A Very Gentle Note on the Construction of Dirichlet Process

Xinhua Zhang Xinhua.Zhang@anu.edu.au

Research School of Information Sciences and Engineering
The Australian National University, Canberra ACT 0200, Australia

Statistical Machine Learning Program
National ICT Australia, Canberra, Australia

Contents

1. Introduction

We will not touch inference issues for Dirichlet processes. This note is based heavily on Teh (2007).

2. Preliminaries

In this section, we pile some fundamentals.

2.1 Exponential family and conjugate prior

Constituents of exponential family: $\mathcal{X}$: sample space; ϕ: $\mathbb{R}^n \mapsto \mathbb{R}^d$ sufficient statistics; $\theta \in \mathbb{R}^d$: natural parameter; $g(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$ log partition function. We will ignore the base measure for convenience: if you are particular, we are just assuming Lebesgue measure for continuously valued distributions and counting measure for discrete distributions.

Suppose the likelihood model is in exponential family:

$$p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta)), \tag{1}$$

where $g(\theta) := \int_{\mathcal{X}} \exp(\langle \phi(x), \theta \rangle) \mathrm{d}x$ takes care of normalization. Implicitly, we assume that θ are only from $\Theta := \left\{\theta | \int_{\mathcal{X}} \exp(\langle \phi(x), \theta \rangle) \mathrm{d}x < \infty\right\}$.

We endow a prior distribution $p(\theta)$, and are interested in the posterior distribution $p(\theta|x)$. It will be desirable if $p(\theta|x)$ is in the same exponential family as $p(\theta)$, because this will allow us to chain the observations $X := \{x_i\}_{i=1}^n$ and to update the posterior $p(\theta|X)$ successively. In other words, we can progressively absorb the information of the observation x_i, and update our belief over the parameter θ. This relationship between $p(x|\theta)$ and $p(\theta)$ is called *conjugate* which is formally defined as follows:

Definition 1 (Conjugate prior/distribution)**.** *Given a class of likelihood functions $p(x|\theta)$, a class of prior probability distributions $p(\theta)$ is said to be conjugate to the class of likelihood functions $p(x|\theta)$ if the resulting posterior distributions $p(\theta|x)$ are in the same family as $p(\theta)$.*

Remark 2. *For a given likelihood $p(x|\theta)$, it has a trivial family of conjugate priors: the whole set of all possible distributions. We are normally interested in the "smallest" conjugate prior, minimal in terms of the dimension of smooth parametrization. Also note that conjugate prior is not unique: in fact linear combinations of conjugate priors are still conjugate priors.*

If $p(x|\theta)$ is in the exponential family, then there is a constructive way to formulate the conjugate prior as follows, which is guaranteed to be minimal, as long as the original likelihood $p(x|\theta)$ is expressed in the minimal form (features in $\phi(x)$ are linearly independent).

Theorem 3. *If likelihood $p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$, then a conjugate prior of $p(x|\theta)$ is:*

$$p(\theta) = \exp(\langle \theta, w \rangle + \rho g(\theta) - h(w, \rho)), \tag{2}$$

where $h(w,\rho) := \int_\Theta \exp(\langle\theta, w\rangle + \rho g(\theta))\mathrm{d}\theta$ *takes care of normalization. And this conjugate prior is a minimal one.*

Generally, if $p(x|\theta) = \exp(\langle\phi(x), \psi(\theta)\rangle - g(\theta))$ *where* ψ *is* $\mathbb{R}^d \mapsto \mathbb{R}^d$*, then the prior is changed to*

$$p(\theta) = \exp(\langle\psi(\theta), w\rangle + \rho g(\theta) - h(w,\rho)). \tag{3}$$

The base measure of $p(x|\theta)$*, if any, does not affect the choice of prior.*

Proof. By simple calculation.

$$p(x,\theta) = p(\theta)p(x|\theta) = \exp(\langle\theta, \phi(x) + w\rangle + (\rho-1)g(\theta) - h(w,\rho)).$$

So

$$p(\theta|x) = \exp(\langle\theta, \phi(x) + w\rangle + (\rho-1)g(\theta) - l(x)) \tag{4}$$

where $l(x) := \int_\Theta \exp(\langle\theta, \phi(x) + w\rangle + (\rho-1)g(\theta))\mathrm{d}\theta$. Obviously, $p(\theta|x)$ in Eq. (4) is in the same class as $p(\theta)$ in Eq. (2). It is also easy to verify the conjugacy when θ is replaced by $\psi(\theta)$. We omit the proof of minimality. □

Remark 4. *This construction is built upon the log partition function* $g(\theta)$*. Usually there is no closed form of it and in general it is NP-hard to compute* $g(\theta)$ *numerically.*

More properties of conjugate prior, especially for exponential families, can be found at wikipedia and links therein.

2.2 Basic properties of Dirichlet Distribution

The Dirichlet distribution of order n is defined over the space of n-dimensional simplex

$$\Delta_n := \left\{x \in \mathbb{R}^n : \sum\nolimits_i x_i = 1, x_i \geq 0\right\}.$$

The distribution is parameterized by n positive parameters $\{\alpha_i\}_{i=1}^n$ $(a_i > 0)$.

Definition 5 (Dirichlet distribution)**.** *A random variable* $x \in \Delta_n$ *is said to have Dirichlet distribution if its probability density function with respect to Lebesgue measure is given by:*

$$p(x_1, \ldots, x_n) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i - 1} \tag{5}$$

and it is denoted as $x \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$ *or simply* $x \sim \mathrm{Dir}(\alpha)$*.*

Remark 6. *It is crucial to note that the above distribution is defined on Δ_n. This is in fact not a good definition. An equivalent and better approach is to define the distribution on $\Delta'_{n-1} := \left\{x \in \mathbb{R}^{n-1} : \sum_i x_i \le 1, x_i \ge 0\right\}$, and*

$$p(x_1, \dots, x_{n-1}) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^{n-1} x_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} x_i\right)^{\alpha_n - 1}. \tag{6}$$

Property 7 (Mean, covariance, mode, marginal)**.** *If $(x_1, \dots, x_n) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_n)$, then*

$$\mathbb{E}\left[(x_1, \dots, x_n)\right] = \left(\frac{\alpha_1}{s}, \dots, \frac{\alpha_n}{s}\right), \quad \text{where } s := \sum_i \alpha_i$$

$$\mathrm{Cov}[x_i x_j] = \frac{-\alpha_i \alpha_j}{s^2(s+1)}$$

$$\text{marginal distribution } x_i \sim \mathrm{Dir}(\alpha_i, \textstyle\sum_{j \ne i} \alpha_i)$$

$$\text{mode } (m_1, \dots, m_n) = \left(\frac{\alpha_1 - 1}{s - n}, \dots, \frac{\alpha_n - 1}{s - n}\right).$$

A well-known and important property of the Dirichlet distribution is that it is a conjugate prior of the multinomial distribution. A multinomial distribution is parametrized by a vector $(\theta_1, \dots, \theta_n)$, where $\sum_i \theta_i = 1$ and $\theta_i \ge 0$. The multinomial distribution of order m is defined over the set of $\{(x_1, \dots, x_n) : \sum_i x_i = m, x_i \in \mathbb{Z}, x_i \ge 0\}$. Then $p(x)$ has multinomial distribution if

$$p(x) = \frac{m!}{\prod_i x_i!} \prod_i \theta_i^{x_i}, \tag{7}$$

and is denoted as $x \sim \mathrm{Multi}(\theta_1, \dots, \theta_n)$ or $x \sim \mathrm{Multi}(\theta)$ in brief.

Remark 8. *Similar to Remark 6, here $\sum_i x_i$ must be strictly m, and an equivalent definition will be over $\left\{(x_1, \dots, x_{n-1}) : \sum_{i=1}^{n-1} x_i \le m, x_i \in \mathbb{Z}, x_i \ge 0\right\}$ by*

$$p(x_1, \dots, x_{n-1}) = \frac{m!}{\prod_{i=1}^{n-1} x_i! (m - \sum_{i=1}^{n-1} x_i)} \left(1 - \sum_{i=1}^{n-1} \theta_i\right)^{m - \sum_{i=1}^{n-1} x_i} \prod_{i=1}^{n-1} \theta_i^{x_i}. \tag{8}$$

Proposition 9. *Dirichlet distribution $\theta \sim Dir(\alpha)$ is a conjugate prior of $x|\theta \sim \mathrm{Multi}(\theta)$.*

Proof. Just by simple calculation.

$$p(\theta) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)} \prod_{i=1}^n \theta_i^{\alpha_i - 1}, \qquad p(x|\theta) = \frac{n!}{\prod_i x_i!} \prod_{i=1}^n \theta_i^{x_i}$$

So

$$p(x, \theta) \propto \frac{1}{\prod_i x_i!} \prod_{i=1}^n \theta_i^{x_i + \alpha_i - 1},$$

which implies that

$$p(\theta|x) \propto \prod_{i=1}^{n} \theta_i^{x_i+\alpha_i-1}.$$

i.e., $\theta|x \sim \mathrm{Dir}(x+\alpha)$. □

Remark 10. *It is also straightforward to carry out the proof by applying Theorem 3. First notice that the base measure of* Multi(x) *(likelihood) is* $\frac{1}{\prod_i x_i!}$*, and Theorem 3 does allow non-counting measure. However, the real problem is that there is a constraint that* $\sum_{i=1}^{n} x_i = m$*, so the exponential family representation in Eq. (7) is not really the whole story. The correct starting point of applying Theorem 3 is Eq. (8), which can be further rewritten into the canonical form of exponential families:*

$$p(x_1, \ldots, x_{n-1}) = \frac{m!}{\prod_{i=1}^{n-1} x_i! \left(m - \sum_{i=1}^{n-1} x_i\right)!} \exp\left(\sum_{i=1}^{n-1} x_i \log \frac{\theta_i}{1 - \sum_{i=1}^{n-1} \theta_i} + m \log\left(1 - \sum_{i=1}^{n-1} \theta_i\right)\right)$$

So the base measure is $\frac{m!}{\prod_{i=1}^{n-1} x_i! \left(m - \sum_{i=1}^{n-1} x_i\right)!}$*, sufficient statistics are* x_i*, natural parameters are* $\log \frac{\theta_i}{1-\sum_{i=1}^{n-1} \theta_i}$*, and the log partition function is* $-m \log\left(1 - \sum_{i=1}^{n-1} \theta_i\right)$*. By Theorem 3, a conjugate prior is:*

$$\begin{aligned} p(\theta) &\propto \exp\left(\sum_{i=1}^{n-1} \alpha_i \log \frac{\theta_i}{1 - \sum_{i=1}^{n-1} \theta_i} + \alpha_0 \log\left(1 - \sum_{i=1}^{n-1} \theta_i\right)\right) \\ &\propto \prod_{i=1}^{n-1} \theta_i^{\alpha_i} \left(1 - \sum_{i=1}^{n-1} \theta_i\right)^{\alpha_0 - \sum_{i=1}^{n-1} \alpha_i} \end{aligned}$$

which is a Dirichlet distribution by Definition in Eq. (6).

Below we collect two less known results, which are still pretty simple. To start with, we first introduce a very useful lemma as a construction of Dirichlet distributions.

Definition 11 (Gamma distribution)**.** *A random variable* x *is said to have Gamma distribution with shape* $k > 0$ *and scale* $\theta > 0$ *if its pdf satisfies:*

$$p(x) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)}$$

for $x > 0$*. And it is denoted as* $x \sim \mathrm{Gamma}(k, \theta)$*.*

Lemma 12 (Construction of Dirichlet distribution via Gamma distribution)**.** *If* $y_i \sim \mathrm{Gamma}(\alpha_i, 1)$ *are independent, then*

1. $v := \sum_{i=1}^{n} y_i \sim \mathrm{Gamma}(\sum_{i=1}^{n} \alpha_i, 1)$

2. $(x_1, \ldots, x_n) := (y_1/v, \ldots, y_n/v) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$.

Proof. By simple brutal-force calculation. □

Proposition 13 (Agglomeration)**.** *If* $(x_1, x_2, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_n)$*, then*

$$(x_1 + x_2, x_3, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_n). \tag{9}$$

And generally, if $(I_1, \ldots, I_s)$ *is a partition of* $\{1, \ldots, n\}$*, then*

$$\left(\sum_{i\in I_1} x_i, \ldots, \sum_{i\in I_s} x_i\right) \sim \mathrm{Dir}\left(\sum_{i\in I_1} \alpha_i, \ldots, \sum_{i\in I_s} \alpha_i\right). \tag{10}$$

Proof. Suppose $y_i \sim \mathrm{Gamma}(\alpha_i, 1)$ and let $s := \sum_{i=1}^n y_i$, $x_i := y_i/s$. Then $(x_1, \ldots, x_n) = (y_1/s, \ldots, y_n/s) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$. On the other hand, $y_1 + y_2 \sim \mathrm{Gamma}(\alpha_1 + \alpha_2, 1)$, hence

$$(x_1 + x_2, x_3, \ldots, x_n) = ((y_1 + y_2)/s, \ldots, y_n/s) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \ldots, \alpha_n).$$

Eq. (10) can be easily derived by repeatedly applying Eq. (9). □

Proposition 14 (Decimation)**.** *If* $(x_1, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$*, and* $(\tau_1, \tau_2) \sim \mathrm{Dir}(\alpha_1\beta_1, \alpha_1\beta_2)$ *where* $\beta_1 + \beta_2 = 1$*, then*

$$(x_1\tau_1, x_1\tau_2, x_2, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, \ldots, \alpha_n). \tag{11}$$

Proof. Let $z_1 \sim \mathrm{Gamma}(\alpha_1\beta_1, 1)$, $z_2 \sim \mathrm{Gamma}(\alpha_1\beta_2, 1)$, $y_i \sim \mathrm{Gamma}(\alpha_i, 1)$ for $i = 2, \ldots, n$. Let $s = z_1 + z_2 + \sum_{i=2}^n y_i$. Then

$$(z_1/s, z_2/s, y_2/s, \ldots, y_n/s) \sim \mathrm{Dir}(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, \ldots, \alpha_n). \tag{12}$$

Define $x_1 := (z_1 + z_2)/s$, $x_i := y_i/s$ for $i = 2, \ldots, n$, and $\tau_i := z_i/(z_1 + z_2)$ for $i = 1, 2$. Since $z_i \sim \mathrm{Gamma}(\alpha_1\beta_i, 1)$ for $i = 1, 2$, so $(\tau_1, \tau_2) \sim \mathrm{Dir}(\alpha_1\beta_1, \alpha_1\beta_2)$. Besides, since $z_1 + z_2 \sim \mathrm{Gamma}(\alpha_1\beta_1 + \alpha_1\beta_2, 1) = \mathrm{Gamma}(\alpha_1, 1)$ and $s = (z_1 + z_2) + \sum_{i=2}^n y_i$, so $(x_1, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$. In sum, our construction meets the preconditions.

Now notice $\frac{z_i}{s} = \frac{z_1+z_2}{s}\frac{z_i}{z_1+z_2} = x_1\tau_i$ for $i = 1, 2$. So substituting into Eq. (12), we arrive at Eq. (11). □

Remark 15. *As a simple generalization, we can prove, by using exactly the same procedure, that if* $(\tau_1, \ldots, \tau_m) \sim \mathrm{Dir}(\alpha_1\beta_1, \ldots, \alpha_1\beta_m)$ *with* $\sum_{i=1}^m \beta_i = 1$*, then*

$$(x_1\tau_1, \ldots, x_1\tau_m, x_2, \ldots, x_n) \sim \mathrm{Dir}(\alpha_1\beta_1, \ldots, \alpha_1\beta_m, \alpha_2, \ldots, \alpha_n).$$

Even more generally, we can split many x_i*'s into (different) fractions.*

3. Theoretical Definition of Dirichlet Process

Dirichlet process (DP) is a distribution over distributions. For a random distribution G to be distributed according to a DP, its marginal distributions have to be Dirichlet distributed. This is similar to the definition of Gaussian processes.

Definition 16 (Dirichlet Process by Ferguson (1973))**.** *Let H be a distribution over Θ and α be a positive real number. Then for any finite measurable partition $A_1, \ldots, A_r$ of Θ, the vector $(G(A_1), \ldots, G(A_r))$ is random since G is random. We say G is a Dirichlet process distributed with base distribution H and concentration parameter α, written $G \sim DP(\alpha, H)$, if*

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)) \tag{13}$$

for every finite measurable partition $A_1, \ldots, A_r$ of Θ.

Property 17. *For any measurable set A, $\mathbb{E}[G(A)] = H(A)$, $\mathrm{Var}[G(A)] = \frac{H(A)(1-H(A))}{\alpha+1}$.*

Proof. For any measurable set A, by Eq. (13),

$$(G(A), G(A^c)) \sim \mathrm{Dir}(\alpha H(A), \alpha H(A^c)) = \mathrm{Dir}(\alpha H(A), \alpha - \alpha H(A))$$

By Property 7, we have

$$\begin{aligned}
\mathbb{E}\left[G(A)\right] &= \frac{\alpha H(A)}{\alpha H(A) + \alpha - \alpha H(A)} = H(A) \\
\mathrm{Var}\left[G(A)\right] &= \frac{\alpha H(A) \cdot (\alpha - \alpha H(A))}{\alpha^2(\alpha+1)} = \frac{H(A)(1-H(A))}{\alpha+1}.
\end{aligned}$$

So the larger α is, the smaller the variance. □

As $\alpha \to \infty$, we have $G(A) \to H(A)$ for any measurable set A, that is $G \to H$ weakly or pointwise. However, this is not equivalent to saying that $G \to H$. As will be shown later, draws from a DP will be discrete distributions with probability 1, even if H is smooth. Thus G and H need not even be absolutely continuous with respect to each other. If smoothness is a concern, one can extend the DP by convolving G with kernels.

3.1 Posterior distributions

Since G is a distribution randomly drawn from DP and is unobservable, we can make better and better estimates of G by drawing samples from G. This is the idea of posterior distributions. Suppose we have observed values $\theta_1, \ldots, \theta_n$. Let $A_1, \ldots, A_r$ be a finite measurable partition of Θ, and let $n_k = \#\{i : \theta_i \in A_k\}$ be the number of observed values in A_k. By Eq. (13), we have $(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$ as prior. The likelihood model is multinomial

because we have partitioned the whole space Θ into a fixed set of subsets and $r < \infty$. By conjugacy of Dirichlet and multinomial distributions, we have

$$(G(A_1), \ldots, G(A_r))|\theta_1, \ldots, \theta_n \sim \text{Dir}(\alpha H(A_1) + n_1, \ldots, \alpha H(A_r) + n_r) \tag{14}$$

Since Eq. (14) is true for all finite measurable partitions, we can guess that the posterior over G is also a DP. Notice the parameters has constant sum: $\sum_{i=1}^{n} \alpha H(A_r) + n_r = \alpha + n$ is constant (does not depend on the partition). This reminds that if we can find a distribution H' and positive real number α' such that for all partitions, $\alpha H(A_i) + n_i = \alpha' H'(A_i)$ for all $i = 1, \ldots, r$, then $G|\theta_1, \ldots, \theta_n$ must also a Dirichlet process. Fortunately, it is easy to see that the posterior DP has updated concentration parameter to $\alpha' = \alpha + n$ and base distribution $H' = \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$, where δ_{θ_i} is point mass located at θ_i (atom). In other words,

$$G|\theta_1, \ldots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right). \tag{15}$$

Now we study the predictive distribution θ_{n+1} after observing $\theta_1, \ldots, \theta_n$ and marginalizing out G:

$$p(\theta_{n+1}) = \int_G p(\theta_{n+1}, G|\theta_1, \ldots, \theta_n).$$

Notice $\theta_{n+1} \perp\!\!\!\perp \theta_1, \ldots, \theta_n | G$. For all measurable set A, we have

$$\Pr(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) = \mathbb{E}\left[G(A)|\theta_1, \ldots, \theta_n\right] = \frac{\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)}{\alpha + n}, \tag{16}$$

where the first equality is by definition and second equality follows from the posterior base distribution of G. Since Eq. (16) holds for arbitrary A, we have

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}. \tag{17}$$

Now we can see that the posterior base distribution given $\theta_1, \ldots, \theta_n$ is also the predictive distribution of θ_{n+1}. This is not that surprising, given the fact that the expectation of $G|\theta_1, \ldots, \theta_n$ equals the posterior base distribution as Property 17 states.

Now a fundamental question arises: is the Definition 16 reasonable, i.e., does there exist such a stuff which satisfies the definition. There are two ways to prove the well-definedness. One is by showing the exchangeability of the process and then resort to the de Finetti's theorem. Another approach, which is much more direct, is to construct a process explicitly and prove that it satisfies the conditions in Definition 16. We start from the first approach.

4. Pólya Urn Scheme

Eq. (17) provides a convenient way to draw samples from G, though G is not observable in its own right. This is the so-called Pólya urn scheme.

Pólya Urn Scheme Suppose each value in Θ is a unique color, and draws $\theta \sim G$ are balls with the drawn value being the color of the ball. In addition we have an urn containing previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from H, i.e. draw $\theta_1 \sim H$, paint a ball with that color, and drop it into the urn. In subsequent steps, say the $(n+1)$st, we will either, with probability $\frac{\alpha}{\alpha+n}$, pick a new color (draw $\theta_{n+1} \sim H$), paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{\alpha+n}$, reach into the urn to pick a random ball out (draw θ_{n+1} from the empirical distribution), paint a new ball with the same color and drop both balls back into the urn.

Remark 18. *One tricky part of the story is the space of color and the distribution H over it. A crucial assumption needed in the analysis is that the probability of two independent random draws from H having the same value be 0. So it definitely does not mean that we have only three colors "red, green, blue" and a discrete distribution over it. Instead, one may assume that color is in $[0,1]^3$ of RGB space and is uniformly distributed.*

It is important to interpret the probabilistic mechanism of drawing balls. Since the values of draws (ball colors) $\{\theta_k\}$ are repeated, let $\theta_1^*, \ldots, \theta_m^*$ be the unique values among $\theta_1, \ldots, \theta_n$, and n_k be the number of repeats of θ_k^*. Then the predictive distribution Eq. (17) can be equivalently rewritten as:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha+n}\left(\alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*}\right).$$

Notice that value θ_k^* will be repeated by θ_{n+1} with probability proportional to n_k, the number of times it has already been observed. The larger n_k is, the higher the probability that it will grow. This is a rich-gets-richer phenomenon, where large clusters (a set of θ_i's with identical values θ_k^* being considered a cluster) grow ever larger. This is both good and bad. On the good side, this leads to the fact that draws G from $\mathrm{DP}(\alpha, H)$ are discrete with probability 1. The number of new colors can also be shown to grow logarithmically in the number of draws. However, on the other side, it may also limit the capacity of this model. Finally, notice that the rich-gets-richer has nothing to do with the base measure H, which is assumed to be smooth (continuous) in most cases.

Also notice that the colors define a partition of $\{1, \ldots, n\}$, or more profoundly, a random permutation. We don't go into the details here.

4.1 Application of de Finetti's Theorem

The Pólya urn scheme has been used to show the existence of DP by Blackwell and MacQueen (1973). The basic tool is the de Finetti's theorem, which guarantees a mixture model provided that the observations are exchangeable, which is clearly satisfied in the Pólya urn scheme. And this mixture model is exactly the DP, hence justifies the definition.

Definition 19 (Exchangeability of random sequences)**.** *A random process $(\theta_1, \theta_2, \ldots)$ is called infinitely exchangeable if for any $n \in \mathbb{N}$ and any permutation σ on $1, \ldots, n$, the probability of generating $(\theta_1, \ldots, \theta_n)$ is equal to the probability of drawing them in a different order $(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)})$:*

$$P(\theta_1, \ldots, \theta_n) = P(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)}).$$

Theorem 20 (de Finetti's Theorem)**.** *Suppose a random process* $(\theta_1, \theta_2, \ldots)$ *is infinitely exchangeable, then the joint probability* $p(\theta_1, \theta_2, \ldots, \theta_N)$ *has a representation as a mixture:*

$$p(\theta_1, \theta_2, \ldots, \theta_N) = \int \left(\prod_{i=1}^{N} G(\theta_i) \right) \mathrm{d}P(G),$$

for some random variable G*.*

Starting from the definition of the Pólya urn scheme, especially Eq. (17), we can construct a distribution over sequences $\theta_1, \theta_2, \ldots$ by iteratively drawing each θ_i given $\theta_1, \ldots, \theta_{n-1}$ by Eq. (17). Notice that the conditional probability Eq. (17) is always well-defined regardless of whether DPs exist. For $n \geq 1$, let

$$p(\theta_1, \ldots, \theta_n) := \prod_{i=1}^{n} p(\theta_i | \theta_1, \ldots, \theta_{i-1})$$

be the joint distribution over the first n observations where $p(\theta_i | \theta_1, \ldots, \theta_{i-1})$ is given by Eq. (17). It is straightforward to verify that this random sequence is infinitely exchangeable. In fact, if there are C colors (i.e., the action of picking a new color $\theta_{n+1} \sim H$ occurred for C times), and n_c balls are drawn for each color θ_c^*, then the statistics $\{n_c\}_c$ do not change with the permutation of $(\theta_1, \ldots, \theta_n)$. Moreover, $p(\theta_1, \ldots, \theta_n)$ depends only on $\{n_c\}_c$ by:

$$p(\theta_1, \ldots, \theta_n) = \frac{\alpha^C \prod_{c=1}^{C} H(\theta_c^*)(n_c - 1)!}{(\alpha + n - 1)(\alpha + n - 2) \ldots \alpha}, \quad \text{where } n := \sum\nolimits_{i=1}^{C} n_c.$$

Now de Finetti's theorem states that there exists a prior over the random distributions, $P(G)$. And this $P(G)$ is exactly the Dirichlet process $\mathrm{DP}(\alpha, H)$, thus establishing existence.

5. Chinese Restaurant Process

It is also well known that DP can also be represented as Chinese restaurant process (CRP). This is evident from the Pólya Urn Scheme.

Chinese restaurant process. Suppose we have a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The first customer enters the restaurant and sits at the first table. The second customer enters and decides either to sit with the first customer, or by herself at a new table. In general, the $(n+1)$st customer either joins an already occupied table k with probability proportional to the number n_k of customers already sitting there, or sits at a new table with probability proportional to α. Identifying customers with integers $1, 2, \ldots$ and tables as clusters, after n customers have sat down the tables define a partition of $[n]$ with the distribution over partitions being the same as the one above. The fact that most Chinese restaurants have round tables is an important aspect of the CRP. This is because it does not just define a distribution over partitions of $[n]$, it also defines a distribution over permutations of $[n]$, with each table corresponding to a cycle of the permutation.

It is easy to establish the correspondence between the Pólya urn scheme and CRP. For example, opening a new table corresponds to drawing a new color of ball, and the relevant probabilities are also clearly the same. The only superficial difference is that the colors in Pólya urn scheme are drawn randomly from H, and opening a new table looks like a deterministic action. So when using CRP as metaphor, sometimes people associate a dish θ_k^* to each table k, and that dish is drawn *iid* from H (well I appreciate people's imaginativeness).

It is interesting to consider the expected number of tables (clusters) among the n customers (observations). Notice that for $i \geq 1$, the probability that the k customer takes on a new table is $\alpha/(\alpha + k - 1)$. Thus the average number of tables m is:

$$\mathbb{E}[m|n] = \sum_{k=1}^{n} \frac{\alpha}{\alpha + k - 1} \in O(\alpha \log n).$$

That is, the number of clusters grows only logarithmically in the number of observations. This slow growth makes sense because of the rich-gets-richer phenomenon. Besides, larger α implies a larger number of clusters a priori.

6. Stick-breaking Construction

It is already intuited that draws from a DP are composed of a weighted sum of point masses. Sethuraman (1994) made this precise by providing a constructive definition of the DP as such, called the stick-breaking construction. This construction is also significantly more straightforward and general than previous proofs of the existence of DPs. It is simply given as follows:

$$\begin{array}{lcllcl} \beta_k & \sim & \text{Beta}(1, \alpha) & \theta_k^* & \sim & H \\ \pi_k & = & \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) & G & = & \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}. \end{array}$$

Table 1: Stick-breaking construction of Dirichlet process.

Then $G \sim \text{Dir}(\alpha, H)$. The construction of π can be understood metaphorically as follows. Starting with a stick of length 1, we break it at β_1, assigning π_1 to be the length of stick we just broke off. Now recursively break the other portion to obtain π_2, π_3 and so forth. The stick-breaking distribution over π is sometimes written $\pi \sim \text{GEM}(\alpha)$, where the letters stand for Griffiths, Engen and McCloskey. Because of its simplicity, the stick-breaking construction has lead to a variety of extensions as well as novel inference techniques for the Dirichlet process.

It is definitely not trivial to bridge this stick-breaking representation to the other three definitions of DP presented in the previous sections. We attempt to do that in one direction: deriving the stick-breaking construction from the original Definition 16.

6.1 From Dirichlet Process to Stick-breaking

In this section, we want to start from the original Definition 16 and motivate the stick-breaking scheme.

Proposition 21. *Suppose G is drawn from $\mathrm{DP}(\alpha, H)$, i.e., for any arbitrary partition $\{A_1, \ldots, A_r\}$ of Θ,*

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)),$$

then G can be equivalently drawn from the stick-breaking scheme in Table 1.

Before presenting the proof, we need a small lemma.

Lemma 22. *Let $(x, 1-x) \sim \mathrm{Dir}(1, \alpha)$, and $(x, (1-x)y_1, \ldots, (1-x)y_n) \sim \mathrm{Dir}(1, \alpha\beta_1, \ldots, \alpha\beta_n)$, where $\sum_i \beta_i = 1$. Then*

$$(y_1, \ldots, y_n) \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_n). \tag{18}$$

Proof. Let independent random variables $z \sim \mathrm{Gamma}(1, 1)$, $w_i \sim \mathrm{Gamma}(\alpha\beta_i, 1)$ for $i = 1, \ldots, n$. So $\sum_i w_i \sim \mathrm{Gamma}(\sum_i \alpha\beta_i, 1) = \mathrm{Gamma}(\alpha, 1)$.

Define: $s := z + \sum_i w_i$, $x := \frac{z}{s}$, $y_i := \frac{w_i}{\sum_i w_i}$. Then we have

$$(x, 1-x) = \left(\frac{z}{s}, \frac{\sum_i w_i}{s}\right) \sim \mathrm{Dir}(1, \alpha).$$

Notice $(1-x)y_i = \frac{\sum_i w_i}{s} \cdot \frac{w_i}{\sum_i w_i} = \frac{w_i}{s}$. So

$$(x, (1-x)y_1, \ldots, (1-x)y_n) = \left(\frac{z}{s}, \frac{w_1}{s}, \ldots, \frac{w_n}{s}\right) \sim \mathrm{Dir}(1, \alpha\beta_1, \ldots, \alpha\beta_n).$$

So all the preconditions are met. Finally, check the conclusion:

$$(y_1, \ldots, y_n) = \left(\frac{w_1}{\sum_i w_i}, \ldots, \frac{w_n}{\sum_i w_i}\right) \sim \mathrm{Dir}(\alpha\beta_1, \ldots, \alpha\beta_n). \qquad \square$$

Now we prove Proposition 21.

Proof. We know that the following facts are equivalent:

$$\begin{array}{rcl} G & \sim & \mathrm{DP}(\alpha, H) \\ \theta|G & \sim & G \end{array} \quad \Leftrightarrow \quad \begin{array}{rcl} \theta & \sim & H \\ G|\theta & \sim & \mathrm{DP}\left(\alpha+1, \frac{\alpha H + \delta_\theta}{\alpha+1}\right) \end{array}. \tag{19}$$

For notational convenience, we denote $G|\theta$ as $G_{|\theta}$. After drawing a sample θ from G, we consider the partition $\{\{\theta\}, \Theta \setminus \{\theta\}\}$ of Θ. By Eq. (19) and Definition 16, we have

$$\begin{aligned} (G_{|\theta}(\theta), G_{|\theta}(\Theta \setminus \theta)) &\sim \mathrm{Dir}\left((\alpha+1)\frac{\alpha H + \delta_\theta}{\alpha+1}(\theta), (\alpha+1)\frac{\alpha H + \delta_\theta}{\alpha+1}(\Theta \setminus \theta)\right) \\ &= \mathrm{Dir}(1, \alpha) \end{aligned} \tag{20}$$

Hence $G_{|\theta}$ has a point mass located at θ:

$$G_{|\theta} = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \text{Beta}(1,\alpha) \text{ according to Eq. (20)} \tag{21}$$

and G' is the (renormalized) probability measure with the point mass removed. To further derive the expression of G', we consider a further partition $\{\theta, A_1, \ldots, A_r\}$ of Θ. By Eq. (19) and Definition 16, we have:

$$(G_{|\theta}(\theta), G_{|\theta}(A_1), \ldots, G_{|\theta}(A_r)) \sim \text{Dir}(1, \alpha H(A_1), \ldots, \alpha H(A_r)). \tag{22}$$

On the other hand, Eq. (21) implies that

$$(G_{|\theta}(\theta), G_{|\theta}(A_1), \ldots, G_{|\theta}(A_r)) = (\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_r)). \tag{23}$$

Combining Eq. (22) and Eq. (23), we arrive at:

$$(\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_r)) \sim \text{Dir}(1, \alpha H(A_1), \ldots, \alpha H(A_r)). \tag{24}$$

Now applying Lemma 22 to Eq. (24), we have

$$(G'(A_1), \ldots, G'(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)),$$

which implies that

$$G' \sim \text{DP}(\alpha, H). \tag{25}$$

Eq. (25) in conjunction with Eq. (21) allows us to telescope and absorb all the observations successively and update the posterior distribution of G. By recursively applying

$$G \sim \text{DP}(\alpha, H) \qquad \text{and} \qquad G_{|\theta_1} = \beta_1\delta_{\theta_1} + (1-\beta_1)\text{DP}(\alpha, H),$$

we conclude that

$$G_{|\theta_1,\ldots,\theta_n,\ldots} = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k},$$

where

$$\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i), \qquad \beta_k \sim \text{Beta}(1,\alpha), \qquad \theta_k \sim H.$$

This is exactly the stick-breaking scheme in Table 1. □

6.2 From Stick-breaking to Dirichlet Process

In this section, we wish to show the opposite direction of Section 6.1. Formally, we want to show:

Proposition 23. *Suppose G is constructed by the stick-breaking scheme in Table 1, then G is actually drawn from $\text{DP}(\alpha, H)$, i.e., for any arbitrary partition $\{A_1, \ldots, A_r\}$ of Θ,*

$$(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Proof. This is not trivial and the proof can be found in Theorem 3.4 of Sethuraman (1994). So far, we have not been able to find a simple proof. □

7. Infinite Mixture Model

Another interpretation of DP is the infinite mixture model, which is also the most common application of DP. Here the nonparametric nature of the DP translates to mixture models with a countably infinite number of components. We model the set of observations $\{x_1, \ldots, x_n\}$ using a set of latent parameters $\{\theta_1, \ldots, \theta_n\}$. Each θ_i is drawn independently and identically from G, while each x_i has distribution $F(\theta_i)$ parametrized by θ_i:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G|\alpha, H &\sim DP(\alpha, H) \end{aligned} \tag{26}$$

Because G is discrete, multiple θ_i's can take on the same value simultaneously, and the above model can be seen as a mixture model, where x_i's with the same value of θ_i belong to the same cluster. The mixture perspective can be made more in agreement with the usual representation of mixture models using the stick-breaking construction Table 1. Let z_i be a cluster assignment variable, which takes on value k with probability π_k. Then Eq. (26) can be equivalently expressed as with

$\pi\|\alpha$	$\sim$	GEM(α)	$\theta_k^*\|H$	$\sim$	H
$z_i\|\pi$	$\sim$	Multi(π)	$x_i\|z_i, \{\theta_k^*\}$	$\sim$	$F(\theta_{z_i})$.

Table 2: Stick-breaking construction of mixture model.

$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ and $\theta_i = \theta_{z_i}^*$ (now you see why we used θ_k^* in Table 2). In mixture modeling terminology, π is the mixing proportion, θ_k^* are the cluster parameters, $F(\theta_k^*)$ is the distribution over the data in cluster k, and H is the prior over cluster parameters.

So DP model can be viewed as an infinite mixture model: a model with a countably infinite number of clusters. However, because the π_k's decrease exponentially quickly, only a small number of clusters will be used to model the data a priori (in fact, as we saw previously, the expected number of components used a priori is logarithmic in the number of observations). This is different than a finite mixture model, which uses a fixed number of clusters to model the data. In the DP mixture model, the actual number of clusters used to model data is not fixed, and can be automatically inferred from data using the usual Bayesian posterior inference framework.

Acknowledgements

The author thanks Marcus Hutter for very helpful comments.

References

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 1973.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:693–650, 1994.

Y. W. Teh. Dirichlet processes. Submitted to Encyclopedia of Machine Learning, 2007.