Introduction to the Dirichlet Distribution and Related Processes

Bela A. Frigyik, Amol Kapila, and Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195
{gupta}@ee.washington.edu

UWEE Technical Report
Number UWEETR-2010-0006
December 2010

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Abstract

This tutorial covers the Dirichlet distribution, Dirichlet process, Pólya urn (and the associated Chinese restaurant process), hierarchical Dirichlet Process, and the Indian buffet process. Apart from basic properties, we describe and contrast three methods of generating samples: stick-breaking, the Pólya urn, and drawing gamma random variables. For the Dirichlet process we first present an informal introduction, and then a rigorous description for those more comfortable with probability theory.

Contents

1 Introduction to the Dirichlet Distribution

An example of a pmf is an ordinary six-sided die - to sample the pmf you roll the die and produce a number from one to six. But real dice are not exactly uniformly weighted, due to the laws of physics and the reality of manufacturing. A bag of 100 real dice is an example of a *random pmf*- to sample this random pmf you put your hand in the bag and draw out a die, that is, you draw a pmf. A bag of dice manufactured using a crude process 100 years ago will likely have probabilities that deviate wildly from the uniform pmf, whereas a bag of state-of-the-art dice used by Las Vegas casinos may have barely perceptible imperfections. We can model the randomness of pmfs with the Dirichlet distribution.

One application area where the Dirichlet has proved to be particularly useful is in modeling the distribution of words in text documents [9]. If we have a dictionary containing k possible words, then a particular document can be represented by a pmf of length k produced by normalizing the empirical frequency of its words. A group of documents produces a collection of pmfs, and we can fit a Dirichlet distribution to capture the variability of these pmfs. Different Dirichlet distributions can be used to model documents by different authors or documents on different topics.

In this section, we describe the Dirichlet distribution and some of its properties. In Sections 1.2 and 1.4, we illustrate common modeling scenarios in which the Dirichlet is frequently used: first, as a conjugate prior for the multinomial distribution in Bayesian statistics, and second, in the context of the compound Dirichlet (a.k.a. Pólya distribution), which finds extensive use in machine learning and natural language processing.

Then, in Section 2, we discuss how to generate realizations from the Dirichlet using three methods: urn-drawing, stick-breaking, and transforming Gamma random variables. In Sections 3 and 6, we delve into Bayesian non-parametric statistics, introducing the Dirichlet process, the Chinese restaurant process, and the Indian buffet process.

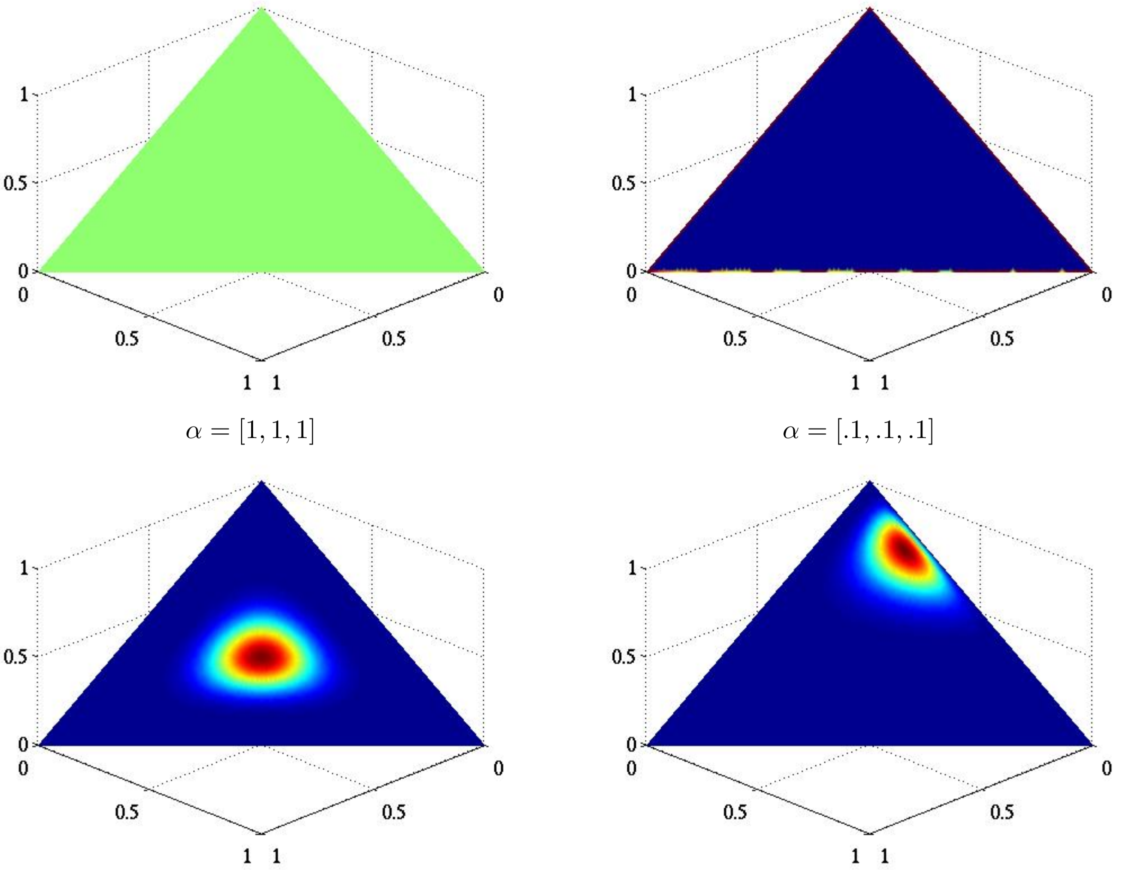
1.1 Definition of the Dirichlet Distribution

A pmf with k components lies on the $(k-1)$-dimensional probability simplex, which is a surface in $\mathbb{R}^k$ denoted by Δ_k and defined to be the set of vectors whose k components are non-negative and sum to 1, that is $\Delta_k = \{q \in \mathbb{R}^k \mid \sum_{i=1}^k q_i = 1, q_i \geq 0 \text{ for } i = 1, 2, \ldots, k\}$. While the set Δ_k lies in a k-dimensional space, Δ_k is itself a $(k-1)$-dimensional object. As an example, Fig. 1 shows the two-dimensional probability simplex for $k = 3$ events lying in three-dimensional Euclidean space. Each point q in the simplex can be thought of as a probability mass function in its own right. This is because each component of q is non-negative, and the components sum to 1. The Dirichlet distribution can be thought of as a probability distribution over the $(k-1)$-dimensional probability simplex Δ_k; that is, as a *distribution over pmfs* of length k.

Dirichlet distribution: Let $Q = [Q_1, Q_2, \ldots, Q_k]$ be a random pmf, that is $Q_i \geq 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^k Q_i = 1$. In addition, suppose that $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k]$, with $\alpha_i > 0$ for each i, and let $\alpha_0 = \sum_{i=1}^k \alpha_i$. Then, Q is said to have a Dirichlet distribution with parameter α, which we denote by $Q \sim \text{Dir}(\alpha)$, if it has[1] $f(q;\alpha) = 0$ if q is not a pmf, and if q is a pmf then

$$f(q;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k q_i^{\alpha_i - 1}, \tag{1}$$

[1] The density of the Dirichlet is positive only on the simplex, which as noted previously, is a $(k-1)$-dimensional object living in a k-dimensional space. Because the density must satisfy $P(Q \in A) = \int_A f(q;\alpha) d\mu(q)$ for some measure μ, we must restrict the measure to being over a $(k-1)$-dimensional space; otherwise, integrating over a $(k-1)$-dimensional subset of a k-dimensional space will always give an integral of 0. Furthermore, to have a density that satisfies this usual integral relation, it must be a density with respect to $(k-1)$-dimensional Lebesgue measure. Hence, technically, the density should be a function of $k-1$ of the k variables, with the k-th variable implicitly equal to one minus the sum of the others, so that all k variables sum to one. The choice of which $k-1$ variables to use in the density is arbitrary. For example, one way to write the density is as follows: $f(q_1, q_2, \ldots, q_{k-1}) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{k-1} q_i\right)^{\alpha_k - 1}$. However, rather than needlessly complicate the presentation, we shall just write the density as a function of the entire k-dimensional vector q. We also note that the constraint that $\sum_i q_i = 1$ forces the components of Q to be dependent.



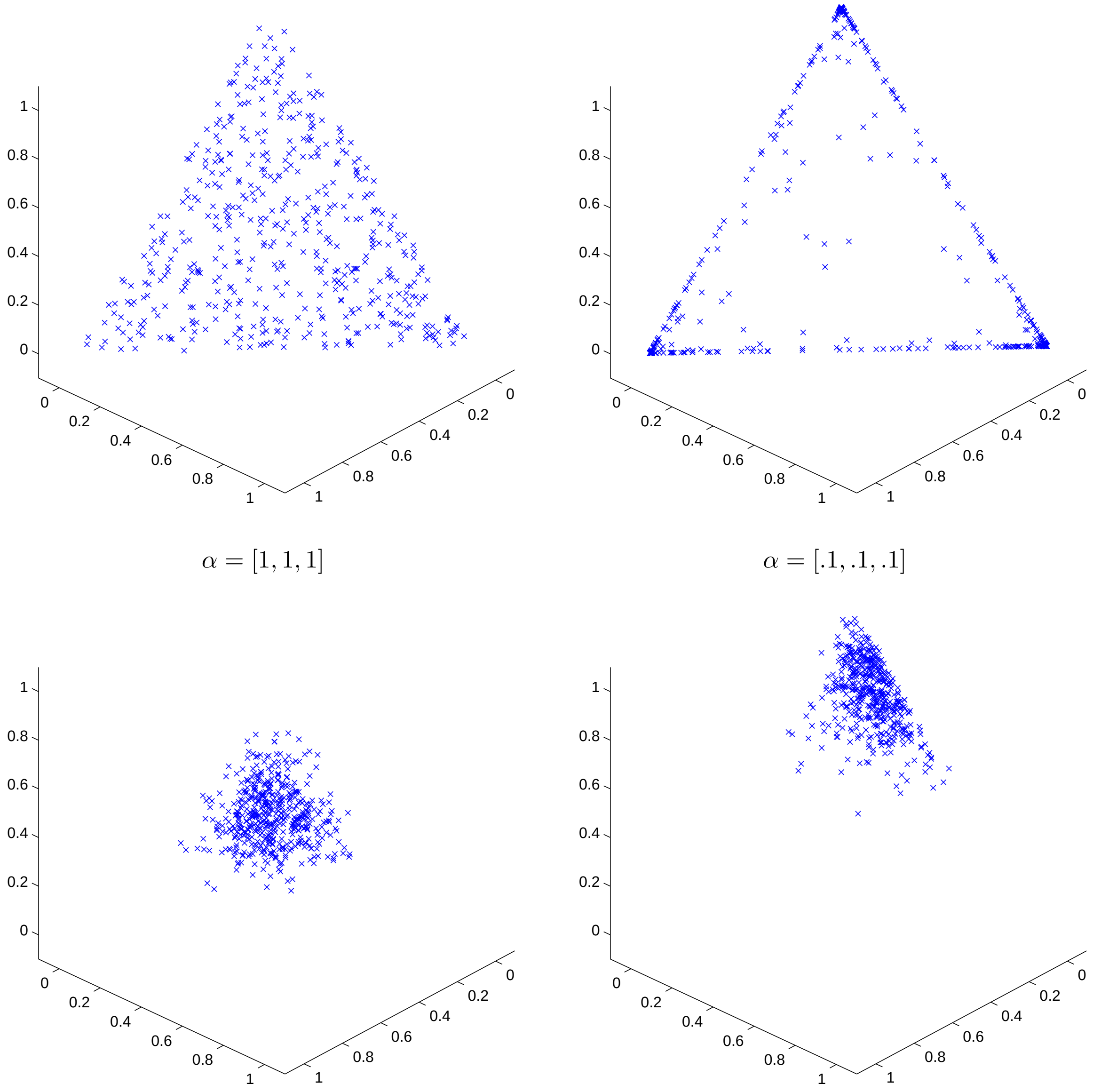
$\alpha = [1, 1, 1]$

$\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$

$\alpha = [2, 5, 15]$

Figure 1: Density plots (blue = low, red = high) for the Dirichlet distribution over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α. When $\alpha = [c, c, c]$ for some $c > 0$, the density is symmetric about the uniform pmf (which occurs in the middle of the simplex), and the special case $\alpha = [1, 1, 1]$ shown in the top-left is the uniform distribution over the simplex. When $0 < c < 1$, there are sharp peaks of density almost at the vertices of the simplex and the density is miniscule away from the vertices. The top-right plot shows an example of this case for $\alpha = [.1, .1, .1]$, one sees only blue (low density) because all of the density is crammed up against the edge of the probability simplex (clearer in next figure). When $c > 1$, the density becomes concentrated in the center of the simplex, as shown in the bottom-left. Finally, if α is not a constant vector, the density is not symmetric, as illustrated in the bottom-right.



$\alpha = [1, 1, 1]$

$\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$

$\alpha = [2, 5, 15]$

Figure 2: Plots of sample pmfs drawn from Dirichlet distributions over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α.

Table 1: Key Notation

$Q \sim \text{Dir}(\alpha)$	random pmf Q coming from a Dirichlet distribution with parameter α
q	pmf, which in this tutorial, will often be a realization of a random pmf $Q \sim \text{Dir}(\alpha)$
q_j	jth component of the pmf q
$q^(i)$	ith pmf of a set of L pmfs
k	number of events the pmf q is defined over, so $q = [q_1, q_2, \ldots, q_k]$
α	parameter of the Dirichlet distribution
α_0	$= \sum_{i=1}^k \alpha_i$
$m = \alpha/\alpha_0$	normalized parameter vector, mean of the Dirichlet
Δ_k	$(k-1)$-dimensional probability simplex living in $\mathbb{R}^k$
v_i	ith entry of the vector v
v_{-i}	the vector v with the i-th entry removed
$\Gamma(s)$	the gamma function evaluated at s, for $s > 0$
$\Gamma(k, \theta)$	Gamma distribution with parameters k and θ
$\overset{D}{=}$	$AD{=}B$ means random variables A and B have the same distribution

where $\Gamma(s)$ denotes the gamma function. The gamma function is a generalization of the factorial function: for $s > 0$, $\Gamma(s+1) = s\Gamma(s)$, and for positive integers n, $\Gamma(n) = (n-1)!$ because $\Gamma(1) = 1$. We denote the mean of a Dirichlet distribution as $m = \alpha/\alpha_0$.

Fig. 1 shows plots of the density of the Dirichlet distribution over the two-dimensional simplex in $\mathbb{R}^3$ for a handful of values of the parameter vector α. When $\alpha = [1, 1, 1]$, the Dirichlet distribution reduces to the uniform distribution over the simplex (as a quick exercise, check this using the density of the Dirichlet in (1).) When the components of α are all greater than 1, the density is monomodal with its mode somewhere in the interior of the simplex, and when the components of α are all less than 1, the density has sharp peaks almost at the vertices of the simplex. Note that the support of the Dirichlet is *open* and does not include the vertices or edge of the simplex, that is, no component of a pmf drawn from a Dirichlet will ever be zero.

Fig. 2 shows plots of samples drawn IID from different Dirichlet distributions.

Table 2 summarizes some key properties of the Dirichet distribution.

When $k = 2$, the Dirichlet reduces to the Beta distribution. The Beta distribution $\text{Beta}(\alpha, \beta)$ is defined on $(0, 1)$ and has density

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}.$$

To make the connection clear, note that if $X \sim \text{Beta}(a, b)$, then $Q = (X, 1 - X) \sim \text{Dir}(\alpha)$, where $\alpha = [a, b]$, and vice versa.

1.2 Conjugate Prior for the Multinomial Distribution

The multinomial distribution is parametrized by an integer n and a pmf $q = [q_1, q_2, \ldots, q_k]$, and can be thought of as follows: If we have n independent events, and for each event, the probability of outcome i is q_i, then the multinomial distribution specifies the probability that outcome i occurs x_i times, for $i = 1, 2, \ldots, k$. For example, the multinomial distribution can model the probability of an n-sample empirical histogram, if each sample is drawn iid from q. If $X \sim \text{Multinomial}_k(n, q)$, then its probability mass function is given by

$$f(x_1, x_2, \ldots, x_k \mid n, q = (q_1, q_2 \ldots, q_k)) = \frac{n!}{x_1!\, x_2! \ldots x_k!} \prod_{i=1}^{k} q_i^{x_i}.$$

When $k = 2$, the multinomial distribution reduces to the binomial distribution.

Table 2: **Properties of the Dirichlet Distribution**

Density	$\frac{1}{B(\alpha)} \prod_{j=1}^{d} q_j^{\alpha_j - 1}$
Expectation	$\frac{\alpha_i}{\alpha_0}$
Covariance	For $i \neq j$, $Cov(Q_i, Q_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}$. and for all i, $Cov(Q_i, Q_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$
Mode	$\frac{\alpha - \mathbf{1}}{\alpha_0 - k}$.
Marginal Distributions	$Q_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$.
Conditional Distribution	$(Q_{-i} \mid Q_i) \sim (1 - Q_i)\,\text{Dir}(\alpha_{-i})$
Aggregation Property	$(Q_1, Q_2, \ldots, Q_i + Q_j, \ldots, Q_k) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_k)$. In general, if $\{A_1, A_2, \ldots, A_r\}$ is a partition of $\{1, 2, \ldots, k\}$, then $\left(\sum_{i \in A_1} Q_i, \sum_{i \in A_2} Q_i, \ldots, \sum_{i \in A_r} Q_i\right) \sim \text{Dir}\left(\sum_{i \in A_1} \alpha_i, \sum_{i \in A_2} \alpha_i, \ldots, \sum_{i \in A_r} \alpha_i\right)$.

The Dirichlet distribution serves as a conjugate prior for the probability parameter q of the multinomial distribution.[2] That is, if $(X \mid q) \sim$ Multinomial$_k(n, q)$ and $Q \sim \text{Dir}(\alpha)$, then $(Q \mid X = x) \sim \text{Dir}(\alpha + x)$.

Proof. Let $\pi(\cdot)$ be the density of the prior distribution for Q and $\pi(\cdot|x)$ be the density of the posterior distribution. Then, using Bayes rule, we have

$$\begin{aligned}
\pi(q \mid x) &= \gamma f(x \mid q)\, \pi(q) \\
&= \gamma \left(\frac{n!}{x_1!\, x_2! \ldots x_k!} \prod_{i=1}^{k} q_i^{x_i} \right) \left(\frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1} \right) \\
&= \tilde{\gamma} \prod_{i=1}^{k} q_i^{\alpha_i + x_i - 1} \\
&= \text{Dir}(\alpha + x).
\end{aligned}$$

Hence, $(Q \mid X = x) \sim \text{Dir}(\alpha + x)$. □

1.3 The Aggregation Property of the Dirichlet

The Dirichlet has a useful fractal-like property that if you lump parts of the sample space together you then have a Dirichlet distribution over the new set of lumped-events. For example, say you have a Dirichlet distribution over six-sided dice with $\alpha \in \mathbb{R}^6_+$, but what you really want to know is what is the probability of rolling an odd number versus the probability of rolling an even number. By aggregation, the Dirichlet

[2] This generalizes the situation in which the Beta distribution serves as a conjugate prior for the probability parameter of the binomial distribution.

distribution over the six dice faces implies a Dirichlet over the two-event sample space of odd vs. even, with aggregated Dirichlet parameter $(\alpha_1 + \alpha_3 + \alpha_5,\ \alpha_2 + \alpha_4 + \alpha_6)$.

In general, the aggregation property of the Dirichlet is that if $\{A_1, A_2, \ldots, A_r\}$ is a partition of $\{1, 2, \ldots, k\}$, then $\left(\sum_{i\in A_1} Q_i, \sum_{i\in A_2} Q_i, \ldots, \sum_{i\in A_r} Q_i\right) \sim \text{Dir}\left(\sum_{i\in A_1}\alpha_i, \sum_{i\in A_2}\alpha_i, \ldots, \sum_{i\in A_r}\alpha_i\right)$.

We prove the aggregation property in Sec. 2.3.1.

1.4 Compound Dirichlet

Consider again a bag of dice, and number the dice arbitrarily $i = 1, 2, \ldots, L$. For the ith die, there is an associated pmf $q^{(i)}$ of length $k = 6$ that gives the probabilities of rolling a one, a two, etc. We will assume that we can model these L pmfs as coming from a $\text{Dir}(\alpha)$ distribution. Hence, our set-up is as follows:

$$\text{Dir}(\alpha) \xrightarrow{iid} \begin{array}{c} q^{(1)} \\ q^{(2)} \\ \vdots \\ q^{(L)} \end{array},$$

where $q^{(1)}, q^{(2)}, \ldots, q^{(L)}$ are pmfs. Further, suppose that we have n_i samples from the ith pmf:

$$\text{Dir}(\alpha) \xrightarrow{iid} \begin{array}{c} q^{(1)} \xrightarrow{iid} x_{1,1}, x_{1,2}, \ldots, x_{1,n_1} \triangleq x_1 \\ q^{(2)} \xrightarrow{iid} x_{2,1}, x_{2,2}, \ldots, x_{2,n_2} \triangleq x_2 \\ \vdots \\ q^{(L)} \xrightarrow{iid} x_{L,1}, x_{L,2}, \ldots, x_{L,n_L} \triangleq x_L. \end{array}$$

Then we say that the $\{x_i\}$ are realizations of a *compound Dirichlet distribution*, also known as a *multivariate Pólya distribution*.

In this section, we will derive the likelihood for α. That is, we will derive the probability of the observed data $\{x_i\}_{i=1}^{L}$, assuming that the parameter value is α. In terms of maximizing the likelihood of the observed samples in order to estimate α, it does not matter whether we consider the likelihood of seeing the samples in the order we saw them, or just the likelihood of seeing those sample-values without regard to their particular order, because these two likelihoods differ by a factor that does not depend on α. Here, we will disregard the order of the observed sample values.

The $i = 1, 2, \ldots, L$ sets of samples $\{x_i\}$ drawn from the L pmfs drawn from the $\text{Dir}(\alpha)$ are conditionally independent given α, so the likelihood of α can be written as the product:

$$p(x \mid \alpha) = \prod_{i=1}^{L} p(x_i \mid \alpha). \tag{2}$$

For each set of the L pmfs, the likelihood $p(x_i \mid \alpha)$ can be expressed using the total law of probability over the possible pmf that generated it:

$$\begin{aligned} p(x_i \mid \alpha) &= \int p(x_i, q^{(i)} \mid \alpha)\, dq^{(i)} \\ &= \int p(x_i \mid q^{(i)}, \alpha)\, p(q^{(i)} \mid \alpha)\, dq^{(i)} \\ &= \int p(x_i \mid q^{(i)})\, p(q^{(i)} \mid \alpha)\, dq^{(i)}. \end{aligned} \tag{3}$$

Next we focus on describing $p(x_i \mid q^{(i)})$ and $p(q^{(i)} \mid \alpha)$ so we can use (3). Let n_{ij} be the number of outcomes in x_i that are equal to j, and let $n_i = \sum_{j=1}^{k} n_{ij}$. Because we are using counts and assuming that order does not matter, we have that $(X_i \mid q^{(i)}) \sim \text{Multinomial}_k(n_i, q^{(i)})$, so

$$p(x_i \mid q^{(i)}) = \frac{n_i!}{\prod_{i=1}^k n_{ij}!} \prod_{j=1}^k \left(q_j^{(i)}\right)^{n_{ij}} . \tag{4}$$

In addition, $(Q^{(i)} \mid \alpha) \sim \text{Dir}(\alpha)$, so

$$p(q^{(i)} \mid \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \left(q_j^{(i)}\right)^{\alpha_j - 1} . \tag{5}$$

Therefore, combining (3), (4), and (5), we have

$$p(x_i \mid \alpha) = \frac{n_i!}{\prod_{j=1}^k n_{ij}!} \frac{\Gamma(\alpha_0)}{\prod_{j=1}^k \Gamma(\alpha_j)} \int \prod_{j=1}^k \left(q_j^{(i)}\right)^{n_{ij}+\alpha_j-1} dq^{(i)}.$$

Focusing on the integral alone,

$$\int \prod_{j=1}^k \left(q^{(i)}\right)^{n_{ij}+\alpha_j-1} dq^{(i)} = \frac{\prod_{j=1}^k \Gamma(n_{ij}+\alpha_j)}{\Gamma(\sum_{j=1}^k (n_{ij}+\alpha_j)} \left(\int \frac{\Gamma(\sum_{j=1}^k (n_{ij}+\alpha_j)}{\prod_{j=1}^k \Gamma(n_{ij}+\alpha_j)} \prod_{j=1}^k \left(q_j^{(i)}\right)^{n_{ij}+\alpha_j-1} dq^{(i)} \right),$$

where the term in brackets on the right evaluates to 1 because it is the integral of the density of the $\text{Dir}(n_i + \alpha - 1)$ distribution.

Hence,

$$p(x_i \mid \alpha) = \frac{n_i!}{\prod_{j=1}^k n_{ij}!} \frac{\Gamma(\alpha_0)}{\Gamma(\sum_{j=1}^k (n_{ij}+\alpha_j))} \prod_{j=1}^k \frac{\Gamma(n_{ij}+\alpha_j)}{\Gamma(\alpha_j)},$$

which can be substituted into (2) to form the likelihood of all the observed data.

In order to find the α that maximizes this likelihood, take the log of the likelihood above and maximize that instead. Unfortunately, there is no closed-form solution to this problem and there may be multiple maxima, but one can find a maximum using optimization methods. See [3] for details on how to use the expectation-maximization (EM) algorithm for this problem and [10] for a broader discussion including using Newton-Raphson. Again, whether we consider ordered or unordered observations does not matter when finding the MLE because the only affect this has on the log-likelihood is an extra term that is a function of the data, not the parameter α.

2 Generating Samples From a Dirichlet Distribution

A natural question to ask regarding any distribution is how to sample from it. In this section, we discuss three methods: (1) a method commonly referred to as Pólya's urn; (2) a "stick-breaking" approach which can be thought of as iteratively breaking off pieces of (and hence dividing) a stick of length one in such a way that the vector of the lengths of the pieces is distributed according to a $\text{Dir}(\alpha)$ distribution; and (3) a method based on transforming Gamma-distributed random variables. We end this section with a discussion comparing these methods of generating samples.

2.1 Pólya's Urn

Suppose we want to generate a realization of $Q \sim \text{Dir}(\alpha)$. To start, put α_i balls of color i for $i = 1, 2, \ldots, k$ in an urn, as shown in the left-most picture of Fig. 3. Note that $\alpha_i > 0$ is not necessarily an integer, so we may have a fractional or even an irrational number of balls of color i in our urn! At each iteration, draw one ball uniformly at random from the urn, and then place it back into the urn along with an additional ball

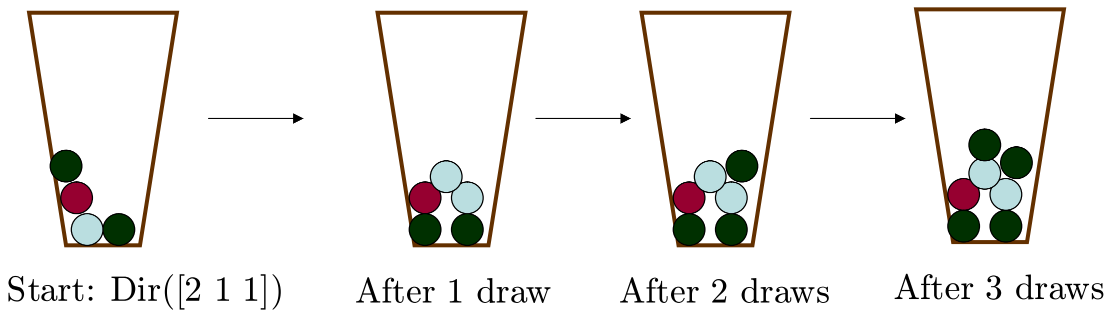


Figure 3: Visualization of the urn-drawing scheme for $Dir([2\ 1\ 1])$, discussed in Section 2.1.

of the same color. As we iterate this procedure more and more times, the proportions of balls of each color will converge to a pmf that is a sample from the distribution $\text{Dir}(\alpha)$.

Mathematically, we first generate a sequence of balls with colors $(X_1, X_2, \ldots)$ as follows:

Step 1: Set a counter $n = 1$. Draw $X_1 \sim \alpha/\alpha_0$. (Note that α/α_0 is a non-negative vector whose entries sum to 1, so it is a pmf.)

Step 2: Update the counter to $n + 1$. Draw $X_{n+1} \mid X_1, X_2, \ldots, X_n \sim \alpha_n/\alpha_{n0}$, where $\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{X_i}$ and α_{n0} is the sum of the entries of α_n. Repeat this step an infinite number of times.

Once you have finished Step 2, calculate the proportions of the different colors: let $Q_n = (Q_{n1}, Q_{n2}, \ldots, Q_{nk})$, where Q_{ni} is the proportion of balls of color i after n balls are in the urn. Then, $Q_n \rightarrow_d Q \sim \text{Dir}(\alpha)$ as $n \rightarrow \infty$, where $\rightarrow_d$ denotes convergence in distribution. That is, $P(Q_{n1} \leq z_1, Q_{n2} \leq z_2, \ldots, Q_{nk} \leq z_k) \rightarrow P(Q_1 \leq z_1, Q_2 \leq z_2, \ldots, Q_k \leq z_k)$ as $n \rightarrow \infty$ for all $(z_1, z_2, \ldots, z_k)$.

Note that this does NOT mean that in the limit as the number of balls in the urn goes to infinity the probability of drawing balls of each color is given by the pmf α/α_0.

Instead, asymptotically, the probability of drawing balls of each color is given by a pmf that is a realization of the distribution $\text{Dir}(\alpha)$. Thus asymptotically we have a sample from $\text{Dir}(\alpha)$. The proof relies on the Martingale Convergence Theorem, which is beyond the scope of this tutorial.

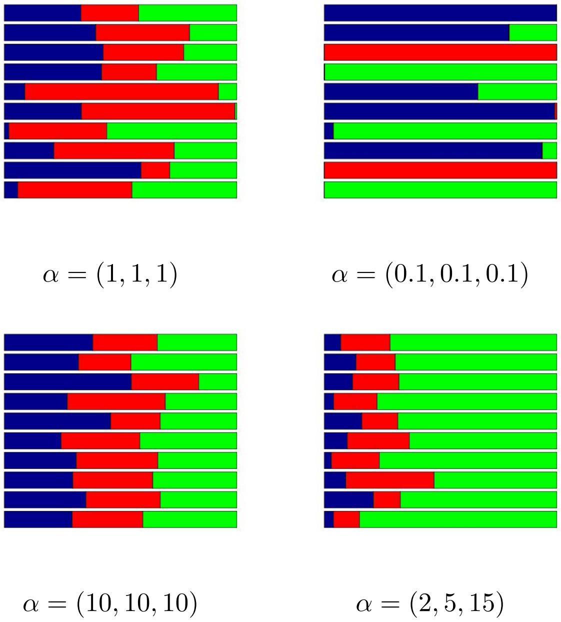
2.2 The Stick-breaking Approach

The stick-breaking approach to generating a random vector with a $\text{Dir}(\alpha)$ distribution involves iteratively breaking a stick of length 1 into k pieces in such a way that the lengths of the k pieces follow a $\text{Dir}(\alpha)$ distribution. Figure 4 illustrates this process with simulation results. We will assume that we know how to generate random variables from the Beta distribution. In the case where α has length 2, simulating from the Dirichlet is equivalent to simulating from the Beta distribution, so henceforth in this section, we will assume that $k \geq 3$.

2.2.1 Basic Idea

For ease of exposition, we will first assume that $k = 3$, and then generalize the procedure to $k > 3$. Over the course of the stick-breaking process, we will be keeping track of a set of intermediate values $\{u_i\}$, which we use to ultimately calculate the realization q. To begin, we generate Q_1 from $\text{Beta}(\alpha_1, \alpha_2 + \alpha_3)$ and set u_1 equal to its value: $u_1 = q_1$. Then, generate $\left(\frac{Q_2}{1-Q_1} \mid Q_1\right)$ from $\text{Beta}(\alpha_2, \alpha_3)$. Denote the result by u_2, and set $q_2 = (1 - u_1)u_2$. The resulting vector $u = (u_1, (1 - u_1)u_2, 1 - u_1 - (1 - u_1)u_2)$ comes from a Dirichlet distribution with parameter vector α. This procedure can be conceptualized as breaking off pieces of a stick of length one in a random way such that the lengths of the k pieces follow a $\text{Dir}(\alpha)$ distribution.

Now, let $k > 3$.



$\alpha = (1, 1, 1)$ $\alpha = (0.1, 0.1, 0.1)$

$\alpha = (10, 10, 10)$ $\alpha = (2, 5, 15)$

Figure 4: Visualization of the Dirichlet distribution as breaking a stick of length 1 into pieces, with the mean length of piece i being α_i/α_0. For each value of α, we have simulated 10 sticks. Each stick corresponds to a realization from the Dirichlet distribution. For $\alpha = (c, c, c)$, we expect the mean length for each color, or component, to be the same, with variability decreasing as $\alpha \to \infty$. For $\alpha = (2, 5, 15)$, we would naturally expect the third component to dominate.

Step 1: Simulate $u_1 \sim \text{Beta}\left(\alpha_1, \sum_{i=2}^{k} \alpha_i\right)$, and set $q_1 = u_1$. This is the first piece of the stick. The remaining piece has length $1 - u_1$.

Step 2: For $2 \leq j \leq k-1$, if $j-1$ pieces, with lengths $u_1, u_2, \ldots, u_{j-1}$, have been broken off, the length of the remaining stick is $\prod_{i=1}^{j-1}(1-u_i)$. We simulate $u_j \sim \text{Beta}\left(\alpha_j, \sum_{i=j+1}^{k} \alpha_i\right)$ and set $q_j = u_j \prod_{i=1}^{j-1}(1-u_i)$. The length of the remaining part of the stick is $\prod_{i=1}^{j-1}(1-u_i) - u_j \prod_{i=1}^{j-1}(1-u_i) = \prod_{i=1}^{j}(1-u_i)$.

Step 3: The length of the remaining piece is q_k.

Note that at each step, if $j-1$ pieces have been broken off, the remainder of the stick, with length $\prod_{i=1}^{j-1}(1-u_i)$, will eventually be broken up into $k-j+1$ pieces with proportions distributed according to a $\text{Dir}(\alpha_j, \alpha_{j+1}, \ldots, \alpha_k)$ distribution.

2.2.2 Neutrality, Marginal, and Conditional Distributions

The reason why the stick-breaking method generates random vectors from the Dirichlet distribution relies on a property of the Dirichlet called *neutrality*, which we discuss and prove below. In addition, the marginal and conditional distributions of the Dirichlet will fall out of our proof of the neutrality property for the Dirichlet.

Neutrality Let $Q = (Q_1, Q_2, \ldots, Q_k)$ be a random vector. Then, we say that Q is *neutral* if for each $j = 1, 2, \ldots, k$, Q_j is independent of the random vector $\frac{1}{1-Q_j} Q_{-j}$. In the case where $Q \sim \text{Dir}(\alpha)$, $\frac{1}{1-Q_j} Q_{-j}$ is simply the vector Q with the j-th component removed, and then scaled by the sum of the remaining elements. Furthermore, if $Q \sim \text{Dir}(\alpha)$, then Q exhibits the neutrality property, a fact we prove below.

Proof of Neutrality for the Dirichlet: Without loss of generality, this proof is written for the case that $j = k$. Let $Y_i = \frac{Q_i}{1-Q_k}$ for $i = 1, 2, \ldots, k-2$, $Y_{k-1} = 1 - \sum_{i=1}^{k-2} Q_i$, and $Y_k = Q_k$. Consider the following transformation T of coordinates between $(Y_1, Y_2, \ldots, Y_{k-2}, Y_k)$ and $(Q_1, Q_2, \ldots, Q_{k-2}, Q_k)$:

$$(Q_1, Q_2, \ldots, Q_{k-2}, Q_k) = T(Y_1, Y_2, \ldots, Y_{k-2}, Y_k) = (Y_1(1-Y_k), Y_2(1-Y_k), \ldots, Y_{k-2}(1-Y_k), Y_k).$$

The Jacobian of this transformation is

$$\begin{pmatrix} 1-Y_k & 0 & 0 & \cdots & 0 & -Y_1 \\ 0 & 1-Y_k & 0 & \cdots & 0 & -Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1-Y_k & -Y_{k-2} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{6}$$

which has determinant with absolute value equal to $(1-Y_k)^{k-2}$.

The standard change-of-variables formula tells us that the density of Y is $f(y) = (g \circ T)(y) \times |\det(T)|$, where

$$g(q) = g(q_1, q_2, \ldots, q_{k-2}, q_k; \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i \neq k-1} q_i^{\alpha_i - 1}\right) \left(1 - \sum_{i \neq k-1} q_i\right)^{\alpha_{k-1}-1} \tag{7}$$

is the joint density of Q. Substituting (7) into our change of variables formula, we find the joint density of the new random variables:

$$f(y;\alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-2} (y_i(1-y_k))^{\alpha_i - 1}\right) y_k^{\alpha_k - 1} \left(1 - \sum_{i=1}^{k-2} y_i(1-y_k) - y_k\right)^{\alpha_{k-1}-1} (1-y_k)^{k-2}.$$

We can simplify one of the terms of the above by pulling out a $(1-y_k)$:

$$\begin{aligned} 1 - \sum_{i=1}^{k-2} y_i(1-y_k) - y_k &= (1-y_k)\left(1 - \sum_{i=1}^{k-2} y_i\right) \\ &= y_{k-1}(1-y_k). \end{aligned}$$

Hence,

$$f(y;\alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-1} y_i^{\alpha_i - 1}\right) y_k^{\alpha_k - 1} (1-y_k)^w,$$

where $w = \sum_{i=1}^{k-2}(\alpha_i - 1) + \alpha_{k-1} - 1 + k - 2 = \sum_{i=1}^{k-1} \alpha_i - 1$, so we have

$$\begin{aligned} f(y;\alpha) &= \left[\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\Gamma(\alpha_k)\Gamma\left(\sum_{i=1}^{k-1} \alpha_i\right)} y_k^{\alpha_k - 1}(1-y_k)^{\sum_{i=1}^{k-1}\alpha_i - 1}\right] \left[\frac{\Gamma\left(\sum_{i=1}^{k-1} \alpha_i\right)}{\prod_{i=1}^{k-1} \Gamma(\alpha_i)} \prod_{i=1}^{k-1} y_i^{\alpha_i - 1}\right] \\ &= f_1(y_k; \alpha_k, \sum_{i=1}^{k-1} \alpha_i) f_2(y_1, y_2, \ldots, y_{k-1}; \alpha_1, \alpha_2, \ldots, \alpha_{k-1}) \\ &= f_1(q_k; \alpha_k, \sum_{i=1}^{k-1} \alpha_i) f_2\left(\frac{q_1}{1-q_k}, \frac{q_2}{1-q_k}, \ldots, \frac{q_{k-1}}{1-q_k}; \alpha_{-k}\right), \end{aligned}$$

where

$$f_1(y_k; \alpha_k, \sum_{i=1}^{k-1} \alpha_i) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\Gamma(\alpha_k)\Gamma\left(\sum_{i=1}^{k-1} \alpha_i\right)} y_k^{\alpha_k - 1}(1-y_k)^{\sum_{i=1}^{k-1}\alpha_i - 1} \tag{8}$$

is the density of a Beta distribution with parameters α_k and $\sum_{i=1}^{k-1}\alpha_i$, while

$$f_2(y_1, y_2, \ldots, y_{k-1}; \alpha_{-k}) = \frac{\Gamma\left(\sum_{i=1}^{k-1}\alpha_i\right)}{\prod_{i=1}^{k-1}\Gamma(\alpha_i)} \prod_{i=1}^{k-1} y_i^{\alpha_i - 1} \tag{9}$$

is the density of a Dirichlet distribution with parameter α_{-k}. Hence, the joint density of Y factors into a density for Y_k and a density for $(Y_1, Y_2, \ldots, Y_{k-1})$, so Y_k is independent of the rest, as claimed.

In addition to proving the neutrality property above, we have proved that the marginal distribution of Q_k, which is equal to the marginal distribution of Y_k by definition, is Beta $\left(\alpha_k, \sum_{i=1}^{k-1}\alpha_i\right)$. By replacing k with $j = 1, 2, \ldots, k$ in the above derivation, we have that the marginal distribution of Q_j is Beta $\left(\alpha_j, \sum_{i\neq j}\alpha_j\right)$. This implies that

$$\begin{aligned} f(y_{-j} \mid y_j) &= \frac{f(y;\alpha)}{f_1(y_j;\alpha)} \\ &= f_2(y_1, y_2, \ldots, y_{k-1}; \alpha_{-k}) = \frac{\Gamma\left(\sum_{i=1}^{k-1}\alpha_i\right)}{\prod_{i=1}^{k-1}\Gamma(\alpha_i)} \prod_{i=1}^{k-1} y_i^{\alpha_i - 1}, \end{aligned}$$

so

$$\begin{aligned} &(Y_{-j} \mid Y_j) \sim \mathrm{Dir}(\alpha_{-j}) \\ &\Rightarrow \left(\left(\frac{Q_{-j}}{1-Q_j}\right) \mid Q_j\right) \sim \mathrm{Dir}(\alpha_{-j}) \\ &\Rightarrow (Q_{-j} \mid Q_j) \sim (1-Q_j)\,\mathrm{Dir}(\alpha_{-j}). \end{aligned} \tag{10}$$

2.2.3 Connecting Neutrality, Marginal Distributions, and Stick-breaking

We can use the neutrality property and the marginal and conditional distributions derived above to rigorously prove that the stick-breaking approach works as advertised. The basic reasoning of the proof (and by extension, of the stick-breaking approach) is that in order to sample from the joint distribution of $(Q_1, Q_2, \ldots, Q_k) \sim \mathrm{Dir}(\alpha)$, it is sufficient to first sample from the marginal distribution of Q_1 under α, then sample from the conditional distribution of $(Q_2, Q_3, \ldots, Q_k \mid Q_1)$ under α. We apply this idea recursively in what follows.

Case 1: $j = 1$: From Section 2.2.2, marginally, $Q_1 \sim$ Beta $\left(\alpha_1, \sum_{i=2}^{k}\alpha_i\right)$, which corresponds to the method for assigning a value to Q_1 in the stick-breaking approach. What remains is to sample from $((Q_2, Q_3, \ldots, Q_k) \mid Q_1)$, which we know from (10), is distributed as $(1-Q_1)\,\mathrm{Dir}(\alpha_2, \alpha_3, \ldots, \alpha_k)$. So, in the case of $j = 1$, we have broken off the first piece of the stick according to the marginal distribution of Q_1, and the length of the remaining stick is $1 - Q_1$, which we break into pieces using a vector of proportions from the $\mathrm{Dir}(\alpha_2, \alpha_3, \ldots, \alpha_k)$ distribution.

Case 2: $2 \leq j \leq k-2$, which we treat recursively: Suppose that $j-1$ pieces have been broken off, and the length of the remainder of the stick is $\prod_{i=1}^{j-1}(1-Q_i)$. This is analogous to having sampled from the marginal distribution of $(Q_1, Q_2, \ldots, Q_{j-1})$, and still having to sample from the conditional distribution $((Q_j, Q_{j+1}, \ldots Q_k) \mid (Q_1, Q_2, \ldots, Q_{j-1}))$, which from the previous step in the recursion is distributed as $\left[\prod_{i=1}^{j-1}(1-Q_i)\right]\mathrm{Dir}(\alpha_j, \alpha_{j+1}, \ldots, \alpha_k)$. Hence, using the marginal distribution in (8), we have that $(Q_j \mid (Q_1, Q_2, \ldots, Q_{j-1})) \sim \left[\prod_{i=1}^{j-1}(1-Q_i)\right]$ Beta $\left(\alpha_j, \sum_{i=j+1}^{k}\alpha_i\right)$, and from (9) and (10), we have that

$$
\begin{aligned}
&((Q_{j+1}, Q_{j+2}, \ldots, Q_k) \mid (Q_1, Q_2, \ldots, Q_j)) \\
&\sim \left[\prod_{i=1}^{j-1}(1-Q_i)\right](1-Q_j)\,\mathrm{Dir}(\alpha_{j+1}, \alpha_{j+2}, \ldots, \alpha_k) \\
&\overset{D}{=} \left[\prod_{i=1}^{j}(1-Q_i)\right]\mathrm{Dir}(\alpha_{j+1}, \alpha_{j+2}, \ldots, \alpha_k),
\end{aligned}
$$

where $=_d$ means equal in distribution. This completes the recursion.

Case 3: $j = k-1, k$: Picking up from the case of $j = k-2$ above, we have that $((Q_{k-1}, Q_k) \mid (Q_1, Q_2, \ldots, Q_{k-2})) \sim \left[\prod_{i=1}^{k-2}(1-Q_i)\right]\mathrm{Dir}(\alpha_{k-1}, \alpha_k)$. Hence, we simply split the remainder of the stick into two pieces by drawing $Q_{k-1} \sim \left[\prod_{i=1}^{k-2}(1-Q_i)\right]\mathrm{Beta}(\alpha_{k-1}, \alpha_k)$ and allowing Q_k to be the remainder.

Note that the aggregation property (see Table 2) is a conceptual corollary of the stick-breaking view of the Dirichlet distribution. We provide a rigorous proof of the aggregation property in the next section.

2.3 Generating the Dirichlet from Gamma RVs

We will argue that generating samples from the Dirichlet distribution using Gamma random variables is more computationally efficient than both the urn-drawing method and the stick-breaking method. This method has two steps which we explain in more detail and prove in this section:

Step 1: Generate gamma realizations: for $i = 1, \ldots, k$, draw a number z_i from $\Gamma(\alpha_i, 1)$.

Step 2: Normalize them to form a pmf: for $i = 1, \ldots, k$, set $q_i = \frac{z_i}{\sum_{j=1}^k z_j}$. Then q is a realization of $\mathrm{Dir}(\alpha)$.

. The Gamma distribution $\Gamma(\kappa, \theta)$ is defined by the following probability density:

$$
f(x; \kappa, \theta) = x^{\kappa-1}\frac{e^{-x/\theta}}{\theta^\kappa \Gamma(\kappa)}. \tag{11}
$$

$\kappa > 0$ is called the *shape* parameter, and $\theta > 0$ is called the *scale* parameter.[3,4] One important property of the Gamma distribution that we will use below is the following: Suppose $X_i \sim \Gamma(\kappa_i, \theta)$ are independent for $i = 1, 2, \ldots, n$; that is, they are on the same scale but can have different shapes. Then, $S = \sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \kappa_i, \theta\right)$.

To prove that the above procedure creating Dirichlet samples from Gamma r.v. draws works, we use the change-of-variables formula to show that the density of Q is the density corresponding to the $\mathrm{Dir}(\alpha)$ distribution. First, recall that the original variables are $\{Z_i\}_1^k$, and the new variables are $Z, Q_1, \ldots, Q_{k-1}$. We relate them using the transformation T:

$$
(Z_1, \ldots, Z_k) = T(Z, Q_1, \ldots, Q_{k-1}) = \left(ZQ_1, \ldots, ZQ_{k-1}, Z\left(1 - \sum_{i=1}^{k-1} Q_i\right)\right).
$$

[3] There is an alternative commonly-used parametrization of the Gamma distribution (denoted by $\Gamma(\alpha, \beta)$) with pdf $f(x; \alpha, \beta) = \beta^\alpha x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$. To switch between parametrizations, we set $\alpha = \kappa$ and $\beta = 1/\theta$. β is called the *rate* parameter. We will use the shape parametrization in what follows.

[4] Note that the symbol Γ (the Greek letter gamma) is used to denote both the gamma function and the Gamma distribution, regardless of the parametrization. However, because the gamma function only takes one argument and the Gamma distribution has two parameters, the meaning of the symbol Γ is assumed to be clear from its context.

The Jacobian matrix (matrix of first derivatives) of this transformation is:

$$J(T) = \begin{pmatrix} Q_1 & Z & 0 & 0 & \cdots & 0 \\ Q_2 & 0 & Z & 0 & \cdots & 0 \\ Q_3 & 0 & 0 & Z & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Q_{k-1} & 0 & 0 & \cdots & 0 & Z \\ 1 - \sum_1^{k-1} Q_i & -Z & -Z & -Z & \cdots & -Z \end{pmatrix}, \tag{12}$$

which has determinant Z^{k-1}.

The standard change-of-variables formula tells us that the density of $(Z, Q_1, \ldots, Q_{k-1})$ is $f = g \circ T \times |\det(T)|$, where

$$g(z_1, z_2, \ldots, z_k; \alpha_1, \ldots, \alpha_k) = \prod_{i=1}^{k} z_i^{\alpha_i - 1} \frac{e^{-z_i}}{\Gamma(\alpha_i)} \tag{13}$$

is the joint density of the original (independent) random variables. Substituting (13) into our change of variables formula, we find the joint density of the new random variables:

$$\begin{aligned} f(z, q_1, \ldots, q_{k-1}) &= \left(\prod_{i=1}^{k-1} (zq_i)^{\alpha_i - 1} \frac{e^{-zq_i}}{\Gamma(\alpha_i)} \right) \left[\left(z \left(1 - \sum_{i=1}^{k-1} q_i \right) \right)^{\alpha_k - 1} \frac{e^{-z\left(1 - \sum_{i=1}^{k-1} q_i\right)}}{\Gamma(\alpha_k)} \right] z^{k-1} \\ &= \frac{\left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)^{\alpha_k - 1}}{\prod_{i=1}^{k} \Gamma(\alpha_i)} z^{\left(\sum_{i=1}^{k} \alpha_i\right) - 1} e^{-z}. \end{aligned}$$

Integrating over z, the marginal distribution of $\{Q_i\}_{i=1}^{k-1}$ is

$$\begin{aligned} f(q) = f(q_1, \ldots, q_{k-1}) &= \int_0^\infty f(z, q_1, \ldots, q_{k-1}) dz \\ &= \frac{\left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)^{\alpha_k - 1}}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int_0^\infty z^{\left(\sum_{i=1}^{k} \alpha_i\right) - 1} e^{-z} dz \\ &= \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} q_i \right)^{\alpha_k - 1}, \end{aligned}$$

which is the same as the Dirichlet density in (1).

Hence, our procedure for simulating from the Dirichlet distribution using Gamma-distributed random variables works as claimed.

2.3.1 Proof of the Aggregation Property of the Dirichlet

We have now introduced the tools needed to prove the Dirichlet's aggregation property as stated in Sec. 1.3.

Proof of the Aggregation Property: We rely on the property of the Gamma distribution that says that if $X_i \sim \Gamma(\kappa_i, \theta)$ for $i = 1, 2, \ldots, n$, then $\sum_{i=1}^{n} X_i \sim \Gamma\left(\sum_{i=1}^{n} \kappa_i, \theta\right)$. Suppose $(Q_1, Q_2, \ldots, Q_k) \sim \mathrm{Dir}(\alpha)$.

Then, we know that $Q = Z/\left(\sum_{i=1}^{k} Z_i\right)$, where $Z_i \sim \Gamma(\alpha_i, \theta)$ are independent. Then,

$$\begin{aligned}
&\left(\sum_{i \in A_1} Q_i, \sum_{i \in A_2} Q_i, \ldots, \sum_{i \in A_r} Q_i\right) \\
&= \frac{1}{\sum_{i=1}^{k} Z_i} \left(\sum_{i \in A_1} Z_i, \sum_{i \in A_2} Z_i, \ldots, \sum_{i \in A_r} Z_i\right) \\
&=_d \frac{1}{\sum_{i=1}^{k} \Gamma(\alpha_i, 1)} \left(\Gamma\left(\sum_{i \in A_1} \alpha_i, 1\right), \Gamma\left(\sum_{i \in A_2} \alpha_i, 1\right), \ldots, \Gamma\left(\sum_{i \in A_r} \alpha_i, 1\right)\right) \\
&=_d \operatorname{Dir}\left(\sum_{i \in A_1} \alpha_i, \sum_{i \in A_2} \alpha_i, \ldots, \sum_{i \in A_r} \alpha_i\right).
\end{aligned}$$

2.4 Discussion on Generating Dirichlet Samples

Let us compare the three methods we have presented to generate samples from a Dirichlet: the Pólya urn, stick-breaking, and the Gamma transform. The Pólya urn method is the least efficient because it depends on a convergence result, and as famed economist John Maynard Keynes once noted, "In the long run, we are all dead." One needs to iterate the urn-drawing scheme many times to get good results, and for any finite number of iterations of the scheme, the resulting pmf is not perfectly accurate.

Both the stick-breaking approach and the Gamma-based approach result in pmfs distributed exactly according to a Dirichlet distribution.[5] However, if we assume that it takes the same amount of time to generate a Gamma random variable as it does a Beta random variable, then the stick-breaking approach is more computationally costly. The reason for this is that at each iteration of the stick-breaking procedure, we need to perform the additional intermediate steps of summing the tail of the α vector before drawing from the Beta distribution and then multiplying by $\prod_{i=1}^{j-1}(1 - Q_i)$. With the Gamma-based approach, all we need to do after drawing Gamma random variables is to divide them all by their sum, once.

3 The Dirichlet Process: An Informal Introduction

We first begin our description of the Dirichlet process with an informal introduction, and then we present more rigorous mathematical descriptions and explanations of the Dirichlet process in Section 4.

3.1 The Dirichlet Process Provides a Random Distribution over Distributions over Infinite Sample Spaces

Recall that the Dirichlet distribution is a probability distribution over pmfs, and we can say a random pmf has a Dirichlet distribution with parameter α. A random pmf is like a bag full of dice, and a realization from the Dirichlet gives us a specific die. The Dirichlet distribution is limited in that it assumes a finite set of events. In the dice analogy, this means that the dice must have a finite number of faces. The Dirichlet process enables us work with an infinite set of events, and hence to model probability distributions over infinite sample spaces.

As another analogy, imagine that we stop someone on the street and ask them for their favorite color. We might limit their choices to `black`, `pink`, `blue`, `green`, `orange`, `white`. An individual might provide a

[5] This is true as long as we can perfectly sample from the Beta and Gamma distributions, respectively. All "random-number generators" used by computers are technically *pseudo*-random. Their output appears to produce sequences of random numbers, but in reality, they are not. However, as long as we don't ask our generator for too many random numbers, the pseudo-randomness will not be apparent and will generally not influence the results of the simulation. How many numbers is *too many*? That depends on the quality of the random-number generator being used.

different answer depending on his mood, and you could model the probability that he chooses each of these colors as a pmf. Thus, we are modeling each person as a pmf over the six colors, and we can think of each person's pmf over colors as a realization of a draw from a Dirichlet distribution over the set of six colors.

But what if we didn't force people to choose one of those six colors? What if they could name any color they wanted? There is an infinite number of colors they could name. To model the individuals' pmfs (of infinite length), we need a distribution over distributions over an infinite sample space. One solution is the Dirichlet process, which is a random distribution whose realizations are distributions over an arbitrary (possibly infinite) sample space.

3.2 Realizations From the Dirichlet Process Look Like a Used Dartboard

The set of all probability distributions over an infinite sample space is unmanageable. To deal with this, the Dirichlet process restricts the class of distributions under consideration to a more manageable set: discrete probability distributions over the infinite sample space that can be written as an infinite sum of weighted indicator functions. You can think of your infinite sample space as a dartboard, and a realization from a Dirichlet is a probability distribution on the dartboard marked by an infinite set of darts of different lengths (weights).

The k^{th} indicator δ_{y_k} marks the location of the k^{th} dart-of-probability such that $\delta_{y_k}(B) = 1$ if $y_k \in B$, and $\delta_{y_k}(B) = 0$ otherwise. Each realization of a Dirichlet process has a different and infinite set of these dart locations. Further, the k^{th} dart has a corresponding probability weight $p_k \in [0, 1]$ and $\sum_{k=1}^{\infty} p_k = 1$. So, for some set B of the infinite sample space, a realization of the Dirichlet process will assign probability $P(B)$ to B, where

$$P(B) = \sum_{k=1}^{\infty} p_k \delta_{y_k}(B). \tag{14}$$

Dirichlet processes have found widespread application to discrete sample spaces like the set of all words, or the set of all webpages, or the set of all products. However, because realizations from the Dirichlet process are atomic, they are not a useful model for many continuous scenarios. For example, say we let someone pick their favorite color from a continuous range of colors, and we would like to model the probability distribution over that space. A realization of the Dirichlet process might give a positive probability to a particular shade of dark blue, but zero probability to adjacent shades of blue, which feels like a poor model for this case. However, in cases the Dirichlet process might be a fine model. Say we ask color professionals to name their favorite color, then it would be reasonable to assign a finite atom of probability to *Coca-cola can red*, but zero probability to the nearby colors that are more difficult to name. Similarly, darts of probability would be needed for *international Klein blue*, *530 nm pure green*, all the Pantone colors, etc., all because of their "nameability." As another example, suppose we were asking people their favorite number. Then, one realization of the Dirichlet process might give most of its weight to the numbers $1, 2, \ldots, 10$, a little weight to π, hardly any weight to $\sin(1)$, and zero weight to 1.00000059483813 (I mean, whose favorite number is that?).

As we detail in later sections, the locations of the darts are independent, and the probability weight associated with the k^{th} dart is independent of its location. However, the weights on the darts are not independent. Instead of a vector α with one component per event in our six-color sample space, the Dirichlet process is parameterized by a function (specifically, a measure) α over the sample space of all possible colors $\mathcal{X}$. Note that α is a finite positive function, so it can be normalized to be a probability distribution β. The locations of the darts y_k are drawn iid from β. The weights on the darts p_k are a decreasing sequence, and are a somewhat complicated function of $\alpha(\mathcal{X})$, the total mass of α.

3.3 The Dirichlet Process Becomes a Dirichlet Distribution For Any Finite Partition of the Sample Space

One of the nice things about the Dirichlet distribution is that we can aggregate components of a Dirichlet-distributed random vector and still have a Dirichlet distribution (see Table 2).

The Dirichlet process is constructed to have a similar property. Specifically, any finite partition of the sample space of a Dirichlet process will have a Dirichlet distribution. In particular, say we have a Dirichlet process with parameter α (which you can think of as a positive function over the infinite sample space $\mathcal{X}$). And now we partition the sample space $\mathcal{X}$ into M subsets of events $\{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \ldots, \mathcal{B}_M\}$. For example, after people tell us their favorite color, we might categorize their answer as one of $M = 11$ major color categories (red, green, blue, etc). Then, the Dirichlet process implies a Dirichlet distribution over our new M-color sample space with parameter $(\alpha(\mathcal{B}_1), \alpha(\mathcal{B}_2), \alpha(\mathcal{B}_M))$, where $\alpha(\mathcal{B}_i)$ for any $i = 1, 2, \ldots, M$ is the integral of the indicator function of $\mathcal{B}_i$ with respect to α:

$$\alpha(\mathcal{B}_i) = \int_{\mathcal{X}} \mathbb{1}_{\mathcal{B}_i}(x) d\alpha(x),$$

where $\mathbb{1}_{\mathcal{B}_i}$ is the indicator function of $\mathcal{B}_i$.

4 Formal Description of the Dirichlet Process

A mathematically rigorous description of the Dirichlet process requires a basic understanding of measure theory. Readers who are not familiar with measure theory can pick up the necessary concepts from the very short tutorial *Measure Theory for Dummies* [6], available online.

4.1 Some Necessary Notation

Let $\mathcal{X}$ be a set, and let $\mathscr{B}$ be a σ-algebra on $\mathcal{X}$; that is $\mathscr{B}$ is a non-empty collection of subsets of $\mathcal{X}$ such that (1) $\mathcal{X} \in \mathscr{B}$; (2) if a set B is in $\mathscr{B}$ then its complement B^c is in $\mathscr{B}$; and (3) if $\{B_i\}_{i=1}^{\infty}$ is a countable collection of sets in $\mathscr{B}$ then their union $\cup_{i=1}^{\infty} B_i$ is in $\mathscr{B}$. We call the couple $(\mathcal{X}, \mathscr{B})$ a measurable space. The function $\mu : \mathscr{B} \to [0, \infty]$ defined on elements of $\mathscr{B}$ is called a measure if it is countably additive and $\mu(\emptyset) = 0$. If $\mu(\mathcal{X}) = 1$, then we call the measure a probability measure.

4.2 Dirichlet Measure

We will denote the collection of all the probability distributions on $(\mathcal{X}, \mathscr{B})$ by $\mathcal{P}$. The Dirichlet Process was introduced by Ferguson, whose goal was to specify a distribution over $\mathcal{P}$ that was manageable, but useful [4]. He achieved this goal with just one restriction: Let α be a finite non-zero measure (just a measure, not necessarily a probability measure) on the original measurable space $(\mathcal{X}, \mathscr{B})$. Ferguson termed P a Dirichlet process with parameter α on $(\mathcal{X}, \mathscr{B})$ if for any finite measurable partition $\{B_i\}_{i=1}^{k}$ of $\mathcal{X}$, the random vector $(P(B_1), \ldots, P(B_k))$ has Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. (We call $\{B_i\}_{i=1}^{k}$ a finite measurable partition of $\mathcal{X}$ if $B_i \in \mathscr{B}$ for all $i = 1, \ldots, k$, $B_i \cap B_j = \emptyset$ if $i \neq j$ and $\cup_{i=1}^{k} B_i = \mathcal{X}$.) If P is a Dirichlet process with parameter α, then its distribution $\mathcal{D}_\alpha$ is called a Dirichlet measure.

As a consequence of Ferguson's restriction, $\mathcal{D}_\alpha$ has support only for atomic distributions with infinite atoms, and zero probability for any non-atomic distribution (e.g. Gaussians) and for atomic distributions with finite atoms [4]. That is, a realization drawn from $\mathcal{D}_\alpha$ is a measure:

$$P = \sum_{k=1}^{\infty} p_k \delta_{y_k}, \tag{15}$$

where δ_x is the Dirac measure on $\mathcal{X}$: $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise; $\{p_k\}$ is some sequence of weights; and $\{y_k\}$ is a sequence of points in $\mathcal{X}$. For the proof of this property, we refer the reader to [4].

4.3 In What Sense is the Dirichlet Process a Random Process?

A process is a collection of random variables indexed by some set, where all the random variables are defined over the same underlying set, and the collection of random variables has a joint distribution.

For example, consider a two-dimensional Gaussian process $\{X_t\}$ over time. That is, for each time $t \in \mathbb{R}$, there is a two-dimensional random vector $X_t \sim N_2(0, I)$. Note that $\{X_t\}$ is a collection of random variables whose index set is the real line $\mathbb{R}$, each random variable X_t is defined over the same underlying set, which is the plane $\mathbb{R}^2$. This collection of random variables has a joint Gaussian distribution. A realization of this Gaussian process would be some function $f : \mathbb{R} \to \mathbb{R}^2$.

Similarly, a Dirichlet process is a collection of random variables whose index set is the σ-algebra $\mathscr{B}$. Just as any time t for the Gaussian process example above corresponds to a Gaussian random variable X_t, any set $B \in \mathscr{B}$ has a corresponding random variable $\tilde{P}(B) \in [0, 1]$. You might think that because the marginals in a Gaussian process are random variables with Gaussian distribution, the marginal random variables $\{\tilde{P}(B)\}$ in a Dirichlet process will be random variables with Dirichlet distributions. However, this is not true, as things are more subtle. Instead, the random vector $[\tilde{P}(B) \quad 1 - \tilde{P}(B)]$ has a Dirichlet distribution with parameters $\alpha(B)$ and $\alpha(\mathcal{X}) - \alpha(B)$. Equivalently, a given set B and its complement set B^C form a partition of $\mathcal{X}$, and the random vector $[\tilde{P}(B) \quad \tilde{P}(B^C)]$ has a Dirichlet distribution with parameters $\alpha(B)$ and $\alpha(B^C)$.

Because of the form of the Dirichlet process, it holds that $\tilde{P}(B) = \sum_{k=1}^{\infty} \tilde{p}_k \delta_{\tilde{y}_k}(B)$, where in this context $\tilde{p}_k$ and $\tilde{y}_k$ denote random variables.

What is the *common underlying set* that the random variables are defined over? Here, it's the domain for the infinity of underlying random components $\tilde{p}_k$ and $\tilde{y}_k$, but because the $\{\tilde{p}_k\}$ are a function of underlying random components $\{\tilde{\theta}_i\} \in [0, 1]$, one says that the *common underlying set* is $([0, 1] \times \mathcal{X})^{\infty}$. We say that the collection of random variables $\{\tilde{P}(B)\}$ for $B \in \mathscr{B}$ has a joint distribution, whose existence and uniqueness is assured by Kolmogorov's existence theorem [8].

Lastly, a realization of the Dirichlet process is a probability measure $P : \mathscr{B} \to [0, 1]$.

5 Generating Samples from a Dirichlet Process

How do we generate samples from a Dirichlet process? In this section we describe stick-breaking, the Pólya urn process, and the Chinese restaurant process, of which the latter two are different names for the same process.

In all cases, we generate samples by generating the sequences $\{p_k\}$ and $\{y_k\}$, then using (15) (or equivalently (14)) to produce a sample measure P. The $\{y_k\}$ are simply drawn randomly from the normalized measure $\alpha/\alpha(\mathcal{X})$. The trouble is drawing the $\{p_k\}$, in part because they are not independent of each other since they must sum to one. Stick-breaking draws the p_k exactly, but one at a time: p_1, then p_2, etc. Since there are an infinite number of p_k, it takes an infinitely long time to generate a sample. If you stop stick-breaking early, then you have the first $k - 1$ coefficients exactly correct.

The Pólya Urn process also takes an infinitely long time to generate the $\{p_k\}$, but does so by iteratively refining an estimate of the weights. If you stop Pólya Urn process after $k - 1$ steps, you have an approximate estimate of up to $k - 1$ coefficients. We describe the Pólya urn and Chinese restaurant process first because they are easier.

5.1 Using the Pólya Urn or Chinese Restaurant Process to Generate a Dirichlet Process Sample

The Pólya sequence and the Chinese restaurant process (CRP) are two names for the same process due to Blackwell-MacQueen, which asymptotically produces a partition of the natural numbers [2]. We can then use this partition to produce the $\{p_k\}$ needed for (15).

The Pólya sequence is analogous to the Pólya urn described in Section 2.1 for generating samples from a Dirichlet distribution, except now there are an infinite number of ball colors and you start with an empty urn. To begin, let $n = 1$, and

Step 1: Pick a new color with probability distribution $\alpha/\alpha(\mathcal{X})$ from the set of infinite ball colors. Paint a new ball that color and add it to the urn.

	Indexing	**Partition Labels**
Pólya Urn	sequence of draws of balls	ball colors
Chinese Restaurant Process	sequence of incoming customers	different dishes (equivalently, different tables)
Clustering	sequence of the natural numbers	clusters

Table 3: **Equivalences between different descriptions of the same process.**

Step 2: With probability $\frac{n}{n+\alpha(\mathcal{X})}$, pick a ball out of the urn, put it back with another ball of the same color, and repeat Step 2. With probability $\frac{\alpha(\mathcal{X})}{n+\alpha(\mathcal{X})}$, go to Step 1.

That is, one draws a random sequence $(X_1, X_2, \ldots)$, where X_i is a random color from the set $\{y_1, y_2, \ldots, y_k, \ldots, y_\infty\}$. The elegance of the Pólya sequence is that we do not need to specify the set $\{y_k\}$ ahead of time.

Equivalent to the above two steps, we can say that the random sequence $(X_1, X_2, \ldots)$ has the following distribution:

$$X_1 \sim \frac{\alpha}{\alpha(\mathcal{X})}$$

$$X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha_n}{\alpha_n(\mathcal{X})}, \qquad \text{where} \qquad \alpha_n = \alpha + \sum_{i=1}^{n} \delta_{X_i}.$$

Since $\alpha_n(\mathcal{X}) = \alpha(\mathcal{X}) + n$, equivalently:

$$X_{n+1}|X_1, \ldots, X_n \sim \sum_{i=1}^{n} \frac{1}{\alpha(\mathcal{X}) + n} \delta_{X_i} + \frac{1}{\alpha(\mathcal{X}) + n} \alpha.$$

The balls of the kth color will produce the weight p_k, and we can equivalently write the distribution of $(X_1, X_2, \ldots)$ in terms of k. If the first n draws result in K different colors $y_1, \ldots, y_K$, and the kth color shows up m_k times, then

$$X_{n+1}|X_1, \ldots, X_n \sim \sum_{k=1}^{K} \frac{m_k}{\alpha(\mathcal{X}) + n} \delta_{y_k} + \frac{1}{\alpha(\mathcal{X}) + n} \alpha.$$

The Chinese restaurant interpretation of the above math is the following: Suppose each y_k represents a table with a different dish. The restaurant opens, and new customers start streaming in one-by-one. Each customer sits at a table. The first customer at a table orders a dish for that table. The nth customer chooses a new table with probability $\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}$ (and orders a dish), or chooses to join previous customers with probability $\frac{n}{\alpha(\mathcal{X})+n}$. If he chooses a new table, he orders a random dish distributed as $\alpha/\alpha(\mathcal{X})$. If the nth customer joins previous customers and there are already K tables, then he joins the table with dish y_k with probability $m_k/(n-1)$, where m_k is the number of customers already enjoying dish y_k.

Note that the more customers enjoying a dish, the more likely a new customer will join them. We summarize the different interpretations in Table 3.

After step N, the output of the CRP is a partition of N customers across K tables, or equivalently a partition of N balls into K colors, or equivalently a partition of the natural numbers $1, 2, \ldots, N$ into K sets.

For the rest of this paragraph, we will stick with the restaurant analogy. To find the expected number of tables, let us introduce the random variables Y_i, where Y_i is 1 if the i^{th} customer has chooses a new table and 0 otherwise. If $T_N = \sum_{i=1}^{N} Y_i$ then T_N is the number of tables occupied by the first N customers. The expectation of T_N is (see [1])

$$\mathbb{E}[T_N] = \mathbb{E}\left[\sum_{i=1}^{N} Y_i\right] = \sum_{i=1}^{N} \mathbb{E}[Y_i] = \sum_{i=1}^{N} \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + i - 1}.$$

If we let $N \to \infty$, then this infinite partition can be used to produce a realization p_k to produce a sample from the Dirichlet process. We do not need the actual partition for this, only the sizes $m_1, m_2, \ldots$. Note that each m_k is a function of N, which we denote by $m_k(N)$. Then,

$$p_k = \lim_{N\to\infty} \frac{m_k(N)}{\alpha(\mathcal{X}) + N}. \tag{16}$$

Blackwell and MacQueen proved that $\frac{\alpha_n}{\alpha_n(\mathcal{X})}$ converges to a discrete probability measure whose distribution is a Dirichlet measure with parameter α [2] .

5.1.1 Using Stick-breaking to Generate a Dirichlet Process Sample

Readers who are not familiar with measure theory can safely ignore the more advanced mathematics of this section and still learn how to generate a sample using stick-breaking.

We know that the Dirichlet realizations are characterized by the atomic distributions of the form (15). So we will characterize the distribution of the $\{\tilde{p}_k, \tilde{y}_k\}$, which will enable us to generate samples from D_α. However, it is difficult to characterize the distribution of the $\{\tilde{p}_k, \tilde{y}_k\}$, and instead we characterize the distribution of a different set $\{\tilde{\theta}_k, \tilde{y}_k\}$, where $\{\tilde{\theta}_k\}$ will allow us to generate $\{\tilde{p}_k\}$.

Consider the countably-infinite random sequence $((\tilde{\theta}_k, \tilde{y}_k))_{k=1}^{\infty}$ that takes values in $([0,1] \times \mathcal{X})^{\infty}$. All the $\{\tilde{\theta}_k\}_{k=1}^{\infty}$ and all the $\{\tilde{y}_k\}_{k=1}^{\infty}$ are independent, each $\tilde{\theta}_k$ has beta distribution $B(1, \alpha(\mathcal{X}))$ and each $\tilde{y}_k$ is distributed according to β, where $\beta = \alpha/\alpha(\mathcal{X})$.

Kolmogorov's existence theorem says that there is a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where $\mathcal{A}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability measure on $\mathcal{A}$, such that the random sequence $((\tilde{\theta}_k, \tilde{y}_k))_{k=1}^{\infty}$ on $(\Omega, \mathcal{A}, \mathbb{P})$ has the joint distribution D_α. This is wonderful because as soon as we connect the $\{\theta_k\}$ to the $\{p_k\}$, we will be able to generate samples from D_α.

The random sequence $((\tilde{\theta}_k, \tilde{y}_k))_{k=1}^{\infty}$ is a map from $([0,1] \times \mathcal{X})^{\infty}$ into $\mathcal{P}$. Draw a realization $((\theta_k, y_k))_{k=1}^{\infty}$ as described above. Then to form a corresponding probability distribution in $\mathcal{P}$ from this sequence we can use the stick-breaking method as follows. Let $p_1 = \theta_1$; that is, break off a θ_1 portion of a unit-long stick. What remains has length $1 - \theta_1$. Break off a θ_2 fraction of the remaining stick; that is, let $p_2 = \theta_2(1 - \theta_1)$. What is left after this step is a $(1-\theta_1)(1-\theta_2)$ long stick. In the kth step, we have a stick of length $\prod_{i=1}^{k-1}(1-\theta_i)$ remaining, and to produce p_k, we break off a θ_k portion of it, so $p_k = \theta_k \prod_{i=1}^{k-1}(1-\theta_i)$. The result is a sequence $(p_i)_{i=1}^{k}$ and we can use this sequence directly to produce a probability distribution P in $\mathcal{P}$:

$$P(((\theta_k, y_k))_{k=1}^{\infty}) = \sum_{k=1}^{\infty} p_k \delta_{y_k}.$$

We can think of P as a stochastic process on $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in $\mathcal{P}$. The index set of this random process is the set $\mathscr{B}$.

5.2 Conditional Distribution of a Dirichlet Process

Let P be a random probability measure on $(\mathcal{X}, \mathscr{B})$. We say that $X_1, \ldots, X_n$ is a sample of size n from P if for any $m = 1, 2, \ldots$ and measurable sets $A_1, \ldots, A_m, C_1, \ldots C_n$, the probability of $X_1 \in C_1, \ldots, X_n \in C_n$

given $P(A_1), \dots, P(A_m), P(C_1), \dots, P(C_n)$ is (see [4])

$$\mathbb{P}\left(X_1 \in C_1, \dots, X_n \in C_n \;\middle|\; P(A_1), \dots, P(A_m), P(C_1), \dots, P(C_n)\right) = \prod_{i=1}^{n} P(C_i) \text{ a.s.}$$

In other words, $X_1, \dots, X_n$ is a sample of size n from P if the events $\{X_1 \in C_1\}, \dots, \{X_n \in C_n\}$ are independent of the rest of the process and they are independent among themselves.

Let P be a Dirichlet process on $(\mathcal{X}, \mathscr{B})$. Then the following theorem proved in [4] gives us the conditional distribution of P given a sample of size n:

Theorem 5.2.1. *Let P be a Dirichlet process on $(\mathcal{X}, \mathscr{B})$ with parameter α, and let $X_1, \dots, X_n$ be a sample of size n from P. Then the conditional distribution of P given $X_1, \dots, X_n$ is a Dirichlet process with paramater $\alpha + \sum_{i=1}^{n} \delta_{X_i}$.*

5.3 Estimating the Dirichlet Process Given Data

The description of a draw from a Dirichlet process requires an infinite sum. This is not practical, so we need to consider ways to approximate the draw by finite sums. There are at least two ways to do this.

The most natural one is the truncation method. Suppose we want to generate the draw P by using the stick-breaking method. But, instead of generating the whole infinite sum, we may stop after finding the first N terms of it. The resulting measure won't necessarily be a probability measure, so to ensure that we use probability measures to approximate P we need to set the last draw θ_N to 1 instead of drawing it from $Beta(1, \alpha(\mathcal{X}))$. If we call the resulting probability measure P_N, then $P_N(g) \to P(g)$ almost surely as $N \to \infty$ for any bounded continuous function $g : \mathcal{X} \to \mathbb{R}$, where $P_N(g) = \int g dP_N$and $P(g) = \int g dP$ [7].

Another way to approximate the draw P is more surprising. It is called the *finite-dimensional Dirichlet prior* and it is constructed in the following way. Let $(q_1, \dots, q_N) \sim Dir\left(\alpha(\mathcal{X})/N, \dots, \alpha(\mathcal{X})/N\right)$ and draw Z_i, $i = 1, \dots, N$ from $\beta = \frac{\alpha}{\alpha(\mathcal{X})}$. Then the finite-dimensional Dirichlet prior P_N is simply

$$P_N = \sum_{i=1}^{N} q_i \delta_{Z_i}.$$

This prior converges to P in a different way as the previous approximation. For any measurable $g : \mathcal{X} \to \mathbb{R}$ that is integrable w.r.t. β we have that $P_N(g) \to P(g)$ in distribution.

5.4 Hierarchical Dirichlet Process (the Chinese Restaurant Franchise Interpretation)

To explain the *hierarchical Dirichlet process* (see [12]) consider the following scenario. Suppose we have several Dirichlet processes that are independent but controlled by the same parameter α. Then in general, we cannot relate the samples drawn from one process to samples drawn from another process. For example, suppose $\mathcal{X}$ is the interval $[0, 1]$ and α is the Lebesgue measure and we have two Chinese restaurant processes on $[0, 1]$. The probability of seeing the same sample in both processes is 0. With the hierarchical Dirichlet process one does not assume a common parameter α, but instead draws a probability measure G_0 from a Dirichlet process with parameter α. We would like to use G_0 as a parameter for a new Dirichlet process, but the parameter of a Dirichlet process is not necessarily a probability measure. We need an additional parameter $\alpha_0 > 0$, and then we can use $\alpha_0 G_0$ as the parameter for a new independent Dirichlet process. Since G_0 is itself a realization of a Dirichlet process, we know its form:

$$G_0 = \sum_{i=1}^{\infty} q_i \delta_{X_i}.$$

Now, if we have several independent Dirichlet processes sharing the same parameter $\alpha_0 G_0$ then the corresponding Chinese restaurant processes can be generated separately and in the same fashion as was described in 5.1.

The Chinese restaurant franchise interpretation of the hierarchical Dirichlet process is the following: Suppose we have a central menu (with dishes specified by the δ darts of G_0) and at each restaurant each table is associated with a dish from this menu. A guest, by choosing a table, chooses a menu item. The popularity of different dishes can differ from restaurant to restaurant, but the probability that two restaurants will offer the same dish (two processes share a sample) is non-zero in this case.

6 Random Binary Matrices and the Indian Buffet Process

In this section, we discuss random binary matrices and random binary measures over matrices.

The *Indian buffet process* (IBP) [5] is a method to generate binary matrices that have a fixed number of rows N, and a variable number of non-trivial columns K. It can be useful if one needs to generate or model the generation of binary matrices with certain constraints (details below). It can also be motivated as producing, after a correction, samples of a *matrix beta distribution*. Why is it called the IBP? How is it related to the CRP? As we describe in this section, the IBP and CRP have in common that: *(i)* the two are often interpreted in terms of customers and dishes; *(ii)* stick-breaking can be used to generate samples from both of them; and *(iii)* they can be used to generate samples of partitions. First, we discuss random binary matrices, then describe the IBP, then review the stick-breaking variant of the IBP.

6.1 He'll Have the Curry Too and an Order of Naan

To begin, let us first construct a *Bernoulli matrix distribution* over binary matrices of fixed size $N \times K$. The N rows can be thought to correspond to objects (or later, customers), and the K columns to features (later, dishes). Each of the N objects can possess any of the K features. Each object has the same probability to possess a given feature: the probability that the ith object possesses the kth feature is π_k. We can use a binary matrix to describe which object possesses which feature. Let Z be a random binary matrix of size $N \times K$ whose entry Z_{ik} is 1 with probability π_k. That is, $Z_{ik} \sim Ber(\pi_k)$, where $Ber(\pi_k)$ is the Bernoulli distribution with parameter π_k, and each Z_{ik} is independent of the other matrix entries. We now have a distribution over binary matrices parameterized by the natural number N and the vector $\pi \in [0,1]^K$.

Next, consider a distribution over the vector parameter π. Let $\alpha > 0$ be a scalar (eventually, this α will be a parameter of the IBP but we are not ready to discuss the IBP yet), and let π be random such that $\pi_k \sim Beta(\alpha/K, 1)$. This *beta-Bernoulli matrix distribution* can be used as a way to generate Bernoulli matrix distributions (and hence, we can say it is a distribution over random matrices), or it can be used in a compound-fashion to generate a sample binary matrix by first drawing a π, and then drawing a binary matrix from the above Bernoulli matrix distribution with parameter π. The probability of seeing a particular $N \times K$ binary matrix z from this *compound beta-Bernoulli matrix distribution* is

$$P(Z = z) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma\left(m_k + \frac{\alpha}{K}\right)\Gamma(N - m_k + 1)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}, \text{ where } m_k = \sum_{i=1}^{N} z_{ik}, \tag{17}$$

that is, the probability of a particular random matrix z only depends on m_k, which is the number of appearances of the kth feature summed over the N rows. Note, we are not yet at the IBP.

Next, note from (17) that the probability of a particular matrix z is the same as the probability of any another matrix z' that was formed by permuting columns of z. Let $z \sim z'$ denote that z' is such a column-permutation of z. Column-permutation is an *equivalence relation* on the set of $N \times K$ binary matrices. Let $[z]$ denote the *equivalence class of matrix* z such that $[z]$ is the set of all matrices that can be formed by permuting columns of z. At this point it is useful to consider the customer-dish analogy for a binary matrix. Let the N rows correspond to N customers, and the K columns correspond to K different dishes. So a column denotes which customer ordered that particular dish. Now, and this is key - in this analogy when we permute the columns the dish label goes with it. The ordering of the K columns is considered to be the *order in which you choose the dishes*. For example, consider a z and a column permutation of it z':

$$z = \left(\begin{array}{cccc} \text{curry} & \text{naan} & \text{aloo} & \text{dahl} \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right) \quad \text{and its permutation } z' = \left(\begin{array}{cccc} \text{naan} & \text{curry} & \text{aloo} & \text{dahl} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right).$$

The matrix z can be interpreted (reading left-to-right and top-to-bottom) as, "The first customer ordered *naan*, the second customer ordered *curry* and then ordered *aloo*, the third customer ordered *curry* and then *naan* and then *aloo* and then *dahl*." The matrix z' swaps the order of the *curry* and *naan* columns, and would be interpreted (reading left-to-right and top-to-bottom) as, "The first customer ordered *naan*, the second customer ordered *curry* and then *aloo*, and the third customer ordered *naan* and then *curry* and then *aloo* and then *dahl*. So the only difference in interpretation of the orders is the ordering of the dishes. With this interpretation, a column permutation does not change *what gets ordered by whom*. So if that is all we care about, then rather than consider the probability of each matrix z, we want to deal with the probability of seeing any matrix from the equivalence class $[z]$, or more directly, deal with the probabilities of seeing each of the possible equivalence classes.

How many matrices are in the equivalence class $[z]$? To answer that important question, let us assign a number to each of the 2^N different binary $\{0,1\}^N$ vectors that could comprise a column. Form the one-to-one correspondence (bijection) h that maps the kth column onto the set of numbers $\{0,\ldots,2^N-1\}$:

$$h(z_k) = \sum_{i=1}^{N} z_{ik} 2^{N-i}.$$

Since h is one-to-one, we can now describe a particular possible column using the inverse mapping of h, for example $h^{-1}(0)$ corresponds to the column $[0\,0\,0\,\ldots 0]^T$. Now suppose a matrix z has K_0 columns of $h^{-1}(0)$, and K_1 number of columns of $h^{-1}(1)$, and so on up to K_{2^N-1} columns of $h^{-1}(2^N-1)$. Then the cardinality of $[z]$ is the ways to arrange those columns:

$$\text{cardinality of } [z] = \frac{K!}{\prod_{b=0}^{2^N-1} K_b!}.$$

Thus the probability of seeing any matrix in the equivalence class of z is

$$\begin{aligned} P(Z \in [z]) &= \frac{K!}{\prod_{b=0}^{2^N-1} K_b!} P(Z = z) \\ &= \frac{K!}{\prod_{b=0}^{2^N-1} K_b!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma\left(m_k + \frac{\alpha}{K}\right)\Gamma(N - m_k + 1)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}. \end{aligned} \tag{18}$$

Equation (18) is a distribution over the equivalence classes (that is, over the column-permutation-equivalent *partitions* of the set of binary matrices), and its left-hand side could equally be written $P([Z] = [z])$.

What happens if there are infinite dishes possible (as is said to be the case at Indian buffets in London)? If we keep the K_b fixed and take the limit $K \to \infty$ in (18), we get that:

$$P_{K\to\infty}(Z \in [z]) = \frac{\alpha^{K_+}}{\prod_{b=1}^{2^N-1} K_b!} e^{-\alpha H_N} \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}, \tag{19}$$

where $K_+ = \sum_{b=1}^{2^N-1} K_b$ is the number of columns of Z that are not completely zero (that is K_+ is the number of dishes that at least one person orders) and $H_N = \sum_{i=1}^{N} \frac{1}{i}$ is the Nth harmonic number. This is an interesting case because we do not have to decide ahead of time how many dishes K_+ are going to be ordered, and yet this distribution assigns positive probabilities to $K_+ = 1, 2, \ldots$.... One begins (correctly) to feel that there is something very Poisson going on here...

6.2 Generating Binary Matrices with the IBP

Finally we are ready to tell you about the IBP. The IBP is a method to generate binary matrices using Poisson random variables, and as we describe shortly, creates samples of equivalence classes that are (after a correction) as if drawn from the beta-Bernoulli matrix partition distribution given by (19).

IBP: Let there be N customers and infinitely many different dishes (a banquet befitting Ganesha himself).

Step 1: The first customer chooses $K^{(1)}$ different dishes, where $K^{(1)}$ is distributed according to a Poisson distribution with parameter α.

Step 2: The second customer arrives and chooses to enjoy each of the dishes already chosen for the table with probability 1/2. In addition, the second customer chooses $K^{(2)} \sim \mathit{Poisson}(\alpha/2)$ new dishes.

Steps 3 through N: The ith customer arrives and chooses to enjoy each of the dishes already chosen for the table with probability m_{ki}/i, where m_{ki} is the number of customers who have chosen the kth dish before the ith customer. In addition, the ith customer chooses $K^{(i)} \sim \mathit{Poisson}(\alpha/i)$ new dishes.

After the N steps one has a $N \times K_+$ binary matrix z that describes the customers' choices, where K_+ can be calculated as $K_+ = \sum_{i=1}^{N} K^{(i)}$. Note there are many binary matrices that cannot be generated by the IBP, such as the example matrix z given above, because the first customer cannot choose to have the second dish (*naan*) but not the first dish (*curry*). Let Υ be the set of all matrices that can be generated by the IBP. The probability of generating a matrix $z \in \Upsilon$ by the IBP is

$$P(Z = z) = \frac{\alpha^{K_+}}{\prod_{i=1}^{N} K^{(i)}!} e^{-\alpha H_N} \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!} \quad \text{for } z \in \Upsilon, \text{ and zero otherwise.} \tag{20}$$

Note the IBP distribution given in (20) is not the beta-Bernoulli matrix distribution we described in the last section, for one the IBP assigns zero probability to many sample matrices would be generated by the beta-Bernoulli. However, the two are related in terms of the distribution of the column-permutation equivalence classes of the binary matrices. For each equivalence class, there is at least one matrix in Υ, but for some equivalence classes there is more than one such matrix in Υ. This unequal sampling by the IBP of the different equivalence classes has to be corrected for if we want to use the IBP realizations as samples from the same probability distribution over the equivalence classes given (19). This correction factor is the relative cardinality of the actual column-permutation equivalence class $[z]$ compared to the cardinality of that equivalence class in Υ. This ratio is

$$\upsilon([z]) = \frac{\frac{K_+!}{\prod_{b=1}^{2^N-1} K_b!}}{\frac{K_+!}{\prod_{i=1}^{N} K^{(i)}!}} = \frac{\prod_{i=1}^{N} K^{(i)}!}{\prod_{b=1}^{2^N-1} K_b!}. \tag{21}$$

The goal here is to create a distribution that is close to one given by (19), because we would like to use it as a prior. To do this we can use the IBP: Generate M matrices $\{z^{(m)}\}_{m=1}^{M}$ using the IBP and calculate the corresponding histogram. The distribution we created this way is not the one we are after. To find the desired distribution, that is to find the probability $P([z])$ for any particular z we need to identify an m such that $[z^{(m)}] = [z]$ and multiply the value we get from the histogram belonging to $z^{(m)}$ by $\upsilon([z])$.

There are two problems with this in practice. First, it is combinatorically-challenging to determine if two matrices are from the same equivalence class, so counting how many of which equivalence classes you have seen given the IBP-generated sample matrices is difficult. Second, you end up with a histogram over the equivalence classes with some number of samples that you do not get to choose (you started with M sample matrices, but you do not end up with M samples of the equivalence classes).

6.3 A Stick-breaking Approach to IBP

Just as there are multiple ways to generate samples from a Dirichlet process (i.e. urn versus stick-breaking), Teh et al. showed that IBP sample matrices could instead by generated by a stick-breaking approach [11]. In the original paper ,the authors integrated out the parameters π_k by using the Beta priors [5]. The stick-breaking approach instead generates each π_k probability that any customer will want the kth dish, and then after generating all the π_k's, draws N iid samples from a Bernoulli corresponding to each π_k to fill in the entries of the binary matrix. The only difficulty is that we assumed the number of features $K \to \infty$, so to use stick-breaking to get exact samples it will take a long time. In practice, we simply stop after some choice of K. As we describe below, this is not so bad an approximation, because more popular dishes tend to show up earlier in the process.

As before, let $\pi_k \sim Beta(\alpha/K, 1)$ and $z_{ik} \sim Ber(\pi_k)$. The density is

$$p_k(\pi_k) = \frac{\alpha}{K}\pi_k^{\frac{\alpha}{K}-1}$$

and the corresponding cumulative distribution function (cdf) is

$$F_k(\pi_k) = \int_{-\infty}^{\pi_k} \frac{\alpha}{K} s^{\frac{\alpha}{K}-1}\mathbb{I}(0 \leq s \leq 1)ds = \pi_k^{\frac{\alpha}{K}}\mathbb{I}(0 \leq \pi_k \leq 1) + \mathbb{I}(1 \leq \pi_k), \tag{22}$$

where $\mathbb{I}(A)$ is the indicator (or characteristic) function of the set A, and the second indicator in the above just confirms the cdf does not grow past $\pi_k = 1$.

Rather than directly generating $\{\pi_k\}$, we will generate the order statistics: let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{K-1} \geq \mu_K$ be the sorted $\{\pi_k\}$ such that $\mu_1 = \max_k \pi_k$ and $\mu_K = \min_k \pi_k$. Warning: we will abuse notation slightly and use μ_k to mean either a random variable or its realization, but each usage should be clear from context and will consider k to be a generic variable.

Since the random variables $\{\pi_1, \ldots, \pi_K\}$ are independent, the cdf of μ_1 is the product of the cdfs of all the π_ks:

$$F(\mu_1) = \prod_{k=1}^{K} F_k(\mu_1) = \left(\mu_1^{\frac{\alpha}{K}}\mathbb{I}(0 \leq \mu_1 \leq 1) + \mathbb{I}(1 \leq \mu_1)\right)^K = \mu_1^{\alpha}\mathbb{I}(0 \leq \mu_1 \leq 1) + \mathbb{I}(1 \leq \mu_1). \tag{23}$$

Since $F(\mu_1)$ is differentiable almost everywhere, the corresponding density exists and can be expressed

$$p(\mu_1) = \alpha\mu_1^{\alpha-1} \tag{24}$$

From (23) and (24), one sees that $\mu_1 \sim Beta(\alpha, 1)$. At this point you should be hearing the faint sounds of stick-breaking in the distance.

Next, say we already had the distributions for $\{\mu_1, \ldots, \mu_m\} \triangleq \mu_{(1:m)}$, and wanted to specify the cdf for μ_{m+1}. The probability of $P(\mu_{m+1} < x|\mu_{(1:m)})$ is 1 if $x \geq \mu_m$ and $\prod_{l \in L} C_l P(\pi_l < x)$ where L is a set of $K - m$ indices and C_l is the normalization constant from the definition of the conditional probability. Since we are working with order statistics we can be sure that for m indices (for those, that belong to $\mu_{(1:m)}$) the condition $\pi_l < x$ won't be satisfied. Since all the π_ks are independent and identically distributed we don't need to worry about the concrete indices only about how many of them are there.

The constant C_l^{-1} is the same for all $l \in L$:

$$\int_0^{\mu_m} \frac{\alpha}{K} s^{\frac{\alpha}{K}-1} ds = \mu_m^{\frac{\alpha}{K}},$$

therefore

$$\prod_{l \in L} C_l P(\pi_l < x) = C_l^{K-m} F_k(x)^{K-m}$$

and

$$F(x|\mu_{(1:m)}) = P(\mu_{m+1} < x|\mu_{(1:m)}) = \mu_m^{-\frac{\alpha(K-m)}{K}} x^{\frac{\alpha(K-m)}{K}} \mathbb{1}(0 \leq x \leq \mu_m) + \mathbb{1}(\mu_m \leq x). \tag{25}$$

Switching to using μ_{m+1} to denote a realization, and substituting (22) into (25), we can write (25) as

$$F(\mu_{m+1}|\mu_{(1:m)}) = \mu_m^{-\frac{\alpha(K-m)}{K}} \mu_{m+1}^{\frac{\alpha(K-m)}{K}} \mathbb{I}(0 \leq \mu_{m+1} \leq \mu_m) + \mathbb{I}(\mu_m \leq \mu_{m+1}). \tag{26}$$

In the limit $K \to \infty$,

$$F_{K\to\infty}(\mu_{m+1}|\mu_{(1:m)}) = \mu_m^{-\alpha} \mu_{m+1}^{\alpha} \mathbb{I}(0 \leq \mu_{m+1} \leq \mu_m) + \mathbb{I}(\mu_m \leq \mu_{m+1}).$$

with corresponding density

$$p(\mu_{m+1}|\mu_{(1:m)}) = \alpha\mu_m^{-\alpha} \mu_{m+1}^{\alpha-1} \mathbb{I}(0 \leq \mu_{m+1} \leq \mu_m)$$

Next, define a new set of random variables $\{\nu_k\}$, where $\nu_1 = \mu_1 \sim Beta(\alpha, 1)$ and $\nu_{k+1} = \frac{\mu_{k+1}}{\mu_k}$. Then $d\nu_{k+1} = \frac{1}{\mu_k} d\mu_{k+1}$ and we can derive their density as follows:

$$\begin{aligned}
\alpha\mu_k^{-\alpha}\mu_{k+1}^{\alpha-1}\mathbb{I}(0 \leq \mu_{k+1} \leq \mu_k)d\mu_{k+1} &= \alpha\mu_k^{-\alpha+1}\mu_{k+1}^{\alpha-1}\mathbb{I}(0 \leq \frac{\mu_{k+1}}{\mu_k} \leq 1)\frac{1}{\mu_k}d\mu_{k+1} \\
&= \alpha\nu_{k+1}^{\alpha-1}\mathbb{I}(0 \leq \nu_{k+1} \leq 1)d\nu_{k+1} \\
&= dF(\nu_{k+1}).
\end{aligned}$$

Thus $\nu_k \sim Beta(\alpha, 1)$ for all k. We can generate the ν_k and $\mu_1 \sim Beta(\alpha, 1)$ using Beta random variables, and then compute $\mu_{k+1} = \nu_{k+1}\mu_k = \prod_{i=1}^{k+1} \nu_i$. This generation mechanism can be interpreted as stick-breaking. Take a stick of length 1 and break off ν_1 portion of it and keep it. Next, break off a ν_2 portion of the piece you keep and throw away the rest. Etc. Since we are only interested in sampling the column-permutation equivalence classes, we can use the $\{\mu_k\}$ in order for each column, and not worry about the original π_k.

Acknowledgements

This work was supported by the United States Office of Naval Research and the United States PECASE award. We thank Evan Hanusa and Kristi Tsukida for helpful suggestions and proofreading.

References

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

[2] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1:353–355, 1973.

[3] Y. Chen and M.R. Gupta. Theory and use of the EM algorithm. *Foundations and Trends in Machine Learning*, 2011.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, March 1973.

[5] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:1–25, 2006.

[6] M. R. Gupta. A measure theory tutorial (measure theory for dummies). Technical report, Technical Report of the Dept. of Electrical Engineering, University of Washington, 2006.

[7] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 30(2):269–283, 2002.

[8] O. Kallenberg. *Foundations of Modern Probability*, pages 115–6. Springer-Verlag, second edition, 2001.

[9] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proc. Intl. Conf. Machine Learning*, 2005.

[10] T. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, Cambridge, 2003.

[11] Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 11, 2007.

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