Dirichlet Processes
Draft

Adalberto Claudio Quiros
a.claudio-quiros.1@research.gla.ac.uk.
School of Computing Science, University of Glasgow

Introduction

This document covers the definition of a Dirichlet process, its properties and different representations (Stick-breaking construction, Blackwell-MacQueen urn, and Chinese restaurant process), the posterior of a Dirichlet process and the predictive posterior, how they are related to the representations, and finally the Dirichlet process mixture model and inference. It also includes an appendix with brief overviews of the Dirichlet distribution, GEM distribution, and concepts of measure theory. There are other related concepts like Size-biased sampling, Normalized Gamma processes, Pitman-Yor processes, and Power laws which are not included here.

This is an introduction to Dirichlet Processes based on different tutorials and reviews from Yee Whye Teh (University of Oxford), Michael I. Jordan (U.C. Berkeley), Tamara Roderick (MIT), Zoubin Ghahramani (University of Cambridge), and Peter Orbanz (UCL) on Bayesian Nonparametrics and Dirichlet processes.

Dirichlet process

1 Definition

The Dirichlet Process (DP) is a stochastic process on probability measures, a distribution over random probability measures G on the measurable space (Θ, Σ). Such that for any finite disjoint set of partitions $\{(A_1, ..., A_K) = \Theta; A_i \cap A_j = \emptyset,\ i \neq j\}$ the random vector $(G(A_1), ..., G(A_K))$ is Dirichlet distributed:

$$(G(A_1), ..., G(A_K)) \sim Dirichlet(\alpha H(A_1), ..., \alpha H(A_K)) \tag{1}$$

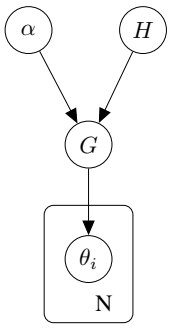


Figure 1: Dirichlet Process graphical model.

The random probability measure G, a draw from the Dirichlet Process, defines a possible countable infinite partition of the space Θ. Each subset of the partition and its associated probability weight behave according to the concentration parameter α and the base measure H.

$$G \sim DP(\alpha, H) \tag{2}$$

Given the definition of a Dirichlet process, we can consider any measurable subset $A_i \subset \Theta$, the marginal distribution of any subset A_i is the $Beta$ distribution:

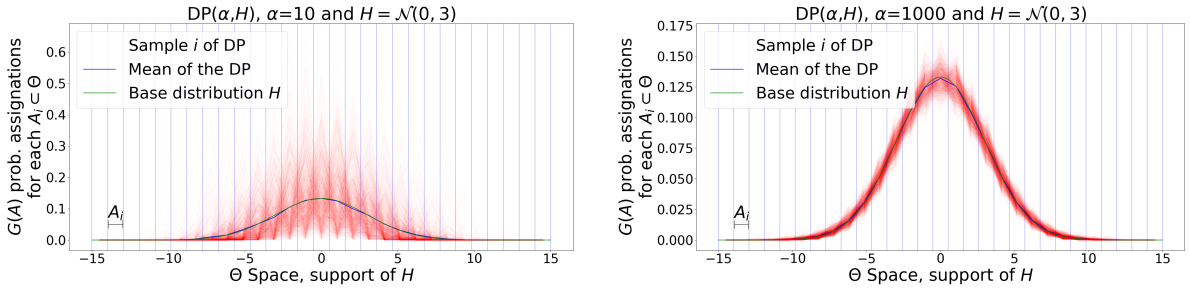
$$G(A_i) \sim Beta(\alpha H(A_i), \alpha H(A_i^c)) \tag{3}$$

Parameters

The Dirichlet process is parameterized by the concentration parameter α and the base measure H. The base measure H serves as a mean to the Dirichlet process, for any measurable subset $A_i \subset \Theta$, $E[G(A_i)] = H(A_i)$, additionally the concentration parameter α describes the consolidation around the mean of the Dirichlet process, it serves as the inverse of the variance

$Var[G(A_i)] = \frac{H(A_i)(1-H(A_i))}{\alpha+1}$. Figure 2 capture how these parameters α, H dictate the behavior of the Dirichlet process.

From the same Figure 2, we may notice that when α tends to higher values, samples G have closer values to the base measure H, since there's less variation around the mean of the Dirichlet process. When $\alpha \longrightarrow \infty,\ G(A_i) \longrightarrow H(A_i)$, although it might seem like $G \longrightarrow H$, this is not true, G is a discrete distribution with probability sum up to one. When $\alpha \longrightarrow \infty$, $G \longrightarrow H$ weakly or pointwise.



(a) (b)

Figure 2: Dirichlet process visualizations for different α values and base measure $H = \mathcal{N}(0, 3)$, (a) $\alpha = 10$ and (b) $\alpha = 1000$. These figures highlight $E[G(A_i)] = H(A_i)$ and how higher α values show more variability along the base measure H. We can also infer that as $\alpha \to \infty, G \to H$ pointwise. These visualizations were constructed with 1000 G samples and A_i measurable sets between vertical blue lines.

Existence

There are different ways to derive the Dirichlet process, different representations that allow intuitive ways to identify its properties:

- 'Fergun's definition'[1]: Starting from a finite Dirichlet distribution $Dir(\alpha/K, ..., \alpha/K)$ and assuming that $K \to \infty$, Ferguson proves the existence of the Dirichlet process. Using Kolmogorov's consistency theorem to show that a collection of finite-dimensional distributions over the measurable space Θ can define a stochastic process, such in Equation

1. Although the Kolmogorov's consistency theorem does not guarantee that the Dirichlet process is a distribution over probability measures, Ferguson[1] proves it through the construction of a Dirichlet process as normalized gamma process.

- Polya and Blackwell-MacQueen urns[2]: These urn models and De Finetti's theorem allow to derive the marginal predictive distribution of the Dirichlet process. This representation along with the Chinese restaurant process gives a description of the exchangeability property of Dirichlet processes.

- Chinese restaurant process[3]: The Chinese restaurant process metaphor describes the clustering property, and how the Dirichlet process induces random partitions over the sample space.

- Stick-breaking construction[4]: Sethuraman defines the Dirichlet process through the stick-breaking construction of G, an atomic distribution of a weighted sum of point masses.

The Blackwell-MacQueen urn and Chinese restaurant process provide a representation to obtain exchangeable sequences of the predictive distribution (marginalizing the random probability measure G), while the Stick-breaking representation gives a direct way of constructing a Dirichlet process and its posterior.

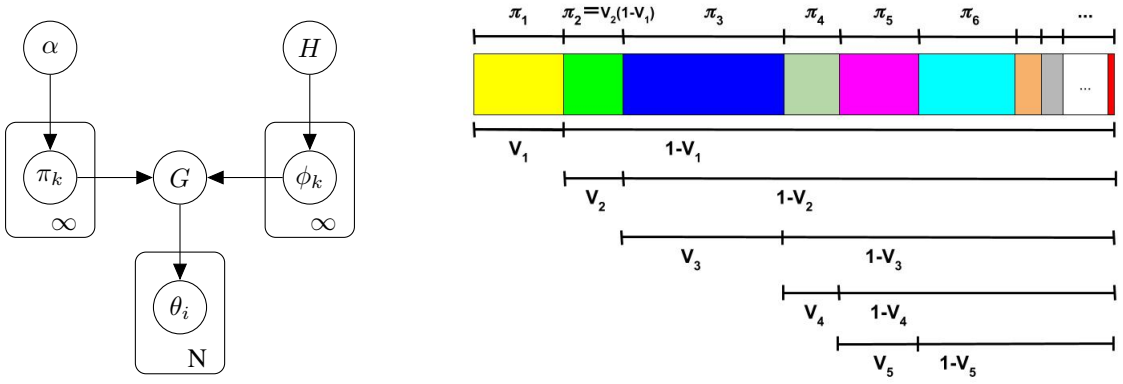
2 Stick-breaking construction

Sethuraman[4] defines the DP existence as a stick-breaking construction, a weighted sum of point masses. The construction goes as follows:

$$\begin{aligned} \pi \sim GEM(\alpha);\ \phi \overset{i.i.d}{\sim} H \\ G = \sum_{k=1}^{\infty} \pi_k \delta(\phi, \phi_k) \\ \theta_i \sim G \end{aligned} \tag{4}$$

The definition of the vector π is the same as the Griffiths-Engen-McCloskey (GEM) distribution[5]. We can construct the random vector π as infinite mixing proportions through an infinite sequence of $Beta$ random variables, Figure 3b. It is important to notice that although the mixing proportions π_k do not strictly decrease, they do in average (Appendix: GEM distribution Figures 18):

$$\begin{aligned} V_k \sim Beta(1, \alpha);\ \ k = 1, 2, ... \\ \pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l);\ \ \sum_{k=1}^{\infty} \pi_k = 1 \end{aligned} \tag{5}$$



(a) (b)

Figure 3: (a) Graphical model for the stick-breaking construction. (b) Representation of the construction of the random vector π.

The random probability measure G is an atomic distribution, a discrete distribution constructed as a weighted sum ($\pi \sim GEM(\alpha)$) of point masses ($\phi \sim H$), the Figure 4a shows an example of the atomicity of G, this type of distribution forces the clustering property from the Dirichlet process where draws values θ_i from G tend to be repeated given π_k.

As mentioned in the section of Ferguson's definition of Dirichlet processes, larger values of the concentration parameter α give more even weights to draws from the base measure H, if these draws from H have more even weights then the density of draws will be dictated more like the base measure, if we were to integrate through a set of the partition of Θ for samples G we could see that $E[G(A_i)] = H(Ai)$, Figure 4d, 4c.

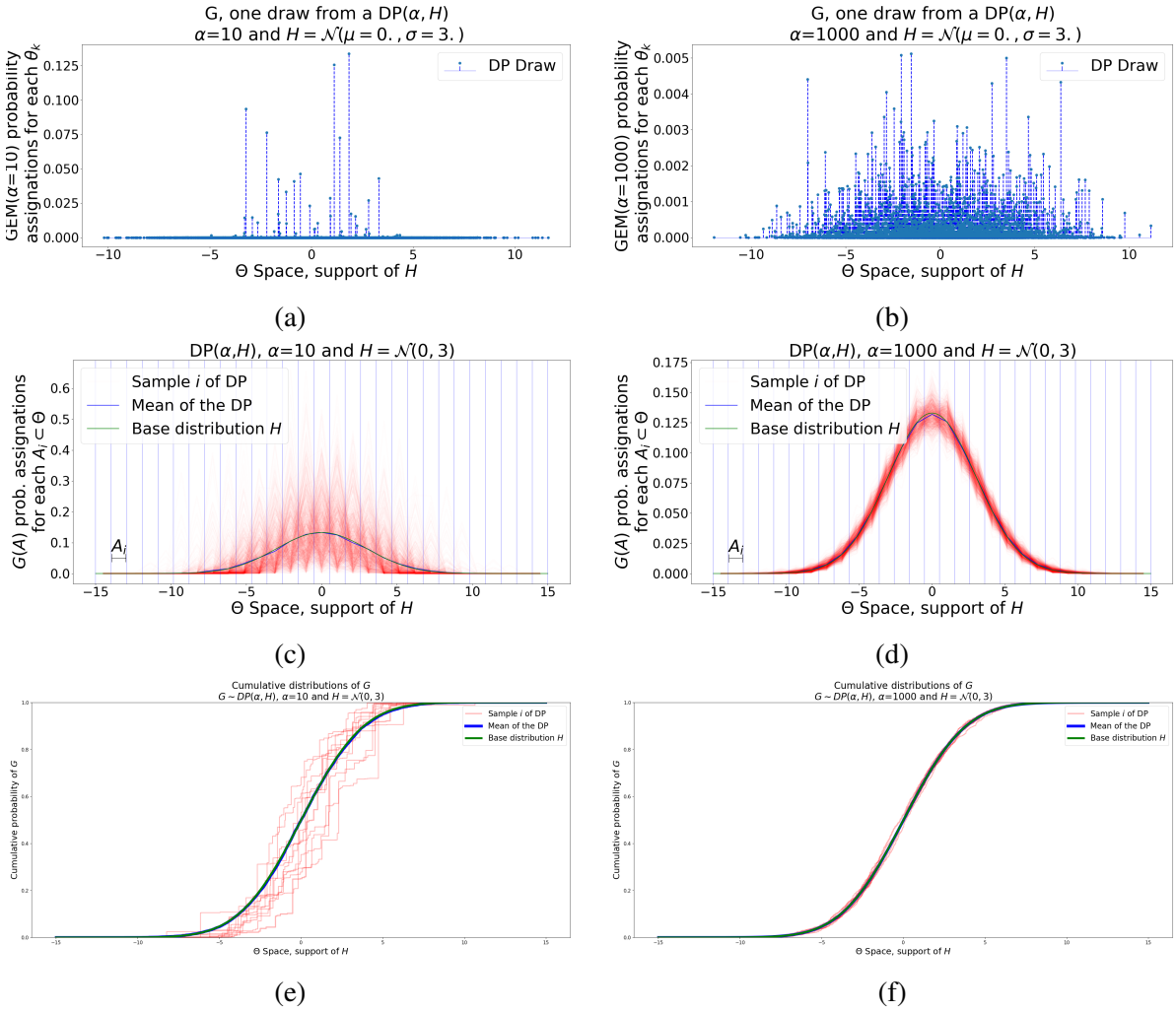


Figure 4: (a-b) One draw of a Dirichlet process, G is a discrete distribution, an atomic distribution of a weighted π_k sum of point masses (ϕ_k). (c-d) Visualization of several draws from a Dirichlet process, its mean $E[G(A_i)]$, and the base measure H (in this particular case a Gaussian distribution). (e-f) Visualization of cumulative distribution of several draws from a Dirichlet process, its mean $E[G(A_i)]$, and the base measure H. Notice the variance of the Dirichlet process increases with smaller α values, as well as most of the probability weight is concentrated in a smaller set of point masses.

3 Exchangeability

An infinite sequence of random variables $(\theta_1, \theta_2, ..., \theta_N, ...)$ with a joint distribution p is exchangeable, if for any subset of N random variables and for every possible permutation of the subset σ, the ordering of the random variables does not affect their joint probability:

$$p(\theta_1, ..., \theta_N) = p(\theta_{\sigma(1)}, ..., \theta_{\sigma(N)}) \tag{6}$$

Random variables that independent and identically distributed are exchangeable but exchangeable random variables are not necessarily independent and identically distributed, the Polya urn is an example of this.

3.1 De Finetti's and Hewitt-Savage theorem

De Finett's theorem explains the mathematical relationship between independence and exchangeability.

Theorem: [6, 7] Given infinite sequence of random variables $\{\theta_i\}_{i=1}^{\infty},\ \theta i \in \Theta$, the joint distribution p of any N random variables $\theta_1, ..., \theta_N$ is exchangeable if and only if there's a random probability measure G with distribution Q:

$$p(\theta_1, ..., \theta_N) = \int_\Theta Q(G, \theta_1, ..., \theta_N) dG = \int_\Theta Q(G) \prod_{i=1}^{N} Q(\theta_i / G) dG \tag{7}$$

According to De Finetti's, the exchangeable sequence $\theta_1, ..., \theta_N$ implies an underlying random probability measure G making them independent and identically distributed.

A more intuitive interpretation can be that an infinite exchangeable sequence of random variables $\{\theta_i\}_{i=1}^{\infty},\ \theta i \in \Theta$, $\theta_1, ..., \theta_N$ are independent and identically distributed given the exchangeable σ-algebra G (a random partition of Θ). In our case the prior $P(G)$ is the Dirichlet process $DP(\alpha, H)$.

3.2 Polya urn

Imagine an urn with two types of color balls, black and white. Initially there are b black and w white balls. We do n draws of balls and for each draw, add an additional ball of the same color as the retrieved to the urn, we put back the ball retrieved.

The probability at the n draw goes as follows:

$$\begin{aligned} p(\theta_n = black/\theta_1, ..., \theta_{n-1}) &= \frac{a_{b,n-1}}{a_{b,n-1} + a_{w,n-1}} \\ p(\theta_n = white/\theta_1, ..., \theta_{n-1}) &= \frac{a_{w,n-1}}{a_{b,n-1} + a_{w,n-1}} \\ a_{b,n-1} &= b + \sum_{i=1}^{n-1} 1\{\theta_i = black\} \\ a_{w,n-1} &= w + \sum_{i=1}^{n-1} 1\{\theta_i = white\} \end{aligned} \tag{8}$$

If we take the number of draws to ∞, the conditional probability for the n_{th} draw will behave as a Beta distribution[8]

$$\lim_{n\to\infty} \frac{a_{b,n-1}}{a_{b,n-1} + a_{w,n-1}} = \rho_b \sim Beta(b, w) \tag{9}$$

Calculating the joint probability of the sequence of draws, we can see that it does not depend on the ordering, it is exchangeable:

$$\begin{aligned} p(\theta_1, ..., \theta_n; n_b, n_w) &= \frac{[b]_1^{n_b}[w]_1^{n_w}}{[b+w]_1^{n_b+n_w}} \\ n_b &\equiv Number\ of\ black\ balls \\ n_w &\equiv Number\ of\ white\ balls \\ [x]_b^a &= x \cdot (x+b) \cdot ... \cdot (x + (a-1) \cdot b) \end{aligned} \tag{10}$$

The Polya urn is a distribution over exchangeable sequences, in which each draw of the sequence is dictated by the ratio of the current balls, this ratio is distributed by an underlying Beta distribution with parameters $\alpha = b$, $\beta = w$. The Polya urn allows to get samples of exchangeable sequences without having to compute samples ρ_b of the beta distribution $Beta(b, w)$, Figure

5:

$$\theta_1, ..., \theta_n \sim p(\theta_1, ..., \theta_n; n_b, n_w) = PolyaUrn(b, w; n_b, n_w)$$
$$equivalent\ to$$
$$\rho_b \sim Beta(b, w)$$
$$\theta_1, ..., \theta_n \overset{i.i.d}{\sim} Cat(\rho_b, 1 - \rho_b) \tag{11}$$

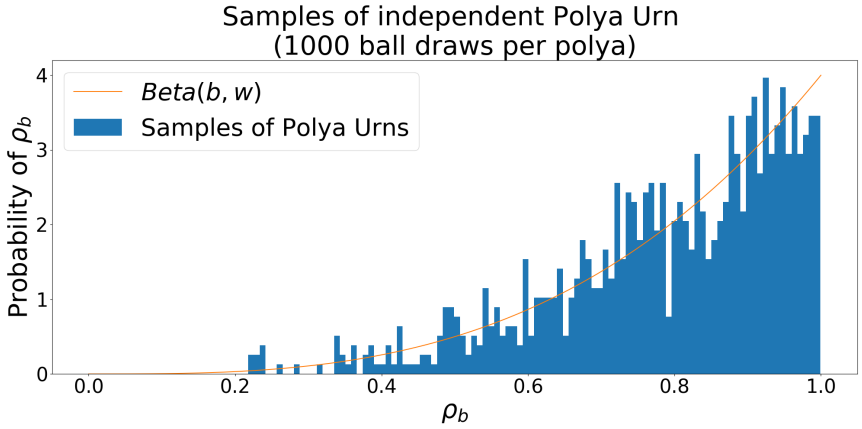


Figure 5: Visualization of 2000 samples of independent Polya urns after 1000 ball draws for each $Polya(b = 4, w = 1)$, when $n\ ball\ draws \to \infty$ in a Polya urn run, the color ball draws are distributed as categorical draws from a sample of a $Beta$ distribution. The Polya urn shows how an underlying probability measure $q(G) = Beta(b, w)$ controls i.i.d draws of color balls given a sample $G \sim q(G)$.

In a multivariate case of a Polya urn, the urn will have balls of multiple colors, if we extrapolate the previous logic to this case:

$$\lim_{n\to\infty} \frac{(a_{b,n-1}, a_{w,n-1}, a_{g,n-1}, a_{o,n-1})}{a_{total}} = (\rho_b, \rho_w, \rho_g, \rho_o) \sim Dirichlet(b, w, g, o)$$
$$a_{b,n-1} = b + \sum_{i=1}^{n-1} 1\{\theta_i = black\}; a_{w,n-1} = w + \sum_{i=1}^{n-1} 1\{\theta_i = white\}$$
$$a_{g,n-1} = b + \sum_{i=1}^{n-1} 1\{\theta_i = green\}; a_{oi,n-1} = w + \sum_{i=1}^{n-1} 1\{\theta_i = orange\} \tag{12}$$
$$a_{total} = a_{b,n-1} + a_{w,n-1} + a_{g,n-1} + a_{o,n-1}$$

The same conclusions apply to here that in the binary case of the Polya urn.

Relation to De Finetti's theorem

From Equation 10, we can see that the probability of sequence $\theta_1, ..., \theta_N$ is exchangeable, so in the context of De Finetti's theorem we can think about an underlying random probability measure that defines the probability of drawing a black ball:

$$p(\theta_1 = c_1, ..., \theta_N = c_N) = \int q(G) \prod_{i=1}^{N} G(C_i) dG \tag{13}$$

Where c_i is a color $black$ or $white$, and G a probability measure over $\{black, white\}$, $q(G)$ is a $Beta(b, w)$ with initial b black balls and w white black balls.

3.3 Blackwell-MacQueen urn

Imagine an urn with just one initial 'magical' ball. We do n ball draws on this urn, for each draw, if we retrieve the 'magic' ball we add a ball of a new color ($\phi \sim H$) and if we retrieve a ball of any color we add another ball of the same color, in both case we put back the ball retrieved.

Consider a sequence of n draws, the probability of draws a balls of a particular color ϕ_1 is given by Equation 14 and the probability of drawing a new color by Equation 15:

$$p(\theta_n = \phi_1 / \theta_1, ..., \theta_{n-1}; \alpha, H) = \frac{\alpha H(\phi_1) + \sum_{i=1}^{n} \delta_{\theta_i}(\phi_1)}{\alpha + n - 1} \tag{14}$$

$$p(\theta_n = new\ \phi / \theta_1, ..., \theta_{n-1}; \alpha, H) = \frac{\alpha H(\Theta \setminus \phi_1, ..., \phi_k)}{\alpha + n - 1} \tag{15}$$

This 'magic' ball approximates remaining infinite color of balls not draw before. The probability of a sequence then is defined as follows:

$$\begin{gathered} p(\theta_1, ..., \theta_n; \alpha, H) = \prod_{i=1}^{n} p(\theta_i / \theta_1, ..., \theta_{i-1}) = \frac{\alpha^K \prod_{k=1}^{K} (n_k - 1)!}{\alpha \cdot (\alpha + 1) \cdot ... \cdot (\alpha + n - 1)} \prod_{k=1}^{K} H(\phi_k) \\ n_k = \sum_{i=1}^{n} 1\{\theta_i = \phi_k\} \end{gathered} \tag{16}$$

As in the case of the Polya urn, the distribution over a sequence is exchangeable, the ordering does not matter in the probability of a sequence. It also provides a way of obtaining samples of exchangeable sequences distributed as $GEM(\alpha)$ without having to sample directly from that distribution.

Relation to Polya urn

As an example, imagine a possible 5 draw scenario as in Figure 6, if we consider the distribution over a binary sequence of draws for 'red' and 'not red', this is the Polya urn, the sequence of draws doesn't matter in the probability of the sequence.

$$(Number\ of\ balls\ of\ \phi\ color,\ Number\ of\ balls\ of\ \neq \phi) = PolyaUrn(1, \alpha) \tag{17}$$

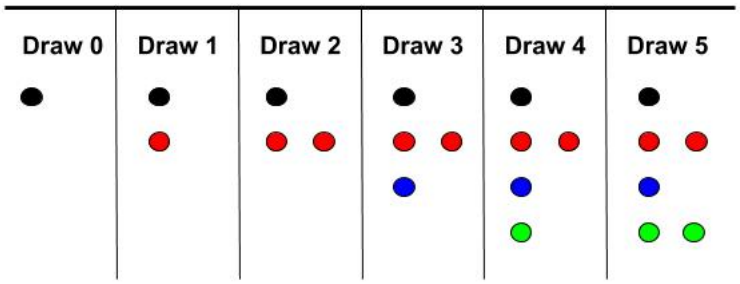


Figure 6: Representation of an exchangeable sequence of draws from a Blackwell-MacQueen urn, where the black ball represents the 'magical' ball from which we draw new colors.

Relation to Stick-breaking construction

Continuing with the previous example in Figure 6, if the probability of a binary sequence of draws 'red'/'not red' behaves like a Polya urn, then the conditional probability is distributed as $V_{red} \sim Beta(1, \alpha)$ when we have $n \to \infty$ draws. Extending this idea to more colors, if we were to draw a 'blue', we will need not to draw a 'red' ball first $(1 - V_r)$ and draw the

new color 'blue' ($\phi_k \sim H$), and with $n \to \infty$ draws, the probability of drawing a blue ball will be $\pi_{blue} = V_{blue}(1 - V_{red})$. Following this idea to infinite colors, this corresponds to the stick-breaking construction.

An alternative approach on describing this relationship will be the following: Take the conditional probability of θ_n given $\theta_1, ..., \theta_{n-1}$ from Blackwell-MacQueen urn and ϕ_k being the unique values that θ's may have. Let's consider $n \to \infty$, we will do ∞ color draws from H and given the nature of this conditional distribution, we will also have repetitions in the color choosing (clustering):

$$p(\theta_n/\theta_1, ..., \theta_{n-1}; \alpha, H) = \frac{\alpha H(.) + \sum_{i=1}^{n} \delta_{\theta_i}(.)}{\alpha + n - 1} \overset{n\to\infty}{\longrightarrow} \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(.)$$
$$\pi_k = \lim_{n\to\infty} n_k/n; \;\; n_k = \sum_{i=1}^{n} \delta(\phi_k, \theta_i) \tag{18}$$

As the number of draws from the urn increases, the initial weight of the base measure H becomes negligible compared to number of repetitions of the unique values ϕ_k (clustering property), this corresponds to the stick-breaking construction.

4 Chinese restaurant process

4.1 Random Partitions

Consider a sample space S = {Ducati, BMW, Triumph, Honda, Harley-Davidson, Royal-Enfield}, we can define this sample space as different partitions ϱ of it, disjoint and non-empty subsets whose union is the sample space:

- ϱ_1 = { {Ducati, BMW}, {Triumph, Honda, Harley-Davidson}, {Royal-Enfield} }
- ϱ_2 = { {Ducati}, {Triumph, Honda, Harley-Davidson, BMW}, {Royal-Enfield} }
- ϱ_3 = { {Ducati, Royal-Enfield}, {Triumph, Honda, Harley-Davidson,BMW} }

Additionally, consider the set of all possible partitions of the sample space S, $\mathcal{P}_S = \{\varrho_1, \varrho_2, ...\}$. We can define random variables that takes values in $\mathcal{P}_S$, these are partitions of the samples space S.

4.2 Definition

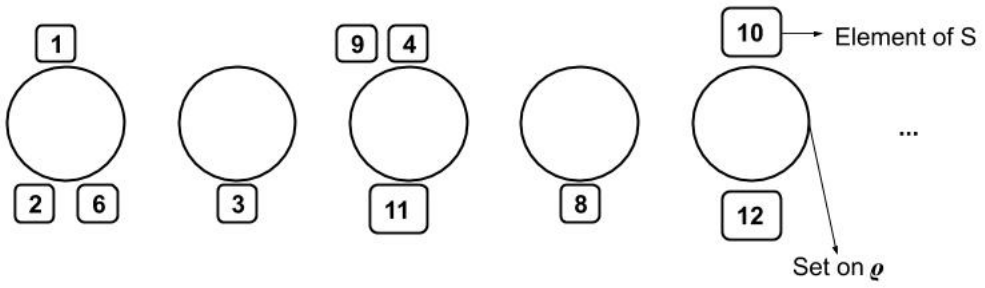


Figure 7: Chinese restaurant process representation with 12 customers and 4 tables, each customer is an element of the sample space S and each table is a set on the possible partition ϱ of the sample space.

Consider a Chinese restaurant that has an infinite number of tables, each of the tables has no limit on the number of customer seated at that table. The first customer seats at the first table, the second customer will seat to the first table with a proportion of $\frac{1}{\alpha+1}$ and choose a new table with the proportion $\frac{\alpha}{\alpha+1}$. Extrapolating this process until the customer n, this customer will seat at any of the occupied tables with the probability of equation 19, and at a new table with probability of equation 20:

$$P(cust_n = table\ c / cust_1, .., cust_{n-1}) = \frac{n_c}{\alpha + \sum\limits_{c\ \in \varrho} n_c} \tag{19}$$

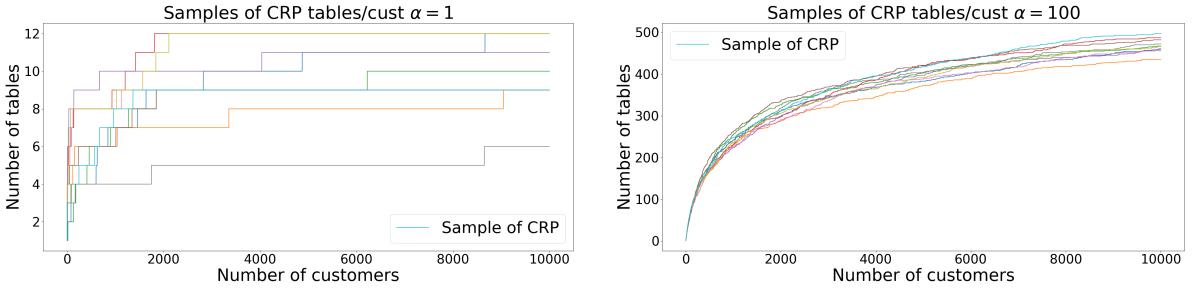
$$P(cust_n = new\ table / cust_1, .., cust_{n-1}) = \frac{\alpha}{\alpha + \sum\limits_{c\ \in \varrho} n_c} \tag{20}$$

$$where \sum_{c\ \in \varrho} n_c = n - 1$$

The Chinese restaurant process defines a prior over partitions of the sample space S, it takes the clustering property of the Dirichlet process without the base measure H, we can see the clustering property in the probability of table assignations, tables with more customers tend to get more customers $P(customer\ n\ at\ table\ c) \propto n_C$ [3].

The number of tables depends on the number of customers n and the concentration parameter α, both the mean and variance are logarithmically proportional to the the number of customer scaled by the concentration parameter, we can see this behavior in the Figure 8.

$$
\begin{aligned}
E[|\varrho|; n, \alpha] &= \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} = \alpha(\psi(\alpha + n) - \psi(\alpha)) \\
&\simeq \alpha log(1 + n/\alpha)\ for\ n, \alpha \gg 0 \\
Var[|\varrho|; n, \alpha] &= \alpha(\psi(\alpha + n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha + n) - \psi'(\alpha)) \\
&\simeq \alpha log(1 + n/\alpha)\ for\ n > \alpha \gg 0 \\
&where\ \psi(.)\ is\ the\ digamma\ function.
\end{aligned}
\tag{21}
$$



(a) (b)

Figure 8: In this Figure we can see how the number of customers n and α influence the mean $E[|\varrho|; n, \alpha]$ and $Var[|\varrho|; n, \alpha]$ variance in the number of tables induced by the Chinese restaurant process. Both increase logarithmically with n scale by α and proportional to α.

4.3 Exchangeability: Exchangeable partition probability function (EPPF)

Following the definition of probabilities for customer assignations to tables, we can define the joint probability of customer assignations, this is equivalent to the probability of a partition of

the sample space, defined as the following:

$$P(\varrho;\alpha) = P(cust_1, cust_2, ..., cust_N;\alpha) = \frac{\alpha^{|\varrho|}\prod_{c\in\varrho}(n_c-1)!}{\alpha\cdot ...\cdot(\alpha+N-1)} = \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(\alpha+N)}\prod_{c\in\varrho}\Gamma(n_c) \quad (22)$$

The Chinese restaurant process satisfies exchangeability. The joint probability of customer assignation to tables does not depend on the order, in other words the probability of a partition ϱ does not depend on the order, it does not depend on the identities of elements in the samples space S either, it only depends on the number of tables ($|\varrho|$) and customer per table (n_c), both conditioned by α and total number of customers N.

In the following example 23, we can see that even though the ordering of the incoming customers is different, both partitions have the same probability. As mentioned before the identities of elements in the samples space S do not matter:

$$\begin{gathered} P(\varrho_1 = \{\{cust_1, cust_2, cust_3\}, \{cust_7, cust_5\}, \{cust_6, cust_4\}\};\alpha) = \\ P(\varrho_2 = \{\{cust_3, cust_7, cust_5\}, \{cust_1, cust_6\}, \{cust_2, cust_4\}\};\alpha) \\ P(\varrho_1;\alpha) = \frac{\alpha}{\alpha}\cdot\frac{1}{\alpha+1}\cdot\frac{2}{\alpha+2}\cdot\frac{\alpha}{\alpha+3}\cdot\frac{\alpha}{\alpha+4}\cdot\frac{1}{\alpha+5}\cdot\frac{2}{\alpha+6} \\ P(\varrho_2;\alpha) = \frac{\alpha}{\alpha}\cdot\frac{\alpha}{\alpha+1}\cdot\frac{\alpha}{\alpha+2}\cdot\frac{1}{\alpha+3}\cdot\frac{1}{\alpha+4}\cdot\frac{1}{\alpha+5}\cdot\frac{2}{\alpha+6} \end{gathered} \quad (23)$$

4.4 Relation to Blackwell-MacQueen urns

As we did on the relation between the Blackwell-MacQueen and Polya urns, we can consider any binary sequence of siting at a particular table or siting at any other, the probability of distribution is the same as in Polya urn.

The differences between the CRP and Blackwell-MacQueen urn are subtle, Blackwell-MacQueen representation gets a new color (table in CRP) from $\phi \sim H$, the CRP creates a partition without labeling the tables or clusters. Additionally, the CRP is a distribution on random partitions (distribution over $\pi = (\pi_1, ..., \pi_k)$), not a distribution over sequences as the

Blackwell-MacQueen urn.

Recall the Blackwell-MacQueen urn probability of a sequence of draws, it is equal to the partition induced by a CRP and probability measures of the 'color' draws:

$$p(\theta_1, ..., \theta_n; \alpha, H) = \frac{\alpha^K \prod_{k=1}^{K}(n_k - 1)!}{\alpha \cdot (\alpha + 1) \cdot ... \cdot (\alpha + n - 1)} \prod_{k=1}^{K} H(\phi_k) = CRP(\varrho; \alpha) \prod_{k=1}^{K} H(\phi_k) \quad (24)$$

4.5 Relation to Stick-breaking process

We can draw the same analogy as in the Blackwell-MacQueen urn, where the probability of siting at the first table when we have $n \to \infty$ costumers behaves as $\pi_1 = V_1; V_1 \sim Beta(1, \alpha)$, second table $\pi_2 = V_1(1 - V_2); V_2 \sim Beta(1, \alpha)$, extending this idea to all possible tables.

5 Posterior distribution

Given the definition of Dirichlet process $DP(\alpha, H)$ and independent observation samples θ from G:

$$\begin{gathered} G \sim DP(\alpha, H) \\ \theta_i \sim G \; for \; i = 1, ..., N \end{gathered} \quad (25)$$

We are interested in computing the posterior distribution of $p(G/\theta)$:

$$p(G/\theta) = \frac{p(\theta/G)p(G)}{p(\theta)};\; p(\theta) = \int_{\Theta} p(\theta/G)p(G)dG \quad (26)$$

For a fixed partition $(A_1, ..., A_K)$ of Θ and the base measure H whose support is Θ, $G(A_i)$ is the random probability measure over the subset A_i by $G \sim DP(\alpha, H)$:

$$(G(A_1), ..., G(A_K)) \sim Dirichlet(\alpha H(A_1), ..., \alpha H(A_K)) \quad (27)$$

Let's consider one observation in the J_{th} segment $\theta_1 \in A_j$, given the Dirichlet-Categorical

conjugancy [9]:

$$(G(A_1),...,G(A_K))/\theta_1 \sim Dirichlet(\alpha H(A_1),...,\alpha H(A_j)+1,...,\alpha H(A_K)) \Rightarrow$$

$$(G(A_1),...,G(A_K))/\theta_1 \sim Dirichlet(\alpha H(A_1)+\delta_{\theta_1}(A_1),...,\alpha H(A_j)+\delta_{\theta_1}(A_j),...,\alpha H(A_K)+\delta_{\theta_1}(A_K))$$

$$\delta_{\theta_1}(A_i) = \begin{cases} 1, & \text{if } \theta_1 \in A_i \\ 0, & \text{Otherwise} \end{cases} \tag{28}$$

For N observations $\theta_1,...,\theta_N$:

$$(G(A_1),...,G(A_K))/\theta_1,...,\theta_N \sim Dirichlet(\alpha H(A_1)+n_1,...,\alpha H(A_K)+n_K)$$

$$n_i = \sum_{i=1}^{N} \delta_{\theta_1}(A_i) \Rightarrow$$

$$(G(A_1),...,G(A_K))/\theta_1,...,\theta_N \sim Dirichlet((\alpha+N)[\frac{\alpha H(A_1)+n_1}{(\alpha+N)}],...,\frac{\alpha H(A_K)+n_K}{(\alpha+N)}]) \Rightarrow$$

$$(G(A_1),...,G(A_K))/\theta_1,...,\theta_N \sim DP(\alpha+N, \frac{\alpha H(.)+\sum_{i=1}^{N}\delta_{\theta_i}(.)}{\alpha+N}) \tag{29}$$

The posterior's base measure $H' = \frac{\alpha H(.)+\sum_{i=1}^{N}\delta_{\theta_i}(.)}{\alpha+N} = \frac{\alpha}{\alpha+N}H + \frac{N}{\alpha+N}\frac{\sum_{i=1}^{N}\delta_{\theta_i}(.)}{N}$ can be seen as a weighted average of the initial base measure H and the empirical distribution of the observations $\frac{\sum_{i=1}^{N}\delta_{\theta_i}(.)}{N}$. The concentration parameter α dictates the mass associated with the prior, if $\alpha \to 0$ the prior has no weight in the posterior depending only on the empirical distribution, at the same time if $N \gg \alpha$ the observations will dominate in the posterior.

5.1 Relation to Stick-breaking construction

Consider one point mass θ_i with respect to the total space Θ, a partition $(\theta_i, \Theta/\theta)$,

$$(G'(\theta_i), G'(\Theta/\theta_i)) \sim Dirichlet((\alpha+1)\frac{\alpha H(\theta_i)+\delta_{\theta_i}(\theta_i)}{\alpha+1}, (\alpha+1)\frac{\alpha H(\Theta/\theta_i)+\delta_{\theta_i}(\Theta/\theta_i)}{\alpha+1}) \Rightarrow$$

$$(G'(\theta_i), G'(\Theta/\theta_i)) \sim Dirichlet(\alpha H(\theta_i)+1, \alpha(1-H(\theta_i))) \Rightarrow$$

$$Total\ measure\ H(\Theta/\theta_i) \simeq 1,\ then\ H(\theta_i) \simeq 0$$

$$(G'(\theta_i), G'(\Theta/\theta_i)) \sim Dirichlet(1, \alpha) \tag{30}$$

Going further in the partition $(\theta_i, A_1, ..., A_K)$ and using the agglomerative and decimative property from Dirichlet distributions:

$$\begin{gathered}(G'(\theta_i), G'(A_1), ..., G'(A_K)) \sim Dirichlet(1, \alpha H(A_1), ..., \alpha H(A_K)) \Rightarrow \\ (V_1', (1 - V_1')G(H(A_1), ..., (1 - V_1')G(H(A_K)) \sim Dirichlet(1, \alpha H(A_1), ..., \alpha H(A_K)) \Rightarrow \\ V_i' \sim Beta(1, \alpha)\end{gathered} \tag{31}$$

Given the stick-breaking construction, we can define G':

$$\begin{gathered}G' = V_i'\delta_\theta + (1 - V_i')G; \;\; V_i' \sim Beta(1, \alpha) \\ G = \sum_{k=1}^{\infty} \pi_k \delta(\phi, \phi_k)\end{gathered} \tag{32}$$

We can see this argument as redimensioning of our prior Dirichlet process G (renormalizing G) when we get our observations $\theta_1, ...\theta_n$, resulting in the posterior Dirichlet process G'. In Figure 9 and 10, we can see the posteriors behavior using the stick-breaking construction, sampling the observations from an assumed 'true' posterior Dirichlet process. We can see that as we get more observations, proportionally to α in the prior, our posterior Dirichlet process moves to closer to the observations' base measure.

6 Predictive posterior distribution

Given our Dirichlet process $DP(\alpha, H)$ and observations sequence $\theta_1, ..., \theta_n \overset{i.i.d}{\sim} G$, we are interested in the predictive posterior θ_{n+1} conditioned on our observations $\theta_1, ..., \theta_n$ with G marginalized out, for a subset $A_i \subset \Theta$:

$$\begin{gathered}p(\theta_{N+1} \in A_i / \theta_1, ..., \theta_N) = \int_\Theta p(\theta_{N+1}/G) p(G/\theta_1, ..., \theta_N) dG = \\ E_{\theta_{N+1} \sim p(\theta_{N+1}/G(A_i))}[G(A_i)/\theta_1, ..., \theta_N]\end{gathered} \tag{33}$$

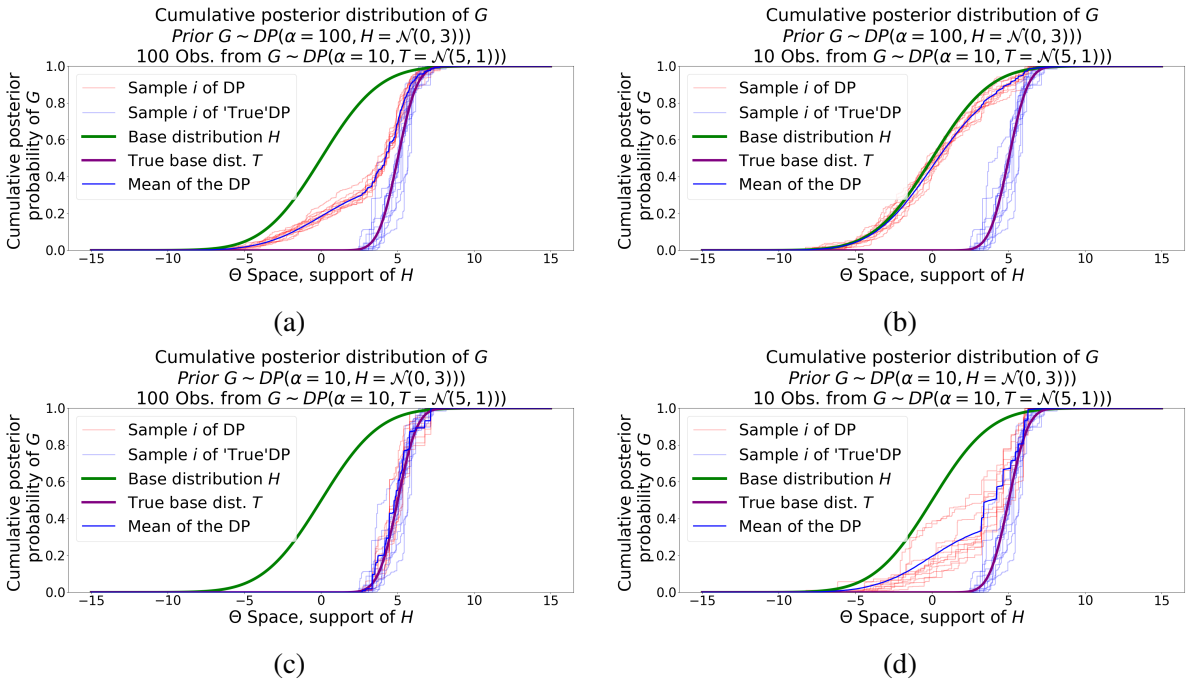


Figure 9: Visualization of cumulative of G samples from posterior of the Dirichlet process for different priors and number of observations. (a) Prior $\alpha = 100$ and 100 Observations, (b) Prior $\alpha = 100$ and 10 Observations, (c) Prior $\alpha = 10$ and 10 Observations, (d) Prior $\alpha = 10$ and 10 Observations. We sample the observations from an assumed 'true' posterior Dirichlet process $DP(\alpha = 10, H = \mathcal{N}(5, 1))$. We can see that as we get more observations, proportionally to α in the prior, our posterior Dirichlet process moves to closer to the observations' base measure.

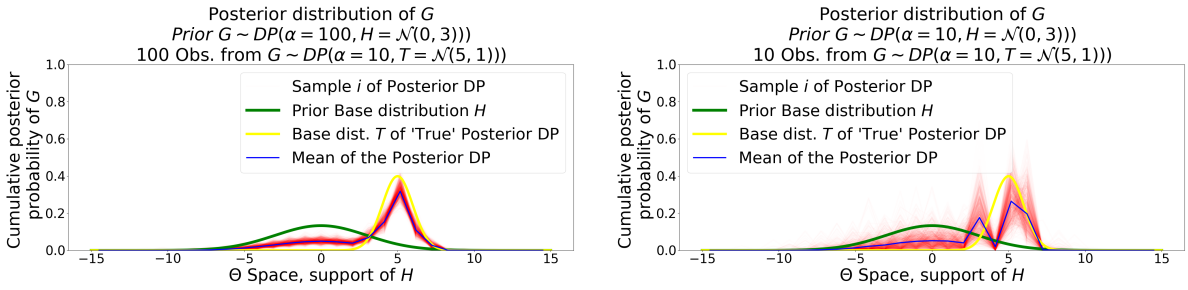
6.1 Relation to Blackwell-MacQueen urn and CRP

The Blackwell-MacQueen urn gives samples of exchangeable sequences distributes as $GEM(\alpha)$ without having to sample directly from that distribution, marginalizing out the random probability measure G:

$$\theta_{N+1}/\theta_1, ..., \theta_N \sim \frac{\alpha H(.) + \sum_{i=1}^{N} \delta_{\theta_i}(.)}{\alpha + N}$$

$$p(\theta_1, ..., \theta_{N+1}; \alpha, H) = \frac{\alpha^{K'} \prod_{k=1}^{K'} (n_k - 1)!}{\alpha \cdot (\alpha + 1) \cdot ... \cdot (\alpha + n - 1)} \prod_{k=1}^{K'} H(\phi_k) = CRP(\varrho; \alpha) \prod_{k=1}^{K'} H(\phi_k)$$

$$n_k = \sum_{i=1}^{n} 1\{\theta_i = \phi_k\} \tag{34}$$



(a) (b)

Figure 10: Visualization of the posterior of a Dirichlet process for different priors and number of observations. (a) Prior $\alpha = 100$ and 100 Observations, (b) Prior $\alpha = 10$ and 10 Observations. We sample the observations from an assumed 'true' posterior Dirichlet process $DP(\alpha = 10, H = \mathcal{N}(5, 1))$. We can see that as we get more observations, proportionally to α in the prior, our posterior Dirichlet process moves to closer to the observations' base measure.

7 Dirichlet Process Mixture Model

Given the properties of the Dirichlet process in clustering data and exchangeability, one of its most common applications is in a mixture model for density estimation. The advantages from this approach as compared to a finite mixture model, come from having a model that increases the complexity with data (number of clusters), and having a exchangeable sequence of observations that makes the model unaffected by the existence of unobserved data.

We can define a Dirichlet process mixture model as a set of observations $(x_1, ..., x_n)$ modeled from a set of latent variables $(\theta_1, ..., \theta_n)$, each latent variable θ_i is independent and identically distributed given the random probability measure G:

$$\begin{aligned} G &\sim DP(\alpha, H) \\ \theta &\overset{i.i.d}{\sim} G \\ x_i &\sim F(./\theta_i) \end{aligned} \tag{35}$$

We can use the different representations of a Dirichlet process to work with this model, Figure 11. The CRP mixture model allows us to work with a finite number of clusters over the theory that there is an infinite number of them, it makes it explicit that this is a clustering

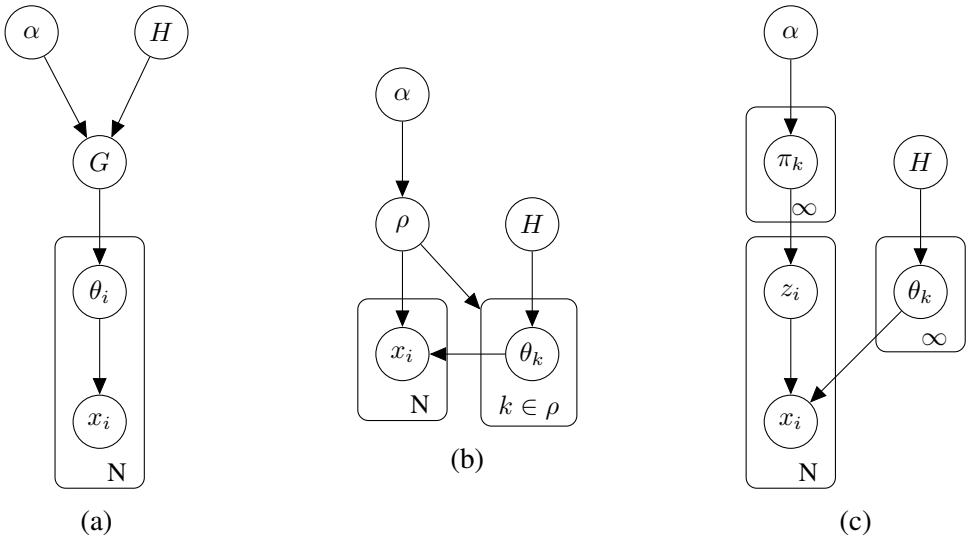


Figure 11: Dirichlet Process Mixture Model graphical models: (a) Dirichlet Process Mixture Model, (b) DPMM through Chinese Restaurant Process, and (c) DPMM through stick-breaking construction.

model. Inference is done through MCMC[10], forcing us to have conjugancy between the base measure H and likelihood F:

$$\begin{aligned} \rho &\sim CRP(|n|, \alpha) \\ \theta_c &\sim H \; for \; each \; cluster \; c \in \rho \\ x_i &\sim F(./\theta_c) \; for \; i \in c \end{aligned} \tag{36}$$

The stick-breaking representation of DPMM uses an explicit representation of G, having an infinite number of elements, it has the advantage of being able to work with non-conjugate likelihood F and base measure H. In this case, inference can be done through MCMC[10] or variational inference[11], but we need to truncate our representation of G to a finite number of elements:

$$\begin{aligned} \pi &\sim GEM(\alpha), \;\; \theta_k \sim H \\ z_i &\sim Categorical(\pi) \\ x_i &\sim F(./\theta, z_i) \end{aligned} \tag{37}$$

7.1 Limit of finite mixture models

Consider a finite mixture model as in Figure 12, where we have the priors $\pi \sim Dirichlet(\alpha)$, $\theta_k \sim H$, latent variables $z_i/\pi \sim Categorical(\pi)$, and likelihood $x_i/z_i, \theta_{z_i} \sim F(./\theta_{z_i})$. Each cluster has its own parameter θ_c, each set of observations x_i assigned to c with mixing probability π_c, resulting into random partitions with less or equal to K clusters.

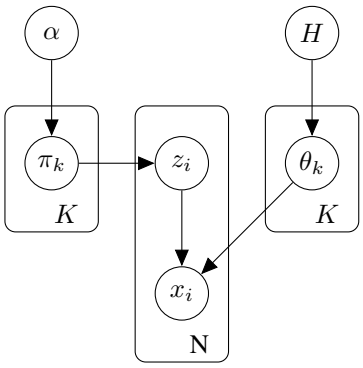


Figure 12: Finite mixture model graphical representation where $\pi \sim Dirichlet(\alpha/K, ..., \alpha/K)$ and $z_i \sim Categorical(\pi)$.

Considering the Dirichlet-Categorical conjugancy, we can derive the joint distribution of z_i and π, posterior distribution, and marginal distribution over z:

$$
\begin{aligned}
p(\pi, z; \alpha) = p(\pi; \alpha) \cdot \prod_{i=1}^{n} p(z_i/\pi) = \frac{\Gamma(\sum_K \alpha_k)}{\prod_K \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \prod_{k=1}^{K} \pi_k^{n_k}; n_k = \sum_{i=1}^{n} 1\{z_i = k\} \\
p(\pi/z; \alpha) \propto p(\pi, z; \alpha) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1 + n_k} \\
p(\pi/z; \alpha) = \frac{\Gamma(\sum_K \alpha_k + n_k)}{\prod_K \Gamma(\alpha_k + n_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k + n_k - 1} \\
p(z; \alpha) = \frac{p(\pi, z; \alpha)}{p(\pi/z; \alpha)} = \frac{\Gamma(\sum_K \alpha_k)}{\Gamma(\sum_K \alpha_k + n_k)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)}
\end{aligned}
\tag{38}
$$

In the following sections, we derive Gibbs sampling for a finite mixture model on the posterior cluster assignments $p(z_i/z_{-i}, x; \alpha, H)$, and assume $K \to \infty$ to show the relationship

between the finite mixture model and Dirichlet process mixture mode, its limitations and why the Chinese restaurant process or Stick-breaking construction offer a better representation.

Gibbs and Collapsed Gibbs sampling

On Gibbs sampling, we derive the complete conditionals for each posterior variable assuming the remaining constant:

$$\begin{aligned} p(z_i = k/z_{-i}, x, \theta, \pi; \alpha, H) &\propto p(z_i = k/\pi; \alpha) \cdot p(x_i/z_i = k, \theta) = \pi_k f(x_i/\theta_k) \\ p(\pi/z_i, x, \theta; \alpha, H) &\propto p(\pi; \alpha) \prod_{i=1}^{n} p(z_i/\pi) \Rightarrow \pi/z_i, x, \theta; \alpha \sim Dir(\frac{\alpha}{K} + n_1, ..., \frac{\alpha}{K} + n_K) \\ p(\theta_k/\theta_{-k}, z, x, \pi; \alpha, H) &\propto p(\theta; H) \prod_{i=1}^{n} p(x_i/z_i = k, \theta) = H(\theta_k) \prod_{j:z_j=k} f(x_j/\theta_k) \end{aligned} \tag{39}$$

This sampling method can be more efficient if we marginalize out the variables π, and θ (Collpased Gibbs sampling), leaving us just with the conditional probability of each latent assignation to a cluster, although we are limited by making the assumption of the likelihood

$f(x/\theta)$ and prior $H(\theta)$ being conjugates:

$$p(z_i/z_{-i}, x; \alpha, H) \propto p(z_i, x/z_{-i}; \alpha, H) = \int_{\theta,\pi} p(z_i, x, \pi, \theta/z_{-i}; \alpha, H) d\theta d\pi =$$
$$\int_{\theta,\pi} p(z_i, \pi/z_{-i}; \alpha, H) p(x, \theta/\pi, z_i, z_{-i}; \alpha, H) d\theta d\pi =$$
$$\int_{\pi} p(z_i, \pi) p(\pi/z_{-i}; \alpha) d\pi \int_{\theta} p(x, \theta/z_i, z_{-i}; \alpha, H) d\theta =$$
$$p(z_i/z_{-i}; \alpha) \int_{\theta} p(x, \theta/z_i, z_{-i}; \alpha) d\theta =$$
$$\prod_{k=1}^{K} \left(\frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \right)^{1\{z_i = k\}} \int_{\theta} p(x, \theta/z_i, z_{-i}; H) d\theta = \tag{40}$$
$$\prod_{k=1}^{K} \left(\frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \right)^{1\{z_i = k\}} \int_{\theta} p(x_i/z_i, \theta) \cdot p(x_{-i}/z_{-i}, \theta) p(\theta; H) d\theta =$$
$$\prod_{k=1}^{K} \left(\frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \right)^{1\{z_i = k\}} \int_{\theta_1, ..., \theta_K} \prod_{k=1}^{K} \left[\prod_{j=1}^{i} f(x_j/\theta_k)^{1\{z_j = k\}} H(\theta_k) \right] d\theta_1, ..., \theta_k =$$
$$\prod_{k=1}^{K} \left(\frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \right)^{1\{z_i = k\}} \prod_{k=1}^{K} \int_{\theta_k} \left[\prod_{j=1}^{i} f(x_j/\theta_k)^{1\{z_j = k\}} H(\theta_k) \right] d\theta_k$$

Infinite limit of mixture model

Taking the complete conditional from the collapsed Gibbs sampler $p(z_i = k/z_{-i}, x; \alpha, H)$:

$$p(z_i = k/z_{-i}, x; \alpha, H) \propto \frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \int_{\theta_k} h(\theta) f(x_i/\theta_k) \prod_{j \neq i: z_j = k} f(x_j/\theta_k) d\theta_k \tag{41}$$

At most, there are $n < K$ occupied clusters, most of the components are empty, lumping the probability of all the empty clusters together:

$$p(z_i = k/z_{-i}, x; \alpha, H) = \frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \int_{\theta_k} h(\theta_k) f(x_i/\theta_k) \prod_{j \neq i: z_j = k} f(x_j/\theta_k) d\theta_k$$
$$p(z_i = k_{empty}/z_{-i}, x; \alpha, H) = (K - K^*) \cdot \frac{\alpha/K}{\alpha + n - 1} \int_{\theta_{k_{empty}}} h(\theta_{k_{empty}}) f(x_i/\theta_{k_{empty}}) d\theta_{k_{empty}}$$
$$where\ K^* \equiv Number\ of\ non\ empty\ clusters;$$
$$(K - K^*) \equiv Number\ of\ empty\ clusters \tag{42}$$

Now assuming that we have $K \to \infty$, the probability for each occupied cluster and the probability of choosing any empty cluster (introducing a new cluster) is defined as:

$$\frac{\alpha/K + n_k^{-i}}{\alpha + n - 1} \overset{K\to\infty}{\longrightarrow} \frac{n_k^{-i}}{\alpha + n - 1} \; and \; \frac{\alpha\frac{K-K^*}{K}}{\alpha + n - 1} \overset{K\to\infty}{\longrightarrow} \frac{\alpha}{\alpha + n - 1}$$

$$p(z_i = k/z_{-i}, x; \alpha, H) = \frac{n_k^{-i}}{\alpha + n - 1} \int_{\theta_k} h(\theta_k) f(x_i/\theta_k) \prod_{j\neq i: z_j = k} f(x_j/\theta_k) d\theta_k \tag{43}$$

$$p(z_i = k_{empty}/z_{-i}, x; \alpha, H) = \frac{\alpha}{\alpha + n - 1} \int_{\theta_{k_{empty}}} h(\theta_{k_{empty}}) f(x_i/\theta_k) d\theta_{k_{empty}}$$

If we consider the prior probability of any of the data points being assigned to any of the cluster when $k \to \infty$, this probability of cluster assignations tends to zero, leaving the mixing proportions close to 0. The Chinese restaurant process and stick-breaking construction offer better defined ways of having an infinite limit mixture model.

Relationship to Chinese restaurant process

In the finite mixture model, the marginal $p(z_1, ..., z_n; \alpha)$ describes a partition of the dataset into clusters, and the labeling of each cluster with a mixture component index:

$$p(z; \alpha) = \frac{p(\pi, z; \alpha)}{p(\pi/z; \alpha)} = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha/K)} \frac{\prod_{k=1}^{K} \Gamma(n_k + \alpha/K)}{\Gamma(n + \alpha)} \tag{44}$$

A Chinese restaurant process induces a distribution over partitions of the dataset without labeling $p(\varrho; \alpha)$, having the following relationship between the two in Equation 45, where the partition ϱ has $|\varrho| = k \leq K$ clusters, each of which can be assigned one of the K labels (without replacement):

$$\begin{gathered} p(z; \alpha) = \frac{1}{K \cdot (K-1) \cdot ... \cdot (K-k-1)} p(\varrho; \alpha) \Rightarrow \\ p(\varrho; \alpha) = [K]_{-1}^{k} \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in \varrho} \frac{\Gamma(n_c + \alpha/K)}{\Gamma(\alpha/K)} \\ where \; [x]_b^a = x \cdot (x + b) \cdot ... \cdot (x + (a-1) \cdot b) \end{gathered} \tag{45}$$

Taking $K \to \infty$ we get a distribution over partitions without labeling as in the Chinese restaurant process:

$$
\begin{aligned}
[K]_{-1}^{k} \overset{K\to\infty}{\longrightarrow} \alpha^{|\varrho|} \; and \; \Gamma(1+\alpha/K) = \frac{\alpha}{K} \cdot \Gamma(\alpha/K) \Rightarrow \\
p(\varrho;\alpha) = \frac{\alpha^{|\varrho|} \cdot \Gamma(\alpha)}{\Gamma(n+\alpha)} \cdot \prod_{c\in\varrho} \frac{\Gamma(n_c+\alpha/K)}{\Gamma(\alpha/K)} \Rightarrow \\
p(\varrho;\alpha) = \frac{\alpha^{|\varrho|} \cdot \Gamma(\alpha)}{\Gamma(n+\alpha)} \cdot \prod_{c\in\varrho} \Gamma(n_c)
\end{aligned}
\tag{46}
$$

The difference between the Chinese restaurant process and a finite mixture model comes from the probability of permutations for labeling the clusters, assuming uniform for labeling and $K \to \infty$. In the infinite mixture model case, the Chinese restaurant process makes no assumptions on the number of clusters and the it's part of the inference process.

7.2 Inference on DPMM

The posterior distributions for the variables θ, z, π in Dirichlet process mixture models cannot be computed analytically, we turn to MCMC techniques [10] or variational inference [11] to do inference in these models. In here, we only review Gibbs sampling and collapsed Gibbs sampling on CRP prior, blocked Gibbs sampling on truncated stick-breaking, and variational inference.

Gibbs sampling for DPMM

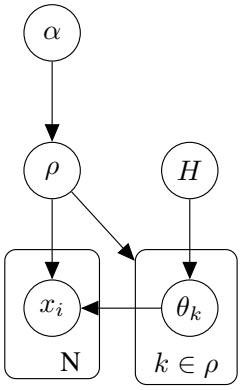


Figure 13: CRP mixture model.

For inference on DPMM with a CRP prior, a common inference algorithm is to use a Gibbs sampling ([10]:Algorithm 2), calculating the complete conditionals of $p(\rho_n/\rho_{-n}, x, \theta; \alpha)$ and $p(\theta_k/x, \rho; H)$.

MCMC algorithms for Dirichlet process mixture models [10, 13] take advantage of the exchangeability property. We can assume that each customer i is the last one to arrive, and iteratively sample cluster

assignations for each customer. Swapping between customers as being the last will not change the probability of the partition ρ.

Gibbs sampling has the drawbacks of limiting us to have conjugancy between the base measure H and the likelihood of our observations F, and being slow because we iterate over each data point sampling cluster assignations:

$$\begin{gathered} p(\rho_n/\rho_{-n}, x, \theta; \alpha, H) \propto p(\rho; \alpha) \prod_{i=1}^{n} p(x_i/\rho, \theta) p(\theta/\rho; H) \propto \\ p(\rho_n/\rho_{-n}; \alpha) \cdot p(x_i/\rho_n, \theta) \Rightarrow \\ p(\rho_n/\rho_{-n}; \alpha) = \begin{cases} \frac{n_k}{n-1+\alpha}, & \text{if } \rho_n = k \in \rho_{-n} \\ \frac{\alpha}{n-1+\alpha}, & \text{if } \rho_n = new \end{cases} \\ p(x_n/\theta, \rho_n) = \begin{cases} F(x_n/\theta_{\rho_n}), & \text{if } \rho_n = k \in \rho_{-n} \\ \int_\theta F(x_n/\theta) H(\theta) d\theta, & \text{if } \rho_n = new \end{cases} \end{gathered} \tag{47}$$

Additionally, we draw samples for new assignations of each cluster θ_k from the posterior distribution:

$$\begin{gathered} p(\theta_k/x, \rho) \propto p(\theta_k, x/\rho) = \prod_{i=1:\rho_i=k}^{n} p(\theta_k, x_i) = p(\theta_k) \prod_{i=1:\rho_i=k}^{n} p(x_i/\theta_k) = \\ H(\theta_k) \prod_{i=1:\rho_i=k}^{n} F(x_i/\theta_k) \end{gathered} \tag{48}$$

The Gibbs sampling algorithm([10]:Algorithm 2) goes as detailed in Algorithm 1.

We could also place a prior on α, typically a gamma distribution since $\alpha > 0$ [12]:

$$\begin{gathered} p(\alpha; a, b) \frac{b^a}{\Gamma(a)} \alpha^{(a-1)} exp\{-b\alpha\} \\ p(\alpha/\rho) \propto p(\alpha) p(\rho/\alpha) = p(\alpha) \frac{\alpha^{|\varrho|} \Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{k \in \varrho} \Gamma(n_k) \end{gathered} \tag{49}$$

In case we place a prior on α or our base measure H and likelihood F were non-conjugates, we could turn to **Gibbs sampling with auxiliary parameters** ([10]:Algorithm 8)

Algorithm 1: Gibbs sampling for Dirichlet process mixture model with CRP prior

Result: Samples from posteriors $p(\rho/x, \theta; \alpha, H),\ p(\theta/x, \rho; H)$

Initialization:

Start current state of Markov Chain: $\rho = (\rho_1, \rho_2, ..., \rho_n)$ and $\theta = (\theta_1, \theta_2, ..., \theta_n)$;

while $p(\rho/x, \theta; \alpha, H),\ p(\theta/x, \rho; H)$ *have not converged* **do**

  for *each data point* $i = 1, ..., n$ **do**

    $k = \rho_i$ current assignation for data point x_i;

    if $n_k = 1$ **then**

      Discard θ_k from current state and $K = K - 1$;

    Draw a new cluster assignation: $p(\rho_i/\rho_{-i}, x, \theta; \alpha, H)$ as in Equation 47;

    if ρ_i *value does not correspond to any cluster* θ_k **then**

      Draw new value θ_k from $p(\theta/x, \rho; H)$ as in Equation 48;

      Add θ_k to the state and $K = K + 1$;

  end

  for *each cluster* $\theta_k;\ \ k = 1, ..., K.$ **do**

    Draw value from $\theta_k \sim p(\theta_k/x, \rho; H)$ as in Equation 48

  end

end

Collapsed Gibbs sampling for DPMM

We could go further and marginalize out the cluster parameters θ_k:

$$p(\rho_n = k/\rho_{-n}, x; \alpha, H) \propto p(\rho_n = k, \rho_{-n}; \alpha) p(x/\rho) \propto p(\rho_n = k/\rho_{-n}; \alpha) p(x/\rho_n = k, \rho_{-n}) \Rightarrow$$

$$p(\rho_n/\rho_{-n}; \alpha) = \begin{cases} \frac{n_k}{n-1+\alpha}, & \text{if } \rho_n = k \in \rho_{-n} \\ \frac{\alpha}{n-1+\alpha}, & \text{if } \rho_n = new \end{cases}$$

$$p(x/\rho_n = k, \rho_{-n}) = \int_\theta p(x/\rho_n = k, \rho_{-n}, \theta) p(\theta/\rho; H) d\theta = \int_\theta p(\theta/\rho; H) \prod_{i=1}^{n} p(x_i/\rho_i \theta) d\theta =$$

$$\int_\theta p(x_n/\rho_n = k, \theta) \prod_{i=1}^{n-1} p(x_i/\rho_i, \theta) \prod_{p=1}^{|\varrho|} p(\theta_p; H) d\theta =$$

$$\int_{\theta_k} p(x_n/\rho_n = k, \theta) \prod_{i=1:\rho_i=k}^{n-1} p(x_i/\rho_i, \theta) p(\theta_k; H) d\theta_k \int_\theta \prod_{i=1:\rho_i \neq k}^{n-1} p(x_i/\rho_i, \theta) \prod_{p=1:p\neq k}^{|\varrho|} p(\theta_p; H) d\theta \propto$$

$$\int_{\theta_k} f(x_n/\theta_k) \prod_{i=1:\rho_i=k}^{n-1} f(x_i/\theta_k) H(\theta_k) d\theta_k$$

$$p(\rho_n = k/\rho_{-n}, x; \alpha, H) \propto \begin{cases} \frac{n_k}{n-1+\alpha} \int_{\theta_k} f(x_n/\theta_k) \prod_{i=1:\rho_i=k}^{n-1} f(x_i/\theta_k) H(\theta_k) d\theta_k, & \text{if } \rho_n = k \in \rho_{-n} \\ \frac{\alpha}{n-1+\alpha} \int_{\theta_k} f(x_n/\theta_k) H(\theta_k) d\theta_k, & \text{if } \rho_n = new \end{cases} \tag{50}$$

The collapsed Gibbs sampling algorithm([10]:Algorithm 3) goes as detailed in Algorithm 2.

Algorithm 2: Collpased Gibbs sampling for Dirichlet process mixture model with CRP prior

Result: Samples from posteriors $p(\rho/x, \theta; \alpha, H)$
Initialization:
    Start current state of Markov Chain: $\rho = (\rho_1, \rho_2, ..., \rho_n)$;
while $p(\rho/x, \theta; \alpha, H)$ *have not converged* **do**
    for *each data point* $i = 1, ..., n$ **do**
        Draw a new cluster assignation: $p(\rho_i/\rho_{-i}, x; \alpha, H)$ as in Equation 50;
    end
end

Blocked Gibbs sampling on truncated DPMM

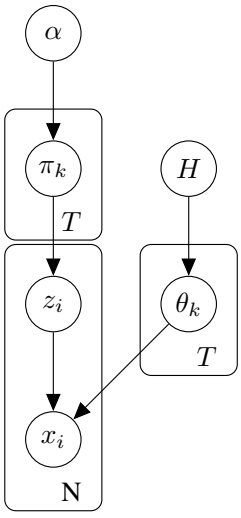


Figure 14: CRP mixture model.

In this representation we use the stick-breaking construction to model our Dirichlet process mixture model, and use a truncation value for the number of clusters T, $\pi_k = 0$ for $k > T$. The truncation value T does not have to be very large to get a good approximation [13]:

$$G = \sum_{k=1}^{T} \pi_k \delta_{\theta_k} \tag{51}$$

In this case we are interested in the posteriors for the variable cluster assignations z_i, cluster variable parameters θ, and Beta variables β that

construct π:

$$
\begin{aligned}
p(z_i = k/z, \pi, \theta, x; \alpha, H) &\propto \pi_k p(x_i \theta_k) \\
p(\pi/z, x, \theta; \alpha, H) \propto p(\pi; \alpha) p(z/\pi) = p(\pi; \alpha) \prod_{i=1}^{n} p(z_i/\pi) &\Rightarrow \\
p(\pi/z, x, \theta; \alpha, H) = Dirichlet(\alpha/T + n_1, ..., \alpha/T + n_T) & \\
p(\theta_k/\theta_{-k}, z, \pi; \alpha, H) = H(\theta_k) \prod_{j:z_j=k} f(x_j/theta_k) &
\end{aligned}
\tag{52}
$$

The blocked Gibbs sampling algorithm [13] is detailed in Algorithm 52.

Algorithm 3: Blocked Gibbs sampling for truncated Dirichlet process mixture model

Result: Samples from posteriors $p(z/\pi, \theta, x; \alpha, H)$, $p(\pi/z, x, \theta; \alpha, H)$, and $p(\theta/z, \pi; \alpha, H)$

Initialization:

Start current state of Markov Chain: $z = (z_1, z_2, ..., z_n)$, π, and $\theta = (\theta_1, \theta_2, ..., \theta_T)$;

while $p(\rho/x, \theta; \alpha, H)$ *have not converged* **do**

  for *each data point* $i = 1, ..., n$ **do**

    Draw a new cluster assignation: $p(z_i/z, \pi, \theta, x; \alpha, H)$ as in Equation 52;

  end

  for *each cluster* $k = 1, ..., T$ **do**

    Compute $n_k = \sum_{i=1}^{n} 1\{z_i = k\}$

  end

  Draw a value of probability vector $\pi \sim Dirichlet(\alpha/T + n_1, ..., \alpha/T + n_T)$;

  for *each cluster* $k = 1, ..., T$ **do**

    Draw a value of cluster parameters: $p(\theta_k/\theta_{-k}, z, \pi; \alpha, H)$ as in Equation 52;

  end

end

Variational inference on DPMM

Variational inference on Dirichlet process mixture models makes three key assumptions, the observable data is drawn from an exponential family, the base distribution is a conjugate prior,

and the variational approximation is truncated to a value T. Additionally, we assume the mean-field approximation and CAVI algorithm, which dictate our variational approximations as the form:

$$q(z_i; \lambda_i) \propto exp\{E_{z_{-i}}[log\ p(z_i/z_{-i}, x)]\} \tag{53}$$

where $z_i$are latent variables and λ_i its corresponding parameters

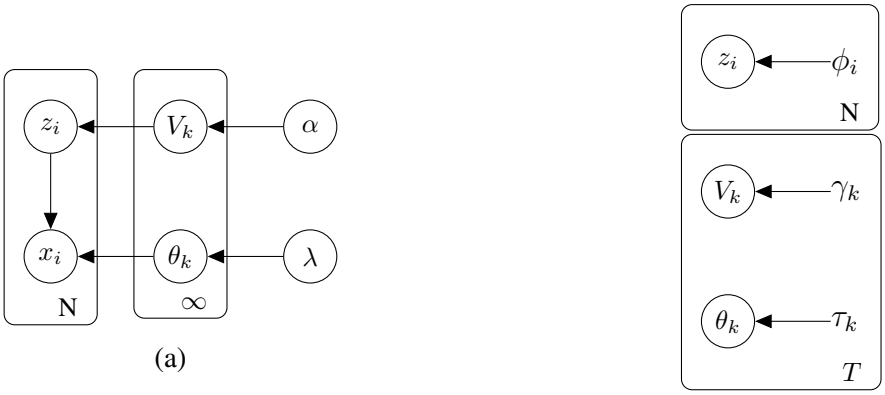


Figure 15: (a)Dirichlet process mixture model graphical model of exponential families, (b) Variational approximation of DPMM.

We can define our model and joint distribution as the following, notice that we truncate the variational approximation, not the model:

$$\begin{gathered}
p(x, z, V, \theta; \alpha, H) = p(V; \alpha) \cdot p(\theta; \lambda) \prod_{i=1}^{n} p(x_i/\theta_{z_i}) p(z_i/V) \\
p(V; \alpha) = \prod_{i=1}^{\infty} p(V_i; \alpha) = \prod_{i=1}^{\infty} \frac{1}{\mathcal{B}(1, \alpha)} (1 - V_i)^{(\alpha - 1)} \\
p(\theta; \lambda) = h(\theta) exp\{\lambda^T t(\theta) - a(\lambda)\};\ \ \lambda = [\lambda_1, \lambda_2];\ t(\theta) = [\theta, -a(\theta)] \\
p(z_i/V) = \prod_{i=1}^{\infty} (\pi_k)^{1\{z_I = k\}} = \prod_{i=1}^{\infty} (V_k \prod_{j=1}^{k-1} (1 - V_j))^{1\{z_i = k\}} \\
p(x_i/\theta_{z_i}) = h(x_i) exp\{\theta_{z_i}^T x_i - a(\theta_{z_i})\} = \prod_{i=1}^{\infty} (h(x_i) exp\{\theta_k^T x_i - a(\theta_k)\})^{1\{z_i = k\}}
\end{gathered} \tag{54}$$

And complete conditionals for the variational approximations:

$$p(\theta_k/\theta_{-k}, V, z, x; \lambda) \propto p(\theta_k; \lambda) \prod_{i=1}^{n} p(x_i/\theta_{z_i})^{1\{z_i=k\}} =$$
$$h(\theta_k) exp\{\lambda^T t(\theta_k) + a(\lambda)\} \prod_{i=1}^{n} [\prod_{j=1}^{\infty} (h(x_i) exp\{\theta_k^T x_i - a(\theta_k)\})^{1\{z_i=j\}}]^{1\{z_i=k\}}$$

$$p(z_i/z_{-i}, \theta, V, x) \propto p(z_i/V) p(x_i/\theta_{z_i}) =$$
$$\prod_{k=1}^{\infty} (V_k \prod_{p=1}^{k-1} (1 - V_p))^{1\{z_i=k\}} \tag{55}$$

$$p(V_k/V_{-k}, z, x, \theta) \propto p(V_j; \alpha) \prod_{i=1}^{n} p(z_i/V)^{1\{z_i \geq k\}} =$$
$$(1 - V_j)^{\alpha} \prod_{i=1}^{n} [\prod_{j=1}^{\infty} (V_j \prod_{p=1}^{j-1} (1 - V_p))^{1\{z_i=k\}}]^{1\{z_i \geq k\}}$$

Now that we have define the complete conditionals, we can derive the variational approximations:

$$q(V, \theta, z; \gamma, \tau, \phi) = \prod_{k=1}^{T-1} q(V_k; \gamma_k) \prod_{k=1}^{T} q(\theta_k; \tau_k) \prod_{i=1}^{n} q(z_i; \phi_i) \tag{56}$$

$$\boldsymbol{q(\theta_k;\tau_k)} \Rightarrow$$

$$E_{\theta_{-j},z,x,V}[log\ p(\theta_k/\theta_{-k},z,x,V)] \propto \lambda^T t(\theta_k) + a(\lambda) + \sum_{i=1}^{n} E_{z_i}[1\{z_i=k\}](\theta_k^T x_i - a(\theta_k)) =$$

$$\lambda_1\theta_k + \lambda_2 a(\theta_k) + \sum_{i=1}^{n} q(z_i=k;\phi_{ik})(\theta_k^T x_i - a(\theta_k)) =$$

$$\theta_k^T(\lambda_1 + \sum_{i=1}^{n} q(z_i=k;\phi_{ik})x_i) + a(\theta_k)(\lambda_2 - \sum_{i=1}^{n} q(z_i=k;\phi_{ik}))) \Rightarrow$$

For $k=1,...,T$ Exponential distributions:

$$q(\theta_k;\tau_k) = h(\theta_k)exp\{\tau_k^T t(\theta_k) + a(\theta_k)\};\ \ t(\theta_k) = [\theta_k, -a(\theta_k)]$$

$$\tau_{k,1} = \lambda_1 + \sum_{i=1}^{n} q(z_i=k;\phi_{ik})x_i$$

$$\tau_{k,2} = \lambda_2 - \sum_{i=1}^{n} q(z_i=k;\phi_{ik}) \tag{57}$$

$$\boldsymbol{q(V_k;\gamma_k)} \Rightarrow$$

$$E_{V_{-k},z,x,\theta}[log\ p(V_k/V_{-k},z,x,\theta)] \propto$$

$$[\alpha - 1 + \sum_{i=1}^{n}\sum_{j=k+1}^{T} q(z_i=j)]log(1-V_k) + \sum_{i=1}^{n} q(z_i=k;\phi_{ik})\cdot logV_j =$$

$$log(1-V_k)^{[\alpha-1+\sum_{i=1}^{n}\sum_{j=k+1}^{T} q(z_i=j)]} + logV_j^{\sum_{i=1}^{n} q(z_i=k;\phi_{ik})} \Rightarrow$$

For $k=1,...,T-1$ Beta distributions: (58)

$$q(V_k;\gamma_k) = \frac{1}{\mathcal{B}(\gamma_{k,1},\gamma_{k,2})} V_i^{(\gamma_{k,1}-1)}(1-V_i)^{(\gamma_{k,2}-1)}$$

$$\gamma_{k,1} = \sum_{i=1}^{n} q(z_i=k;\phi_{ik})$$

$$\gamma_{k,2} = \alpha + \sum_{i=1}^{n}\sum_{j=k+1}^{T} q(z_i=j;\phi_{ij})$$

$$\boldsymbol{q(z_i;\phi_i)} \Rightarrow$$

$$E_{z-i,V,x,\theta}[log\ p(z_i/z_{-i},\theta,V,x)] \propto$$

$$\sum_{k=1}^{T} 1\{z_i=k\}[E_{V_k}[logV_k] + \sum_{p=1}^{k-1} E_{V_p}[log(1-V_p)] + E_{\theta_k}[\theta_k]x_i - E_{\theta_k}[a(\theta_k)]] \Rightarrow$$

$$\textit{For } i=1,...,n \textit{ Categorical distributions:} \tag{59}$$

$$q(z_i;\phi_i) = \frac{1}{\sum_{k=1}^{T}\phi_{i,k}} \prod_{k=1}^{T} \phi_{i,k}^{1\{z_i=k\}}$$

$$\phi_{i,k} = exp\{E_{V_k}[logV_k] + \sum_{p=1}^{k-1} E_{V_p}[log(1-V_p)] + E_{\theta_k}[\theta_k]x_i - E_{\theta_k}[a(\theta_k)]]\}$$

Finally, we can use the truncated variational approximations to compute the predictive distribution:

$$p(x_{n+1}/x;\alpha,\lambda) = \int_{V,\theta} p(x_{n+1}/\theta,V)p(V,\theta/x;\alpha,\lambda)dVd\theta =$$

$$\int_{V,\theta} [\sum_{k=1}^{\infty} \pi_k(V)p(x_{n+1}/\theta_k)]p(V,\theta/x;\alpha,\lambda)dVd\theta \Rightarrow \tag{60}$$

$$p(x_{n+1}/x;\alpha,\lambda) \simeq \sum_{k=1}^{T} E_{V\sim q}[\pi_k(V)]E_{\theta_k\sim q}[p(x_{n+1}/\theta_k)]$$

As mentioned in the Gibbs sampling sections, we could place a prior on $\alpha \sim Gamma(\alpha;b,c)$, in this case we use a stick-breaking construction and the $Gamma$ distribution is conjugate to

the $Beta$ distribution:

$$p(\alpha; b, c) = \frac{1}{\alpha} exp\{-c \cdot \alpha + b \cdot log(\alpha) - a(b, c)\}$$

$$a(b, c) = log(\Gamma(b)) - b \cdot log(c)$$

$$p(\alpha/V; b, c) \propto p(\alpha; b, c) \prod_{k=1}^{T-1} p(V_k/\alpha) = p(\alpha; b, c) \prod_{k=1}^{T-1} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} (1 - V_k)^{(\alpha-1)} \Rightarrow$$

$$log\, p(\alpha/V; b, c) \propto log(1/\alpha) - c \cdot \alpha + b \cdot log(\alpha) + \sum_{k=1}^{T-1} [log(\alpha) + (\alpha - 1) log(1 - V_k)] \propto$$

$$log(1/\alpha) + log(\alpha)[b + T - 1] + \alpha[-c + \sum_{k=1}^{T-1} log(1 - V_k)]$$

$$\boldsymbol{q(\alpha; w_1, w_2) \Rightarrow}$$

$$q(\alpha; w_1, w_2) = \frac{1}{\alpha} exp\{-w_2 \cdot \alpha + w_1 \cdot log(\alpha) - a(w_1, w_2)\}$$

$$w_1 = b + T - 1$$

$$w_2 = -c + \sum_{k=1}^{T-1} E_{V_k}[log(1 - V_k)] \tag{61}$$

Additionally, we need to replace α with $E_\alpha[\alpha] = w_1/w_2$ in the updates for the variational approximation $q(V_k; \gamma_k)$.

Appendix

Dirichlet distribution

Multivariate probability density function parametrized by a vector $\alpha = (\alpha_1, ..., \alpha_k); \alpha_i > 0$, multivariate generalization of the $Beta$ distribution. Often used as a prior for categorical or

multinomial distributions.

$$\begin{aligned} \pi = (\pi_1, \pi_2, ..., \pi_k) &\sim Dirichlet(\alpha); \\ p(\pi_1, \pi_2, ..., \pi_k) &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}; \\ Support \;\; S_k &= \{\pi = (\pi_1, \pi_2, ..., \pi_k); \pi_i \geq 0; \sum_{i=1}^{k} \pi_k = 1\} \end{aligned} \tag{62}$$

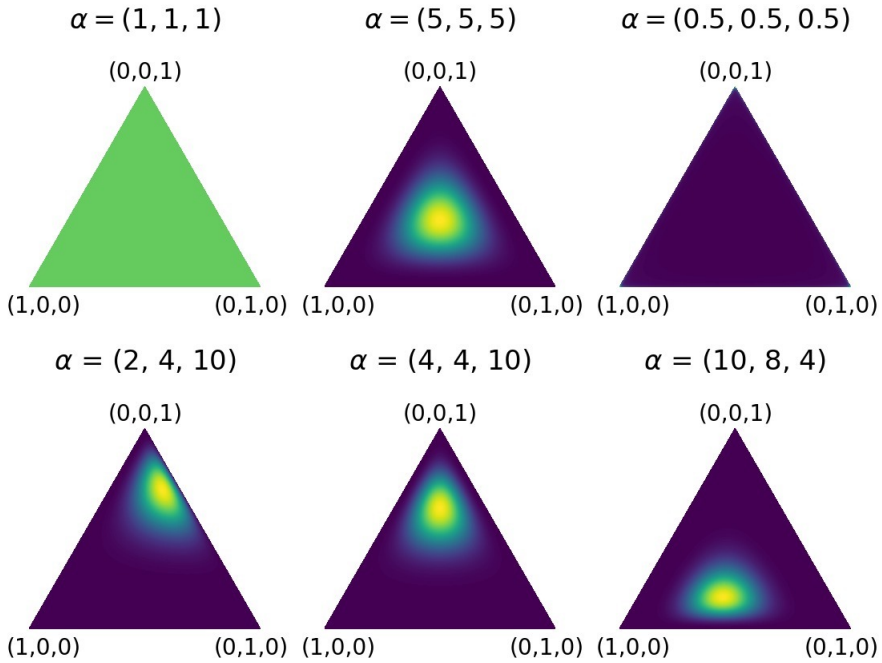


Figure 16: Density plots of $Dirichlet$ distributions for different values of vector α.

We can see the Dirichlet distribution as a probability distribution over a set of K-dimensional discrete distributions. Although the support S_k is defined in a K-dimensional space, because of the constrain $\pi_i \geq 0;\ \sum_{i=1}^{k} \pi_k = 1$, it lies in a $K-1$-dimensional object. We show an example if the Figure 16, where the samples π are 3-dimensional but all of them lie in the 2-dimensional plane.

Marginal of a Dirichlet process

Marginal distributions $\pi_i \sim p(\pi_i)$ from a Dirichlet distribution are $Beta$ distributions:

$$(\pi_1, ..., \pi_K) \sim Dirichlet(\alpha_1, ..., \alpha_K); \;\; \mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$p(\pi_j) = \frac{\pi_j^{\alpha_j - 1}}{\mathcal{B}(\alpha)} = \int_0^{1-\pi_j} \pi_1^{\alpha_1 - 1} \int_0^{1-\pi_j-\pi_1} \pi_2^{\alpha_2 - 1} ... \int_0^{1-\sum_{i=1}^{K-2}\pi_i} \pi_{K-1}^{\alpha_1 - 1}(1 - \sum_{i=1}^{K-1}\pi_i)^{\alpha_K - 1} d\pi_{K-1} \cdot ... \cdot d\pi_1 \tag{63}$$

To solve this integral we can do a change of variable:

$$\begin{gathered}
\pi_{K-1} = z(1 - \sum_{i=1}^{K-2}\pi_i) \Rightarrow d\pi_{K-1} = dz(1 - \sum_{i=1}^{K-2}\pi_i); \\
\int_0^{1-\sum_{i=1}^{K-2}\pi_i} \pi_{K-1}^{\alpha_1 - 1}(1 - \sum_{i=1}^{K-1}\pi_i)^{\alpha_K - 1} d\pi_{K-1} = \\
(1 - \sum_{i=1}^{K-2}\pi_i)^{\alpha_K + \alpha_{K-1} - 1} \int_0^1 z^{\alpha_{K-1} - 1}(1-z)^{\alpha_K - 1} dz = \\
(1 - \sum_{i=1}^{K-2}\pi_i)^{\alpha_K + \alpha_{K-1} - 1} \cdot \mathcal{B}(\alpha_{K-1}, \alpha_K) \int_0^1 Beta(z; \alpha_{K-1}, \alpha_K) dz = \\
(1 - \sum_{i=1}^{K-2}\pi_i)^{\alpha_K + \alpha_{K-1} - 1} \cdot \mathcal{B}(\alpha_{K-1}, \alpha_K)
\end{gathered} \tag{64}$$

Doing this for remaining π_i variables and noticing that $\mathcal{B}(\alpha_{K-1}, \alpha_K) \cdot \mathcal{B}(\alpha_{K-2}, \alpha_{K-1} + \alpha_K) = \mathcal{B}(\alpha_{K-2}, \alpha_{K-1}, \alpha_K)$:

$$\begin{gathered}
p(\pi_j) = \frac{\mathcal{B}(\alpha_{-j})}{\mathcal{B}(\alpha)} \pi_j^{\alpha_j - 1}(1 - \pi_j)^{\sum_{i=1:i\neq j}^{K}\alpha_i} \\
p(\pi_j) = \frac{1}{\mathcal{B}(\alpha_j, \sum_{i=1:i\neq j}^{K}\alpha_i)} \pi_j^{\alpha_j - 1}(1 - \pi_j)^{\sum_{i=1:i\neq j}^{K}\alpha_i} \\
p(\pi_j) = Beta(\pi_j; \alpha_j, \sum_{i=1:i\neq j}^{K}\alpha_i)
\end{gathered} \tag{65}$$

Agglomerative Property

$$(\pi_1, \pi_2, ..., \pi_k) \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_k) \Rightarrow (\pi_1 + \pi_2, ..., \pi_k) \sim Dirichlet(\alpha_1 + \alpha_2, ..., \alpha_k)$$

Generally if $(I_1, ...,_j)$ *is a partition (1, ..., k):*

$$(\sum_{i \in I_1} \pi_i, \sum_{i \in I_2} \pi_i, ..., \sum_{i \in I_j} \pi_i) \sim Dirichlet(\sum_{i \in I_1} \alpha_i, \sum_{i \in I_2} \alpha_i, ..., \sum_{i \in I_k} \alpha_i) \tag{66}$$

Decimative Property

$$\begin{gathered} (\pi_1, \pi_2, ..., \pi_k) \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_k) \\ (\tau_1, \tau_2) \sim Dirichlet(\alpha_1 \cdot \beta_1, \alpha_1 \cdot \beta_2) \textit{ with } \beta_1 + \beta_2 = 1 \Rightarrow \\ (\tau_1 \cdot \pi_1, \tau_2 \cdot \pi_2, ..., \pi_k) \sim Dirichlet(\alpha_1 \cdot \beta_1, \alpha_1 \cdot \beta_2, \alpha_2, ..., \alpha_k) \end{gathered} \tag{67}$$

Dirichlet-Categorical conjugancy

Dirichlet distributions are conjugate to categorical and multinomial distributions, in this case we are covering the caterogical likelihood, although it can be easily generalized to the multinomial case. Additionally, we assume symmetrical Dirichlet distribution for simplicity, $\alpha_i = \alpha/K; \sum_{i=1}^{K} \alpha_i = \alpha$, we capture the graphical model in Figure 17:

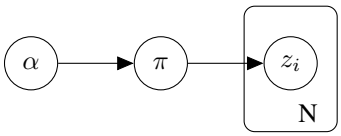


Figure 17: Dirichlet-Categorical conjugancy model where $\pi \sim Dir(\alpha)$ and $z_i \sim \ Cat(\pi)$.

Likelihood $p(z/\pi)$**:**

$$\begin{gathered} p(z/\pi) = \prod_{i=1}^{n} p(z_i/\pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{1\{z_i=k\}} = \prod_{k=1}^{K} \pi_k^{n_k} \\ n_k = \sum_{i=1}^{n} 1\{z_i = k\} \end{gathered} \tag{68}$$

Posterior $p(\pi/z;\alpha)$**:**

$$p(\pi/z;\alpha) \propto p(\pi;\alpha)\prod_{i=1}^{n} p(z_i/\pi) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\alpha/K)}\prod_{k=1}^{K}\pi_k^{\alpha/K-1}\prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{1\{z_i=k\}} =$$
$$\frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\alpha/K)}\prod_{k=1}^{K}\pi_k^{\alpha/K+n_k-1} \tag{69}$$
$$p(\pi/z;\alpha) = \frac{\Gamma(\alpha+n)}{\prod_{k=1}^{K}\Gamma(\alpha/K+n_k)}\prod_{k=1}^{K}\pi_k^{\alpha/K+n_k-1}$$

Marginal $p(z_1,...,z_{i-1},z_i;\alpha)$**:**

$$p(z;\alpha) = \frac{p(\pi,z;\alpha)}{p(\pi/z;\alpha)} = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\prod_{k=1}^{K}\frac{\Gamma(\alpha/K+n_k)}{\Gamma(\alpha/K)}; \Gamma(\alpha+1)=\alpha\Gamma(\alpha) \Rightarrow$$
$$p(z;\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha)\prod_{i=1}^{n}(\alpha-i-1)}\prod_{k=1}^{K}\frac{\Gamma(\alpha/k)\prod_{i=1}^{n_k}(\alpha/k-i-1)}{\Gamma(\alpha/k)} = \tag{70}$$
$$p(z;\alpha) = \frac{\prod_{k=1}^{K}\prod_{i=1}^{n_k}(\alpha/k-i-1)}{\prod_{i=1}^{n}(\alpha-i-1)}$$

Conditional $p(z_i/z_1,...,z_{i-1})$**:**

$$p(z_i,z_{-i};\alpha) = p(z_i/z_{-i};\alpha)\cdot p(z_{-i};\alpha)$$
$$p(z_i/z_{-i};\alpha) = \frac{\prod_{i=1}^{n-1}(\alpha-i-1)}{\prod_{i=1}^{n}(\alpha-i-1)}\cdot\frac{\prod_{k=1}^{K}(\alpha/k-n_K^{-i}+1-1)^{1\{z_i=k\}}\prod_{i=1}^{n_k^{-i}}(\alpha/k-i-1)}{\prod_{k=1}^{K}\prod_{i=1}^{n_k^{-i}}(\alpha/k-i-1)}$$
$$p(z_i/z_{-i};\alpha) = \frac{\prod_{k=1}^{K}(\alpha/k-n_K^{-i}+1-1)^{1\{z_i=k\}}}{\alpha-n-1} \tag{71}$$

Griffiths-Engen-McCloskey (GEM) distribution

The Griffiths-Engen-McCloskey distribution $GEM(\alpha)$ is defined as follows:

$$\rho = (\rho_1,\rho_2,...) \sim GEM(\alpha)\ where$$
$$\rho_k = V_k(\prod_{i=1}^{k-1}(1-V_i);\ \ V_k \overset{i.i.d}{\sim} Beta(1,\alpha) \tag{72}$$

We can draw the analogy of the GEM distribution as an infinite Dirichlet distribution. The parameter α dictates how the mixing proportions are distributed, smaller α values produce more

disproportionate partition of the total probability 1, where the first mixing probabilities account for most of the total probability (Figures 18a, 18c), on the contrary, large α values produce probability vectors with more distributed mixing proportions (Figures 18b, 18d). The mixing proportions of the vector samples $\rho \sim GEM(\alpha)$ decrease in average but not strictly (Figures 18c, 18d).

Concepts of measure theory

Σ-Algebra

Collection of subsets Σ of the sample space Θ with the following properties:

- $\Theta \in \Sigma$
- If $A \in \Sigma$ then $A^c \in \Sigma$
- If $A_{n i=1}^{\infty} \subset \Sigma$ then $\cup_{i=1}^{\infty} A_n \in \Sigma$

Then (Θ, Σ) is a measurable space.

Example 1:

In the context of coin flips, we can have a sample space for 1 coin flip $\Theta = (Head,\ Tails)$ and Σ-Algebra $= \{\{\}, \{H\}, \{T\}, \{H,\ T\}\}$

Example 2:

In the context of coin flips, we can have a sample space for 3 coin flips $\Theta = (HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH, TTT)$. We could create different partitions of Θ:

- Partition 1, $A_1 = \{HHH,\ HHT,,\ THH,\ THT\}, A_2 = \{HTH,\ HTT,\ TTH, TTT\}$, $\Theta = A_1 \cup A_2$: $\Sigma_1 = \{\{\}, A_1, A_2, \{A_1, A_2\}\}$
- Partition 2, $B_0 = \{HHH\}, B_1 = \{HHT,\ THH,\ HTH\}, B_2 = \{HTT,\ TTH\}, B_3 = \{TTT\}$, $\Theta = B_0 \cup B_1 \cup B_2 \cup B_3$: $\Sigma_2 = \{\{\}, \{B_0\}, \{B_1\}, \{B_2\}, \{B_3\}, \{B_0, B_1\}, \{B_0, B_2\},$

$\{B_0, B_3\}, \{B_1, B_2\}, \{B_1 B_3\}, \{B_1, B_2\}, \{B_0, B_1, B_2\}, \{B_0, B_1, B_3\}, \{B_0, B_2, B_3\}, \{B_1, B_2, B_3\},$
$\{B_0, B_1, B_2, B_3\}\}$

Measure

A measure μ over a measurable space (Θ, Σ) is a function $\mu : \Sigma \to [0, \infty]$ with the following properties:

- $\mu(\emptyset) = 0$
- If $\{A_n\}_{n=1}^{\infty} \subset \Sigma$ is disjoint then $\mu(\cup_{n=1}^{\infty} A_n) = \cup_{n=1}^{\infty} \mu(A_n)$

Then (Θ, Σ, μ) is a measurable space.

Probability measure:

Measure μ in measurable space (Θ, Σ, μ) such that $\mu(\Theta) = 1$.

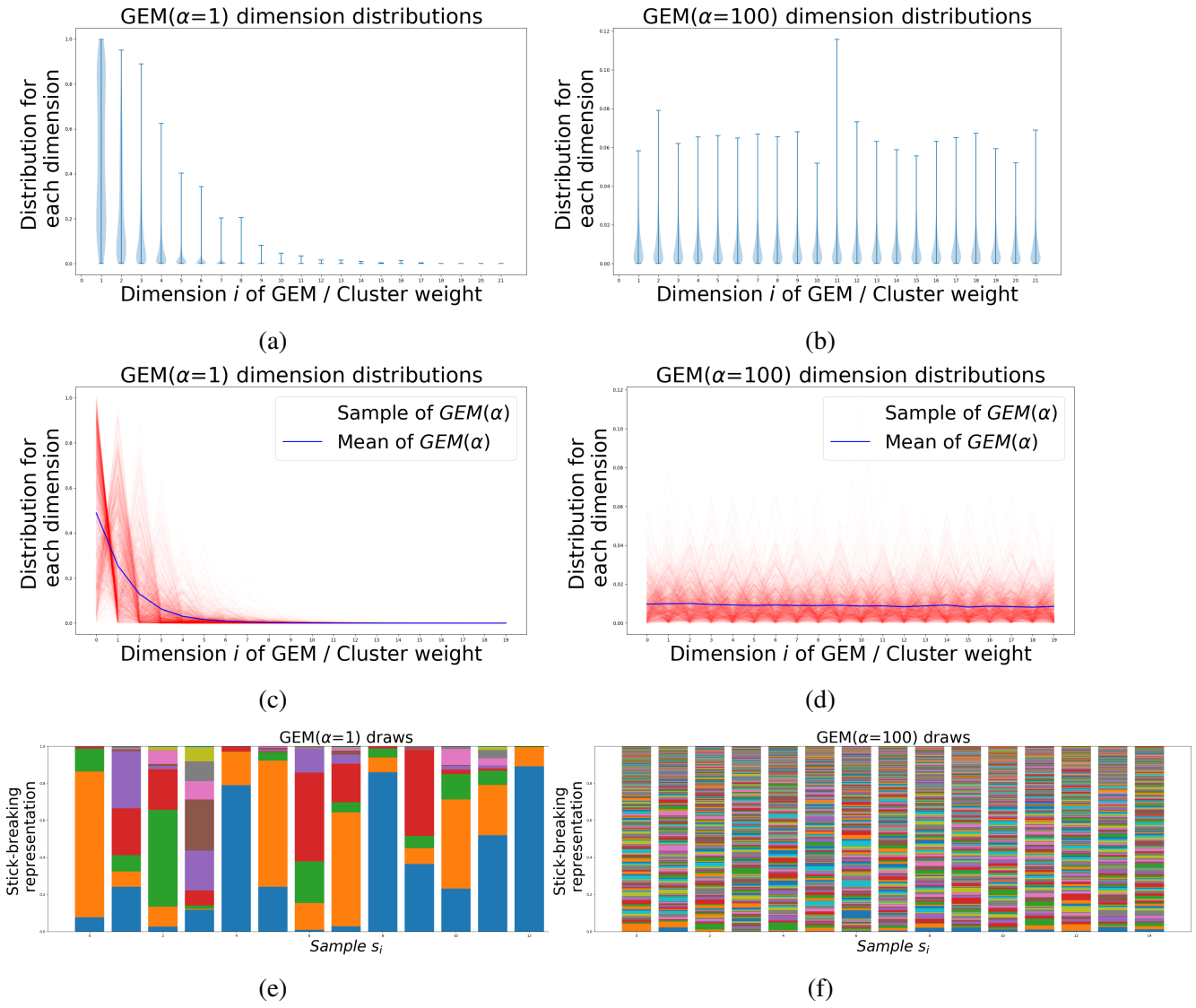


Figure 18: (a-b) Distribution of the mixing proportions ρ_k for 1000 samples of GEM distributions $\alpha = 1, 100$. (c-d) Visualization of samples ρ and mean of 1000 samples of a GEM for $\alpha = 1, 100$. (e-f) Stick-breaking representations for samples of GEM $\alpha = 1, 100$, each color represent a mixing proportion.

References

[1] T. S. Ferguson, *The Annals of Statistics* **1**, 209 (1973).

[2] D. Blackwell, J. B. MacQueen, *The Annals of Statistics* **1**, 353 (1973).

[3] D. J. Aldous (1985), vol. 1117 of *Lecture Notes in Math.*.

[4] J. Sethuraman, *Statistica Sinica* **4**, 639 (1994).

[5] J. Pitman, *Combinatorial stochastic processes*, vol. 1875 of *Lecture Notes in Mathematics* (Springer-Verlag, 2006).

[6] B. de Finetti, *Atti della R. Accademia Nazionale dei Lincei, Ser. 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturali 4* pp. 251–299 (1931).

[7] E. S. Hewitt, L. J. Savage (1955).

[8] D. Blackwell, D. Kendall, *Journal of Applied Probability* **1**, 284 (1964).

[9] M. Schervish, *Theory of Statistics*, Springer Series in Statistics (Springer New York, 1996).

[10] R. M. Neal, *Journal of Computational and Graphical Statistics* **9**, 249 (2000).

[11] D. M. Blei, M. I. Jordan, *Bayesian Anal.* **1**, 121 (2006).

[12] M. D. Escobar, M. West, *Journal of the American Statistical Association* **90**, 577 (1995).

[13] H. Ishwaran, L. F. James, *Journal of the American Statistical Association* **96**, 161 (2001).