[illegible]

[illegible]

wray.buntine@nicta.com.au
NICTA and Australian National University
Locked Bag 8001, Canberra ACT 2601, Australia

[illegible]

marcus.hutter@anu.edu.au
Australian National University and NICTA
RSISE, Daley Road, Canberra ACT 0200, Australia

15 February 2012

Abstract

The two parameter Poisson-Dirichlet Process (PDP), a generalisation of the Dirichlet Process, is increasingly being used for probabilistic modelling in discrete areas such as language technology, bioinformatics, and image analysis. There is a rich literature about the PDP and its derivative distributions such as the Chinese Restaurant Process. This article reviews some of the basic theory and then the major results needed for Bayesian modelling of discrete problems including details of priors, posteriors and computation.

The PDP is a generalisation of the Dirichlet distribution that allows one to build distributions over partitions, both finite and countably infinite. The PDP has two other remarkable properties: first it is partially conjugate to itself, which allows one to build hierarchies of PDPs, and second using a marginalised relative the Chinese Restaurant Process (CRP), one gets fragmentation and clustering properties that lets one layer partitions to build trees. This article presents the basic theory for understanding the notion of partitions and distributions over them, the PDP and the CRP, and the important properties of conjugacy, fragmentation and clustering, as well as some key related properties such as consistency and convergence. This article also presents a Bayesian interpretation of the Poisson-Dirichlet process: it is based on an improper and infinite dimensional Dirichlet distribution. This interpretation requires technicalities of priors, posteriors and Hilbert spaces, but conceptually, this means we can understand the process as just another Dirichlet and thus all its sampling properties emerge naturally.

The theory of PDPs is usually presented for continuous distributions (more generally referred to as non-atomic distributions), however, when applied to discrete distributions its remarkable conjugacy property emerges. This context and basic results are also presented, as well as techniques for computing the second order Stirling numbers that occur in the posteriors for discrete distributions.

Keywords

Pitman-Yor process; Dirichlet; two-parameter Poisson-Dirichlet process; Chinese Restaurant Process; Consistency; (non)atomic distributions; improper prior; hierarchical models; Bayesian interpretation.

1 [illegible]

The *two-parameter Poisson-Dirichlet process* (PDP), also known as the Pitman-Yor process (named so in [IJ01]), is an extension of the *Dirichlet process* (DP). Related is a marginalisation known as the Chinese Restaurant Process (CRP) which gives an elegant analogy of incremental sampling of partitions. These models have proven useful in a number of ways as tools for non-parametric and hierarchical Bayesian modelling, especially in discrete domains such as with language and images where one wants to develop hierarchical models, or be flexible with dimension.

In language domains, PDPs are proving useful for full probability modelling of various phenomena including n-gram modelling and smoothing [Teh06b, GGJ06, MS08], dependency models for grammar [JGG07, WSM08], and for data compression [WAG+09]. The PDP-based n-gram models correspond well to versions of Kneser-Ney smoothing [Teh06b], the state of the art method in applications. These models are intriguing from the probability perspective, as well as sometimes being competitive in terms of performance. More generally, the models are also being used for clustering [GR01, Ras00], and for related tasks such as image segmentation [SJ09], relational modelling [XTYK06], and exemplar-based clustering [TZF08].

PDPs and their associated distributions are basically a tool for modelling two kinds of objects: mixture models and partitions. They can then be used to model trees and other hierarchical structures. Section 2 introduces the interrelated concepts of a simple mixture model and a partition along with some of their statistical complications. The definition of a PDP is then presented in Section 3. The theory is well developed for the more general context of continuous distributions, and that theory is reviewed here. Details of statistics, consistency and the forms of posteriors are given in Section 4. Statistical analysis of partitions is given in Section 5 which arises when one marginalises the posterior of the PDP to obtain the CRP, which can be further marginalised to obtain a distribution on partition sizes. Results on fragmentation and grouping of partitions in Section 6 allow one to extend the CRP distribution to trees (*i.e.*, nested partitions).

A new Bayesian interpretation and definition of the PDP is then given in Section 7. This uses the methodology of improper priors too show that the distribution on the infinite probability vector underlying the PDP is in fact an infinite improper Dirichlet. From this one can readily obtain all the standard sampling and additivity properties for the PDP.

With the use of PDPs increasing in computer science applications, where sophisticated discrete probabilistic modelling is required, this article reviews and summarises the basic theory of PDPs in the discrete context in Section 8. This context has distinct properties where the distribution on partition size, introduced above, is needed to express posteriors using Stirling numbers of the second kind. These play the role of the Gamma function that occurs in the posterior for a Dirichlet distribution. Basic conjugacy results for the PDP are then presented that make it such an important distribution for hierarchical Bayesian reasoning.

We have attempted where possible to follow conventions used in the mathematical statistics community, and to provide some pointers to that literature.

2 [illegible]

Before introducing the PDP, we introduce the basic context of its use, a simple mixture model. The kinds of mixture models we consider require as input a *base* probability distribution $H(\cdot)$ on a measurable space $\mathcal{X}$, and yield a discrete distribution on a finite or countably infinite subset of $\mathcal{X}$. This means the output distribution is a weighted set of impulses at points in $\mathcal{X}$.

[illegible] *Given a probability distribution $H(\cdot)$ on a measurable space $\mathcal{X}$, assume the values $X_k^* \in \mathcal{X}$ for $k = 1, ..., \infty$ are independently and identically distributed according to $H(\cdot)$. Also an infinite dimensional probability vector $\vec{p}$ is sampled from a distribution $Q(\cdot)$ independently of each X_k^* so $0 \leq p_k \leq 1$ and $\sum_{k=1}^{\infty} p_k = 1$. Then*

$$\sum_{k=1}^{\infty} p_k \delta_{X_k^*}(\cdot) \tag{1}$$

is an impulse mixture model *with* base distribution $H(\cdot)$ *and* probability distribution $Q(\cdot)$. *Note $\delta_{X_k^*}(\cdot)$ is a discrete measure concentrated at X_k^*.*

An example of an impulse mixture model is given in Figure 1 where the base distribution is a Gaussian and the probability vector $\vec{p}$ was generated by a so-called stick-breaking method (introduced later).

These kinds of models have a long history and appear in various forms. In [Pit96] the species sampling model is developed in the context of the PDP and attributed to R.A. Fisher for simple models of animal species. An alternative derivation takes the view that the $\vec{p}$ are initially unnormalised, so-called random measures, which one then normalises [JLP09]. There are various schemes for sampling the probability vector $\vec{p}$ including normalized random measures [JLP09] and general species sampling schemes [IJ03]. The Poisson-Dirichlet distribution we present here is a particular version that is partially conjugate and thus admits convenient Bayesian computation.

When the base distribution is continuous, such as with the Gaussian, one can see that almost surely no two values X_k^* and X_l^* (for $k \neq l$) would be the same. The general property is described as *non-atomic*, which means $H(X) = 0$ for all $X \in \mathcal{X}$, then samples from $H(\cdot)$ are almost surely distinct. The counter property is where the distribution is *discrete*, so $H(X) > 0$ for all $X \in \mathcal{X}$ and thus it is always possible that $X_k^* = X_l^*$ for some $l \neq k$. This distinction, non-atomic versus discrete, has important consequences for posterior analysis as explained in Section 8. Mixed base distributions are not considered.

When sampling from an impulse mixture model, one would observe a sequence of N data values $X_1, X_2, X_3, ..., X_N$, and one may also assume corresponding indices,

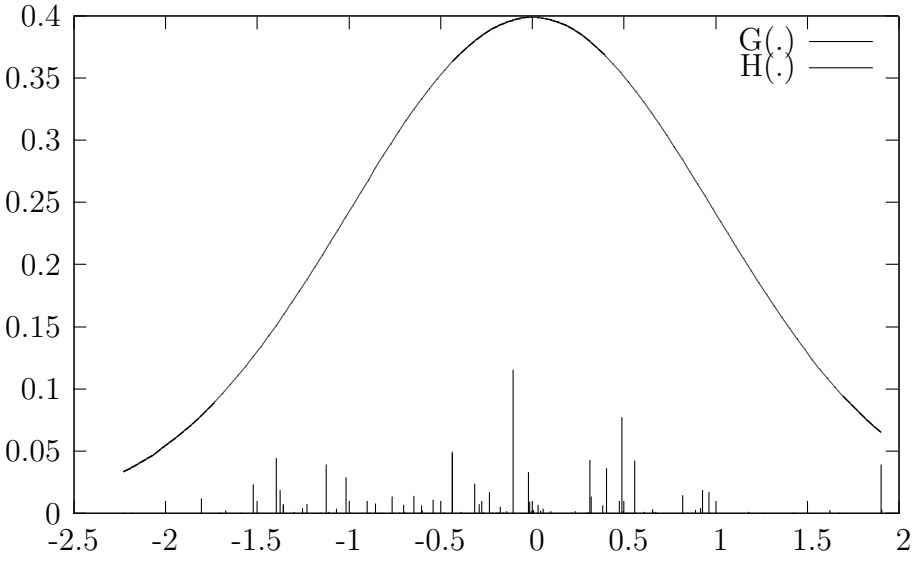

Figure 1: Gaussian base distribution $H(\cdot)$ and a resultant impulse mixture model $G(\cdot)$. The infinite number of smaller impulses do not show.

the k in Formula (1), also exist, so these might be $k_1, k_2, k_3, ..., k_N$. Since the actual indices are latent and not part of the observed data for the model of Formula (1), the latent indices can be arbitrarily relabelled and converted to a normal form. Such a normal form where the labels are irrelevant is called a *partition*: it defines the grouping of items in the sequence, not the actual indices assigned.

Definition 2 *A* partition *P of a countable set X is a mutually exclusive and exhaustive set of subsets of X. The* partition size *of P is given by the number of sets $|P|$.*

In our case, we make a partition of the data sequence $\{X_1, X_2, X_3, ..., X_N\}$. Partitions are important because, not only do they provide the assignments in a mixture model, they can also be used as a primitive in the generation of many data structures. For instance, random trees can be created using partitions either by a top-down process of fragmentation, illustrated in Figure 2, or by a bottom-up process of coagulation, illustrated in Figure 3. When storing and sampling partitions, however, one may need to order them. Most importantly, when sampling infinite partitions or infinite probability vectors $\vec{p}$, one needs to order them roughly by size: there is no point in generating a million infinitesimal probabilities before generating one with significant size.

One obvious ordering condition to try is by size. One could make the probability vector $\vec{p}$ ordered so that $p_{k+1} \leq p_k$ for all k, or order bins in the sampled partition by their decreasing number of entries. From a statistical perspective, however, this

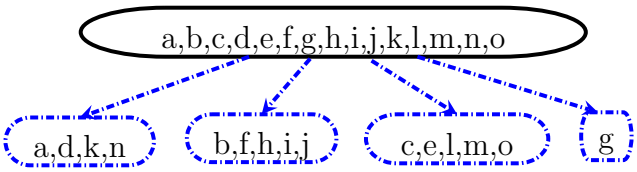

Figure 2: Shows a partition of a set of letters $\{a, b, ..., o\}$, illustrating a *fragmentation*. The top node is the full set and the bottom row is the partition displayed in order of least elements, *i.e.* a, b, c, g.

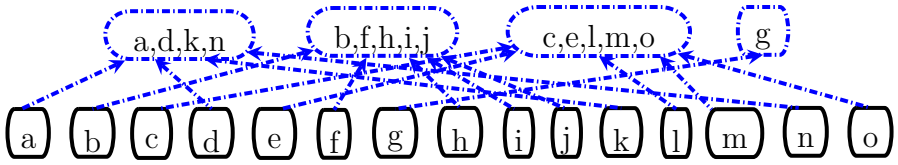

Figure 3: Shows a partition applied to a set of sets of letters $\{\{a\}, \{b\}, ..., \{o\}\}$, illustrating a *coagulation*. The bottom set of nodes lists the sets, and the top row applies the partition to the set of sets to get subsets of sets which are then unioned. The top row is displayed in order of least elements.

is impractical because one does not know the value of each p_k, and the sizes of the bins will vary during sampling.

The *size-biased order* and the *order of least elements* are two ways of ordering sets that order partitions by their order of first occurrence in a sequence. The sets in the partition can then be enumerated according to this order. For instance, assuming the data sequence is 'a', 'b', 'c', ..., then the top partition in Figure 3 is listed left to right in order of least elements. We define slightly different versions of these depending in the representation of the the set being ordered.

Definition 3 *Size-biased orders are defined for index sequences, probability vectors and partitions:*

- *An* index sequence I *of length* N *given by* $k_1, k_n, ..., k_N$ *is in* size-biased order *if* $k_1 = 1$ *and* $k_n \leq 1 + \max_{1 \leq i < n} k_i$ *for* $n = 2, ..., N$.
- *An* infinite probability vector $\vec{p}$ *is in* size-biased order *if the elements* p_k *have been reordered to be in their order of first occurrence in a random sample from* $\vec{p}$.
- *A* partition P *is in* order of least elements *if members are listed in order of their least element.*

Note these three definitions are related as follows: if we take an infinite random sample from a probability vector $\vec{p}$, and then renumber the index sequence so they

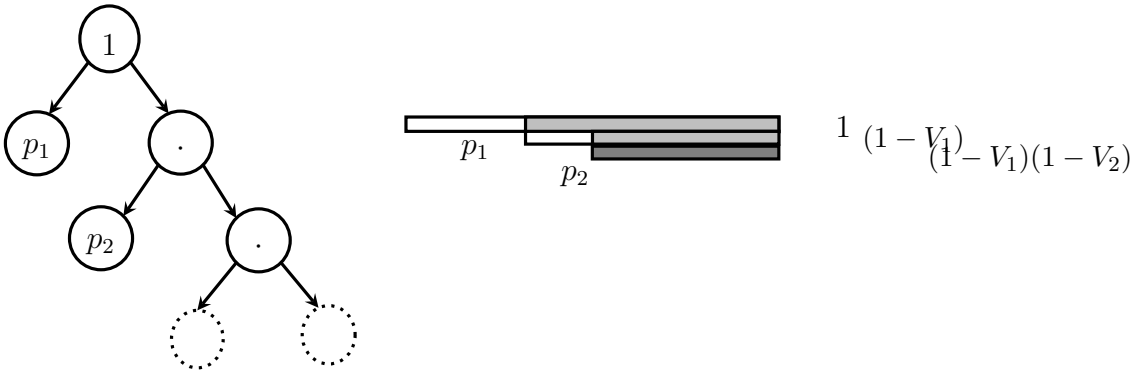

Figure 4: The stick-breaking analogy for a GEM. Left shows the tree of probabilities being generated. Middle shows the stick lengths, with p_1 and p_2 already broken off, and the remainder, $(1 - V_1)(1 - V_2)$, about to be broken.

are in size-biased order, then the corresponding renumbering converts the probability vector $\vec{p}$ to size-biased order. If we take an infinite random sample from an impulse mixture model with probability vector $\vec{p}$ and non-atomic base distribution over $\mathcal{X}$, partition the sample according the values from $\mathcal{X}$, and then label the entries by the order of occurrence, $1, 2, 3, ...$, then the order of least elements of the partition also yields a size-biased order of the probability vector $\vec{p}$.

A normal form for index sequences can be derived by renumbering indices so that the sequence is in size-biased order: indices are renumbered to 1,2,3,... according to their first occurrence in the sequence. So a size-biased ordered *renumbering* of the indices $12, 435, 7198, 12, 12, 35, 7198$ is $1, 2, 3, 1, 1, 4, 3$. Also, a partition of a data sequence can be represented by a size-biased order of the indices. The top partition in Figure 3, for instance, can be represented by the sequence "1 2 3 1 3 2 4 2 2 2 1 3 3 1 3".

3 [redacted]

For an impulse mixture model, we not only need a base distribution $H(\cdot)$ but also need to specify the probability vector $\vec{p}$. Within the PDP literature, $\vec{p}$ follows a two parameter Poisson-Dirichlet distribution [PY97] when the probabilities p_k are ordered by decreasing size, or equivalently by the Griffiths-Engen-McCloskey (GEM) distribution [Pit06] when the probabilities $\vec{p}$ have a size-biased order.

The GEM distribution is defined via the so-called "stick-breaking" model which goes as follows:

1. We take a stick of length one and randomly break it into two parts with proportions V_1 and $1 - V_1$. The first broken stick has length V_1.

2. We then take the remaining part, of length $1 - V_1$ and apply the same process to randomly break into proportions V_2 and $1 - V_2$. This second broken stick is the first part, of length $(1 - V_1)V_2$.

3. Again, we take the remaining part, of length $(1 - V_1)(1 - V_2)$ and apply the same process to randomly partition into proportions V_3 and $1 - V_3$. This third broken stick is the first part, of length $(1 - V_1)(1 - V_2)V_3$.

4. ...

Formally, this goes as follows:

Definition 4 *For $0 \le a < 1$ and $b > -a$, suppose that independent random variables V_k are such that V_k has Beta$(1-a, b+k\,a)$ distribution. Let*

$$p_1 \;=\; V_1, \qquad p_k \;=\; (1 - V_1) \cdots (1 - V_{k-1})V_k \qquad k \ge 2 \; .$$

Define the Griffiths-Engen-McCloskey distribution *with parameters a, b, abbreviated GEM(a, b) to be the resultant distribution of $(p_1, p_2, ...)$.*

Definition 5 *Let $(\tilde{p}_1, \tilde{p}_2, ...) \;\sim\; GEM(a, b)$ and define $\vec{p} = (p_1, p_2, ...)$ to be their sorted values so that $p_1 \ge p_2 \ge \cdots$. Then $\vec{p}$ follows the* Poisson-Dirichlet distribution *with parameters a, b, abbreviated PDD(a, b).*

Here the parameter a is usually called the *discount parameter* in the literature, and b is called the *concentration parameter*. The term concentration when used in the statistics community usually means a quantity that behaves like the inverse of a variance.

Why do we need two definitions, a PDD and a GEM with the same parameters modelling different orderings of the same distribution?

- When estimating the probability $\vec{p}$ for the mixture model of Formula (1) and using the GEM distribution, the sorting is useful for sampling efficiency [KWT07]. One acts like one is using a PDD.

- The PDD gives a canonical form for the distribution. Different sorts of the one underlying distribution can be generated with a GEM.

- The GEM has a convenient sampling form (the stick-breaking model) that allows for simpler analysis.

- When used inside a mixture model or partition model, they are indistinguishable, so we use which ever is convenient.

Samples of $\vec{p} \sim$ PDD(a, b) for different parameter settings are given in Figure 5, showing both probabilities and log scale probabilities. One can see that for the initial values (the first 20 say), the effects of the discount a are not that great,

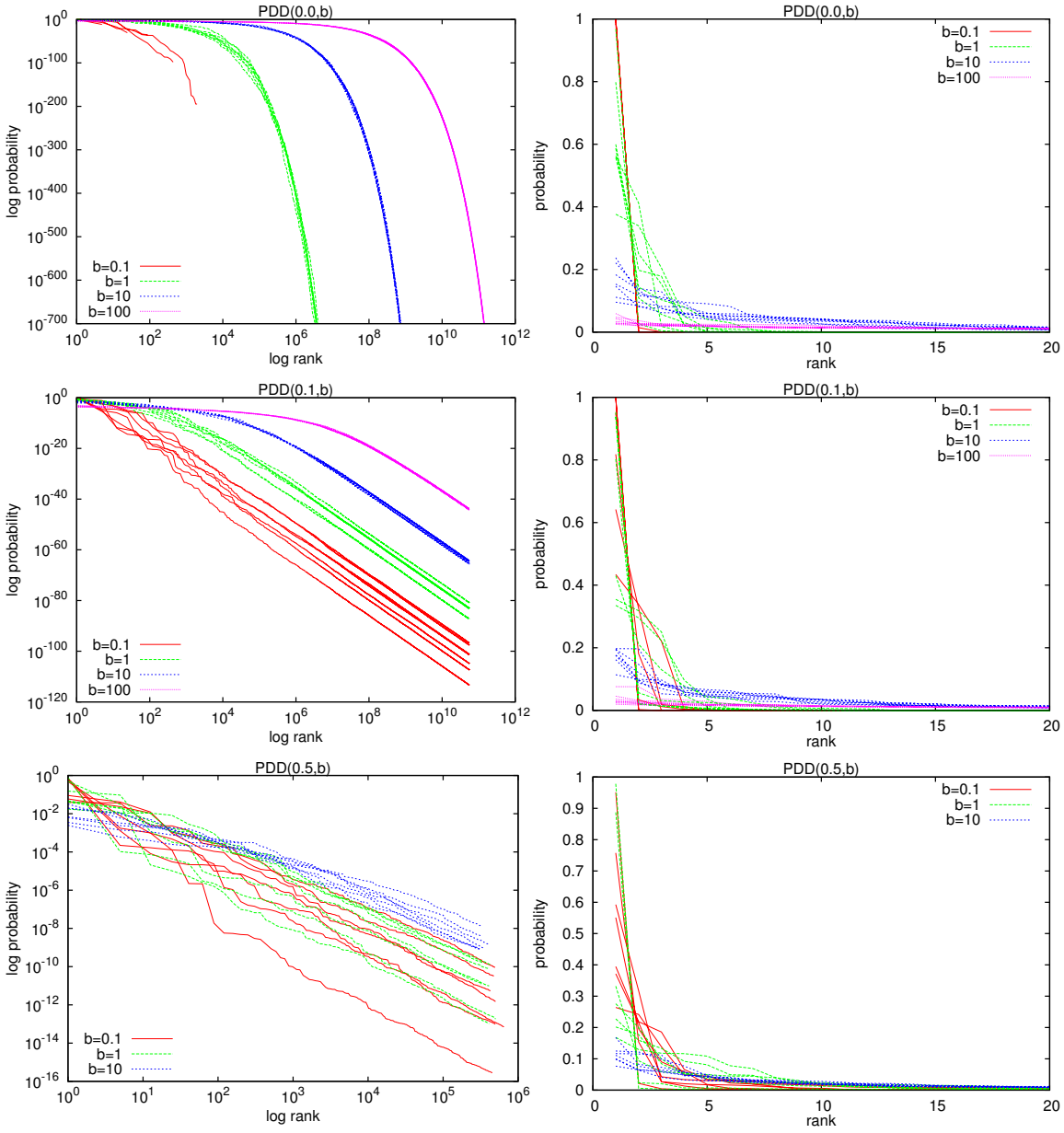

Figure 5: The PDD with different parameter settings. Each line of each plot shows $\vec{p} \sim \text{PDD}(a, b)$ for a fixed discount a and a choice of concentrations b. The left plots are log-log scale, showing hundreds of values. The right plots are normal scale showing just 25 values (subsequent ones are effectively zero).

whereas the concentration b effectively changes the onset of the decrease. We show later that vector values for the PDD with discount $a > 0$ are Zipfian, behaving roughly like $p_k \propto k^{-1/a}$, which makes the slope on the log-log plots $-1/a$. Whereas for the case with discount $a = 0$ the values are geometric, behaving roughly like $p_k \propto e^{-k/b}$.

A common definition of a Poisson-Dirichlet process is that it extends the Poisson-Dirichlet (or GEM) distribution. This definition presents the PDP as a functional on distributions: it takes as input a measurable space with domain $\mathcal{X}$, and a distribution over it called the *base distribution*, usually represented here as $H(\cdot)$, and yields as output a discrete distribution with a finite or countable set of possible values on the domain $\mathcal{X}$.

Definition 6 *Let $H(\cdot)$ be a distribution over some measurable space $\mathcal{X}$. For $0 \le a < 1$ and $b > -a$, suppose that $\vec{p}$ is drawn from a Poisson-Dirichlet (or GEM) distribution with parameters a, b. Moreover, let X_k^* for $k = 1, 2, ...$ be a sequence of independent samples drawn according to $H(\cdot)$ and independent of $\vec{p}$. Then $\vec{p}$ and X_k^* for $k = 1, 2, ...$ define a discrete distribution on $\mathcal{X}$ given by the formula*

$$\sum_{k=1}^{\infty} p_k \delta_{X_k^*}(\cdot) \ . \tag{2}$$

This distribution is a Poisson-Dirichlet Process *with parameters a, b and* base distribution *$H(\cdot)$, denoted PDP$(a, b, H(\cdot))$.*

The Dirichlet Process (DP) is the special case where $a = 0$, and has some quite distinct properties as shown later. Note strict requirements on the sorting or ordering as in Definition 5 are not needed in the definition of the PDP due to the effect of the summation.

The PDP is also called a *stochastic process* because it can be defined as a sequence of values $X_1, X_2, ... \in \mathcal{X}$ from some *base probability distribution* $H(\cdot)$ indexed by integer valued time as $1, 2, 3, ...$ The stochastic process is the sequential sample from this PDP distribution. The conditional distribution with $\vec{p}$ marginalised out for this, as long as $H(\cdot)$ is non-atomic, is as follows:

$$p(X_{N+1} \,|\, X_1, ..., X_N, a, b, H(\cdot)) \ = \ \frac{b + Ma}{b + N} H(\cdot) + \sum_{m=1}^{M} \frac{n_m - a}{b + N} \delta_{X_m^*}(\cdot) \ , \tag{3}$$

where there are M distinct values in the sequence $X_1, ..., X_N$ denoted by $X_1^*, ..., X_M^*$ and their occurrence counts respectively are $n_1, ..., n_M$, so $\sum_{m=1}^{M} n_m = N$.

The Chinese restaurant analogy for this sequential sampling process goes as follows:

- A *customer* walks into the restaurant and sees M occupied *tables* where n_m others sit at table m enjoying the *menu item* X_m^*.

- He can start his own table with probability $\frac{b+Ma}{b+N}$ and receive a new item X^*_{M+1} from menu $H(\cdot)$ by sampling.

- Otherwise, he goes to one of the existing M tables with probability $\frac{n_m-a}{b+N}$ and enjoys the item X^*_m.

The Chinese Restaurant Process (CRP) is defined over the partition so generated, represented with a size-biased order of indices.

Definition 7 (CRP) *For $0 \le a < 1$ and $b > -a$, generate a sequence of integers $k_1, k_2, ...$ as follows:*

$$p(k_{N+1}|k_1, ..., k_N, a, b) \;=\; \frac{b+Ma}{b+N}\,\delta_{k_{N+1}=M+1} + \sum_{m=1}^{M} \frac{n_m - a}{b+N}\delta_{k_{N+1}=m}\;,$$

where M and the counts $n_1, n_2, ..., n_M$ are derived as for Equation (3).

If indices are sampled according to the CRP then one gets a size-biased ordered index sequence from a PDD(a, b) [Pit95, IJ01], and thus the CRP can serve as an alternative sampler for the PDP where the probability vector $\vec{p}$ is unknown.

4 [illegible]

Before getting onto discrete domains, we review basic properties of the PDD, and of the PDP in non-atomic domains. Some of these results will be used subsequently to address discrete domains.

For the sample from the distribution of Formula (2), a latent sequence of indices exists $I_N := (k_1, ..., k_N)$, however, these remain hidden (only the corresponding data values are known, not the indexes). For a non-atomic base distribution the indices are irrelevant and we can renumber them by size-biased ordering. Each index corresponds to a table in the CRP, and the number of distinct indices in the sample is the number of tables active at the restaurant.

Our notation for statistics is as follows:

Definition 8 (Statistics) *When sampling independently and identically from a discrete distribution in the form of Formula (2), one gets a sequence of latent indices, an* index sequence *of length N given by $I_N = k_1, k_2, ..., k_N$. In I_N one index value k can occur multiple times. Sort and count the N points of I_N. Suppose there are M distinct values in I_N, k^*_m for $m = 1, ..., M$ that occur n_m times respectively, so $\sum_{m=1}^{M} n_m = N$. Call M the* partition size *and note it depends on the sample I_N and the sample size N. Moreover, use a size-biased ordering of the indices and renumber them according to this. The corresponding size-biased ordered index sequence is denoted I^*_N.*

Consider a sample with $N = 7$ points with latent indices $I_N = 12, 435, 7198, 12, 12, 35, 7198$. Then the size-biased ordered renumbering is $I^*_N = 1, 2, 3, 1, 1, 4, 3$. Thus the partition size $M = 4$ and occurrence counts $n_1 = 3$, $n_2 = 1$, $n_3 = 2$ and $n_4 = 1$. Note the partition size M for a sequence corresponds to the number of active tables in the CRP terminology.

Definition 9 *When sampling from the discrete distribution of Formula (3) or via a PDP, one gets a* data sequence *of length N given by $S_N = X_1, X_2, ..., X_N$. Sort and count the N points of S_N. Suppose there are M distinct values in S_N, X^*_m for $m = 1, ..., M$ that occur n_m times respectively, so $\sum_{m=1}^{M} n_m = N$. For non-atomic base distributions $H(\cdot)$, it is safe to associate index m with data value X^*_m, which is the unique size-biased ordered renumbering. The corresponding index sequence is denoted I^*_N.*

For discrete base distributions $H(\cdot)$, a unique size-biased ordered renumbering for the indices does not exist because if two data items in a sample are equal, one cannot be certain their latent indices are the same.

4 The DP

The Dirichlet Process (DP) is a special case of the PDP when the discount parameter $a = 0$. It has quite distinct properties as subsequent analysis will show. It is usually defined in a completely different manner to the PDP as follows. Let $\mathcal{X}$ be a measurable space. For a random probability distribution $G(\cdot)$ to be distributed according to a DP, its marginal distributions have to be Dirichlet distributions too. Ferguson [Fer73] gave a formal definition of the DP as follows.

Definition 10 (DP) *Let $H(\cdot)$ be a random measure on $\mathcal{X}$ and $b > 0$ be positive real number. We say a random probability measure $G(\cdot)$ on $\mathcal{X}$ is a Dirichlet process with a base measure $H(\cdot)$ and concentration parameter b, i.e. $G(\cdot) \sim DP(b, H(\cdot))$, if for any finite measurable partition $(B_1, B_2, \ldots, B_k)$ of $\mathcal{X}$, the random vector $(G(B_1), G(B_2), \ldots, G(B_k))$ is Dirichlet distributed with parameter $(bH(B_1), bH(B_2), \ldots, bH(B_k))$:*

$$(G(B_1), G(B_2), \ldots, G(B_k)) \sim Dirichlet(bH(B_1), bH(B_2), \ldots, bH(B_k)) \ .$$

The DP is an extension of a Dirichlet distribution, which is defined for a finite set. The following simple corollary of the definition above demonstrates this.

Corollary 11 *According to Definition 10, if $H(\cdot)$ is a categorical distribution over a finite space, represented by the probability vector $\vec{\theta}$ say, then the following holds*

$$DP(b,\ H(\cdot)) = Dirichlet(b\vec{\theta}) \ .$$

This is important because posterior analysis of hierarchical Dirichlets is intrinsically difficult, so posterior analysis of hierarchical DPs can help.

2.2 [illegible]

The PDD can be used to approximate a broader class of distributions, not just those sampled from a PDD(a, b). The following lemma derived from James [Jam08, Proposition 2.2] shows this. This supposes a "true" probability vector $\vec{q}$ gives a distribution of integers and then shows a sufficient property required of $\vec{q}$ so that a PDD distribution can approximate it based on samples.

[illegible] *Suppose an integer sequence I of length N is sampled independently and identically according to the probabilities $\vec{q}$ where $0 \leq q_k \leq 1$ for $k = 1, 2, ...$ and $\sum_{k=1}^{\infty} q_k = 1$ and use the notation of Definition 8. If it is assumed the $\vec{q}$ is PDD(a, b) for $0 \leq a < 1$ and $b > -a$, then the posterior distribution on $\vec{q}$ given I converges weakly to $\vec{q}$ if $\mathbb{E}_{I|\vec{q},N}[M/N] \to 0$ as $N \to \infty$ where M is the partition size defined in Definition 8.*

Basically, we have some "true" model over samples given by the probability vector $\vec{q}$. From this we compute the expected partition size $\mathbb{E}_{I|\vec{q},N}[M]$ for sample sequences I of size N, and then check this grows slower than N as $N \to \infty$. If this holds for $\vec{q}$, then the distribution $\vec{q}$ can be learnt using Bayesian methods that assume $\vec{q}$ is PDD(a, b). We show later that if $\vec{q} \sim$ PDD$(0, b)$, then almost surely $\mathbb{E}_{I|\vec{q},N}[M]$ is $O(\log N)$ and if $\vec{q} \sim$ PDD(a, b) for $a > 0$, then almost surely $\mathbb{E}_{I|\vec{q},N}[M]$ is $O(N^a)$.

As a warning more than anything else, it is important to realise the PDP should not be used to approximate continuous distributions. This is made precise by the consistency result due to James [Jam08, Proposition 2.1].

[illegible] *Suppose data is sampled independently and identically from a Polish space $\mathcal{X}$ according to a continuous distribution $P_0(\cdot)$, and let $H(\cdot)$ be another distribution on $\mathcal{X}$ where $H(\cdot)$ is non-atomic. Then the posterior of the PDP$(a, b, H(\cdot))$ distribution given the sample converges weakly to point mass at the distribution*

$$aH(\cdot) + (1 - a)P_0(\cdot)$$

Hence the posterior is consistent only if either $H(\cdot) = P_0(\cdot)$ or $a = 0$.

Note discrete distributions cannot be continuous since they have finite mass concentrated at points. Thus the above lemma does not apply to the discrete case. When the "true" distribution $P_0(\cdot)$ is discrete, weak converge does hold.

2.3 [illegible]

One can derive the probability of evidence or data given the model, a useful diagnostic in Bayesian analysis. Various versions of this are well known, see [PY97, Appendix] and [Pit95, Proposition 9], and easily proven by induction using the CRP.

Lemma 14 *Consider finite samples $S_N = X_1, X_2, ..., X_N$ from PDP$(a, b, H(\cdot))$, where the base distribution $H(\cdot)$ is non-atomic. Use the notation of Definition 9. Then the probability of evidence given the model PDP$(a, b, H(\cdot))$ is*

$$p(X_1, X_2, ..., X_N|a, b) = \frac{(b|a)_M}{(b)_N} \prod_{m=1}^{M} H(X_m^*) \prod_{m=1}^{M} (1-a)_{n_m-1} \ ,$$

where $(x)_N$ denotes the Pochhammer symbol $x(x+1)...(x+N-1) = \Gamma(x+N)/\Gamma(x)$ and $(x|y)_N$ denotes $x(x+y)...(x+(N-1)y)$, the Pochhammer symbol with increment y, and $(x|0)_N = x^N$.

5 [illegible]

One can easily see the posterior in Lemma 14, with the term for the base distribution ($\prod_{m=1}^{M} H(X_m^*)$) removed represents a distribution on a partition. This section presents various properties of this distribution.

5.1 [illegible]

For a partition represented as a size-biased order I_N^*, the probability is a function of the partition size M and the occurrence counts $n_1, ..., n_M$, where by default the counts are listed in size-biased order. The resultant probability on I_N^* is then a neat function $f(n_1, ..., n_M)$. This is called an *exchangeable partition probability function*. This goes under various names in the literature, so the term *Chinese Restaurant Distribution* is used here to differentiate it from the CRP from which it is derived.

Definition 15 *Given a set P of size $N = |P|$, represent the partitions of P by the set of size-biased orderings of length N, where one is denoted I_N^*. Define*

$$p(I_N^*|a, b) = \begin{cases} \frac{(b|a)_M}{(b)_N} \prod_{m=1}^{M} (1-a)_{n_m-1} & \text{for } a > 0 \ , \\ \frac{b^M}{(b)_N} \prod_{m=1}^{M} (n_m - 1)! & \text{for } a = 0 \ . \end{cases}$$

where M and the occurrence counts $n_1, ..., n_M$ follow Definition 8. Call this the Chinese Restaurant Distribution *with parameters (a, b), abbreviated CRD(P, a, b), and note its samples are a partition of P.*

Note this distribution runs over all possible partitions, so for instance if $N = 3$ the possible partitions of $\{a, b, c\}$ represented using an ordering of least elements are given in Table 1.

partition	$\{\{a,b,c\}\}$	$\{\{a,b\},\{c\}\}$	$\{\{a,c\},\{b\}\}$	$\{\{a\},\{b,c\}\}$	$\{\{a\},\{b\},\{c\}\}$
partition I_N^*	$(1,1,1)$	$(1,1,2)$	$(1,2,1)$	$(1,2,2)$	$(1,2,3)$
size M	1	2	2	2	3
counts $\vec{n}$	(3)	$(2,1)$	$(2,1)$	$(1,2)$	$(1,1,1)$
$p(I_N^* \mid a>0,b)$	$\frac{(1-a)(2-a)}{(b+1)(b+2)}$	$\frac{(b+a)(1-a)}{(b+1)(b+2)}$	$\frac{(b+a)(1-a)}{(b+1)(b+2)}$	$\frac{(b+a)(1-a)}{(b+1)(b+2)}$	$\frac{(b+a)(b+2a)}{(b+1)(b+2)}$
$p(I_N^* \mid a=0,b)$	$\frac{2}{(b+1)(b+2)}$	$\frac{b}{(b+1)(b+2)}$	$\frac{b}{(b+1)(b+2)}$	$\frac{b}{(b+1)(b+2)}$	$\frac{b^2}{(b+1)(b+2)}$

Table 1: Space of partitions over $\{a,b,c\}$.

2 [illegible]

The key characteristic of a sample from the PDD or a CRD is the *partition size* M from Definition 8. This is related to the expected posterior probability of starting a new bin (for the PDD or CRP) or a new data value from $\mathcal{X}$ in the non-atomic case of the PDP. This is given by the formula for the unseen part of the CRP,

$$p(k_{N+1} \notin I_N^* \,|\, I_N^*, M, a, b) \;=\; p(X_{N+1} \notin S_N \,|\, S_N, M, a, b) \;=\; \frac{b + M\,a}{N+b}\;.$$

The posterior distribution for the partition size given just the sample size introduces a significant function, $S^N_{M,a}$, which is a generalised Stirling number. It was applied to the task by Pitman [Pit99, Equation (89)] where it was represented as $a(N,M,a)$ and by Teh [Teh06a] in the form $s_a(N,M)$, as a generalised Stirling number of type $(-1,-a,0)$ attributed to Hsu and Shiue, where it was applied to the analysis of hierarchical PDPs. The case for $a=0$ was first presented by Antoniak [Ant74, p1161].

[illegible] *Consider the size-biased ordering of indices I_N^* for a sample of size N from a PDD with parameters (a,b). The probability distribution for M given just N and integrating over all possible partitions I_N^* of size M is*

$$p(M \,|\, N, a, b) \;=\; \frac{(b|a)_M}{(b)_N} S^N_{M,a}, \quad \textit{where} \tag{4}$$

$$S^N_{M,a} \;:=\; N! \sum_{\sum_1^M n_m = N,\; n_m \geq 1} \prod_{m=1}^{M} \left(\frac{(1-a)_{n_m-1}}{n_m!} \frac{n_m}{\left(N - \sum_{i=1}^{m-1} n_i\right)} \right)\;, \tag{5}$$

for $M \leq N$ and 0 else.

The following expressions are useful for computing $S^N_{M,a}$.

[illegible] $S^N_{M,a}$)

$$
\begin{aligned}
&\textit{(i) Linear recursion:} && S^{N+1}_{M,a} = S^N_{M-1,a} + (N - Ma) S^N_{M,a} \\
&\textit{Boundary cond.:} && S^N_{M,a} = 0 \textit{ for } M > N, \quad S^N_{0,a} = \delta_{N,0}. \\
&\textit{(ii) Mult. recursions:} && S^N_{M,a} = \sum_{n=m}^{N-M+m} \frac{\binom{N}{n}}{\binom{M}{m}} S^n_{m,a} S^{N-n}_{M-m,a} = \sum_{n=1}^{N-M+1} \binom{N-1}{n-1} S^n_{1,a} S^{N-n}_{M-1,a} \\
& && S^N_{1,a} = \Gamma(N-a)/\Gamma(1-a). \qquad \textit{Any } 0 < m < M. \\
&\textit{(iii) Explicit expression:} && S^N_{M,a} = \frac{1}{M!\, a^M} \sum_{m=0}^{M} \binom{M}{m} (-)^m \prod_{h=0}^{N-1} (h - am) \qquad \textit{for } a > 0 \\
& && S^N_{M,0} = \frac{(-)^{M-1}}{(M-1)!} \frac{\partial^{M-1}}{\partial a^{M-1}} \left(\frac{\Gamma(N-a)}{\Gamma(1-a)} \right) \Bigg|^{a=0} \\
&\textit{(iv) Asymptotic expr.:} && S^N_{M,a} \simeq \frac{1}{\Gamma(1-a)} \frac{1}{\Gamma(M)\, a^{M-1}} \frac{\Gamma(N)}{N^a} \textit{ for } a > 0 \\
&\textit{(v) Expr. for } a = 0\textit{:} && S^N_{M,0} = |s^{(M)}_N| = \textit{unsigned Stirling\# of 1st kind [AS74]}
\end{aligned}
$$

The asymptotic expression holds for $N \to \infty$ and fixed M and a.

The explicit closed form (*iii*) was developed in [Tos39] and presented in [Pit06]. This explicit form becomes unstable for large values of M since it is effectively an M-point interpolation to a partial derivative, thus has a lot of differences of similar values. It remains effective for small M. Some examples are:

$$
\begin{aligned}
S^N_{2,0} &= S^N_{1,0} \left(\psi_0(N) - \psi_0(1) \right) & S^N_{3,0} &= \frac{1}{2} S^N_{1,0} \left(\psi_1(N) - \psi_1(1) + (\psi_0(N) - \psi_0(1))^2 \right) \\
S^N_{2,a} &= \frac{1}{a} \left(S^N_{1,a} - S^N_{1,2a} \right) & S^N_{3,a} &= \frac{1}{2a^2} \left(S^N_{1,a} - 2 S^N_{1,2a} + S^N_{1,3a} \right) .
\end{aligned}
$$

Figure 6 illustrates the shape of the distributions and their location for different values of a and b and fixed $N = 1000$. Similar looking plots are produced when $N = 10000$. Note the distribution does reflect a Poisson in some ways, being skewed both at the lower boundary $M = 0$ and the upper boundary $M = N$, and being fairly symmetric in other cases. Figure 7 illustrates the shape of the distributions and their location for different values of a as N grows, for $b = 50$. The figure for $a = 0.9$ has a different horizontal scale. Note also how the spread of M increases as the sample size N increases.

3 [illegible]

It is well known that expected partition size for the DP (PDP with $a = 0$) is $O(\log N)$ and for the PDP it is $O(N^a)$. Here the exact rates are presented along with their expected variance [YS00]. Further details of moments for the PDD are also given by Ishwaran and James [IJ01].

[illegible] *In the context of Definition 8, if a partition sequence I^*_N has the probability vector $\vec{p}$ distributed* a priori *according to PDD(a, b), the expected* a posteriori *M for a sample of size N denoted $\mathbb{E}_{\vec{p}|a,b,N}[M]$*

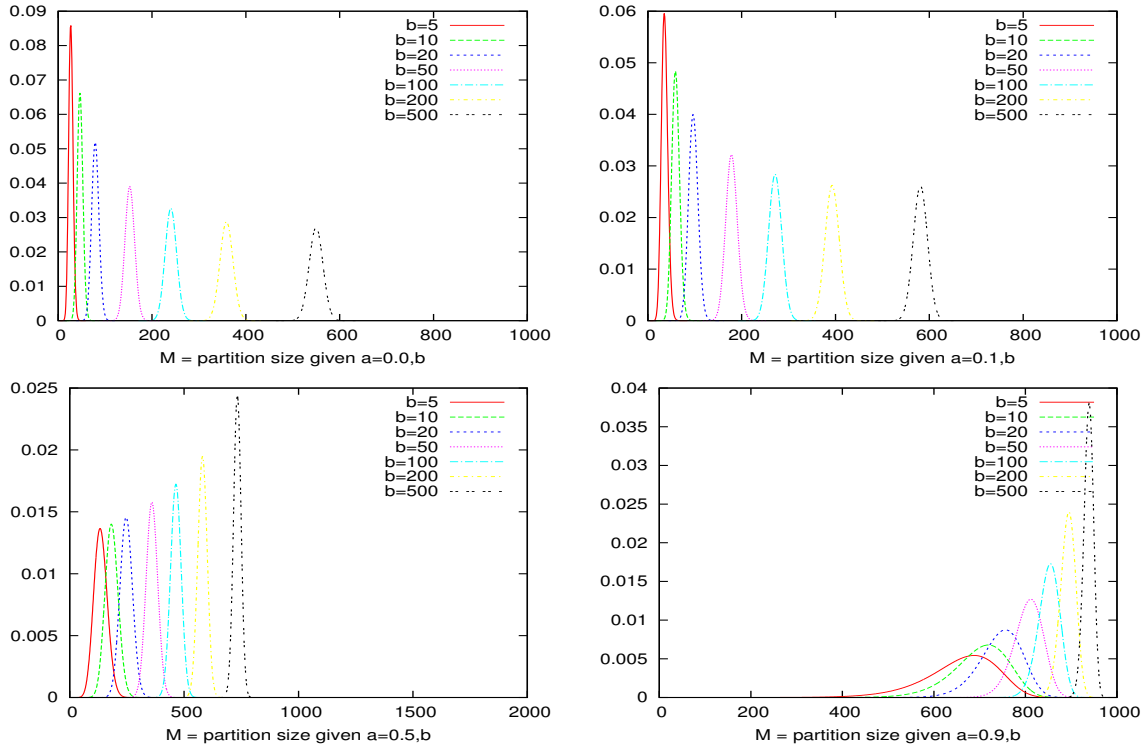

Figure 6: Posterior probability on M given $N = 1000$ and different a.

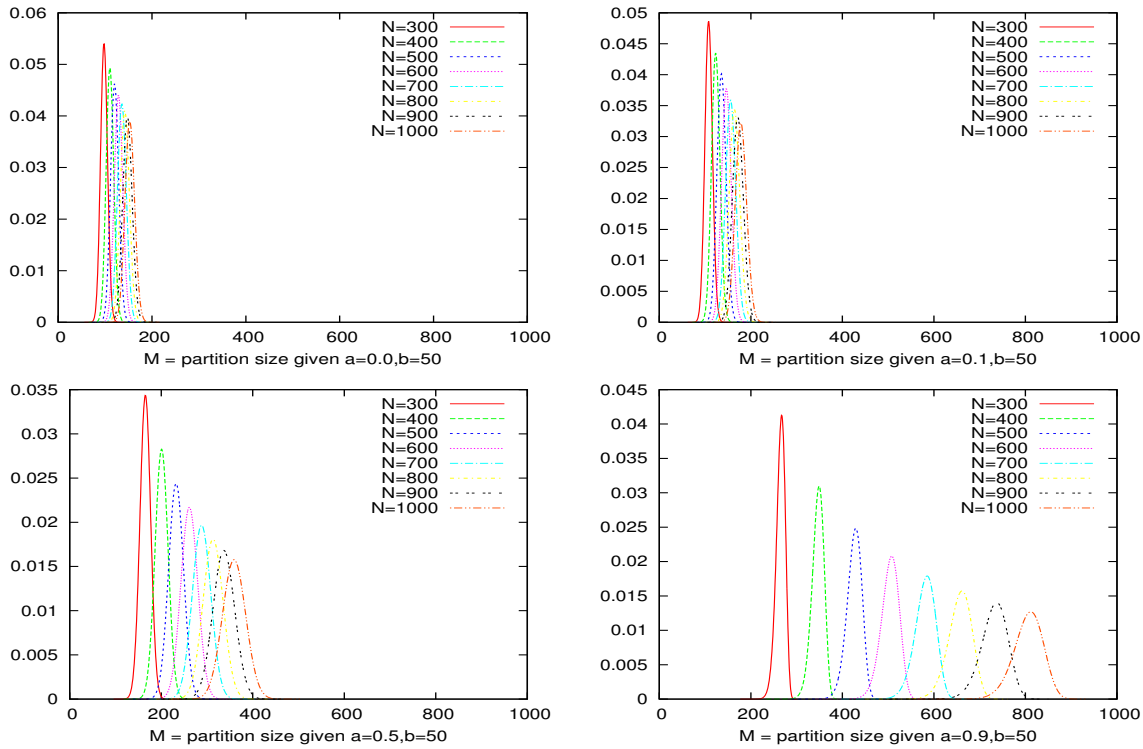

Figure 7: Posterior probability on M for increasing N and fixed $b = 50$.

(and note the actual sample is unknown here, just its size N is known), when $a > 0$ is given by

$$\begin{aligned}
\mathbb{E}_{\vec{p}|a,b,N}[M] &= \frac{b}{a}\frac{(b+a)_N}{(b)_N} - \frac{b}{a}\,, \\
&= b\left(\phi_0(b+N) - \phi_0(b)\right) + O(ab\log^2 N) \\
&\simeq \frac{b}{a}\left(1+\frac{N}{b}\right)^a \exp\left(\frac{aN}{2b(b+N)}\right) - \frac{b}{a} \qquad \textit{for } N, b \gg a\,,
\end{aligned}$$

where $(x)_N$ denotes the Pochhammer symbol $x(x+1)...(x+N-1) = \Gamma(x+N)/\Gamma(x)$. The a posteriori *variance of M for a sample of size N, denoted $Var_{\vec{p}|a,b,N}[M]$, when $a > 0$ is given by*

$$\begin{aligned}
Var_{\vec{p}|a,b,N}[M] &= \frac{b(a+b)}{a^2}\frac{(b+2a)_N}{(b)_N} - \frac{b}{a}\frac{(b+a)_N}{(b)_N} - \left(\frac{b}{a}\frac{(b+a)_N}{(b)_N}\right)^2 \\
&\simeq \frac{b}{a}\left(1+\frac{N}{b}\right)^{2a} \exp\left(\frac{aN}{b(b+N)}\right) \qquad \textit{for } N, b \gg a\,.
\end{aligned}$$

In the context where $a = 0$,

$$\mathbb{E}_{\vec{p}|a,b,N}[M] = b(\psi_0(b+N) - \psi_0(b))$$

$$\begin{aligned} &\simeq b \log\left(1 + \frac{N}{b}\right) \qquad \textit{for} \quad N, b \gg 0 \ , \\ Var_{\vec{p}|a,b,N}[M] &= b(\psi_0(b+N) - \psi_0(b)) \\ &\quad + b^2(\psi_1(b+N) - \psi_1(b)) \\ &\simeq b \log\left(1 + \frac{N}{b}\right) \qquad \textit{for} \quad N > b \gg 0 \ , \end{aligned}$$

where $\psi_0(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the 1-st order polygamma function, the derivative of the digamma function.

Thus for $0 \leq a < 1$ and b fixed, $\mathbb{E}_{\vec{p}|a,b,N}[M]$ is almost surely sublinear in N as described in Section 4.2.

Note $\mathbb{E}_{\vec{p}|a,b,N}[M]$ is roughly linear in b in all cases. For the DP case (when $a = 0$) and $N \gg b \gg 0$, the *a posteriori* standard deviation of M is approximately the square root of $\mathbb{E}_{\vec{p}|a,b,N}[M]$, so M is somewhat Poisson in its behaviour. For the PDD $a > 0$ and $N \gg b \gg a$, the *a posteriori* standard deviation of M is approximately $\mathbb{E}_{\vec{p}|a,b,N}[M] / \sqrt{b/a}$, so is smaller than $\mathbb{E}_{\vec{p}|a,b,N}[M]$ for $b \gg a$.

To compare convergence of PDD distributions with known series, we use the following lemma.

[illegible] *Suppose a partition sequence I_N^* of length N is sampled independently and identically according to the probabilities $\vec{q}$ where $0 \leq q_k \leq 1$ for $k = 1, 2, ...$ and $\sum_{k=1}^{\infty} q_k = 1$ and use the notation of Definition 8. If $\vec{q}$ takes the form of a geometric series, $q_k = r^{k-1}(1-r)$, then*

$$\mathbb{E}_{I_N^*|\vec{q}}[M] \ \leq \ \frac{\log N}{\log 1/r} + \frac{1 + 2\log 1/r + \log\log 1/r}{\log 1/r} \ .$$

If $\vec{q}$ takes the form of a Dirichlet series $q_k = k^{-s}\zeta(s)$ for $s > 1$ (where $\zeta(s)$ is the Riemann zeta function), then

$$\mathbb{E}_{I_N^*|\vec{q}}[M] \ \leq \ 3/2 + \frac{s}{(s-1)} \left(\frac{N}{\zeta(s)}\right)^{1/s} \ .$$

The bounds are often quite good. Experimental evaluation shows the geometric series bound is close to about 20% except where r approaches 1, and the Dirichlet series bound is close to about 20% except where s approaches 1.

Comparing the expected partition sizes of Lemma 18 with the different convergent series above, one can see that the PDD case for $a > 0$ behaves more like a Dirichlet series with exponent $s = 1/a$, whereas the DP case (for $a = 0$) behaves more like a geometric series with factor $r = \exp(-1/b)$.

4 [illegible]

It is instuctive to compare the CRD with the Dirichlet-multinomial model obtained by marginalising out parameters from a multinomial posterior. This is a posterior

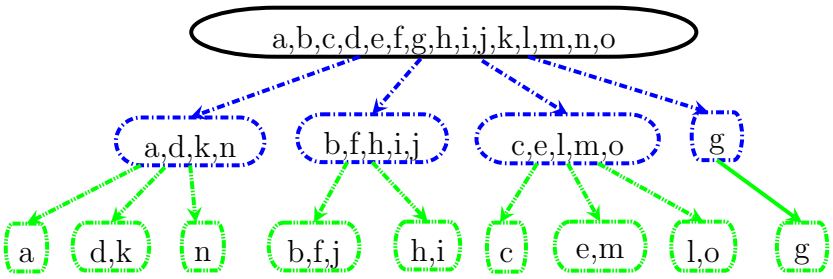

Figure 8: Shows two *fragmentations* of a set of letters $\{a, b, ..., o\}$. The top node is the full set, the middle row is after the first fragmentation, and the bottom row is after the second fragmentation. Each row is in size-biased order.

on assignments of N data points to K classes, rather than a partition of N data points. So given N data in total assigned to K classes with counts $\vec{n} = (n_1, ..., n_K)$, and using a prior of Dirichlet$_K(\vec{\alpha})$:

$$\begin{aligned} p(\vec{n}|N, \alpha) &= \frac{\Gamma(b)}{\Gamma(N+b)} \frac{\prod_k \Gamma(n_k + \alpha_k)}{\prod_k \Gamma(\alpha_k)} \\ &= \frac{1}{(b)_N} \prod_k (\alpha_k)_{n_k} \end{aligned}$$

where $b = \sum_k \alpha_k$. Here the second line represents the formula using Pochhammer symbols to bring out the correspondence with Definition 15. Comparisons are as follows:

- For the Dirichlet-multinomial, the number of classes K is fixed, for the CRD the size of the partition (M is sometimes used) is varied as well, and has the posterior given in Lemma 16.
- The $n_k - 1$ subscript in the Pochhammer symbol in Definition 15 loses 1 due to starting off the new bin in the partition.

6 [illegible]

Fragmentation is the term used when a partition created by one distribution is further split using partitions created by a second distribution. In a simple finite case, the technique is illustrated in Figure 8. Fragmentation works as follows: every set in the partition is further partitioned. A complementary process is bottom-up, called coagulation, and is illustrated in Figure 9. Coagulation makes a partition of the set of sets forming the original partition, and then merges entries in the one bin. The use of the second partition here is a bit more indirect. So, for instance, at the bottom row of Figure 9, the initial set is $\{\{a\}, \{b, f, j\}, \{c\}, \{d, k\}, \{e, m\}, \{g\}, \{h, i\}, \{l, o\}, \{n\}\}$. Partition this set into 4 parts (1) $\{\{a\}, \{d, k\}, \{n\}\}$, (2) $\{\{b, f, j\}, \{h, i\}\}$, (3)

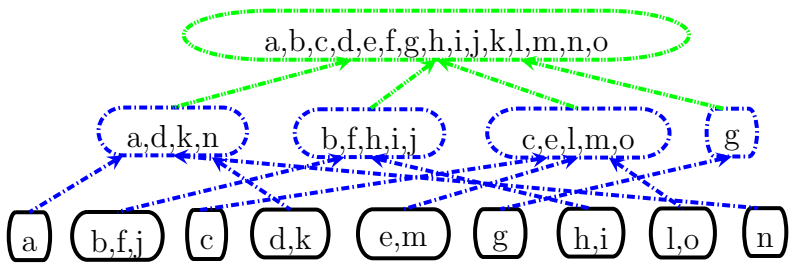

Figure 9: Shows the coagulations reversing the fragmentations in Figure 8. The bottom row is the initial partition. The middle row is a coagulation of the bottom row, and the top row is a coagulation of the middle row.

$\{\{c\},\{e,m\},\{l,o\}\}$ and (4) $\{\{g\}\}$, and then flatten these sets to get a partition of $\{a,d,k,n\}\}$, $\{b,f,h,i,j\}$, $\{c,e,l,m,o\}$, and $\{g\}$ as shown in the second row.

Definitions below are given in terms of sets. Conversion to size-biased ordering is detailed but not difficult.

Definition 20 *Consider a partition P represented as a set of sets $\{P_1,P_2,...,P_M\}$, and a sequence of partitions $Q_1,Q_2,...,Q_M$ of the sets $P_1,P_2,...,P_M$ respectively. Then the* fragmentation *of P using $Q_1,Q_2,...,Q_M$ is $\bigcup_{m=1}^{M} Q_m$.*

Definition 21 *Consider a partition P represented as a set of sets $\{P_1,P_2,...,P_M\}$, and a second partition Q of $\{1,2,...,M\}$ where $M=|P|$. Then the* coagulation *of P using Q is $\{\bigcup_{m\in q} P_m : q\in Q\}$.*

As seen in the figures, fragmentation and coagulation would seem to be complementary in their way of changing a partition, one splits and one merges, but both driven by partition templates.

6 [illegible]

In some cases, the two operations are also statistically complementary when applied to samples from the Chinese Restaurant distribution. This is usually defined in terms of fragmentation and coagulation functionals over distributions [Pit06]. For simplicity we present the operations directly in terms of Definitions 20 and 21 as follows:

Proposition 22 *For $0<a_1<1$, $0\le a_2<1$ and $b>-a_1a_2$ and a set P, if a partition Q sampled from $CRD(P,a_1a_2,b)$ is fragmented with partitions sampled from $CRD(Q_m,a_1,-a_1a_2)$, for $Q_m\in Q$, then the resultant fragmented partition is distributed as $CRD(P,a_1,b)$.*

Theorem 2 (Coagulation and Fragmentation) *For $0 < a_1 < 1$, $0 \leq a_2 < 1$ and $b > -a_1 a_2$ and a set P, let $I \sim CRD(P, a_1, b)$, the partition size of I is M, and $J \sim CRD(I, a_2, b/a_1)$ then a coagulation of I with J is $CRD(P, a_1 a_2, b)$.*

Thus one coagulates to convert a partition from $\mathrm{CRD}(P, a_1, b)$ to be a partition from $\mathrm{CRD}(P, a_1 a_2, b)$, and conversely, one fragments to convert a partition from $\mathrm{CRD}(P, a_1 a_2, b)$ to be a partition from $\mathrm{CRD}(P, a_1, b)$.

Note that if we traverse down a tree levelwise, it corresponds to performing a sequence of fragmentations. Likewise, if all leaves occur at the same depth (a condition that can be enforced by the insertion of dummy nodes), then one can traverse up the tree levelwise, and it corresponds to performing a sequence of coagulations. If we have a schedule of strictly increasing discounts $a_1, a_2, a_3, ...$ and a maximum tree depth, then we can generate a tree with leaves in $\{1, 2, ..., N\}$ levelwise as given in Algorithm 1. For the partitions generated, it is readily seen that

Algorithm 1 Sampling a tree with N nodes using schedule $a_1, a_2, ..., a_{maxdepth}$.

```
root = {1, 2, ..., N};
tree = (); nodes = ();
list ~ CRD(root, a_1, b);
for m ∈ list do
    push(tree, parent(m, root));
    push(nodes, (m, 1));
end for
while (node, depth) = pop(nodes) do
    if depth < maxdepth then
        if |node| > 1 then
            list ~ CRD(node, a_{depth+1}, −a_{depth});
            for m ∈ list do
                push(tree, parent(m, node));
                push(nodes, (m, depth + 1));
            end for
        else
            {Copy through to next depth.}
            push(tree, (node, depth + 1));
            push(nodes, (node, depth + 1));
        end if
    end if
end while
{Tree is represented as a list of parent(·, ·) relations.}
return tree;
```

- the partition at depth D (such that $1 \leq D \leq maxdepth$) as generated will be distributed as $\mathrm{CRD}(root, a_D, b)$;

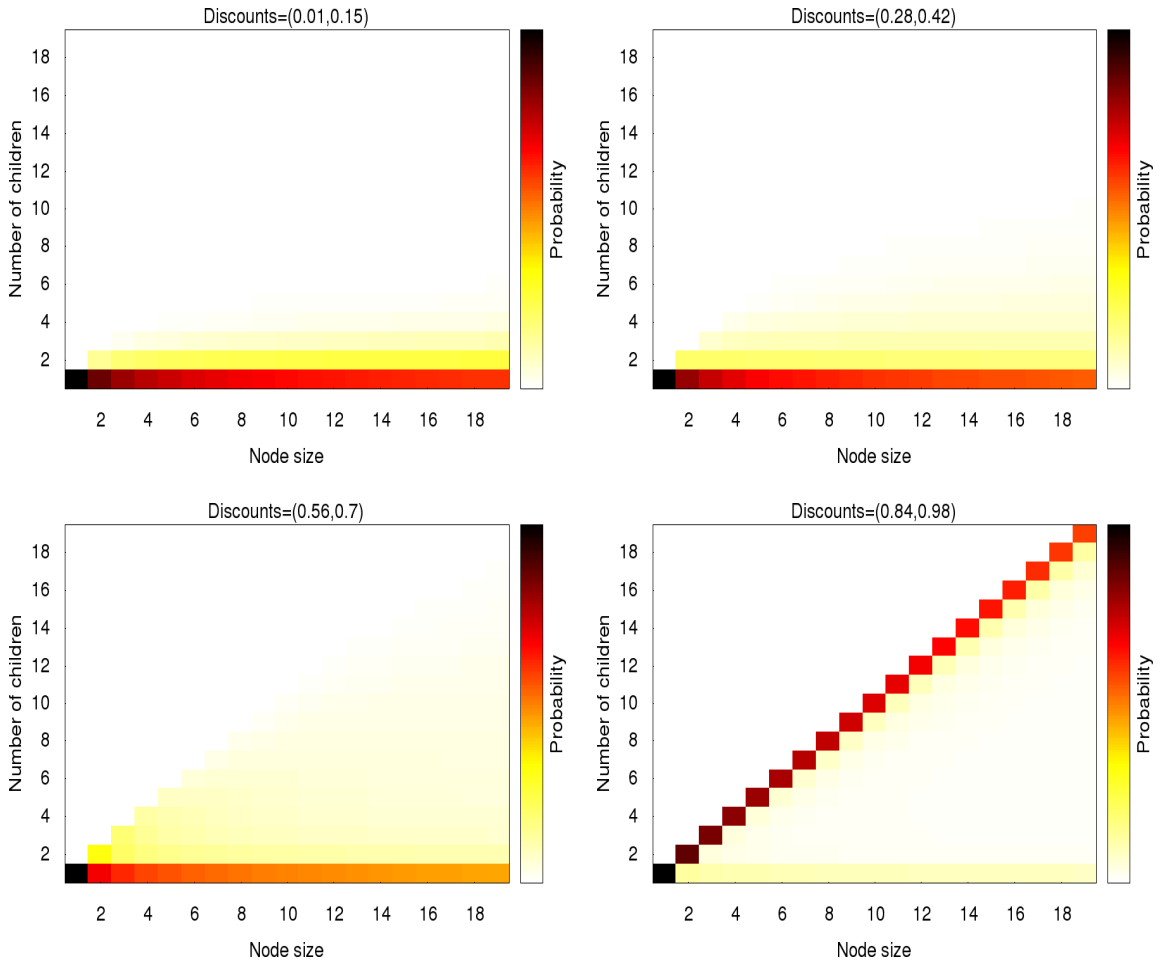

Figure 10: Distribution over the number of children for internal nodes based on discount schedule, thus using CRD $(node, a_{D+1}, -a_D)$.

- the partition applied to a single node (excepting the root node) at level D is CRD $(node, a_{D+1}, -a_D)$.

The expected partition sizes can then be inferred from Lemma 18. For instance, if $a_1 = 0$, then the first level below the root will have on average approximately $b \log(1 + N/b)$ nodes. Subsequent node partitioning will occur independently of b as the subsequent CRD parameters are set by the discount schedule only. This is illustrated in Figure 10. For instance, when splitting at a level where the discount is 0.56 and the discount for the next level down is 0.70, then look at the third plot with discounts labelled $(0.56, 0.70)$ which represents applying a CRD$(node, 0.70, -0.56)$. By the orange colour, about 50% of nodes of size 10 remain unpartitioned, but the remained are split into 2-8 say children. Whereas for the fourth plot with discounts labelled $(0.84, 0.98)$, 80% of nodes are maximally partitioned into single element nodes (leaves) and some small percentage will remain unpartitioned. One can see from this that by judicious use of a discount schedule, one can generate trees that

split with increasing propensities at lower levels.

2 [illegible]

Fragmentation and coagulation also apply to the distributions on infinite partitions represented by PDDs (or GEMs) and their counterpart the PDP.

The basic construction is most easily understood in terms of the basic summation form of Formula (2). We start with a simple draw from a PDP, $\sum_{k=1}^{\infty} p_k \delta_{X_k^*}(\cdot)$. However, we now replace the draws from the base distribution X_k^* by another draw from a PDP with probability vector $\vec{q}_k$, so we get

$$\sum_{k=1}^{\infty} p_k \sum_{j=1}^{\infty} q_{k,j} \delta_{X_{k,j}^*}(\cdot) \;=\; \sum_{k,j=1}^{\infty} (p_k q_{k,j}) \delta_{X_{k,j}^*}(\cdot) \;. \tag{6}$$

In some circumstances the terms $p_k q_{k,j}$ can be shown to follow a PDD distribution. The general result [Pit06] reworded for simplicity is as follows:

[illegible] *For $0 < a_1 < 1$, $0 \le a_2 < 1$ and $b > -a_1 a_2$, if $\vec{p} \sim PDD(a_1 a_2, b)$ and $\vec{q}_k \sim PDD(a_1, -a_1 a_2)$ for each $k = 1, ..., \infty$ then a sort of the resultant $\{p_k q_{k,j} : j, k = 1, ..., \infty\}$ is $PDD(a_1, b)$.*

[illegible] *Let $H(\cdot)$ be a distribution over some measurable space $\mathcal{X}$. For $0 < a_1 < 1$, $0 \le a_2 < 1$ and $b > -a_1 a_2$ introduce a latent distribution $Q(\cdot)$*

$$Q(\cdot) \;\sim\; PDP(a_1 a_2, b, PDP(a_1, -a_1 a_2, H(\cdot))) \;.$$

A sample from $R(\cdot)$ is taken by first drawing a sample $Q(\cdot)$ by the nested PDP above, and then taking a sample from $Q(\cdot)$. Then it follows that

$$R(\cdot) \;\sim\; PDP(a_1, b, H(\cdot)) \;.$$

Coagulation is the inverse of fragmentation so one follows a similar explanation. Start with a simple draw from a PDP, $\sum_{k=1}^{\infty} p_k \delta_{Y_k^*}(\cdot)$. However, we now wish to cluster some of the values $\delta_{Y_k^*}(\cdot)$. To do this we obtain a second draw from another PDP, $Q(\cdot) = \sum_{j=1}^{\infty} q_j \delta_{X_j^*}(\cdot)$, and replace each Y_k^* by a sample $X_{j_k}^* \sim Q(\cdot)$. So we get

$$\sum_{k=1}^{\infty} p_k \delta_{X_{j_k}^*}(\cdot) \;=\; \sum_{j=1}^{\infty} \left(\sum_{k\,:\,j_k = j} p_k \right) \delta_{X_j^*}(\cdot) \;. \tag{7}$$

In some circumstances the terms $\sum_{k\,:\,j_k=j} p_k$ can be shown to follow a PDD distribution. The general result [Pit06] again is as follows:

[illegible] *For $0 < a_1 < 1$, $0 \le a_2 < 1$ and $b > -a_1 a_2$, if $\vec{p} \sim PDD(a_1, b)$ and $\vec{q} \sim PDD(a_2, b/a_1)$ and for $k = 1, ..., \infty$, $j_k \sim \vec{q}$, then a sort of the resultant $\left\{ \sum_{k\,:\,j_k=j} p_k \;:\; j = 1, ..., \infty \right\}$ is $PDD(a_1 a_2, b)$.*

[illegible] *Let $H(\cdot)$ be a distribution over some measurable space $\mathcal{X}$. For $0 < a_1 < 1$, $0 \leq a_2 < 1$ and $b > -a_1 a_2$, sample a distribution $R(\cdot)$ using a latent distribution $Q(\cdot)$ as follows:*

$$Q(\cdot) \sim PDP(a_2, b/a_1, H(\cdot)) \qquad R(\cdot) \sim PDP(a_1, b, Q(\cdot)) \ .$$

Marginalising out $Q(\cdot)$, it follows that

$$R(\cdot) \sim PDP(a_1 a_2, b, H(\cdot)) \ .$$

The intrinsic nature of fragmentation and coagulation is typified by Equations (6) and (7) respectively and they demonstrate a duality: for $0 \leq a_2 < a_1 < 1$ and $b >= -a_2$,

$$\mathrm{PDD}(a_1, b) \underset{coagulation}{\overset{fragmentation}{\rightleftharpoons}} \mathrm{PDD}(a_2, b)$$

Moreover, fragmentation is an operator for simplifying nested PDPs whereas coagulation is for simplifying hierarchical PDPs (as used in [WAG+09]).

7 [illegible]

A distribution or prior is called *proper* if it integrates (or sums) to one. The Bayesian theory of *improper priors* allows one to extend the space of reasonable priors. The idea is that if the posteriors from the prior are always proper, then perhaps one can represent the improper prior as a sequence of proper priors. The limit of this sequence may not be proper, but at least its posteriors all are. In this section we develop an improper prior that corresponds to the PDD.

With any L_d distance for $d \geq 1$, the infinite-dimensional probability vector $\vec{p}$ of Formula (2) defines a Hilbert space[1]. It is difficult to define a prior probability on such a space because not only does one require a measure be defined for the infinite vector, it must be normalised, so the total measure is 1. Some theories just give priors for finite linear projections of the full Hilbert space, for instance the cylindrical measures of Minlos [Min01]. This is sufficient according to Carathodory's extension theorem, see Bogachev [Bog07], to define the prior on the full space[2]. For the PDD model, an additional problem is the projections of the prior on finite vector subspaces appear to be improper as well. Thus, the best one can do is define a prior in terms of a measure for all finite sub-vectors as follows:

[1]Only when $d \geq 1$ is the subsequent distance guaranteed to be finite for any two members of the space.

[2]The cylinders form a semi-ring, and we have a countably additive (pre-)measure on the semi-ring, this implies a unique extension on the generated ring, the sigma-algebra is generated by the cylinder sets, and Caratheodory's extension theorem shows that there exists a unique extension of the (pre-)measure to the sigma-algebra.

Definition 28 *Given parameters (a, b), where $0 \leq a < 1$ and $b > -a$, define the* improper prior for PDDs *(an unnormalised measure) as follows. Take any reordering of the infinite-dimensional probability vector $\vec{p}$, and then for every sub-vector $p_1, p_2, ..., p_M$ of the reordering, use the following measure:*

$$p(p_1, p_2, ..., p_M, p_M^+) \ := \ \left(p_M^+\right)^{b+M\,a-1} \prod_{m=1}^{M} p_m^{-a-1} \ ,$$

where $p_M^+ = 1 - \sum_{m=1}^{M} p_m$.

Note this applies to every sub-vector, so ordering of the probabilities is not needed as in Definition 5. The measure $p(p_1, p_2, ..., p_M, p_M^+)$ in the definition is an instance of an $M+1$-dimensional improper Dirichlet with parameters $(-a, -a, ..., -a, b+M\,a)$, denoted here informally as

$$\text{Dirichlet}_{M+1}(-a, -a, ..., -a, b + M\,a) \ .$$

Moreover, note that we believe this measure has no corresponding limit form as $M \to \infty$ on the full infinite-dimensional probability vector $\vec{p}$. Given an improper prior measure, one can infer a posterior measure using an unnormalised version of Bayes theorem. If the posterior measure can be normalised, then the posterior is now a correct probability.

It is shown next that the definition is consistent in the sense that the measures for different sub-vectors are natural extensions of one another. This property is called additivity for proper Dirichlets and is well-known. It is plausible that it should hold for the improper measure too, but the standard proofs cannot be transferred since the involved integrals no longer exist. Here we check additivity does transfer to improper Dirichlets.

Theorem 29 *In the context of Definition 28, if the prior measure for $p_1, ..., p_M$ is projected down to some sub-vector, say $p_1, ..., p_L$ for $L < M$, then the projected measure is consistent with Definition 28.*

We must now show the improper prior for PDDs is well defined. This is done using the L_1 (total variation) distance defined for probability density functions $H(\cdot)$ and $G(\cdot)$ as follows

$$L_1(H, G) \ = \ \int_{\vec{p}} |H(\vec{p}) - G(\vec{p})| \, \mathrm{d}\, \vec{p} \ . \tag{8}$$

The theorem below says that a sequence of proper priors exist that can approximate the improper prior for PDDs arbitrarily closely in the sense that their posteriors given any sequence sample can be made arbitrarily close to the corresponding proper posterior of the improper prior. Closeness here is measured by total variation distance.

Theorem 32 *Using the notation of Definitions 28 and 8, there exists a set of proper priors G_δ for $\delta > 0$ such that for any $\epsilon > 0$ and any sample I_N there exists a δ_0 such that for all $0 < \delta < \delta_0$ the proper posterior (I) given I_N^* of Lemma 31 is within ϵ by the L_1 distance of the posterior of G_δ given I_N^*.*

Because the improper prior is well defined, one can justifiably obtain posteriors and sampling results from the prior. Now these are identical to those for the PDD and PDP, as we detail below, however they were derived from the improper prior, not from any of the standard definitions for PDDs or PDPs.

Theorem 33 *Using the improper prior for PDDs with parameters (a, b) and non-atomic base distribution $H(\cdot)$, the following holds:*

(I) *Using the notation of Definitions 8 and 28. The posterior distribution given I_N^* is*

$$\left(p_1, ..., p_M, p_M^+\right) \mid I_N^* \;\sim\; \mathit{Dirichlet}(n_1 - a, ..., n_M - a, b + Ma) \;, \tag{9}$$

where $p_M^+ = 1 - \sum_{m=1}^M p_m$.

(II) *Using the notation of Definition 9, in the case of arbitrary samples S_N, the posterior distribution given S_N is*

$$\begin{aligned} X_{N+1} \mid p_1, ..., p_M, p_M^+, S_N &\;\sim\; p_M^+ \, H(\cdot) + \sum_{m=1}^M p_m \delta_{X_m^*}(\cdot) \\ \left(p_1, ..., p_M, p_M^+\right) \mid S_N &\;\sim\; \mathit{Dirichlet}(n_1 - a, ..., n_M - a, b + Ma) \;, \end{aligned} \tag{10}$$

where $p_M^+ = 1 - \sum_{m=1}^M p_m$.

(III) *Using the notation of Definition 9, if we marginalise out the probability vector $\vec{p}$, then the posterior distribution in the next sample X_{N+1}, $p(X_{N+1} \mid S_N)$, is*

$$\frac{b + Ma}{b + N} \, H(\cdot) + \sum_{m=1}^M \frac{n_m - a}{b + N} \delta_{X_m^*}(\cdot) \;.$$

(IV) *A stick-breaking like construction holds for the posteriors (I) and (II) above. That is, for $1 \leq m \leq M$*

$$p_m \;=\; V_m \prod_{i=1}^{m-1} (1 - V_i)$$

where each V_m is independent $\mathit{Beta}\left(n_m - a, b + ma + \sum_{i=m+1}^M n_i\right)$. Since the $\left(n_m, \sum_{i=m+1}^M n_i\right)$ are count terms, one can say each V_m has an improper prior $\mathit{Beta}(-a, b + ma)$.

Lemma *The prior on a size-biased ordering from the improper prior for PDDs with parameters (a, b) follows a GEM(a, b).*

The posterior formulation for PDPs corresponding to Equation (10) is attributed to Pitman [Pit96] by Ishwaran and James [IJ01, Section 4.4]. The sampling result is the standard Chinese Restaurant Process for the PDP from Ishwaran and James [IJ01, Section 2.2]. The stick-breaking result here is different to the standard PDP [IJ01, Section 2.1], which has stick priors Beta$(1 - a, b + ma)$ (that is, it is proper), see [PY97]. Here we use improper priors Beta$(-a, b+ma)$, which matches the sampling of the CRP as described above.

Now the PDD(a, b) distribution is defined with sorting, whereas the improper prior for PDDs with parameters a, b is not. Therefore they do not correspond directly unless some sorting is done. So we can say that sorting the $\vec{p}$ for an improper prior for PDDs yields a PDD(a, b) distribution.

8 [illegible]

Now consider the case of discrete base distributions. In this case, $H(\cdot)$ is a probability function, not a probability density function, so for each sample X_k from $H(\cdot)$ its probability is finite and thus identical draws can be repeated. This makes the evidence calculation of Lemma 14 invalid whenever the PDP is used in discrete or hierarchical contexts. Here we present the techniques used to get around this.

If we have not been given the index sequences for a sample from an impulse mixture model, we can only guess what the indices might be. In this case, the detail of the data partition is partially hidden. So for instance, consider the sample of words:

"from","apple","to","from","from","cat","to",...

This can have a size-biased ordered index sequence of $I_N^* = 1, 2, 3, 1, 1, 4, 3$. However, since $H(\cdot)$ is discrete, it could be that the three instances of "from" come from different indices in Formula (2). Some other size-biased orderings of indices compatible with this sequence of words are as follows:

1, 2, 3, 1, 1, 4, 3, ..., 1, 2, 3, 1, 4, 5, 3, ...,
1, 2, 3, 1, 1, 4, 5, ..., 1, 2, 3, 4, 5, 6, 3, ...,

We are unable to say which is correct, however, they have a single coarsest version. Thus we introduce additional latent variables to account for the uncertainty, introduced next.

8 [illegible]M

The definition below, *multiplicity*, measures the cardinality of the (unknown) set of indices contributing to one observation X that occurs multiple times in the data. In

Word map	Index sequence I_N	Multiplicities $\vec{t}$
"from","apple","to","cat"	1, 2, 3, 1, 1, 4, 3	1,1,1,1
"from","apple","to","from","cat"	1, 2, 3, 1, 4, 5, 3	2,1,1,1
"from","apple","to","cat","to"	1, 2, 3, 1, 1, 4, 5	1,2,1,1
"from","apple","to","from","from","cat"	1, 2, 3, 4, 5, 6, 3	3,1,1,1

Table 2: Multiplicities from different index sequences of the sequence $S_7 =$ {"from","apple","to","from","from","cat","to"}.

the Chinese restaurant analogy, the multiplicity for the data value X is the number of active tables with the particular menu item X.

Definition 9 *Consider Definition 9, but now assume that the base distribution $H(\cdot)$ is discrete. For a given sample S_N, let I_N be a size-biased ordered index sequence matching S_N, and assume they are represented as $S_N = (X_1, ..., X_N)$ and $I_N = (k_1, ..., k_N)$. The* multiplicity *of the value $X \in S_N$ is defined as the size of the set $\{k_n : n = 1, ..., N,\ X_n = X\}$.*

Multiplicities are statistics from the latent indices I_N and are thus themselves latent. Continuing the words example at the start of Section 8, some of the potential size-biased ordered index sequences are illustrated in Table 2. The first column gives the map from index to word, the second column gives the resultant index sequence, and the third column lists the multiplicities for the words "from","apple","to","cat" respectively (*i.e.*, in size-biased order). Note, for instance, in the last row, the word "from" appears 3 times in the map and thus has multiplicity 3.

For the discrete base distribution, we must consider the situation where the multiplicities can be greater than one, so a more general probability of evidence result is needed, for instance for Lemma 14, since PDP(a, b, H) returns values from $\mathcal{X}$, but no indices. The following corollary of Lemmas 14 and 16 is a special case of [Teh06a, Equation (31)], there proven directly for the hierarchical PDP.

Corollary 17 *Consider the probability of evidence for a finite sample $X_1, X_2, ..., X_N$ from PDP(a, b, H) with discrete base distribution $H(\cdot)$. Use Definition 9, and let t_m be the latent multiplicity of X^*_m in the sample, and let their total $\sum_{m=1}^{M} t_m = T$. Note they must satisfy the constraints $0 \le t_m \le n_m$ and $t_m = 0$ if and only if $n_m = 0$. Then the joint probability of the sample and the multiplicities is:*

$$p(X_1, X_2, ..., X_N, t_1, ..., t_M \,|\, a, b, H(\cdot)) \;=\; \frac{(b|a)_T}{(b)_N} \prod_{m=1}^{M} \left(H(X^*_m)^{t_m} \, S^{n_m}_{t_m,a} \right) \;,$$

where $S^N_{M,a}$ is defined in (5).

Notice if one is Gibbs sampling with the latent multiplicites t_m, then one needs to sample t_m for all values from 1 up to n_m. This can be a problem if $n_m = 1000000$.

8 [illegible]

A second representation developed in [CDB11] stores a table indicator for each data item in a way that makes it exchangeable. For this, let the table indicator r_n be zero if the data X_n does not contribute a new table, and one if it does contribute a new table. The multiplicity t_k for the data value X_k^* is then computed as the sum for those with the same data value, so $t_k = \sum_{n=1}^{N} r_n 1_{X_n = X_k^*}$. Since we are invariant as to which of the data starts a table, and there are $C_{t_k}^{n_k}$ choices, the above posterior is modified to yield:

[illegible] *Following the situation of Corollary 33,*

$$p(X_1, X_2, ..., X_N, r_1, ..., r_N \,|\, a, b, H(\cdot)) \;=\; \frac{(b|a)_T}{(b)_N} \prod_{m=1}^{M} \left(H(X_m^*)^{t_m} S_{t_m,a}^{n_m} \frac{1}{\binom{n_m}{t_m}} \right) \;, \quad (11)$$

where the t_m are derived from the indicators r_n.

Note the table indicators do not appear explicitly in this other than through t_m and thus we can "forget" the table indicators and randomly resample them as needed at any stage of Gibbs sampling since by symmetry their probability of being 1 is t_m/n_m.

This representation has two major advantages: one only needs to incrementally change the t_m, not explore all possible values between 1 and n_m, and it only requires the use of ratios of Stirling numbers (for instance $S_{t_m+1,a}^{n_m}/S_{t_m,a}^{n_m}$) which can be easier to compute.

8 [illegible]

Useful quantities to understand the application of the PDP to a discrete base distribution, especially for the hierarchical case, are its moments. We give them here so we can properly interpret the discrete case.

[illegible] *Assume the discrete base distribution $H(\cdot)$ is over the integers $\mathbb{N}$, with probability vector $\vec{\theta}$, so there is probability θ_k for the value k. Let $\vec{p} \sim PDP(a, b, H)$. Then the mean, variance, covariance and third order moments of $\vec{p}$ according to this prior are given by*

$$\begin{aligned} \mathbb{E}[\vec{p}] &= \vec{\theta}\,. \\ Var[p_k] &= \frac{1-a}{b+1}\theta_k(1-\theta_k) \\ Cov[p_{k_1}, p_{k_2}] &= -\frac{1-a}{b+1}\theta_{k_1}\theta_{k_2} \end{aligned}$$

$$\mathbb{E}\left[(p_{k_1}-\theta_{k_1})(p_{k_2}-\theta_{k_2})(p_{k_3}-\theta_{k_2})\right]$$
$$= \begin{cases} 2\frac{(1-a)(2-a)}{(b+1)(b+2)}\theta_{k_1}\theta_{k_2}\theta_{k_3} & \textit{when } k_1,k_2,k_3 \textit{ disjoint} \\ \frac{(1-a)(2-a)}{(b+1)(b+2)}(2\theta_{k_1}-1)\theta_{k_1}\theta_{k_2} & \textit{when } k_1=k_2\neq k_3 \\ \frac{(1-a)(2-a)}{(b+1)(b+2)}\theta_{k_1}(1-\theta_{k_1})(1-2\theta_{k_1}) & \textit{when } k_1=k_2=k_3 \end{cases}$$

Now consider the case where $H(\cdot)$ has domain $1,...,K$, and probability vector $\vec{\theta}$. Denote this by discrete($\vec{\theta}$). Consider a K dimensional Dirichlet distribution with parameters given by $\alpha\vec{\theta}$. This has corresponding moments

$$\begin{aligned} \mathbb{E}\left[\vec{p}\right] &= \vec{\theta}\,. \\ \text{Var}\left[p_k\right] &= \frac{1}{\alpha+1}\theta_k(1-\theta_k) \\ \text{Cov}\left[p_{k_1},\, p_{k_2}\right] &= -\frac{1}{\alpha+1}\theta_{k_1}\theta_{k_2} \end{aligned}$$

$$\mathbb{E}\left[(p_{k_1}-\theta_{k_1})(p_{k_2}-\theta_{k_2})(p_{k_3}-\theta_{k_3})\right]$$
$$= \begin{cases} \frac{4}{(\alpha+1)(\alpha+2)}\theta_{k_1}\theta_{k_2}\theta_{k_3} & \text{when } k_1,k_2,k_3 \text{ disjoint} \\ \frac{2}{(\alpha+1)(\alpha+2)}(2\theta_{k_1}-1)\theta_{k_1}\theta_{k_2} & \text{when } k_1=k_2\neq k_3 \\ \frac{2}{(\alpha+1)(\alpha+2)}\theta_{k_1}(1-\theta_{k_1})(1-2\theta_{k_1}) & \text{when } k_1=k_2=k_3 \end{cases}$$

Thus we can conclude following: When $0 < a \ll 1$, then we have that PDP$(a, b, \text{Categorical}(\vec{\theta}))$ is approximated by a Dirichlet $\left(\frac{a+b}{1-a}\vec{\theta}\right)$ (and it is already known equality holds when $a = 0$). The two distributions differ by a factor of $O(a^2)$ in all the moments of order one to three. Thus the PDP applied to finite discrete distributions is approximated by a proper Dirichlet.

8 [illegible]

To work with the discrete case, one needs to sample the multiplicities t_m. These can be sampled using Gibbs sampling and precomputed tables of the Stirling numbers $S^n_{t,a}$. When used in sampling, the Stirling numbers $S^n_{t,a}$ need to be tabulated or cached for the required discount parameter a. Because they rapidly become very large, they need to be stored and computed in log format. The recursion must be used rather than the exact formulation, and becomes

$$\log S^{n+1}_{t,a} = \log S^n_{t,a} + \log\left(\exp\left(\log S^n_{t-1,a} - \log S^n_{t,a}\right) + (n - t\,a)\right)\,. \tag{12}$$

The log() and exp() functions can make the evaluation slow if implemented naively. Moreover, memory requirements for the full table of $S^n_{t,a}$ with a fixed discount a is $n(n+1)/2$ floats for $t \leq n$. When keeping the discount a fixed, we can reduce memory with two tricks: (1) placing a maximum value on t say 100 or 1000 to limit the cache and (2) striping the cache for higher values of t, so only every L-th value is stored, entries of $S^n_{t,a}$ for $n = N, N+L, N+2L, ...$ and $t < n$. Computational time in this case becomes $O(2^L/L)$. These two techniques save considerable memory at a small factor in computational time or sampling accuracy.

Table 3: Comparative values for ratio versus log calculations for $a = 0.5$

(n, t)	double $V^n_{t,a}$	float $V^n_{t,a}$	float $\exp\left(\log S^n_{t,a} - \log S^n_{t-1,a}\right)$
(10000,10)	0.222133	0.222124	0.180696
(10000,100)	0.0201025	0.0201025	0.0262359
(10000,1000)	0.00189684	0.00189685	0.00181349

5 [illegible]

When sampling with the table indicator representation, one repeatedly needs ratios of Stirling numbers. Denote the Stirling number ratios

$$U^n_{t,a} = \frac{S^{n+1}_{t,a}}{S^n_{t,a}} \qquad V^n_{t,a} = \frac{S^n_{t,a}}{S^n_{t-1,a}} \; . \tag{13}$$

These have the advantage that they do not need to be stored in log space. The first ratio, $U^n_{t,a}$, is readily computed from the second, $V^n_{t,a}$, so is not stored:

$$\begin{aligned} U^n_{1,a} &= n - a && \text{for } n \geq 1 \\ U^n_{t,a} &= \frac{1}{V^n_{t,a}} + (n - t\,a) && \text{for } n \geq t > 1 \; . \end{aligned}$$

The second formula follows directly from the linear recursion for $S^n_{t,a}$. The second ratio, $V^n_{t,a}$, has the following recursion

$$\begin{aligned} V^n_{n,a} &= \frac{1}{U^{n-1}_{n-1,a}} && \text{for } n \geq 2 \\ V^{n+1}_{t,a} &= \frac{1}{U^n_{t-1,a}} \left(1 + (n - t\,a) V^n_{t,a}\right) && \text{for } n \geq t \geq 2 \; . \end{aligned} \tag{14}$$

Thus a simple recursion yields tables for $V^n_{t,a}$ from which $U^n_{t,a}$ can be computed without any resort to the log storage of the Stirling numbers or the use of transcendental functions.

Moreover, these ratios yield more accurate computation because the direct log computation of the Stirling numbers leads to some loss of precision. Values in Table 3 computed with discount parameter $a = 0.5$ presents the results using floating point (32 bit) calculation versus the double precision results (64 bit) as a proxy for evaluating round-off error. One can see that the linear recursion in log space, Equation (12), yields considerably less accurate results than the ratio recursion of Equation (14) when computation is done with floats rather than doubles. Computation using the ratios can also be upto 5 times faster.

9 [illegible]

For the non-atomic case of the two parameter Poisson-Dirichlet distribution, consistency, convergence and posterior results have been presented, mostly drawn from the

literature, though some proofs are given in the Appendix. We have augmented these results with a number of plots to illustrate the nature of the underlying distributions. Most significantly, in practice we recommend fitting one of the two parameters a or b of the PDP or PDD when performing inference with them.

[illegible] The distribution on partitions induced by the two parameter Poisson-Dirichlet distribution, called here the Chinese Restaurant distribution, has also been presented, together with a summary of its use in a scheme for generating trees. The fragmentation and coagulation duality for the Chinese Restaurant distribution and the two parameter Poisson-Dirichlet distribution followed: coagulation allows simplification of some hierarchical Poisson-Dirichlet distributions, and fragmentation allows simplification of some nested Poisson-Dirichlet distributions.

[illegible] The infinite probability vector underlying the Poisson-Dirichlet distribution was shown to be in the form of an improper Dirichlet on any finite sub-vector. As soon as data is presented, the posterior becomes proper (or normalised). Alternatively, if a size-biased ordering of the probabilities in the infinite vector is made, the knowledge implicit in making a size-biased order turns the improper Dirichlet into a GEM distribution, the standard "stick-breaking" distribution most commonly used to define a Poisson-Dirichlet distribution.

[illegible] For the discrete case, not well covered in the Probability and Statistics literature, posterior results have also been presented. Moreover, it has been shown that the two parameter Poisson-Dirichlet Process with discount $a > 0$ and concentration b on a discrete base distribution behaves rather like a Dirichlet distribution with concentration $\frac{a+b}{1+a}$. The Dirichlet Process on a discrete base distribution is identical to a Dirichlet distribution.

[illegible] While the posteriors presented for the discrete case of the PDP introduced latent values, multiplicities or indicators, they are conjugate to the multinomial family. This means the Poisson-Dirichlet Process and the Dirichlet Process can be used hierarchically and remarkably a form of semi-conjugacy applies (*i.e.*, conjugacy at the expense of introducing latent variables).

[illegible] Computation of the Stirling numbers needed to perform Gibbs sampling was also presented. The closed form for computing Stirling numbers looks like a difference approximation to an M-th order derivative, thus it is intrinsically unstable to compute for higher values of M. The standard linear recursion therefore seems necessary, although double precision is needed for larger values. Ratios of Stirling numbers, however, can be more accurately computed.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

[Ant74] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover publications, 1974.

[Bog07] V. I. Bogachev. *Measure Theory*. Springer, 2007.

[CDB11] C. Chen, L. Du, and W. Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Athens/Greece, September 2011.

[Fer73] T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

[GGJ06] S. Goldwater, T. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 459–466. MIT Press, Cambridge, MA, 2006.

[GR01] P.J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandanavian Journal of Statistics*, 28:355–375, 2001.

[HS88] L. Hsu and P. Shiue. A unified approach to generalized Stirling numbers. *Advances in Applied Mathematics*, 20:366–384, 1988.

[IJ01] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of ASA*, 96(453):161–173, 2001.

[IJ03] H. Ishawaran and L.F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211–1235, 2003.

[Jam08] L.F. James. Large sample asymptotics for the two-parameter Poisson-Dirichlet process. In B. Clarke and S. Ghosal, editors, *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3. IMS, 187–199, 2008.

[JGG07] M. Johnson, T.L. Griffiths, and S. Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA, 2007.

[JLP09] L.F. James, A. Lijoi, and I. Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36:76–97, 2009.

[KWT07] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 20, 2007.

[Min01] R.A. Minlos. Cylindical measure. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, 2001.

[MS08] D. Mochihashi and E. Sumita. The infinite Markov model. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1017–1024. MIT Press, Cambridge, MA, 2008.

[Pit95] J. Pitman. Exchangable and partially exchangeable random partitions. *Probab. Theory Relat. Fields*, 102:145–158, 1995.

[Pit96] Jim Pitman. Some developments of the Blackwell-Macqueen urn scheme. In *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30, pages 245–267. Institute of Mathematical Statistics, 1996.

[Pit99] J. Pitman. Brownian motion, bridge, excursion, and meander characterized by sampling at independent uniform times. *Electronic Journal of Probability*, 4:paper 11, 1999.

[Pit06] J. Pitman. *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[PY97] J. Pitman and M. Yor. The two-parameter Poisson-Diriclet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[Ras00] C.E. Rasmussen. The infinite Gaussian mixture model. In S.A. et al. Solla, editor, *Advances in information processing systems 12*, pages 554–560. MIT Press, 2000.

[SJ09] E.B. Sudderth and M. Jordan. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*. MIT Press, 2009.

[Teh06a] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.

[Teh06b] Y.W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 985–992, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[Tos39] L. Toscano. Numeri di stirling generalizzati operatori differenziali e polinomi ipergeometrici. *Commentationes Pontificia Academica Scientarum*, 3:721–757, 1939.

[TZF08] D. Tarlow, R. Zemel, and B. Frey. Flexible priors for exemplar-based clustering. In *Uncertainty in Artificial Intelligence (UAI)*, 2008.

[WAG+09] F. Wood, C. Archambeau, J. Gasthaus, L.F. James, and Y.W. Teh. A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*, 2009.

[WSM08] H. Wallach, C. Sutton, and A. McCallum. Bayesian modeling of dependency trees using hierarchical Pitman-Yor priors. In *Proceedings of the Workshop on Prior Knowledge for Text and Language (in conjunction with ICML/UAI/COLT)*, pages 15–20, 2008.

[XTYK06] Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, Cambridge, MA, USA, July 2006.

[YS00] H. Yamato and M. Sibuya. Moments of some statistics of Pitman sampling formula. *Bulletin of Informatics and Cybernetics*, 32:1–10, 2000.

A F

A F

A F

Build on the result from Lemma 14 using Definition 8. The formula of Lemma 14 also applies to I_N^*, the indices I_N with size-biased ordering applied, so $p(I_N^*)$, but remove the terms in $H(X_k^{\#})$. Using the notation of Definition 8 and this lemma, we get the form

$$p(I_N^*) \;=\; \frac{(b|a)_M}{(b)_N} \prod_{m=1}^{M} \frac{\Gamma(n_m - a)}{\Gamma(1-a)} \;. \tag{15}$$

Now one can marginalise out the entires in I_N^* but keeping the constraint that there are M distinct k's in there, which will affect the last product of M terms only.

The indexes $1, ..., M$ occur in the sequence I_N^* of size N. Ignoring the ordering constraints of size-biased ordering, there are N choose n_1, ..., n_M, $C^N_{n_1,\ldots,n_M}$ ways the indexes can occur in I_N^*. Now adjust this for the ordering constraints. For every sequence starting with 1 there exists some starting with 2, ..., M. By symmetry, n_1/N of the sequences start with 1. Now how many of these have the second integer appearing in sequence being 2? Again by symmetry, $n_2/(N-n_1)$ of the sequences starting with 1 have 2 as the next integer in sequence. Likewise, of those sequences with 1, 2 being the first two occurring integers respectively, $n_3/(N-n_1-n_2)$ have 3 occurring next. Thus, the number of sequences by size-biased ordering with counts n_1, ..., n_M are

$$C^N_{n_1,\ldots,n_M} \prod_{m=1}^{M} \frac{n_m}{N - \sum_{i=1}^{m-1} n_i} \;.$$

Inspection shows this evaluates to an integer since each term $N - \sum_{i=1}^{m-1} n_i$ divides into $N!$.

To marginalise out the indexes I_N^* in (15) then, one does

$$\begin{aligned} p(M|N) \;&=\; \sum_{\sum_{m=1}^{M} n_m = N,\; n_m \geq 1} p(I_N^*)\; C^N_{n_1,\ldots,n_M} \prod_{m=1}^{M} \frac{n_m}{N - \sum_{i=1}^{m-1} n_i} \;. \\ &=\; \frac{(b|a)_M}{(b)_N} \; N! \sum_{\sum_{m=1}^{M} n_m = N,\; n_m \geq 1} \prod_{m=1}^{M} \left(\frac{\Gamma(n_m - a)}{\Gamma(n_m+1)\Gamma(1-a)} \frac{n_m}{N - \sum_{i=1}^{m-1} n_i} \right) \;. \end{aligned}$$

The full summation formula for $S^N_{M,a}$ follows.

A F

We need the following expressions for generalized Stirling numbers. All but the explicit expression (iii) are due to [HS88].

Theorem 6 *The following expressions all define the same generalized Stirling numbers $S(n,k;\alpha,\beta,r)$, where the parameters $\alpha,\beta,r \in \mathbb{R}$ have been suppressed when constant.*

(o) Implicit: $$(t|-\alpha)_n = \sum_{k=0}^{n} S(n,k)(t-r|-\beta)_k$$

Both sides are polynomials in t of degree n. $(z|a)_n := z(z+a)...(z+(n-1)a)$.

(i) Linear recursion: $S(n+1,k) = S(n,k-1) + (k\beta - n\alpha + r)S(n,k)$

Boundary cond.: $S(n,k) = 0$ *for* $k > n$, $\quad S(n,0) = (r|-\alpha)_n$

(ii) Mult. recursion: $\binom{N}{K} S(N,K,\alpha,\beta,R) =$ *(any k,r)*

$$= \sum_{n=0}^{N} \binom{N}{n} S(n,k;\alpha,\beta,r) S(N-n,K-k;\alpha,\beta,R-r)$$

(iii) Explicit expression: $$S(n,k) = \frac{1}{k!\,\beta^k} \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} (\beta j + r|-\alpha)_n \qquad (\beta \neq 0)$$

(iv) Generative fct.: $$\sum_{n=0}^{\infty} S(n,k)\frac{t^n}{n!} = \frac{(1+\alpha t)^{r/\alpha}}{k!} \left(\frac{(1+\alpha t)^{\beta/\alpha} - 1}{\beta} \right)^k \qquad (\alpha\beta \neq 0)$$

Proof The generalized Stirling numbers are defined in [HS88] by (o). Hsu and Liu derive expressions (i),(ii), and (iv) from (o): It is easy to verify that recursion (i) satisfies definition (o). Using (i), one can see that the generating function $g_k(t) := \sum_{n=0}^{\infty} S(n,k)t^n/n!$ satisfies the differential equation system

$$(1+\alpha t)\frac{d}{dt} g_k(t) = g_{k-1}(t) + (k\beta + r) g_k(t) \quad \text{with} \quad g_k(0) = 0 \quad \text{and} \quad g_0(t) = (1+\alpha t)^{r/\alpha},$$

which has a unique solution. Substituting (iv) into this dgl shows that (iv) is a solution. If we take a product of the generating functions of $S(n,k;\alpha,\beta,r)$ and $S(N-n,K-k;\alpha,\beta,R-r)$, use (iv), and identify the coefficients of t^n in both sides, we arrive at the multiplicative recursion (ii).

Interestingly, Hsu and Liu do *not* derive the explicit expression (iii), although it easily follows by Taylor expanding the r.h.s. of (iv) and by identifying the coefficients of t^n as follows: The binomial identity gives

$$((1+\alpha t)^{\beta/\alpha} - 1)^k \;=\; \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} (1+\alpha t)^{\beta j/\alpha}$$

Exploiting this, (iv), and $(1+z)^\gamma = \sum_{n=0}^{\infty} \binom{\gamma}{n} z^n$, where $\binom{\gamma}{n} = \frac{\Gamma(\gamma+1)}{n!\,\Gamma(\gamma-n+1)}$, we get

$$\begin{aligned}
\sum_{n=0}^{\infty} S(n,k)\frac{t^n}{n!} &= \frac{1}{k!\,\beta^k} \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} (1+\alpha t)^{\frac{\beta j + r}{\alpha}} \\
&= \frac{1}{k!\,\beta^k} \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} \sum_{n=0}^{\infty} \binom{\frac{\beta j + r}{\alpha}}{n} (\alpha t)^n \\
&= \sum_{n=0}^{\infty} \frac{1}{k!\,\beta^k} \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} (\beta j + r|-\alpha)_n \frac{t^n}{n!}
\end{aligned}$$

A.1.4

The proof is based on (a) a recursion for $p(M|N)$, (b) the expressions for the generalized Stirling numbers in Appendix A.1.2, and of course (c) the definition (5) of $S^N_{M,a}$. In order to distinguish between different M as the sample size N increases, use M_N to denote the value at sample size N.

(i) We exploit recursion

$$p(M_{N+1} = m|M_N) \;=\; \mathbb{1}_{M_N=m-1}\frac{b+(m-1)a}{b+N} + \mathbb{1}_{M_N=m}\frac{N-ma}{b+N}\;,$$

which easily follows from the predictive sampling distribution (3). Multiplying each side by $p(M_N)$, and summing over M_N this becomes

$$p(M_{N+1} = m) \;=\; p(M_N = m-1)\frac{b+(m-1)a}{b+N} + p(M_N = m)\frac{N-ma}{b+N}$$

Inserting the explicit expression $p(M_N = m) = S^N_{m,a}(b|a)_m/(b)_N$ of Lemma 16 into this recursion and canceling all common factors we get

$$S^{N+1}_{m,a} \;=\; S^N_{m-1,a} + (N-ma)S^N_{m,a}.$$

The boundary conditions $S^N_{m,a} = 0$ for $m > N$ and $S^N_{0,a} = \delta_{N,0}$ follow from the explicit expression in Definition (5) or simply by reflecting on the meaning of $p(M_N = m)$.

(i) and (i) Consider the generalized Stirling numbers $S(n,k;\alpha,\beta,r)$ for the special choice of parameters $(\alpha,\beta,r) = (-1,-a,0)$. For this choice, recursion (i) of Theorem 36 reduces to recursion (i) of Theorem 17, including the boundary conditions. Hence $S^N_{M,a} = S(N,M;-1,-a,0)$.

It is easy to see that also (ii) and (iii) of Theorem 36 reduce to the first expression of (ii) and (iii) of Theorem 17 for $(\alpha,\beta,r) = (-1,-a,0)$, which shows that the expressions are equivalent.

The special case for (iii) when $a = 0$ holds by noting that (iii) for $a > 0$ is in fact an M-point interpolation to an $M-1$-th partial derivative. As $a \to 0$, this becomes the partial derivative.

The last expression in (ii) follows from the definition of $S^N_{M,a}$ in (5) by splitting the sum into $\sum_{n_1=1}^{N-M+1}$ and $\sum_{n_2+\ldots+n_M=N-n_1}$ and the product into $m = 1$ and $m > 1$, and identifying the terms with $\binom{N-1}{n_1-1}$, $S^{n_1}_{1,a}$ and $S^{N-n_1}_{M-1,a}$.

Note that the first expression in (ii) does *not* reduce to the second expression for $m = 1$. Nevertheless, the (very different!) derivations of the two expressions show that they must be equal.

(iv) Using $\Gamma(N+x)/\Gamma(N+y) \simeq N^{x-y}$ for large N, we see that the m-contribution in (iii) is asymptotically proportional to

$$\prod_{h=0}^{N-1}(h-am) \;=\; \frac{\Gamma(N-am)}{\Gamma(-am)} \;=\; \frac{-am\Gamma(N)}{\Gamma(1-am)}\frac{\Gamma(N-am)}{\Gamma(N)} \;\stackrel{N\to\infty}{\simeq}\; \frac{-am\Gamma(N)}{\Gamma(1-am)}\frac{1}{N^{am}}$$

Due to the factor m, the $m=0$ term does not contribute. So the dominant contribution is from $m=1$, followed by $m=2$, etc. The $m=1$ term yields

$$S^N_{M,a} \;\simeq\; \frac{1}{M!a^M}M\frac{a\Gamma(N)}{\Gamma(1-a)}\frac{1}{N^a} \;=\; \frac{1}{\Gamma(1-a)}\frac{1}{\Gamma(M)\,a^{M-1}}\frac{\Gamma(N)}{N^a}$$

The relative accuracy is $O(M/N^a)$, i.e. the approximation is good for $M \ll N^a$. The smaller a, the larger N needs to be to reach a reasonable accuracy. Higher m-terms may be added, but the alternating sign indicates cancelations and hence potential numerical problems.

(v) follows from $S^N_{M,0} = S(n,k;-1,0,0) = |S(n,k;1,0,0)|$ and the fact that $S(n,k;1,0,0)$ are Stirling numbers of the first kind from [HS88].

A [illegible]

We need to differentiate the different M that results from the partition sample I^*_N as N increases. Subscript M as M_N so we can differentiate it as N changes. When M_N is known, the following series relation holds:

$$\mathbb{E}_{M_N}[M_{N+1}] \;=\; \frac{b+M_Na}{N+b}+M_N \;=\; \frac{b}{N+b}+\frac{a+b+N}{N+b}M_N\ .$$

Taking expected values across M_N yields the recursive form

$$\mathbb{E}[M_{N+1}] \;=\; \frac{b}{N+b}+\frac{a+b+N}{N+b}\mathbb{E}[M_N]$$

The equation for $\mathbb{E}[M_N]$ given in the lemma is proven from this by induction, with the value 1 when $N=1$. Note the derivation of the solution to the above recursive formula was made by unfolding the recursion into a summation, and then simplifying the summation using hypergeometric functions.

The approximation for $\mathbb{E}[M_N]$ given in the lemma is derived for $N, b \gg a$ as follows:

$$\begin{aligned}
\frac{(a+b)_N}{(b)_N} &= \exp\left(\log\Gamma(a+b+N)-\log\Gamma(b+N)-(\log\Gamma(a+b)-\log\Gamma(b))\right)\\
&\simeq \exp\left(a\left(\psi_0(b+N)-\psi_0(b)\right)\right)\\
&\simeq \left(1+\frac{N}{b}\right)^a\exp\left(\frac{-a}{2}\left(\frac{1}{b+N}-\frac{1}{b}\right)\right)\\
&= \left(1+\frac{N}{b}\right)^a\exp\left(\frac{aN}{2b(b+N)}\right)\ .
\end{aligned}$$

The first approximation step makes a first order Taylor expansion since a is small, $0 < a < 1$, and the second approximation step uses an approximation for $\psi_0(b)$ with error $O(1/b^2)$.

For the expected variance, a similar strategy is used but the steps are more complicated. The following series relation holds:

$$\begin{aligned}\mathbb{E}_{M_N}\left[M_{N+1}^2\right] &= \frac{b + M_N a}{N+b}(M_N+1)^2 + \frac{N - M_N a}{N+b}M_N^2 \\ &= \frac{b}{N+b} + \frac{2b+a}{N+b}M_N + \frac{2a+b+N}{N+b}M_N^2 \ ,\end{aligned}$$

where $\mathbb{E}\left[M_N^2\right] = 1$ when $N = 1$. Taking expected values over M_N yields the recursive form

$$\mathbb{E}\left[M_{N+1}^2\right] = \frac{b}{N+b} + \frac{2b+a}{N+b}\mathbb{E}\left[M_N\right] + \frac{2a+b+N}{N+b}\mathbb{E}\left[M_N^2\right] \ .$$

Evaluation of this recursive formula can be made as before, and the result is the formula

$$\mathbb{E}\left[M_N^2\right] = \frac{b(a+b)}{a^2}\frac{(2a+b)_N}{(b)_N} - \frac{b(2b+a)}{a^2}\frac{(a+b)_N}{(b)_N} + \frac{b^2}{a^2} \ .$$

The result then comes from evaluating $\mathbb{E}\left[M_N^2\right] - \left(\mathbb{E}\left[M_N\right]\right)^2$ and simplifying terms. The approximation proceeds as before.

To handle the case where $a = 0$ the same recursive formula for $\mathbb{E}\left[M_N\right]$ and $\mathbb{E}\left[M_N^2\right]$ can be used, but are evaluated differently since $a = 0$. The closed form formula for $\mathbb{E}\left[M_N\right]$ follows clearly by induction on N. The closed form formula for $\mathbb{E}\left[M_N^2\right]$, readily proven by induction, is

$$b(\psi_0(b+N) - \psi_0(b)) + b^2\left(\psi_0(b+N) - \psi_0(b)\right)^2 + b^2(\psi_1(b+N) - \psi_1(b)) \ .$$

Subtracting off $\left(\mathbb{E}\left[M_N\right]\right)^2$ yields the result for $\mathrm{Var}\left[M_N\right]$. The approximations for both $\mathbb{E}\left[M_N\right]$ and $\mathrm{Var}\left[M_N\right]$ follow by taking the first order terms of $\psi_0(\cdot)$ and $\psi_1(\cdot)$, the log and the inverse respectively.

[illegible]

The value M is equal to the number of indices that have a non-zero count in the sample of size N. Given probability vector $\vec{q}$, the probability that index k has a non-zero count after N samples is $1 - (1-q_k)^N$. Summing these over all k gives an upper bound.

To generate bounds on $1 - (1-q_k)^N$, note $1 - (1-q_k)^N \leq 1$, and the bound is closer the larger q_k. Second, by Taylor expansion

$$1 - (1-q_k)^N = Nq_k - \frac{N(N-1)}{2}q_k'^2$$

for some $0 \leq q'_k \leq q_k$. So $1-(1-q_k)^N \leq Nq_k$, and the bound is closer the smaller q_k, especially when $Nq_k \ll 1$. Put these two bounds together and we get for any positive integer m

$$\sum_{k=1}^{\infty} 1-(1-q_k)^N \;\leq\; m + N \sum_{k=m+1}^{\infty} q_k \;.$$

For the geometric series, $q_k = r^{k-1}(1-r)$, the sum in the bound evaluates to r^m, so we seek to minimise $m + Nr^m$. This bound can be modified if we let $m \in \mathbb{R}^+$ to

$$\min_{m \in \mathbb{N}^+} m + Nr^m \;\leq\; \min_{m \in \mathbb{R}^+} 1 + m + Nr^{m-1}$$

Differentiating yields a minima at $r^{m-1} = \frac{1}{N \log 1/r}$. The result follows by substitution.

For the Dirichlet series, $q_k = \frac{k^{-s}}{\zeta(s)}$, the sum in the bound can be bounded by an integral

$$\begin{aligned}
\sum_{k=m+1}^{\infty} q_k &\leq \int_{m+1/2}^{\infty} \frac{1}{(k)^s \zeta(s)} \;, \\
&= \frac{-1}{(k)^{s-1}\zeta(s)(s-1)}\Big|_{m+1/2}^{\infty} \\
&= \frac{1}{(m+1/2)^{s-1}\zeta(s)(s-1)}
\end{aligned}$$

As before, modifying the bounds yields the formula to minimise

$$m + 1 + \frac{N}{(m-1/2)^{s-1}\zeta(s)(s-1)} \;.$$

Differentiation gives a minimum at $N = (m-1/2)^s \zeta(s)$ and so the bound follows.

2 [illegible]

2.1 [illegible]

One sees the samples generated by the fragmentation process. Given a partition with partition count L with occurrences $n_1, ..., n_L$ totalling N, one needs to assign every entry l to a latent cluster k_l. This results in a coarser partition with count K bins, and with total occurrence $m_k = \sum_{l\,:\,k_l=k} n_l$ (where $m_k = |P_m|$ for the latent bin P_m) from $t_k = \sum_{l\,:\,k_l=k} 1$ original bins, totalling $\sum_{k=1}^{K} t_k = L$. The full probability for the fragmentation components is (note a_2 could also be zero here, but these terms correctly cancel, so the parallel equations for $a_2 = 0$ are not given):

$$\frac{(b|a_1 a_2)_K}{(b)_N} \prod_{k=1}^{K} (1-a_1 a_2)_{m_k - 1} \;\cdot\; \prod_{k=1}^{K} \frac{(-a_1 a_2|a_1)_{t_k}}{(-a_1 a_2)_{m_k}} \prod_{l\,:\,k_l=k} (1-a_1)_{n_l - 1}$$

$$= \frac{1}{(b)_N}\prod_{l=1}^{L}(1-a_1)_{n_l-1}(b|a_1a_2)_K\prod_{k=1}^{K}\frac{(-a_1a_2|a_1)_{t_k}}{-a_1a_2}$$
$$= \frac{1}{(b)_N}a_1^L\prod_{l=1}^{L}(1-a_1)_{n_l-1}\cdot(b/a_1|a_2)_K\prod_{k=1}^{K}(1-a_2)_{t_k-1}\ .$$

We need to marginalise over all possible coarser partitions or assignments to k_l. Now the second term is in the form of a CRD $(\{1,...,l\},a_2,b/a_1)$ so marginalising out the t_k yields $(b/a_1)_L$, which multiplied by a_1^L gives $(b|a_1)_L$. The result is that the above probability becomes

$$\frac{(b|a_1)_L}{(b)_N}\prod_{l=1}^{L}(1-a_1)_{n_l-1}\ .$$

This is the CRD $(\{1,...,N\},a_1,b)$.

2A 6FB2

Consider a sample partition with a bin j with n_j entries, however we do not know which sub-partitions (from the unknown index k) the entry comes from, nor how many there were. So now each bin j would be sub-partitioned into L_j bins with occurrence counts $m_{j,l}$ each where $\sum_{l=1}^{L_j}m_{j,l}=n_j$. In total, there are $L=\sum_{j=1}^{J}L_j$ bins. With this assignment done, the probability for the coagulation becomes:

$$\frac{(b|a_1)_L}{(b)_N}\prod_{j,l}(1-a_1)_{m_{j,l}-1}\cdot\frac{(b/a_1|a_2)_J}{(b/a_1)_L}\prod_{j}(1-a_2)_{L_j-1}$$
$$= \frac{(b|a_1a_2)_J}{(b)_N}\frac{1}{(-a_1a_2)^J}\prod_{j=1}^{J}(-a_1a_2|a_1)_{L_j}\prod_{l=1}^{L_j}(1-a_1)_{m_{j,l}-1}\ .$$

Note the last formula is undefined for $a_2=0$, however, the zero terms cancel and a parallel proof for the $a_2=0$ case is readily seen to hold. To construct the marginalised probability, we have to marginalise over all possible sub-partitions (L_j bins with occurrence counts $m_{j,l}$). The term inside the $\prod_{j=1}^{J}$ is the form of a CRD $(\{1,...,L_j\},a_1,-a_1a_2)$ so marginalising out the partitions (represented by L_j and $m_{j,l}$) yields $(-a_1a_2)_{n_j}=(1-a_1a_2)_{n_j-1}(-a_1a_2)$ (again, noting a similar argument applies for the $a_2=0$ case). The probability simplifies then to

$$\frac{(b|a_1a_2)_J}{(b)_N}\prod_{j=1}^{J}(1-a_1a_2)_{n_j-1}\ .$$

This is then CRD $(\{1,...,N\},a_1a_2,b)$ as required.

2A 6FB2

To prove the result, it is sufficient to show for all samples, the two have equivalent marginalised posteriors. Consider a sample partition, for a given bin each entry

would be drawn with a probability $p_k q_{j,k}$, however we do not know the j or k associated. We have to assign the k, whereas the j term is simply a sub-partition so we do not need to know it. Given a sample partition like this of L bins with sizes $n_1, ..., n_L$ totalling N, one needs to assign every entry l to a cluster k_l. This results in a coarser partition with K bins, with total count $m_k = \sum_{l\,:\,k_l=k} n_l$ from $t_k = \sum_{l\,:\,k_l=k} 1$ original bins, totalling $\sum_{k=1}^{K} t_k = L$. But with the clusters k_l assigned, the posterior with the marginalised terms for $\vec{p}$ and $\vec{q}_k$'s can be given. This posterior corresponds to the situation in Theorem 22, so apply that result and this converts the posterior into a probability for CRD $(\{1, ..., N\}, a_1, b)$. This is the required posterior for the sample from PDD (a_1, b).

2A [illegible]

As before, to prove the result, it is sufficient to show for all samples, the two have equivalent marginalised posteriors. Consider a sample partition with a bin j with n_j entries, each entry would be drawn with a probability $\sum_{k\,:\,j_k=j} p_k$, however we do not know which sub-partitions (from the unknown index k) the entry comes from, nor how many there were. So now each bin j would be sub-partitioned into L_j bins with occurrence counts $m_{j,l}$ each where $\sum_{l=1}^{L_j} m_{j,l} = n_j$. In total, there are $L = \sum_{j=1}^{J} L_j$ bins. With this assignment done, the marginalised posterior can be written down from the term for $\vec{p}$ and the term for $\vec{q}$, however this corresponds to the situation from Theorem 23. Applying that result yields a marginalised posterior in the form of a CRD $(\{1, ..., N\}, a_1 a_2, b)$. This is the required posterior for the sample from PDD $(a_1 a_2, b)$.

3A [illegible]

1A [illegible]

Consider the prior measure for $p_1, ..., p_M, p_M^+$. Do a change of variables to $p_1, ..., p_{M-1}, q_M, p_{M-1}^+$ where $q_M = p_M / p_{M-1}^+$ and $p_{M-1}^+ = p_M + p_M^+$. The Hessian of this change is $1/p_{M-1}^+$, and the domain goes from the constraint set $\{p_1 \geq 0, ..., p_M \geq 0, p_M^+ \geq 0\}$ to $\{p_1 \geq 0, ..., p_{M-1} \geq 0, p_{M-1}^+ \geq 0, 0 \leq q_M \leq 1\}$. The prior measure can thus be converted to

$$\left(q_M^{-a-1} \left(1 - q_M\right)^{b+Ma-1}\right) \left(p_{M-1}^+\right)^{b+(M-1)\,a-1} \prod_{m=1}^{M-1} p_m^{-a-1} \ ,$$

under the new constraint set. Note the prior measure on sub-vector $p_1, ..., p_{M-1}$, as given in Definition 28, appears in the second half of this measure. The initial part involves only q_M, but its constraints are simply $0 \leq q_M \leq 1$ which are independent of the remaining variables. Thus one is left with a measure on $p_1, ..., p_{M-1}$. The measure on the sub-vector is now consistent with Definition 28. We can repeat this process recursively to verify consistency for any other sub-vector.

2A [illegible]

Consider Definitions 28 and 8. Define G_δ in terms of its projection on the finite sub-spaces $\{p_1, p_2, ..., p_M\}$ for all M. Let

$$p(p_1, p_2, ..., p_M, p_M^+) \;\propto\; \left(p_M^+\right)^{b+M\,a-1} \prod_{m=1}^{M} p_m^{-a-1} \;, \tag{16}$$

where $p_M^+ = 1 - \sum_{m=1}^{M} p_m$ and the domain is constrained to be $p_m > (1 - \sum_{i=1}^{m-1} p_i)\delta$ for $m = 1, ..., M$, and $p_M^+ \geq 0$. Note that by Definition 28, $b > -a$, and thus $b + Ma > 0$. Exploiting $p_m > \delta$ for $m = 1, ..., M$, we show below that the proportionality constant, i.e. the integral over the constrained simplex, is finite. To show G_δ is proper, we need to show that the finite priors for each M are proper and that consistency holds between these priors for different M.

The normalization is done as follows. Use the same change of variables as in the proof of Lemma 29, however now the domain is different. The constraint set for the initial variables is

$$C_{p,M} \;=\; \left\{ p_1 \geq \delta, ..., p_m \geq \left(1 - \sum_{i=1}^{m-1} p_i\right)\delta, ..., p_M \geq \left(1 - \sum_{i=1}^{M-1} p_i\right)\delta, p_M^+ \geq 0 \right\} \;.$$

By the change of variables this gets mapped to

$$C_{q,M} \;=\; \left\{ p_1 \geq \delta, ..., p_{M-1} \geq \left(1 - \sum_{i=1}^{M-2} p_i\right)\delta, p_{M-1}^+ \geq 0, \delta \leq q_M \leq 1 \right\} \;.$$

For the purposes of integration, denote the initial and changed variable sets as $\vec{p}$ and $\vec{q}$ respectively. Thus the integration works as follows:

$$\begin{aligned}
Z_{a,b,M,\delta} \;&:=\; \int_{C_{p,M}} \left(p_M^+\right)^{b+M\,a-1} \prod_{m=1}^{M} p_m^{-a-1} \mathrm{d}\,\vec{p} \\
&=\; \int_{C_{q,M}} \left(p_{M-1}^+\right)^{b+(M-1)\,a-1} \prod_{m=1}^{M-1} p_m^{-a-1} \left(1 - q_M\right)^{b+M\,a-1} q_M^{-a-1} \mathrm{d}\,q \\
&=\; Z_{a,b,M-1,\delta} \int_{\delta}^{1} (1-q)^{b+M\,a-1}\, q^{-a-1} \mathrm{d}\,q \\
&=\; ... \;=\; \prod_{m=1}^{M} \int_{\delta}^{1} (1-q)^{b+m\,a-1}\, q^{-a-1} \mathrm{d}\,q \;.
\end{aligned}$$

Note this is bounded above by bounding the q^{-a-1} terms from inside the integral with δ^{-a-1}, and extending the integrals to the range $[0, 1]$. This yields the upper bound $\delta^{-M(a+1)} \prod_{m=1}^{M} \frac{\Gamma(b+ma)}{\Gamma(b+ma+1)}$.

Now we prove consistency. We need to show that the projection from the subset $m = 1, ..., M$ down to some smaller subset $m = 1, ..., M' < M$ is consistent. The change of variables above handled the case where $p(p_1, p_2, ..., p_M, p_M^+)$ was projected down to $p(p_1, p_2, ..., p_{M-1}, p_{M-1}^+)$. Clearly, the projected prior is equivalent

to the direct definition above (see Lemma 29 for details). Thus by induction, one can project the prior from the subset $m = 1, ..., M$ down to a any smaller subset $m = 1, ..., M' < M$, and get the same prior. By this condition, and Kolmogorov's Consistency Theorem, it follows that the prior G_δ exists and is proper for the full Hilbert space of $\vec{p}$.

Now consider the posteriors for a given sample I_N. The posterior for I_N using the improper prior on PDDs is given in Lemma 31. To deal with the proper prior G_δ, the notion of partition size is needed, as given in Definition 8. Let M_N be the partition size for a I_N, then $p(p_1, p_2, ..., p_M, p_M^+ | G_\delta, I_N)$ is proportional to

$$\left(p_M^+\right)^{b+M\,a-1} \prod_{m=M_N+1}^{M} p_m^{-a-1} \prod_{m=1}^{M_N} p_m^{n_m-a-1} \ ,$$

where the constraints $C_{p,M}$ hold as before. This is the same form as the posterior Dirichlet distribution (I) given in Lemma 31 where the probabilities are further constrained by $C_{p,M}$. The normalizing constant can be worked out as above to be

$$Z_{a,b,M_N,\delta} \ = \ \prod_{m=1}^{M_N} B_{1-\delta}(b + ma, n_m - a)$$

where $B_x(u, v) = \int_0^x t^{u-1}(1-t)^{v-1}dt$ is the incomplete Beta function defined for $u, v > 0$. In our case, $n_m > 0$ for all $1 \le m \le M_N$ and $a + b > 0$, so the Beta function and incomplete Beta function are well defined. Note the normalizing constant for the posterior Dirichlet distribution (I) given in Lemma 31 is $Z_{a,b,M_N,0}$.

Now consider the L_1 distance between the two posteriors, $p(p_1, p_2, ..., p_M, p_M^+ | G_\delta, I_N)$ and the posterior Dirichlet distribution (I) given in Lemma 31. Note these differ only in domain. Denote them by P_δ and P_0 respectively. Using $P_\delta \ge P_0$ on G_δ, and $P_\delta = 0$ on $G_0 \setminus G_\delta$, and $\int_{G_\delta} P_\delta d\vec{p} = 1$, we get

$$\begin{aligned}
\frac{1}{2} d_1(P_\delta, P_0) \ &:= \ \sup_A |P_\delta[A] - P_0[A]| \\
&= \ \frac{1}{2}\int_{G_0} |P_\delta - P_0| d\vec{p} \\
&= \ \frac{1}{2}\int_{G_\delta} |P_\delta - P_0| d\vec{p} + \frac{1}{2}\int_{G_0 \setminus G_\delta} |P_\delta - P_0| d\vec{p} \\
&= \ \frac{1}{2}\int_{G_\delta} P_\delta d\vec{p} - \frac{1}{2}\int_{G_\delta} P_0 d\vec{p} + \frac{1}{2}\int_{G_0 \setminus G_\delta} P_0 d\vec{p} \\
&= \ 1 - \int_{G_\delta} P_0 d\vec{p} \ = \ 1 - \frac{Z_{a,b,M_N,\delta}}{Z_{a,b,M_N,0}} \int_{G_\delta} P_\delta d\vec{p} \\
&= \ 1 - \frac{Z_{a,b,M_N,\delta}}{Z_{a,b,M_N,0}} \to 0 \quad \text{for} \quad \delta \to 0
\end{aligned}$$

This implies convergence in distribution.

3.4 [illegible]

[illegible] Note for the Proper Posteriors II claim, since $H(\cdot)$ is non-atomic, each distinct data X^*_m has a corresponding distinct index k^*_m, thus for the purposes of analysis, assume the indices are given and w.l.o.g. they follow size-biased ordering, so $k^*_m = m$. Thus to prove the Proper Posteriors I and II claim about posteriors for $\vec{p}$, multiply the prior measure for $(p_1, ..., p_M)$ of Definition 28 by the likelihood, which is in terms of the same sub-vector, and the posterior measure clearly is proportional to the corresponding posterior Dirichlet in this lemma. The remaining part of the Proper Posteriors II claim follows from the model family.

To prove the Sampling claim, note that this just takes the expected value of the posterior in Proper Posteriors II. To prove the Stick Breaking claim, note this follows directly from the posterior by standard properties of the Dirichlet.

The size-biased sampling claim is developed sequentially. First note p_1 has the improper prior Beta$(-a, b + a)$. The value $p_2/(1 - p_1)$ is *apriori* independent of p_1 and has a the improper prior Beta$(-a, b + 2a)$. The value $p_3/(1 - p_1 - p_2)$ is *apriori* independent of p_1 and p_2 and has a the improper prior Beta$(-a, b+3a)$, *etc.* However, because p_1 is size-biased we know it is the first in the sample, so we add one to get a Beta$(1-a, b+a)$. Likewise, we also know p_2 appears first in the sample (after p_1), since it is sized-biased, so again we add one getting Beta$(1 - a, b + 2a)$. Repeating this yields the standard stick-breaking definition of the GEM distribution.

4 [illegible]

4.1 [illegible]

[illegible] In the general case where each draw from $H(\cdot)$ is not necessarily almost surely distinct, the formula of Lemma 14 also applies to $p(X_1, X_2, ..., X_N, k_1, ..., k_N)$. Now one can marginalise out the $k_1, ..., k_N$, which will affect the last product of M terms only.

Given the constraints that t_m represents the multiplicity of X^*_k and n_k represents the total count of X^*_k, then all values for $k_1, ..., k_N$ must be included that satisfy the constraints. Each n_k will be partitioned into t_k different indices, each occurring at least once, and totaling n_k. Thus the problem of marginalising out the indices $k_1, ..., k_N$ to the multiplicities $t_1, ..., t_M$ is equivalent to the summation over configurations by size-biased ordering, done for Lemma 16, and an identical result can be applied.

4.2 [illegible]

Let $\vec{p} \sim \text{PDP}(a, b, H)$. Let $\vec{q} \sim \text{PDD}(a, b)$ be the underlying PDD, and let the corresponding independent samples from $H(\cdot)$ be $X_l \in \mathcal{N}$. From the definition of a PDP,

$$p_k = \sum_l q_l 1_{X_l = k}$$

Taking the expected value of this over $\vec{X}$, yields $\sum_l q_l \theta_k$, and hence θ_k irrespective of $\vec{q}$.

Now consider any moment. We present one case, and others can be treated similarly. For k_1, k_2, k_3 three indices

$$\begin{aligned} & \mathbb{E}_{\vec{q},\vec{X}} \left[\left(\sum_l q_l 1_{X_l=k_1} - \theta_{k_1} \right) \left(\sum_l q_l 1_{X_l=k_2} - \theta_{k_2} \right) \left(\sum_l q_l 1_{X_l=k_3} - \theta_{k_3} \right) \right] \\ = \; & \mathbb{E}_{\vec{q},\vec{X}} \left[\sum_{l_1,l_2,l_3} q_{l_1} q_{l_2} q_{l_3} \left(1_{X_{l_1}=k_1} - \theta_{k_1} \right) \left(1_{X_{l_2}=k_2} - \theta_{k_2} \right) \left(1_{X_{l_3}=k_3} - \theta_{k_3} \right) \right] \end{aligned}$$

Now X_{l_1} is independent of X_{l_2} whenever $l_1 \neq l_2$. So we have to express the sum $\sum_{l_1,l_2,l_3}$ into different equal and unequal parts so that the expected value over $\vec{X}$ can be applied. This would be

$$\sum_{l_1,l_2,l_3} \cdot \; = \sum_{l_1,l_2,l_3 \text{ disjoint}} \cdot + \sum_{l_1=l_2\neq l_3} \cdot + \sum_{l_1=l_3\neq l_2} \cdot + \sum_{l_2=l_3\neq l_1} \cdot + \sum_{l_1=l_2=l_3} \cdot$$

Any sum which has a term with one index, l_1 say, not equal to the others, will contain the expression

$$\mathbb{E}_{X_{l_1}} \left[1_{X_{l_1}=k_1} - \theta_k \right] \; = \; \theta_{k_1} - \theta_{k_1} \; = \; 0 \; ,$$

and hence can be discarded. Thus for the first three central moments, the expansion of sums that remains non-zero are

$$\begin{aligned} \sum_{l_1,l_2} \cdot \; &= \sum_{l_1=l_2} \cdot \\ \sum_{l_1,l_2,l_3} \cdot \; &= \sum_{l_1=l_2=l_3} \cdot \\ \sum_{l_1,l_2,l_3,l_4} \cdot \; &= \sum_{l_1=l_2\neq l_3=l_4} \cdot + \sum_{l_1=l_3\neq l_2=l_4} \cdot + \sum_{l_1=l_4\neq l_2=l_3} \cdot + \sum_{l_1=l_2=l_3=l_4} \cdot \end{aligned}$$

Applying these summations to the three moments leads to:

$$\begin{aligned} & \mathbb{E}_{\vec{q}} \left[\sum_l q_l^2 \right] \mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_2} - \theta_{k_2} \right) \right] \\ & \mathbb{E}_{\vec{q}} \left[\sum_l q_l^3 \right] \mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_2} - \theta_{k_2} \right) \left(1_{X=k_3} - \theta_{k_3} \right) \right] \\ & \mathbb{E}_{\vec{q}} \left[\sum_l q_l^4 \right] \mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_2} - \theta_{k_2} \right) \left(1_{X=k_3} - \theta_{k_3} \right) \left(1_{X=k_4} - \theta_{k_4} \right) \right] \\ & + \left(\left(\mathbb{E}_{\vec{q}} \left[\sum_l q_l^2 \right] \right)^2 - \mathbb{E}_{\vec{q}} \left[\sum_l q_l^4 \right] \right) \\ & \quad \Big(\mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_2} - \theta_{k_2} \right) \right] \mathbb{E}_X \left[\left(1_{X=k_3} - \theta_{k_3} \right) \left(1_{X=k_4} - \theta_{k_4} \right) \right] \\ & \quad \; \mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_3} - \theta_{k_3} \right) \right] \mathbb{E}_X \left[\left(1_{X=k_2} - \theta_{k_2} \right) \left(1_{X=k_4} - \theta_{k_4} \right) \right] \\ & \quad \; \mathbb{E}_X \left[\left(1_{X=k_1} - \theta_{k_1} \right) \left(1_{X=k_4} - \theta_{k_4} \right) \right] \mathbb{E}_X \left[\left(1_{X=k_2} - \theta_{k_2} \right) \left(1_{X=k_3} - \theta_{k_3} \right) \right] \Big) \end{aligned}$$

The expected sum of powers of $\vec{q}$ we solve for below. The expectation of X is for the multivariate discrete (or a multinomial with $N = 1$), so the values are known for the various cases of $k_1, k_2, ...$ For example, when $k_1 \neq k_2$, $\mathbb{E}_X\left[\left(1_{X=k_1} - \theta_{k_1}\right)\left(1_{X=k_2} - \theta_{k_2}\right)\right] = -\theta_{k_1}\theta_{k_2}$.

The expected sum of powers of $\vec{q}$ is obtained as follows. From first principles, it can be seen that

$$\mathbb{E}_{\vec{q}}[M] \;=\; \mathbb{E}_{\vec{q}}\left[1 - (1 - q_k)^N\right]$$

For $N = 2$ and rearranging terms we get

$$\mathbb{E}_{\vec{q}}[M_2] \;=\; 2 - \mathbb{E}_{\vec{q}}\left[\sum_l q_l^2\right]$$

Applying Lemma 18 one gets the closed form expression for the left-hand side. Likewise, we get:

$$\begin{aligned}
\mathbb{E}_{\vec{q}}\left[\sum_l q_l^2\right] &= \frac{1-a}{1+b} \\
\mathbb{E}_{\vec{q}}\left[\sum_l q_l^3\right] &= \frac{(1-a)(2-a)}{(1+b)(2+b)} \\
\mathbb{E}_{\vec{q}}\left[\sum_l q_l^4\right] &= \frac{(1-a)(2-a)(3-a)}{(1+b)(2+b)(3+b)} \\
\mathbb{E}_{\vec{q}}\left[\sum_l q_l^5\right] &= \frac{(1-a)(2-a)(3-a)(4-a)}{(1+b)(2+b)(3+b)(4+b)}
\end{aligned}$$

Combining the resultant formula yields the cases in the lemma.