A Bayesian Review of the Poisson-Dirichlet Process

Wray Buntine

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

Marcus Hutter

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

1 July 2010

Abstract

The two parameter Poisson-Dirichlet process is also known as the Pitman-Yor Process and related to the Chinese Restaurant Process, is a generalisation of the Dirichlet Process, and is increasingly being used for probabilistic modelling in discrete areas such as language and images. This article reviews the theory of the Poisson-Dirichlet process in terms of its consistency for estimation, the convergence rates and the posteriors of data. This theory has been well developed for continuous distributions (more generally referred to as non-atomic distributions). This article then presents a Bayesian interpretation of the Poisson-Dirichlet process: it is a mixture using 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 fit naturally. Finally, this article also presents results for the discrete case which is the case seeing widespread use now in computer science, but which has received less attention in the literature.

Keywords

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

1 Introduction

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 particular interpretation of the model known as the Chinese Restaurant Process (CRP) which gives an elegant analogy of incremental sampling for these models. The models have proven useful in a number of ways as tools for non-parametric and hierarchical Bayesian modelling, especially in discrete domains such as language and images where one wants to be flexible with dimensions.

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 with performance based approaches. 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].

The theory is well developed for the more general context of continuous distributions, and that theory is reviewed here in detail. Details of convergence, consistency and the forms of posteriors are given in Section 4.

A new interpretation and definition of the PDP is then given in Section 5. This uses the methodology of Bayesian improper priors too show that the distribution on positive integers (an infinite probability vector $\vec{p}$) underlying the PDP is in fact an infinite improper Dirichlet, which, conceptually, is what we understand anyway with all its sampling and additivity properties.

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 6. This context is quite different from the continuous distributions of standard theory.

2 Infinite Mixture Models

Before introducing the PDP, we introduce the basic context of its use, an infinite mixture model.

The kinds of models we consider require as input a *base* probability distribution $H(\cdot)$ on a measurable space $\mathcal{X}$, and yields a discrete distribution on a finite or countably infinite subset of $\mathcal{X}$. A distribution of this form can be represented as

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

where $\vec{p}$ is a probability vector so $0 \leq p_k \leq 1$ and $\sum_{k=1}^{\infty} p_k = 1$, and $\delta_{X_k^*}(\cdot)$ is a discrete measure concentrated at X_k^*. We assume the values $X_k^* \in \mathcal{X}$ are independently and identically distributed according to $H(\cdot)$, which is referred to as the *base distribution*. When the base distribution is non-atomic[1], for instance a probability density function such as a Gaussian, it follows that almost surely $X_k^* \neq X_l^*$ whenever $k \neq l$.

Consider a sample, where $H(\cdot)$ is the uniform distribution on $[0, 1]$, and the real numbers presented here are rounded to four places, such as $0.4674\cdot$. In this case, a sequence of indices k and values X_k^* are sampled from Formula (1) as $(k_1, X_{k_1}^*), ..., (k_n, X_{k_n}^*), ...$ This might be:

$$(12, 0.4674\cdot), (435, 0.3925\cdot), (7198, 0.1937\cdot), (12, 0.4674\cdot),$$
$$(12, 0.4674\cdot), (35, 0.3947\cdot), (7198, 0.1937\cdot), ...$$

This says the X_{12}^* appears at the 1st, 4th and 5th position in the sample, X_{7198}^* appears at the 3rd and 7th, etc. One can view just the indices from Formula (1), the k's, used in generating this sample. In the above case this is the *index sequence* $12, 435, 7198, 12, 12, 35, 7198, ...$, whereas the *data sequence* for this is

$$0.4674\cdot, 0.3925\cdot, 0.1937\cdot, 0.4674\cdot, 0.4674\cdot, 0.3947\cdot, 0.1937\cdot, ...$$

Since the actual indices are latent and not part of the observed data for the model of Formula (1), the indices observed can be converted to a normal form. A *standard renumbering of the indices* appearing in the sample is

1, 2, 3, 1, 1, 4, 3, ...,

where each new index (the k in Formula (1)) is given the next free integer, starting at 1[2]. Informally, we refer to the sequence of indices under the standard renumbering as the *partition structure*: it defines the grouping of items in the sequence, not the actual k's assigned.

This standard renumbering of the indices is an ordering under *sized-biased sampling* [Pit95], but we do not formally cover the theory here for brevity. Our treatment of indices and partition structure is motivated by Pitman [Pit95], but our notation is different; Pitman considers a far richer set of research questions, whereas we are restricting our analysis to the infinite mixture model of Formula (1).

To complete a definition of a family of models following the infinite mixture model of Formula (1), we need to specify the probability vector $\vec{p}$. Within the PDP literature, $\vec{p}$ follows a two parameter Poisson-Dirichlet distribution [PY97]. One definition for it is 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.

[1] So $H(X) = 0$ for all $X \in \mathcal{X}$, thus samples from $H(\cdot)$ are almost surely distinct.

[2] So the sequence $k_1, ..., k_n, ...$ has the property that $k_n \leq 1 + \max_{1 \leq i < n} k_i$.

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 1 (Poisson-Dirichlet distribution) *For $0 \leq a < 1$ and $b > -a$, suppose that a probability $P_{a,b}$ governs independent random variables V_k such that V_k has Beta$(1 - a, b + k\,a)$ distribution. Let*

$$\tilde{p}_1 \;=\; V_1, \qquad \tilde{p}_k \;=\; (1 - V_1) \cdots (1 - V_{k-1})V_k \qquad k \geq 2 \;,$$

and let $p_1 \geq p_2 \geq \cdots$ be the ranked (sorted) values of the $\tilde{p}_k$. Define the Poisson-Dirichlet distribution *with parameters a, b, abbreviated PDD(a, b) to be the $P_{a,b}$ distribution of p_n.*

Here our a parameter is usually called the *discount parameter* in the literature, and b is called the *concentration parameter*.

Note when estimating the probability $\vec{p}$ for the mixture model of Formula (1) and using this Poisson-Dirichlet distribution, the sorting is important for sampling efficiency [KWT07], but it is not always necessary in the theory and does not appear in some definitions [IJ01].

3 Poisson-Dirichlet Process

One definition of a Poisson-Dirichlet process is that it extends the Poisson-Dirichlet 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*, commonly represented 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 2 (Poisson-Dirichlet Process) *Let $H(\cdot)$ be a distribution over some measurable space $\mathcal{X}$. For $0 \leq a < 1$ and $b > -a$, suppose that $\vec{p}$ is drawn from a Poisson-Dirichlet distribution with parameters a, b. Moreover, let X_k^* for $k = 1, 2, ...$ be a sequence of independent samples drawn according to $H(\cdot)$. 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.

The PDP is also called a *stochastic process* because it is often 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 output distribution. The conditional distribution with $\vec{p}$ marginalised out for this, as long as $H(\cdot)$ is non-atomic, is as follows:

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

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 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 enjoy the item X^*_m.

If data is sampled according to this Chinese Restaurant Process (CRP) and $H(\cdot)$ is non-atomic, then it is a *Poisson-Dirichlet Process* with parameters a, b and *base distribution* $H(\cdot)$ [Pit95, IJ01], thus the CRP can serve as an alternative definition of the PDP.

In discrete applications common in computer science the non-atomicity of the base distribution does not hold, the domain is countable already, for instance it might represent the space of possible English language words. Technically, $H(\cdot)$ is discrete when $H(X) > 0$ for all $X \in \mathcal{X}$. In this case, the standard theory needs to be modified. Teh [Teh06b] presented some modifications in this case, and we expand on these in Section 6. Alternatively, some authors avoid the domain $\mathcal{X}$ by dealing with the space of underlying indices or partitions that result from the sampling, such as Pitman and Yor [Pit95, PY97] and this yields the *two-parameter Poisson-Dirichlet distribution* discussed in the previous section.

4 Basic Properties

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) of $S_N := (X^*_{k_1}, ..., X^*_{k_N})$, a corresponding sample of natural numbers exists $I_N := (k_1, ..., k_N)$, however, these

remain hidden (only the values of each $X^*_{k_i}$ are known, not the indexes k_i). For a non-atomic base distribution the indices are irrelevant and we can renumber the indices by the standard renumbering. 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 3 (Index Statistics) *When sampling independently and identically from the discrete distribution of Formula (2), one gets a* data sequence *of length N given by $S_N = X_1, X_2, ..., X_N$. Associated with this are the latent indices, an* index sequence *of length N given by $I_N = k_1, k_2, ..., k_N$. Alternatively, one could sample independently and identically from a PDD to obtain such an index sequence. In I_N the 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 and sample size N. Moreover, retain the order of occurrence so that k^*_1 occurs first in I_N, k^*_2 occurs second, and so forth. Note, without loss of generality, one could apply the standard renumbering to the indices, which means setting $k^*_m = m$ for $m = 1, \ldots, M$, and the corresponding index sequence is denoted I^*_N.*

In the running example of Section 2 with $N = 7$ points, $I_N = 12, 435, 7198, 12, 12, 35, 7198$ and $I^*_N = 1, 2, 3, 1, 1, 4, 3$. Thus $M = 4$ 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 *table count* in the CRP terminology. Note also that for a discrete base distribution, the "true" but hidden indices I_N may form a finer partition than I^*_N as given in the definition.

Definition 4 (Data Statistics) *When sampling from the discrete distribution of Formula (3), 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 X^*_m, so assigning $k^*_m = m$, which is the standard renumbering. The corresponding index sequence is denoted I^*_N.*

For the running example again, $X^*_1 = 0.4674\cdot$, $X^*_2 = 0.3925\cdot$, $X^*_3 = 0.1937\cdot$ and $X^*_4 = 0.3947\cdot$ with I^*_N, M and the n_m the same as before.

4.1 Consistency results

The PDD can be used to learn a broader class of distributions, not just those that are from a given 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 learn $\vec{q}$ based on integer samples.

Lemma 5 *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 3. 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 3.*

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 learn continuous distributions. This is made precise by the consistency result due to James [Jam08, Proposition 2.1].

Lemma 6 (PDP posterior convergence) *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 Poisson-Dirichlet process with parameters $0 \leq a < 1$ and $b > -a$ and base distribution $H(\cdot)$ 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.

4.2 Posteriors

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 7 (Probability of evidence) *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 4. Then the probability of evidence given the model PDP$(a, b, H(\cdot))$ is*

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

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* $(x|0)_N = x^N$*.*

The key characteristic of the PDD is the *partition size* M from Definition 3. This is also related to the expected posterior probability of seeing a new index (for the PDD) or a new data value from $\mathcal{X}$ in the non-atomic case, 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].

Lemma 8 (Probability on partition size) *Consider the 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* I_N *is*

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

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

for $M \leq N$ *and 0 else.*

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

Theorem 9 (Expressions for $S^N_{M,a}$)

(i) Linear recursion: $S^{N+1}_{M,a} = S^N_{M-1,a} + (N - Ma) S^N_{M,a}$
Boundary cond.: $S^N_{M,a} = 0$ *for* $M > N$, $\quad S^N_{0,a} = \delta_{N,0}$.

(ii) Mult. recursions: $S^N_{M,a} = \sum_{n=m}^{N-M+m} \frac{\binom{N}{n}}{\binom{M}{m}} S^n_{m,a} S^{N-n}_{M-m,a} = \sum_{n=1}^{N-M+1} \binom{N-1}{n-1} S^n_{1,a} S^{N-n}_{M-1,a}$
$S^N_{1,a} = \Gamma(N-a)/\Gamma(1-a)$. *Any* $0 < m < M$.

(iii) Explicit expression: $S^N_{M,a} = \frac{1}{M!\,a^M} \sum_{m=0}^{M} \binom{M}{m} (-)^m \prod_{h=0}^{N-1} (h - am)$

(iv) Asymptotic expr.: $S^N_{M,a} \simeq \frac{1}{\Gamma(1-a)} \frac{1}{\Gamma(M)\,a^{M-1}} \frac{\Gamma(N)}{N^a}$ *for* $a > 0$

(v) Expr. for $a = 0$*:* $S^N_{M,0} = |s^{(M)}_N|$ = *unsigned Stirling# of 1st kind [AS74]*

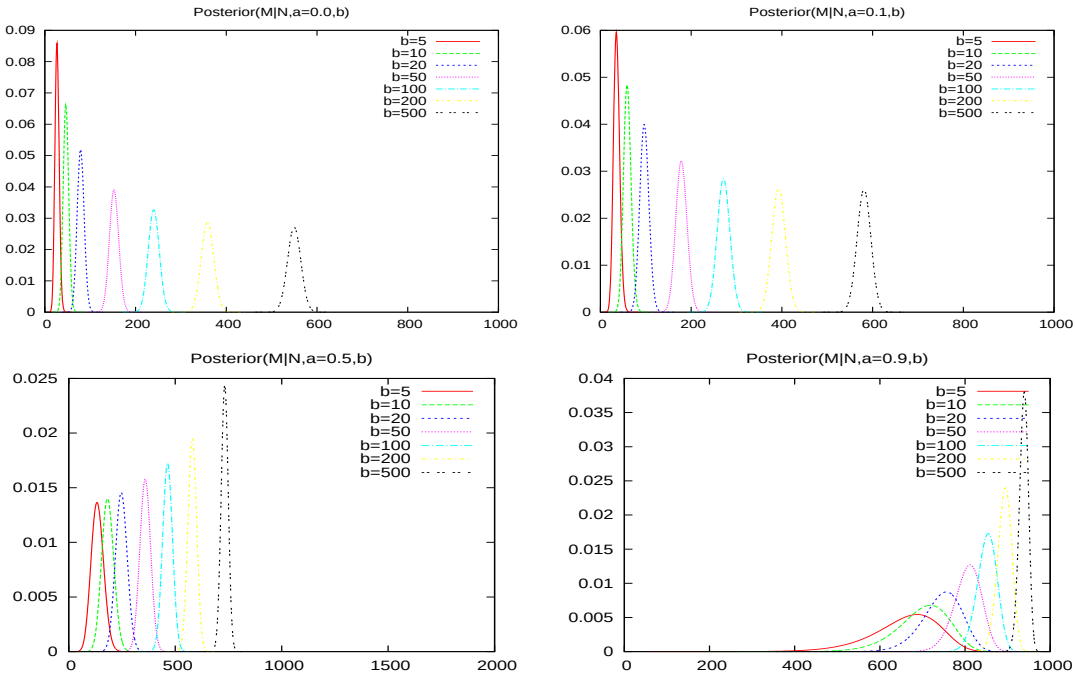

Figure 1: Posterior distribution on M given $N = 1000$ and different a.

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

The explicit closed form (*iii*) is new. Figure 1 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 2 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.

4.3 Convergence results

It is well known that expected partition size for the DP (PDP with $a = 0$) is $O(\log N)$ and for the PDP it has been shown to be $O(N^a)$ by [Teh06a]. 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].

Lemma 10 (Expected partition size) *In the context of Definition 3, if a sample I_N has the probability vector $\vec{p}$ distributed* a priori *according to PDD(a, b), the*

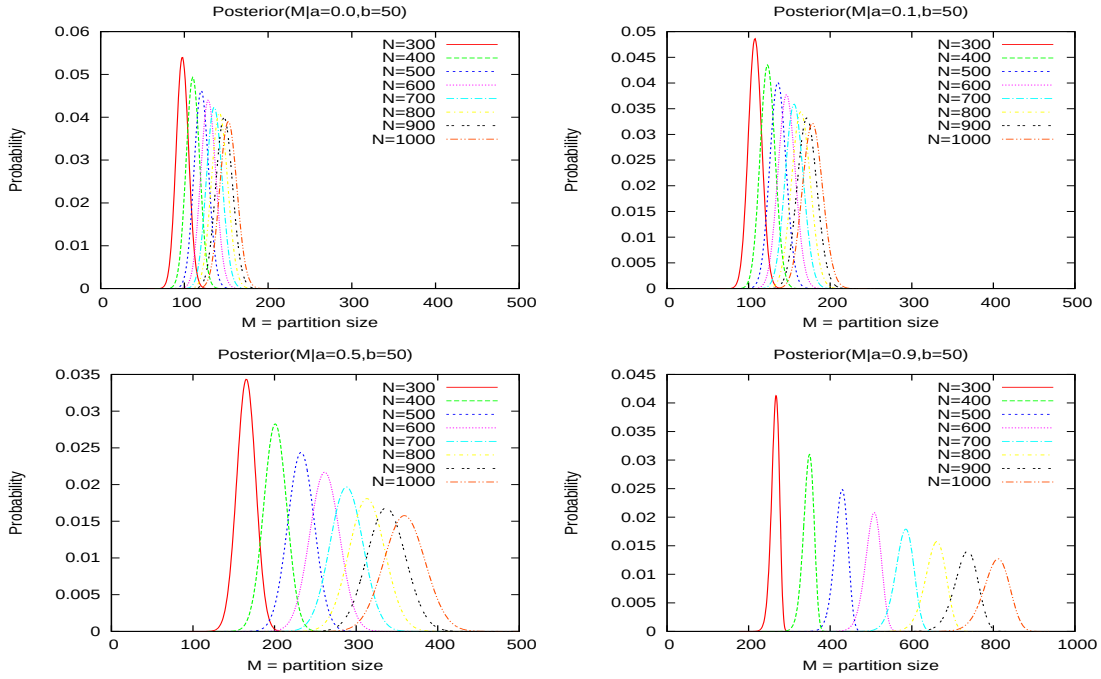

Figure 2: Posterior distribution on M for increasing N and fixed $b = 50$.

expected a posteriori M *for a sample of size* N *denoted* $\mathbb{E}_{\vec{p}|a,b,N}[M]$ *(and note the actual sample is unknown here, just its size* N *is known), when* $a > 0$ *is given by*

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

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 $a = 0$,

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

$$
\begin{aligned}
& + b^2(\psi_1(b+N) - \psi_1(b)) \\
\simeq \;& b \log\left(1 + \frac{N}{b}\right) \qquad\qquad \textit{for} \quad N > b \gg 0 \;,
\end{aligned}
$$

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 4.1.

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.

Lemma 11 (Upper bound on expected partition size) *Suppose an integer 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 3. 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 10 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 $r = \exp(-1/b)$.

5 Improper Priors

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[3]. 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[4]. 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:

Definition 12 (Improper prior for PDDs) *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 1. 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

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

[4]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.

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.

Lemma 13 (Consistency of projections) *In the context of Definition 12, 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 12.*

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

$$L_1(H, G) \;=\; \int_{\vec{p}} |H(\vec{p}) - G(\vec{p})| \, \mathrm{d}\, \vec{p} \, . \tag{6}$$

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 sample I_N can be made arbitrarily close to the corresponding proper posterior of the improper prior. Closeness here is measured by total variation distance.

Theorem 14 (Justifying improper prior) *Using the notation of Definitions 12 and 3, 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 15 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.

Lemma 15 (Some properties) *Using the improper prior for PDDs with parameters (a, b) and non-atomic base distribution $H(\cdot)$, the following holds:*

Proper posteriors (I): *Using the notation of Definitions 3 and 12, in the case of arbitrary samples I_N, and I_N^* be result of applying the standard renumbering. The posterior distribution given I_N^* is*

$$\left(p_1, ..., p_M, p_M^+\right) \,|\, I_N^* \;\sim\; \mathit{Dirichlet}(n_1 - a, ..., n_M - a, b + Ma) \; , \tag{7}$$

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

Proper posteriors (II): *Using the notation of Definition 4, in the case of arbitrary samples S_N, the posterior distribution given S_N is*

$$\begin{aligned} X_{N+1} \,|\, p_1, ..., p_M, p_M^+, S_N \;&\sim\; p_M^+ \, H(\cdot) + \sum_{m=1}^M p_m \delta_{X_m^*}(\cdot) \\ \left(p_1, ..., p_M, p_M^+\right) \,|\, S_N \;&\sim\; \mathit{Dirichlet}(n_1 - a, ..., n_M - a, b + Ma) \; , \end{aligned} \tag{8}$$

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

Sampling: *Using the notation of Definition 4, 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*

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

Stick-breaking: *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 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 Beta$(-a, b + ma)$.

The posterior formulation for PDPs corresponding to Equation (8) 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 p_k's in $\vec{p}$ for an improper prior for PDDs yields a PDD(a, b) distribution. Alternatively, one can replace the use of PDDs in Definition 2 with the improper prior for PDDs.

6 The Discrete Case

Now consider the case of discrete base distributions. If we have not been given the index sequences (the k's) in a sample, and the distribution $H(\cdot)$ is discrete, we can only guess what the k's might be. In this case, the partition structure is partially hidden. So for instance, consider the sample of words:

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

This can have the same partition structure as the example in Section 2, which has $I^*_N = 1, 2, 3, 1, 1, 4, 3$ with $N = 7$ However, since $H(\cdot)$ is discrete, it could be that $X^*_2 = X^*_{12}$ in Formula (2) and both are equal to "from". Some standard renumberings 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 are all finer than the standard one I^*_N. The definition below, *multiplicity*, measures the cardinality of the (unknown) set of $k's$ contributing to one observation X that occurs multiple times in the data.

Definition 16 (Multiplicity) *Consider Definitions 3 and 4, but now assume that the base distribution $H(\cdot)$ is discrete, so it may be that $X_k = X_l$ for $k \neq l$. For a given sample S_N, and consider the corresponding latent indices I_N. The* multiplicity *of the value $X \in S_N$ is defined as the size of the set $\{k_m : m = 1, ..., M,\ X^*_m = X\}$.*

Multiplicities are statistics from the latent indices I_N and are thus themselves latent. Continuing the example at the start of this section, suppose the latent I_N were given as $I_N = 12, 435, 7198, 12, 13, 35, 7198$, then $N = 7$, $M = 5$ and the distinct indices are $12, 435, 7198, 13, 35$. The counts $n_1 = n_3 = 2$ and all others are 1. All multiplicities are 1, except for the token "from" which has multiplicity 2 due to the indexes 12 and 13.

In the fully discrete 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 equal draws can be repeated. In these cases, 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 7, since $\text{PDP}(a, b, H)$ returns values from $\mathcal{X}$, but no indices. The following corollary of Lemmas 7 and 8 is a special case of [Teh06a, Equation (31)], there proven directly for the hierarchical PDP.

Corollary 17 (Evidence for discrete case) *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 4, 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 $0 \leq t_m \leq n_m$ and $t_m = 0$ if and only if $n_m = 0$. Then the joint probability of the sample and the multiplicities is:*

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

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

Notice the terms relevant in the formula for each t_m (after some simplifying)

$$b^{t_m \mathbb{1}_{a\equiv 0}} \left(a^{t_m} \Gamma(b/a + T) \right)^{\mathbb{1}_{a>0}} H(X^*_m)^{t_m} S^{n_m}_{t_m,a} \;, \tag{9}$$

where $\mathbb{1}_A$ has the value 1 if A is true, and 0 otherwise. This reflects the functional form for the posterior probabilities for the partition size M Equation (4), thus the analysis for that can be borrowed. A key distinction is term $H(X^*_m)^{t_m}$ which has the effect of discouraging multiplicity since invariably $H(X^*_k) \ll 1$. It is this term that would keep the multiplicity small.

To work with the discrete case, one needs to approximate the t_m. These can be sampled using Gibbs sampling and precomputed tables of the Stirling numbers $S^n_{t,a}$. Note it has been reported by Teh [Teh06b] that the multiplicities t_m in practice are quite small, and this holds for a near zero and b small. Figure 1 showed the posterior distribution of M for a given (a, b). It can be see that the expected range of M is rather narrow, thus in practice one may well want to fit hyperparameters (a, b) during training.

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.

Lemma 18 (Moments for the discrete case) *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*

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

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

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

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

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

Thus we can conclude following:

- When $a = 0$ for a finite discrete distribution and $b > 0$, we have that $\text{PDP}(0, b, \text{discrete}(\vec{\theta}))$ is well approximated by the $\text{Dirichlet}(b\vec{\theta})$. The two distributions agree in all the moments of order one to three.

- When $b = 0$ and $a \ll 1$, then we have that $\text{PDP}(a, 0, \text{discrete}(\vec{\theta}))$ is approximated by the $\text{Dirichlet}(a\vec{\theta})$. The two distributions differ by a factor of $O(a^2)$ in all the moments of order one to three..

Thus remarkably, in these cases, the PDP applied to finite discrete distributions is approximated by a proper Dirichlet. It is shown in the proof, however, that the approximation breaks down at the fourth order moments.

7 Conclusion

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, we recommend fitting one of the two parameters a or b of the PDP or PDD in practice.

For the infinite distribution on positive integers $\vec{p}$ underlying the PDP, which takes the form of a Poisson Dirichlet distribution, we showed that it corresponds to an infinite improper version of a regular Dirichlet distribution. This is a conceptual contribution: this means we can understand the distribution as having all the additivity and sampling properties we expect of a Dirichlet.

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 distribution on a discrete base distribution behaves rather like a Dirichlet distribution. This means that if $\vec{\mu} \sim \text{Dirichlet}_K(b\vec{\theta})$, then one can approximate it as $\vec{\mu} \sim \text{PDP}(0, b, \text{discrete}(\vec{\theta}))$ or $\vec{\mu} \sim \text{PDP}(b, 0, \text{discrete}(\vec{\theta}))$ and then remarkably the posterior with $\vec{\mu}$ integrated out is conjugate to the Dirichlet.

Acknowledgements. 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.

A Proof of Lemma 8

Build on the result from Lemma 7 using Definition 3. The formula of Lemma 7 also applies to I_N^*, the indices I_N with the standard renumbering applied, so

$p(k_1^*, k_2^*, ..., k_N^*)$, but remove the terms in $H(X_k^{\#})$. Using the notation of Definition 3 and this lemma, we get the form

$$p(k_1^*, ..., k_N^*) \;=\; \frac{(b|a)_M}{(b)_N} \prod_{m=1}^{M} \frac{\Gamma(n_m - a)}{\Gamma(1-a)} \;. \tag{10}$$

Now one can marginalise out the $k_1^*, ..., k_N^*$ 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 the standard renumbering, there are N choose $n_1, ..., n_M$, $C^N_{n_1,...,n_M}$ 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 with the standard renumbering with counts $n_1, ..., n_M$ are

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

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 (10) then, one does

$$\begin{aligned} p(M|N) \;&=\; \sum_{\sum_{m=1}^{M} n_m = N,\; n_m \geq 1} p(k_1^*, ..., k_N^*)\; C^N_{n_1,...,n_M} \prod_{m=1}^{M} \frac{n_m}{N - \sum_{i=1}^{m-1} n_i} \;. \\ &=\; \frac{(b|a)_M}{(b)_N} \; N! \sum_{\sum_{m=1}^{M} n_m = N,\; n_m \geq 1} \prod_{m=1}^{M} \left(\frac{\Gamma(n_m - a)}{\Gamma(n_m+1)\Gamma(1-a)} \frac{n_m}{N - \sum_{i=1}^{m-1} n_i} \right) \;. \end{aligned}$$

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

B Generalized Stirling Numbers

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

Theorem 19 (Expressions for Generalized Stirling Numbers) *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: $S(n,k) = \frac{1}{k!\,\beta^k} \sum_{j=0}^{k} \binom{k}{j} (-)^{k-j} (\beta j + r|-\alpha)_n$ *($\beta \neq 0$)*

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

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

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

C Proof of Theorem 9

The proof is based on (a) a recursion for $p(M|N)$, (b) the expressions for the generalized Stirling numbers in Appendix B, 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 8 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 $p(M_N = m)$.

(ii) and (iii) 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 19 reduces to recursion (i) of Theorem 9, 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 19 reduce to the first expression of (ii) and (iii) of Theorem 9 for $(\alpha,\beta,r) = (-1,-a,0)$, which shows that the expressions are equivalent.

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

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

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].

D Proof of Lemma 10

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:

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

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.

E Proof of Lemma 11

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.

F Proof of Lemma 13

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 12, 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 12. We can repeat this process recursively to verify consistency for any other sub-vector.

G Proof of Lemma 15

Proof sketch. 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 are $k^*_m = m$. 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 12 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.

H Proof of Theorem 14

Consider Definitions 12 and 3. Define G_δ in terms of its projection on the finite sub-spaces $\{p_1, p_2, ..., p_M\}$ for all M. Let

$$p(p_1, p_2, ..., p_M, p^+_M) \;\propto\; \left(p^+_M\right)^{b+M\,a-1} \prod_{m=1}^{M} p_m^{-a-1} \;, \tag{11}$$

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 12, $b > -a$, 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 13, 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:

$$Z_{a,b,M,\delta} \;:=\; \int_{C_{p,M}} \left(p^+_M\right)^{b+M\,a-1} \prod_{m=1}^{M} p_m^{-a-1} \mathrm{d}\,\vec{p}$$

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

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 13 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 15. To deal with the proper prior G_δ, the notion of partition size is needed, as given in Definition 3. 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 15 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 15 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 15. 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

$$
\frac{1}{2} d_1(P_\delta, P_0) \ := \ \sup_A |P_\delta[A] - P_0[A]|
$$

$$
\begin{aligned}
&= \frac{1}{2}\int_{G_0} |P_\delta - P_0| d\vec{p} \\
&= \frac{1}{2}\int_{G_\delta} |P_\delta - P_0| d\vec{p} + \frac{1}{2}\int_{G_0 \setminus G_\delta} |P_\delta - P_0| d\vec{p} \\
&= \frac{1}{2}\int_{G_\delta} P_\delta d\vec{p} - \frac{1}{2}\int_{G_\delta} P_0 d\vec{p} + \frac{1}{2}\int_{G_0 \setminus G_\delta} P_0 d\vec{p} \\
&= 1 - \int_{G_\delta} P_0 d\vec{p} \;=\; 1 - \frac{Z_{a,b,M_N,\delta}}{Z_{a,b,M_N,0}} \int_{G_\delta} P_\delta d\vec{p} \\
&= 1 - \frac{Z_{a,b,M_N,\delta}}{Z_{a,b,M_N,0}} \to 0 \quad \text{for} \quad \delta \to 0
\end{aligned}
$$

This implies convergence in distribution.

I Proof of Corollary 17

Proof sketch. In the general case where each draw from $H(\cdot)$ is not necessarily almost surely distinct, the formula of Lemma 7 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 of the standard renumbering, done for Lemma 8, and an identical result can be applied.

J Proof of Lemma 18

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

$$\mathbb{E}_{\vec{q},\vec{X}}\left[\left(\sum_l q_l 1_{X_l=k_1} - \theta_{k_1}\right)\left(\sum_l q_l 1_{X_l=k_2} - \theta_{k_2}\right)\left(\sum_l q_l 1_{X_l=k_3} - \theta_{k_3}\right)\right]$$

$$= \mathbb{E}_{\vec{q},\vec{X}}\left[\sum_{l_1,l_2,l_3} q_{l_1} q_{l_2} q_{l_3} \left(1_{X_{l_1}=k_1} - \theta_{k_1}\right)\left(1_{X_{l_2}=k_2} - \theta_{k_2}\right)\left(1_{X_{l_3}=k_3} - \theta_{k_3}\right)\right]$$

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:

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

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 $k_1, k_2, \ldots$. 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 10 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.

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.

[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.

[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.

[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. *Lecture Notes-Monograph Series*, 30:245–267, 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.

[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.

[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.