which by Eq. (24) uniquely determines the sequence of moments of the measure μ. In turn, μ is a bounded measure on $[0,1]$ and hence uniquely determined by its moments. And the parameters $(\lambda, p_1, p_2, \ldots)$ are uniquely determined by μ. □

Appendix 4: Proof of Lemma 21

For $n = 1, 2, \ldots$, we assume $\#_n$ has the extended Poisson-binomial distribution with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$. We further assume μ_n has the spike size-location measure with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$.

(2) ⇒ (1). Suppose the μ_n converge weakly to some finite measure μ on $[0,1]$. We want to show that $\#_n$ converges in distribution to a finite-valued limit random variable.

In Appendix 3, we noted that we can express the probability generating function of an extended Poisson-binomial distribution in terms of a spike size-location measure with the same parameters. In particular, by Eq. (23), we can write the negative log of the probability generating function of $\#_n$ as

$$-\log \mathbb{E}s^{\#_n} = \int_{[0,1]} f_s(x)\ \mu_n(dx),$$

where

$$f_s(x) := \sum_{j=1}^{\infty} \frac{1}{j}(1-s)^j x^{j-1} = \begin{cases} -x^{-1}\log[1-(1-s)x] & x > 0 \\ 1-s & x = 0 \end{cases}. \tag{25}$$

Since $f_s(x)$ is bounded in x for each fixed s with $0 < s \leq 1$, we have by the assumption of weak convergence of μ_n that

$$\lim_{n\to\infty} -\log \mathbb{E}s^{\#_n} = \int_{[0,1]} f_s(x)\ \mu(dx).$$

Moreover, since μ is finite by assumption, we have that the result is finite for each s with $0 < s \leq 1$. It follows that $\#_n$ converges in distribution to a finite random variable #, with probability generating function given by

$$\mathbb{E}s^{\#} = \exp\left\{-\int_{[0,1]} f_s(x)\ \mu(dx)\right\}. \tag{26}$$

Assume (1). Now suppose the $\#_n$ converge in distribution to a finite random variable $\#$. The next two parts of the proof will rely on an intermediate step: showing that μ_n has bounded total mass in this case.

To show that μ_n has bounded total mass, first note that $\mathbb{E}\#_n$ is exactly the total mass of μ_n:

$$\mathbb{E}\#_n = \lambda_n + \sum_{k=1}^{\infty} p_{n,k} =: \Sigma_n,$$
$$\text{and } \mathrm{Var}\#_n = \lambda_n + \sum_{k=1}^{\infty} p_{n,k}(1 - p_{n,k}).$$

Noting that $\mathrm{Var}\#_n \leq \Sigma_n$ allows us to apply Chebyshev's inequality to find

$$\begin{aligned} 1/4 &\geq \mathbb{P}(|\#_n - \mathbb{E}\#_n| \geq 2\sqrt{\mathrm{Var}\#_n}) \\ 3/4 &\leq \mathbb{P}(|\#_n - \Sigma_n| \leq 2\sqrt{\mathrm{Var}\#_n}) \\ &\leq \mathbb{P}(|\#_n - \Sigma_n| \leq 2\sqrt{\Sigma_n}) \\ &\leq \mathbb{P}(\#_n \geq \Sigma_n - 2\sqrt{\Sigma_n}). \end{aligned}$$

Since $\#_n$ converges in distribution by assumption, the sequence $\#_n$ is tight. Choose ϵ such that $1/2 > \epsilon > 0$. Then there exists some N_ϵ such that, for all $n \geq 1$, we have $\mathbb{P}(\#_n \leq N_\epsilon) > 1 - \epsilon > 1/2$. It follows that, for all $n \geq 1$,

$$1/4 \leq \mathbb{P}(N_\epsilon \geq \Sigma_n - 2\sqrt{\Sigma_n}).$$

Since Σ_n is non-random, it must be that $\mathbb{P}(N_\epsilon \geq \Sigma_n - 2\sqrt{\Sigma_n}) = 1$. That is, the total mass of μ_n is bounded.

Assume (1) and (2). Suppose $\#_n$ converges in distribution to some finite-valued limit random variable $\#$ and that μ_n converges weakly to some finite measure μ. We want to show that $\#$ has an extended Poisson-binomial distribution, that μ is a spike size-location measure, and that $\#$ and μ have the same parameters.

We start by showing that μ is discrete. Choose any $\epsilon > 0$. Since the mass of μ_n is bounded across n by the previous part of the proof ("Assume (1)"), the number of atoms of μ_n greater than ϵ is bounded across n. It follows that the number of atoms of μ has the same bound. So μ is discrete. Since μ_n converges weakly to μ, we see that μ must have atoms with sizes and locations $p_1, p_2, \ldots$ such that

$$1 \geq p_1 \geq p_2 \geq \ldots$$

as well as a potential atom, with size we denote by λ, at zero. That is, μ is a spike size-location measure with parameters $(\lambda, p_1, p_2, \ldots)$.

In a previous part of the proof ("(2) $\Rightarrow$ (1)"), we expressed the probability generating function of $\#$ as a function of μ (Eq. (26)). With this relation in hand, we can reverse the series of equations presented in Appendix 3 and ending in

Eq. (23) to find the form of the probability generating function for # (Eq. (22)). In particular, Eq. (22) tells us that # is an extended Poisson-binomial random variable with parameters $(\lambda, p_1, p_2, \ldots)$. In particular, we emphasize that # has the same parameters as μ, which we have already shown above is a spike size-location measure.

(1) $\Rightarrow$ (2) Now step back and assume that $\#_n$ converges in distribution to a finite-valued limit random variable; call it #. We wish to show that μ_n converges weakly to some finite measure on $[0, 1]$.

By a previous part of this proof ("Assume (1)"), the mass of μ_n is bounded across n. Moreover, by construction, all of the mass for each μ_n is concentrated on $[0, 1]$. So it must be that the sequence μ_n is tight. It follows that if every weakly convergent subsequence μ_{n_j} has the same limit μ, then μ_n converges weakly to μ.

Consider a subsequence $(n_j)_j$ of $\mathbb{N}$. We know $\#_{n_j}$ converges in distribution to # by the assumption that $\#_n$ converges in distribution to #. The previous part of this proof ("Assume (1) and (2)") gives that the form of the limit of μ_{n_j} is determined by #; namely, the limit is a spike size-location measure with parameters shared by #. In particular, then, the limit μ must be the same for every subsequence, and the desired result is shown. □