Scribe Notes

11/30/07

Lauren Hannah reviewed salient points from "Some Developments of the Blackwell-MacQueen Urn Scheme" by Jim Pitman. The paper links existing ideas on Dirichlet Processes (Chinese Restaurant Process, stick breaking construction, etc.) to the Poisson-Dirichlet distribution. Lauren reviewed measure theory, described the Poisson-Dirichlet distribution, and reviewed theorems from the paper. Peter Frazier then presented a novel extension from a theorem in the paper to Dirichlet hyperparameters.

1 Measure theory review

E is a state space and $\mathcal{E}$ is a σ-algebra (a collection of subsets) on E if

- $A \in \mathcal{E} \Rightarrow E \setminus A \in \mathcal{E}$, i.e. the collection of subsets $\mathcal{E}$ is closed under complements.
- $A_1, A_2, \ldots \in \mathcal{E} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$, i.e. the collection of subsets $\mathcal{E}$ is closed under countable unions.

Examples

- $\mathcal{E} = \{\emptyset, E\}$, i.e. the trivial σ-algebra
- $\mathfrak{B}_{()}$, i.e. the Borel σ-algebra generated by the collection of all open sets, most often on the real line $\mathbb{R}$, denoted by $\mathfrak{B}_{\mathbb{R}}$

$(E, \mathcal{E})$ is a measurable space. μ is the function $\mathcal{E} \mapsto \bar{\mathbb{R}}_+ : [0, \infty]$. μ is a measure if

- $\mu(\emptyset) = 0$

- $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$ if all subsets A_i and A_j are pairwise disjoint $\forall i \neq j$.

1.1 Lebesgue Measure

$\mathbb{R}$ = length of interval. $\mathbb{R}^2$ = area. Usually denoted by Leb(A)

1.2 Dirac Measure

$Dirac : \delta_x(A) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$, i.e. a point measure [or delta measure] at point x.

2 Random measures

Recall that a random variable X is defined as a [deterministic] function that maps the outcome to a state space, e.g. $X = f : \Omega \mapsto \mathbb{R}$. Assume state space $(E, \mathcal{E})$ and outcome [or sample] space $(\Omega, \mathcal{H})$. $M : \Omega \times \mathcal{E} \mapsto \bar{\mathbb{R}}_+$ is a random measure if

- For fixed $A \in \mathcal{E}$, $\omega \mapsto M(\omega, A)$ is a random variable.
- For fixed $\omega \in \Omega$, $A \mapsto M(\omega, A)$ is a measure on E.

To understand the equations above it helps to temporarily drop all notions of randomness. In a purely deterministic world, we have two subsets ω and A. We assign a mapping from ω, a "state of the world" to A, our subset of $\mathcal{E}$, the σ-algebra on the state space that we care about in a random experiment. Furthermore, we additionally have another function $M(\omega, A)$ that maps any combination of ω and A to the positive portion of the real line, $\bar{\mathbb{R}}_+$. In formal measure-theory terms, M is a *transition kernel* from the product space $\Omega \times E$ to $\bar{\mathbb{R}}_+$. See figs. 1, 2, and 3.

"Randomness" comes about by stating that some other agent fixed ω in advance, yielding the observed A and corresponding $M(\omega, A)$. [1]

[1] or rather, as Prof. Cinlar likes to say, the Greek Goddess of chance, Tyche, fixes ω for you. "Doesn't it seem unnecessarily complex and unproductive to describe the source of randomness in a model through ω and then map it to another variable of interest, A? Why not just represent probabilities, etc. directly on the values of the variable of interest,

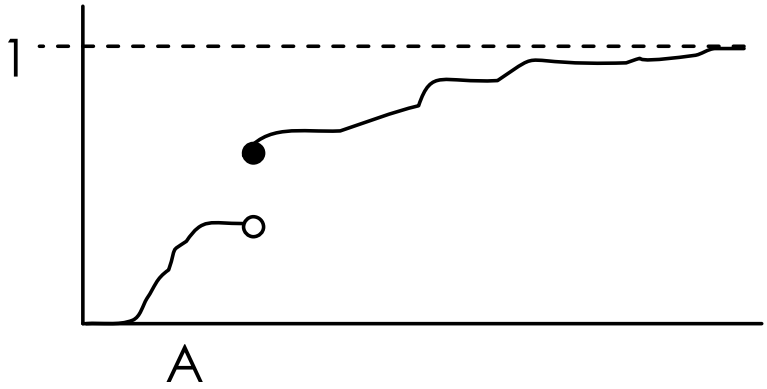

Figure 1: Fixing $\omega \in \Omega$. The resulting cdf of $M(\omega, \cdot)$ on $(E, \mathcal{E})$ is depicted here. The cdf is a deterministic measure dependent on ω.

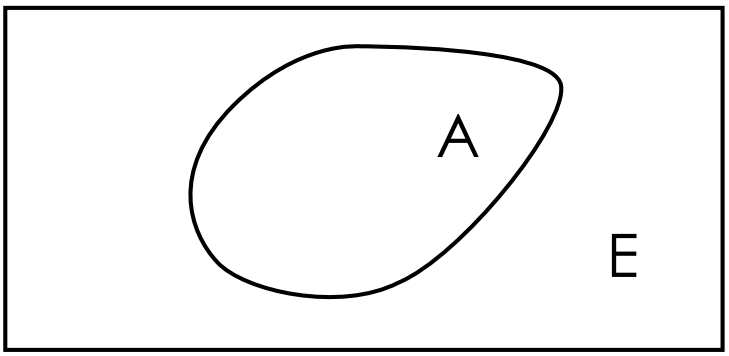

Figure 2: Fixing $A \in \mathcal{E}$. The area of A is a random variable dependent on the underlying measure, which itself depends on ω.

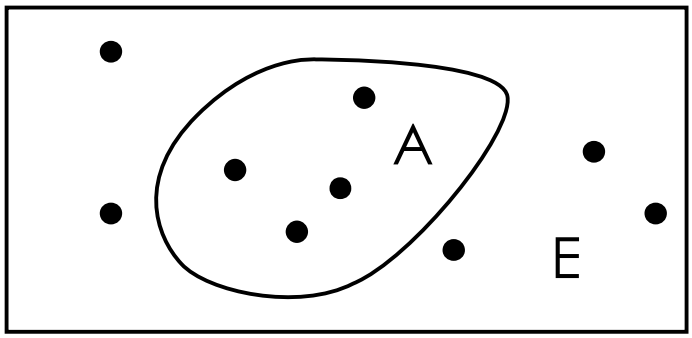

Figure 3: Fixing $A \in \mathcal{E}$. Let $M(\omega, A) = \text{Poisson}(\lambda)$. The measure of A [# dots in A, here 4] is a Poisson randomly distributed variable. The dot pattern on E is a realization of outcome ω.

An example of a random measure that we've seen is a Dirichlet Random Measure *DRM*. In a *DRM*, realizations of M are almost surely purely atomic measures on E. Fixing ω while varying A, we obtain a $\bar{\mathbb{R}}_+$-valued scalar $M(\omega, A)$ for each $A \in \mathcal{E}$, which is a deterministic measure, here a probability distribution, on $(E, \mathcal{E})$. This would look like fig. 4. Similarly [as is often done in Poisson random measures, although not commonly with Dirichlet Processes] we can isolate a specific subset $A \in \mathcal{E}$, i.e. restrict ourselves to

e.g. through a probability density function?" Well, the relationship between states of the world and experiment realizations may be more complicated than just a 1:1 bijection, e.g. consider an experiment where the cardinality of $\mathcal{H}$ and $\mathcal{E}$ differ substantially: many states of the world influencing the outcome of a binary-valued variable of interest vs. only two states of the world influencing the outcome. As more complex issues arise, one finds additional comforts modeling a random experiment in a measure-theoretical terms. Joint probabilities and conditional expectations may be interpreted as the sequential composition of functions, mapping a value from one space to another, and to the next, and so on. Thus, by positing the existence of ω c. 1920, the forefathers of probability theory (Kolmogorov et al.) allowed us to retain deterministic tools of math to analyze seemingly random or undeterministic systems. Flowery notions of "chance" and "dice throws" were violently usurped, seemingly overnight, by the strict formalism of measure theory and calculus - the very elements of probability theory as we know them today. One can only imagine the bewilderment of haggardly gambling-types upon initial confrontation with a presumably equally-bewildered camp of geeky Russian mathematicians. This mixing phenomenon still resonates within the halls of the Princeton Graduate School, a.k.a "Where the Nerd World Meets the Third World."

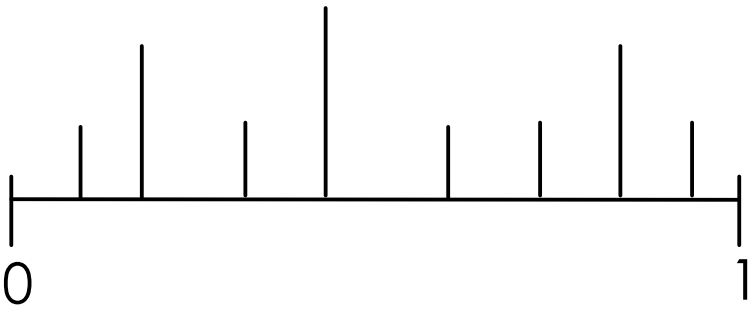

Figure 4: DRM realization fixing $\omega \in \Omega$. Discrete stick-valued probability atoms sitting on $A \in \mathcal{E}$ having $\sum = 1$.

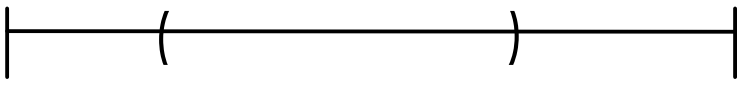

Figure 5: Fixing $A \in \mathcal{E}$.

only a portion of our state space, while varying ω. This yields a $\bar{\mathbb{R}}_+$-valued scalar random variable $\omega \mapsto M(\omega, A)$. See fig. 5.

3 Poisson random measures

Let ν be a measure on $(E, \mathcal{E})$. A random measure N is a *Poisson random measure* with mean measure ν if

- $A \in \mathcal{E}, N(A)$ is a Poisson random variable.
- $\mathbb{P}(N(A) = k) = \dfrac{e^{-\nu(A)}(\nu(A))^k}{k!}$
- when $A_1, A_2, \ldots, A_n$ are disjoint then random variables $N(A_1), N(A_2), \ldots, N(A_n)$ are independent.

Examples of Poisson random measures

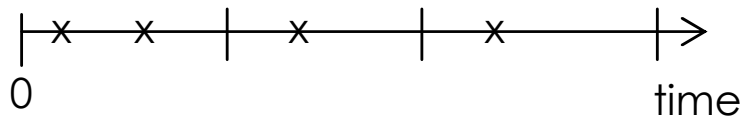

Figure 6: Poisson arrival process.

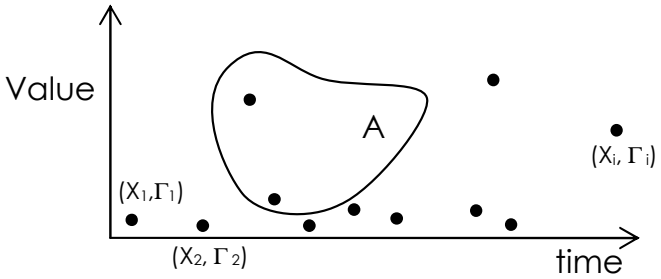

Figure 7: Compound Poisson Process.

- Poisson Process. See fig. 6. $T_1 \sim \exp(\lambda)$. The number of arrivals in each disjoint subinterval of time space is independent. A is an arbitrary subset, e.g. $A = \{(X_i, \Gamma_i) : \Gamma_i^2 > X_i - 2\}$.

- Compound Poisson Process. See fig. 7. X_i Poisson, $\Gamma_i \perp\!\!\!\perp$ gamma distributed. A is a region of the x-y space. # dots in $A \sim \text{Poisson}(\int_A dt\lambda dx a x^{-1} e^{-cx})$, or in other words, a Poisson random variable with parameter p where p is the integral of the mean measure over the region A. Here, $A \in [0, \infty] \times [0, \infty]$, i.e. a Lebesgue measure on x-axis [time] and a gamma measure on the y-axis [a value]. For a pictoral understanding of the y-axis, see fig. 8.

4 Theorems from the paper

We now review some theorems from the paper.

4.1 Theorem 1 and 2

Let (X_n) be a sample from F. Then

$$F|X_1, \ldots, X_n \sim \text{Dirichlet}(\mu_n)$$

$$\mu_n = \mu + \sum_{i=1}^{n} \delta(x_i)$$

$$\text{Measure } 0 < \mu(S) < \infty, \nu = \frac{\mu}{\theta}, \theta = \mu(S).$$

Note: here, $\mu = \alpha G_0$ from before.

4.2 Theorem 3

Let $\Gamma_{(1)} > \Gamma_{(2)} > \ldots$ be the points of a Poisson random measure on $(0, \infty)$ with mean measure $\theta x^{-1} e^{-x} dx$. To further aid our understanding of what a gamma measure might look like, we isolate the y-axis from fig. 7 above. Plotting the y-values now horizontally, we observe a dense clustering of points [raindrops] nearer the origin, thinning out as $x \to \infty$. See fig. 8. If we were to count the # of points falling in each subinterval of the x-axis, or in other

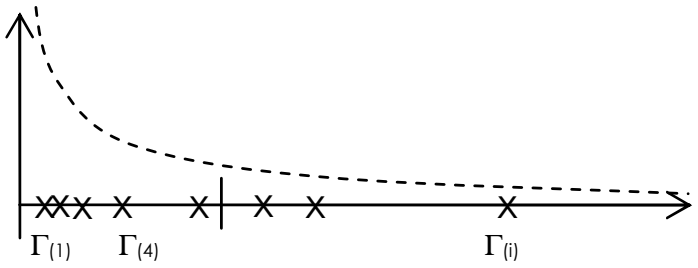

Figure 8: y-values, Γ_i in fig. 7 are reordered and labeled $\Gamma_{(i)}$, here plotted horizontally. The dotted line, the resulting pdf of $\Gamma_{(i)}$ is a gamma distribution.

words integrating mean measure $\theta x^{-1}e^{-x}dx$ over subintervals $[0,\epsilon]$ and $[\epsilon,\infty]$ we get

$$\int_{\epsilon}^{\infty} \theta x^{-1}e^{-x}dx \leq \frac{\theta}{\epsilon}\int_{\epsilon}^{\infty} e^{-x}dx < \infty. \tag{1}$$

$$\int_{0}^{\epsilon} \theta x^{-1}e^{-x}dx \geq \theta e^{-\epsilon}\int_{0}^{\epsilon} \frac{1}{x}dx = \infty. \tag{2}$$

(1) reveals that there are a finite # of points as $x \to \infty$ while (2) reveals that there an infinite # of points on $[0,\epsilon]$. Note that the ∞ on the right-hand side of (2) does not render the measure invalid since we are not using the Lebesgue measure but instead the gamma measure. Also note that θx^{-1} gives us (1) $\Rightarrow$ (2) since as $x \to \infty$ the mean measure $\theta x^{-1}e^{-x}dx \to 0$ [i.e. the intensity of "rain drops" decreases]. Peter illustrated how we would simulate these raindrops on a computer:

1. Choose an interval of the x-axis $[\epsilon,\infty]$. For example, $[1,\infty]$.

2. To determine the # of points in the subinterval, draw a sample N from a Poisson distribution with rate $\theta x^{-1}e^{-x}dx$. That is, compute $N = \int_1^{\infty} \theta x^{-1}e^{-x}dx$. Let's assume that $N = 3$.

3. Obtain the x-values for these N points by taking N # of samples from this gamma distribution, conditioned on the interval $[1,\infty]$. Here, we sample 3 values from the conditional gamma distribution. We then plot the 3 points along the x-axis at those values.

4. Choose the interval of the x-axis $[\frac{\epsilon}{2}, \epsilon)$, e.g. $[\frac{1}{2}, 1)$. Repeat step 2.

As $\epsilon \to 0$, we achieve our x-axis raindrop plot as depicted above. Thus, the density parameter θx^{-1} controls both the number of points and location of points in each subinterval. Generating a sequence of (x_i, y_i) values using the recipe above for y_i while drawing $x_i \sim \exp(\lambda)$ yields the Compound Poisson process depicted above.

D. Blei provided some intuition on relating the θ parameter in this Compound Poisson process with the α parameter as we traditionally have seen it. θ, α large = more points having a smaller height. In fig. 7 this would correspond to many raindrops clustered around $y = 0$ which, in a Dirichlet stick-breaking figure, looks like many low-probability stick lengths spread out evenly along the x-axis. θ, α small = fewer points having a larger height. In fig. 7 this would correspond to fewer raindrops around $y = 0$ and more raindrops around higher y-values which, in fig. 4 , would look like a few highly-varying probability stick lengths with a larger degree of clustering.
Continuing with the theorem, put

$$P_i = \frac{\Gamma_{(i)}}{\Sigma}, \Sigma = \sum_{i=1}^{\infty} \Gamma_{(i)}$$

Finally, define

$$F = \sum_{i=1}^{\infty} P_i \delta(\hat{X}_i) \tag{3}$$

where the $\hat{X}_i$ are i.i.d. (ν), independent also of the $\Gamma_{(i)}$. It may be worth nothing that we only use the Poisson random measure to generate the gamma random variables (rather than just drawing gamma random variables) to get *an ordering* of the gamma random variables. Otherwise, try to answer the question: have we drawn the largest random variable yet? The second largest? etc. Then F has Dirichlet$(\theta\nu)$ distribution, independently of Σ which has gamma(θ) distribution.

How does a gamma distribution emerge from a Poisson random measure? Specifically, how do we have $\Sigma \sim$ gamma$(\theta, 1)$? Let

$$f(x) = x$$

And

$$\begin{aligned} Nf &= \int_{\bar{\mathbb{R}}_+} f(x)N(\omega, dx) \text{ [} dx \text{ is a measure]} \\ &= \int_0^\infty x(\sum_{i=1}^\infty \mathbb{1}_{\Gamma_i}(dx)) \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}e^{-r\Sigma} &= \mathbb{E}e^{-rNf} && (4) \\ &= \exp\left(-\int_0^\infty dx\theta x^{-1}e^{-x}(1-e^{-rx})\right) && (5) \\ &= \exp(-\theta\log(1+r)) && (6) \\ &= \left(\frac{1}{1+r}\right)^\theta \Rightarrow Nf \sim \text{gamma}(\theta, 1) && (7) \end{aligned}$$

Note: $4 \Rightarrow 5$ is a property of Poisson random measure, $\mathbb{E}e^{-rNf} = \exp\left(-\int \nu(1-e^{-rf})\right)$.

Thus using this setup we are able to order our stick breaks such that they follow a gamma distribution. Each data point X_i has an associated value Γ_i drawn from a Poisson random measure with mean measure $\theta x^{-1}e^{-x}dx$. Reordering Γ_i yields a gamma distribution with scale parameter θ as a result as shown in eqn. 7. Integrating $\Gamma_{(i)}$ over $\omega \in \Omega$ and normalizing such that $\sum = 1$ gives the posterior probability stick-lengths. The scale of the obtained gamma distribution θ is exactly the hyperparameter α of the posterior Dirichlet distribution.

5 Theorem 5

[Definition 4 and 6, as well as Theorem 5 and associated corollaries from the paper were written].

6 Species sampling

We demonstrate how the previous definition can be applied to *species sampling* models. First, let

(P_i) be the frequencies (of each species);

$(\hat{X}_i)$ be the tags (of each species);

X_n be the species of the nth observation;

$\tilde{X}_j$ be the jth species to be observed;

N_{jn} be the number of times the jth species appears in the sample $X_1, \ldots, X_n$, i.e. $\sum_{m=1}^{n} \mathbb{1}_(X_m = \tilde{X}_j)$;

K_n be the number of distinct species in the first n observations, i.e. $\max\{j \mid N_{jn} > 0\}$.

Further assume that ν is drawn from a uniform distribution $U[0,1]$. Using the Blackwell-MacQueen prediction rule, the predictive distribution can be written as

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n, K_n = k) = \sum_{j=1}^{k} \frac{N_{jn}}{n+\theta} \mathbb{1}_(\tilde{X}_j \in A) + \frac{\theta}{n+\theta} \nu(A). \quad (8)$$

Observe that the terms $\frac{N_{jn}}{n+\theta}$ and $\frac{\theta}{n+\theta}$ can be seen as functions of the partition. This leads to generalizations of Equation 8 wherein the aforementioned terms are replaced by other functions of the partition.

Define

$$\mathbf{N}_n \equiv (N_{1n}, N_{2n}, \ldots) \equiv (n_1, \ldots, n_k) \equiv \mathbf{n},$$

and

$$p_j(\mathbf{n}) = \mathbb{P}(X_{n+1} = \tilde{X}_j \mid \mathbf{N}_n = \mathbf{n}) \qquad 1 \leq j \leq k(\mathbf{n}) + 1 \quad (9)$$

$$\mathbb{P}(X_1 \in A) = \nu(A). \quad (10)$$

With these definitions in place, (X_n) now has a distribution determined by $p_j(\mathbf{n})$. The Blackwell-MacQueen rule can then be seen as a special case of this formulation, with

$$p_j(\mathbf{n}) = \frac{n_j}{n+\theta} \mathbb{1}_(1 \leq j \leq k) + \frac{\theta}{n+\theta} \mathbb{1}_(j = k+1) \quad (11)$$

A statement about this more general case is given by Proposition 11. It assumes that

1. (X_n) is an exchangeable sequence;

2. (X_n) obeys the predictive rules given by Equation 9 and Equation 10.

Then the predictive distribution converges (in total variation norm) a.s. as $n \to \infty$ to

$$F = \sum_j \tilde{P}_j \delta(\tilde{X}_j) + \left(1 - \sum_j \tilde{P}_j\right) \nu, \tag{12}$$

where $\tilde{X}_j$ is drawn i.i.d. from ν and

$$\tilde{P}_j = \lim_{n\to\infty} \frac{N_{jn}}{n}. \tag{13}$$

This proposition implies that

- the number of values may be finite;
- F is almost surely discrete iff $\sum_j \tilde{P}_j = 1$;
- the form of F is not specified.

7 Exchangeable partition probability functions (EPPFs)

Let $[n] = \{1, \ldots, n\}$ be partitioned into k non-empty subsets $A_1, \ldots, A_k$. If (X_n) is exchangeable then

$$\mathbb{P}\left(\bigcap_{j=1}^{k} (X_l = \tilde{X}_j, \forall l \in A_j)\right) = p(\#A_1, \ldots, \#A_k), \tag{14}$$

where p is symmetric and $\#A_j$ denotes the number of elements of A_j. p is called the *exchangeable partition probability function* (EPPF) derived from the exchangeable sequence (X_n). Now denote

$$\mathbf{n} = (n_1, \ldots, n_k, 0, 0, \ldots) \tag{15}$$

$$\mathbf{n}^{j+} = (n_1, \ldots, n_j + 1, \ldots, n_k, \ldots). \tag{16}$$

An EPPF must then satisfy

$$p(1) = 1 \tag{17}$$

$$p(\mathbf{n}) = \sum_{j=1}^{k(\mathbf{n})+1} p(\mathbf{n}^{j+}). \tag{18}$$

Therefore, a p defined in such a way is an EPPF of an exchangeable sequence (X_n). Such a p can thus be used to define the predictive rules given in Equation 9 and Equation 10 in such a way as to satisfy the condition for Proposition 11(Equation 12). The p_j governing the prediction rule can be written in terms of p as

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})} \qquad 1 \leq j \leq k(\mathbf{n}) + 1, p(\mathbf{n}) > 0. \tag{19}$$

The previous statement can be clarified by Proposition 13,

> Corresponding to each pair (p, ν) where p is an EPPF and ν is a diffuse probability distribution, there is a unique distribution for a sampling sequence (X_n) such that p is the EPPF of (X_n) and ν is the distribution of X_1.

The implication of all this is the stronger Theorem 14 which states that

> Given a diffuse probability distribution ν and a sequence of functions p_j which satisfy Equation 15 and Equation 16, let (X_n) be a sequence governed by Equation 9 and Equation 10. (X_n) is exchangeable iff there exists a non-negative, symmetric function p defined such that Equation 19 holds. Then (X_n) is a sample from F as in Equation 12 and the EPPF of (X_n) is the uniquely p.

Finally, we can once again return to the Blackwell-MacQueen Urn Scheme we are familiar with. For a given θ, we can write the EPPF p_θ as

$$p_\theta(\mathbf{n}) = \frac{\theta^{k-1} \prod_{i=1}^{k} (n_i - 1)!}{[1 + \theta]_{n-1}}, \tag{20}$$

where $[x]_m = \prod_{j=1}^{m} (x + j - 1)$. The result of Theorem 14 is that p_θ is the EPPF of a sample from a Dirichlet process with parameter $\theta\nu$.