An introduction to Bayesian nonparametrics

Daniel Selsam
Outline

- Dirichlet processes (DPs)
- Hierarchical Dirichlet processes (HDPs)
- Pitman-Yor processes (PYs and HPYs)
Dirichlet processes (DPs)

- Definition
- Posterior distribution
- Predictive distribution and the Blackwell-MacQueen urn scheme
- Chinese restaurant process (CRP) view of the DP
- Stick-breaking construction of the DP
- Application: Dirichlet process mixture models
Definition

Let $H$ be a distribution over $\Theta$ and $\alpha$ a positive real number. We say that $G$ is a Dirichlet process distributed with base distribution $H$ and concentration parameter $\alpha$, written

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

if for any finite measurable partition $A_1, \ldots, A_r$ of $\Theta$, the vector $(G(A_1), \ldots, G(A_r))$ is distributed as

$$(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$$
The role of the parameters

- The parameters $H$ and $\alpha$ play intuitive roles in the definition of the DP.
- For any measurable set $A \subset \Theta$, we have
  \[E[G(A)] = H(A)\]
  \[\text{Var}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}\]
- We can view $H$ as the mean of the DP, and we can view the concentration parameter $\alpha$ as an inverse variance.
The posterior distribution

Let $G \sim \text{DP}(\alpha, H)$, and let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from $G$. Let $A_1, \ldots, A_r$ be a finite measurable partition of $\Theta$, and let $n_k$ be the number of observed values in $A_k$.

Then by the conjugacy between the Dirichlet and the multinomial distributions, we have:

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

Let $G \sim \text{DP}(\alpha, H)$, and let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from $G$. Let $A_1, \ldots A_r$ be a finite measurable partition of $\Theta$, and let $n_k$ be the number of observed values in $A_k$. Then by the conjugacy between the Dirichlet and the multinomial distributions, we have:

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

Recall that $G \sim \text{DP}(\alpha, H)$ if for any finite measurable partition $A_1, \ldots A_r$ of $\Theta$, the vector $(G(A_1), \ldots, G(A_r))$ is distributed as

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

What is the posterior distribution of $G$?
The posterior distribution

- A little algebra shows that the posterior is distributed as:

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

- Notice that the posterior base distribution is a weighted average of the prior base distribution \( H \) and empirical distribution. Thus we can interpret \( \alpha \) as the strength of the prior.

- As the number of observations increases, the posterior is dominated by the empirical distribution.
The predictive distribution

Recall that

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

and that for any measurable set \( A \subset \Theta \), we have

\[ \mathbf{E} \left[ G(A) \right] = H(A) \]
The predictive distribution

Let \( G \sim \text{DP}(\alpha, H) \), and let \( \theta_1, \ldots, \theta_n \) be a sequence of independent draws from \( G \). For any measurable set \( A \subset \Theta \), we have:

\[
\text{Prob}(\theta_{n+1} \in A|\theta_1, \ldots, \theta_n) = E[G(A)|\theta_1, \ldots, \theta_n] = ?
\]
The predictive distribution

Let \( G \sim DP(\alpha, H) \), and let \( \theta_1, \ldots, \theta_n \) be a sequence of independent draws from \( G \). For any measurable set \( A \subset \Theta \), we have:

\[
\text{Prob} (\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = \mathbf{E} [G(A) | \theta_1, \ldots, \theta_n] \\
= \frac{1}{\alpha + n} \left( \alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A) \right)
\]
The predictive distribution

Let $G \sim \text{DP}(\alpha, H)$, and let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from $G$. For any measurable set $A \subset \Theta$, we have:

$$\text{Prob}(\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = \mathbf{E}[G(A) | \theta_1, \ldots, \theta_n]$$

$$= \frac{1}{\alpha + n} \left( \alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A) \right)$$

Thus $\theta_{n+1}$ is distributed according to the posterior base measure:

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left( \alpha H + \sum_{i=1}^{n} \delta_{\theta_i} \right)$$
The Blackwell-MacQueen urn scheme

- Recall that:

\[
\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left( \alpha H + \sum_{i=1}^{n} \delta_{\theta_i} \right)
\]

- Helpful metaphor number one: the urn scheme. We can sample from the Dirichlet process \( \text{DP}(\alpha, H) \) in the following manner: consider an urn containing \( \alpha \) colorless balls. To generate a sample, we draw a ball at random from the urn.
  - If the ball is colorless, we select a new color from \( H \).
  - If the ball is colored, we select another ball with the same color.

Either way, we put the new ball and the old ball back into the urn, and continue.
The Chinese restaurant process (CRP)

- Recall that:

\[
\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left( \alpha H + \sum_{i=1}^{n} \delta_{\theta_i} \right)
\]

- Let \( \theta_1^*, \ldots, \theta_m^* \) be the unique values among \( \theta_1, \ldots, \theta_n \), and \( n_k \) the number of repeats of \( \theta_k^* \). The predictive distribution can be equivalently written as:

\[
\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left( \alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*} \right)
\]
The Chinese restaurant process (CRP)

- Let $\theta_1^*, \ldots, \theta_m^*$ be the unique values among $\theta_1, \ldots, \theta_n$, and $n_k$ the number of repeats of $\theta_k^*$. The predictive distribution can be equivalently written as:

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

- Notice that the value $\theta_k^*$ will be repeated by $\theta_{n+1}$ with probability proportional to $n_k$, the number of times it has already been observed. The larger $n_k$ is, the higher the probability that it will grow.

- This is a rich-gets-richer phenomenon, where large clusters (a set of $\theta_i$'s with identical values $\theta_k^*$ being considered a cluster) grow larger faster.
The Chinese restaurant process (CRP)

The unique values of $\theta_1, \ldots, \theta_n$ induce a partition of the set $\{1, \ldots, n\}$ into clusters such that within each cluster, all draws take on the same value. Given that $\theta_1, \ldots, \theta_n$ are random, this induces a random partition of $\{1, \ldots, n\}$.

Thus, we can sample from the Dirichlet process $\text{DP}(\alpha, H)$ in the following manner: first, draw a random partition, and for each cluster in the partition draw a value from $H$. (It is not hard to prove this).
The Chinese restaurant process (CRP)

- This brings us to our second helpful metaphor: the Chinese restaurant process.
- Imagine a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers.
  - The first customer enters the restaurant and sits at the first table.
  - The second customer enters and decides either to sit with the first customer, or by herself at a new table.
  - In general, the \( n + 1 \) customer either joins an already occupied table with probability proportional to the number of customers already sitting there, or sits at a new table with probability proportional to \( \alpha \).
- The dish served at each table will be an i.i.d. draw from the base distribution \( H \).
Let’s consider the distribution of the number of clusters among \( n \) observations.

\[
E[m|n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \\
= \alpha (\psi(\alpha + n) - \psi(\alpha)) \\
\approx \alpha \log \left(1 + \frac{n}{\alpha}\right)
\]

Note that the number of clusters grows (approximately) linearly in \( \alpha \), and only logarithmically in the number of observations.
The stick-breaking construction

- We can construct a draw $G \sim \text{DP}(\alpha, H)$ by first drawing a distribution over a countably infinite number of elements, drawing a countably infinite number of atoms from the base distribution, and writing $G$ as a weighted sum of these atoms.
- This is a mouthful, but it will make sense after we go over the details. In many ways, this turns out to be the most intuitive view of the Dirichlet process.
The stick-breaking construction

- First we will consider the weights, before considering the atoms.
- For \( k = 1, 2, \ldots \), suppose we draw an infinite sequence

\[
\beta_k \sim \text{Beta}(1, \alpha),
\]

and then assign

\[
\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)
\]

- This brings us to our third helpful metaphor: the stick-breaking construction.
The stick-breaking construction

\[ \beta_k \sim \text{Beta}(1, \alpha), \quad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \]

- Starting with a stick of length 1, we break it at \( \beta_1 \), assigning \( \pi_1 \) to be the length of stick we just broke off. Now recursively break the other portion to obtain \( \pi_2, \pi_3 \) and so forth.
- The stick-breaking distribution over \( \pi \) is sometimes written

\[ \pi \sim \text{GEM}(\alpha) \]

- Informally, a draw from \( \text{GEM}(\alpha) \) can be thought of as a countably infinite multinomial distribution.
The stick-breaking construction

- To sample $G \sim \text{DP}(\alpha, H)$, we first sample a countably infinite vector of weights via stick-breaking:
  
  $$\beta_k \sim \text{Beta}(1, \alpha), \quad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

  i.e.
  
  $$\pi \sim \text{GEM}(\alpha)$$

- Next, we sample a countably infinite number of atoms from the base distribution:
  
  $$\theta_k^* \sim H$$

- Finally, we can write $G$ as a weighted sum of these atoms:
  
  $$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$
The stick-breaking construction

To sample $G \sim \text{DP}(\alpha, H)$, we first sample a countably infinite vector of weights via stick-breaking:

$$
\beta_k \sim \text{Beta}(1, \alpha), \quad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)
$$

$$
\theta_k^* \sim H
$$

$$
G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
$$

What is the role of $\alpha$?
Application: Dirichlet process mixture models

- The most common application of the Dirichlet process is in clustering data using mixture models. Here the nonparametric nature of the Dirichlet process translates to mixture models with a countably infinite number of components.
- We model a set of observations \( \{x_1, \ldots, x_n\} \) using a set of latent parameters \( \{\theta_1, \ldots, \theta_n\} \), with each drawn from a Dirichlet process \( G \sim \text{DP}(\alpha, H) \):

\[
\begin{align*}
  G | \alpha, H & \sim \text{DP}(\alpha, H) \\
  \theta_i | G & \sim G \\
  x_i | \theta_i & \sim F(\theta_i)
\end{align*}
\]
Inference for Dirichlet process mixture models

- Collapsed Gibbs
- Block Gibbs
- Mean field
Inference for Dirichlet process mixture models

- The simplest approach to doing inference in Dirichlet process mixture models is via collapsed Gibbs.
- The algorithm is very similar to the collapsed Gibbs algorithm you implemented for the Gaussian mixture model earlier in the course. The only difference is that every time you sample one of the labels, there is a nonzero chance that it does not join any of the existing clusters.

\[
p(z_n = k|z_{-n}, x) \propto p(x_n|x_{-n}, z_{-n}, z_n = k)p(z_n = k|z_{-n})
\]

\[
p(z_n = k|z_{-n}) \propto \begin{cases} 
     n_{k,-i} & \text{if } k \text{ is an existing cluster} \\
     \alpha & \text{if } k \text{ is the next new cluster}
\end{cases}
\]
Inference for Dirichlet process mixture models

- A second approach involves actually sampling (a truncated version of) $G$.
- If we assume that $G$ is only a sum of a finite number of atoms (i.e. that $\beta_K = 1$ for some fixed $K$), we can perform block Gibbs sampling over the following variables:
  - The stick-breaking weights:
    \[ V = \{V_1, \ldots, V_{K-1}\} \]
  - The atoms of $G$:
    \[ \theta = \{\theta_1, \ldots, \theta_K\} \]
  - The labels:
    \[ Z = \{Z_1, \ldots, Z_N\} \]
Each of these updates can be done easily, at least in the case of conjugate priors.

Note that we call this block Gibbs, since each of the three groups can be sampled all at once independently, when conditioned on the other two.
A third method that is applicable to the Dirichlet process mixture model is mean field.

The standard approach is to choose our family of distributions $q$ to be fully factored distributions over the variables used in the block Gibbs scheme:

$$q(v, \theta, z) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\theta_t) \prod_{n=1}^{N} q_{\phi_n}(z_n)$$

where $q_{\gamma_t}(v_t)$ are Beta distributions, $q_{\tau_t}(\theta_t)$ are exponential family distributions with natural parameters $\tau_t$, and $q_{\phi_n}(z_n)$ are multinomial distributions.
Inference for Dirichlet process mixture models

\[ q(v, \theta, z) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\theta_t) \prod_{n=1}^{N} q_{\phi_n}(z_n) \]

- The algebra is surprisingly manageable!
- The updates are essentially deterministic versions of the stochastic updates of the block Gibbs sampler.
Inference for Dirichlet process mixture models

Let’s review a general result about the mean field updates. We will use more general notation for this section.

Suppose

\[ p(w_i|w_{-i}, x, \theta) \propto h_i(w_i) \exp \{ \langle g_i(w_{-i}, x, \theta) \rangle \tau_i(w_i) \} \]

Further suppose that we parameterize our family in terms of natural parameters:

\[ q_{\nu}(w) \propto \prod_{i=1}^{M} h_i(w_i) \exp \{ \langle \nu_i, \tau_i(w_i) \rangle \} \]

Then the mean field algorithm involves iteratively updating each \( \nu_i \) in sequence:

\[ \nu_i := \mathbb{E}_q [g_i(W_{-i}, x, \theta)] \]
Inference for Dirichlet process mixture models

Let’s compare the Gibbs algorithm to the mean field algorithm.

▶ In Gibbs, we iteratively sample each latent variable $w_i$ in sequence:

$$p(w_i|w_{-i}, x, \theta) \propto h_i(w_i) \exp \left\{ \langle g_i(W_{-i}, x, \theta), \tau_i(w_i) \rangle \right\}$$

▶ In mean field, we iteratively update each natural parameter $\nu_i$ in sequence:

$$\nu_i := \mathbb{E}_q [g_i(W_{-i}, x, \theta)]$$
Outline

- Dirichlet processes (DPs)
- Hierarchical Dirichlet processes (HDPs)
- Pitman-Yor processes (PYs and HPYs)
Hierarchical Dirichlet processes (HDPs)

- The model in terms of the DP
- Stick-breaking construction
- Chinese restaurant franchise (CRF) view of the HDP
- Application: Information retrieval
- Application: Topic modeling (HDP-LDA)
Hierarchical Dirichlet processes (HDPs)

Consider an indexed collection of DPs, \( \{ G_j \} \) defined on a common probability space. The hierarchical Dirichlet process ties these random measures probabilistically by letting them share their base measure and letting this base measure be random:

\[
\begin{align*}
G_0 | \gamma, H & \sim \text{DP}(\gamma, H) \\
G_j | \alpha, G_0 & \sim \text{DP}(\alpha, G_0), \quad \forall j \in J
\end{align*}
\]
Hierarchical Dirichlet processes (HDPs)

\[ G_0 \mid \gamma, H \sim \text{DP}(\gamma, H) \]
\[ G_j \mid \alpha, G_0 \sim \text{DP}(\alpha, G_0), \quad \forall j \in J \]

What might be the problem with simply drawing each \( G_j \) directly from \( H \)? In other words, why is the hierarchy important?
Hierarchical Dirichlet processes (HDPs)

\[ G_0 | \gamma, H \sim \text{DP}(\gamma, H) \]
\[ G_j | \alpha, G_0 \sim \text{DP}(\alpha, G_0), \quad \forall j \in J \]

The individual Dirichlet processes would not have any atoms in common, and therefore there would be no sharing of clusters among them. We will explore this in more detail.
Stick-breaking construction

Recall that a Dirichlet process can be written as a weighted sum of atoms drawn from its base distribution. Thus in the HDP, each of the child Dirichlet processes can be written as a (differently) weighted sum of the atoms of the shared parent:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}, \quad \forall j \in J$$

If the shared parent was not discrete, the child Dirichlet processes would not contain the same atoms.
Chinese restaurant franchise

The Chinese restaurant process describes the marginal probabilities of the DP in terms of a random partition obtained from sequence of customers sitting at tables in a restaurant. There is an analogous representation for the HDP which we refer to as the Chinese restaurant franchise.
In a CRF, the metaphor of a Chinese restaurant is extended to a set of restaurants, one for each index in $\mathcal{J}$. The customers in the $j$th restaurant sit at tables in the same manner as for the CRP, and this is done independently in the restaurants. The coupling among restaurants is achieved via a franchise-wide menu.

The first customer to sit at a table in a restaurant chooses a dish from the menu and all subsequent customers who sit at that table inherit that dish. Dishes are chosen with probability proportional to the number of tables (franchise-wide) which have previously served that dish.
Chinese restaurant franchise

Notation:

- $\theta_{ji} \sim G_j$: the dish eaten by the $i$th customer in the $j$th restaurant.
- $\theta^*_{jk} \sim G_0$: the dish served at the $k$th table in the $j$th restaurant.
- $\theta^{**}_k \sim H$: the $k$th (unique) dish sampled from $H$.
- $t_{ji}$: the table at which the $i$th customer in the $j$th restaurant sits at.
- $n_{jtk}$: the number of customers in restaurant $j$ seated at table $t$ and eating dish $k$.
- $m_{jk}$: the number of tables in restaurant $j$ serving dish $k$.
- $K$: the number of unique dishes served in the entire franchise.
Chinese restaurant franchise

Quiz:

- $\theta_{ji} \sim G_j$
- $\theta_{jk}^* \sim G_0$
- $\theta_{k}^{**} \sim H$
- $t_{ji}$
- $n_{jtk}$
- $m_{jk}$
- $K$
Using our knowledge of Dirichlet processes, we can integrate out the random measures $G_j$ and $G_0$ in turn from the HDP. We start by integrating out the child measures. This yields a set of conditional distributions for the $\theta_{ji}$ described by the urn scheme:

$$
\theta_{ji}|\theta_{j1}, \ldots, \theta_{j,i-1}, \alpha, G_0 \sim \frac{1}{\alpha + nj.} \left( \alpha G_0 + \sum_{t=1}^{m_j.} n_{jt.}\delta_{\theta_{jt}^*} \right)
$$
Chinese restaurant franchise

\[
\theta_{ji} | \theta_{j1}, \ldots, \theta_{j,i-1}, \alpha, G_0 \sim \frac{1}{\alpha + n_j.} \left( \alpha G_0 + \sum_{t=1}^{m_j.} n_{jt.} \delta_{\theta^*_{jt}} \right)
\]

We can integrate out \( G_0 \) as well, obtaining an urn scheme for the \( \theta^*_j \):

\[
\theta^*_{jt} | \theta^*_{11}, \ldots, \theta^*_1, m_1., \ldots, \theta^*_j, t-1, \gamma, H \sim \ldots
\]

What should this distribution be?
Chinese restaurant franchise

\[
\theta_{ji}|\theta_{j1}, \ldots, \theta_{j,i-1}, \alpha, G_0 \sim \frac{1}{\alpha + n_{j.}} \left( \alpha G_0 + \sum_{t=1}^{m_j} n_{jt} \delta_{\theta_{jt}^*} \right)
\]

We can integrate out \( G_0 \) as well, obtaining an urn scheme for the \( \theta_{jt}^* \):

\[
\theta_{jt}^*|\theta_{11}^*, \ldots, \theta_{1,m_1}^*, \ldots, \theta_{j,t-1}^*, \gamma, H \sim \frac{1}{\gamma + m_.} \left( \gamma H + \sum_{k=1}^{K} m_. k \delta_{\theta_{k}^{**}} \right)
\]
The CRF is useful in understanding the scaling properties of the clustering induced by the HDP.

If we assume that there are $J$ groups and that the groups (i.e. the number of customers in the different restaurants) have roughly the same size $N$, then the total number of clusters satisfies:

$$K = O \left( \gamma \log \frac{\alpha}{\gamma} + \gamma \log J + \gamma \log \log \frac{N}{\alpha} \right)$$
Chinese restaurant franchise

- If we assume that there are $J$ groups and that the groups (i.e. the number of customers in the different restaurants) have roughly the same size $N$, then the total number of clusters satisfies:

  $$K = O \left( \gamma \log \frac{\alpha}{\gamma} + \gamma \log J + \gamma \log \log \frac{N}{\alpha} \right)$$

- The number of clusters scales linearly in $\gamma$, logarithmically in $\alpha$ and $J$, and only doubly logarithmically in the size of each group.

- The HDP expresses a prior belief grows extremely slowly in $N$.

- If this prior belief is inappropriate for a given problem, there are alternatives such as the hierarchical Pitman-Yor process (HPY) which we will cover shortly.
Consider the following nonparametric generalization of the LDA model:

\[ G_0|\gamma, H \sim \text{DP}(\gamma, H) \]  \hspace{1cm} (1)

\[ G_j|\alpha, G_0 \sim \text{DP}(\alpha, G_0) \quad \text{for each document } j \in J \]  \hspace{1cm} (2)

\[ \theta_{ji}|G_j \sim G_j \quad \text{for each word } i = 1, \ldots, n_j \]  \hspace{1cm} (3)

\[ x_{ji}|\theta_{ji} \sim F_{\theta_{ji}} \]  \hspace{1cm} (4)

where \( H \) is the prior distribution over topics, \( G_j \) is the distribution over topics representing the \( j \)th document. \( \theta_{ji} \) is the topic of the \( i \)th word in the \( j \)th document, and \( x_{ji} \) is the \( i \)th word in the \( j \)th document.
Application: HDP-LDA

Perplexity on test abstracts of LDA and HDP mixture

- LDA
- HDP Mixture

Perplexity vs. Number of LDA topics
Application: HDP-HMM

Figure 3: HDP hidden Markov model.
Outline

- Dirichlet processes (DPs)
- Hierarchical Dirichlet processes (HDPs)
- Pitman-Yor processes (PYs and HPYs)
As discussed earlier, the HDP expresses a prior belief that the number of clusters grows extremely slowly with the number of customers per restaurant. For some problems, this prior belief may be inappropriate. There are alternative models such as the Pitman-Yor process, that may be more appropriate in these cases.
The Pitman-Yor process is a two-parameter generalization of the DP, with a discount parameter \(0 \leq d < 1\) and \(\alpha > -d\). When \(d = 0\), the Pitman-Yor process reduces to a DP with concentration parameter \(\alpha\).
Pitman-Yor processes

\[ G \sim PY(d, \alpha, H) \]

Suppose \( H \) is a smooth distribution and let \( \theta_1, \theta_2, \ldots \) be i.i.d draws from \( G \). Marginalizing out \( G \), the distribution of \( \theta_i \) conditioned on \( \theta_1, \ldots, \theta_{i-1} \) follows a generalization of the urn scheme:

\[ \theta_i|\theta_1, \ldots, \theta_{i-1}, d, \alpha, H \sim \frac{1}{\alpha + i - 1} \left( (\alpha + Kd)H + \sum_{t=1}^{K} (n_t - d)\delta_{\theta_t^*} \right) \]
Pitman-Yor processes

\[ \theta_i | \theta_1, \ldots, \theta_{i-1}, d, \alpha, H \sim \frac{1}{\alpha + i - 1} \left( (\alpha + Kd)H + \sum_{t=1}^{K} (n_t - d)\delta_{\theta_t^*} \right) \]

There are two salient properties.

- The rich-gets-richer property of the original Chinese restaurant process is preserved, which means that there are a small number of large tables.
- There are also a large number of small tables, since the probabilities of occupying new tables grows along with the number of occupied tables, and the discount \( d \) decreases the probabilities of new customers sitting at small tables.
Hierarchical Pitman-Yor processes

The Hierarchical Pitman-Yor (HPY) process is defined in the obvious manner:

\[ G_0 | \eta, \gamma, H \sim PY(\eta, \gamma, H) \quad (5) \]
\[ G_j | d, \alpha, G_0 \sim PY(d, \alpha, G_0) \quad \text{for } j \in J \quad (6) \]

This has many applications!
Acknowledgements

All content and even some entire paragraphs were taken from the following papers:

- **D.M. Blei and M.I. Jordan.**  
  Variational inference for dirichlet process mixtures.  

- **Y.W. Teh.**  
  Dirichlet process.  

- **Y.W. Teh and M.I. Jordan.**  
  Hierarchical bayesian nonparametric models with applications.  
  *Bayesian Nonparametrics, 2008.*

- **Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei.**  
  Hierarchical dirichlet processes.  