CS547Q Statistical Modeling with Stochastic Processes **Winter 2011**

Part 2: Basics of Dirichlet processes

Lecturer: Alexandre Bouchard-Côté *Scribe(s): Liangliang Wang*

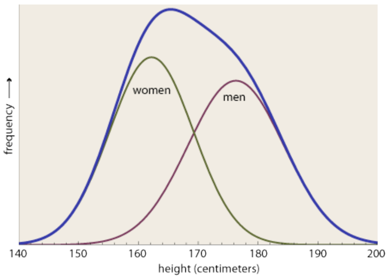
Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Last update: May 17, 2011

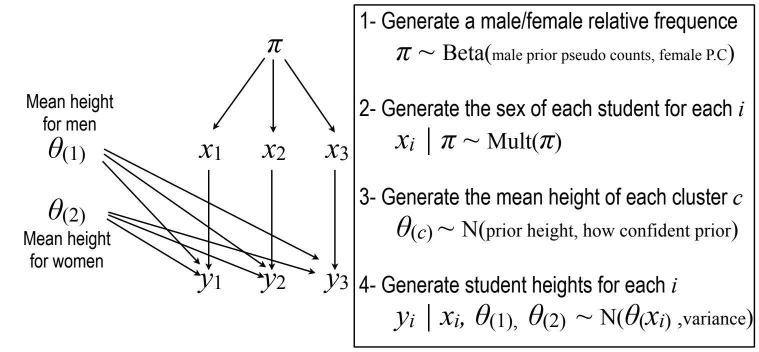
2.1 Motivation

To motivate the Dirichlet process, let us consider a simple density estimation problem: modeling the height of UBC students. We are going to take a Bayesian approach to this problem, considering the parameters as random variables. In one of the most basic models, one would define a mean and variance random parameter $\theta = (\mu, \sigma^2)$, and a height random variable normally distributed conditionally on θ, with parameters θ.[1]

Using a single normal distribution is clearly a defective approach, since for example the male/female sub-populations create a skewness in the distribution, which cannot be capture by normal distributions:



The solution suggested by this figure is to use a mixture of two normal distributions, with one set of parameters θ_c for each sub-population or cluster $c \in \{1, 2\}$. Pictorially, the model can be described as follows:



1- Generate a male/female relative frequence
$\pi \sim \text{Beta}(\text{male prior pseudo counts, female P.C})$

2- Generate the sex of each student for each i
$x_i \mid \pi \sim \text{Mult}(\pi)$

3- Generate the mean height of each cluster c
$\theta_{(c)} \sim \text{N}(\text{prior height, how confident prior})$

4- Generate student heights for each i
$y_i \mid x_i, \theta_{(1)}, \theta_{(2)} \sim \text{N}(\theta(x_i)\text{ ,variance})$

[1] Yes, this has many problems (heights cannot be negative, normal assumption broken, etc). But this is only an illustration. Moreover, some of these problems will be addressed soon.

Here Mult is the categorical or multinomial distribution[2], $x_i \dot\in \{1, 2\}$ are cluster membership variables[3], $y_i \dot\in (0, \infty)$ are the observed heights, and π is the (unknown) frequency of one of the populations.

This approach can now capture non-zero skewness, but it might still not be enough. For example, other factors such as the age of the student may also affect the distribution, and more clusters would be required to capture these finer sub-population differences. Fortunately, the model can be extended to more clusters: then more parameters θ_c are used, and π becomes a Dirichlet distribution.

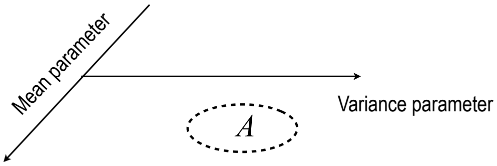
This leads to the question of determining the number of clusters. This is an important question: too many clusters will create over-fitting (good fit on the training data in terms of likelihood, but poor generalization on new data points), too few clusters will not fully exploit the information and structure present in the training data.

Many approaches exist for determining the complexity (the number of parameters, i.e. the dimensionality of a continuous parameterization of the model) of a model: cross-validation, AIC, BIC, etc. In this course, we will take another route: using nonparametric Bayesian priors. Informally, a nonparametric Bayesian prior is a distribution over models such that the complexity of the model is also random. There is no a priori bounds on the complexity in the model, but since we put a distribution on model complexities, as the complexity of the models increases, one eventually gets in the tail of a distribution, which penalizes models of high complexity. However as more data points are used to train the model, the posterior over complexity will shift towards more complex models.

To make this idea more concrete, let us go back to the UBC height density estimation. In this case, the Dirichlet process (a popular nonparametric prior) will remove the need to set a fixed number of clusters in the model.

2.2 Dirichlet process as a prior for density estimation

To understand the role of Dirichlet processes in density estimation, we will start by looking at the two-clusters model from a different point of view. Let Ω denote the set of parameters of the likelihood model (in this example, since we have a mean and a variance parameter, $\Omega = \mathbb{R} \times \mathbb{R}^+$), and let $\mathcal{F}_\Omega$ be a σ-algebra (i.e. a set of events) on that set. If $A \in \mathcal{F}_\Omega$, A is a set of parameters. For example in the UBC height example, a set A would look like:



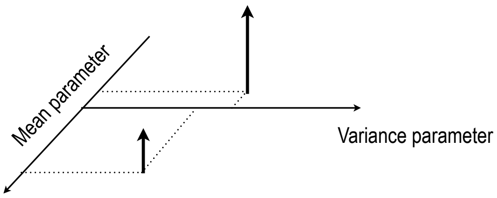
Now let $\delta_\theta : \mathcal{F}_\Omega \to \{0, 1\}$ denote a Dirac delta, which is defined as follows:

$$\delta_\theta(A) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{o.w.} \end{cases},$$

[2] in this course, unless specified otherwise, we always use multinomial distributions with the parameter n, the number of trials, set to 1

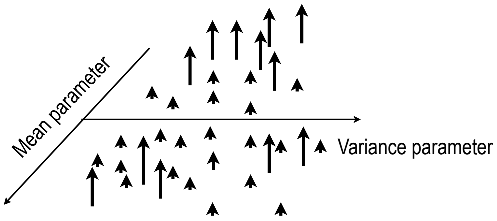
[3] In this course, we do not strictly follow the convention of using capital letters for random variable, since most objects under study will be random variables. Also, we use the notation $x \dot\in S$ to denote that the random variable x is S-valued.

and let $G = \pi\delta_{\theta_1} + (1-\pi)\delta_{\theta_2}$. Note that $G : \mathcal{F}_\Omega \to [0,1]$ and that $G(\Omega) = 1$, in other words, G is distribution. Since it is constructed from random variables, it is a random distribution. Here is an example of a realization of G:



As it is apparent from this figure, the prior over G has support over discrete distributions with two point masses.

This is where the Dirichlet process comes in: it is a prior, denoted by DP, with a support over discrete with a countably infinite number of point masses. If $G \sim \mathrm{DP}$, this means that a realization from G will look like:



In the next section, we define Dirichlet process more formally.

2.3 First definition of Dirichlet processes

A Dirichlet process has two parameters:

1. A positive real number $\alpha_0 > 0$, called the *concentration parameter*. We will see later that it can be interpreted as a precision (inverse variance) parameter.
2. A distribution $G_0 : \mathcal{F}_\Omega \to [0,1]$ called the *base measure*. We will see later that it can be interpreted as a mean parameter.

Recall that by the Kolmogorov consistency theorem, in order to guarantee the existence of a stochastic process on a probability space $(\Omega', \mathcal{F}_{\Omega'})$, it is enough to provide a consistent definition of what the marginals of this stochastic process are. As the name suggest, in the case of a Dirichlet process, the marginals are Dirichlet distributions:

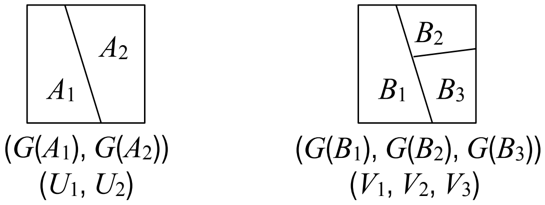
Definition 2.1 (Dirichlet Process [2]) *Let α_0, G_0 be of the types listed above. We say that $G : \mathcal{F}_{\Omega'} \to (\mathcal{F}_\Omega \to [0,1])$ is distributed according to the Dirichlet process distribution, denoted by $G \sim \mathrm{DP}(\alpha_0, G_0)$, if for all measurable partitions of Ω, $(A_1, \ldots, A_K)$ (this means that A_k are events, $A_k \in \mathcal{F}$, that they are disjoint, and that their union is equal to Ω), we have:*

$$(G(A_1), G(A_2), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \alpha_0 G_0(A_2), \ldots, \alpha_0 G_0(A_K)).$$

Here $G(A) : \mathcal{F}_{\Omega'} \to [0,1]$ denotes the random measure of a fixed set: $\big(G(A)\big)(\omega) = \big(G(\omega)\big)(A)$, for $\omega \in \Omega'$.

We now need to check that these marginals are indeed consistent. The main step can be illustrated by this example:

Proposition 2.2 *: Let (A_1, A_2) and (B_1, B_2, B_3) be the following two measurable partitions of Ω:*



I.e. $A_1 = B_1$, and $(B2, B_3)$ is a partition of A_2. Let U_1, U_2 and V_1, V_2, V_3 be the random variables as defined in the figure above ($U_1 = G(A)$, etc.). Then: $(U_1, U_2) \overset{d}{=} (V_1, V_2 + V_3)$, where the special equality symbol denotes equality in distribution.

In order to prove this, we use an important tool in the study of Dirichlet processes: gamma representation of Dirichlet distributions:

Lemma 2.3 *If $Y_1, \ldots, Y_K \sim \text{Gamma}(\alpha_i, \theta)$ are independent, where α_i is a shape parameter and θ is the scale parameter, then:*

$$\left(\frac{Y_1}{\sum_k Y_k}, \ldots, \frac{Y_K}{\sum_k Y_k}\right) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K).$$

Proof: A standard change of variable problem. See the wikipedia page on the Dirichlet distribution. ■

We now turn to the proof of proposition 2.2:

Proof: Let $\theta > 0$ be an arbitrary scale parameter, and $Y_i \sim \text{Gamma}(\alpha_0 G_0(B_i), \theta)$ be independent. We have from Lemma 2.3:

$$\begin{aligned}(V_1, V_2 + V_3) &\overset{d}{=} \left(\frac{Y_1}{\sum_k Y_k}, \frac{Y_2 + Y_3}{\sum_k Y_k}\right) \\ &\overset{d}{=} \left(\frac{Y_1}{\sum_k Y_k}, \frac{Y'}{\sum_k Y_k}\right) \\ &\overset{d}{=} (U_1, U_2),\end{aligned}$$

where $Y' \sim \text{Gamma}(G_0(B_2) + G_0(B_3), \theta) = \text{Gamma}(G_0(A_2), \theta)$ by standard properties of Gamma random variables. ■

The full proof would consider any finite number of blocks, but follows the same argument. Invariance under permutations is obvious. Therefore, we indeed have a stochastic process.

2.4 Stick breaking construction

This previous definition has the disadvantage of being non-constructive. We present in this section an alternative, constructive definition. We prove in the next section that the two definitions are equivalent.

The alternative definition, known as the stick breaking or GEM process goes as follows:

Definition 2.4 (Stick breaking construction [3]) *Let* $\beta_c \sim \mathrm{Beta}(1, \alpha_0)$ *be independent. Define the* stick lengths $\pi \sim \mathrm{GEM}(\alpha_0)$ *as follows:*

$$\pi_1 = \beta_1 \tag{2.1}$$

$$\pi_c = \beta_c \prod_{1 \le c' < c} (1 - \beta_{c'}) \quad \textit{for } c > 1. \tag{2.2}$$

The process can be understood as starting with a stick of length 1. At each step c a proportion β_c is broken and used as a stick length, and the rest is kept for the remaining ones. If at some point the stick has length L then the new stick will have length $\beta_c L$.

Since we will use π as a distribution, we need to make sure that the sum of the sticks is one a.s.:

Proposition 2.5 *If* $\pi \sim \mathrm{GEM}(\alpha_0)$*, then* $\sum_{c=1}^{\infty} \pi_c = 1$ *(a.s.).*

Proof: We have:

$$\sum_{c: c \le n} \pi_c = 1 - \prod_{j: j \le n} \underbrace{(1 - \beta_j)}_{\text{prop. not used at step } j}$$

We therefore need to prove the following:

$$\mathbb{P}\left(\lim_{n \to \infty} \prod_{j: j \le n} (1 - \beta_j) = 0 \right) = 1.$$

Let us fix $\epsilon > 0$, and define $E_n = (1 - \beta_j < 1/2)$. Since the E_n's are independent and $\sum_n \mathbb{P}(E_n) = \infty$, we can apply the second Borel-Cantelli lemma, and obtain:

$$\begin{aligned} \mathbb{P}\left(\lim_{n \to \infty} \prod_{j: j \le n} (1 - \beta_j) = 0 \right) &\le \mathbb{P}(E_n \quad \text{i.o.}) \\ &= 0. \end{aligned}$$

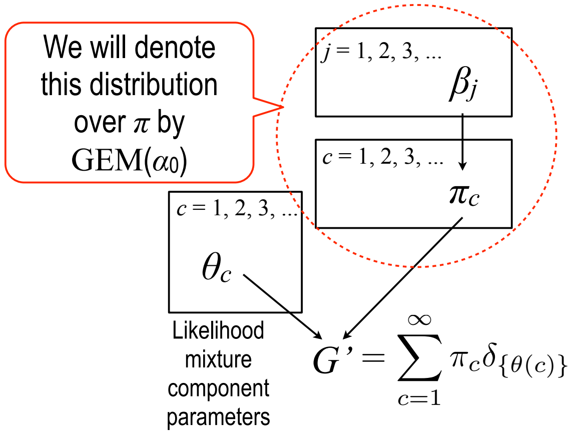
■

If we generate independently an infinite list of *stick locations*, $\theta_c \sim G_0$, and form

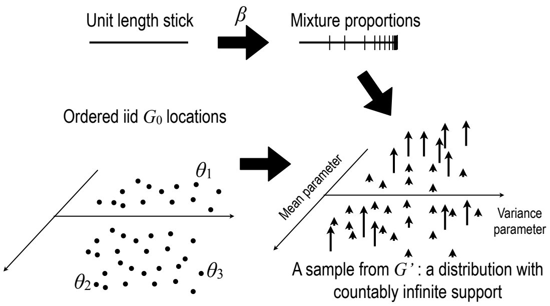
$$G' = \sum_{c=1}^{\infty} \pi_c \delta_{\theta_c}, \tag{2.3}$$

then we get an alternative definition for the Dirichlet process.

The high-level picture of this construction can also be understood from the following graphical model:


We will denote this distribution over π by GEM(α_0)
j = 1, 2, 3, ...
β_j
c = 1, 2, 3, ...
π_c
c = 1, 2, 3, ...
θ_c
Likelihood mixture component parameters
$G' = \sum_{c=1}^{\infty} \pi_c \delta_{\{\theta(c)\}}$

or from the following cartoon:



2.5 Equivalence of stick breaking and Kolmogorov consistency definitions

In this section, we present the proof of [3] that the two definitions for the Dirichlet process (stick breaking and by Kolmogorov consistency) introduced in the previous lecture are equivalent.

Let G is a Dirichlet Process defined by Kolmogorov consistency. Let G' be the constructed process using stick-breaking construction. That is,

$$G' = \sum_{c=1}^{\infty} \pi_c \delta_{\theta(c)}.$$

The goal is to show $G = G'$ in distribution.

The strategy is to show that for all partitions $(A_1, \cdots, A_K)$, the constructed process G' has finite Dirichlet marginals:

$$(G'(A_1), \cdots, G'(A_k)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_K)).$$

The key observation that will make the argument possible is a self-similarity property in the stick breaking definition. In order to explain this self-similarity in more detail, we first need to define:

Definition 2.6 *Let f be the (deterministic) map that transforms the random locations and beta variables into a Dirichlet process, as defined in the last lecture, i.e.:*

$$G' = f(\beta, \theta) = \sum_{c=1}^{\infty} \pi_c \delta_{\theta(c)}$$

Definition 2.7 *For any infinite sequence $\beta = \beta_1, \beta_2, \dots$, let β^* denote the infinite shifted suffix:*

$$\beta^* = (\beta_1, \beta_2, \cdots)^* = (\beta_2, \beta_3, \cdots).$$

We can now express the self-similarity equation as follows:

$$\begin{aligned} G' &= \pi_1 \delta_{\{\theta(1)\}} + (1 - \pi_1) f(\beta^*, \theta^*) \\ &= \pi_1 \delta_{\{\theta(1)\}} + (1 - \pi_1) G'', \end{aligned}$$

where $G'' \overset{d}{=} G'$.

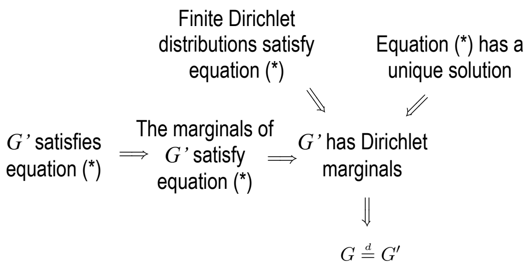
Definition 2.8 *We will use the following special equality sign to mean that there is are two random variables $G'' \overset{d}{=} G'$ such that, by plugging-in one in the left hand side and one in the right hand side, we get standard equality:*

$$G' \overset{st}{=} \pi_1 \delta_{\{\theta(1)\}} + (1 - \pi_1) G'. \tag{2.4}$$

Note that if we have a partition $(B_1, \dots, B_K)$ of the space Ω (as in the Kolmogorov definition of DPs), we can apply the left-hand and right-hand sides of the measures in Equation (2.4) to each of these sets, and get a self-similarity equation in terms of finite dimensional vectors:

$$\begin{bmatrix} G'(B_1) \\ \vdots \\ G'(B_K) \end{bmatrix} \overset{st}{=} \pi_1 \begin{bmatrix} \delta_{\theta(1)}(B_1) \\ \vdots \\ \delta_{\theta(1)}(B_K) \end{bmatrix} + (1 - \pi_1) \begin{bmatrix} G'(B_1) \\ \vdots \\ G'(B_K) \end{bmatrix}, \qquad (*) \tag{2.5}$$

We use identity (*) to show that if there is a distribution that satisfies this equation, it is unique; and that the finite Dirichlet distribution satisfies it. To summarize the high-level plan of this proof:


Finite Dirichlet distributions satisfy equation (*)
Equation (*) has a unique solution
G' satisfies equation (*)
The marginals of G' satisfy equation (*)
G' has Dirichlet marginals
$G \overset{d}{=} G'$

We have shown the two leftmost items in this graph already. We now show uniqueness.

To simplify the notation, let us introduce the random variables U, V, W as follows:

$$\underbrace{G'}_{V} = \underbrace{\pi_1 \delta_{\theta(1)}}_{U} + \underbrace{(1 - \pi_1)}_{W} G''$$

so that we can write:

$$V \overset{st}{=} U + WV \tag{2.6}$$

Note that G'' is independent of (U, W), and that W is non-degenerate, in the sense that $\mathbb{P}(0 < W < 1/2) > 0$.

Claim 2.9 *Under these conditions, the solution of (2.6) is unique.*

Proof: Suppose V and V' both satisfy (2.6) but have different distributions. Let (W_n, V_n) be independent copies of (W, V) and define

$$\begin{cases} V_1 &= V \\ V_1' &= V' \end{cases} \qquad \begin{cases} V_{n+1} &= U_n + W_n V_n \\ V_{n+1}' &= U_n + W_n V_n' \end{cases}$$

For all n,

$$\begin{aligned} V_n &\overset{d}{=} V \\ V_n' &\overset{d}{=} V'. \end{aligned}$$

Using the property $\mathbb{P}(0 < W < 1/2) > 0$ and the independence statements, we can apply the second Borel-Cantelli lemma (see Proposition 2.5 for an example of applying the second Borel-Cantelli in more details), to obtain:

$$\begin{aligned} |V_{n+1} - V_{n+1}'| &= |W_n||V_n - V_n'| \\ &= \prod_{m=1}^{n} |W_m||V_1 - V_1'| \xrightarrow{a.s.} 0. \end{aligned}$$

Hence, the solution of (2.6) is unique. ■

The final step is to show that (finite) Dirichlet distributions satisfy equation (*) as well:

Claim 2.10 *If* $Z \sim \mathrm{Dir}(\alpha_0 G_0(B_1), \ldots, \alpha_0 G_0(B_K))$ *is a Dirichlet-distributed random variables, then the following holds:*

$$\pi_1 \begin{bmatrix} \delta_{\theta(1)}(B_1) \\ \vdots \\ \delta_{\theta(1)}(B_K) \end{bmatrix} + (1 - \pi_1) \begin{bmatrix} Z_1 \\ \vdots \\ Z_K \end{bmatrix} \sim \mathrm{Dir}(\alpha_0 G_0(B_1), \ldots, \alpha_0 G_0(B_K))$$

We will need the following two easy lemmas:

Lemma 2.11 *Let* $U \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$*,* $V \sim \mathrm{Dir}(\gamma_1, \ldots, \gamma_K)$*, and* $W \sim \mathrm{Beta}(\alpha_0, \gamma_0)$ *be independent random variables, where* $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ *and similarly for* γ_0*. Then:*

$$WU + (1 - W)V \sim \mathrm{Dir}(\alpha + \gamma).$$

Lemma 2.12 *Let* e_j *denote a* $K-$ *dimensional unit vector* ($e_j(k) = 1$ *for* $k = j$ *and* 0 *otherwise),* $\bar{\gamma}_j = \gamma_j/\gamma_0$*. Then:*

$$\sum_{j=1}^{K} \bar{\gamma}_j \, \mathrm{Dir}(\gamma + e_j) \sim \mathrm{Dir}(\gamma).$$

We can now prove Claim 2.10:

Proof: Let us denote the random vector $(\delta_{\theta(1)}(B_1), \ldots, \delta_{\theta(1)}(B_K)$ by X, and observe that

$$X \sim \mathrm{Mult}(G_0(B_1), \ldots, G_0(B_K)).$$

Conditioning on $X = e_j$, we have:

$$\begin{aligned}\big(\pi_1 X + (1-\pi_1)Z | X = e_j\big) &\overset{d}{=} \pi_1 \,\mathrm{Dir}(e_j) + (1-\pi_1)Z \\ &= \mathrm{Dir}(\gamma + e_j) \quad \text{(by Lemma 2.11)},\end{aligned}$$

where, γ is defined as follows:

$$\frac{\gamma_j}{\gamma_0} = \mathbb{P}(X = e_j) = G_0(B_j)$$

Finally, we sum over the possible values of X and use Lemma 2.12 to get:

$$\pi_1 X + (1-\pi_1)Z \sim \mathrm{Dir}(\gamma).$$

That is, Z satisfies Equation (*). Hence, $G = G'$ in distribution.

■

2.6 Main properties of Dirichlet Processes

2.6.1 Moments

In this section, we derive the first and second moments of $G(A)$, for $G \sim \mathrm{DP}(\alpha_0, G_0)$ and $A \subset \Omega$. To do that, we use the Kolmogorov definition and consider the partition (A, A^c) of Ω. We get:

$$(G(A), G(A^c)) \sim \mathrm{Dir}(\alpha_0 G_0(A), \alpha_0 G_0(A^c)).$$

This implies that:

$$G(A) \sim \mathrm{Beta}(F, G),$$

where x denotes $\alpha_0 G_0(A)$, and y denotes $\alpha_0 G_0(A^c)$.

The first moment of $G(A)$ is therefore

$$\mathbb{E}[G(A)] = \frac{x}{x+y} = \frac{\alpha_0 G_0(A)}{\alpha_0 G_0(A) + \alpha_0 G_0(A^c)} = G_0(A),$$

and the second moment of $G(A)$ is

$$\begin{aligned}\mathbf{V}ar[G(A)] &= \frac{xy}{(x+y)^2(1+x+y)} \\ &= \frac{\alpha_0^2 G_0(A)(1-G_0(A))}{\alpha_0^2(\alpha_0+1)} \\ &= \frac{G_0(A)(1-G_0(A))}{\alpha_0+1}.\end{aligned}$$

This gives an interpretation of α_0 as a precision parameter for the Dirichlet process.

2.6.2 Conjugacy

Let $G \sim \mathrm{DP}(\alpha_0, G_0)$. Recall that since G is a measure-valued random variable, we can sample random variables $\underline{\theta}$ from realizations of G, i.e. we define $\underline{\theta}$ by $\underline{\theta}|G \sim G$.[4] Note that we are using the underscore to differentiate the random variables θ_i used in the stick breaking construction from the samples $\underline{\theta}$ from the Dirichlet process. Note that $\underline{\theta} = \theta_x$ for a random x sampled from a multinomial distribution with parameters given by the sticks $x \sim \mathrm{Mult}(\pi)$.[5]

In this section we show that the Dirichlet process is conjugate in the following sense: $G|\underline{\theta} \sim \mathrm{DP}(\alpha_0', G_0')$ for α_0', G_0' defined below.

To prove this result, we first look at the posterior of the finite dimensional distributions:

Lemma 2.13 *If $(B_1, \dots, B_K)$ is a measurable partition of Ω, then:*

$$(G(B_1), \dots, G(B_K))|\underline{\theta} \sim \mathrm{Dir}\left(\alpha_0 G_0(B_1) + \delta_{\{\underline{\theta}\}}(B_1), \dots, \alpha_0 G_0(B_K) + \delta_{\{\underline{\theta}\}}(B_K)\right).$$

Proof: Define the random variables $Z = (G(B_1), \dots, G(B_K))$ and $X = (\delta_{\underline{\theta}(1)}(B_1), \dots, \delta_{\underline{\theta}(1)}(B_K)$. We have:

$$\begin{aligned} Z &\sim \mathrm{Dir}(\alpha_0 G_0(B_1), \dots, \alpha_0 G_0(B_K)) \\ X|Z &\sim \mathrm{Mult}(Z). \end{aligned}$$

The result therefore follows by standard Dirichlet-multinomial conjugacy. ■

Since this result is true for all partitions, this means that the posterior is a Dirichlet process as well by the Kolmogorov consistency definition.

We can now obtain the parameters α_0', G_0' of the updated Dirichlet process. To get α_0, we take the sum of the parameters of any finite dimensional Dirichlet distribution, obtaining $\alpha_0' = \alpha_0 + 1$. To get G_0', we normalize the expression in the conclusion of Lemma 2.13 to get:

$$\begin{aligned} G_0' &= \frac{\alpha_0}{\alpha_0 + 1} G_0 + \frac{1}{\alpha_0 + 1} \delta_{\{\underline{\theta}\}} \\ \alpha_0' &= \alpha_0 + 1. \end{aligned}$$

This formula can be generalized to the case of multiple observations, by applying it n times:

Proposition 2.14 *Suppose $G \sim \mathrm{DP}(\alpha_0, G_0)$ and $\underline{\theta}_i|G \sim G$ for $i \in \{1, \dots, n\}$, iid given G. Then the posterior has the following distribution:*

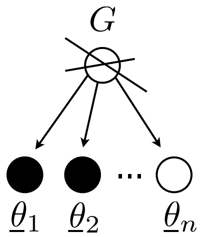
$$G|\underline{\theta}_1, \dots, \underline{\theta}_n \sim \mathrm{DP}\left(\alpha_0 + n, \frac{\alpha_0}{\alpha_0 + n} G_0 + \frac{1}{\alpha_0 + n} \sum_{i=1}^{n} \delta_{\{\underline{\theta}_i\}}\right).$$

2.6.3 Predictive distribution and the Chinese Restaurant Process (CRP)

Using the same notation as the previous section, we now seek to find an expression for the predictive distribution $\underline{\theta}_{n+1}|\underline{\theta}_1, \dots, \underline{\theta}_n$. Pictorially (using crosses to indicate marginalization, and shading for conditioning):

[4] Recall that the notation $\underline{\theta}|G \sim G$ means: for all bounded h, $\mathbb{E}[h(\underline{\theta})|G] = \int h(x) G(\,\mathrm{d}x) = \sum_{c=1}^{\infty} \pi_c h(\theta_c)$ for $\pi \sim \mathrm{GEM}(\alpha_0)$ and $\theta_c \sim G_0$ independent.

[5] Note that finite multinomial distributions over $\{1, 2, \dots, K\}$ can be extended to distributions over the infinite list $\{1, 2, \dots\}$, in which case they take as parameters an infinite list of non-negative real numbers that sum to one (e.g.: sticks from the stick breaking construction). We will show in the next lecture how sampling from such generalized multinomials can be implemented in practice.



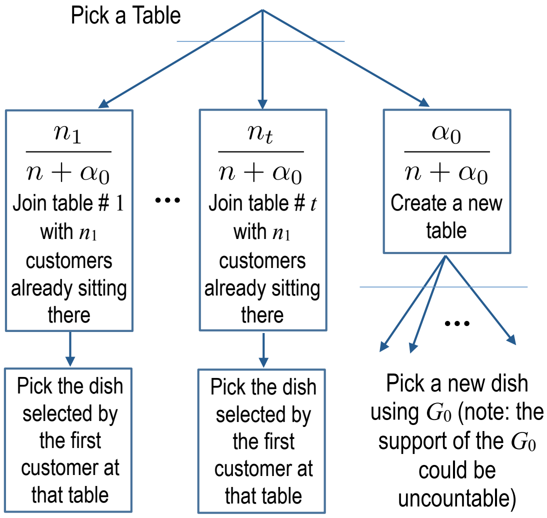
This will be useful for designing MCMC samplers over a finite state space (i.e. samplers that avoid the difficulty of representing infinite objects such as π).

Proposition 2.15 *Let $A \subset \Omega$ be measurable, $G \sim \mathrm{DP}(\alpha_0, G_0)$, and $\underline{\theta}_i|G \sim G$. Then the predictive distribution is characterized by:*

$$\mathbb{P}(\underline{\theta}_{n+1} \in A|\underline{\theta}_1, \cdots, \underline{\theta}_n) = \frac{\alpha_0}{\alpha + n} G_0(A) + \frac{1}{\alpha_0 + n} \sum_{i=1}^{n} \delta_{\underline{\theta}_i}(A).$$

A popular metaphor to interpret this formula is the *Chinese Restaurant Process* (CRP), in which customers (data points) enter sequentially in a restaurant and sit down at a table (which either contains other customers already, or is empty). When the table is empty, the customer samples a dish $\theta \sim G_0$, and when there are already customers, the dish selected by the first customer at that table is shared by the new customer.

Suppose customer $n+1$ enters the restaurant, and we want to define a probability over where he will sit given the seating arrangement of the first n customers. The formula can be interpreted as the following decision diagram over the possible outcomes:


Pick a Table

$\frac{n_1}{n + \alpha_0}$ Join table # 1 with n_1 customers already sitting there

$\frac{n_t}{n + \alpha_0}$ Join table # t with n_1 customers already sitting there

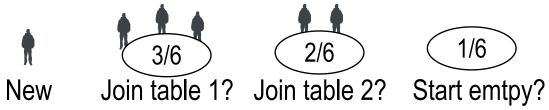
$\frac{\alpha_0}{n + \alpha_0}$ Create a new table

Pick the dish selected by the first customer at that table

Pick the dish selected by the first customer at that table

Pick a new dish using G_0 (note: the support of the G_0 could be uncountable)

Here is an example, with $\alpha_0 = 1$ of numerical probabilities of the table choices derived from the previous decision diagram:



Given α_0 and G_0, this process defines a distribution over dishes eaten by customers. We will denote the outcome of this process by $(\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_n) \sim \mathrm{CRP}(\alpha_0, G_0)$, where we denote the dish eaten by customer i by $\underline{\theta}_i$. The distribution over $(\underline{\theta}_1, \ldots, \underline{\theta}_n)$ has the following important property:

Proposition 2.16 *The distribution over the list of dishes eaten by customers in a CRP is exchangeable, i.e. if* $\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_n \sim \mathrm{CRP}(\alpha_0, G_0)$ *and* $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ *is a permutation, then*

$$(\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_n) \overset{d}{=} (\underline{\theta}_{\sigma(1)}, \underline{\theta}_{\sigma(2)}, \ldots, \underline{\theta}_{\sigma(n)}). \tag{2.7}$$

The proof follows directly from the fact that CRP emerged as the predictive distribution of draws from a Dirichlet process. It can also be done directly, which we leave as an exercise.

One can also ignore the dishes and view this process as a distribution over seating arrangements of customers. There are actually two ways of defining these seating arrangements, which we will respectively call labeled and unlabeled customer partitions. Labeled partitions correspond to the case where we can distinguish customers from each other (e.g. when there are observations attached to them and we are computing a posterior). Unlabeled partitions correspond to the case where we cannot distinguish customer from each other, in which case for example the partition $\{\{1, 2, 3\}, \{4, 5\}\}$ is deemed equivalent to the partition $\{\{1, 2, 4\}, \{3, 5\}\}$.

Since we will be interested in attaching likelihood models and observation, which will generally make customers distinguishable,[6] we will focus on a the distribution induced by this process on labeled partitions, which we will denote by $\mathrm{CRP}(\alpha_0)$.[7] If $\rho = \{\{1, 2, 3\}, \{4, 5\}\}$ for example, $\mathrm{CRP}(\rho; \alpha_0)$ for $\alpha_0 = 1$ is 1/60, which comes from applying the conditional probabilities of the CRP decision diagram 5 times:

$$\begin{aligned}
\mathrm{CRP}(\{\{1,2,3\},\{4,5\}\};\alpha_0) =& \,\mathrm{CRP}(\{\{1\}\}|\{\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{1,2\}\}|\{\{1\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{1,2,3\}\}|\{\{1,2\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{1,2,3\},\{4\}\}|\{\{1,2,3\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{1,2,3\},\{4,5\}\}|\{\{1,2,3\},\{4\}\};\alpha_0) \\
=& 1 \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{4} \times \frac{1}{5} \\
=& \frac{1}{60}.
\end{aligned}$$

Here we denote the conditional probability of a new seating arrangement given the previous one by $\mathrm{CRP}(\rho'|\rho; \alpha_0)$.

By Proposition 2.16, this is a well defined probability distribution since the order at which we construct the partition does not matter. For example, using the customer order 4 → 5 → 3 → 2 → 1 instead of 1 → 2 → 3 → 4 → 5, we get the same result:

$$\begin{aligned}
\mathrm{CRP}(\{\{1,2,3\},\{4,5\}\};\alpha_0) =& \,\mathrm{CRP}(\{\{4\}\}|\{\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{4,5\}\}|\{\{4\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{4,5\},\{3\}\}|\{\{4,5\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{4,5\},\{2,3\}\}|\{\{4,5\},\{3\}\};\alpha_0)\times \\
& \mathrm{CRP}(\{\{1,2,3\},\{4,5\}\}|\{\{4,5\},\{2,3\}\};\alpha_0) \\
=& 1 \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4} \times \frac{2}{5} \\
=& \frac{1}{60}.
\end{aligned}$$

[6] Except when G_0 has a countable support, in which case there may be identical observations with positive probability.

[7] Note that the notation for the dish-less version of the CRP can be distinguished from the dish version by the absence of a base measure parameters in the former.

Finally, it is also possible to define a distribution over the seating arrangement viewed as an unlabeled partition. Note that an unlabeled partition of n customers can be viewed as a histogram that tells you for each table size s, how many tables there are of that size, a_s, or equivalently, as a list of non-negative integers $a_1, a_2, \ldots, a_n$ such that $a_1 + 2a_2 + 3a_3 + \cdots + na_n = n$.

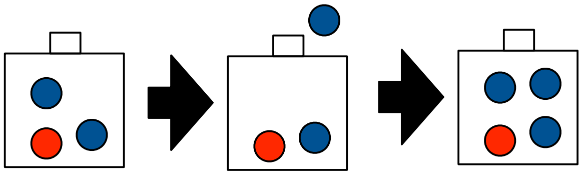
The distribution over unlabeled partitions that arise from the CRP is called Ewen's Sampling Formula (ESF), and we denote it by $(a_1, a_2, \ldots, a_n) \sim \mathrm{ESF}(\alpha_0)$. Since a relabeling ρ' of a labeled partition ρ does not change $\mathrm{CRP}(\rho; \alpha_0)$ (e.g. $\mathrm{CRP}(\{\{1,2\},\{3\}\}; \alpha_0) = \mathrm{CRP}(\{\{1,3\},\{2\}\}; \alpha_0)$), Ewen's formula can be obtained by multiplying the expression for the CRP by the number of labeled partitions corresponding to an unlabeled one. As an exercise, you can show that this leads to the following formula:

$$\mathrm{ESF}(a_1, a_2, \ldots, a_n; \alpha_0) = \frac{n!}{\alpha_0(\alpha_0+1)\cdots(\alpha_0+n-1)} \prod_{j=1}^{n} \frac{\alpha_0^{a_j}}{j^{a_j} a_j!},$$

2.6.4 Application: Polya urn

In this section, we use the result of the previous section to find the limit of the following process:

Definition 2.17 (Polya urn) *Consider an opaque urn with R red balls and B blue balls initially. At each step, one ball is drawn at random from the urn. If the ball drawn is red, an extra red ball is added to the urn; if the drawn ball is blue, an extra blue ball is added. The drawn ball is also reinserted (so there is a total of one more ball at each iteration):*



We are interested in finding an expression for the asymptotic ratio of red:blue balls in the urn as the number of steps go to infinity. In contrast to the Markov chain results, this asymptotic ratio is random (because the process is not Markovian). As an exercise, use the result of the previous section and a Dirichlet process with $\alpha_0 = R + B$ and $G_0 = \mathrm{Mult}(R/R+B, B/R+B)$ to find the distribution of the asymptotic ratio. See [1] for more on the connection between Polya urns and Dirichlet processes.

References

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

[2] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. In *Annals of Statistics*, 1973.

[3] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1994.