Appendix G

Finite-Dimensional Dirichlet Distribution

This Dirichlet distribution on the finite-dimensional unit simplex is the basis for the definition of the Dirichlet process, and therefore of central importance to Bayesian nonparametric statistics. This chapter states many of its properties and characterizations.

A random vector $X = (X_1, \ldots, X_k)$ with values in the k-dimensional unit simplex $\mathbb{S}_k := \{(s_1, \ldots, s_k): s_j \geq 0, \sum_{j=1}^k s_j = 1\}$ is said to possess a *Dirichlet distribution* with parameters $k \in \mathbb{N}$ and $\alpha_1, \ldots, \alpha_k > 0$ if it has density proportional to $x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1}$ with respect to the Lebesgue measure on $\mathbb{S}_k$.

The unit simplex $\mathbb{S}_k$ is a subset of a $(k-1)$-dimensional affine space, and so "its Lebesgue measure" is to be understood to be $(k-1)$-dimensional Lebesgue measure appropriately mapped to $\mathbb{S}_k$. The normalizing constant of the Dirichlet density depends on the precise construction. Alternatively, the vector X may be described through the vector $(X_1, \ldots, X_{k-1})$ of its first $k-1$ coordinates, as the last coordinate satisfies $X_k = 1 - \sum_{i=1}^{k-1} X_i$. This vector has a density with respect to the usual $(k-1)$-dimensional Lebesgue measure on the set $\mathbb{D}_{k-1} = \{(x_1, \ldots, x_{k-1}): \min_i x_i \geq 0, \sum_{i=1}^{k-1} x_i \leq 1\}$. The inverse of the normalizing constant is given by the *Dirichlet form*

$$\int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-x_1-\cdots-x_{k-2}} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_{k-1}^{\alpha_{k-1}-1} \times (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}\, dx_{k-1} \cdots dx_2\, dx_1. \tag{G.1}$$

The Dirichlet distribution takes its name from this integral, which by successive integrations and scalings to beta integrals can be evaluated to $\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)/\Gamma(\alpha_1 + \cdots + \alpha_k)$.

Definition G.1 (Dirichlet distribution) The *Dirichlet distribution* $\mathrm{Dir}(k; \alpha)$ with parameters $k \in \mathbb{N} \setminus \{1\}$ and $\alpha = (\alpha_1, \ldots, \alpha_k) > 0$ is the distribution of a vector $(X_1, \ldots, X_k)$ such that $\sum_{i=1}^k X_i = 1$ and such that $(X_1, \ldots, X_{k-1})$ has density

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_{k-1}^{\alpha_{k-1}-1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}, \quad x \in \mathbb{D}_{k-1}. \tag{G.2}$$

The Dirichlet distribution with parameters k and $\alpha \geq 0$, where $\alpha_i = 0$ for $i \in I \subsetneq \{1, \ldots, k\}$, is the distribution of the vector $(X_1, \ldots, X_k)$ such that $X_i = 0$ for $i \in I$ and such that $(X_i: i \notin I)$ possesses a lower-dimensional Dirichlet distribution, given by a density of the form (G.2).

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

For $k = 2$, the vector (X_1, X_2) is completely described by a single coordinate, where $X_1 \sim \mathrm{Be}(\alpha_1, \alpha_2)$ and $X_2 = 1 - X_1 \sim \mathrm{Be}(\alpha_2, \alpha_1)$. Thus the Dirichlet distribution is a multivariate generalization of the beta distribution. The $\mathrm{Dir}(k; 1, \ldots, 1)$-distribution is the uniform distribution on $\mathbb{S}_k$.

Throughout the section we write $|\alpha|$ for $\sum_{i=1}^k \alpha_i$.

There are several handy structural characterizations of the Dirichlet distribution.

Proposition G.2 (Representations) *For random variables $Y_1, \ldots, Y_k$ and $Y = \sum_{i=1}^k Y_i$,*

(i) *If $Y_i \overset{ind}{\sim} \mathrm{Ga}(\alpha_i, 1)$, then $(Y_1, \ldots, Y_k)/Y \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$, and is independent of Y.*
(ii) *If $Y_i \overset{ind}{\sim} \mathrm{Be}(\alpha_i, 1)$, then $((Y_1, \ldots, Y_k)|\, Y = 1) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$.*
(iii) *If $Y_i \overset{ind}{\sim} \mathrm{Exp}(\alpha_i)$, then $((e^{-Y_1}, \ldots, e^{-Y_k})|\, \sum_{i=1}^k e^{-Y_i} = 1) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$.*

Proof We may assume that all α_i are positive.

(i). The Jacobian of the inverse of the coordinate transformation

$$(y_1, \ldots, y_k) \mapsto (y_1/y, \ldots, y_{k-1}/y, y) =: (x_1, \ldots, x_{k-1}, y)$$

is given by y^{k-1}. The density of the $\mathrm{Ga}(\alpha_i, 1)$-distribution is proportional to $e^{-y_i} y_i^{\alpha_i - 1}$. Therefore the joint density of $(Y_1/Y, \ldots, Y_{k-1}/Y, Y)$ is proportional to

$$e^{-y} y^{|\alpha|-1} x_1^{\alpha_1 - 1} \cdots x_{k-1}^{\alpha_{k-1}-1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}.$$

This factorizes into a Dirichlet density of dimension $k - 1$ and the $\mathrm{Ga}(|\alpha|, 1)$-density of Y.

(ii). The Jacobian of the transformation $(y_1, \ldots, y_k) \mapsto (y_1, \ldots, y_{k-1}, y)$, for $y = \sum_{i=1}^k y_i$, is equal to 1. The $\mathrm{Be}(\alpha_i, 1)$-density is proportional to $y_i^{\alpha_i - 1}$. Therefore the joint density of $(Y_1, \ldots, Y_{k-1}, Y)$ is proportional to $y_1^{\alpha_1 - 1} \cdots y_{k-1}^{\alpha_{k-1}-1} (y - y_1 - \cdots - y_{k-1})^{\alpha_k - 1}$, for $0 < y_i < 1$ for $i = 1, \ldots, k-1$ and $0 < y - \sum_{i=1}^{k-1} y_i < 1$. The conditional density of $(Y_1, \ldots, Y_{k-1})$ given $Y = 1$ is proportional to this expression with y taken equal to 1, i.e. a Dirichlet density (G.2).

(iii). This follows from (ii), since $e^{-Y_i} \overset{\text{ind}}{\sim} \mathrm{Be}(\alpha_i, 1)$, $i = 1, \ldots, k$. □

The gamma representation in Proposition G.2(i) leads to several important properties. The first result states that the conditional distribution of a collection of coordinates given their sum inherits the Dirichlet structure and is independent of the total mass that is conditioned.

Proposition G.3 (Aggregation) *If $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ and $Z_j = \sum_{i \in I_j} X_i$ for a given partition $I_1, \ldots, I_m$ of $\{1, \ldots, k\}$, then*

(i) *$(Z_1, \ldots, Z_m) \sim \mathrm{Dir}(m; \beta_1, \ldots, \beta_m)$, where $\beta_j = \sum_{i \in I_j} \alpha_i$, for $j = 1, \ldots, m$.*
(ii) *$(X_i/Z_j\colon i \in I_j) \overset{ind}{\sim} \mathrm{Dir}(\#I_j; \alpha_i, i \in I_j)$, for $j = 1, \ldots, m$.*
(iii) *$(Z_1, \ldots, Z_m)$ and $(X_i/Z_j\colon i \in I_j)$, $j = 1, \ldots, m$, are independent.*

Conversely, if X is a random vector such that (i)–(iii) hold, for a given partition $I_1, \ldots, I_m$ and $Z_j = \sum_{i \in I_j} X_i$, then $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$.

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

Proof In terms of the gamma representation $X_i = Y_i/Y$ of Proposition G.2 we have

$$Z_j = \frac{\sum_{i \in I_j} Y_i}{Y}, \qquad \text{and} \qquad \frac{X_i}{Z_j} = \frac{Y_i}{\sum_{i \in I_j} Y_i}.$$

Because $W_j := \sum_{i \in I_j} Y_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\beta_j, 1)$ for $j = 1, \ldots, m$, and $\sum_j W_j = Y$, the Dirichlet distributions in (i) and (ii) are immediate from Proposition G.2. The independence in (ii) is immediate from the independence of the groups $(Y_i : i \in I_j)$, for $j = 1, \ldots, m$. By Proposition G.2 W_j is independent of $(Y_i/W_j : i \in I_j)$, for every j, whence by the independence of the groups the variables W_j, $(Y_i/W_j : i \in I_j)$, for $j = 1, \ldots, m$, are jointly independent. Then (iii) follows, because $(X_i/Z_j : i \in I_j, j = 1, \ldots, m)$ is a function of $(Y_i/W_j : i \in I_j, j = 1, \ldots, m)$ and $(Z_1, \ldots, Z_m)$ is a function of $(W_j : j = 1, \ldots, m)$.

The converse also follows from the gamma representation. □

Corollary G.4 (Moments) *If* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$, *then* $X_i \sim \mathrm{Be}(\alpha_i, |\alpha| - \alpha_i)$. *In particular,* $\mathrm{E}(X_i) = \alpha_i/|\alpha|$ *and* $\mathrm{var}(X_i) = \alpha_i(|\alpha| - \alpha_i)/(|\alpha|^2(|\alpha| + 1))$. *Furthermore,* $\mathrm{cov}(X_i, X_j) = -\alpha_i\alpha_j/(|\alpha|^2(|\alpha| + 1))$ *and, with* $r = r_1 + \cdots + r_k$,

$$\mathrm{E}(X_1^{r_1} \cdots X_k^{r_k}) = \frac{\Gamma(\alpha_1 + r_1) \cdots \Gamma(\alpha_k + r_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \times \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + r)}. \tag{G.3}$$

In particular, if $r_1, \ldots, r_k \in \mathbb{N}$, *then the expression in* (G.3) *is equal to* $\alpha_1^{[r_1]} \cdots \alpha_k^{[r_k]}/|\alpha|^{[r]}$, *where* $x^{[m]} = x(x+1)\cdots(x+m-1)$, $m \in \mathbb{N}$, *stands for the ascending factorial.*

Proof The first assertion follows from Proposition G.3 by taking $m = 2$, $I_i = \{i\}$, $I_2 = I \setminus \{i\}$, for $I = \{1, \ldots, k\}$. Next the expressions for expectation and variance follow by the properties of the beta distribution.

For the second assertion, we take $m = 2$, $I_1 = \{i, j\}$ and $I_2 = I \setminus I_1$ in Proposition G.3 to see that $X_i + X_j \sim \mathrm{Be}(\alpha_i + \alpha_j, |\alpha| - \alpha_i - \alpha_j)$. This gives $\mathrm{var}(X_i + X_j) = (\alpha_i + \alpha_j)(|\alpha| - \alpha_i - \alpha_j)/(|\alpha|^2(|\alpha| + 1))$, and allows to obtain the expression for the covariance from the identity $2\,\mathrm{cov}(X_i, X_j) = \mathrm{var}(X_i + X_j) - \mathrm{var}(X_i) - \mathrm{var}(X_j)$.

For the derivation of (G.3), observe that the mixed moment is the ratio of two Dirichlet forms (G.1) with parameters $(\alpha_1 + r_1, \ldots, \alpha_k + r_k)$ and $(\alpha_1, \ldots, \alpha_k)$. □

The *stick-breaking* representation of a vector $(X_1, \ldots, X_k)$ is given by

$$X_j = V_j \prod_{i<j}(1 - V_i), \qquad V_j = \frac{X_j}{1 - \sum_{i<j} X_i}, \qquad j = 1, \ldots, k-1.$$

(If the vector takes its values in the unit simplex, then X_k is redundant, but one may define $V_k = 1$ to make the first equation true also for $j = k$.) The stick lengths V_j of a Dirichlet vector possess a simple characterization.

Corollary G.5 (Stick-breaking) *A vector* $(X_1, \ldots, X_k)$ *is* $\mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*-distributed if and only if the stick lengths* $V_1, \ldots, V_{k-1}$ *are independent with* $V_j \sim \mathrm{Be}(\alpha_j, \sum_{i>j} \alpha_i)$, *for* $j = 1, \ldots, k-1$.

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

Proof Since the map between the vector $(X_1, \dots, X_{k-1})$ and the stick lengths is a bimeasurable bijection, it suffices to derive the distribution of the stick lengths from the Dirichlet distribution of X. By (iii) of Proposition G.3 applied with the partition $I_1 = \{1\}$ and $I_2 = \{2, \dots, k\}$ the variable $V_1 = X_1$ and vector $(X_2, \dots, X_k)/(1 - X_1)$ are independent, by (i) the first is $\mathrm{Be}(\alpha_1, \sum_{i>1} \alpha_i)$-distributed and by (ii) the second possesses a $\mathrm{Dir}(k-1; \alpha_2, \dots, \alpha_k)$-distribution. We repeat this argument on the vector $(X_2, \dots, X_k)/(1-X_1)$, giving that $V_2 = X_2/\sum_{i>2} X_i$ is independent of $(X_3, \dots, X_k)/(1-X_1 - X_2)$ (where the factor $1 - X_1$ has cancelled out upon computing the quotients), and this variable and vector possess $\mathrm{Be}(\alpha_2, \sum_{i>2} \alpha_i)$- and $\mathrm{Dir}(k-2; \alpha_3, \dots, \alpha_k)$-distributions. Further repetitions lead to $V_1, \dots, V_{k-2}$ being independent beta variables independent of $(X_{k-1}, X_k)/(1 - \sum_{i<k-1} X_i)$, from which we extract the distribution of its first coordinate V_{k-1}. ☐

Dirichlet distributions of fixed dimension depend continuously on their parameter vector α. If the sum of the parameters tends to zero or infinity, then the weak limit points are discrete, and supported on the vertices $e_1, \dots, e_k$ of the unit simplex or at a single point, respectively.

Proposition G.6 (Limits) (i) *If* $\alpha \to \beta \in [0, \infty)^k$, $\beta \neq 0$, *then* $\mathrm{Dir}(k; \alpha) \rightsquigarrow \mathrm{Dir}(k; \beta)$.
(ii) *If* $|\alpha| \to 0$ *such that* $\alpha/|\alpha| \to \rho$, *then* $\mathrm{Dir}(k; \alpha) \rightsquigarrow \sum_{i=1}^k \rho_i \delta_{e_i}$.
(iii) *If* $|\alpha| \to \infty$ *such that* $\alpha/|\alpha| \to \rho$, *then* $\mathrm{Dir}(k; \alpha) \rightsquigarrow \delta_\rho$.

Proof For convenience of notation let $(X_1, \dots, X_k) \sim \mathrm{Dir}(k; \alpha)$.

(i). If all coordinates of β are positive, then the result follows by the (pointwise) convergence of densities, since the gamma function is continuous. If some coordinate β_i is zero, but $\beta = \lim \alpha > 0$, then $\mathrm{E}X_i = \alpha_i/|\alpha| \to 0$, and hence X_i tends to zero in distribution. After eliminating these indices, convergence of the vector of the remaining coordinates follows by convergence of densities.

(ii). From the form of the beta densities of the coordinates it follows that $\mathrm{P}(\epsilon < X_i < 1 - \epsilon) \to 0$ for every $\epsilon > 0$. Therefore every weak limit point $(Y_1, \dots, Y_k)$ has $Y_i \in \{0, 1\}$ a.s. Because $\sum_{i=1}^k Y_i = 1$, it follows that $(Y_1, \dots, Y_k) \in \{e_1, \dots, e_k\}$. Also by convergence of moments $\mathrm{E}(Y_1, \dots, Y_k) = \rho$.

(iii). By assumption $\mathrm{E}X_i = \alpha_i/|\alpha| \to \rho_i$, while $\mathrm{var}\, X_i \to 0$ for every i. ☐

We may include the limit distributions in (ii) and (iii) into an extended family of *generalized Dirichlet distributions* $\overline{\mathrm{Dir}}(k; M, \rho)$ parameterized by three parameters, and defined to be the ordinary Dirichlet distribution $\mathrm{Dir}(k; M\rho)$ if $M \in (0, \infty)$ and $\rho \in [0, \infty)^k \setminus \{0\}$, to be $\sum_{i=1}^k \rho_i \delta_{e_i}$ if $M = 0$, and to be δ_ρ if $M = \infty$. This family is then closed under weak convergence, and is characterized by conditional independence properties in the following proposition. The parameter ρ is the mean vector, while M is a shape parameter (which is identifiable provided we do not allow $\rho = e_i$ if $M = 0$).

Let $(X_1, \dots, X_n)$ be a vector of nonnegative variables X_i with $\sum_{i=1}^n X_i \leq 1$, and write $S_j = \sum_{i=1}^j X_j$ and $S_I = \sum_{i \in I} X_i$. The vector $(X_i\colon i \in I)$ is said to be *neutral* in $(X_1, \dots, X_n)$ if it is independent of the vector $(X_j/(1 - S_I)\colon j \notin I)$. The vector $(X_1, \dots, X_n)$ is said to be *completely neutral* if the variables $X_1, X_2/(1 - S_1), X_3/(1 - S_2), \dots, X_n/(1 - S_{n-1})$ are mutually independent. To give a meaning to the independence

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

conditions if the variables $S_I, S_1, \ldots, S_{n-1}$ can assume the value one, each quotient A/B should be replaced by a variable Y such that $A = BY$ a.s. and the independence condition made to refer to the Ys. An equivalent definition of complete neutrality is that there exist independent variables $Y_1, \ldots, Y_n$ such that $X_1 = Y_1, X_2 = (1 - Y_1)Y_2, X_3 = (1 - Y_1)(1 - Y_2)Y_3, \ldots$

If $X_1, \ldots, X_n$ are probabilities of sets in a partition $\{B_1, \ldots, B_n\}$ of a measurable space, then complete neutrality expresses a stick-breaking mechanism to distribute the total mass. The variable X_1 is the mass of the first set B_1, the variable X_2 is the part of the remaining probability assigned to B_2, etc., and all these conditional probabilities are independent. A bit in contradiction to its name, "complete neutrality" is dependent on the ordering of the variables. The masses may well be distributed in complete neutrality in one direction, but not in other "directions."

The following proposition only considers the symmetric case.

Proposition G.7 (Characterizations) *Let $(X_1, \ldots, X_k)$ be a random vector that takes values in $\mathbb{S}_k$ for $k \geq 3$ and be such that none of its coordinates vanishes a.s. Then the following statements are equivalent:*

(i) *X_i is neutral in $(X_1, \ldots, X_{k-1})$, for every $i = 1, \ldots, k-1$.*
(ii) *$(X_j\colon j \neq i, k)$ is neutral in $(X_1, \ldots, X_{k-1})$, for every $i = 1, \ldots, k-1$.*
(iii) *$(X_1, \ldots, X_{k-1})$ is completely neutral and X_{k-1} is neutral in $(X_1, \ldots, X_{k-1})$.*
(iv) *$(X_1, \ldots, X_k)$ possesses a Dirichlet distribution or a weak limit of Dirichlet distributions.*

Proof The equivalence of (i), (ii) and (iv) is the main result of Fabius (1964). The equivalence of (iii) and (iv) is Theorem 2 of James and Mosimann (1980). □

An important reason why the Dirichlet distribution is important in Bayesian inference is its conjugacy with respect to the multinomial likelihood.

Proposition G.8 (Conjugacy) *If $p \sim \mathrm{Dir}(k; \alpha)$ and $N|\, p \sim \mathrm{M}_k(n; p)$, then $p|\, N \sim \mathrm{Dir}(k; \alpha + N)$.*

Proof If some coordinate α_i of α is zero, then the corresponding coordinate p_i of p is zero with probability one, and hence so is the coordinate N_i of N. After removing these coordinates we can work with densities. The product of the Dirichlet density and the multinomial likelihood is proportional to

$$p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \times p_1^{N_1} \cdots p_k^{N_k} = p_1^{\alpha_1 + N_1 - 1} \cdots p_k^{\alpha_k + N_k - 1}.$$

This is proportional to the density of the $\mathrm{Dir}(k; \alpha_1 + N_1, \ldots, \alpha_k + N_k)$-distribution. □

The following three results give a mixture decomposition and two regenerative properties with respect to taking certain random convex combinations.

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

Proposition G.9 *For $k \in \mathbb{N}$ and any α with $|\alpha| > 0$,*

$$\sum_{i=1}^{k} \frac{\alpha_i}{|\alpha|} \mathrm{Dir}(k; \alpha_1, \ldots \alpha_{i-1}, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k) = \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k). \tag{G.4}$$

Proof If $p \sim \mathrm{Dir}(k; \alpha)$ and $N|\, p \sim \mathrm{MN}_k(1; p)$, then $\mathrm{P}(N = i) = \alpha_i/|\alpha|$, and $p|\,\{N = i\} \sim \mathrm{Dir}(k; \alpha + e_i)$, where e_i is the ith unit vector by Proposition G.8. The assertion now follows from decomposing the marginal distribution of p over its conditionals, given N. □

Proposition G.10 *If $X \sim \mathrm{Dir}(k; \alpha)$, $Y \sim \mathrm{Dir}(k; \beta)$, and $V \sim \mathrm{Be}(|\alpha|, |\beta|)$ are independent random vectors, then $VX + (1 - V)Y \sim \mathrm{Dir}(k; \alpha + \beta)$. In particular, if $X \sim \mathrm{Dir}(k; \alpha)$ and $V \sim \mathrm{Be}(|\alpha|, |\beta|)$, then $VX + (1 - V)e_i \sim \mathrm{Dir}(k; \alpha + \beta e_i)$, where e_i is the ith unit vector in $\mathbb{R}^k$, and $i \in \{1, \ldots, k\}$.*

Proof Let $W_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_i, 1)$, $R_i \overset{\text{ind}}{\sim} \mathrm{Ga}(\beta_i, 1)$, $i = 1, \ldots, k$, be independent collections and represent $X_i = W_i/W$, $Y_i = R_i/R$, $V = W/(W + R)$, where $W = \sum_{i=1}^k W_i$, $R = \sum_{i=1}^k R_i$. Then the ith component of the convex combination $VX + (1 - V)Y$ is given by $(W_i + R_i)/(W + R)$, and hence the proof of the first assertion follows from the gamma representation, Proposition G.2(i).

The second assertion follows from the first by viewing $Y = e_i$ as a random vector distributed as $\mathrm{Dir}(k; \beta e_i)$. □

Proposition G.11 *If $p \sim \mathrm{Dir}(k; \alpha)$, $N \sim \mathrm{MN}_k(1; \alpha)$ and $V \sim \mathrm{Be}(1, |\alpha|)$ are independent, then $VN + (1 - V)p \sim \mathrm{Dir}(k; \alpha)$.*

Proof If $Y_0, Y_1, \ldots, Y_k \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_i, 1)$, $i = 0, 1, \ldots, k$, where $\alpha_0 = 1$, then the vector (Y_0, Y) for $Y = \sum_{i=1}^k Y_i$ is independent of $p := (Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}(k, \alpha_1, \ldots, \alpha_k)$ and $V = Y_0/(Y_0 + Y) \sim \mathrm{Be}(1, |\alpha|)$, by Propositions G.3 and G.2. Furthermore

$$(V, (1 - V)p) = (Y_0, Y_1, \ldots, Y_k)/(Y_0 + Y) \sim \mathrm{Dir}(k + 1; 1, \alpha).$$

Next merging the 0th cell with the ith, for any $i = 1, \ldots, k$, we obtain, again from Proposition G.3, that

$$(Ve_i + (1 - V)p) \sim \mathrm{Dir}(k; \alpha + e_i), \quad i = 1, \ldots, k. \tag{G.5}$$

The distribution of the left side with e_i replaced by N is the mixture at the left-hand side of (G.4). □

Lemma G.12 *If $V_a \sim \mathrm{Be}(a, a)$, then as $a \to \infty$,*

(i) $\mathrm{P}(|2V_a - 1| > x) \leq 2\pi^{-1/2} a^{1/2} e^{-x^2 a}$.
(ii) $a\,\mathrm{E}\log(2V_a) = O(1)$.
(iii) $\sqrt{a}\,\mathrm{E}|\log(2V_a)| \to \pi^{-1/2}$ *and* $a\,\mathrm{E}(\log(2V_a))^2 \to 1/2$.

Proof (i). Because the integrand is decreasing on $(1/2, 1]$, the integral $\int_{1/2+x/2}^1 v^{a-1}(1 - v)^{a-1}\, dv$ is bounded by $(1/2 + x/2)^{a-1}(1/2 - x/2)^{a-1}(1 - 1/2 - x/2) \leq 2^{-2a+1}(1 -$

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

$x^2)^a$. Combine this with the inequality $B(a,a) \geq \pi^{1/2}2^{-2a+1}/\sqrt{a}$, which follows from the *duplication property* of the gamma function $\Gamma(2a) = \pi^{-1/2}2^{2a-1}\Gamma(a)\Gamma(a+\frac{1}{2})$ and the estimate $\Gamma(a+\frac{1}{2}) \leq \Gamma(a)\sqrt{a}$.

(ii) The moment $\mathrm{E}V_a^{-m}$ can be expressed as the ratio of two beta functions, and thus be seen to tend to 2^m, as $a \to \infty$. Because $|\log v| \lesssim v^{-1}$ for $v < 1/4$, we see that $\mathrm{E}|\log(2V_a)|\mathbb{1}_{V_a<1/4} \leq (\mathrm{E}(V_a^{-1}))^{1/2}\mathrm{P}(V_a \leq 1/4)^{1/2}$, which tends to zero exponentially fast by (i). Since $|\log(2x) - 2(x-1/2)| \lesssim (x-1/2)^2$ for $x > 1/4$, we find that

$$\begin{aligned}|\mathrm{E}(\log(2V_a))| &= |\mathrm{E}(\log(2V_a) - 2(V_a - 1/2))| \\ &\lesssim o(a^{-1}) + \mathrm{E}(V_a - 1/2)^2\mathbb{1}\{V_a > 1/4\} = O(a^{-1}).\end{aligned}$$

(iii) By Proposition G.2 and Corollary G.4 we can represent V_a as $Y_a/(Y_a+Z_a)$ for Y_a, Z_a i.i.d. $\mathrm{Ga}(a,1)$-variables, whence

$$\log(2V_a) = -\log\Big(1 + \frac{Z_a/a - Y_a/a}{2Y_a/a}\Big).$$

By the central limit theorem, $\sqrt{a}(Y_a/a - 1)$ and $\sqrt{a}(Z_a/a - 1)$ converge jointly in distribution to independent standard normal variables. From this it can be shown by the delta-method that $\sqrt{a}\log(2Y_a)$ tends in distribution to an $\mathrm{Nor}(0,2)$-distribution. If it can be shown that all absolute moments of these variables are bounded, then all moments converge to the corresponding $\mathrm{Nor}(0,2)$-moment, from which the assertion follows. Now by the same argument as under (ii) $\mathrm{E}|\sqrt{a}\log(2V_a)|^m\mathbb{1}\{V < 1/4\} \lesssim a^{m/2}\sqrt{\mathrm{E}(V_a)}\mathrm{P}(V_a \leq 1/4)^{1/2}$ is exponentially small, while $\mathrm{E}|\sqrt{a}\log(2V_a)|^m\mathbb{1}\{V > 1/4\} \leq \mathrm{E}|\sqrt{a}(V_a - 1/2)|^m = O(1)$, in view of (i). ☐

The following lemma plays a fundamental role in estimating prior probabilities of neighborhoods in relation to Dirichlet process and Dirichlet process mixture priors.

Lemma G.13 (Prior mass) *If $(X_1,\ldots,X_k) \sim \mathrm{Dir}(k;\alpha_1,\ldots,\alpha_k)$, where $A\epsilon^b \leq \alpha_j \leq M$, $Mk\epsilon \leq 1$, and $\sum_{j=1}^k \alpha_j = m$ for some constants A, ϵ, b, M and $M \geq m$, then there exist positive constants c and C depending only on A, M, m and b such that for any point $(x_1,\ldots,x_k)$ in the k-simplex $\mathbb{S}_k$,*

$$P\Big(\sum_{i=1}^k |X_i - x_i| \leq 2\epsilon, \min_{1\leq i\leq k} X_i > \frac{\epsilon^2}{2}\Big) \geq Ce^{-ck\log_-\epsilon}.$$

Proof First assume that $M = 1$, so that $\epsilon < k^{-1}$. There is at least one index i with $x_i \geq k^{-1}$; by relabeling, we can assume that $i = k$, and then $\sum_{i=1}^{k-1} x_i = 1 - x_k \leq (k-1)/k$. If $(y_1,\ldots,y_k)$ is contained in the k-simplex and $|y_i - x_i| \leq \epsilon^2$ for $i = 1,\ldots,k-1$, then

$$\sum_{i=1}^{k-1} y_i \leq \sum_{i=1}^{k-1} x_i + (k-1)\epsilon^2 \leq (k-1)(k^{-1}+\epsilon^2) \leq 1 - \epsilon^2 < 1.$$

Furthermore, $\sum_{i=1}^k |y_i - x_i| \leq 2\sum_{i=1}^{k-1}|y_i - x_i| \leq 2\epsilon^2(k-1) \leq 2\epsilon$ and $y_k > \epsilon^2 > \epsilon^2/2$ in view of the preceding display. Therefore the probability on the left side of the lemma is bounded below by

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

$$\mathrm{P}\Big(\max_{1\le i\le k-1}|X_i-x_i|\le\epsilon^2\Big)\ge\frac{\Gamma(m)}{\prod_{i=1}^k\Gamma(\alpha_i)}\prod_{i=1}^{k-1}\int_{\max((x_i-\epsilon^2),0)}^{\min((x_i+\epsilon^2),1)}1\,dy_i,$$

because $\prod_{i=1}^k y_i^{\alpha_i-1}\ge 1$ for every $(y_1,\ldots,y_k)\in\mathbb{S}_k$, as $\alpha_i-1\le 0$ by assumption. Since each interval of integration contains an interval of at least length ϵ^2, and $\alpha\Gamma(\alpha)=\Gamma(\alpha+1)\le 1$ for $0<\alpha\le 1$, the last display is bounded from below by

$$\Gamma(m)\epsilon^{2(k-1)}\prod_{i=1}^k\alpha_i\ge\Gamma(m)\epsilon^{2(k-1)}(A\epsilon^b)^k\ge Ce^{-ck\log_-\epsilon}.$$

This concludes the proof in the case that $M=1$.

We may assume without loss of generality that a "general" M is an integer, and represent the jth component of the Dirichlet vector as block sums of $(X_{j,m}: m=1,\ldots,M)$, where the whole collection $\{X_{j,m}:\ j=1,\ldots,k,\ \ m=1,\ldots,M\}$ is Dirichlet distributed with parameters 1 and $\alpha_{j,m}=\alpha_j/M$, $j=1,\ldots,k, m=1,\ldots,M$, and it satisfies the conditions of the lemma with $M=1$ and k replaced by Mk. Clearly, the event on the left side of the lemma contains

$$\Big\{\sum_{j=1}^k\sum_{m=1}^M|X_{j,m}-\frac{x_j}{M}|\le 2\epsilon,\ \min_{1\le j\le k-1,1\le m\le M}X_{j,m}>\epsilon^2/2\Big\}.$$

The result now follows from the special case. □

Problems

G.1 Show that if $\alpha\ne\beta$ are nonnegative k-vectors, then $\mathrm{Dir}(k;\alpha)\ne\mathrm{Dir}(k;\beta)$, unless both α and β are degenerate at the same component.

G.2 Prove the converse part of Proposition G.3.

G.3 If $Y\sim\mathrm{Be}(a,b)$, show that $\mathrm{E}(\log Y)=\Psi(a)-\Psi(a+b)$, where $\Psi(x)=\frac{d}{dx}\log\Gamma(x)$ is the digamma function.

G.4 If $U\sim\mathrm{Be}(a_0,b_0)$, $V\sim\mathrm{Be}(a_1,b_1)$, $W\sim\mathrm{Be}(a_2,b_2)$ are independently distributed, and $a_0=a_1+b_1$, $b_0=a_2+b_2$, then $(UV, U(1-V), (1-U)W, (1-U)(1-W))\sim\mathrm{Dir}(4; a_1,b_1,a_2,b_2)$. In particular, $UV\sim\mathrm{Be}(a_1,b_0+b_1)$.

G.5 (Connor and Mosimann 1969) For nonnegative random variables $X_1,\ldots,X_n$ which satisfy $\sum_{i=1}^n X_i\le 1$, the following assertions are equivalent:

(a) $(X_1,\ldots,X_j)$ is neutral in $(X_1,\ldots,X_n)$, for every $j=1,\ldots,n$.

(b) There exist independent random variables $Y_1,\ldots,Y_n$ such that $X_j=Y_j(1-Y_1)\cdots(1-Y_{j-1})$, for every $j=1,\ldots,n$.

If either (a) or (b) is true, then the vector $(X_1,\ldots,X_n)$ is called *completely neutral*. [Hint: Condition (a) means the existence of a random vector $W_{>j}$ independent of $(X_1,\ldots,X_j)$ such that $(X_{j+1},\ldots,X_n)=W_{>j}(1-S_j)$, for every j. The variables $Y_1,Y_2,\ldots$ are the consecutive relative lengths in the stick-breaking algorithm, $1-S_j=(1-Y_1)\cdots(1-Y_j)$ are the remaining stick lengths, and $W_{>j}$ are relative lengths of all the remaining sticks after j cuts. Given the Ys we can define

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022

$W_{>j,k} = (1 - Y_{j+1}) \cdots (1 - Y_{j+k-1})Y_{j+k}$, for $k = 1, \ldots, n - j$, which is a function of the $(Y_{j+1}, \ldots, Y_n)$ and hence is independent of $(X_1, \ldots, X_j)$, which is a function of $(Y_1, \ldots, Y_j)$. Given the Ws, we can take $Y_1 = X_1$, $Y_2 = W_{>1,1}$, $Y_3 = W_{>2,1}$, etc., which can be seen to be independent by arguing that Y_j is independent of $(Y_1, \ldots, Y_j)$ for every j.]

Downloaded from https://www.cambridge.org/core. University of Cambridge Centre of International Studies, on 19 Mar 2019 at 18:15:33, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781139029834.022