A Tutorial on the Dirichlet Process for Engineers

Technical Report

John Paisley
Department of Electrical & Computer Engineering
Duke University, Durham, NC
jwp4@ece.duke.edu

Abstract

This document provides a review of the Dirichlet process originally given in the author's preliminary exam paper and is presented here as a tutorial. No motivation is given (in what I've excerpted here), and this document is intended to be a mathematical tutorial that is still accessible to the engineer.

I. The Dirichlet Distribution

Consider the finite, D-dimensional vector, $\boldsymbol{\pi}$, having the properties $0 \leq \pi_i \leq 1$, $i = 1, \ldots, D$ and $\sum_{i=1}^{D} \pi_i = 1$ (i.e., residing on the $(D-1)$-dimensional simplex in $\mathbb{R}^D$, or $\boldsymbol{\pi} \in \Delta_D$). We view this vector as the parameter for the multinomial distribution, where samples, $X \sim Mult(\boldsymbol{\pi})$, take values $X \in \{1, \ldots, D\}$ with probability $P(X = i|\boldsymbol{\pi}) = \pi_i$. When the vector $\boldsymbol{\pi}$ is unknown, it can be inferred in the Bayesian setting using its conjugate prior, the Dirichlet distribution.

The Dirichlet distribution of dimension D is a continuous probability measure on Δ_D having the density function

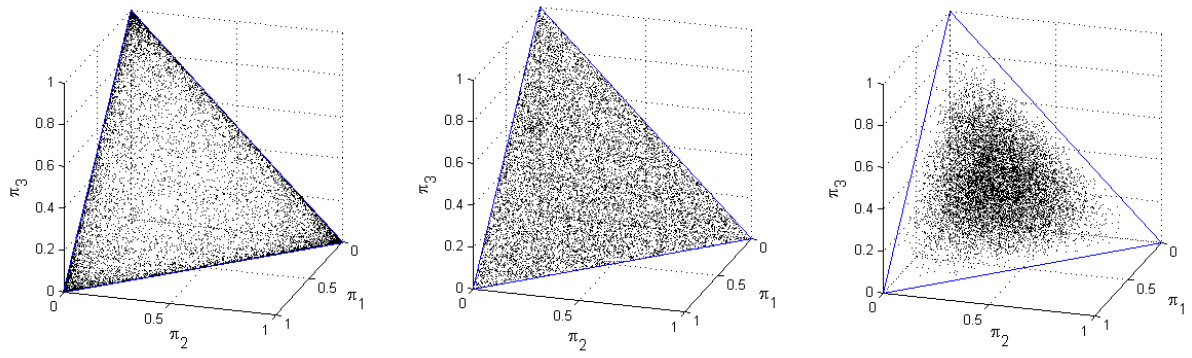
$$p(\boldsymbol{\pi}|\beta_1, \ldots, \beta_D) = \frac{\Gamma(\sum_i \beta_i)}{\prod_i \Gamma(\beta_i)} \prod_{i=1}^{D} \pi_i^{\beta_i - 1} \tag{1}$$

where the parameters $\beta_i \geq 0$, $\forall i$. It will be useful to reparameterize this distribution as follows:

$$p(\boldsymbol{\pi}|\alpha g_{01}, \ldots, \alpha g_{0D}) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha g_{0i})} \prod_{i=1}^{D} \pi_i^{\alpha g_{0i} - 1} \tag{2}$$

where we have defined $\alpha \equiv \sum_i \beta_i$ and $g_{0i} \equiv \beta_i / \sum_i \beta_i$. We refer to this distribution as $Dir(\alpha g_0)$. With this parameterization, the mean and variance of an element in $\boldsymbol{\pi} \sim Dir(\alpha g_0)$ is

$$\mathbb{E}[\pi_i] = g_{0i}, \qquad \mathbb{V}[\pi_i] = \frac{g_{0i}(1 - g_{0i})}{(\alpha + 1)} \tag{3}$$



(a) (b) (c)

Fig. 1. 10,000 samples from a 3-dimensional Dirichlet distribution with g_0 uniform and (a) $\alpha = 1$ (b) $\alpha = 3$ (c) $\alpha = 10$ As can be seen, when (a) $\alpha < 3$, the samples concentrate near vertices and edges of Δ_3. When (b) $\alpha = 3$, the density is uniform and when (c) $\alpha > 3$, the density begins to center on g_0.

so g_0 functions as a prior guess at $\boldsymbol{\pi}$ and α as a strength parameter, controlling how tight the distribution is around g_0. In the case where $g_{0i} = 0$, it follows that $P(\pi_i = 0) = 1$ and when $g_{0j} = 1$ and $g_{0i} = 0$, $\forall i \neq j$, it follows that $P(\pi_j = 1) = 1$.

Figure 1 contains plots of 10,000 samples each from a 3-dimensional Dirichlet distribution with g_0 uniform and $\alpha = 1, 3, 10$, respectively. When $\alpha = D$, or the dimensionality of the Dirichlet distribution, we see that the density is uniform on the simplex; when $\alpha > D$, the density begins to cluster around g_0. Perhaps more interesting, and more relevant to the Dirichlet process, is when $\alpha < D$. We see that as α becomes less than the dimensionality of the distribution, most of the density lies on the corners and faces of the simplex. In general, as the ratio of α to D shrinks, draws of $\boldsymbol{\pi}$ will be *sparse*, where most of the probability mass will be contained in a subset of the elements of $\boldsymbol{\pi}$. This phenomenon will be discussed in greater detail later and is a crucial element of the Dirichlet process. Values of $\boldsymbol{\pi}$ can be drawn from $Dir(\alpha g_0)$ in a finite number of steps using the following two methods (two infinite-step methods will be discussed shortly).

1) A Function of Gamma Random Variables [31]: Gamma random variables can be used to sample values from $Dir(\alpha g_0)$ as follows: Let $Z_i \sim Ga(\alpha g_{0i}, \lambda)$, $i = 1, \ldots, D$, where αg_{0i} is the shape parameter and λ the scale parameter of the gamma distribution. Then the vector $\boldsymbol{\pi} = \left(\frac{Z_1}{\sum_i Z_i}, \ldots, \frac{Z_D}{\sum_i Z_i}\right)$ has a $Dir(\alpha g_0)$ distribution. The parameter λ can be set to any positive, real value, but must remain constant. That is to say, the vector $\boldsymbol{\pi}$ is *free* of λ.

2) A Function of Beta Random Variables [9]: A stick-breaking approach can also be employed to draw from $Dir(\alpha g_0)$. For $i = 1, \ldots, D-1$ and $V_0 \equiv 0$, let

$$\begin{aligned} V_i &\sim Beta\left(\alpha g_{0i}, \alpha \sum_{j=i+1}^{D} g_{0j}\right) \\ \pi_i &= V_i \prod_{j<i}(1 - V_j) \\ \pi_D &= 1 - \sum_{j=1}^{D-1} \pi_j \end{aligned} \tag{4}$$

then the resulting vector, $\boldsymbol{\pi}$, has a $Dir(\alpha g_0)$ distribution. This approach derives its name from an analogy to breaking a proportion, V_i, from the remainder of a unit-length stick, $\prod_{j<i}(1 - V_j)$. Though called a "stick-breaking" approach, this method is distinct from that of Sethuraman [27], which will be discussed later and referred to as the "Sethuraman construction" to avoid confusion.

A. Calculating the Posterior of $\boldsymbol{\pi}$

As indicated above, the Dirichlet distribution is conjugate to the multinomial, meaning that the posterior can be solved analytically and is also a Dirichlet distribution. Using Bayes theorem,

$$p(\boldsymbol{\pi}|X = i) = \frac{p(X = i|\boldsymbol{\pi})p(\boldsymbol{\pi})}{p(X = i)} = \frac{p(X = i|\boldsymbol{\pi})p(\boldsymbol{\pi})}{\int_{\boldsymbol{\pi}\in\Delta_D} p(X = i|\boldsymbol{\pi})p(\boldsymbol{\pi})d\boldsymbol{\pi}} \propto p(X = i|\boldsymbol{\pi})p(\boldsymbol{\pi})$$

we can calculate the posterior distribution of $\boldsymbol{\pi}$ given data, X. For a single datum, the likelihood $p(X = i|\boldsymbol{\pi}) = \pi_i$, so the posterior can be seen to be Dirichlet as well

$$p(\boldsymbol{\pi}|X = i) \propto \pi_i^{(\alpha g_{0i}+1)-1} \prod_{j\neq i} \pi_j^{\alpha g_{0j}-1} \tag{5}$$

which, when normalized, equals $Dir(\alpha g_0 + \mathbf{e}_i)$, where we define $\mathbf{e}_i$ to be a D-dimensional vector of zeros with a one in the i^{th} position. We see that the i^{th} parameter of the Dirichlet distribution has simply been incremented by one. For N observations, this extends to

$$p(\boldsymbol{\pi}|X_1 = x_1, \ldots, X_N = x_N) \propto \prod_{i=1}^{D} \pi_i^{\alpha g_{0i}+n_i-1} \tag{6}$$

where $n_i = \sum_{j=1}^{N} \mathbf{e}_{X_j}(i)$, or the number of observations taking value i, and $N = \sum_{i=1}^{D} n_i$. When normalized, the posterior equals $Dir(\alpha g_{01} + n_1, \ldots, \alpha g_{0D} + n_D)$. Therefore, when used as a conjugate prior to the multinomial, the posterior distribution of $\boldsymbol{\pi}$ is Dirichlet where the parameters have been updated with the "counts" from the observed data. The interplay between the prior and the data can be seen in the posterior expectation of an element, π_i,

$$\mathbb{E}[\pi_i|X_1 = x_1, \ldots, X_N = x_N] = \frac{n_i + \alpha g_{0i}}{\alpha + N} = \frac{n_i}{\alpha + N} + \frac{\alpha g_{0i}}{\alpha + N} \tag{7}$$

where the second expression clearly shows the tradeoff between the prior and the data. We see that, as $N \to \infty$, the posterior expectation converges to a point mass located at $\left(\frac{n_1}{N}, \ldots, \frac{n_D}{N}\right)$, the empirical distribution of the observations. We can also see more clearly the effect of α, which acts as a "prior count," dictating the transition from prior to data. As a Bayes estimator, this parameter can be viewed as controlling how quickly we are willing to trust the empirical distribution of the data over our prior guess, while making this transition smoothly. Moreover, as $N \to \infty$, $\mathbb{V}[\pi_i] \to 0$, $\forall i$, meaning samples of $\boldsymbol{\pi} \sim Dir(\alpha g_{01} + n_1, \ldots, \alpha g_{0D} + n_D)$ will equal this expectation with probability 1, or the Dirichlet distribution converges to a delta function at the expectation.

B. Two Infinite Sampling Methods for the Dirichlet Distribution

Two other methods requiring an infinite number of random variables, the Pólya urn scheme and Sethuraman's constructive definition, also exist for sampling $\boldsymbol{\pi} \sim Dir(\alpha g_0)$. These methods, though of little practical value for the finite Dirichlet distribution, become very useful for sampling from the infinite Dirichlet process. Because they are arguably easier to understand in the finite setting, they are given here, which will allow for a more intuitive extension to the infinite-dimensional case.

1) The Pólya Urn Scheme [17]: The Pólya urn scheme is a sequential sampling process accompanied by the following story: Consider an urn initially containing α balls that can take one of D colors, with αg_{01} balls of the first color, αg_{02} of the second, etc. A person randomly selects a ball, X, from this urn, replaces the ball, and then adds a ball of the *same* color. After the first draw, the second ball is therefore drawn from the distribution

$$p(X_2|X_1 = x_1) = \frac{1}{\alpha + 1}\delta_{x_1} + \frac{\alpha}{\alpha + 1}g_0$$

and, inductively,

$$p(X_{N+1}|X_1 = x_1, \ldots, X_N = x_N) = \sum_{i=1}^{D} \frac{n_i}{\alpha + N}\delta_i + \frac{\alpha}{\alpha + N}g_0 \tag{8}$$

where δ_i is a delta function at the color having index i. If this seems reminiscent of the posterior expectation of $\boldsymbol{\pi}$ given in (7), that's because it is. Using the fact that $\int_{\boldsymbol{\pi} \in \Delta_D} p(X|\boldsymbol{\pi})p(\boldsymbol{\pi})d\boldsymbol{\pi} = p(X)$, we can write that

$$p(X_{N+1}|X_1 = x_1, \ldots, X_N = x_N) = \int_{\boldsymbol{\pi} \in \Delta_D} p(X_{N+1}|\boldsymbol{\pi})p(\boldsymbol{\pi}|X_1 = x_1, \ldots, X_N = x_N)d\boldsymbol{\pi} \tag{9}$$

where, conditioned on $\boldsymbol{\pi}$, X_{N+1} is independent of $X_1, \ldots, X_N$. We observe that $p(X_{N+1} = i|\boldsymbol{\pi}) = \pi_i$, and so

$$p(X_{N+1}|X_1 = x_1, \ldots, X_N = x_N) = \mathbb{E}[\boldsymbol{\pi}|X_1 = x_1, \ldots, X_N = x_N] \tag{10}$$

which is given for one element in (7). In this process, we are integrating out, or *marginalizing*, the random pmf, $\boldsymbol{\pi}$, and are therefore said to be drawing from a *marginalized* Dirichlet distribution. By the law of large numbers [31], as the number of samples $N \to \infty$, the empirical distribution of the urn converges to a discrete distribution, $\boldsymbol{\pi}$. That the distribution of $\boldsymbol{\pi}$ is $Dir(\alpha g_0)$ can be seen using the theory of exchangeability [1].

A sequence of random variables is said to be exchangeable if, for any permutation, ρ, of the integers $\{1, \ldots, N\}$, $p(X_1, \ldots, X_N) = p(X_{\rho_1}, \ldots, X_{\rho_N})$. By selecting the appropriate values from (8) and using the chain rule, $p(X_1, \ldots, X_N) = \prod_{i=1}^{N} p(X_i | X_{j<i})$, we see that

$$p(X_1 = x_1, \ldots, X_N = x_N) = \frac{\prod_{i=1}^{D} \prod_{j=1}^{n_i} (\alpha g_{0i} + j - 1)}{\prod_{j=1}^{N} (\alpha + j - 1)} \tag{11}$$

Because this likelihood is unchanged for all permutations, ρ, the sequence $(X_1, \ldots, X_N)$ is exchangeable. As a result of this exchangeability, de Finetti's theorem [12] states that there exists a discrete pmf, $\boldsymbol{\pi}$, having the yet-to-be-determined distribution, $p(\boldsymbol{\pi})$, conditioned on which the observations, $(X_1, \ldots, X_N)$, are independent as $N \to \infty$. The sequence of equalities follows:

$$\begin{aligned} p(X_1 = x_1, \ldots, X_N = x_N) &= \int_{\boldsymbol{\pi} \in \Delta_D} p(X_1 = x_1, \ldots, X_N = x_N | \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &= \int_{\boldsymbol{\pi} \in \Delta_D} \prod_{j=1}^{N} p(X_j = x_j | \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &= \int_{\boldsymbol{\pi} \in \Delta_D} \prod_{i=1}^{D} \pi_i^{n_i} p(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ &= \mathbb{E}_p \left[\prod_{i=1}^{D} \pi_i^{n_i} \right] \end{aligned} \tag{12}$$

Because a distribution is uniquely defined by its moments, the only distribution, $p(\boldsymbol{\pi})$, having moments

$$\mathbb{E}_p \left[\prod_{i=1}^{D} \pi_i^{n_i} \right] = \frac{\prod_{i=1}^{D} \prod_{j=1}^{n_i} (\alpha g_{0i} + j - 1)}{\prod_{j=1}^{N} (\alpha + j - 1)} \tag{13}$$

is the $Dir(\alpha g_0)$ distribution. Therefore, as $N \to \infty$, the sequence $(X_1, X_2, \ldots)$ can be viewed as iid samples from $\boldsymbol{\pi}$, where $\boldsymbol{\pi} \sim Dir(\alpha g_0)$ and is equal to $\left(\frac{n_1}{N}, \ldots, \frac{n_D}{N}\right)$.

2) The Sethuraman Construction of a Dirichlet Prior [27]: A second method for drawing the vector $\boldsymbol{\pi} \sim Dir(\alpha g_0)$ using an infinite sequence of random variables was proven to exist by Sethuraman:

$$\begin{aligned} \boldsymbol{\pi} &= \sum_{j=1}^{\infty} \left(V_j \prod_{k<j} (1 - V_k) \right) \mathbf{e}_{Y_j} \\ V_j &\sim Beta(1, \alpha) \\ Y_j &\sim Mult(g_0) \end{aligned} \tag{14}$$

where $Y_j \in \{1, \ldots, D\}$ and $V_0 \equiv 0$. Defining $p_j \equiv V_j \prod_{k<j}(1 - V_k)$, we see that $\pi_i = \sum_{j=1}^{\infty} p_j \mathbf{e}_{Y_j}(i)$. The weight vector, $\boldsymbol{p}$, is often said to be constructed via a stick-breaking process due to the analogy previously discussed. A visualization for two breaks of this process is shown in Figure 2. Note that this analogy and accompanying visualization does not take into consideration the location vector, $\boldsymbol{Y}$. For this draw to be meaningful, the values (V_i, Y_i) must always be considered as a pair, with neither being of any value without the other.

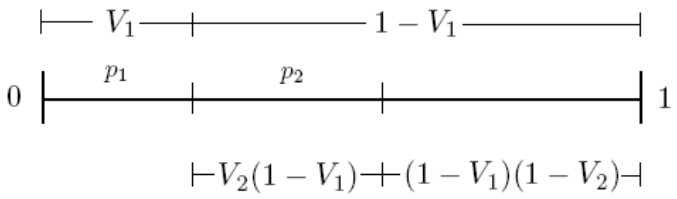


Fig. 2. A visualization of an abstract stick-breaking process after two breaks have been made. With respect to the Sethuraman construction, the random variable, $V_i \sim Beta(1, \alpha)$, must always be considered with its corresponding Y_i as a pair.

This distribution is said to be "constructed" due to the fact that the sequence $\pi_{i,n}$, being the value of π_i after break n, is stochastically increasing in value by stochastically decreasing values of p_n for which $Y_n = i$, converging to π_i as $n \to \infty$. That the resulting vector, $\boldsymbol{\pi}$, is a proper probability mass function can be seen from the fact that $p_j \in [0, 1]$, $\forall j$ because $V_j \in [0, 1]$, $\forall j$, as well as the fact that $\sum_{j=1}^{\infty} p_j = 1$ because $1 - \sum_{j=1}^{\infty} p_j = \prod_{j=1}^{\infty}(1 - V_j) \to 0$ as $j \to \infty$. The convergence of each π_i can be argued using similar reasoning.

The proof of Sethuraman's constructive definition relies on two Lemmas regarding the Dirichlet distribution. These Lemmas are given below, followed by their use in proving (14).

Lemma 1: Consider a random vector, $\boldsymbol{\pi} \in \Delta_D$, drawn according to the following process

$$\begin{aligned} \boldsymbol{\pi} &\sim Dir(\alpha g_0 + \mathbf{e}_Y) \\ Y &\sim Mult(g_0) \end{aligned} \tag{15}$$

where $Y \in \{1, \ldots, D\}$. Then $\boldsymbol{\pi}$ has a $Dir(\alpha g_0)$ distribution. Another way to express this is that

$$Dir(\alpha g_0) = \sum_{i=1}^{D} g_{0i} Dir(\alpha g_0 + \mathbf{e}_i) \tag{16}$$

So a Dirichlet distribution is also a specially parameterized mixture of Dirichlet distributions with the component $Dir(\alpha g_0 + \mathbf{e}_i)$ having mixing weight g_{0i}.

Proof: Basic probability theory allows us to write that

$$\begin{aligned} p(\boldsymbol{\pi}) &= \sum_{i=1}^{D} p(X=i)p(\boldsymbol{\pi}|X=i) \\ p(X=i) &= \int_{\boldsymbol{\pi}\in\Delta_D} p(X=i|\boldsymbol{\pi})p(\boldsymbol{\pi})d\boldsymbol{\pi} \end{aligned} \tag{17}$$

We observe that $p(X=i) = \mathbb{E}[\pi_i] = g_{0i}$ and that $p(\boldsymbol{\pi}|X=i) = Dir(\alpha g_0 + \mathbf{e}_i)$ from the posterior calculation of a Dirichlet distribution. Replacing these two values in (17) yields the desired result.

Lemma 2: Let $\boldsymbol{\pi}$ be constructed according to the following linear combination,

$$\begin{aligned} \boldsymbol{\pi} &= VW_1 + (1-V)W_2 \\ W_1 &\sim Dir(\omega_1, \ldots, \omega_D) \\ W_2 &\sim Dir(\upsilon_1, \ldots, \upsilon_D) \\ V &\sim Beta\left(\sum_{i=1}^{D}\omega_i, \sum_{i=1}^{D}\upsilon_i\right) \end{aligned} \tag{18}$$

Then the vector $\boldsymbol{\pi}$ has the $Dir(\omega_1+\upsilon_1, \ldots, \omega_D+\upsilon_D)$ distribution.

Proof: The proof of this Lemma relies on the representation of $\boldsymbol{\pi}$ as a function of gamma random variables. First, let $W_1 = \left(\frac{Z_1}{\sum_i Z_i}, \ldots, \frac{Z_D}{\sum_i Z_i}\right)$, $W_2 = \left(\frac{Z'_1}{\sum_i Z'_i}, \ldots, \frac{Z'_D}{\sum_i Z'_i}\right)$ and $V = \frac{\sum_i Z_i}{\sum_i Z_i + \sum_i Z'_i}$, where $Z_i \sim Ga(\omega_i, \lambda)$ and $Z'_i \sim Ga(\upsilon_i, \lambda)$. Then $W_1 \sim Dir(\omega_1, \ldots, \omega_D)$, $W_2 \sim Dir(\upsilon_1, \ldots, \upsilon_D)$ and $V \sim Beta(\sum_i \omega_i, \sum_i \upsilon_i)$. However, it remains to show that these are drawn *independently*, or that $p(W_1, W_2, V) = p(W_1)p(W_2)p(V)$. For this, we appeal to Basu's theorem [3], which states that since W_1 and W_2 are *free* of λ, W_1 and W_2 are also independent of any complete sufficient statistic of λ. In this case, $\sum_i Z_i$ and $\sum_i Z'_i$ are complete sufficient statistics for this parameter. Therefore, W_1 is independent of $\sum_i Z_i$ and W_2 is independent of $\sum_i Z'_i$ and, by extension, W_1 and W_2 are both independent of V.

Given that W_1, W_2 and V are constructed as above, we can see in their linear combination that $\boldsymbol{\pi} = VW_1 + (1-V)W_2 = \left(\frac{Z_1+Z'_1}{\sum_i Z_i + \sum_i Z'_i}, \ldots, \frac{Z_D+Z'_D}{\sum_i Z_i + \sum_i Z'_i}\right)$. The distribution of the sum of two gamma random variables allows us to say that $Z_i + Z'_i \sim Ga(\omega_i + \upsilon_i, \lambda)$ and therefore $\boldsymbol{\pi} \sim Dir(\omega_1+\upsilon_1, \ldots, \omega_D+\upsilon_D)$.

An immediate result of this Lemma is that the vector, $\boldsymbol{\pi}$, constructed as,

$$\begin{aligned} \boldsymbol{\pi} &= VW_1 + (1-V)W_2 \\ W_1 &\sim Dir(\mathbf{e}_i) \\ W_2 &\sim Dir(\alpha g_0) \\ V &\sim Beta(1,\alpha) \end{aligned} \tag{19}$$

has the distribution $Dir(\alpha g_0 + \mathbf{e}_i)$. Furthermore, we observe that, using a property of the Dirichlet distribution mentioned above, $p(W_1 = \mathbf{e}_i) = 1$. We now use these two Lemmas in considering the distribution of $\boldsymbol{\pi}$ in the following,

$$\begin{aligned} \boldsymbol{\pi} &= V\mathbf{e}_Y + (1-V)\boldsymbol{\pi}' \\ Y &\sim Mult(g_0) \\ V &\sim Beta(1,\alpha) \\ \boldsymbol{\pi}' &\sim Dir(\alpha g_0) \end{aligned} \tag{20}$$

First, we add a layer to (20) that seemingly contains no new information: $\mathbf{e}_Y \sim Dir(\mathbf{e}_Y)$, (a probability 1 event, which is why we use the same variable notation). Using Lemma 2, however, this allows us to collapse (20) and write

$$\begin{aligned} \boldsymbol{\pi} &\sim Dir(\alpha g_0 + \mathbf{e}_Y) \\ Y &\sim Mult(g_0) \end{aligned} \tag{21}$$

which, using Lemma 1, tells us that $\boldsymbol{\pi} \sim Dir(\alpha g_0)$. This fact allows us to decompose $\boldsymbol{\pi}$ into a linear combination of an infinite set of pairs of random variables. For example, since $\boldsymbol{\pi}' \sim Dir(\alpha g_0)$, this random variable can be expanded in the same manner as (20), producing

$$\begin{aligned} \boldsymbol{\pi} &= V_1\mathbf{e}_{Y_1} + V_2(1-V_1)\mathbf{e}_{Y_2} + (1-V_2)(1-V_1)\boldsymbol{\pi}'' \\ Y_i &\overset{iid}{\sim} Mult(g_0) \\ V_i &\overset{iid}{\sim} Beta(1,\alpha) \\ \boldsymbol{\pi}'' &\sim Dir(\alpha g_0) \end{aligned} \tag{22}$$

for $i = 1, 2$. Letting $i \to \infty$ produces (14). Since this process cannot be realized in practice, it must be truncated, where (V_i, Y_i) are drawn for $i = 1, \ldots, K$ with the remainder, $\epsilon = \prod_{i=1}^{K}(1 - V_i)$, arbitrarily assigned across $\boldsymbol{\pi}$. This truncation will be discussed in greater detail later.

From this definition of the Dirichlet distribution, the effect of α is perhaps more clear vis-à-vis Figure 1. The distribution on the values of $\boldsymbol{p} = \phi(\boldsymbol{V})$, where we will use $\phi(\cdot)$ to represent the stick-breaking function, is only dependent upon α and indicates how many breaks of the stick are necessary (probabilistically speaking) to utilize a given fraction of the probability weight. This number is independent of D. However, the sparsity of the final vector, $\boldsymbol{\pi}$, is dependent upon D (via g_0), as it controls the placement of elements from $\boldsymbol{p}$ into $\boldsymbol{\pi}$ (via $\boldsymbol{Y}$).

The Dirichlet distribution therefore has several means by which one can draw $\boldsymbol{\pi} \sim Dir(\alpha g_0)$. The two finite methods allow one to obtain the vector $\boldsymbol{\pi}$ *exactly*. The two infinite approaches, due to a necessary truncation (dictated by the data in the Pólya urn case and by choice in the Sethuraman construction), are only approximate. For this reason, in the finite setting ($D < \infty$), finite sampling methods are perhaps more reasonable and the two infinite methods discussed above merely interesting facts. However, in the infinite, nonparametric setting, the two infinite approach become the only viable means for working with the Dirichlet process.

II. The Dirichlet Process

To motivate the following discussion of the Dirichlet process, consider a Dirichlet distribution with g_0 uniform, $\boldsymbol{\pi} \in \Delta_k$ and

$$\boldsymbol{\pi} \sim Dir(\alpha/k, \ldots, \alpha/k) \tag{23}$$

What happens when $k \to \infty$? Looking at the mean and variance of π_i in (3), we see that both $\mathbb{E}[\pi_i], \mathbb{V}[\pi_i] \to 0$, leading to questions as to whether this distribution remains well-defined and can be sampled from. Additionally, the definition of the Dirichlet distribution in the previous section was a general analysis of the functional form. Each probability, π_i, was never defined to correspond to anything, and the prior vector, g_0, was arbitrarily defined. However, what if we define $\boldsymbol{\pi}$ over a continuous, measurable space, $(\Theta, \mathcal{B})$ equipped with a measure, αG_0, with which we replace αg_0?

In the following, we will attempt to give a sense of how these problems are approached and can be solved via the use of measure theory. However, our focus is mainly on developing the intuition; the reader is referred to the original papers for rigorous proofs [11][5][2][27]. We begin by giving the definition of the Dirichlet process.

Definition of the Dirichlet Process: Let Θ be a measurable space and $\mathcal{B}$ the σ-field of subsets of Θ. Let α be a positive scalar and G_0 a probability measure on $(\Theta, \mathcal{B})$. Then for all pairwise disjoint partitions $\{B_1, \ldots, B_k\}$ of Θ, where $\bigcup_{i=1}^{k} B_i = \Theta$, the random probability measure, G, on $(\Theta, \mathcal{B})$ is said to be a

Dirichlet process if $(G(B_1), \ldots, G(B_k)) \sim Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_k))$.

This process is denoted $G \sim DP(\alpha G_0)$. Because the nonparametric Dirichlet process is an infinite-dimensional prior on a continuous space, we focus on the infinite-dimensional analogues to the two infinite sampling methods detailed in Section I. Below, we review these two representations as they appear in the literature, noting that the infinite-dimensional analogue to the Pólya urn scheme is called the "Chinese restaurant process." Also in keeping with the literature, we replace the symbol X, as used above, with θ in what follows.

The Chinese Restaurant Process [1]: Given the observations, $\{\theta_1, \ldots, \theta_N\}$, sampled from a Chinese restaurant process (CRP), a new observation, θ_{N+1}, is sampled according to the distribution

$$\theta^*_{N+1} \sim \sum_{i=1}^{K} \frac{n_i}{\alpha + N} \delta_{\theta_i} + \frac{\alpha}{\alpha + N} G_0 \tag{24}$$

where K is the number of unique values among $\{\theta^*_1, \ldots, \theta^*_N\}$, θ_i those specific values and n_i the number of observations taking the value θ_i. As $N \to \infty$, the samples $\{\theta^*_i\}_{i=1}^{\infty}$ are iid from G, where $G \sim DP(\alpha G_0)$ and is equal to the expression in (24). As is evident, θ^*_{N+1} takes either a previously observed value with probability $\frac{N}{\alpha+N}$ or a new value drawn from G_0 with probability $\frac{\alpha}{\alpha+N}$. This value is new almost surely because we assume G_0 to be a continuous probability measure.

The Stick-Breaking (Sethuraman) Construction [27]: Let $\theta \sim G$ and G be constructed as follows:

$$\begin{aligned} G &= \sum_{i=1}^{\infty} \left(V_i \prod_{j<i} (1 - V_j) \right) \delta_{\theta_i} \\ V_i &\sim Beta(1, \alpha) \\ \theta_i &\sim G_0 \end{aligned} \tag{25}$$

Then $G \sim DP(\alpha G_0)$. Notice that G is constructed explicitly in (25), while only implicitly in (24). We now provide a general overview of how measure theory [4][16] relates to the Dirichlet process.

A. Measure Theory and the Dirichlet Process

In the following discussion, we will focus on the real line, $\theta \in \mathbb{R}$, with $\mathcal{B}$ the Borel σ-field generated by sets $A \in \mathcal{A}$ of the form $A = \{\theta : \theta \in (-\infty, a]\}$, where $a \in Q$, the set of rational numbers. This σ-field is generated by $\mathcal{A}$, or $\mathcal{B} = \sigma(\mathcal{A})$, by including in $\mathcal{B}$ all complements, unions and intersections of

the sets in $\mathcal{A}$. For our purposes, this means that all disjoint sets of the form $B_i = \{\theta : \theta \in (a_{i-1}, a_i]\}$, are included in $\mathcal{B}$ since $B_i = A_i \cap A_{i-1}^c$, where $A^c = \{\theta : \theta \notin A\}$, the complement of A and letting A_i be the set for which $a = a_i$, with $-\infty < a_{i-1} \leq a_i < \infty$.

This is necessary because the Dirichlet process is parameterized by a measure, G_0, on the Borel sets of Θ. To accomplish this, we let the probability density function $p(\theta)$ be associated with the probability measure, G_0, such that

$$G_0(B_i) = \int_{a_{i-1}}^{a_i} p(\theta)\ d\theta = P(a_i) - P(a_{i-1}) \tag{26}$$

For example, we could define $P(\theta) \equiv \mathcal{N}(\theta; \mu, \sigma^2)$. This is illustrated in Figure 3 for three partitions. We observe that, in this illustration, the measure is simply the area under the curve of the region $(a_{i-1}, a_i]$. In essence, measure theory is introduced because it allows us to partition the real line in any way we choose, provided a is rational, and lets us refine these partitions to infinitesimally small regions, all in a rigorous, well-defined manner.

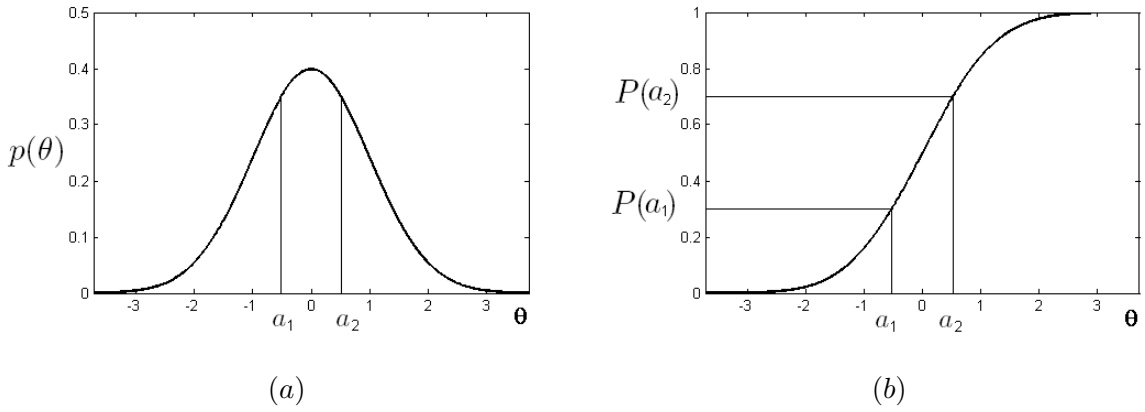


Fig. 3. (a) A partition of $\mathbb{R}$ into three sections, $((-\infty, a_1], (a_1, a_2], (a_2, \infty))$. (b) The corresponding probabilities of the three sections, $(P(a_1), P(a_2) - P(a_1), 1 - P(a_2))$. In measure theoretic notation, this corresponds to the disjoint sets (B_1, B_2, B_3) and their respective measures $(G_0(B_1), G_0(B_2), G_0(B_3))$.

In Ferguson's definition of the Dirichlet process [11], a prior measure, G_0, on the Borel sets in $\mathcal{B}$ is used along with a strength parameter, α, to parameterize the Dirichlet distribution. This applies to any set of and any number of disjoint partitions, which motivates the generality of the definition. In machine learning applications, we are concerned with the infinite limit, where Θ is partitioned in a specific way that we will discuss later. First, we observe that the function of α in the Dirichlet process is identical to that in Section I and, because $G_0(B_i)$ is simply a number (an area under a curve), G_0 performs the same

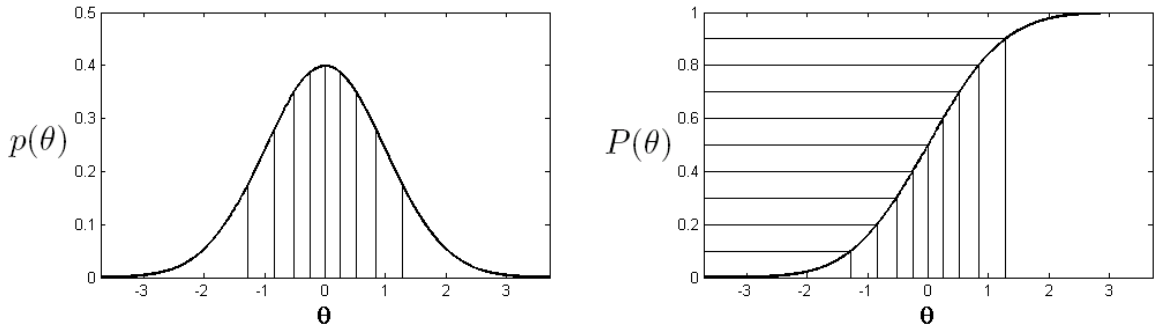
function as g_0. The crucial difference is that we have definied G over a measurable space and defined the prior for G using a probability measure, G_0, on that same space. Functionally speaking, $G(B_1)$ and π_1 are equivalent in that they are both probabilities corresponding to the first dimension of the Dirichlet distribution, but $G(B_1)$ contains the additional information that the measure is associated with the set B_1. Indeed, to emphasize this similarity, this could be written in the following way:

$$\begin{aligned} G(B) &= \sum_{i=1}^{k} \pi_i \delta_{B_i}(B) \\ \boldsymbol{\pi} &\sim Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)) \end{aligned} \tag{27}$$

where $G(B)$ is a real-valued set function that takes as its argument $B \in \{B_1, \ldots, B_k\}$ and outputs the π_i for which $\delta_{B_i}(B) = 1$, which occurs when $B = B_i$ and is 0 otherwise. The summation, though uninteresting in this case, is to be taken literally. For example, $G(B_1) = \pi_1 \cdot 1 + \sum_{i=2}^{k} \pi_i \cdot 0 = \pi_1$.

It will be noticed that discussion of the Dirichlet process in the machine learning and signal processing literature never mentions sets of the form of B, but rather specific values of θ. This is due to taking the limit as $k \to \infty$ and motivates the following discussion of the inverse-cdf method for obtaining random variables having a density $p(\theta)$. Figure 4 contains a plot of ten partitions of the space Θ according to a function, $p(\theta)$, such that the measure of any partition is $G_0(B_i) = 1/10$. This is clearly visible in Figure 4(b). According to this picture, one can see that a partition, B_i, could be selected according to this measure by first drawing a uniformly distributed random variable, $\eta \sim U(0, 1)$, finding the appropriate cell on the y-axis and inverting to obtain the randomly selected partition along the θ-axis. The partition, B_i, will have been selected uniformly, but the values, $\theta \in B_i$, will span various widths of Θ according to $P(\theta)$. If we define new sets, $\{C_1, \ldots, C_k\}$, of the form, $C_i = \{\eta : \eta \in (\frac{i-1}{k}, \frac{i}{k}]\}$, then B_i is the inverse image of C_i, or $B_i = \{\theta : \theta = P^{-1}(\eta), \forall \eta \in C_i\}$. Allowing $k \to \infty$, uniformly sampling among sets, C_i, and finding the inverse image converges to sampling $\theta \sim G_0$. In other words, if we sample $\eta \sim U(0, 1)$ and calculate $\theta = P^{-1}(\eta)$, then $\theta \sim P(\theta)$.

We mention this because this is implicitly taking place in the Dirichlet process. When we sample $G \sim DP(\alpha G_0)$, we are assigning to each dimension of $Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_k))$ a partition of measure $G_0(B_i) = 1/k$ derived according to the distribution function $P(\theta)$. In a manner identical to the finite dimensional case, (8) and (14), where we sample from g_0 (i.e. $Mult(g_0)$), we sample uniformly among partitions B_i according to the measure G_0. Though each measure, $G_0(B_i)$, is equal, their locations along Θ are distributed according to the distribution $P(\theta)$. Therefore, by letting $G_0(B_i) = 1/k$ for $i = 1, \ldots, k$ and allowing $k \to \infty$, we can use the mathematical derivations of Section I explicitly, as we will discuss in the next section.



(a) (b)

Fig. 4. A partition of $\mathbb{R}$ into $k = 10$ disjoint sets, $(B_1, \ldots, B_{10}) = (\{(-\infty, a_1]\}, \{(a_1, a_2]\}, \ldots, \{(a_8, a_9]\}, \{(a_9, \infty)\})$, of equal measure, $G_0(B_i) = \frac{1}{10}$, $\forall i$ (or $P(a_i) - P(a_{i-1}) = \frac{1}{10}$). The uniform partition of $P(\theta)$ in (b) can be seen to lead to partitions of θ approximating the density $p(\theta)$ in (a). As the partitions are refined, or $k \to \infty$, uniformly sampling among partitions converges to drawing $\theta \sim G_0$.

Though significantly more discussion is needed to make this rigorous (e.g., that it satisfies Kolmogorov's consistency condition [11]), the above discussion is essentially what is required for an intuitive understanding of the differences between the Dirichlet distribution and the Dirichlet process. Furthermore, this intuition can be extended to spaces of higher dimension, allowing $\theta \sim G_0$ to indicate the drawing of multiple variables in a higher dimensional space.

B. Drawing from the Infinite Dirichlet Process

We examine this uniform sampling of B_i more closely and how it directly relates to the Chinese Restaurant Process (CRP) and Sethuraman construction. First, consider a finite number of partitions, in which case the CRP can be written as a finite-dimensional Pólya urn model (8):

$$B^*_{N+1} \sim \sum_{i=1}^{k} \frac{n_i}{\alpha + N} \delta_{B_i} + \frac{\alpha}{\alpha + N} G_0(B) = \sum_{i=1}^{k} \frac{n_i + \alpha G_0(B_i)}{\alpha + N} \delta_{B_i} \tag{28}$$

Letting $G_0(B_i) = 1/k$ and $k \to \infty$ as discussed, we see that $\alpha G_0(B_i) \to 0$, thus disappearing from all components for which $n_i > 0$. However, there are an infinite number of remaining, unused components with locations distributed according to G_0. Notice that we do not need to change G_0 to reflect the removal of θ values, as they are all of measure zero. This means that the weight on G_0 is also unchanged and remains α. As a result, though we cannot explicitly write the remaining, unused components, we can "lump" them together under G_0 with a probability of selecting from G_0 proportional to α. Whereas for the finite-dimensional case there was a nonzero probability that a value selected from G_0 would equal

one previously seen, for the continuous case this is a probability zero event. Therefore, by allowing the number of components to tend to infinity, we obtain a continuum of colors as proven in [5], also known as the Chinese restaurant process [1].

Identical reasoning is used for the Sethuraman construction. For $k \to \infty$, consider the step

$$Y_i \sim Mult(G_0(B_1), \ldots, G_0(B_k)) \tag{29}$$

in light of Figure 4(b), in which case $Y \in \{B_1, \ldots, B_k\}$. Sampling partitions from this infinite-dimensional uniform multinomial distribution is precisely equivalent to drawing $\theta \sim G_0$ for reasons previously discussed regarding the inverse-cdf method. In this case and that of the CRP, a change of notation has concealed the fact that the same fundamental underlying process is taking place as that of the finite-dimensional case – a notable difference being that repeated values are here probability zero events. Using the notation of Section I-B2, this difference implies for the Sethuraman construction that a one-to-one correspondence exists between values in $\boldsymbol{p}$ and the desired vector, $\boldsymbol{\pi}$. For clarity, we will continue to use $\boldsymbol{p}$ when discussing the Dirichlet process, leaving $\boldsymbol{\pi}$ to represent values drawn from the finite-dimensional Dirichlet distribution. Also, we emphasize that G is a symbol (a set function) that incorporates both $\boldsymbol{p}$ and θ as a complete measure. Just as V_i was meaningless without the Y_i in the finite case, p_i is meaningless without its corresponding location, θ_i.

Finally, it is hopefully clear why it is sometimes rhetorically asked in the literature how one can go about drawing from the Dirichlet process. We are attempting to draw an infinite-dimensional pmf that covers every point of a continuous space. Using the two "finite" methods (Section I), so labeled because they produced samples $\boldsymbol{\pi} \sim Dir(\alpha g_0)$ in a finite number of steps, we need an infinite length of time to obtain $\boldsymbol{\pi}$. The two infinite methods are now the only feasible means and each solves this problem elegantly in its own way. With the Chinese restaurant process, we never need to draw G. More remarkably, the Sethuraman construction allows one to explicitly draw G by drawing the important components *first*, stopping when a desired portion of the unit mass is assigned and implicitly setting the probability of the infinitely remaining components to zero. We will see later that this prior can be applied to drawing directly from the Beta process as well, a result we believe to be new.

C. Dirichlet Process Mixture Models

Having established the means to draw $G \sim DP(\alpha G_0)$, we next discuss the use of the values $\theta \sim G$. The main reason we replaced X with θ in our discussion of the Dirichlet process is that the symbol X is generally reserved for the observed data. In Section I, this was appropriate because we were only

interested in observing samples from the discrete $\boldsymbol{\pi}$. However, the primary use of the Dirichlet process in the machine learning community is in nonparametric mixture modeling [2][10], where $\theta \sim G$ is not the observed data, but rather a parameter (or set of parameters) for some distribution, $F(\theta)$, from which X is drawn. This can be written as follows:

$$\begin{aligned} X_j &\sim F(\theta_j^*) \\ \theta_j^* &\overset{iid}{\sim} G \\ G &\sim DP(\alpha G_0) \end{aligned} \tag{30}$$

where θ_j^* is a specific value selected from G and associated with observation X_j. An example of this process is the Gaussian mixture model, where $\theta = \{\mu, \Lambda\}$, the mean and precision matrix for a multivariate normal distribution, and G_0 the normal-Wishart conjugate prior. Whereas samples from the discrete G will repeat with high probability, samples from $F(\theta)$ are again from a continuous distribution. Therefore, values that share parameters are *clustered* together in that, though not exactly the same, they exhibit similar statistical characteristics according to some distribution function, $F(\theta)$.

III. Inference for the Dirichlet Process Mixture Model

To show the Dirichlet process in action, we outline a general method for performing Markov chain Monte Carlo (MCMC) [13] inference for Dirichlet process mixture models. We let $f(x|\theta)$ be the likelihood function of an observation, x, given parameters, θ, and $p(\theta)$ the prior density of θ, corresponding to the probability measure G_0. We also use an auxiliary variable, $c \in \{1, \ldots, K+1\}$, which acts as an indicator of the parameter value, θ_{c_i}, associated with observation x_i. We will discuss the value of K and why we write $K+1$ shortly. To clarify further the function of c, we rewrite (30) utilizing this auxiliary variable:

$$\begin{aligned} X_j &\sim F(\theta_{c_j}) \\ c_j &\overset{iid}{\sim} Mult(\phi_K(\boldsymbol{V})) \\ V_k &\overset{iid}{\sim} Beta(1, \alpha) \\ \theta_k &\overset{iid}{\sim} G_0 \end{aligned} \tag{31}$$

where $\boldsymbol{p} = \phi_K(\boldsymbol{V})$ represents the stick-breaking portion of the Sethuraman construction that has been truncated to length $K+1$, with $p_{K+1} = \prod_{i=1}^{K}(1 - V_i)$. We use subscript K because the $(K+1)$-dimensional vector, $\boldsymbol{p}$, is a function of K parameters. Therefore, only the first K values of V_k and the first $K+1$ values of θ_k are used. For clarity, following each iteration only utilized components are

retained (as indicated by $\{c_j\}_{j=1}^N$), and a new, proposal components is drawn from the base distribution. See [23] for further discussion. The number of utilized components for any given iteration is represented by the variable K, and we here only propose a $K+1^{st}$ component. Below is this sampling process.

Initialization: Select a truncation level, $K+1$, and initialize the model, G, by sampling $\theta_k \sim G_0$ for $k = 1, \ldots, K+1$ and $V_k \sim Beta(1, \alpha)$ for $k = 1, \ldots, K$ and construct $\boldsymbol{p} = \phi_K(\boldsymbol{V})$. For now, α is set *a priori*, but we will discuss inference for this parameter as well.

Step 1: Sample the indicators, $c_1, \ldots, c_N$, independently from their respective conditional posteriors, $p(c_j|x_j) \propto f(x_j|\theta_{c_j})p(\theta_{c_j}|G)$,

$$c_j \sim \sum_{k=1}^{K+1} \frac{p_k f(x_j|\theta_k)}{\sum_l p_l f(x_j|\theta_l)} \delta_k \tag{32}$$

Set K to be the number of unique values among $c_1, \ldots, c_N$ and relabel from 1 to K.

Step 2: Sample $\theta_1, \ldots, \theta_K$ from their respective posteriors conditioned on $c_1, \ldots, c_N$ and $x_1, \ldots, x_N$,

$$\begin{aligned} \theta_k &\sim p\left(\theta_k|\{c_j\}_{j=1}^N, \{x_j\}_{j=1}^N\right) \\ p\left(\theta_k|\{c_j\}_{j=1}^N, \{x_j\}_{j=1}^N\right) &\propto \prod_{j=1}^N \left(f(x_j|\theta)^{\delta_{c_j}(k)}\right) p(\theta) \end{aligned} \tag{33}$$

where $\delta_{c_j}(k)$ is a delta function equal to one if $c_j = k$ and zero otherwise, simply picking out which x_j belong to component k. Sample $\theta_{K+1} \sim G_0$.

Step 3: For the Sethuraman construction, construct the $(K+1)$-dimensional weight vector, $\boldsymbol{p} = \phi_K(\boldsymbol{V})$, using $V_1, \ldots, V_K$ sampled from their Beta-distributed posteriors conditioned on $c_1, \ldots, c_N$,

$$V_k \sim Beta\left(1 + \sum_{j=1}^N \delta_{c_j}(k), \alpha + \sum_{l=k+1}^K \sum_{j=1}^N \delta_{c_j}(l)\right) \tag{34}$$

Set $p_{K+1} = \prod_{k=1}^K (1 - V_k)$. For the Chinese restaurant process, this step instead consists of setting $p_k = \frac{n_k}{\alpha+N}$ and $p_{K+1} = \frac{\alpha}{\alpha+N}$, as discussed in Section II.

Repeat Steps 1 – 3 for a desired number of iterations. The convergence of this Markov chain can be assessed [13], after which point uncorrelated samples (properly spaced out in the chain) of the values in Steps 1 – 3 are considered iid samples from the posterior, $p\left(\{\theta_k\}_{k=1}^K, \{V_k\}_{k=1}^K, \{c_j\}_{j=1}^N, K|\{x_j\}_{j=1}^N\right)$. Additional inference for α can be done for the CRP using a method detailed in [10] and for the Sethuraman construction using a conjugate, gamma prior.

Step 4: Sample α from its posterior gamma distribution conditioned on $V_1, \ldots, V_K$

$$\alpha \sim Ga\left(a_0 + K, b_0 - \sum_{k=1}^{K} \ln(1 - V_k)\right) \tag{35}$$

where a_0, b_0 are the prior hyperparameters for the gamma distribution.

As can be seen, inference for the Dirichlet process is fairly straightforward and, when $p(\theta)$ is conjugate to $f(x|\theta)$, fully analytical. Other MCMC methods exist for performing this inference [20] as does a fast, variational inference method [7][6] that, following initialization, deterministically converges to a local optimal approximation to the full posterior distribution.

IV. Extensions of the Dirichlet Process

With the advances of computational resources and MCMC (and later, VB) inference methods, the Dirichlet process has had a significant role in Bayesian hierarchical modeling. This can be seen in the many developments of the original idea. To illustrate, we briefly review two of these developments common to the machine learning literature: the hierarchical Dirichlet process (HDP) [28] and the nested Dirichlet process (nDP) [26]. We then present our own extension that, to our knowledge, is new to the machine learning and signal processing communities. This extension, called the *Dirichlet process with product base measure* (DP-PBM), extends the Dirichlet process to the modeling of multiple modalities, m, that each call for unique distribution functions, $F_m(\theta_m)$.

A. The Hierarchical Dirichlet Process [28]

In our discussion of the Dirichlet distribution in Section I, we defined α to be a positive scalar and g_0 to be a D-dimensional pmf, with the new pmf $\boldsymbol{\pi} \sim Dir(\alpha g_0)$. Though our discussion assumed g_0 was uniform, in general there are no restrictions on g_0, so long as it is in Δ_D. This leads to the following possibility

$$\boldsymbol{\pi}' \sim Dir(\beta\boldsymbol{\pi}), \qquad \boldsymbol{\pi} \sim Dir(\alpha g_0) \tag{36}$$

where β is a value having the same restrictions as α. In essence, this is the hierarchical Dirichlet process (HDP). A measure, G_i', is drawn from an HDP, denoted $G_i' \sim HDP(\alpha, \beta, G_0)$, as follows:

$$G_i' \sim DP(\beta G), \qquad G \sim DP(\alpha G_0) \tag{37}$$

Using the measure theory outlined in Section II-A, this process can be understood as the infinite-dimensional analogue to (36) extended to a measurable space $(\Theta, \mathcal{B})$. We note that, in the case where G

is constructed according to a stick-breaking process and truncated, G_i' is a finite-dimensional Dirichlet process for which the prior expectation, $\boldsymbol{p}$, and locations, θ, are determined for G_i' by G. A key benefit of the HDP is that draws of $G_1', \ldots, G_n' \sim DP(\beta G)$ will utilize the same subset of atoms, since G is a discrete measure with probability one. Therefore, this method is useful for multitask learning [8][21], where atoms might be expected to share across tasks, but with a different probability of usage for each task.

B. The Nested Dirichlet Process [26]

In our discussion of the Dirichlet process, we assumed G_0 to be a continuous distribution, though this isn't necessary, as we've seen with the HDP. Consider the case where G_0 is a Dirichlet distribution. Extending this base Dirichlet to its own measurable space, the base distribution becomes a Dirichlet process as well. This is called the nested Dirichlet process (nDP). A way to think of a draw $G \sim nDP(\alpha, \beta, G_0)$ is as a mixture model of mixture models. This can be seen in the Sethuraman construction,

$$\begin{aligned} G &= \sum_{j=1}^{\infty} \left(V_j \prod_{k<j} (1 - V_k) \right) \delta_{G_j} \\ V_i &\sim Beta(1, \alpha) \\ G_j &\sim DP(\beta G_0) \end{aligned} \tag{38}$$

Now, rather than drawing $\theta_i^* \sim G$, an entire mixture, $G_i^* \sim G$, is drawn, from which is then drawn an atom, $\theta_i^* \sim G_i^*$. As with the HDP, this prior can be useful in multitask learning, where a task, t, selects an indicator from the "top-level" DP, G, and each observation within that task utilizes atoms sampled from a "second-level" DP, G_t^*. In the case of multitask learning with the HDP, all tasks share the same atoms, but with different mixing weights. The nDP for multitask learning is "all-or-nothing;" two tasks either share a mixture model exactly, or nothing at all.

C. The Dirichlet Process with a Product Base Measure

We here propose another extension of the Dirichlet process that incorporates a product base measure, denoted DP-PBM, where rather than drawing parameters for one parametric distribution, $\theta \sim G_0$, parameters are drawn for *multiple* distributions, $\theta_m \sim G_{0,m}$ for $m = 1, \ldots, M$. In other words, rather than having a connection between data, $\{X_i\}_{i=1}^N$, and their respective parameters, $\{\theta_i^*\}_{i=1}^N$, through a parametric distribution, $\{F(\theta_i^*)\}_{i=1}^N$, *sets* of data, $\{X_{1,i}, \ldots, X_{M,i}\}_{i=1}^N$ have respective *sets* of parameters, $\{\theta_{1,i}^*, \ldots, \theta_{M,i}^*\}_{i=1}^N$,

used in inherently different and incompatible distribution functions, $\{F_1(\theta^*_{1,i}), \ldots, F_M(\theta^*_{M,i})\}$. The DP-PBM is so called because it utilizes a product base measure to achieve this end, $G_0 = G_{0,1} \times G_{0,2} \times \cdots \times G_{0,M}$, where, in this case, M modalities are considered. The space over which this process is defined is now $\left(\prod_{m=1}^{M} \Theta_m, \bigotimes_{m=1}^{M} \mathcal{B}_m, \prod_{m=1}^{M} G_{0,m}\right)$. Though this construction implicitly takes place in all mixture models that attempt to estimate multiple parameters, for example the multivariate Gaussian mixture model, we believe our use of these parameters in multiple, incompatible distributions (or modalities) is novel. The fully generative process can be written as follows:

$$\begin{aligned} X_{m,i} &\sim F_m(\theta^*_{m,i}) \qquad m = 1, \ldots, M \\ \{\theta^*_{m,i}\}_{m=1}^{M} &\sim G \\ G &= \sum_{j=1}^{\infty} \left(V_j \prod_{k<j} (1 - V_k) \right) \prod_{m=1}^{M} \delta_{\theta_{m,j}} \\ V_j &\overset{iid}{\sim} Beta(1, \alpha) \\ \theta_{m,j} &\sim G_{0,m} \qquad m = 1, \ldots, M \end{aligned} \tag{39}$$

where $\theta_{m,j}$ are drawn iid from $G_{0,m}$ for a fixed m and independently under varying m. To make a clarifying observation, we note that if each $G_{0,m}$ is a univariate normal-gamma prior, this model reduces to a multivariate GMM with a forced diagonal covariance matrix. As previously stated, we are more interested in cases where each X_m is inherently incompatible, but is still linked by the structure of the data set.

For example, consider a set of observations, $\{O_i\}_{i=1}^{N}$, where each $O_i = \{X_{1,i}, X_{2,i}\}$ with $X_1 \in \mathbb{R}^d$ and X_2 a sequence of time-series data. In this case, a single density function, $f(X_1, X_2|\theta_1, \theta_2)$ cannot analytically accommodate O_i, making inference difficult. However, if these densities can be considered as *independent*, that is $f(X_1, X_2|\theta_1, \theta_2) = f(X_1|\theta_1) \cdot f(X_2|\theta_2)$, then this problem becomes analytically tractable and, furthermore, no more difficult to solve than for the standard Dirichlet process. One might choose to model X_1 with a Gaussian distribution, with $G_{0,1}$ the appropriate prior and X_2 by an HMM [25], with $G_{0,2}$ its respective prior [22] . In this case, this model becomes a hybrid Gaussian-HMM mixture, where each component is *both* a Gaussian *and* a hidden Markov model. We develop a general MCMC inference method below to allow us to look more deeply at the structure of our framework. We also observe that this idea can be easily extended to the nDP and HDP frameworks.

D. Inference for the DP-PBM Mixture Model

MCMC inference for the DP-PBM mixture model entails a few simple alterations to the sampling scheme outlined in Section III. These changes alone are given below:

Initialization: For each $j = 1, \ldots, K+1$, sample $\theta_{m,j} \sim G_{0,m}$ for $m = 1, \ldots, M$, where M is the number of "modalities," or distinct subsets of data for each observation.

Step 1: Sample the indicators, $c_1, \ldots, c_N$, independently from their conditional posteriors.

$$c_j \sim \sum_{k=1}^{K+1} \frac{p_k \prod_{m=1}^{M} f_m(x_{m,j}|\theta_{m,k})}{\sum_l p_l \prod_{m=1}^{M} f_m(x_{m,j}|\theta_{m,l})} \delta_k \tag{40}$$

We note that the likelihood term has been factorized into a product of the likelihood for each modality.

Step 2: Sample $\{\theta_{m,1}\}_{m=1}^{M}, \ldots, \{\theta_{m,K}\}_{m=1}^{M}$ from their conditional posteriors given $c_1, \ldots, c_N$ and $\{x_{m,1}\}_{m=1}^{M}, \ldots, \{x_{m,N}\}_{m=1}^{M}$.

$$\begin{aligned} \theta_{m,k} &\sim p\left(\theta_{m,k}|\{c_j\}_{j=1}^{N}, x_{m,1}, \ldots, x_{m,N}\right) \\ p\left(\theta_{m,k}|\{c_j\}_{j=1}^{N}, x_{m,1}, \ldots, x_{m,N}\right) &\propto \prod_{j=1}^{N} \left(f(x_{m,j}|\theta_m)^{\delta_{c_j}(k)}\right) p(\theta_m) \end{aligned} \tag{41}$$

Sample $\theta_{m,K+1} \sim G_{0,m}$ for $m = 1, \ldots M$. The essential difference here is that each component, $k = 1, \ldots, K$ now has M posteriors to calculate. These M posteriors are calculated independently of one another given the relevant data for that modality extracted from the observations assigned to that component. We stress that when an "observation" is assigned to a component (via the indicator, c) it is actually *all* of the data that comprise that observation that is being assigned to the component.

E. Experimental Result with Synthesized Data

To further illustrate our model, we implement it on a toy problem of two modalities. For this problem, each observation is of the form, $O = \{X_1, X_2\}$, where $X_1 \sim \mathcal{N}(\mu, \Sigma)$ and $X_2 \sim HMM(A, B, \pi)$. We define three, two-dimensional Gaussian distributions with respective means $\mu_1 = (-3, 0)$, $\mu_2 = (3, 0)$ and $\mu_3 = (0, 5)$ and each having the identity as the covariance matrix. Two hidden Markov models are defined as below,

$$\mathbf{A_1} = \begin{bmatrix} 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \\ 0.9 & 0.05 & 0.05 \end{bmatrix} \quad \mathbf{A_2} = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix} \quad \mathbf{B_1}, \mathbf{B_2} = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$$

with the initial state vector $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 = [1/3, 1/3, 1/3]$. Data was generated as follows: We sampled 300 observations, 100 from each Gaussian, constituting $X_{1,i}$ for $i = 1, \ldots, 300$. For each sample, if the observation was on the right half of its respective Gaussian, a sequence of length 50 was drawn from HMM 1, if on the left, from HMM 2. We performed fast variational inference [6], truncating to 50 components, to obtain results for illustration. This precisely defined data set allows the model to clearly display the purpose of its design. If one were to build a Gaussian mixture model on the X_1 data alone, three components would be uncovered, as the posterior expectation shows in Figure 5(a). If an HMM mixture were built alone on the X_2 data, only two components would be uncovered. Using *all* of the data, that is, mixing on $\{O_i\}_{i=1}^{300}$ rather than just $\{X_{1,i}\}_{i=1}^{300}$ or $\{X_{2,i}\}_{i=1}^{300}$ alone, the correct number of six components was uncovered, as shown in Figure 5(b).

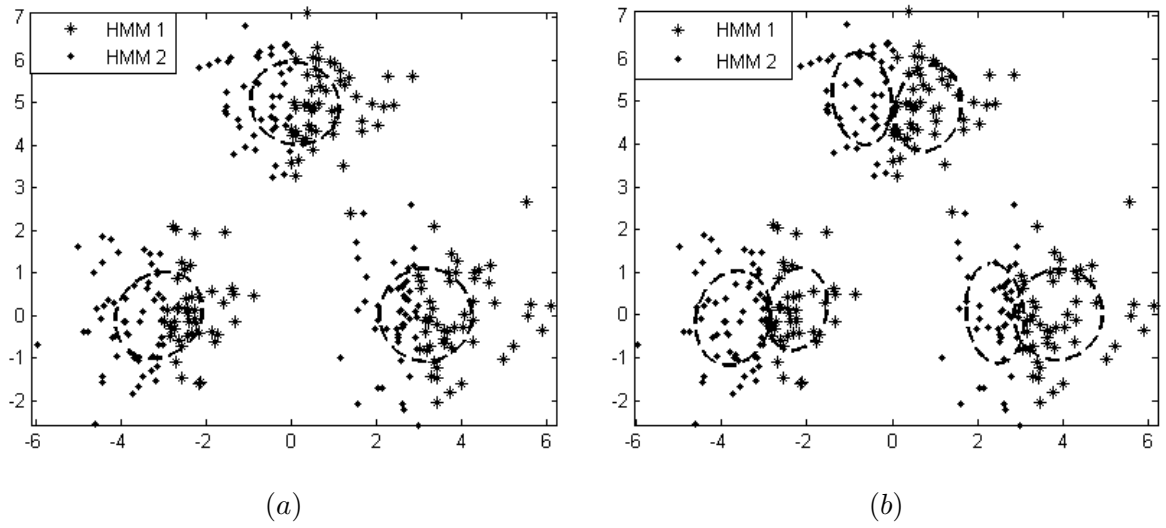


Fig. 5. An example of a mixed Gaussian-HMM data set where observations on the right side of each Gaussian has an associated sequence from HMM 1 and the left side from HMM 2. (a) Gaussian mixture model results without using sequential information. (b) Gaussian-HMM mixture results using both spatial and sequential information. Each ellipse corresponds to a cluster. Because there are six Gaussian-HMM combinations, there are six clusters.

The results show that, as was required by the data, the Dirichlet process prior uncovered six distinct groups of data. In other words, though no two observations were exactly the same, it was inferred that six distinct sets of parameters were required to properly characterize the *statistical* properties of the data. We recall that we initialized to 50 components. Upon convergence, 44 of them contained no data. The DP-PBM framework therefore allows for the incorporation of *all* information of a data set to be included, thus providing more precise clustering results.

REFERENCES

[1] D. Aldous (1985). Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983* 1-198 Springer, Berlin.

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

[3] D. Basu (1955). On statistics independent of a complete sufficient statistic. *Sankhya: The Indian Journal of Statistics*, 15:377-380.

[4] P. Billingsley (1995). *Probability and Measure, 3rd edition*. Wiley Press, New York.

[5] D. Blackwell and J.B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics,* 1:353-355.

[6] M.J. Beal (2003). *Variational Algorithms for Approximate Bayesian Inference* PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

[7] D. Blei and M.I. Jordan (2006). Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis,* 1(1):121-144.

[8] R. Caruana (1997). Multitask learning. *Machine Learning*, 28:41-75.

[9] R.J. Connor and J.E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association,* 64:194-206.

[10] M.D. Escobar and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association.* 90(430):577-588.

[11] T. Ferguson (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics,* 1:209-230.

[12] B. de Finetti (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré,* 7:1-68.

[13] D. Gamerman and H.F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*, Chapman & Hall.

[14] P. Halmos (1944). Random Alms. *The Annals of Mathematical Statistics,* 15:182-189.

[15] E. Hewitt and L.J. Savage (1955). Symmetric measures on Cartesian products. *Trans. American Math Association,* 80:470-501.

[16] J. Jacod and P. Protter (2004). *Probability Essentials, 2nd ed.* Springer.

[17] N. Johnson and S. Kotz (1977). *Urn Models and Their Applications*. Wiley Series in Probability and Mathematical Statistics.

[18] J.F.C. Kingman (1978). Uses of exchangeability. *Annals of Probability,* 6(2):183-197.

[19] E. Kreyszig (1999). *Advanced Engineering Mathematics,* 8^{th} *edition.* John Wiley & Sons, Inc., New York.

[20] R.M. Neal (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics,* 9(2):249-265.

[21] K. Ni, J. Paisley, L. Carin and D. Dunson (2008). Multi-task learning for analyzing and sorting large databases of sequential data. *IEEE Trans. on Signal Processing* to appear.

[22] J. Paisley and L. Carin (2008). Hidden Markov Models with Stick Breaking Priors. *submitted to IEEE Trans. on Signal Processing*.

[23] O. Papaspiliopoulos and G.O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika,* 95(1):169-186.

[24] J. Pitman (1996). Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability and Game Theory. Papers in honor of David Blackwell,* T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds. 245-267.

[25] L.R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* Vol. 77, No 2 pp. 257-286.

[26] A. Rodriquez, D.B. Dunson, and A.E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association,* to appear.

[27] J. Sethuraman (1994). A constructive definition of Dirichlet priors. *Statistica Sinica,* 4:639-650.

[28] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association,* 101(476):1566-1581.

[29] Y. Xue, X. Liao, L. Carin and B. Krishnapuram (2007). Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research,* 8:35-63.

[30] M. West, P. Müller and M.D. Escobar (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pp 363-386.

[31] S.S. Wilks (1962). *Mathematical Statistics*. John Wiley, New York.