Introduction to the Dirichlet Process

Billy Fang

12, 19 October 2016

The following are rough notes for a two-hour reading group discussion on Chapters 1-6 of [1]. Any errors are mine.

1 The Chinese Restaurant Process

1.1 Definition

The **Chinese Restaurant Process (CRP)** is a sequence of distributions indexed by N. For a fixed $\alpha > 0$ and for each N, $\mathrm{CRP}(\alpha, N)$ is a distribution over all partitions of the labeled set $[N] := \{1, 2, \ldots, N\}$. For example, $\pi_{[5]} = \{\{1, 3\}, \{2\}, \{4, 5\}\}$ is a partition of $[5]$. The ordering of the subsets of the partition and the ordering of the elements within each subset do not matter. We sometimes think of each subset of a partition as a table in a restaurant, and customers $1, \ldots, N$ arrive sequentially and sit down at tables. We will use the terms "subset of a partition," "cluster," and "table" interchangeably.

The distribution is defined recursively. Given a partition $\pi_{[n]}$ (i.e., n people have already sat down), the destination of the next person $n + 1$ has the following distribution.

$$P(n+1 \text{ joins table } c \mid \pi_{[n]}) = \frac{|c|}{n+\alpha}$$
$$P(n+1 \text{ starts a new table} \mid \pi_{[n]}) = \frac{\alpha}{n+\alpha}.$$

That is, whenever a new person arrives at the restaurant, she starts a new table with probability proportional to α, or joints an occupied table with probability proportional to the number of people already at that table. The process begins with the first person starting her own table with probability 1, and then the new customers each join according the above distribution conditioned on the previous customers.

For example, the probability of the partition $\{\{1, 3\}, \{2\}, \{6, 4, 5\}\}$ under $\mathrm{CRP}(\alpha, 6)$ is

$$\frac{\alpha}{\alpha} \cdot \frac{\alpha}{\alpha+1} \cdot \frac{1}{\alpha+2} \cdot \frac{\alpha}{\alpha+3} \cdot \frac{1}{\alpha+4} \cdot \frac{2}{\alpha+5}.$$

In general, we see that the probability of a given partition $\pi_{[N]} \sim \mathrm{CRP}(\alpha, N)$ is

$$\begin{aligned} P(\pi_{[N]}) &= \frac{1}{\alpha(\alpha+1)\cdots(\alpha+N-1)} \prod_{c \in \pi_{[N]}} \alpha(|c|-1)! \\ &= \frac{\alpha^K}{\alpha(\alpha+1)\cdots(\alpha+N-1)} \prod_{c \in \pi_{[N]}} (|c|-1)!, \end{aligned}$$

where K represents the number of clusters c in $\pi_{[N]}$. From this equation we see that the CRP is **exchangeable** in the sense that only the sizes of the clusters affect the probability, and not the labeling. In other words, the probability of any final seating configuration of N people is the same, even if we had run the process with a different ordering of the N customers.

1.2 The CRP mixture model

As our terminology suggests, we can use the CRP to define a mixture model for clustering. The **CRP mixture model** is defined as follows.

$$\pi_{[N]} \sim \mathrm{CRP}(\alpha, N) \tag{1}$$

$$(\phi_c \mid \pi_{[N]}) \overset{\text{i.i.d.}}{\sim} G_0, \qquad c \in \pi_{[N]} \tag{2}$$

$$(x_i \mid \phi, \pi_{[N]}) \overset{\text{ind}}{\sim} F(\phi_c) \qquad i \in c$$

In words, we first draw a partition $\pi_{[N]} \sim \mathrm{CRP}(\alpha, N)$. Then, for each cluster $c \in \pi_{[N]}$ we draw a parameter ϕ_c i.i.d. from some base distribution G_0. Finally, if i is assigned to a cluster c according to $\pi_{[N]}$, then x_i is drawn from some distribution $F(\phi_c)$ parameterized by the corresponding parameter ϕ_c.

For a concrete example, consider $G_0 = N(0, 1)$, and $F(\phi_c) = N(\phi_c, 1)$. Then the cluster assignments are given by $\pi_{[N]} \sim \mathrm{CRP}(\alpha, N)$, the cluster centers $\phi - c$ are drawn i.i.d. from $N(0, 1)$, and the data x_i, conditioned on i belonging to cluster c, are drawn i.i.d. from a normal distribution centered at the cluster center ϕ_c.

As an example, consider $\pi_{[6]} = \{\{1, 3\}, \{2\}, \{6, 4, 5\}\}$, and for convenience we label the three clusters a, b, c respectively. Then we draw the cluster centers ϕ_a, ϕ_b, ϕ_c i.i.d. from $N(0, 1)$. Then, x_1 and x_3 are drawn i.i.d. from $N(\phi_a, 1)$, x_2 is drawn from $N(\phi_b, 1)$, and x_4, x_5, x_6 are drawn i.i.d. from $N(\phi_c, 1)$. [Picture omitted.]

At first glance, this model does not seem any different from a finite mixture model. However, the difference lies in the behavior as N grows: here, the number of clusters (and hence cluster parameters ϕ_c) will grow with N, which does not occur in a finite mixture model. This is the essence of the "nonparametric" aspect of "Bayesian nonparametrics."

1.3 Gibbs sampling in the CRP mixture model

We give a few quick remarks about inference in the CRP mixture model. The main goal of clustering is to find the posterior distribution of the cluster assignments

$$p(\pi_{[N]} \mid x) = \frac{p(x \mid \pi_{[N]}) p(\pi_{[N]})}{\sum_{\pi'_{[N]}} p(x \mid \pi_{[N]}) p(\pi'_{[N]})}.$$

Computing this is intractable due to the sum in the denominator: the number of partitions (known as the Bell number) grows as $O(N^N)$.

The standard way to cope with this is through sampling approaches. In the interest of time, I will not go into the details of Gibbs sampling for the CRP mixture model, but I will outline the structure so that we may compare it to the Gibbs sampler via the Dirichlet process later.

If we denote the densities of G_0 and $F(\phi_c)$ by g_0 and $f(\cdot \mid \phi_c)$, we have

$$\begin{aligned} p(\pi_{[N]} \mid x) &\propto p(\pi_{[N]}, x) \\ &= \int p(\pi_{[N]}, \phi, x)\, d\phi \\ &= p(\pi_{[N]}) \int \prod_{c \in \pi_{[N]}} \left[g_0(\phi_c) \prod_{i \in c} f(x_i \mid \phi_c) \right] d\phi \\ &= p(\pi_{[N]}) \prod_{c \in \pi_{[N]}} h(x_c), \end{aligned}$$

where

$$h(x_c) := \int \left[g_0(\phi_c) \prod_{i \in c} f(x_i \mid \phi_c) \right] d\phi_c \,.$$

The integration over ϕ results in the integrals defining $h(x_c)$, which can be computed if g_0 is the conjugate prior for the likelihoods $f(x_i \mid \phi_c)$, but otherwise the algorithm below cannot be used.

The Gibbs sampler performs a sort of "stochastic coordinate ascent" on the space of partitions. Given the previous sample $\pi_{[N]}$, a person i is removed from the partition, and then re-added to the partition according to the above distribution $p(\pi_{[N]} \mid x)$ to form a new sample $\pi'_{[N]}$. It turns out that the procedure for re-adding person i into the partition is some "mixture" of the prior CRP process and the new likelihoods h. See [1, §3.3] for details.

2 The Blackwell-MacQueen urn

The **Blackwell-MacQueen (BM) urn** is a generalization of the Pólya urn that essentially captures the first two parts (1) and (2) of the CRP mixture model above.

Let $(\theta_1, \ldots, \theta_N)$ be the cluster parameters ϕ_c for each x_i in the CRP mixture model. In our earlier example, we would have $(\theta_1, \ldots, \theta_6) = (\phi_a, \phi_b, \phi_a, \phi_c, \phi_c, \phi_c)$. We can incorporate the CRP and the base measure G_0 to describe the distribution of θ succinctly using the recursion

$$(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}.$$

Here, δ_{θ_i} is the Dirac delta measure and denotes a point mass at θ_i. [Note that the sum on the right-hand side can be rewritten as $\sum_{c \in \pi_{[n]}} |c| \delta_{\phi_c}$ by collecting terms.] We denote the distribution on $(\theta_1, \ldots, \theta_N)$ defined by the above recursion as $\theta \sim \mathrm{BM}(\alpha, G_0, N)$.

Thus, the CRP mixture model can be rewritten as

$$\begin{aligned} \theta &\sim \mathrm{BM}(\alpha, G_0, N) \\ (x_i \mid \theta_i) &\overset{\text{ind}}{\sim} F(\theta_i) \qquad\qquad i = 1, \ldots, N. \end{aligned}$$

The BM urn is exchangeable! This follows by exchangeability of the elements in the CRP model combined with the i.i.d. draws of ϕ_c from the base measure G_0. Unlike the notion of exchangeability in the CRP, the notion of exchangeability of the BM urn is compatible with de Finetti's theorem, since exchangeability here refers to the sequence $(\theta_1, \theta_2, \ldots,)$. One could also think of the BM urn as a reformulation of the CRP into a sequence form compatible with de Finetti's theorem.

De Finetti's theorem implies the existence of a *random* probability measure G such that for any N,

$$P(\theta_1 \in A_1, \ldots, \theta_N \in A_N) = \int \left[\prod_{n=1}^{N} G(A_n) \right] Q(dG)\,.$$

The statement of the theorem is sometimes rephrased as the θ_i being conditionally i.i.d. given G:

$$P(\theta_1 \in A_1, \ldots, \theta_N \in A_N \mid G) = \prod_{n=1}^{N} G(A_n).$$

We find this G in the next section.

3 The Dirichlet process

3.1 Definition

We seek the random measure G such that if $\theta \sim \mathrm{BM}(\alpha, G_0)$, then

$$(\theta_i \mid G) \overset{\text{i.i.d.}}{\sim} G. \tag{3}$$

Consider a measure of the form

$$G = \sum_{k=1}^{\infty} w_k \delta_{\phi_k}. \tag{4}$$

If the w_k and ϕ_k are fixed and $\sum_{k=1}^{\infty} w_k = 1$, then this is a valid probability measure. However, if the w_k and ϕ_k are random, then G becomes a random probability measure.

It turns out the G satisfying (3) comes from the **Dirichlet process** $\mathrm{DP}(\alpha, G_0)$, where G takes the form (4) and where

$$\begin{aligned} w &\sim \mathrm{GEM}(\alpha), \\ \phi_k &\overset{\text{i.i.d.}}{\sim} G_0. \end{aligned}$$

It remains to define the **Griffiths-Engen-McCloskey** $\mathrm{GEM}(\alpha)$ distribution.

$$\beta_k \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1,\alpha) \qquad\qquad k = 1, 2, \ldots$$
$$w_k := \beta_k \prod_{j<k} (1-\beta_j).$$

As mentioned above, we must have $\sum_k w_k = 1$, i.e., w can be viewed as a partition of the interval $[0,1]$ into countably many pieces, just as the Dirichlet distribution can be thought of a partition of the interval into finitely many pieces. Thus the $\mathrm{GEM}(\alpha)$ distribution can be viewed as a "stick-breaking" procedure. The first break point $\beta_1 \sim \mathrm{Beta}(1,\alpha)$ is drawn, and we have $w_1 = \beta_1$. We then "break off" a piece of proportion β_1 from our stick of length 1, and keep the remaining piece of length $1-\beta_1$. We then draw $\beta_2 \sim \mathrm{Beta}(1,\alpha)$, and break off a β_2 proportion of the remaining stick, which has length $w_2 = \beta_2(1-\beta_1)$.

When α is large, the β_k tend to be close to zero, so the cluster sizes are very small, so the draws θ_i resemble draws from G_0. When α is very small, the cluster sizes are larger, and our draws θ_i tend to join existing clusters.

3.2 Intuition behind stick-breaking

Why should this construction be the random measure G guaranteed by de Finetti's theorem for the BM urn? The intuition for the above construction of G is that each w_k represents the probability of being in a particular table/cluster. Let us start with w_1, the probability of being at the first table. Let $Z_1, \ldots, Z_N$ be indicators for the customers $2, \ldots, N+1$ sitting at the first table with customer 1. Then, letting $S_n = \sum_{k=1}^n Z_k$, we have

$$P(Z_1 = 1) = \frac{1}{\alpha+1}$$
$$P(Z_{n+1} = 1 \mid Z_1, \ldots, Z_n) = \frac{S_n}{n+\alpha}.$$

In particular,

$$P(Z_1 = z_1, \ldots, Z_N = z_N) = \frac{S_N! \cdot \alpha(\alpha+1)\cdots(\alpha+N-s-1)}{(\alpha+1)(\alpha+2)\cdots(\alpha+N)} = \frac{\Gamma(S_N+1)\Gamma(\alpha+N-S_N)\Gamma(\alpha+1)}{\Gamma(\alpha+N+1)\Gamma(\alpha)}$$

Note that the Z_i are exchangeable. One can verify directly that the Z_i have the same distribution as if they were instead drawn according to the model $(Z_i \mid \beta_1) \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\beta_1)$ and $\beta_1 \sim \mathrm{Beta}(1,\alpha)$, then it turns out that their joint distribution is precisely the one above, so is the random measure guaranteed by de Finetti's theorem in the Bernoulli case. Thus $w_1 = \beta_1 \sim \mathrm{Beta}(1,\alpha)$. Another way to state this is that the limiting fraction $\beta_1 = \lim_{N\to\infty} \frac{S_N}{N}$ of customers going to the first table follows the $\mathrm{Beta}(1,\alpha)$ distribution.

Having allocated probability w_1 to the first table, we have $1-w_1$ probability dedicated to other seatings. If we perform the same analysis to compute the conditional probability that customers [after the first occupant of the second table] sit at or not at the second table, given they are not at the first table, then we get the same result $\beta_2 \sim \mathrm{Beta}(1,\alpha)$. Multiplying by the probability of not being at the first table gives $w_2 = \beta_2(1-w_1)$. [One can also note that if we consider the Chinese restaurant process starting at the first occupant of the second table, and ignore the customers that go to the first table, we just have another Chinese restaurant process.]

Generalizing, we have

$$w_k = \beta_k \left(1 - \sum_{j<k} w_j\right) = \beta_k \prod_{j<k} (1-\beta_j).$$

In short, one should think of the w_k as the limiting proportion $w_k = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{\infty} \mathbf{1}_{\{\theta_i = \phi_k\}}$ of draws of the Blackwell-MacQueen urn that are labeled/colored by ϕ_k (the kth cluster).

One might be a little suspicious about thinking of w_k and ϕ_k as being associated with the kth cluster that appears, since if we were to sample i.i.d. from G, we would not have any guarantee that ϕ_1 is the first observation we see, and ϕ_2 is the second new observation, and so on. However, it turns out that both of these are equal in distribution. More precisely, suppose w, ϕ_k, and $\mathrm{DP}(\alpha, G_0)$ are as defined above, and we observe $(\theta \mid G) \overset{\text{i.i.d.}}{\sim} G$. If we let $\widetilde{w}_k$ and $\widetilde{\phi}_k$ be

the weight and value of the kth new atom we see in this sample θ, then

$$\sum_{k=1}^{\infty} w_k \delta_{\phi_k} \stackrel{d}{=} \sum_{k=1}^{\infty} \widetilde{w}_k \delta_{\widetilde{\phi}_k}.$$

See Corollaries 9 and 10 of [2] for more detail.

3.3 Gibbs sampling using stick-breaking

Consider the **Dirichlet process mixture model**.

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ (\theta_i \mid G) &\overset{\text{i.i.d.}}{\sim} G && i = 1, \ldots, N \\ (x_i \mid \theta_i) &\overset{\text{ind}}{\sim} F(\theta_i) && i = 1, \ldots, N. \end{aligned}$$

We describe the Gibbs sampler before describing the advantages of this model over the CRP/BM mixture model.

One hurdle we encounter is sampling from G, which has countably many atoms. Here, we discuss a truncation approach. There is also an exact approach that generates atoms "on the fly" as needed, and exploits the fact that only finitely many atoms are needed in practice [1, §6.4].

In the truncation, we approximate the infinite sum using $G = \sum_{k=1}^{K_{\max}} w_k \delta_{\phi_k}$. We sample $\beta_k \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha)$ for $k - 1, \ldots, K_{\max} - 1$ as before, but then set $\beta_{K_{\max}} = 1$. The definition $w_k = \beta_k \prod_{j<k}(1 - \beta_j)$ remains the same, and we have $\sum_{k=1}^{K_{\max}} w_k = 1$.

For the sampling algorithm it is convenient to include cluster assignment variables z_i to indicate which cluster $k \in \{1, \ldots, K_{\max}\}$ contains x_i. Then we can rewrite the model as

$$\begin{aligned} (z_i \mid G) &\overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(w_1, \ldots, w_{K_{\max}}) && i = 1, \ldots, N \\ (x_i \mid z_i = k, G) &\overset{\text{ind}}{\sim} F(\phi_k) && i = 1, \ldots, N \end{aligned}$$

Note that the joint distribution decomposes as

$$p(x, z, \beta, \phi) = \left[\prod_{k=1}^{K_{\max}} g_0(\phi_k)\right] \left[\prod_{k=1}^{K_{\max}-1} p(\beta_k)\right] \left[\prod_{i=1}^{N} p(z_i \mid \beta) p(x_i \mid z_i, \phi)\right].$$

We perform Gibbs sampling over the state space $\{\beta_k, \phi_k\}_{k=1}^{K_{\max}}$ and $\{z_i\}_{i=1}^{n}$. At each iteration, we choose one of these variables and re-sample it from its conditional distribution given all the other variables.

$$\begin{aligned} p(z_i = k \mid x, z_{-i}, \beta, \phi) &\propto p(z_i = k \mid \beta) p(x_i \mid z_i = k, \phi) \\ &= w_i f(x_i \mid \phi_k) && i = 1, \ldots, N \\ p(\phi_k \mid x, z, \beta, \phi_{-k}) &\propto g_0(\phi_k) \prod_{i=1}^{N} p(x_i \mid z_i, \phi) \\ &\propto g_0(\phi_k) \prod_{i : z_i = k} f(x_i \mid \phi_k) && k = 1, \ldots, K_{\max} \\ p(\beta_k \mid x, z, \beta_{-k}, \phi) &\propto p(\beta_k) \prod_{i=1}^{N} p(z_i \mid \beta) \\ &\propto (1 - \beta_k)^{\alpha - 1} \prod_{i : z_i \geq k} w_{z_i} \\ &\propto \beta_k^{n_k} (1 - \beta_k)^{n_{>k} + \alpha - 1}. && k = 1, \ldots, K_{\max} \end{aligned}$$

where $n_k = \sum_{i=1}^n \mathbf{1}[z_i = k]$ is the number of observations in cluster k, and $n_{>k} = \sum_{i=1}^n \mathbf{1}[z_i > k]$ is the number of observations in clusters $\ell > k$. The last step used the fact that $w_k \propto \beta_k$ and $w_\ell \propto 1 - \beta_k$ for $\ell > k$.

The conditional distribution for β_k is simply $\mathrm{Beta}(n_k + 1, n_{>k} + \alpha)$. The conditional of ϕ_k has a closed form if G_0 is conjugate to the likelihood, but if not, one can use an MCMC update like Metropolis Hastings.

The two main advantages for Gibbs sampling that we gain from using this model over the BM/CRP model are the following.

1. There is no integral, which diminishes the need for conjugacy of G_0 with the likelihood.
2. The decoupling (conditional indepedence) of the θ_i given G renders many parts of the sampling algorithm parallelizable. Specifically, the cluster assignments z can be updated in parallel. Also, the updates to β and ϕ can also be done in parallel.

3.4 The posterior Dirichlet process

In this section we state some results without proof. See §6.2 and appendices C and D of [1] for more detail.

Our Dirichlet process provides a discrete distribution over objects and take i.i.d. samples from this distribution. Analogous to the beta-binomial and Dirichlet-multinomial conjugacy, we suspect the posterior distribution of the Dirichlet process, after observe samples, is also a Dirichlet process. We will make this precise.

Suppose we have a partition $(A_1, \dots, A_K)$ of Θ. The vector $(\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K))$ is an indicator vector for the index k such that $\theta_i \in A_k$, and this event (conditioned on G) has probability $G(A_k)$. Thus, (conditioned on G) this vector is a categorical/multinoulli random variable with parameters $(G(A_1), \dots, G(A_K))$.

$$((\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K)) \mid G) \sim \mathrm{Cat}(G(A_1), \dots, G(A_K)).$$

Moreover, the random parameters follow a Dirichlet distribution [1, Appendix C]:

$$(G(A_1), \dots, G(A_K)) \sim \mathrm{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_K)), \tag{5}$$

for any partition $(A_1, \dots, A_K)$ of Θ. This is the primary reason for the name "Dirichlet process," and in fact the latter condition (5) is traditionally taken to be the *definition* of the Dirichlet process. The Dirichlet process borrows nice consistency properties from the Dirichlet distribution, such as $G(\Theta) = 1$ and the aggregation property

$$(G(A_1), \dots, G(A_i) + G(A_{i+1}), \dots, G(A_K)) \stackrel{d}{=} (G(A_1), \dots, G(A_i \cup A_{i+1}), \dots, G(A_K)).$$

Since the Dirichlet distribution is conjugate to the multinomial distribution, we have the posterior

$$\begin{aligned}&\big((G(A_1), \dots, G(A_K)) \mid \{(\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K)\}_{i=1}^N\big) \\ &\sim \mathrm{Dir}\left(\alpha G_0(A_1) + \sum_{i=1}^N \delta_{\theta_i}(A_1), \dots, \alpha G_0(A_K) + \sum_{i=1}^N \delta_{\theta_i}(A_K)\right)\end{aligned}$$

This result holds for any partition $(A_1, \dots, A_K)$, so the posterior process is a Dirichlet process

$$\big(G \mid \{(\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K)\}_{i=1}^N\big) \sim \mathrm{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^N \delta_{\theta_i}\right).$$

However, we want the posterior given $\theta_1, \dots, \theta_N$, not given $\{(\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K)\}_{i=1}^N$. A priori, θ could provide more information about G than $\{(\delta_{\theta_i}(A_1), \dots, \delta_{\theta_i}(A_K)\}_{i=1}^N$ since the latter information only gives information about which subset A_k contains θ_i and not the precise location of θ_i. However, this is not the case: the conditional distribution of G given θ is the same. This is known as the "tail-free" property of the Dirichlet process; see [1, Appendix D].

$$(G \mid \theta_1, \dots, \theta_N) \sim \mathrm{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^N \delta_{\theta_i}\right). \tag{6}$$

A final side remark: One can also write this posterior as a mixture of the prior Dirichlet process and point masses on the distinct points $\theta_1^*, \dots, \theta_K^*$ generated from G; see [1, §6.2].

$$(G \mid \theta) = w'G' + \sum_{k=1}^{K} w_k \delta_{\theta_k^*},$$

where

$$(w_1, \ldots, w_K, w') \sim \mathrm{Dir}(n_1, \ldots, n_K, \alpha)$$
$$(G' \mid \theta) \sim \mathrm{DP}(\alpha, G_0).$$

3.5 Back to the Blackwell-MacQueen urn and the Chinese Restaurant Process

Here we check that G is indeed the random measure we seek. If θ is generated according to (3) with $G \sim \mathrm{DP}(\alpha, G_0)$, then

$$\begin{aligned}
&P(\theta_{N+1} \in A \mid \theta_1, \ldots, \theta_N) \\
&= \mathbb{E}_{G \sim \mathrm{DP}(\alpha, G_0)}[P(\theta_{N+1} \in A \mid \theta_1, \ldots, \theta_N, G) \mid \theta_1, \ldots, \theta_N] && \text{tower property} \\
&= \mathbb{E}_{G \sim \mathrm{DP}(\alpha, G_0)}[P(\theta_{N+1} \in A \mid G) \mid \theta_1, \ldots, \theta_N] && \text{cond. indep. given } G \\
&= \mathbb{E}_{G \sim \mathrm{DP}(\alpha, G_0)}[G(A) \mid \theta_1, \ldots, \theta_N] && (\theta_{N+1} \mid G) \sim G \\
&= \frac{\alpha}{\alpha + N} G_0(A) + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}(A). && \text{posterior D.P., see (6)}
\end{aligned}$$

References

[1] Michael I. Jordan and Yee Whye Teh. A gentle introduction to the dirichlet process, the beta process, and bayesian nonparametrics. Draft, 2016.

[2] Jim Pitman. *Some developments of the Blackwell-MacQueen urn scheme*, volume Volume 30 of *Lecture Notes–Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, CA, 1996.