On the Dirichlet Process as a Prior in Bayesian Nonparametrics

Philipp Burckhardt
Heinz College, Department of Statistics
Carnegie Mellon University
Pittsburgh, USA
pgb@andrew.cmu.edu

1 Introduction

In machine learning and statistics, Bayesian methods offer an alternative to frequentist procedures and have proven to be useful in many scenarios in which prior information can be incorporated into the analysis (e.g. probabilistic expert systems) and domains such as natural language processing because of the built-in regularization of Bayesian techniques.

However, only recently Bayesian techniques have started to become widely adopted in nonparametric problems. Since the Bayesian machinery offers a coherent framework to update one's belief in face of incoming data, nonparametric models with infinite-dimensional parameter spaces are very appealing in comparison to parametric models with their fixed size of parameters: Since a Bayesian nonparametric model will adapt the number of used parameters depending on the observed data such that the model complexity is allowed to grow as more data comes in. For example, in the non-parametric version of a Gaussian mixture model, the Dirichlet Process Mixture Model (DPM), the number of clusters grows as the sample size increases.

In his seminal paper, Ferguson (1973) identifies the problem of devising practical prior distributions on the parameter space as the main reason why the Bayesian approach has not been as successful in treating nonparametric problems. According to Ferguson, a nonparametric prior distribution has to fulfill two goals: First, the support of the prior should be large and second, the posterior distribution should be tractable analytically. A class of distributions which address both of these opposed goals were discovered by Ferguson (1973) who called them Dirichlet process priors. Today, Dirichlet processes are the de facto standard prior for many nonparametric problems involving categorical variables, whereas Gaussian processes (GP) are often used for continuous variables. The reason why Dirichlet processes are appealing in categorical problems is that samples from the Dirichlet process are discrete distributions with probability 1 (Blackwell, 1973).

However, the class of all possible prior processes is much larger than Gaussian processes and Dirichlet processes, and the interested reader is referred to Phadia (2013) for an exhaustive overview.

In this article, we will exclusively deal with the Dirichlet process and its application in clustering. Probably the best way to introduce the Dirichlet process is to first cover in depth the Dirichlet distribution (the finite-dimensional equivalent of the Dirichlet process) as many properties carry over to the infinite-dimensional case.

1.1 The Dirichlet Distribution

The Dirichlet distribution is very popular in Bayesian statistics in models involving categorical data due to its conjugacy to the multinomial family of distributions. Recall that a prior distribution is *conjugate* to a likelihood function if the posterior distribution given the data lies in the same family of distributions as the prior.

The Dirichlet is often described as a distribution over distributions, as draws from a Dirichlet are distributions over a discrete probability space.

To properly define the Dirichlet distribution, we have to introduce some notation: Let $\mathbf{g} = \{g_1, \ldots, g_k\}$ be a discrete probability distribution on the space $\boldsymbol{\theta} = \{\theta^1, \ldots, \theta^k\}$ with random variable Θ defined on $\boldsymbol{\theta}$ which takes the value θ^k with probability $P(\Theta = \theta^k) = g_i$. Prior beliefs about $\mathbf{g}$ can be encoded by placing a Dirichlet prior on the distribution $\mathbf{g}$. As a running example, let us consider the case of a factory which produces loaded six-sided dice - to be sold to people engaged in fraudulent gambling[1]. In this case, the sample space is $\boldsymbol{\theta} = \left\{\theta^1 = 1, \ldots, \theta^6 = 6\right\}$, i.e. θ^i are the labels on the sides of the die. A fair die would have $\mathbf{g} = \left\{\frac{1}{6}, \ldots, \frac{1}{6}\right\}$. The production of a single die equals a draw from the prior distribution placed on $\mathbf{g}$. Since our factory produces loaded dice, the probabilities g_i of each individual draw will differ from the fair die. This can be encoded by two parameters in the Dirichlet distribution $\mathrm{Dir}(\alpha_0, \boldsymbol{\alpha})$[2], a base distribution $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ which gives the mean values for the elements in $\mathbf{g}$ ($\mathbb{E}(g_k) = a_k$) and a concentration parameter α_0 which specifies how much each draw from $\mathrm{Dir}(\alpha_0, \boldsymbol{\alpha})$ varies around $\boldsymbol{\alpha}$. In this parametrization, the density of the Dirichlet distribution can then be written as

$$P(\mathbf{g}; \alpha_0, \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} g_i^{a_0 a_k - 1}. \tag{1}$$

The Dirichlet distribution satisfies the following important property:

Theorem 1 (Agglomerative Property). *If Θ is distributed as $\mathrm{Dir}(\alpha_0, \boldsymbol{\alpha})$, then for any partition $B_1, \ldots, B_m$ of $\boldsymbol{\theta}$ the vector $P(B_1), P(B_2), \ldots, P(B_m)) = \left(\sum_{i \in B_1} g_i, \ldots, \sum_{i \in B_m} g_i\right)$ is distributed as $\mathrm{Dir}\left(\alpha_0, \left(\sum_{i \in B_1} a_i, \ldots, \sum_{i \in B_m} a_i\right)\right)$.* □

The agglomerative property of the Dirichlet shows that combining the atoms of the probability vector preserves the distributional assumption. In our dice example, we could define $B_1 = \left\{\theta^1, \theta^3, \theta^5\right\}$ and $B_2 = \left\{\theta^2, \theta^4, \theta^6\right\}$ such that B_1 stands for all odd numbers and B_2 for all even numbers a die can land on. Then B_1 and B_2 clearly form a partition of the sample space $\boldsymbol{\theta}$ as $B_1 \cup B_2 = \boldsymbol{\theta}$ and $B_1 \cap B_2 = \emptyset$. The vector $(P(B_1), P(B_2))$ would then be again Dirichlet distributed with concentration parameter α_0 and base distribution $(\sum_{i \in B_1} a_i, \sum_{i \in B_2} a_i)$. In case that all $\alpha_i = \frac{1}{6}$, we would thus have $\left(\frac{1}{2}, \frac{1}{2}\right)$, meaning that *on average* the produced dice have equal probability of showing even and odd numbers (recall that we still operate a factory producing loaded dice, such that each individual draw from $\mathbf{g}$ will be a die with unequal probabilities for each side, as drawn from the Dirichlet).

Taking a different perspective, we can view the $\boldsymbol{\alpha}$ in terms of the non-negative set-function $\alpha(A) = \sum_{i \in A} \alpha_i$ over sample space $\boldsymbol{\theta}$. Given this perspective, it is easy to show that $\alpha(A)$ fulfills all requirements of a measure [3].

This view will be particularly helpful when transitioning to the Dirichlet process. The Dirichlet distribution in Theorem 1 can hence equivalently be described as $\mathrm{Dir}(\alpha_0, (\alpha(B_1), \ldots, \alpha(B_m)))$. The following property of the Dirichlet distribution is stated here for future use:

Theorem 2 (Tail Free Property). *Let $B_1, \ldots, B_m$ be a partition of $\boldsymbol{\theta}$. For $i = 1, 2, \ldots, m$ with $\alpha(B_i) > 0$, let $P(\cdot \mid B_i)$ be the conditional probability given B_i defined by*

$$P(\theta^j \mid M_i) = \frac{P(\theta^j)}{P(M_i)} \text{ for } \theta^j \in B_i.$$

[1] This scenario is adapted from Tresp (2007). Although the example might look like contrived not to mention the moral issues at stake, it neatly captures the main points we are going to make.

[2] This is a slightly unconventional parametrization we picked up from a highly recommended video lecture on Bayesian nonparametrics by Tresp (2007). The reason for parametrization will become clear when we outline the connections to the Dirichlet process. Compared to the usual parametrization with parameters $a_1^\star, \ldots, a_k^\star$, we define $a_0 = \sum_{i=1}^{k} a_i^\star$ and $a_i = \frac{a_i^\star}{a_o}$ $\forall i \in \{1, \ldots, k\}$.

[3] Recall that a measure μ is a non-negative set function defined on measurable space $(\Omega, \mathcal{F})$ such that $\mu(\emptyset) = 0$ and for every sequence $\{B_m\}_{m=1}^{\infty}$ of mutually disjoint elements of $\mathcal{F}$ we have $\mu(\cup_{m=1}^{\infty} B_m) = \sum_{m=1}^{\infty} \mu(B_m)$ (see Ash and Doléans-Dade (2000)).

1.2 Dirichlet-Multinomial Sampling

Previously, we calculated the probability mass function for a single loaded dice as $P\left(\Theta = \theta^k\right) = g_k$. Due to increasing demand for loaded dice, we instead consider the likelihood of a whole sample $\Theta_1, \ldots, \Theta_n$. Let us denote with N_k the number of times a loaded dice lands on side θ^k. The likelihood function is then given by

$$P(\Theta_1, \ldots, \Theta_n \mid \mathbf{g}) = \prod_{k=1}^{K} g_k^{N_k}. \tag{2}$$

Consequently, the posterior distribution of $\mathbf{g}$ evaluates to

$$P(\mathbf{g} \mid \Theta_1, \ldots, \Theta_n) \propto P(\Theta_1, \ldots, \Theta_n \mid \mathbf{g}) P(\mathbf{g} | \alpha_0, \boldsymbol{\alpha}) \tag{3}$$

$$= \prod_{k=1}^{K} g_k^{\alpha_0 \alpha_k + N_k - 1}, \tag{4}$$

which we recognize as the kernel of a Dirichlet distribution. Thus,

$$\mathbf{g} \mid \Theta_1, \ldots, \Theta_n \sim \mathrm{Dir}\left(\alpha_0 + n, \frac{1}{\alpha_0 + n}\left(\alpha_0 \boldsymbol{\alpha} + \sum_{k=1}^{K} N_k \delta_{\theta^k}\right)\right), \tag{5}$$

where δ_{θ^k} is a degenerate distribution centered at θ^k.

Observe that the base distribution of the posterior is a mixture distribution

$$\frac{\alpha_0}{\alpha_0 + n} \boldsymbol{\alpha} + \sum_{k=1}^{K} \frac{N_k}{\alpha + n} \delta_{\theta^k},$$

where with probability $\frac{\alpha_0}{\alpha_0 + n}$ the new observation is drawn from the base distribution $\boldsymbol{\alpha}$ and with probability proportional to the number of observations $\Theta_1, \ldots, \Theta_n$ assigned to class k.

This formula for the predictive distribution given a sample $\Theta_1, \ldots, \Theta_n$ will make a reappearance in our treatment of the Dirichlet process and might serve as a reminder that the Dirichlet process is in essence just a generalization of the Dirichlet distribution for infinite-dimensional sample spaces where the discrete distribution $\boldsymbol{\alpha}$ is replaced by a continuous base measure G_0. In the Dirichlet process setting, the predictive distribution is visualized via the metaphor of customers entering a Chinese restaurant, the so-called Chinese Restaurant Process (CRP), which is equivalent to to the Polya Urn scheme, both of which we introduce in the next section.

2 The Dirichlet Process

2.1 Of Polya Urns and the Chinese Restaurant Process (CRP)

Consider the following scenario: We have an urn with α_0 balls of which α_0 / K balls have the color k, where the color $k \in \{1, \ldots, K\}$ is deterministically linked to θ^k. We now draw a sample of balls $c_1, \ldots, c_n$ from the urn according to the following rule: We draw the first ball at random from the

urn. After drawing each ball, we put the ball itself and an exact copy with the same color back into the urn. The probabilities at each iteration are

$$\begin{aligned} P(c_1 = k) &= \frac{a_0/K}{a_0} = \frac{1}{K} \\ P(c_2 = k|c_1) &= \frac{a_0/K + \delta(c_1 = k)}{a_0 + 1} \\ &\vdots \\ P(c_n = k|c_1, \ldots, c_{n-1}) &= \frac{a_0/K + \sum_{i=1}^{n-1} \delta(c_i = k)}{a_0 + n - 1} \end{aligned} \tag{6}$$

Notice that the probabilities at each iteration equal the predictive distributions obtained from Dirichlet-multinomial sampling under a $\mathrm{Dir}(a_0, (1/K, \ldots, 1/K))$ prior. Hence, we could either generate $c_1, \ldots, c_n$ according to above procedure or by drawing them from the Dirichlet-Multinomial. The predictive distribution for c_n is equivalent to the predictive distribution for Θ_n, as $c_n = k$ signals that the n-th observation has label θ^k.

Let us now see what happens if we take the limit as $K \to \infty$ (Neal, 2000). We get

$$P(c_n = k|c_1, \ldots, c_{n-1}) = \frac{m_k}{a_0 + n - 1}, \tag{7}$$

where $m_k = \sum_{i=1}^{n-1} \delta(c_i = k)$ is the number of balls from type k drawn till iteration n and

$$P(c_n \neq c_j \; \forall j < n) = \frac{a_0}{a_0 + n - 1}. \tag{8}$$

The case when $K \to \infty$ corresponds to the so-called Chinese Restaurant Process (CRP). The metaphor for the CRP works as follows: Imagine a Chinese restaurant with an infinite number of tables. Customers enter the restaurant and either start a new table or sit down at an already occupied table. At each iteration, they start a new table with probability (8) or sit down at an already occupied table with probability (7). If a new table is started, the customer orders a dish from the menu $\boldsymbol{\theta}$, where in contrast to the Dirichlet distribution with its discrete base measure $\boldsymbol{\alpha}$ we now draw it from a continuous base measure G_0 as the space $\boldsymbol{\theta}$ is now assumed to be infinite-dimensional, i.e. there is an infinite number of dishes available on the menu. One possible explanation why the metaphor

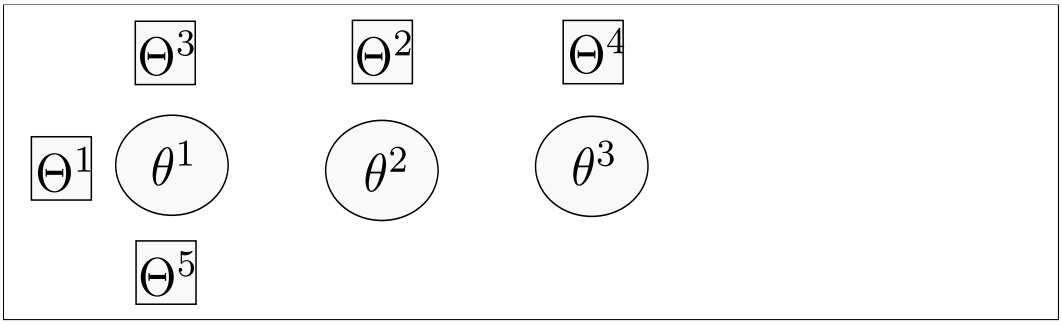

Figure 1: Visualization of the CRP. In this example, five customers Θ_i have entered the restaurant and three tables are occupied in total. The dishes ordered at each table are θ_1, θ_2 and θ_3, which are shared by all customers sitting at the same table. Right now, guest 1, 3 and 5 all share the same dish θ_1.

is about a *Chinese* restaurant is this: Not every guest orders his or her dish individually, but a single order is placed for each table such that all guests sitting together at table share a dish among each other. Such sharing of dishes is common in Chinese cuisine, where often several dishes are ordered at once and then shared among all guests of a dinner.

Given this specification, the predictive distribution for Θ_{n+1} given all previous observations can be written as

$$\Theta_{n+1} \mid \Theta_1, \ldots, \Theta_n \sim \frac{1}{a_0 + n} \left(a_0 G_0 + \sum_{k=1}^{K} N_k \delta_{\theta^k} \right),$$

where δ_{θ^k} is a point-mass centered at θ^k and K denotes the number of occupied tables when the $n+1$-th customer enters the restaurant. Observe that this mixture distribution exactly corresponds to the CRP: When a new customer comes in, he either starts a new table and orders a new dish from base distribution G_0 with probability proportional to α_0 or he sits down at an already occupied table with probability proportional to the number of people already eating at that table (N_k) and eats from the same dish θ^k as everyone else sitting at that table. Notice the similarity between the predictive distributions of the Dirichlet process and distribution (see section 1.2).

The joint probability of drawing a sample of balls from the urn can be factored according to the chain rule as follows:

$$P(c_1, \ldots, c_n) = P(c_1) P(c_2 \mid c_1) \cdots P(c_n \mid c_1, \ldots, c_{n-1}).$$

Using the results in Equations (7) and (8), it can be shown (Gershman and Blei, 2012, p.6) that the joint probability can also be written as

$$P(c_1, \ldots, c_n) = \frac{\alpha_0^K \prod_{k=1}^{K} (N_k - 1)!}{\prod_{i=1}^{n} (i - 1 + \alpha_0)},$$

where N_k is the number of balls with label k after n draws. The important thing to notice is that this probability does not depend on the order in which the balls for each type were drawn! Hence, random samples generated according the Polya urn scheme are exchangeable, which means that for any permutation of the indices $1, 2, \ldots, n-1$ the joint probability distribution of the permuted sequence is the same as that of the original sequence.

An important result from probability theory, De Finetti's representation theorem, states informally that for any (infinite) exchangeable sequence, there exists a latent variable drawn from some underlying measure which renders all variables in the sequence independent. A precise formulation of the theorem given by Schervish (1997) is:

Theorem 3 (De Finetti's representation theorem). *Let $(S, \mathcal{A}, \mu)$ be a probability space, and let $(\mathbb{R}, \mathcal{B})$ be a Borel space. For each n, let $X_n : S \mapsto \mathbb{R}$ be measurable. The sequence $(X_n)_{n=1}^{\infty}$ is exchangeable if and only if there is a random probability measure P on $(\mathbb{R}, \mathcal{B})$ such that, conditional on P, $\{X_n\}_{n=1}^{\infty}$ are iid with distribution P. Furthermore, if the sequence is exchangeable, then the distribution of P is unique, and $P_n(B)$ converges to $P(B)$ almost surely for each $B \in \mathcal{B}$.* □

For the CRP, this underlying measure P is the Dirichlet process. That is, we can generate a sequence of random variables $\Theta_1, \ldots, \Theta_n$ either by using the CRP or by drawing

1. Draw $G \sim \mathrm{DP}(G_0, \alpha_0)$, a random variable drawn from a Dirichlet process with base distribution G_0 and concentration parameter α_0
2. Draw $\Theta_1, \ldots, \Theta_n \mid G \overset{iid}{\sim} G$.

With a method to generate samples from a Dirichlet process at our disposal, we will now examine the properties of the Dirichlet process and its applications in the next sections.

2.2 Key Results

Recall that in contrast to the case of a Dirichlet distribution defined over a finite space $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ specifying different categories, the Dirichlet process is defined over an infinite-dimensional space, which we denote by Ω. Let $\mathcal{F}$ be an associated σ-algebra of events of that space such that together $(\Omega, \mathcal{F})$ form a measurable space. This formulation gives rise to the following definition of the Dirichlet process:

Definition 4 (Definition of Dirichlet Process). *Let G_0 be a non-null finite measure on $(\Omega, \mathcal{F})$. We say G is a Dirichlet process on $(\Omega, \mathcal{F})$ if for every $i = 1, \ldots, k$ and measurable partition $(B_1, \ldots, B_k)$ of $\mathcal{F}$, the distribution of $(G(B_1), \ldots, G(B_k))$ is Dirichlet:*

$$(G(B_1), \ldots, G(B_k)) \sim \mathrm{Dir}\left(\alpha_0 G_0(B_1), \ldots, \alpha_0 G_0\left(B_k\right)\right) \tag{9}$$

The existence of the Dirichlet process was first proven in Ferguson (1973). It is sketched in section 3, as are all results presented in this part of the report. The following properties can be derived for the Dirichlet process:

Theorem 5 (Properties of Dirichlet Process). *Let G be distributed according to the Dirichlet process. Let $A \in \mathcal{F}$ be any measurable set. Then*

1. $\mathbb{E}\left[G(A)\right] = G_0(A)$.
2. $\mathrm{Var}\left[G(A)\right] = \frac{G_0(A)(1-G_0(A^c))}{\alpha_0+1}$ □

Observe that as $\alpha_0 \to \infty$, the variance of the Dirichlet process goes to zero: A draw from the Dirichlet will be much more concentrated around G_0. For very large α_0, drawing a sample from G will, loosely speaking, be the same as just drawing from the base measure G_0 in the first place.

One other noteworthy result is that the conditional distribution of G given a sample is again a Dirichlet process. This property makes the Dirichlet process particularly attractive as an element for Bayesian nonparametrics, as it allows repeatedly updating the beliefs about G in face of incoming data.

Theorem 6 (Posterior Distribution). *Suppose $G \sim \mathrm{DP}\left(G_0, \alpha_0\right)$ and $\Theta_i \mid G \sim G$ for all $i \in \{1, \ldots, n\}$. Then the posterior distribution of G is given by*

$$G \mid \Theta_1, \ldots, \Theta_n \sim \mathrm{DP}\left(\frac{1}{\alpha_0 + n}\left(\alpha_0 G_0 + \sum_{k=1}^{K} N_k \delta_{\theta_k}\right), \alpha_0 + n\right). \tag{10}$$

As already noted in the introduction, the DP is very useful for clustering problems because samples drawn from it are discrete distributions. Precisely:

Theorem 7 (Discreteness of Samples from Dirichlet Process). *If $P \sim \mathrm{DP}\left(\alpha_0, G_0\right)$, then every realization P is a discrete probability measure on $(\Omega, \mathcal{F})$ with probability 1.* □

3 Proof Outlines of Results

Existence of Dirichlet process. The existence of the probability measure defined in definition 4 can be verified using the following theorem:

Theorem 8 (Kolmogorov's Extension theorem). *For each t in arbitrary index set T, let $\Omega_t = \mathbb{R}$ and $\mathcal{F}_t$ be the Borel sets of $\mathbb{R}$. Assume that for each nonempty set v of t, we are given a probability measure P_v on $\mathcal{F}_v$. If the P_v are consistent, i.e.*

$$\pi_u\left(P_v\right) = P_u \text{ for each nonempty } u \subset v,$$

where $\pi_u\left(P_v\right) = P_v\left(x \in \Omega_v : x_u \in B \text{ for } B \in \mathcal{F}_u\right)$, then there exists a unique probability measure P defined on the product space $\left(\prod_{t\in T} \Omega_t, \otimes_{t\in T} \mathcal{F}_t\right)$ such that $\pi_v\left(P\right) = P_v$ for all v. □

The consistency requirement for the existence of P follows from the tail-free property of the Dirichlet distribution of theorem 2.

Proof of theorem 5. (1): Consider any partition $\{B, B^c\}$ of Ω. By the definition of the Dirichlet process, we have

$$(G(B), G(B^c)) \sim \mathrm{Beta}\left(\alpha_0 G_0(B), \alpha_0 G_0(B^c)\right).$$

with mean $\mathbb{E}\left[G(B)\right] = \frac{\alpha_0 G(B)}{\alpha_0 G(B) + \alpha_0 G(B^c)} = \frac{G_0(B)}{G_0(\Omega)} = G_0(B)$. Likewise, (2) follows because of the variance formula for the Beta distribution[4]. □

[4]Recall that a $\mathrm{Beta}\left(\alpha, \beta\right)$ random variable has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Proof of theorem 6. Consider an arbitrary measurable partition $A_1, \ldots, A_k$ of sample space Ω (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{k} A_i = \Omega$). From the definition of the Dirichlet process, we know that

$$X = (G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}\left(\alpha_0 G_0\left(A_1\right), \ldots, \alpha_0 G_0\left(A_k\right)\right). \tag{11}$$

In addition,

$$Y \sim \mathrm{Mutinomial}\left(X\right). \tag{12}$$

By the conjugacy of the Dirichlet to the Multinomial, it follows that the posterior distribution of X is

$$X \mid Y \sim \mathrm{Dir}\left(\alpha_0 G_0\left(A_1\right) + \sum_{i=1}^{n} \delta_{\theta_i}\left(A_1\right), \ldots, \alpha_0 G_0\left(A_k\right) + \sum_{i=1}^{n} \delta_{\theta_i}\left(A_k\right)\right)$$

Since this holds for all partitions, we can conclude that the posterior of G is again a Dirichlet process according to the Kolmogorov extension theorem. Normalizing by multiplying and dividing each coefficient of the Dirichlet with $\alpha_0 + n$ allows us to read off the values for the concentration parameter α_0' and the base distribution G_0' for the Dirichlet process as

$$\begin{aligned} \alpha_0' &= \alpha_0 + n \\ G_0' &= \frac{1}{\alpha_0 + n}\left(\alpha_0 G_0 + \sum_{k=1}^{K} N_k \delta_{\theta_k}\right). \end{aligned}$$

Proof of theorem 7. The following sketch of a proof, originally due to Basu and Tiwari (1982), is taken from Ghosh and Ramamoorthi (2003), adapted to the notation of this paper. Another proof is given in Blackwell (1973) and the seminal paper by Ferguson (1973).
Consider the pair (P, X) of random quantities with $P \sim \mathrm{DP}\left(\alpha_0, G_0\right)$ and $X \mid P \sim P$, that is conditional on P the distribution of X is P. Denote their joint distribution as Q and let $\tilde{E} = \{(P, x) : P\{x\} > 0\}$. Define the x-section of E as $\tilde{E}_x = \{P : P\{x\} > 0\}$ and the P-section as $\tilde{E}_P = \{x : P(x) > 0\}$. Under distribution $\mathrm{DP}\left(\alpha_0 + 1, G_0 + \delta_x\right)$, the random variable $P\{x\}$ is positive with probability one as the $G_0 + \delta_x$ measure of the set $\{x\}$ is positive. Since $P\{x\} \sim \mathrm{Beta}\left(\alpha_0 G_0\left(\{x\}\right), \alpha_0 G_0\left(\Omega \setminus \{x\}\right)\right)$ and $G_0\left(\{x\}\right) > 0$, we can conclude that $P\left(\{x\}\right) > 0$ with probability one. Thus

$$Q(E) = E_Q\left(Q\left(\tilde{E}\right) \mid P\right) = E_Q\left(P\left(\tilde{E}_P\right)\right) = 1.$$

From this, it follows that $P\left(\tilde{E}_P\right) = 1$ almost surely. □

4 Conclusion

In this report, we have given a summary of the Dirichlet process, a prior process which can serve as a building-block in many Bayesian nonparametric problems. Developed by Ferguson (1973) in his seminal paper, the Dirichlet process is only one of many prior processes used in Bayesian nonparametrics. Ferguson (1974) identifies two desirable properties for such priors: (1) the support of the prior on the space of probability measures should be large and (2) the posterior distribution given a sample should be analytically tractable. Both these issues are addressed by the Dirichlet process, with its posterior being again Dirichlet distributed as demonstrated in theorem 6.

We hope that by stressing the similarities of the Dirichlet process to its finite-dimensional counterpart, multinomial sampling with a Dirichlet prior, we have given a comparatively easy introduction to this intruiging but also unintuitive and complex subject matter.

Because of the discreteness of samples drawn from a Dirichlet process (recall theorem 7), it is most widely used as a prior in classification or clustering problems. They have been successfuly used in mixture models for clustering, going back to Antoniak (1974). A hierarchical version of the DP has been developed by Teh et al. (2005) to allow for sharing of statistical strength by linking

grouped observations together and used in topic modeling and extert systems, for which sharing of information across groups is usually necessary. However, it has been noted that the Dirichlet process might not be ideal in all of these use cases, and various other processes have been proposed in order to deal with the perceived deficiencies (see Teh and Jordan (2010)).

In order to understand the various existing prior processes better, future research should investigate their relationships and provide guidance for the practitioner by describing their properties and assessing their strengths and weaknesses when used in various contexts.

References

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* URL http://projecteuclid.org/euclid.aos/1176342871.

ASH, R. B. and DOLÉANS-DADE, C. (2000). *Probability and measure theory*. URL http://books.google.com/books?hl=en&lr=&id=GkqQoRpCO2QC&oi=fnd&pg=PR7&dq=Probability+and+Measure+Theory&ots=EnQHMGFPFC&sig=ABhDd3CniC3447u-p15wVCS86ug.

BASU, D. and TIWARI, R. (1982). A Note on the Dirichlet Process. In *DasGupta (ed.), Sel. Work. Debabrata Basu* (A. DasGupta, ed.). Springer New York, New York, NY, 355–369. URL http://link.springer.com/10.1007/978-1-4419-5825-9.

BLACKWELL, D. (1973). Discreteness of Ferguson selections. *Ann. Stat.*, **1** 356–358. URL http://www.jstor.org/stable/2958021.

FERGUSON, T. (1974). Prior distributions on spaces of probability measures. *Ann. Stat.* URL http://www.jstor.org/stable/2958401.

FERGUSON, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Stat.*, **1** 209–230.

GERSHMAN, S. and BLEI, D. (2012). A tutorial on Bayesian nonparametric models. *J. Math. Psychol.* 1–28. arXiv:1106.2697v2, URL http://www.sciencedirect.com/science/article/pii/S002224961100071X.

GHOSH, J. and RAMAMOORTHI, R. (2003). *Bayesian Nonparametrics*. URL http://www.amazon.com/Bayesian-Nonparametrics-Springer-Series-Statistics/dp/0387955372/ref=cm_rdp_product.

NEAL, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. Comput. Graph. Stat.*, **9** 249–265.

PHADIA, E. G. (2013). *Prior Processes and Their Applications: Nonparametric Bayesian Estimation*. Springer; 2013 edition.

SCHERVISH, M. (1997). *Theory of Statistics*. Springer s ed. Springer. URL http://www.amazon.com/Theory-Statistics-Springer-Series/dp/0387945466.

TEH, Y. and JORDAN, M. (2010). Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics Princ. ...* 1–47. URL http://books.google.com/books?hl=en&lr=&id=0GUzMF59AsgC&oi=fnd&pg=PA158&dq=Hierarchical+Bayesian+Nonparametric+Models+with+Applications+%E2%88%97&ots=SUwNLJPC_T&sig=AFvMJumzaYblrhG-Bdh6z1RiOPQ.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2005). Hierarchical Dirichlet Processes 1–30.

TRESP, V. (2007). Dirichlet Processes and Nonparametric Bayesian Modelling. URL http://videolectures.net/mlss06au_tresp_dpnbm/.