STA5904: Bayesian Modeling and Inference **Topic 3**

The Dirichlet Process

1 Preliminaries: The Dirichlet Distribution

The Dirichlet distribution has a special role in Bayesian statistics as the conjugate prior to multinomial sampling. Consider $Y \sim \text{Multinomial}(N, \theta)$ where $\theta \in \mathbb{S}_{C-1} = \{\theta : \theta_j \geq 0, \sum_{j=1}^{C} \theta_j = 1\}$. The likelihood of θ is given by

$$L(\theta) = \frac{N!}{\prod_{j=1}^{C} Y_j!} \prod_{j=1}^{C} \theta_j^{Y_j}.$$

The conjugate prior for θ in this setting is the *Dirichlet distribution*, $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_C)$. When $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_C)$, the first $C-1$ components of θ admit a density with respect to Lebesgue measure on $\mathbb{S}_{C-1}$

$$\pi(\theta) = \frac{\Gamma(\sum_{j=1}^{C} \alpha_j)}{\prod_{j=1}^{C} \Gamma(\alpha_j)} \prod_{j=1}^{C} \theta_j^{\alpha_j - 1}.$$

The posterior distribution of $[\theta \mid Y]$ is easily derived as

$$\pi(\theta \mid Y) \propto \prod_{j=1}^{C} \theta_j^{\alpha_j + Y_j - 1}$$

so $\theta \sim \text{Dirichlet}(\alpha_1 + Y_1, \ldots, \alpha_C + Y_C)$.

The name "Dirichlet distribution" comes from the famous Dirichlet integral equation

$$\int \prod_{j=1}^{C} \theta_j^{\alpha_j} \, \lambda(d\theta) = \frac{\prod_{j=1}^{C} \Gamma(\alpha_j)}{\Gamma(\prod_{j=1}^{C} \alpha_j)},$$

where $\lambda(d\theta)$ denotes Lebesgue measure on $\mathbb{S}_{C-1}$. A heatmap of the Dirichlet density on $\mathbb{S}_2$ is given in Figure 1.

1.1 The Role of the Dirichlet Distribution in Bayesian Nonparametrics

The Dirichlet distribution is particularly important in Bayesian nonparametrics, essentially because many nonparametric approaches can be understood as applying the logic of multinomial sampling.

Example 1 (A Bayesian Histogram)**.** Suppose we are given samples $Y_1, \ldots, Y_N \sim f$ which takes values in $[0, 1]$. Our goal is to estimate $f(y)$. To obtain a flexible model for $f(y)$, we model it as a mixture of uniform distributions

$$f(y) = \sum_{k=1}^{K} \frac{\theta_k}{x_k - x_{k-1}} I(x_{k-1} \leq y < x_k),$$

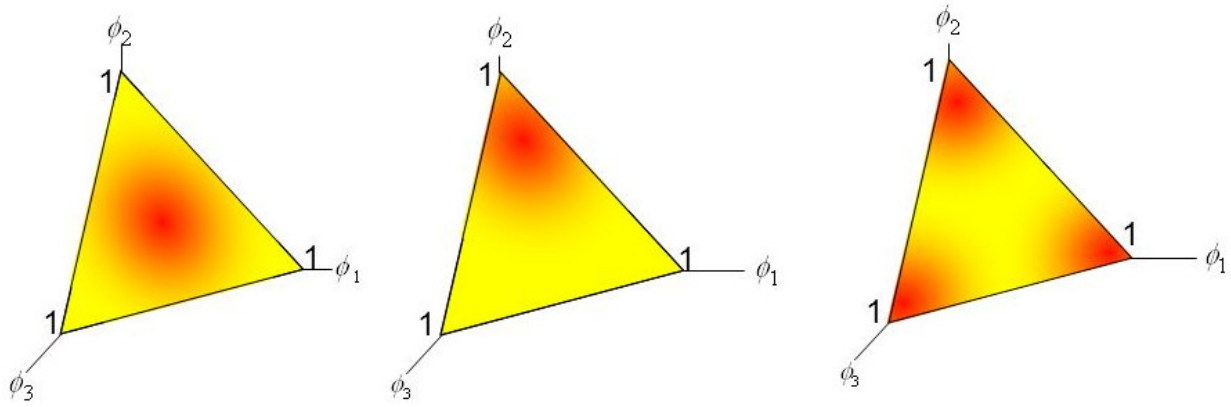

Figure 1: From left to right: $\phi \sim \text{Dirichlet}(5,5,5), \text{Dirichlet}(0.2,5,0.2), \text{Dirichlet}(0.5,0.5,0.5)$; red regions correspond to areas where the density is large.

for some grid $0 = x_0 < x_1 < \ldots < x_K = 1$ and $\theta \in \mathbb{S}_{K-1}$. The density $f(y)$ is effectively a histogram, with binwidth $x_k - x_{k-1}$. Because exactly one of the terms of the summation will be 1, and the result will be 0, we can rewrite this as

$$f(y) = \prod_{k=1}^{K} \left[\frac{\theta_k}{x_k - x_{k-1}} \right]^{I(x_{k-1} \le y < x_k)},$$

so that the likelihood is

$$L(\theta) \propto \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{I(x_{k-1} \le Y_i < x_k)} = \prod_{k=1}^{K} \theta_k^{N_k}$$

where $N_k = \sum_i I(x_{k-1} \le Y_i < x_k)$. Now if we set $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ we have the posterior $\theta \sim \text{Dirichlet}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$.

Example 2. See `Notes-03-Bayes-Hist.R`

1.2 Moments and Other Properties

Relationship to Beta Distribution: The Dirichlet distribution is a generalization of the Beta distribution. Suppose $\theta \sim \text{Beta}(\alpha, \beta)$. Then it can be shown that $(\theta, 1-\theta) \sim \text{Dirichlet}(\alpha, \beta)$. So, in some sense, the Beta distribution is the same as a Dirichlet distribution on $\mathbb{S}_1$.

Moments of the Dirichlet distribution: The Dirichlet distribution has the following moments. Let $\alpha_\bullet = \sum_k \alpha_k$.

$$E(\theta_j) = \frac{\alpha_j}{\alpha_\bullet}, \qquad \text{Var}(\theta_j) = \frac{\alpha_j(\alpha_\bullet - \alpha_j)}{\alpha_\bullet^2(\alpha_\bullet + 1)}, \qquad \text{Cov}(\theta_j, \theta_k) = \frac{-\alpha_j \alpha_k}{\alpha_\bullet^2(\alpha_\bullet + 1)}.$$

Construction as Normalized Gammas: Recall from your distribution theory course that the beta and gamma distributions have the following relationship: suppose $X \sim \text{Gam}(\alpha, 1)$ and $Y \sim \text{Gam}(\beta, 1)$. Then

$$\frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta) \qquad \text{and} \qquad X + Y \sim \text{Gam}(\alpha + \beta, 1).$$

Additionally, these two random variables are *independent*. The Dirichlet distribution has a similar construction as a normalization of gamma random variables.

Theorem 3. *Suppose* $X_1, X_2, \ldots, X_C$ *are independent random variables such that* $X_i \sim \text{Gam}(\alpha_i, 1)$. *Define* $\theta_j = X_j / \sum_{k=1}^{C} X_k$, $S = \sum_{k=1}^{C} X_k$, *and* $\theta = (\theta_1, \ldots, \theta_C)$. *Then* $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_C)$ *and* $S \sim \text{Gam}(\sum_k \alpha_k, 1)$. *Additionally,* θ *and* S *are independent.*

This construction as a normalization of independent gamma random variables is immediately useful for establishing many interesting properties of the Dirichlet distribution. For example we have the following result.

Theorem 4. *Suppose* $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_C)$ *and let* $A_1, \ldots, A_r$ *denote a partition of* $\{1, \ldots, C\}$ *(i.e.* $\bigcup_j A_j = \{1, \ldots, C\}$ *and* $A_i \cap A_j = \emptyset$*). Then*

$$\left(\sum_{j \in A_1} \theta_j, \ldots, \sum_{j \in A_r} \theta_j \right) \sim \text{Dirichlet} \left(\sum_{j \in A_1} \alpha_j, \ldots, \sum_{j \in A_r} \alpha_j \right).$$

We refer to this as the *aggregation property* of the Dirichlet distribution. It implies several other results.

Exercise 1. Suppose $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_C)$. Show that

(a) $\theta_j \sim \text{Beta}(\alpha_j, \sum_{k \neq j} \alpha_k)$; and

(b) $\theta_j + \theta_k \sim \text{Beta}(\alpha_j + \alpha_k, \sum_{i \neq j,k} \alpha_i)$.

2 The Dirichlet Process

The Dirichlet distribution is convenient to work with when the sample space of the data $Y_1, \ldots, Y_N$, denoted by $\mathcal{Y}$, is finite. Many problems in statistics do not consider $\mathcal{Y}$ finite, however.

Example 5 (Estimating a Distribution Function)**.** Suppose we have samples $Y_1, \ldots, Y_n$ which are iid with distribution function G and our goal is to estimate $G(x)$ and quantities of the form $\int f(x)\, G(dx)$. We may also be interested in estimating the quantiles q such that $G(q) = p$ for some specific value of p (when $p = 0.5$, this is the median).

What we need to be able to do to solve these problems, which set $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$, is *place a prior directly on the distribution* G. In this case, the distribution G is to be regarded as the parameter θ. Rather than being a Euclidean parameter, G is referred to as a *random probability measure*.

While priors on probability measures had been studied earlier by Freedman, the first major breakthrough was due to Thomas Ferguson. He proposed a prior, called a Dirichlet process, with the following highly desirable properties:

1. If $G \sim \text{Dirichlet}(\alpha, H)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} G$ given G, then the posterior distribution of $[G \mid Y_1, \ldots, Y_n]$ is still a Dirichlet process. Additionally, the Dirichlet process is analytically tractable.
2. The Dirichlet process has *large support* in the sense that any probability distribution Q which is absolutely continuous with respect to H is in the support of the prior.

The first property means that the Dirichlet process is, in some sense, conjugate to iid sampling, and because the Dirichlet process also has nice analytic properties this makes it easy to work with. The second property means that the Dirichlet process can approximate essentially any distribution Q arbitrarily well.

2.1 Formal Definition

Consider a *Polish space* $(\mathcal{Y}, \mathscr{B})$ (i.e., a complete, separable, metric space equipped with its Borel σ-field). We introduce the set $\mathscr{M}$ of probability measures on $(\mathcal{Y}, \mathscr{B})$; when equipped with the topology induced by weak convergence, $\mathscr{M}$ is also a Polish space, with Borel σ-field $\mathscr{F}$. The Dirichlet process is a probability distribution on $(\mathscr{M}, \mathscr{F})$.

Definition 6. A *random measure* G, taking values in $\mathscr{M}$, is said to have a *Dirichlet process* prior with *mean* $F \in \mathscr{M}$ and *concentration parameter* $\alpha > 0$ if, for every *measurable partition* $A_1, \ldots, A_k$ of $\mathcal{Y}$, we have

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dirichlet}(\alpha F(A_1), \ldots, \alpha F(A_k)).$$

This is written $G \sim \text{Dirichlet}(\alpha, F)$

Example 7. Suppose that we have real valued random variables $X_1, \ldots, X_n \overset{\text{iid}}{\sim} G$. We set $G \sim \text{Dirichlet}(\alpha, F)$ where F is a Normal$(0, 1)$ distribution. Then, for example, we have the following.

- Let $A = (-\infty, 0]$ and $B = (0, 1]$ and $C = (1, \infty)$. This is a measurable partition of $(-\infty, \infty)$. Hence $(G(A), G(B), G(C))$ has a

$$\text{Dirichlet}\left\{\alpha\Phi(0), \alpha(\Phi(1) - \Phi(0)), \alpha(1 - \Phi(1))\right\}$$

  distribution.
- This implies that $G(A) \sim \text{Beta}(\alpha/2, \alpha/2)$.

The formal definition can also be used to approximately sample from the prior. The file `Notes-03-Sim-Dir.R` illustrates two approaches for approximately sampling from the prior. One approach is to consider a very fine partition $A_1, \ldots, A_k$ where k is large and each A_j is a small interval (x_{j-1}, x_j) and $-\infty = x_0 < x_1 < \cdots < x_{k-1} < x_k = \infty$. We can then sample $(G(A_1), \ldots, G(A_k))$ from the associated Dirichlet distribution and approximate the cdf $G(x)$ as $\sum_{j=1}^{k} G(A_k) I(x \leq x_j)$.

2.2 Basic Properties

First, we note the following result about the moments of the Dirichlet process.

Theorem 8. *Suppose that* $G \sim \text{Dirichlet}(\alpha, F)$. *Then, for any measurable sets* A *and* B *we have*

$$\begin{aligned} E(G(A)) &= F(A), \\ \text{Var}(G(A)) &= \frac{F(A)\,F(A^c)}{\alpha + 1}, \\ \text{Cov}\{G(A), G(B)\} &= \frac{F(A \cap B) - F(A)F(B)}{\alpha + 1}. \end{aligned}$$

Exercise 2. Prove the above theorem.

The above result gives some interpretation for the parameters F and α. First, notice that G is centered at F, so that F is the prior mean of G. Second, note that the variance decreases to 0 as $\alpha \to \infty$. Hence, for very large values of α, we expect G to be very close to F, whereas for small values of α we allow for more uncertainty.

For example, we might think that G is likely to be well represented by a Normal$(0, 1)$ distribution (say, we are working with centered and scaled data, and think the distribution should be normal). To guard against possible misspecification, we might set $G \sim \text{Dirichlet}(\alpha, \text{Normal}(0, 1))$ with (say) $\alpha = 1$ to represent a small amount of prior information.

2.3 Conjugacy

An important property of the Dirichlet process is the following *conjugacy* result.

Theorem 9. *Suppose that* $X_1, \ldots, X_n \overset{iid}{\sim} G$ *given* G *and* $G \sim \text{Dirichlet}(\alpha, F)$. *Then the posterior distribution of* G *given* $X_1, \ldots, X_n$ *is*

$$G \sim \text{Dirichlet}\left(\alpha + n, \frac{\alpha F + \sum_{i=1}^n \delta_{X_i}}{\alpha + n}\right),$$

or equivalently $G \sim \text{Dirichlet}(\alpha + n, \alpha/(\alpha+n) \cdot F + n/(\alpha+n)\mathbb{F}_n)$ *where* $\mathbb{F}_n$ *is the empirical distribution of* $X_1, \ldots, X_n$.

We will show this later using the Pölya Urn characterization of the Dirichlet process.

2.3.1 The Bayesian Bootstrap

We observe an interesting phenomenon when we let $\alpha \to 0$ in the posterior of the Dirichlet process: we get a posterior distribution of

$$G \sim \text{Dirichlet}(n, \mathbb{F}_n).$$

This is free of the prior mean F and can be thought of as the posterior of a "non-informative" version of the Dirichlet process prior $\text{Dirichlet}(0, F)$. Of course, there is no proper $\text{Dirichlet}(0, F)$ distribution, so this strategy seems dubious; nevertheless, using this posterior reproduces many aspects of the (frequentist) Bootstrap procedure, and hence is referred to as the *Bayesian Bootstrap*. Recall that the bootstrap bases inference on a discrete measure $\frac{1}{n}\sum_{i=1}^n \omega_i \delta_{X_i}$ where the weight ω is sampled from a $\text{Multinomial}(n, 1/n, \ldots, 1/n)$ distribution (corresponding to resampling the data with replacement).

Now, consider the partition $\{\{X_1\}, \{X_2\}, \ldots, \{X_n\}, A\}$ where A is every possible value of X not realized in the sample. Appealing to the definition of the Dirichlet process, the mass assigned to each of these sets has a $\text{Dirichlet}(1, \ldots, 1, 0)$ prior. Hence, we can express samples of G from the posterior as

$$G = \frac{1}{n}\sum_{i=1}^n \omega_i \delta_{X_i}, \qquad \omega \sim \text{Dirichlet}(1, \ldots, 1).$$

This is the same as the bootstrap, but with Dirichlet weights instead of multinomial weights. An example of the use of the Bayesian bootstrap, and a comparison with the bootstrap, is given in `Notes-03-DP-Correlations.R`. This revisits the point spread data to perform inference about the correlation between true and predicted spreads in football games (see Example 3 of the second notes).

2.4 Existence

Given the definition described for the Dirichlet process, it is not obvious at all that such a random measure G *even exists*. That is, saying "let G be a random measure with these properties" may be as nonsensical as saying "let n be an integer such that $1 < n < 2$."

There are several different mechanisms one can use to show that the Dirichlet process exists.

2.4.1 The Stick-Breaking Construction

One way to show that Dirichlet processes exist is to construct one, explicitly, using quantities that we already know exist. Such a construction is given by the *stick-breaking construction* we define below. This construction was introduced by Sethuraman (1994, Statistica Sinica). Define random variables

$$\omega_k = V_k \prod_{j<k}(1 - V_j), \qquad V_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha),$$

and let $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} F$, where $\alpha > 0$ and F is a distribution. Then define

$$G = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \tag{1}$$

is such that $G \sim \text{Dirichlet}(\alpha, F)$. Here, the δ_{θ_k} is a point-mass distribution at θ_k. We can sample from G in two steps:

1. Sampling $Z \sim \text{Categorical}(\omega_1, \omega_2, \ldots)$; and
2. Given $Z = k$, set $Y = \delta_{\theta_k}$.

In addition to guaranteeing the existence of the Dirichlet process, the stick-breaking construction is rich in both insights and applications. It is immediate from (1) that samples from the Dirichlet process prior are discrete and have a countably-infinite support. This implies, for example, that we should expect to see ties in the data:

$$\Pr_G(X_1 = X_2) = \sum_k \Pr_G(X_1 = \theta_k, X_2 = \theta_k) = \sum_k \omega_k^2 > 0.$$

The stick-breaking construction also suggests a mechanism for sample from the Dirichlet process prior — for some large K, if we set $V_K = 1$, then we see that $\omega_{K+1} = \omega_{K+2}, \ldots = 0$. Setting $V_K = 1$ is referred to as *truncating* the stick-breaking representation, and for the Dirichlet process it can be shown that one does not need to take K too large before this is a very good approximation.

A schematic depiction of the stick-breaking construction is given in Figure 2.

2.4.2 Existence through De Finetti's Theorem

A second proof of the existence of the Dirichlet process can be obtained through De Finetti's theorem.

Definition 10. Let $Y_1, Y_2, \ldots$ be a sequence of random variables. The sequence is called *exchangeable* if, for any bijection $\rho : \mathbb{N} \to \mathbb{N}$,

$$(Y_1, \ldots, Y_n) \overset{d}{=} (Y_{\rho(1)}, \ldots, Y_{\rho(n)}).$$

In words, exchangeability implies that the order of the observations does not matter, and that relabeling (say) (Y_1, Y_2, Y_3) as (Y_3, Y_2, Y_1) does not change anything.

Example 11. Let $Y_1, Y_2, \ldots$ be sampled iid from a (fixed) distribution F. Then the sequence is exchangeable.

Example 12. Suppose that $Y_1, Y_2, \ldots, \overset{\text{iid}}{\sim} f_\theta$ given θ, and suppose $\theta \sim \pi(\theta)$. Then $Y_1, Y_2, \ldots$ are exchangeable. Note that $Y_1, Y_2, \ldots$ are not iid in this case; they are dependent through θ, which is random.

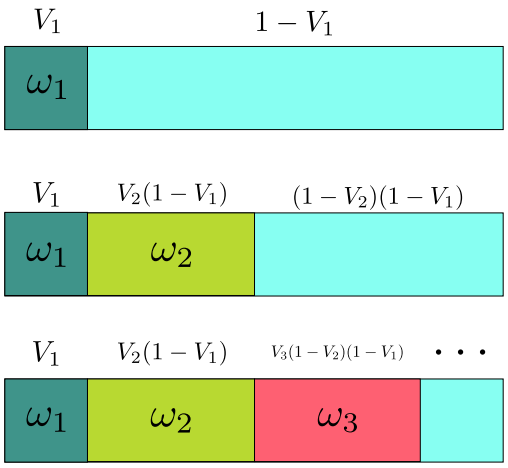

Figure 2: Stick-breaking construction; at each stage, from top to bottom, we break of V_k of the remaining stick to get ω_k.

De Finetti's theorem essentially states that Example 12 is *necessary and sufficient* for exchagenability.

Theorem 13. *Suppose $Y_1, Y_2, \ldots$ is an infinite exchangeable sequence of random variables taking values in $\mathcal{Y}$, and let $\mathscr{M}$ denote the collection of distributions on $\mathcal{Y}$. Then there exists a probability distribution Π on $\mathscr{M}$ such that*

$$Y_1, Y_2, \ldots \overset{iid}{\sim} G \text{ given } G$$
$$G \sim \Pi.$$

De Finetti's theorem will give us another way to show that the Dirichlet process exists: we will construct an infinite exchangeable sequence, and show that the associated G satisfies the definition of the Dirichlet process.