Bayesian nonparametrics

1 Some preliminaries

1.1 de Finetti's theorem

We will start our discussion with this foundational theorem. We will assume throughout all variables are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Theorem 1.1 (de Finetti's theorem[1]). *Let $(X_i)_{i\geq 1}$ be an $(\mathcal{X}, \mathcal{B})$-valued stochastic process, where $(\mathcal{X}, \mathcal{B})$ is Polish. We say $(X_i)_{i\geq 1}$ is exchangeable if*

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_{\sigma(1)} \in B_1, \ldots, X_{\sigma(n)} \in B_n)$$

for any permutation σ of $1, \ldots, n$, all $B_1, \ldots, B_n \in \mathcal{B}$, and any $n \geq 1$. The process is exchangeable if and only if there exists a random measure P, measurable in the tail σ-field of $(X_i)_{i\geq 1}$, such that given P, the variables $X_1, X_2, \ldots$ are independent and distributed according to P. Furthermore, the empirical distribution of $X_1, \ldots, X_n$ converges a.s. to P as $n \to \infty$ at every $B \in \mathcal{B}$.

If we interpret probability as a degree of belief, the exchangeability hypothesis in this theorem is a relatively weak assumption about the variables $X_1, X_2, \ldots$. For example, suppose that these variables represent the outcome of a clinical trial which is repeated on a sequence of patients; if the statistician has no information about the patients, then it is fair to assign an exchangeable prior to $X_1, X_2, \ldots$. In English, we have no reason to believe that $X_1 =$ success, $X_2 =$ failure is more likely than $X_1 =$ failure, $X_2 =$ success. A prior which assumes the variables independent is quite different. If the variables were independent, then observing 10 failures in a row would not affect our belief about the outcome of the 11th observation, a much stronger assumption.

Independence implies exchangeability, but the opposite is not true. What de Finetti's theorem tells us is that if the prior is exchangeable, then this is equivalent to assuming that the variables are independent *conditional* on a hidden probability distribution P on the space of outcomes. Therefore, for a Bayesian who believes exchangeability, putting a prior on the observables $X_1, X_2, \ldots$ is equivalent to putting a prior on the distribution P, which can never be fully observed.

[1] This is Theorem 1.49 in *Theory of Statistics* by Schervish (1995).

1.2 Parametric vs. nonparametric priors and posteriors

Suppose the space $(\mathcal{X}, \mathcal{B})$ is infinite, for example, the real line with the Borel σ-algebra. A *parametric* prior for P would assume that the distribution falls in a restricted family. For example, we might assume that P is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, and we would put a prior on the parameters μ and σ. However, de Finetti's theorem does not require that P is Gaussian; this would be a prior assumption, and quite a strong one. A *nonparametric* analysis would put a prior on P which is supported on a large set of distributions on the real line.

The parametric analysis above is simple because all the information about P is captured by the parameters μ and σ. These parameters and the observables $X_1, X_2, \ldots$ have a joint distribution which has a density. Thus, the posterior distribution has a density:

$$f_{\mu,\theta|X_1,\ldots,X_n}(m, t \mid x_1, \ldots, x_n) = \frac{f_{\mu,\theta,X_1,\ldots,X_n}(m, t, x_1, \ldots, x_n)}{\int f_{\mu,\theta,X_1,\ldots,X_n}(m, t, x_1, \ldots, x_n)\, dm\, dt}.$$

In a nonparametric analysis, the distribution P and the observables $X_1, X_2, \ldots$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, but their distribution need not have a density. Nonetheless, we can still define the posterior of P given $X_1, \ldots, X_n$ as a *Regular Conditional Probability Distribution (RCPD)*. Formally, P lives on the space of probability measures on $(\mathcal{X}, \mathcal{B})$, which we will call $\mathcal{M}(\mathcal{X})$; let $\mathcal{T}$ be the sigma algebra on $\mathcal{M}(\mathcal{X})$ generated by the open sets in the weak-$*$ topology. The posterior of P given $X_1, \ldots, X_n$ is a function $f_{P|X_1,\ldots,X_n} : (\Omega, \mathcal{T}) \to [0, 1]$, satisfying:

1. For every fixed $A \in \mathcal{T}$, $f_{P|X_1,\ldots,X_n}(\omega, A)$ is a version of the conditional expectation $E(\mathbb{1}(P \in A) \mid \sigma(X_1, \ldots, X_n))$.

2. For every fixed ω in a set with probability one in $\mathbb{P}$, $f_{P|X_1,\ldots,X_n}(\omega, A)$ is a probability measure on $(\mathcal{M}(\mathcal{X}), \mathcal{T})$.

Remark. A sufficient condition for the existence of an RCPD is that the random variables take values in a Polish space (or a *Borel isomorphic* space). This explains the condition that the space is Polish in de Finetti's theorem. A Polish space is a topological space which is complete, separable and metrizable. Most spaces of interest in Bayesian nonparametrics will be Polish, including $\mathbb{R}^n$, $\mathbb{C}^n$, any separable Banach space, such as L_2, the space of probabilities on a Polish space with the weak-$*$ topology (the topology of convergence in distribution), the space of continuous real functions with the topology of compact convergence, etc.

We will rarely think of nonparametric posteriors as RCPDs. Instead, we typically work with finite-dimensional projections of the posterior which have nice densities. This allows us to apply Bayes theorem for continuous variables as we would in a parametric analysis. The full posterior, if it is of interest, can be defined through limiting arguments.

2 The Dirichlet Process

2.1 Projective limits of probabilities

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let $\mathcal{I}$ be the set of measurable partitions of $\mathcal{X}$. Let $\mathbf{B} = \{B_1, \ldots, B_n\}$ and $\mathbf{C} = \{C_1, \ldots, C_m\}$ be two measurable partitions; we say $\mathbf{C}$ is a *coarsening* of B if each set C_i is the union of a subset of $\mathbf{B}$.

For every measurable partition $\mathbf{B}$, let $P_{\mathbf{B}}$ be a random probability measure on the finite σ-algebra $\sigma(\mathbf{B})$. We call $\{P_{\mathbf{B}}; \mathbf{B} \in \mathcal{I}\}$ a *projective family* if whenever $\mathbf{C}$ is a coarsening of $\mathbf{B}$, then $P_{\mathbf{B}}$ evaluated at $\sigma(\mathbf{C})$ is identical in distribution to $P_{\mathbf{C}}$. This is a statement about the distribution of each element in the family, so we can consider the random measures to be defined on different probability spaces.

We can think of a projective family as a set of marginal distributions. The property defined above is sometimes referred to as *consistency*. The following theorem asserts that there exists a random measure on the entire space $(\mathcal{X}, \mathcal{B})$ such that the projective family is the family of marginal distributions.

Theorem 2.1 (P. Orbanz[2])**.** *Suppose $(\mathcal{X}, \mathcal{B})$ is Polish and $\{P_{\mathbf{B}}; \mathbf{B} \in \mathcal{I}\}$ is a projective family. Furthermore, suppose the means $\{E(P_{\mathbf{B}}); \mathbf{B} \in \mathcal{I}\}$ are also projective. Then, there exists a random probability measure P such that, for any $\mathbf{B} \in \mathcal{I}$, the restriction of P to the sigma algebra $\sigma(\mathbf{B})$ has the same distribution as $P_{\mathbf{B}}$. The probability P lives on the measure space $(\mathcal{M}(\mathcal{X}), \mathcal{T})$.*

2.2 Construction of the Dirichlet Process

Lemma 2.2. *Let α be a diffuse, finite measure on $(\mathcal{X}, \mathcal{B})$. Define a family of random probabilities with distribution*

$$P_{\mathbf{B}}(B_1), \ldots, P_{\mathbf{B}}(B_n) \sim \textit{Dirichlet}(\alpha(B_1), \ldots, \alpha(B_n)),$$

[2] P. Orbanz. *Projective limit random probabilities in Polish spaces.* Electron. J. Stat. 5 (2011), 1354–1373.

for every $\mathbf{B} \in \mathcal{I}$*. This is a projective family.*

This lemma is a corollary of the lumping property of the Dirichlet distribution. We can apply Theorem 2.1 to conclude that there exists a random probability P on $(\mathcal{X}, \mathcal{B})$ with the Dirichlet marginals defined in the previous lemma. We call this the *Dirichlet Process with base measure* α.

Ferguson was the first to define the Dirichlet Process using a limiting argument like the one above. The original paper from 1973 appeals to Kolmogorov's extension theorem. However, this argument was flawed in that it does not guarantee that the random probability P is measurable on $(\mathcal{M}(\mathcal{X}), \mathcal{T})$. It is necessary to impose topological constraints like the requirement that $(\mathcal{X}, \mathcal{B})$ is Polish in Theorem 2.1.

2.3 Parameters of the Dirichlet Process

The only parameter of the Dirichlet Process is the measure α. We can represent this measure by the pair (β, H), where $\beta = \alpha(\mathcal{X})$ is known as the *concentration* parameter, and $H = \alpha/\alpha(\mathcal{X})$ is the base probability distribution.

H is the mean of the Dirichlet Process in the sense that $E(P(B)) = H(B)$ for all $B \in \mathcal{B}$. As its name implies β modulates how concentrated P is around H. As β goes to infinity $P(B)$ converges a.s. to $H(B)$ for every measurable B. This is a form of pointwise convergence; as we'll see later, P is a.s. discrete while H is diffuse, no matter the value of β. As β goes to 0, P tends to concentrate on a single point mass.

2.4 Posterior distribution

The simplest application of the Dirichlet Process is the model

$$\begin{aligned} P &\sim \mathrm{DP}(\alpha) \\ X_1, X_2, \cdots \mid P &\overset{iid}{\sim} P. \end{aligned}$$

That is, $X_1, X_2, \ldots$ is a sequence of exchangeable observations, and we give a Dirichlet Process prior to the latent distribution P. The following lemma states that the Dirichlet Process is a conjugate prior.

Lemma 2.3. *The posterior distribution of* P *given* n *observations* $X_1, \ldots, X_n$ *is a Dirichlet Process with base measure* $\alpha + \sum_{i=1}^{n} \delta_{X_i}$ *where* δ_x *is a point mass as* x*.*

Proof. We claim that for any measurable partition $B_1, \ldots, B_m$,

$$P(B_1), \ldots, P(B_n) \mid X_1, \ldots, X_n \sim \text{Dirichlet}(\alpha(B_1) + K_1, \ldots, \alpha(B_m) + K_m),$$

where K_i is the number of samples among $X_1, \ldots, X_n$ that fall in B_i. This is the marginal distribution of a Dirichlet Process with base measure $\alpha + \sum_{i=1}^n \delta_{X_i}$, and since this holds for any measurable partition, we can define the full posterior by an application of Theorem 2.1.

To prove the claim, let

$$Q_B(A) = \frac{P(A)}{P(B)} \qquad \forall A \subset B, A \in \mathcal{B}$$

be the restriction of P to B, and note that $(P(B_1), \ldots, P(B_m))$ together with $Q_{B_1}, \ldots, Q_{B_m}$ specify P. Recall that

$$P(B_1), \ldots, P(B_m) \sim \text{Dirichlet}(\alpha(B_1), \ldots, \alpha(B_m)).$$

Furthermore, due to properties of the Dirichlet distribution, $(P(B_1), \ldots, P(B_m))$ and $Q_{B_1}, \ldots, Q_{B_m}$ are mutually independent. Since we are dealing with a finite-dimensional marginal of P, the posterior has a density:

$$\begin{aligned} f_{P_{\mathbf{B}}|X_{1:n}}(p_{1:m} \mid x_{1:n}) \propto f_{X_{1:n}|P_{\mathbf{B}},K_{1:m}}(x_{1:n} \mid p_{1:m}, k_{1:m}) \times \\ f_{K_k|P_{\mathbf{B}}}(k_{1:m} \mid p_{1:m}) \times f_{P_{\mathbf{B}}}(p_{1:m}), \end{aligned}$$

where we use the notation $a_{1:\ell} = (a_1, \ldots, a_\ell)$ for vectors. Given $K_{1:m}$, the distribution of $X_1, \ldots, X_n$ depends only on $Q_{B_1}, \ldots, Q_{B_n}$, therefore, the first factor on the right hand side does not depend on $P_{\mathbf{B}}$ and

$$f_{P_{\mathbf{B}}|X_{1:n}}(p_{1:m} \mid x_{1:n}) \propto f_{K_k|P_{\mathbf{B}}}(k_{1:m} \mid p_{1:m}) f_{P_{\mathbf{B}}}(p_{1:m}),$$

where on the right hand side we have the product of a multinomial likelihood and a Dirichlet density. The claim follows from the Dirichlet-multinomial conjugacy. □

Remark. We can write the base measure of the posterior

$$\alpha(\cdot) + \sum_{i=1}^n \delta_{X_i}(\cdot) = (\beta + n) \left[\frac{\beta}{\beta + n} H(\cdot) + \frac{n}{\beta + n} \sum_{i=1}^n \frac{\delta_{X_i}(\cdot)}{n} \right].$$

Here, $\beta + n$ is the concentration parameter of the posterior, and the base distribution between brackets is a mixture of the prior base distribution H and the empirical distribution of $X_1, \ldots, X_n$. The first term captures the influence of the prior, while the second captures the influence of the data. The concentration of the prior, β, modulates the weight of each contribution.

2.5 Predictive distribution

The predictive distribution is the distribution of X_{n+1} given the first n observations $X_1, \ldots, X_n$. By the tower property of conditional expectation

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = E(P(A)|X_1, \ldots, X_n),$$

and Lemma 2.3 shows that the posterior of P is a Dirichlet process. Therefore, the conditional expectation above is the mean of the Dirichlet Process posterior (the base probability distribution):

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{\beta}{\beta + n} H(A) + \frac{n}{\beta + n} \sum_{i=1}^{n} \frac{\delta_{X_i}(A)}{n}.$$

Since the right hand side is a mixture of two distributions, we can rephrase the equation as a sampling algorithm: with probability $\beta/(\beta+n)$, draw X_{n+1} from H, and otherwise, let X_{n+1} be equal to one of the previous samples, chosen uniformly.

This means that in $X_1, X_2, \ldots$, it is possible to observe the same point multiple times; in fact, every observable is repeated infinitely often.

Lemma 2.4. *With probability 1, the measure P is discrete, i.e. a mixture of point masses in $\mathcal{X}$.*

Proof. From the predictive distribution, the probability that X_n is never repeated in $X_{n+1}, X_{n+2}, \ldots$ is

$$\prod_{i=1}^{\infty} \frac{C_1}{C_2 + i}$$

for two constants C_1, C_2. This product is equal to 0, therefore, this observable is repeated with probability 1. Since the sequence is countable, with probability 1 every observable is repeated. If P had a diffuse component with some positive probability p, then the event that all observables are repeated would have probability at most $1 - p$, which contradicts the previous fact. □

The predictive distribution above matches that of the following urn scheme, which is a straightforward generalization of the Pólya urn scheme with a mechanism to increase the number of colors.

Blackwell-MacQueen urn. Start with an urn containing a single black ball with weight β. At every step, draw a ball from the urn with probability proportional to its weight. Then,

1. If the ball is black, return it to the urn along with another ball of weight 1, with a new color sampled from H.

2. If the ball is colored, return it to the urn along with another ball of weight 1 of the same color.

The sequence of colors for the balls added to the urn at each step has the same distribution as $X_1, X_2, \ldots$.

This urn scheme defines a mechanism to determine which of the observables are identical or how a sequence $X_1, \ldots, X_n$ is partitioned into classes in which all observables are equal. The colors can be understood as labels for each class and are sampled independently from H. This is even more apparent in the following story.

The Chinese restaurant process. A sequence of customers walk into a chinese restaurant where each table eats a single dish. When the $n+1$-th customer walks into the restaurant, there are M_n tables occupied.

1. With probability $\beta/(\beta+n)$, she sits at a new table and samples a dish $\theta_{M_n+1} \sim H$ for the table.

2. Otherwise, she sits at a table which is already occupied with probability proportional to the number of customers on the table.

Again, this defines a distribution for the partition of customers (or observables) into classes. Each class is labeled by a dish and the sequence of labels $\theta_1, \theta_2, \ldots$ is drawn independently from H.

The Chinese restaurant process is clearly a *rich get richer* scheme, as tables with more customers tend to attract even more customers. The moments of the number of tables occupied after n customers can be derived explicitly.

Lemma 2.5.

$$E[M_n] = \sum_{i=1}^{n} \frac{\beta}{\beta + i + 1} = \beta \log\left(1 + \frac{n}{\beta}\right) + o(1)$$

$$Var[M_n] = \sum_{i=1}^{n} \frac{\beta(i+1)}{(\beta + i + 1)^2} = \beta \log\left(1 + \frac{n}{\beta}\right) + o(1).$$

It is worth noting that the predictive distribution of the Dirichlet Process was discovered independently in the field of mathematical genetics, where it is known as the Ewens sampling formula.

2.6 Construction through de Finetti's theorem

Above, we constructed the Dirichlet Process by defining a projective family of marginal distributions. We could have also started our discussion with the predictive distribution of $X_1, X_2, \ldots$ defined in the previous section. If we prove that $X_1, X_2, \ldots$ are exchangeable, de Finetti's theorem implies the existence of a hidden random measure P, which we call the Dirichlet Process.

Lemma 2.6. *The process $X_1, X_2, \ldots$ defined by the Blackwell-McQueen urn is exchangeable.*

Proof. Let $k_1, k_2, \ldots, k_{M_n}$ be the sizes of each equivalence class in $x_1, \ldots, x_n$, and let $\theta_1, \theta_2, \ldots, \theta_{M_n}$ be the unique labels in $\mathcal{X}$ applied to the classes. If f_H is the density of H, the joint density of the first n observables can be written

$$f_{X_{1:n}}(x_{1:n}) = \mathbb{P}(k_{1:m}) \prod_{j=1}^{M_n} f_H(\theta_j) = \frac{\beta^{M_n} \prod_{j=1}^{M_n} (1)_{k_j}}{(\beta)_n} \prod_{j=1}^{M_n} f_H(\theta_j),$$

where we use the Pochhammer symbol $(x)_n = x(x+1)\ldots(x+n-1)$. This can be proven easily by induction on n. The lemma follows from the observation that the density is invariant to permuting $x_1, \ldots, x_n$. □

2.7 Stick breaking representation

This representation of the Dirichlet Process was discovered by Sethuraman in 1994[3].

Theorem 2.7. *Let $\gamma_1, \gamma_2, \ldots$ be i.i.d. Beta$(1, \beta)$ random variables, and $\theta_1, \theta_2, \ldots$ be i.i.d. from H. Define,*

$$\pi_k = \gamma_k \prod_{i=1}^{k-1} (1 - \gamma_i).$$

The random measure $P = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ is a Dirichlet Process with base measure $\alpha = \beta H$.

Remark. In English, we start with a stick of length 1, and we iterate the following: break off a random Beta$(1, \beta)$ fraction of what is left of the stick and assign this to a point mass at a location drawn at random from H.

[3] J. Sethuraman. *A constructive definition of Dirichlet priors.* Stat. Sinica 4 (1994), 639-650.

Proof. For the full proof, see Sethuraman's paper. Here, we only give a schematic. Let $\theta_i^* = \theta_{i+1}$ and $\gamma_i^* = \gamma_{i+1}$ for $i = 1, 2, \ldots$. And note that the random measure P^*, generated using these sequences, is independent of (θ_1, γ_1) and has the same distribution as P. We can summarize this result by

$$P \stackrel{d}{=} \gamma_1 \delta_{\theta_1} + (1 - \gamma_1) P.$$

For any measurable partition of $\mathcal{X}$, $\mathbf{B} = \{B_1, \ldots, B_m\}$, equality in distribution clearly holds as well for the marginals:

$$P_{\mathbf{B}}(\cdot) \stackrel{d}{=} \gamma_1 \mathbb{1}(\theta_1 \in \cdot) + (1 - \gamma_1) P_{\mathbf{B}}(\cdot).$$

The proof is completed by noting that this equality in distribution characterizes the law of the random vector $P_{\mathbf{B}}$, and in particular, the Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_m))$ satisfies the equality (the former is proven by Sethuraman, while the latter is a consequence of basic properties of the Dirichlet distribution). Since this holds for any measurable partition $\mathbf{B}$, P is a Dirichlet Process with base measure α. □

The distribution of the sequence of weights π, in the specific order defined in the above theorem, is known as the *GEM(β)* distribution, for Griffiths, Ewens, and McCloskey.

3 Mixture models

A mixture model has the following form:

$$\begin{aligned} X_i \mid \phi_i &\sim F(\phi_i) \\ \phi_1, \phi_2, \cdots \mid P &\overset{iid}{\sim} P \\ P &\sim DP(\alpha). \end{aligned} \tag{1}$$

Here, F is some distribution with parameter ϕ; for example, F could be a normal distribution and ϕ a vector of length 2 with its mean and variance. The parameters for each observable $X_1, X_2, \ldots$ are drawn from a Blackwell-MacQueen urn, so certain observables will have the same parameters or belong to the same *cluster*.

There is a dual interpretation of a Bayesian analysis of mixture models. On the one hand, the posterior of P solves a density estimation problem. On the other hand, we can infer how the samples cluster in a typical posterior sample, which solves a clustering problem.

The above model can be rewritten using the stick breaking representation,

$$\begin{aligned} X_i \mid Z_i, (\theta_j)_{j\geq 1} &\sim F(\theta_{Z_i}) \\ Z_1, Z_2, \cdots \mid \pi &\overset{iid}{\sim} \pi \\ \pi &\sim \text{GEM}(\beta) \\ \theta_1, \theta_2, \ldots &\overset{iid}{\sim} H \end{aligned} \tag{2}$$

Here, the variable Z_i is an integer which represents to which cluster X_i belongs. The parameter of this cluster is θ_{Z_i}.

A Dirichlet Process mixture model is a nonparametric generalization of a finite mixture model of the form:

$$\begin{aligned} X_i \mid Z_i, (\theta_j)_{j\geq 1} &\sim F(\theta_{Z_i}) \\ Z_1, Z_2, \cdots \mid \pi &\overset{iid}{\sim} \pi \\ \pi_1, \ldots, \pi_K &\sim \text{Dirichlet}(\beta/K, \ldots, \beta/K) \\ \theta_1, \ldots, \theta_k &\overset{iid}{\sim} H \end{aligned} \tag{3}$$

In fact, you will prove in Homework 1 that a Dirichlet Process mixture model is the limit of this parametric model as $K \to \infty$. The reasons to use a nonparametric extension are the following:

1. As we will see, the computation is not more difficult.
2. The complexity of the model is captured by how spread out the distribution π is, or how many clusters are "important". In a Dirichlet Process mixture, this complexity is not constrained. A priori, it is modulated by the parameter β.

Mixture models are convenient because it is possible to sample the posterior distribution of the cluster memberships $Z_1, \ldots, Z_n$ given the observations $X_1, \ldots, X_n$ through Markov Chain Monte Carlo (MCMC). This can be done most easily when the likelihood F and the prior H are conjugate pairs. Radford Neal wrote a popular review of the most common strategies for MCMC[4]. We will call these algorithms *marginal Gibbs samplers*. Ishwaran and James[5] introduced the *blocked Gibbs sampler* which often works better in practice and relies on the stick breaking representation of the Dirichlet Process.

[4] R. Neal. *Markov chain sampling methods for Dirichlet Process mixture models.* J. Comp. Graph. Stat. 9 (2000), 249-265.

[5] H. Ishwaran, L. James. *Gibbs sampling methods for stick-breaking priors.* J. Amer. Stat. Assoc., 96 (2001), 161-173.

3.1 Marginal Gibbs samplers

For now, let us restrain our attention to the case when the likelihood F and the prior H are conjugate. For example, $F(\theta)$ could be a normal distribution with mean θ, and H a normal distribution. Any conjugate pair (Poisson-gamma, gamma-gamma, etc.) is valid.

The first algorithm we consider uses the representation in (1).

Algorithm 1. Define a Gibbs sampler with state space $(\phi_1, \ldots, \phi_n)$. At each iteration, sample

$$\phi_i \mid \phi_{-i}, X_1, \ldots, X_n \quad \text{for } i = 1, \ldots, n,$$

where $\phi_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$.

The first thing to note is that, given ϕ_{-i}, ϕ_i is independent of X_{-i}. The prior for $\phi_1, \ldots, \phi_n$ is a Blackwell-MacQueen urn, which is exchangeable, therefore, the distribution of $\phi_i \mid \phi_{-i}$ is the predictive rule of the urn,

$$\phi_i \mid \phi_{-i} \sim \frac{1}{n-1+\beta} \sum_{j \neq i} \delta_{\phi_j} + \frac{\beta}{n-1+\beta} H.$$

We derive the distribution of $\phi_i \mid \phi_{-i}, X_i$ using Bayes rule, multiplying the above distribution by the likelihood $X_i \mid \phi_i$:

$$\phi_i \mid \phi_{-i}, X_i \sim \frac{b}{n-1+\beta} \sum_{j \neq i} F(X_i; \phi_j) \delta_{\phi_j} + \frac{b\beta}{n-1+\beta} \left[\int F(X_i; \phi) H(d\phi) \right] H_i.$$

where $H_i(\phi) \propto H(\phi) F(X_i; \phi)$, and b is a normalizing constant. If we choose to make ϕ_i identical to the parameter ϕ_j of an existing cluster, the likelihood of X_i is simply $F(X_i; \phi_j)$. On the other hand, if we let ϕ_i be different from all other parameters, we integrate the likelihood of X_i with respect to the prior H on the parameter.

Here we assume that one can compute the integral $\int F(X_i; \phi) H(d\phi)$, and that the posterior H_i is easy to sample. These assumptions hold when (F, H) is a conjugate pair. Algorithm 1 can be very inefficient, because at each step we change the parameter associated to a single data point.

We now propose a second algorithm, which uses the stick breaking representation of the mixture model in (2).

Algorithm 2. Define a Gibbs sampler with state space $((\theta_i)_{i \geq 1}, Z_1, \ldots, Z_n)$. At each iteration, sample

$$Z_i \mid Z_{-i}, (\theta_{Z_j})_{j \neq i}, X_1, \ldots, X_n \quad \text{for } i = 1, \ldots, n,$$

and

$$\theta_{Z_1}, \ldots, \theta_{Z_n} \mid Z_1, \ldots, Z_n, X_1, \ldots, X_n.$$

The second step in this sampler is straight-forward, because we are conditioning on the partition of samples into clusters and the data for each cluster. The conditional distribution of the parameter θ_i is independent from the conditional distribution of the parameter θ_j for any other cluster. These posteriors are in the same family as H if (F, H) is a conjugate pair. Clearly, we need only sample the posterior of those clusters to which data are assigned.

The first step resembles the iteration of Algorithm 1. We have

$$\mathbb{P}(Z_i = z \mid Z_{-i}, X_i, (\theta_i)) = \begin{cases} b\frac{n_{-i,z}}{n-1+\beta} F(X_i; \theta_z) & \text{if } z = Z_j \text{ for some} j \neq i \\ b\frac{\beta}{n-1+\beta} \int F(X_i; \theta) H(d\theta) & \text{otherwise,} \end{cases}$$

where $n_{-i,z}$ is the number of samples among X_{-i} in cluster z.

Finally, we can define an algorithm in which the vector of parameters $(\theta_i)_{i \geq 1}$ is marginalized.

Algorithm 3. Define a Gibbs sampler with state space $(Z_1, \ldots, Z_n)$. At each iteration, sample

$$Z_i \mid Z_{-i}, X_1, \ldots, X_n \quad \text{for } i = 1, \ldots, n.$$

As before, $\mathbb{P}(Z_i = z \mid Z_{-i})$ corresponds to the prediction rule of a Blackwell-MacQueen urn. Bayes rule has us multiply this by the marginal probability that X_i is in cluster z. When z is already represented in Z_{-i}, we must now integrate over the posterior $H_{z,-i}$ of θ given the samples already in the cluster. Formally,

$$\mathbb{P}(Z_i = z \mid Z_{-i}, X_1, \ldots, X_n) = \begin{cases} b\frac{n_{-i,z}}{n-1+\beta} \int F(X_i; \theta) H_{z,-i}(d\theta) & \text{if } z = Z_j \text{ for some} j \neq i \\ b\frac{\beta}{n-1+\beta} \int F(X_i; \theta) H(d\theta) & \text{otherwise.} \end{cases}$$

where $H_{z,-i}(\theta) \propto \prod_{j \neq i; Z_j = z} F(X_j; \theta) H(\theta)$.

Remark. In deriving these algorithms, the only property of the Dirichlet Process that we used was the formula for its predictive distribution. There are other priors on the random measure P, such as the Pitman-Yor process, which have simple predictive distributions. This allows us to apply the same sampling algorithms.

3.2 Non-conjugate mixtures

It is possible to modify Algorithm 2 to work in the case when F and H are not conjugate. This algorithm requires sampling the posterior distribution of the parameter for a cluster given samples in the cluster. This is proportional to $\prod_{i;Z_i=z} F(X_i;\theta)H(\theta)$. When this distribution cannot be sampled directly, we can instead take a step of a Metropolis-Hastings Markov chain which preserves the distribution. This defines a valid sampler for the posterior distribution of interest.

Sampling $Z_i \mid Z_{-i}, X_1, \ldots, X_n, (\theta_i)$ presents another challenge in the non-conjugate case. In particular, computing the probability that Z_i is in a different cluster from all other samples requires the marginal probability that X_i is in a new cluster with a prior H on the cluster parameter. This integral is not readily available in the non-conjugate case.

Radford Neal reviews several ways to work around this problem. They involve including the parameters of a number of *phantom* clusters, which are not necessarily populated by any samples, in the sample space of the Gibbs sampler. The sampler only proposes moves to these empty clusters conditioning on their parameters. For details on this trick, see the review.

3.3 Blocked Gibbs sampler

The main idea is to include the stick breaking weights π in the state space of the Gibbs sampler. We consider the finite mixture model in which the weights $\pi_1, \ldots, \pi_K$ are sampled with a stick breaking scheme truncated at the Kth step; that is

$$\pi_i = \gamma_i(1-\gamma_{i-1})\ldots(1-\gamma_1)$$

for $i = 1, \ldots, K-1$, and $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$, where $\gamma_i \sim \text{Beta}(1,1)$ are independent.

Blocked Gibbs sampler. The state space of the Markov chain is

$$(\theta = (\theta_1, \ldots, \theta_K), \pi = (\pi_1, \ldots, \pi_K), Z_1, \ldots, Z_n).$$

We iterate sampling the following distributions:

1. $\theta \mid Z_1, \ldots, Z_n, X_1, \ldots, X_n$,
2. $Z_1, \ldots, Z_n \mid \theta, \pi, X_1, \ldots, X_n$,
3. $\pi \mid Z_1, \ldots, Z_n$.

Sampling the distribution in (1) is trivial if (F, H) is a conjugate pair. If not, one can use a step of Metropolis-Hastings which preserves this distribution.

Sampling the joint distribution in (2) is simple as well because we are conditioning on π. We sample each Z_i independently from $\sum_{k=1}^{K} \pi_{k,i}\delta_k(\cdot)$, where $\pi_{k,i} \propto \pi_k F(X_i; \theta_k)$.

Finally, the posterior of π given $Z_1, \ldots, Z_n$ is also a stick breaking process,

$$\pi_i = V_i(1 - V_{i-1}) \ldots (1 - V_1)$$

for $i = 1, \ldots, K-1$, and $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$, where $V_i \sim \text{Beta}(1 + n_i, 1 + \sum_{k=i+1}^{K} n_k)$ are independent for $i = 1, \ldots, K-1$, and $n_1, \ldots, n_K$ are the sizes of each cluster in $Z_1, \ldots, Z_n$.

The reason that this algorithm often converges faster than marginal Gibbs samplers is that in step (2), it resamples all the cluster memberships at once, which makes it easier to radically alter the clustering in one step.

As we let $K \to \infty$, this model converges to a Dirichlet Process mixture model. We defined a different finite mixture model in (3) which also converges to a Dirichlet Process mixture. The difference between these two models is that in the truncated stick breaking scheme, the weights $\pi_1, \pi_2, \ldots, \pi_K$ tend to decrease in size. This makes it possible to define Gibbs samplers for the posterior of the nonparametric model. We will not discuss these in detail, but we note that the stick breaking representation is critical.

Remark. The prior on π was only relevant in step (3), and we only used the fact that it is a stick breaking process where each variable γ_i is independent and Beta-distributed. The parameters of the Beta distribution for each γ_i can be modified. This allows us to define more flexible priors for π and apply the blocked Gibbs sampler. The paper by Ishwaran and James discusses other examples, including the Pitman-Yor process.