Forty-Eighth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 1, 2010

Redundancy of Exchangeable Estimators

(Invited Paper)

N. P. Santhanam
Department of Electrical Engineering
University of Hawaii at Manoa
2540 Dole Street
Honolulu, HI 96822, USA
nsanthan@hawaii.edu

M. M. Madiman
Department of Statistics
Yale University
24 Hillhouse Avenue
New Haven, CT 06511, USA
mokshay.madiman@yale.edu

A. D. Sarwate
ITA CALIT2
UC San Diego
9500 Gilman Dr. MC 0447
La Jolla, CA 92093, USA
asarwate@ucsd.edu

***Abstract*—Exchangeable random partition processes provide a framework for statistical inference in large alphabet scenarios from a Bayesian perspective. On the other hand, the notion of the *pattern* of a sequence provides a framework for data compression in large alphabet scenarios. Owing to the relationship between data compression and parameter estimation, both these approaches are related. Motivated by the possibilities of cross-fertilization, we examine the redundancy of Bayes estimators (specifically those that emerge from the "Chinese restaurant processes") in the setting of unknown discrete alphabets from a universal compression point of view. In particular, we identify relations between alphabet sizes and sample sizes where the redundancy is small– and hence, characterize useful regimes for these estimators.**

I. Introduction

For a number of statistical inference problems of significant contemporary interest, such as text classification, language modeling, and DNA microarray analysis, there is a necessity to perform inference based on observed sequences of symbols, where the sequence length or sample size is comparable or even smaller than the set of symbols, the *alphabet*. For instance, language models for speech recognition estimate distributions over English words using text examples much smaller than the vocabulary.

Inference in this setting has received a lot of attention, from Laplace [1], [2], [3] in the 18th century, to Good [4] in the mid-20th century, to an explosion of work in the statistics [5], [6], [7], [8], [9], [10], [11], [12], [13], information theory [14], [15], [16], [17], [18], [19] and machine learning [20], [21], [22], [23] communities in the last few decades. While a major strand in the information theory literature on the subject has been based on the notion of *patterns*, a major strand in the statistical literature has been based on the notion of *exchangeability*. Our goal in this note is to study the redundancy (an information theoretic criterion) of exchangeable estimators that naturally arise in the nonparametric Bayesian literature.

In probability and nonparametric Bayesian statistics, Kingman advocated the use of *exchangeable random partitions* to accommodate the analysis of data from an alphabet that is not bounded or known in advance [11]. The most popular exchangeable random partition process is the Chinese restaurant process. In this paper we analyze the redundancy of pattern probability estimators derived from popular Bayesian models. In particular, we investigate estimators derived from the "Chinese restaurant process" and the "Poisson-Dirichlet priors".

If the alphabet size can be arbitrarily large with respect to the sample size n, we show that the estimators do not have $o(n)$ redundancy. However, for sample patterns whose number of unique elements is bounded, we can derive much tighter bounds on the redundancy. In this setting the two-parameter Poisson-Dirichlet (or Pitman-Yor) estimator is superior to the estimator derived from the Chinese restaurant process. However, we show that a mixture of Chinese restaurant process estimators is weakly universal.

In order to describe our results, a variety of notions from the literature of diverse communities is required. In Section II, we describe this preliminary material and place it in context, and in Section III we describe our main results.

II. Preliminaries

Let $\mathcal{I}_k$ denote the set of all probability distributions on alphabets of size k, $\mathcal{I}_\infty$ be all probability distributions on countably infinite alphabets, and let

$$\mathcal{I} = \mathcal{I}_\infty \cup \bigcup_{k\geq 1} \mathcal{I}_k \tag{1}$$

be the set of all discrete distributions irrespective of support and support size.

For a fixed p, let $x_1^n = (x_1, x_2, \ldots, x_n)$ be a sequence drawn *i.i.d.* according to p. We denote the *pattern* of x_1^n by ψ^n. The pattern is formed by taking $\psi_1 = 1$ and

$$\psi_i = \begin{cases} \psi_j & x_i = x_j,\ j < i \\ 1 + \max_{j<i} \psi_j & x_i \neq x_j,\ \forall j < i \end{cases} \tag{2}$$

For example, the pattern of $x_1^7 = \mathsf{FEDERER}$ is $\psi_1^7 = 1232424$. Let ψ^n be the set of all patterns of length n. We write $p(\psi^n)$ for the probability that a length-n sequence generated by p has pattern ψ^n.

A. Exchangeable partition processes

An *exchangeable random partition* refers to a sequence $(C_n : n \in \mathbb{N})$, where C_n is a random partition of the set $[n] = \{1, 2, \ldots, n\}$, satisfying the following conditions: (i) the probability that C_n is a particular partition depends only on

the vector $(s_1, s_2, \ldots, s_n)$, where s_k is the number of parts in the partition of size k, and (ii) the realizations of the sequence are consistent in that all the parts of C_n are also parts of the partition C_{n+1}, except that the new element $n+1$ may either be in a new part of C_{n+1} by itself or has joined one of the existing parts of C_n.

If one has a data sequence $X_1, \ldots, X_n$ from a discrete alphabet, one can partition the set $[n]$ into component sets corresponding to the symbols of the alphabet that have appeared, where each part or component of the partition corresponds to the set of locations at which a particular symbol appears. When such partitions are generated from *i.i.d.* data X_i, the corresponding sequence of random partitions is called a *paintbox process*.

The remarkable Kingman representation theorem [8] then asserts that the probability measure induced by any exchangeable random partition is a mixture of paintbox processes, where the mixture is taken using a probability measure ("prior" in Bayesian terminology) on the class of paintbox processes. Since each paintbox process corresponds to a discrete probability measure (the one such that *i.i.d.* X_i drawn from it produced the paintbox process), the prior may be viewed as living on the set of probability measures on a countable alphabet. (Actually, for technical reasons, the alphabet is assumed to be hybrid, with a discrete part as well a continuous part, and also one needs to work with the space of *ordered* probability vectors, but we ignore this for ease of description.)

B. Dirichlet priors and Chinese restaurant processes

Not surprisingly, special classes of priors give rise to special classes of exchangeable random partitions. One particularly nice class of priors on the set of probability measures on a countable alphabet is that of the Poisson-Dirichlet priors [24], [5], [25] (sometimes called Dirichlet processes since they live on the infinite-dimensional space of probability measures and generalize the usual finite-dimensional Dirichlet distribution).

The *Chinese restaurant process* (or CRP) is related to the so-called Griffiths-Engen-McCloskey (GEM) distribution with parameter θ, denoted by $\mathsf{GEM}(\theta)$. Consider $W_1, W_2, \ldots$ drawn i.i.d. according to a $\mathsf{Beta}(1, \theta)$ distribution, and set

$$p_1 = W_1 \tag{3}$$

$$p_i = W_i \prod_{j<i} (1 - W_i) \qquad \forall i > 1 \tag{4}$$

This can be interpreted as follows: take a stick of unit length and break it into pieces of size W_1 and $1 - W_1$. Now take the piece of size $1 - W_1$ and break off a W_2 fraction of that. Continue in this way. The resulting lengths of the sticks create a distribution on a countably infinite set. The distribution of the sequence $p = (p_1, p_2, \ldots)$ is the $\mathsf{GEM}(\theta)$ distribution.

Remark 1: Let π denote the elements of p sorted in decreasing order so that $\pi_1 \geq \pi_2 \geq \cdots$. Then the distribution of π is the Poisson-Dirichlet distribution $\mathsf{PD}(\theta)$ as defined by Kingman.

Another popular class of distributions on probability vectors is the *Pitman-Yor family* of distributions [26], also known as the two-parameter Poisson-Dirichlet family of distributions $\mathsf{PD}(\alpha, \theta)$. The two parameters here are a discount parameter $\alpha \in [0, 1]$, and a strength parameter $\theta > -\alpha$. The distribution $\mathsf{PD}(\alpha, \theta)$ can be generated in a similar way as the Poisson-Dirichlet distribution $\mathsf{PD}(\theta) = \mathsf{PD}(0, \theta)$ described earlier. Let $W_1, W_2, \ldots$ be drawn i.i.d. according to a $\mathsf{Beta}(1-\alpha, \theta+n\alpha)$ distribution, and again set

$$\tilde{p}_1 = W_1 \tag{5}$$

$$\tilde{p}_i = W_i \prod_{j<i} (1 - W_i) \qquad \forall i > 1 \tag{6}$$

A similar "stick-breaking" interpretation holds here as well. Now let p be equal to the sequence $\tilde{p}$ sorted in descending order. The distribution of p is $\mathsf{PD}(\alpha, \theta)$.

C. Patterns and partitions

The following interesting observation (which is rather easy to make once one is aware of the definitions of the relevant terms) emerged through discussions among participants in the American Institute of Mathematics workshop on "Permanents and modeling probability distributions" in September 2009.

Observation 1: There is a bijection between Kingman's paintbox processes and patterns of i.i.d. processes.

To put this observation in context, let us observe that the information-theoretic approach to large alphabet problems using patterns has been tackled primarily from a frequentist perspective, whereas the statistics community has approached the analogous problems primarily from a Bayesian point of view. There are hence differences in the kinds of estimators that have been proposed.

In this paper, we consider some implications of Observation 1. In particular, motivated by the correspondence between paintbox processes and patterns, we consider a family of mixture codes for patterns that correspond to the so-called Ewens sampling formula, and show that inspite of its usefulness for some inference tasks, the redundancies it yields for data compression are much larger than necessary.

D. Pattern probability estimators

Given a sample x_1^n with pattern ψ^n we would like to produce an *pattern probability estimator*. This is a function of the form $q(\psi_{n+1}|\psi^n)$ that assigns a probability of seeing a symbol previously seen in ψ^n as well as a probability of seeing a new symbol. In this paper we will investigate two different pattern probability estimators based on Bayesian models.

The *Ewens sampling formula* [27], with origins in theoretical population genetics, is a formula for the probability mass function of a marginal of a Chinese restaurant process corresponding to a fixed population size. In other words, it specifies the probability of an exchangeable random partition of $[n]$ that is obtained when one uses the Poisson-Dirichlet $\mathsf{PD}(\theta)$ prior to mix paintbox processes. Due to the equivalence between patterns and exchangeable random partitions, it implies that

the probability of a pattern ψ_1^n is

$$q_\theta^{CRP}(\psi_1,\ldots,\psi_n) = \frac{\theta^m}{\theta(\theta+1)\cdots(\theta+n-1)}\prod_{\mu=1}^{n}[(\mu-1)!]^{\phi_\mu} \quad (7)$$

where ϕ_μ is the number of symbols that appear μ times in ψ_1^n and $m = \sum \phi_\mu$ is the number of distinct symbols in ψ_1^n. In particular, the predictive distribution associated to the Ewens sampling formula or Chinese restaurant process is

$$q_\theta^{CRP}(\psi|\psi_1,\ldots,\psi_n) = \begin{cases} \frac{\mu}{n+\theta} & \psi \text{ appeared } \mu \text{ times} \\ & \text{in } \psi_1,\ldots,\psi_n; \\ \frac{\theta}{n+\theta} & \psi \text{ corresponds to new.} \end{cases} \quad (8)$$

More generally, one can define the Pitman-Yor predictor as

$$q_{\alpha,\theta}^{PY}(\psi|\psi_1,\ldots,\psi_n) = \begin{cases} \frac{\mu-\alpha}{n+\theta} & \psi \text{ appeared } \mu \text{ times} \\ & \text{in } \psi_1,\ldots,\psi_n; \\ \frac{\theta+m\alpha}{n+\theta} & \psi \text{ corresponds to new.} \end{cases} \quad (9)$$

where m is the number of distinct symbols in ψ_1^n. In this case, the probability assigned to a pattern ψ_1^n is

$$q^{PY}(\psi_1,\ldots,\psi_n) = \frac{(\theta+\alpha)(\theta+2\alpha)\cdots(\theta+m\alpha)}{\theta(\theta+1)\cdots(\theta+n-1)}\prod_{\mu=1}^{n}\left(\frac{\Gamma(\mu-\alpha-1)}{\Gamma(1-\alpha)}\right)^{\phi_\mu}. \quad (10)$$

E. Worst-case and average redundancy

How should we measure the quality of a pattern probability predictor q? We investigate two criterion here: the worst-case and the average-case redundancy. The *redundancy* of q on a given pattern ψ^n is

$$R(q) \stackrel{\text{def}}{=} \sup_{p\in\mathcal{I}} \log \frac{p(\psi^n)}{q(\psi^n)}, \quad (11)$$

The *worst-case redundancy* of q is defined to be

$$\hat{R}(q) \stackrel{\text{def}}{=} \sup_{p\in\mathcal{I}} \max_{\psi^n\in\Psi^n} \log \frac{p(\psi^n)}{q(\psi^n)}, \quad (12)$$

Recall that $p(\psi^n)$ just denotes the probability that a length-n sequence generated by p has pattern ψ^n—it is unnecessary to specify the support here.

The *average-case redundancy* replaces the $\max$ over patterns with an expectation over p:

$$\bar{R}(q) \stackrel{\text{def}}{=} \sup_{p\in\mathcal{I}} \mathbb{E}_p\left[\log \frac{p(\psi^n)}{q(\psi^n)}\right] \quad (13)$$

$$= \sup_{p\in\mathcal{I}} D\left(p \parallel q\right). \quad (14)$$

That is, the average-case redundancy is nothing but the worst-case Kullback-Leibler divergence between the distribution p and the predictor q.

A pattern probability estimator is considered "good" if the worst-case or average-case redundancies are sublinear in n. Succintly put, it can be proved that if the redundancy grows sublinear in n, the underlying probability of a sequence can be estimated accurately almost surely.

III. Redundancy results

We now describe our main results on the redundancy of estimators derived from the prior distributions on $\mathcal{I}$. Proof outlines are given here– full proofs are deferred to the journal version of this work.

A. Chinese restaurant predictors

Previously [28] it was shown by some of the authors that the worst-case and average-case redundancies for the CRP estimator are both $\Omega(n\log n)$, which means it is neither strongly nor weakly universal. However, this estimator is not quite as bad as this result might suggest, and a simple twist yields an estimator that is weakly universal.

Our first new result is for the CRP estimator when we have a bound on the number m of distinct elements in the pattern ψ_1^n.

Theorem 1 (Redundancy): Consider the estimator $q_\theta^{CRP}(\psi_1^n)$. Then for sufficiently large n and for patterns ψ_1^n whose number of distinct symbols m satisfies

$$m \le C\cdot\frac{n}{\log n}(\log\log n)^2, \quad (15)$$

the redundancy of the predictor $q_\theta^{CRP}(\psi_1^n)$ with $\theta = m/\log n$ satisfies:

$$\log\frac{p(\psi_1^n)}{q_\theta^{CRP}(\psi_1^n)} \le 3C\cdot\frac{n(\log\log n)^3}{\log n} = o(n). \quad (16)$$

Proof idea: This follows from a series of combinatorial upper bounds and by using the assumptions $\theta = m/\log n$ and $m = \Theta\left(\frac{n}{\log n}(\log\log n)^2\right)$. ∎

This theorem is slightly dissatisfying, since it requires us to have a bound on m. It turns out that by taking mixtures of CRP estimators we can arrive at an estimator that is weakly universal. That is, let $\tilde{q}_{m,n}^{CRP}$ be the CRP estimator with $\theta = m/\log n$. Then define

$$q^*(\cdot) = \sum_{n,m} c_{n,m}\tilde{q}_{m,n}^{CRP}(\cdot) \quad (17)$$

for a set of positive coefficients $c_{n,m}$ that sum to 1:

$$\sum_{n,m} c_{n,m} = 1. \quad (18)$$

We can, for example, choose $c_{m,n} = \frac{36}{m^2n^2\pi^4}$ It is clear that q^* is a pattern probability estimator.

Lemma 1: For all discrete i.i.d. processes P with entropy rate H, let M_n be the random variable counting the number of distinct symbols in a sample of length n drawn from P. The following bound holds

$$\mathbb{E}\left[M_n\right] \le \frac{nH}{\log n} + 1. \quad (19)$$

Proof idea: This a direct bound from the definition of M_n. ∎

Theorem 2 (Weak universality): For all discrete i.i.d. processes $p \in \mathcal{I}$ with finite entropy rate,

$$\lim_{n\to\infty} D\left(p \parallel q^*\right) = 0. \quad (20)$$

That is, q^* is weakly universal.

Proof idea: The result follows from repeating the analysis of Theorem 1 and using Markov's inequality applied to Lemma 1. ∎

What the preceding theorem shows is that the mixture of CRP estimators q^* is weakly universal. However, q^* is not itself a CRP estimator.

B. Pitman-Yor predictors

We now turn to the more general class of Pitman-Yor predictors. We can obtain a similar result as for the CRP estimator, but we can handle patterns with $m = o(n)$.

Theorem 3 (Worst-case redundancy): Consider the estimator $q^{PY}_{\alpha,\theta}(\psi_1^n)$. Then for sufficiently large n and for patterns ψ_1^n whose number of distinct symbols m satisfies $m = o(n)$, the redundancy of the predictor $q^{PY}(\psi_1^n)$ with $\theta = m/\log n$ satisfies:

$$\log \frac{p(\psi_1^n)}{q^{PY}_{\alpha,\theta}(\psi_1^n)} = o(n). \tag{21}$$

It is well known that the Pitman-Yor process can produce patterns whose relative frequency is 0, e.g. the pattern $1^k 23\cdots(n-k)$. Therefore it is not surprising that the worst-case redundancy and average case redundancies can be bad. However, the redundancy of $\Theta(n)$ is significantly better than the lower bound of $\Omega(n \log n)$ proved in [28] for Chinese restaurant processes, as the following Theorem shows.

Theorem 4 (Redundancies): Consider the estimator $q^{PY}_{\alpha,\theta}(\psi_1^n)$. Then for sufficiently large n, the worst-case redundancy and average case redundancy satisfy:

$$\hat{R}(q^{PY}_{\alpha,\theta}) = \Theta(n) \tag{22}$$

$$\bar{R}(q^{PY}_{\alpha,\theta}) = \Theta(n). \tag{23}$$

That is, $q^{PY}_{\alpha,\theta}$ is neither strongly nor weakly universal.

Proof ideas: The lower bound follows from considering the singleton and a uniform distribution. The upper bound is standard. ∎

IV. Conclusions and future work

In this note we investigated the worst-case and average-case redundancies of pattern probability estimators derived from priors on $\mathcal{I}$ that are popular in Bayesian statistics. Both the CRP and Pitman-Yor estimators give a vanishing redundancy per symbol for patterns whose number of distinct symbols m is sufficiently small. The Pitman-Yor estimator requires only that $m = o(n)$, which is an improvement on the CRP. However, when m can be arbitrarily large (or the alphabet size is arbitrarily large) the worst-case and average-case redundancies do not scale like $o(n)$. Here again, the Pitman-Yor estimator is superior, in that the redundancies scale like $\Theta(n)$ as opposed to the $\Omega(n \log n)$ for the CRP estimator. While these results show that these estimators are not strongly universal, we constructed a mixture of CRP process (which is not itself a CRP estimator) that is weakly universal.

On the other hand, one of the estimators derived in [16] is exchangeable and has near optimal worst case redundancy, growing as $O(\sqrt{n})$. From Kingman's results, this estimator can be obtained using a prior on $\mathcal{I}$—however, this prior is yet unknown. Finding this prior may potentially reveal new interesting classes of priors other than the Poisson-Dirichlet priors.

Acknowledgments

The authors thank Persi Diaconis, Miroslav Dudik, Fan Chung Graham, Ron Graham, Susan Holmes, Olgica Milenkovic, Alon Orlitsky, Ofer Shayevitz, Krishna Viswanathan, Aaron Wagner and Junan Zhang for helpful conversations related to the paper, as well as the American Institute of Mathematics and NSF for sponsoring a workshop on probability estimation topics. In particular, the authors also thank Alon Orlitsky and Krishna Viswanathan for being co-organizers of the above workshop along with one of the authors.

N. Santhanam was supported by a startup grant from the University of Hawaii and NSF Grant CCF-1018984. M. Madiman was supported by a Junior Faculty Fellowship from Yale University. A.D. Sarwate was supported by the California Institute for Telecommunications and Information Technology (CALIT2) at the University of California, San Diego.

References

[1] P. Laplace, *Philosphical essays on probabilities*, Translated by A. Dale from the 5th (1825) ed. Springer Verlag, New York, 1995.

[2] A. DeMorgan, *An Essay on Probabilities, and on their Application to Life Contingencies and Insurance Offices,*. Longman et at., London., 1838.

[3] A. D. Morgan, *Encyclopedia Metropolitana, Vol 2 Pure Mathematics, B.* Fellowes et. al., London, 1845, ch. Theory of Probabilities, pp. 393–490.

[4] I. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237—264, December 1953.

[5] D. Blackwell and J. B. MacQueen, "Ferguson distributions via pólya urn schemes," *Ann. Statist*, vol. 1, pp. 353–355, 1973.

[6] J. Kingman, "Random discrete distributions," *Journal of the Royal Statistical Society*, vol. B37, pp. 1–22, 1975.

[7] ——, "Random partitions in population genetics," *Proceedings of the Royal Society*, vol. A361, pp. 1–20, 1978.

[8] ——, "The representation of partition structures," *Journal of the London Mathematical Society*, vol. 18, pp. 374–380, 1978.

[9] P. Diaconis and D. Freedman, *R. C. Jeffrey (ed.), Studies in Inductive Logic and Probability, Vol. 2*. University of California Press, Berkeley and Los Angeles, 1980, ch. De Finetti's Generalizations of Exchangeability, pp. 233–50.

[10] D. J. Aldous, *Lecture Notes in Mathematics 1117*. P. L. Hennequin (ed.), Ecole d' tg de Probabilitds de Saint-Flour XII1 - 1983, 1985, ch. Exchangeability and Related Topics, pp. 1–198.

[11] S. Zabell, "Predicting the unpredictable," *Synthese*, vol. 90, pp. 205–232, 1992.

[12] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probability Theory and Related Fields*, vol. 102, pp. 145–158, 1995.

[13] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," Dept. Statistics, U.C. Berkeley, Tech. Rep. Technical Report 433, 1995, also in The Annals of Probability.

[14] B. Clarke and A. Barron, "Information theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453—471, May 1990.

[15] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37—60, 1994.

[16] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469—1481, July 2004.

[17] ——, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427—431, October 17 2003, see also *Proceedings of the* 44th *Annual Symposium on Foundations of Computer Science*, October 2003.

[18] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the Good-Turing estimator," IEEE Int. Symp. Inf. Theor. Proc., pp. 2526–30, 2006.

[19] B. Ryabko, "Compression based methods for non-parametric online prediction, regression, classification and density estimation," *Festschrift in Honor of Jorma Rissanen on the occasion of his 75th birthday*, pp. 271–288, 2008.

[20] A. Nadas, "Good, Jelinek, Mercer, and Robins on Turing's estimate of probabilities," *American Journal of Mathematical and Management Sciences*, vol. 11, pp. 229—308, 1991.

[21] W. Gale and K. Church, "What is wrong with adding one?" in *Corpus based research into language*, N. Oostdijk and P. de Haan, Eds. Rodopi, Amsterdam, 1994, pp. 189—198.

[22] D. McAllester and R. Schapire, "On the convergence rate of Good Turing estimators," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[23] E. Drukh and Y. Mansour, "Concentration bounds on unigrams language model," in *17th Annual Conference on Learning Theory*, 2004, pp. 170—185.

[24] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[25] R. Ramamoorthi and K. Srikanth, *Encyclopedia of Statistical Sciences*. John Wiley and sons, New York, 2007, ch. Dirichlet processes.

[26] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annnals of Probability*, vol. 25, no. 2, pp. 855–900, 1997.

[27] W. J. Ewens, "The sampling theory of selectively neutral alleles," *Theoretical Population Biology*, vol. 3, pp. 87–112, 1972.

[28] N. Santhanam and M. Madiman, "Patterns and exchangeability," in *Proceedings of the IEEE Symposium on Information Theory*, Austin, Texas, USA, June 2010.