A Simple Proof of the Stick-Breaking Construction of the Dirichlet Process

John Paisley

Department of Computer Science
Princeton University, Princeton, NJ

jpaisley@princeton.edu

Abstract

We give a simple proof of Sethuraman's construction of the Dirichlet distribution and discuss its extension to infinite-dimensional spaces.

1 Introduction

The K-dimensional Dirichlet distribution is a distribution on vectors, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, where $0 \leq \pi_k \leq 1$ for all k, and $\sum_{k=1}^{K} \pi_k = 1$. Vectors of this form are said to reside in the simplex of $\mathbb{R}^K$, denoted Δ_K. Parameters for the Dirichlet distribution are $\alpha > 0$ and $g_0 \in \Delta_K$. The density function of a Dirichlet distribution is

$$p(\boldsymbol{\pi}|\alpha g_0) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha g_{0k})} \prod_{k=1}^{K} \pi_k^{\alpha g_{0k} - 1} \tag{1}$$

and a vector having this distribution is denoted $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$. There are several methods for sampling the vector $\boldsymbol{\pi}$, for example, using gamma-distributed random variables [6]. A finite stick-breaking approach also exists [2], where for $k = 1, \dots, K-1$,

$$\begin{aligned} V_k &\sim \text{Beta}\left(\alpha g_{0k}, \alpha \sum\nolimits_{\ell=k+1}^{K} g_{0\ell}\right) \\ \pi_k &= V_k \prod_{\ell=1}^{k-1} (1 - V_\ell) \\ \pi_K &= 1 - \sum_{k=1}^{K-1} \pi_k \end{aligned} \tag{2}$$

The resulting vector $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$.

Sethuraman's constructive definition [7] is a third method that requires the sampling of an infinite collection of random variables, from which $\boldsymbol{\pi}$ is constructed piece-by-piece. Though also called a stick-breaking construction, this method is distinct from (2).

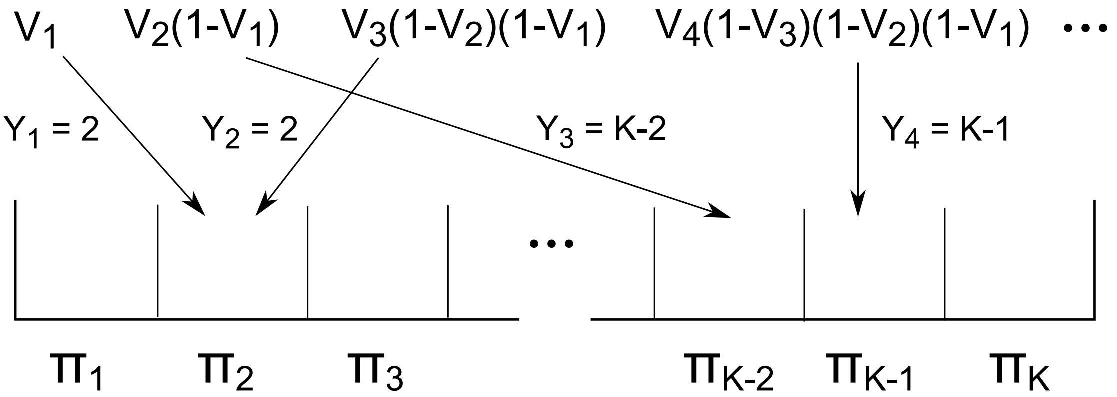

Figure 1: An illustration of the infinite stick-breaking construction of a K-dimensional Dirichlet distribution. Weights are drawn according to a Beta(1,α) stick-breaking process, with corresponding locations taking value k with probability g_{0k}.

2 Constructing the Finite-Dimensional Dirichlet Distribution

The constructive definition of a Dirichlet prior [7] states that, if $\boldsymbol{\pi}$ is constructed according to the following function of random variables, then $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha g_0)$.

$$\begin{aligned}
\boldsymbol{\pi} &= \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\mathbf{e}_{Y_i} \\
V_i &\overset{iid}{\sim} \mathrm{Beta}(1,\alpha) \\
Y_i &\overset{iid}{\sim} \mathrm{Mult}(\{1,\dots,K\}, g_0)
\end{aligned} \tag{3}$$

Using this multinomial parametrization, $Y \in \{1,\dots,K\}$ with $\mathbb{P}(Y = k|g_0) = g_{0k}$. The vector $\mathbf{e}_Y$ is a K-dimensional vector of zeros, except for a one in position Y. The values $V_i \prod_{j=1}^{i-1}(1-V_j)$ are often called "stick-breaking" weights because, at step i, the proportion V_i is "broken" from the remainder, $\prod_{j=1}^{i-1}(1-V_j)$, of a unit-length stick. Since $V \in [0,1]$, the product $V_i \prod_{j=1}^{i-1}(1-V_j) \in [0,1]$ for all i, and it can be shown that $\sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j) = 1$.

In Figure 1, we illustrate this process for $i = 1,\dots,4$. The weights are broken as mentioned, and the random variables $\{Y_i\}_{i=1}^{4}$ indicate the elements of the vector $\boldsymbol{\pi}$ to which each weight is added. In the limit, the value of the k^{th} element is

$$\pi_k = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\mathbb{I}(Y_i = k)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

2.1 Proof of the Construction

We begin with the random vector

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha g_0) \tag{4}$$

The proof that this random vector has the same distribution as the random vector in (3) requires two lemmas concerning general properties of the Dirichlet distribution. These lemmas are used to decompose (4) into hierarchical processes that are equal in distribution to (4).

Lemma 1: Let $Z \sim \sum_{k=1}^{K} g_{0k} \mathrm{Dir}(\alpha g_0 + \mathbf{e}_k)$. Values of Z can be sampled from this distribution by first sampling $Y \sim \mathrm{Mult}(\{1, \dots, K\}, g_0)$, and then sampling $Z \sim \mathrm{Dir}(\alpha g_0 + \mathbf{e}_Y)$. It then follows that $Z \sim \mathrm{Dir}(\alpha g_0)$.

The proof of this lemma is in the appendix. Therefore, the hierarchical process

$$\begin{array}{rcl} \boldsymbol{\pi} & \sim & \mathrm{Dir}(\alpha g_0 + \mathbf{e}_Y) \\ Y & \sim & \mathrm{Mult}(\{1, \dots, K\}, g_0) \end{array} \tag{5}$$

produces a random vector $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha g_0)$. A second lemma is next applied to this result.

Lemma 2: Let the random vectors $W_1 \sim \mathrm{Dir}(w_1, \dots, w_K)$, $W_2 \sim \mathrm{Dir}(v_1, \dots, v_K)$ and $V \sim \mathrm{Beta}(\sum_{k=1}^{K} w_k, \sum_{k=1}^{K} v_k)$. Define the linear combination,

$$Z := VW_1 + (1 - V)W_2 \tag{6}$$

then $Z \sim \mathrm{Dir}(w_1 + v_1, \dots, w_K + v_K)$.

The proof of this lemma is in the appendix. In words, this lemma states that, if one wished to construct the vector $Z \in \Delta_K$ according to the function of random variables (W_1, W_2, V) given in (6), one could equivalently bypass this construction and directly sample $Z \sim \mathrm{Dir}(w_1 + v_1, \dots, w_K + v_K)$.

This lemma is applied to the random vector $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha g_0 + \mathbf{e}_Y)$ in (5), with the result that this vector can be represented by the following process,

$$\begin{array}{rcl} \boldsymbol{\pi} & = & VW + (1 - V)\boldsymbol{\pi}' \\ W & \sim & \mathrm{Dir}(\mathbf{e}_Y) \\ \boldsymbol{\pi}' & \sim & \mathrm{Dir}(\alpha g_0) \\ V & \sim & \mathrm{Beta}\left(\sum_{k=1}^{K} \mathbf{e}_{Y,k}, \sum_{k=1}^{K} \alpha g_{0k}\right) \\ Y & \sim & \mathrm{Mult}(\{1, \dots, K\}, g_0) \end{array} \tag{7}$$

The result is still a random vector $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$. Also, $\sum_{k=1}^{K} \mathbf{e}_{Y,k} = 1$ and $\sum_{k=1}^{K} \alpha g_{0k} = \alpha$. We observe that the distribution of W is degenerate, with only one of the K parameters in the Dirichlet distribution being nonzero. Therefore, since $\mathbb{P}(\pi_k = 0 | g_{0k} = 0) = 1$, we can say that $\mathbb{P}(W = \mathbf{e}_Y | g_0 = \mathbf{e}_Y) = 1$. Modifying (7), the following generative process for $\boldsymbol{\pi}$ produces a random vector $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$.

$$\begin{aligned}
\boldsymbol{\pi} &= V\mathbf{e}_Y + (1-V)\boldsymbol{\pi}' \\
\boldsymbol{\pi}' &\sim \text{Dir}(\alpha g_0) \\
V &\sim \text{Beta}\,(1, \alpha) \\
Y &\sim \text{Mult}(\{1, \dots, K\}, g_0) \qquad (8)
\end{aligned}$$

We observe that the random vectors $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ have the same distribution.

Since $\boldsymbol{\pi}' \sim \text{Dir}(\alpha g_0)$ in (8), this vector can be decomposed according to the same process by which $\boldsymbol{\pi}$ is decomposed in (8). Thus, for $i = 1, 2$ we have

$$\begin{aligned}
\boldsymbol{\pi} &= V_1\mathbf{e}_{Y_1} + V_2(1-V_1)\mathbf{e}_{Y_2} + (1-V_1)(1-V_2)\boldsymbol{\pi}'' \\
V_i &\overset{iid}{\sim} \text{Beta}(1, \alpha) \\
Y_i &\overset{iid}{\sim} \text{Mult}(\{1, \dots, K\}, g_0) \\
\boldsymbol{\pi}'' &\sim \text{Dir}(\alpha g_0) \qquad (9)
\end{aligned}$$

which can proceed letting $i \to \infty$. Each decomposition produces the vector $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$. In the limit as $i \to \infty$, (9)→(3), since $\lim_{i\to\infty} \prod_{j=1}^{i}(1 - V_j) = 0$, concluding the proof. Therefore, (3) arises as an infinite number of decompositions of the Dirichlet distribution, each taking the form of (8).

2.2 The Extension to Infinite-Dimensional Spaces

In this section we briefly make the connection between the stick-breaking representation of the finite Dirichlet distribution and the infinite Dirichlet process. For finite-dimensional vectors, $\boldsymbol{\pi} \in \Delta_K$, the constructive definition of (3) may seem unnecessary, since the infinite sum cannot be carried out in practice, and $\boldsymbol{\pi}$ can be constructed exactly using only K gamma-distributed random variables. The primary use of the stick-breaking representation of the Dirichlet distribution is the case where $K \to \infty$.

For example, consider a K-component mixture model [5], where observations in a data set, $\{X_n\}_{n=1}^N$, are generated according to $X_n \sim F(X|\theta_n^*)$ and $\theta_n^* \overset{iid}{\sim} G_K$, where

$$\begin{aligned}
G_K &= \sum_{k=1}^{K} \pi_k \delta_{\theta_k} \\
\boldsymbol{\pi} &\sim \text{Dir}(\alpha g_0) \\
\theta_k &\overset{iid}{\sim} G_0, \quad k = 1, \dots, K \qquad (10)
\end{aligned}$$

The atom θ_n^* associated with observation X_n contains parameters for some distribution, $F(x|\theta)$, with $\mathbb{P}(\theta_n^* = \theta_k|\boldsymbol{\pi}) = \pi_k$. The $\mathrm{Dir}(\alpha g_0)$ prior is often placed on $\boldsymbol{\pi}$ as shown, and G_0 is a non-atomic base distribution (i.e., continuous everywhere). For a Gaussian mixture model [3], $\theta_k = \{\mu_k, \Sigma_k\}$ and G_0 is often a conjugate Normal – inverse-Wishart prior distribution on the mean vector μ and covariance matrix Σ.

Following the proof of Section 2.1, we can use (3) to construct $\boldsymbol{\pi}$ in (10), producing

$$\begin{aligned} G_K &= \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\delta_{\theta_{Y_i}} \\ V_i &\overset{iid}{\sim} \mathrm{Beta}(1,\alpha) \\ Y_i &\overset{iid}{\sim} \mathrm{Mult}\left(\{1,\dots,K\}, g_0\right) \\ \theta_k &\overset{iid}{\sim} G_0, \quad k=1,\dots,K \end{aligned} \tag{11}$$

Sampling Y_i from the integers $\{1,\dots,K\}$ according to g_0 provides an index of the atom with which to associate mass $V_i \prod_{j=1}^{i-1}(1-V_j)$. Ishwaran and Zarepour [5] showed that, when $g_0 = (\frac{1}{K},\dots,\frac{1}{K})$ and $K \to \infty$, $G_K \to G$, where G is a Dirichlet process with continuous base measure G_0 on the infinite space $(\Theta, \mathcal{B})$, as defined in [4]. Since in the limit as $K \to \infty$, $\mathbb{P}(Y_i = Y_j | i \neq j) = 0$ and $\mathbb{P}(\theta_{Y_i} = \theta_{Y_j} | i \neq j) = 0$, there is a one-to-one correspondence between $\{Y_i\}_{i=1}^{\infty}$ and $\{\theta_i\}_{i=1}^{\infty}$. Let the function $\sigma(Y_i) = i$ reindex the subscripts on θ_{Y_i}. Then

$$\begin{aligned} G &= \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1-V_j)\delta_{\theta_i} \\ V_i &\overset{iid}{\sim} \mathrm{Beta}(1,\alpha) \\ \theta_i &\overset{iid}{\sim} G_0 \end{aligned} \tag{12}$$

Finally, we note that this representation is not as different from (3) as first appears. For example, let G_0 be a *discrete* measure on the first K positive integers. In this case $\theta \in \{1,\dots,K\}$ and (12) and (3) are essentially the same (the difference being that one is presented as a measure on the first K integers, while the other is a vector in Δ_K). Also consider finite partitions of Θ, $\{B_1,\dots,B_K\}$, with $\bigcup_k B_k = \Theta$. In this case, one can let $g_{0k} := G_0(B_k) = \int_{B_k} \theta \; G_0(d\theta)$ and the measure $G(B_k) = \pi_k$, where $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha g_0)$.

References

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

[2] 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.

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

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

[5] H. Ishwaran and M. Zarepour (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica 12* pp. 941-963.

[6] H. Ishwaran and M. Zarepour (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics,* 30, 269-283.

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

3 Appendix

Proof of Lemma 1: Let $Y \sim \text{Mult}(\{1,\dots,K\},\boldsymbol{\pi})$ and $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$. Basic probability theory allows us to write that

$$\begin{aligned} p(\boldsymbol{\pi}|\alpha g_0) &= \sum_{k=1}^{K} \mathbb{P}(Y=k|\alpha g_0)p(\boldsymbol{\pi}|Y=k,\alpha g_0) && (13)\\ \mathbb{P}(Y=k|\alpha g_0) &= \int_{\boldsymbol{\pi}\in\Delta_K} \mathbb{P}(Y=k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha g_0)\, d\boldsymbol{\pi} && (14)\end{aligned}$$

The second equation can be written as $\mathbb{P}(Y=k|\alpha g_0) = \mathbb{E}[\pi_k|\alpha g_0] = g_{0k}$. The first equation uses the posterior of a Dirichlet distribution given observation $Y=k$, which is $p(\boldsymbol{\pi}|Y=k,\alpha g_0) = \text{Dir}(\alpha g_0 + \mathbf{e}_k)$. Replacing these two equalities in the first equation, we obtain $p(\boldsymbol{\pi}|\alpha g_0) = \sum_{k=1}^{K} g_{0k}\text{Dir}(\alpha g_0 + \mathbf{e}_k)$.

Proof of Lemma 2: We use the representation of $\boldsymbol{\pi}$ as a function of gamma-distributed random variables. That is, if $\gamma_k \sim \text{Gamma}(\alpha g_{0k},\lambda)$ for $k=1,\dots,K$, and we define $\boldsymbol{\pi} := \left(\frac{\gamma_1}{\sum_k \gamma_k},\dots,\frac{\gamma_K}{\sum_k \gamma_k}\right)$, then $\boldsymbol{\pi} \sim \text{Dir}(\alpha g_0)$. Using this definition, let

$$W_1 := \left(\frac{\gamma_1}{\sum_k \gamma_k},\dots,\frac{\gamma_K}{\sum_k \gamma_k}\right),\ W_2 := \left(\frac{\gamma'_1}{\sum_k \gamma'_k},\dots,\frac{\gamma'_K}{\sum_k \gamma'_k}\right),\ V := \frac{\sum_k \gamma_k}{\sum_k \gamma_k + \sum_k \gamma'_k} \qquad (15)$$

where $\gamma_k \sim \text{Gamma}(w_k,\lambda)$ and $\gamma'_k \sim \text{Gamma}(v_k,\lambda)$. Then it follows that

$$W_1 \sim \text{Dir}(w_1,\dots,w_K),\ W_2 \sim \text{Dir}(v_1,\dots,v_K),\ V \sim \text{Beta}(\sum_k w_k, \sum_k v_k) \qquad (16)$$

where the distribution of V arises because $\sum_k \gamma_k \sim \text{Gamma}(\sum_k w_k,\lambda)$. Furthermore, Basu's theorem [1] indicates that V is independent of W_1 and W_2, or $p(V|W_1,W_2) = p(V)$. Performing the multiplication $Z = VW_1 + (1-V)W_2$ produces the gamma-distributed representation of $Z \sim \text{Dir}(w_1+v_1,\dots,w_K+v_K)$.