An Introduction to Bayesian Nonparametric Modelling

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

October, 2009 / Toronto
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Regression with Basis Functions

- Supervised learning of a function $f^* : X \rightarrow Y$ from training data $\{x_i, y_i\}_{i=1}^n$. 

![Graph showing supervised learning with basis functions]
Regression with Basis Functions

- Assume a set of basis functions $\phi_1, \ldots, \phi_K$ and parametrize a function:

$$f(x; \mathbf{w}) = \sum_{k=1}^{K} w_k \phi_k(x)$$

Parameters $\mathbf{w} = \{w_1, \ldots, w_K\}$.

- Find optimal parameters

$$\text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left| y_i - f(x_i; \mathbf{w}) \right|^2 = \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left| y_i - \sum_{k=1}^{K} w_k \phi_k(x_i) \right|^2$$

- We will be Bayesian in this lecture, so we need to rephrase using probabilistic model with priors on parameters:

$$y_i | x_i, \mathbf{w} = f(x_i; \mathbf{w}) + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$w_k \sim \mathcal{N}(0, \tau^2)$$

- Computer posterior $p(\mathbf{w} | \{x_i, y_i\})$. 
Regression with Basis Functions

\[ f(x; w) = \sum_{k=1}^{K} w_k \phi_k(x) \]

- What basis functions to use?
- How many basis functions to use?
- Do we really believe that the true \( f^*(x) \) can be expressed as \( f^*(x) = f(x; w^*) \) for some \( w^* \)?

\[ \epsilon_i \sim \mathcal{N}(0, \sigma^2) \]

- Do we believe the noise process is Gaussian?
Density Estimation with Mixture Models

- Unsupervised learning of a density $f^*(x)$ from training samples $\{x_i\}$.

- Perhaps use an exponential family distribution, e.g. Gaussian?
  \[
  \mathcal{N}(x; \mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)
  \]
  Unimodal, restrictive shape, light tail...

- Use a mixture model instead,
  \[
  f(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)
  \]

- Do we believe that the true density is a mixture of $K$ components?
- How many mixture components to use?
Latent Variable Modelling

- Say we have $n$ vector observations $x_1, \ldots, x_n$.
- Model each observation as a linear combination of $K$ latent sources:
  \[ x_i = \sum_{k=1}^{K} \Lambda_k y_{ik} + \epsilon_i \]

  - $y_{ik}$: activity of source $k$ in datum $i$.
  - $\Lambda_k$: basis vector describing effect of source $k$.
- Examples include principle components analysis, factor analysis, independent components analysis.
- How many sources are there?
- Do we believe that $K$ sources is sufficient to explain all our data?
- What prior distribution should we use for sources?
Infer topics from a document corpus, topics being sets of words that tend to co-occur together.

Using (Bayesian) latent Dirichlet allocation:

\[
\pi_j \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \\
\theta_k \sim \text{Dirichlet}(\frac{\beta}{W}, \ldots, \frac{\beta}{W}) \\
z_{ji} | \pi_j \sim \text{Multinomial}(\pi_j) \\
x_{ji} | z_{ji}, \theta_{zi} \sim \text{Multinomial}(\theta_{zi})
\]

How many topics can we find from the corpus?
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Modelling Data

- Models are almost never correct for real world data.
- How do we deal with model misfit?
  - Quantify closeness to true model, and optimality of fitted model;
  - Model selection or averaging;
  - Increase the flexibility of your model class.
- Bayesian nonparametrics are good solutions from the second and third perspectives.
Model Selection and Model Averaging

- Data $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Model $M_k$ parametrized by $\theta_k$, for $k = 1, 2, \ldots$.
- Marginal likelihood:

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k, M_k)\,d\theta_k$$

- Model selection and averaging:

$$M = \arg\max_{M_k} p(\mathbf{x}|M_k) \quad \text{or} \quad p(k, \theta_k|\mathbf{x}) = \frac{p(k)p(\theta_k|M_k)p(\mathbf{x}|\theta_k, M_k)}{\sum_{k'} p(k')p(\theta_{k'}|M_{k'})p(\mathbf{x}|\theta_{k'}, M_{k'})}$$

- Model selection and averaging is to prevent overfitting and underfitting, and are usually expensive to compute.
- But reasonable and proper Bayesian methods should not overfit anyway [Rasmussen and Ghahramani 2001].
Nonparametric Modelling

- What is a nonparametric model?
  - A really large parametric model;
  - A parametric model where the number of parameters increases with data;
  - A model over infinite dimensional function or measure spaces.
  - A family of distributions that is dense in some large space.

- Why nonparametric models in Bayesian theory of learning?
  - broad class of priors that allows data to “speak for itself”;
  - side-step model selection and averaging.

- How do we deal with the very large parameter spaces?
  - Marginalize out all but a finite number of parameters;
  - Define infinite space implicitly (akin to the kernel trick) using either Kolmogorov Consistency Theorem or de Finetti’s Theorem.
Gaussian Processes

- A Gaussian process (GP) is a random function $f : \mathbb{X} \rightarrow \mathbb{R}$ such that for any finite set of input points $x_1, \ldots, x_n$,

$$
\begin{bmatrix}
  f(x_1) \\
  \vdots \\
  f(x_n)
\end{bmatrix}
\sim \mathcal{N}
\begin{pmatrix}
  \begin{bmatrix}
    m(x_1) \\
    \vdots \\
    m(x_n)
  \end{bmatrix},
  \begin{bmatrix}
    c(x_1, x_1) & \cdots & c(x_1, x_n) \\
    \vdots & \ddots & \vdots \\
    c(x_n, x_1) & \cdots & c(x_n, x_n)
  \end{bmatrix}
\end{pmatrix}
$$

where the parameters are the mean function $m(x)$ and covariance kernel $c(x, y)$.

- Note: a random function $f$ is a stochastic process. It is a collection of random variables $\{f(x)\}_{x \in \mathbb{X}}$ one for each possible input value $x$.

- Can also be expressed as

$$
f(x) = \sum_{k=1}^{K} w_k \phi_k(x) \quad \text{as } K \rightarrow \infty.
$$

[Rasmussen and Williams 2006]
Posterior and Predictive Distributions

How do we compute the posterior and predictive distributions?

Training set \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) and test input \(x_{n+1}\).

Out of the (uncountably infinitely) many random variables \(\{f(x)\}_{x \in \mathbb{X}}\) making up the GP only \(n + 1\) has to do with the data: 

\[ f(x_1), f(x_2), \ldots, f(x_{n+1}) \]

Training data gives observations \(f(x_1) = y_1, \ldots, f(x_n) = y_n\). The predictive distribution of \(f(x_{n+1})\) is simply

\[ p(f(x_{n+1})|f(x_1) = y_1, \ldots, f(x_n) = y_n) \]

which is easy to compute since \(f(x_1), \ldots, f(x_{n+1})\) is Gaussian.

This can be generalized to noisy observations \(y_i = f(x_i) + \epsilon_i\) or non-linear effects \(y_i \sim D(f(x_i))\) where \(D(\theta)\) is a distribution parametrized by \(\theta\).
Consistency and Existence

- The definition of Gaussian processes only give finite dimensional marginal distributions of the stochastic process.

- Fortunately these marginal distributions are consistent. For every finite set $x \subset X$ we have a distinct distribution $p_x([f(x)]_{x \in x})$. These distributions are said to be consistent if

$$p_x([f(x)]_{x \in x}) = \int p_{x \cup y}([f(x)]_{x \in x \cup y}) d[f(x)]_{x \in y}$$

for disjoint and finite $x, y \subset X$.

- The marginal distributions for the GP are consistent because Gaussians are closed under marginalization.

- The Kolmogorov Consistency Theorem guarantees existence of GPs, i.e. the whole stochastic process $\{f(x)\}_{x \in X}$.

- Further information in Peter Orbanz’ Bayesian nonparametric tutorial.
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

**Infinite Mixture Models**

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Bayesian Mixture Models

Let’s be Bayesian about mixture models, and place priors over our parameters (and to compute posteriors).

First, introduce variable $z_i$ indicator which component $x_i$ belongs to.

$$z_i | \pi \sim \text{Multinomial}(\pi)$$
$$x_i | z_i = k, \mu, \Sigma \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Second, introduce conjugate priors for parameters:

$$\pi \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$
$$\mu_k, \Sigma_k = \theta^*_k \sim H = \mathcal{N} - \mathcal{IW}(0, s, d, \Phi)$$

[Rasmussen 2000]
Gibbs Sampling for Bayesian Mixture Models

- All conditional distributions are simple to compute:

\[
p(z_i = k | \text{others}) \propto \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)
\]

\[
\pi | \mathbf{z} \sim \text{Dirichlet}(\frac{\alpha}{K} + n_1(\mathbf{z}), \ldots, \frac{\alpha}{K} + n_K(\mathbf{z}))
\]

\[
\mu_k, \Sigma_k | \text{others} \sim \mathcal{N}-\mathcal{IW}(\nu', s', d', \Phi')
\]

- Not as efficient as collapsed Gibbs sampling which integrates out \(\pi, \mu, \Sigma\):

\[
p(z_i = k | \text{others}) \propto \frac{\alpha}{K} + n_k(\mathbf{z}_{-i}) \times \frac{1}{\alpha + n - 1} \times p(x_i | \{x_{i'} : i' \neq i, z_{i'} = k\})
\]

- Demo: fm_demointeractive.
Infinite Bayesian Mixture Models

- We will take $K \to \infty$.
- Imagine a very large value of $K$.
- There are at most $n < K$ occupied components, so most components are empty. We can lump these empty components together:

  Occupied clusters:

  $$p(z_i = k|\text{others}) \propto \frac{\alpha K + n_k(z_i)}{n - 1 + \alpha} p(x_i|x_i^k)$$

  Empty clusters:

  $$p(z_i = k_{\text{empty}}|z_i^--) \propto \frac{\alpha K - K^*}{n - 1 + \alpha} p(x_i|\emptyset)$$

- Demo: dpm_demointeractive.
Infinite Bayesian Mixture Models

- We will take $K \to \infty$.
- Imagine a very large value of $K$.
- There are at most $n < K$ occupied components, so most components are empty. We can lump these empty components together:

  Occupied clusters:
  \[
  p(z_i = k|\text{others}) \propto \frac{n_k(z_{-i})}{n - 1 + \alpha} p(x_i|x_k^{-i})
  \]

  Empty clusters:
  \[
  p(z_i = k_{\text{empty}}|z_{-i}) \propto \frac{\alpha}{n - 1 + \alpha} p(x_i|\emptyset)
  \]

- Demo: dpm_demointeractive.
Infinite Bayesian Mixture Models

- The actual infinite limit of finite mixture models does not make sense: any particular component will get a mixing proportion of 0.
- In the Gibbs sampler we bypassed this by lumping empty clusters together.
- Other better ways of making this infinite limit precise:
  - Look at the prior clustering structure induced by the Dirichlet prior over mixing proportions—**Chinese restaurant process**.
  - Re-order components so that those with larger mixing proportions tend to occur first, before taking the infinite limit—**stick-breaking construction**.
- Both are different views of the **Dirichlet process** (DP).
- DPs can be thought of as infinite dimensional Dirichlet distributions.
- The $K \to \infty$ Gibbs sampler is for DP mixture models.
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
A Tiny Bit of Measure Theoretic Probability Theory

- A \(\sigma\)-algebra \(\Sigma\) is a family of subsets of a set \(\Theta\) such that
  - \(\Sigma\) is not empty;
  - If \(A \in \Sigma\) then \(\Theta \setminus A \in \Sigma\);
  - If \(A_1, A_2, \ldots \in \Sigma\) then \(\bigcup_{i=1}^{\infty} A_i \in \Sigma\).

- \((\Theta, \Sigma)\) is a measure space and \(A \in \Sigma\) are the measurable sets.

- A measure \(\mu\) over \((\Theta, \Sigma)\) is a function \(\mu : \Sigma \rightarrow [0, \infty]\) such that
  - \(\mu(\emptyset) = 0\);
  - If \(A_1, A_2, \ldots \in \Sigma\) are disjoint then \(\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)\).
  - Everything we consider here will be measurable.
  - A probability measure is one where \(\mu(\Theta) = 1\).

- Given two measure spaces \((\Theta, \Sigma)\) and \((\Delta, \Phi)\), a function \(f : \Theta \rightarrow \Delta\) is measurable if \(f^{-1}(A) \in \Sigma\) for every \(A \in \Phi\).
A Tiny Bit of Measure Theoretic Probability Theory

- If $p$ is a probability measure on $(\Theta, \Sigma)$, a *random variable* $X$ taking values in $\Delta$ is simply a measurable function $X : \Theta \rightarrow \Delta$.
  - Think of the probability space $(\Theta, \Sigma, p)$ as a black-box random number generator, and $X$ as a function taking random samples in $\Theta$ and producing random samples in $\Delta$.
  - The probability of an event $A \in \Phi$ is $p(X \in A) = p(X^{-1}(A))$.

- A *stochastic process* is simply a collection of random variables $\{X_i\}_{i \in I}$ over the same measure space $(\Theta, \Sigma)$, where $I$ is an index set.
  - What distinguishes a stochastic process from, say, a graphical model is that $I$ can be infinite, even uncountably so.
  - This raises issues of how do you even define them and how do you ensure that they can even exist (mathematically speaking).

- Stochastic processes form the core of many Bayesian nonparametric models.
  - Gaussian processes, Poisson processes, gamma processes, Dirichlet processes, beta processes...
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Dirichlet Distributions

- A Dirichlet distribution is a distribution over the $K$-dimensional probability simplex:

$$\Delta_K = \{ (\pi_1, \ldots, \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1 \}$$

- We say $(\pi_1, \ldots, \pi_K)$ is Dirichlet distributed,

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\lambda_1, \ldots, \lambda_K)$$

with parameters $(\lambda_1, \ldots, \lambda_K)$, if

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_k \lambda_k)}{\prod_k \Gamma(\lambda_k)} \prod_{k=1}^n \pi_k^{\lambda_k-1}$$

- Equivalent to normalizing a set of independent gamma variables:

$$(\pi_1, \ldots, \pi_K) = \frac{1}{\sum_k \gamma_k} (\gamma_1, \ldots, \gamma_K)$$

$$\gamma_k \sim \text{Gamma}(\lambda_k) \quad \text{for } k = 1, \ldots, K$$
Dirichlet Distributions

Dir(1,0,1,1,0)  Dir(2,0,2,2,0)  Dir(5,5,5,5,0)

Dir(5,0,5,0,2,0)  Dir(9,0,2,0,2,0)  Dir(0,7,7,7,7)
A Dirichlet Process (DP) is a random probability measure $G$ over $(\Theta, \Sigma)$ such that for any finite set of measurable partitions $A_1 \cup \ldots \cup A_K = \Theta$,

$\left( G(A_1), \ldots, G(A_K) \right) \sim \text{Dirichlet}(\lambda(A_1), \ldots, \lambda(A_K))$

where $\lambda$ is a base measure.

The above family of distributions is consistent (next slide), and Kolmogorov Consistency Theorem can be applied to show existence (but there are technical conditions restricting the generality of the definition).

[Ferguson 1973, Blackwell and MacQueen 1973]
If we have two partitions \((A_1, \ldots, A_K)\) and \((B_1, \ldots, B_J)\) of \(\Theta\), how do we see if the two Dirichlets are consistent?

Because Dirichlet variables are normalized gamma variables and sums of gammas are gammas, if \((I_1, \ldots, I_j)\) is a partition of \((1, \ldots, K)\),

\[
\left( \sum_{i \in I_1} \pi_i, \ldots, \sum_{i \in I_j} \pi_i \right) \sim \text{Dirichlet} \left( \sum_{i \in I_1} \lambda_i, \ldots, \sum_{i \in I_j} \lambda_i \right)
\]
Consistency of Dirichlet Marginals

Form the common refinement \((C_1, \ldots, C_L)\) where each \(C_\ell\) is the intersection of some \(A_k\) with some \(B_j\). Then:

By definition, \((G(C_1), \ldots, G(C_L)) \sim \text{Dirichlet}(\lambda(C_1), \ldots, \lambda(C_L))\)

\[(G(A_1), \ldots, G(A_K)) = \left(\sum_{C_\ell \subset A_1} G(C_\ell), \ldots, \sum_{C_\ell \subset A_K} G(C_\ell)\right)\]

\[\sim \text{Dirichlet}(\lambda(A_1), \ldots, \lambda(A_K))\]

Similarly, \((G(B_1), \ldots, G(B_J)) \sim \text{Dirichlet}(\lambda(B_1), \ldots, \lambda(B_J))\)

so the distributions of \((G(A_1), \ldots, G(A_K))\) and \((G(B_1), \ldots, G(B_J))\) are consistent.

Demonstration: DPgenerate.
Parameters of Dirichlet Processes

- Usually we split the \( \lambda \) base measure into two parameters \( \lambda = \alpha H \):
  - *Base distribution* \( H \), which is like the *mean* of the DP.
  - *Strength parameter* \( \alpha \), which is like an *inverse-variance* of the DP.

- We write:

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

if for any partition \((A_1, \ldots, A_K)\) of \( \Theta \):

\[
(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))
\]

- The first and second moments of the DP:

  **Expectation:**
  \[
  \mathbb{E}[G(A)] = H(A)
  \]

  **Variance:**
  \[
  \text{Var}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}
  \]

where \( A \) is any measurable subset of \( \Theta \).
Representations of Dirichlet Processes

- Draws from Dirichlet processes will always place all their mass on a countable set of points:

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_k} \]

where \( \sum_k \pi_k = 1 \) and \( \theta^*_k \in \Theta \).

- What is the joint distribution over \( \pi_1, \pi_2, \ldots \) and \( \theta^*_1, \theta^*_2, \ldots \)?

- Since \( G \) is a (random) probability measure over \( \Theta \), we can treat it as a distribution and draw samples from it. Let

\[ \theta_1, \theta_2, \ldots \sim G \]

be random variables with distribution \( G \).

- What is the marginal distribution of \( \theta_1, \theta_2, \ldots \) with \( G \) integrated out?

- There is positive probability that sets of \( \theta_i \)'s can take on the same value \( \theta^*_k \) for some \( k \), i.e. the \( \theta_i \)'s cluster together. How do these clusters look like?

- For practical modelling purposes this is sufficient. But is this sufficient to tell us all about \( G \)?
Stick-breaking Construction

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

There is a simple construction giving the joint distribution of \( \pi_1, \pi_2, \ldots \) and \( \theta_1^*, \theta_2^*, \ldots \) called the stick-breaking construction.

\[ \theta_k^* \sim H \]
\[ \pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i) \]
\[ v_k \sim \text{Beta}(1, \alpha) \]

Also known as the GEM distribution, write \( \pi \sim \text{GEM}(\alpha) \).

[Sethuraman 1994]
Posterior of Dirichlet Processes

- Since $G$ is a probability measure, we can draw samples from it,

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

$$\theta_1, \ldots, \theta_n | G \sim G$$

What is the posterior of $G$ given observations of $\theta_1, \ldots, \theta_n$?

- The usual Dirichlet-multinomial conjugacy carries over to the nonparametric DP as well:

$$G | \theta_1, \ldots, \theta_n \sim \text{DP}(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n}\delta_{\theta_i}}{\alpha + n})$$
Pólya Urn Scheme

\[ \theta_1, \theta_2, \ldots \sim G \]

- The marginal distribution of \( \theta_1, \theta_2, \ldots \) has a simple generative process called the *Pólya urn scheme*.

\[ \theta_n|\theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1} \]

- Picking balls of different colors from an urn:
  - Start with no balls in the urn.
  - With probability \( \propto \alpha \), draw \( \theta_n \sim H \), and add a ball of color \( \theta_n \) into urn.
  - With probability \( \propto n - 1 \), pick a ball at random from the urn, record \( \theta_n \) to be its color and return two balls of color \( \theta_n \) into urn.

- Pólya urn scheme is like a “representer” for the DP—a finite projection of an infinite object \( G \).

- Also known as the *Blackwell-MacQueen urn scheme*.

[Blackwell and MacQueen 1973]
Chinese Restaurant Process

- $\theta_1, \ldots, \theta_n$ take on $K < n$ distinct values, say $\theta^*_1, \ldots, \theta^*_K$.
- This defines a partition of $(1, \ldots, n)$ into $K$ clusters, such that if $i$ is in cluster $k$, then $\theta_i = \theta^*_k$.
- The distribution over partitions is a *Chinese restaurant process* (CRP).
- Generating from the CRP:
  - First customer sits at the first table.
  - Customer $n$ sits at:
    - Table $k$ with probability $\frac{n_k}{\alpha + n - 1}$ where $n_k$ is the number of customers at table $k$.
    - A new table $K + 1$ with probability $\frac{\alpha}{\alpha + n - 1}$.
  - Customers $\Leftrightarrow$ integers, tables $\Leftrightarrow$ clusters.
The CRP exhibits the *clustering property* of the DP.

- *Rich-gets-richer* effect implies small number of large clusters.
- Expected number of clusters is $K = O(\alpha \log n)$. 
Exchangeability

- Instead of deriving the Pólya urn scheme by marginalizing out a DP, consider starting directly from the conditional distributions:

\[
\theta_n|\theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}
\]

- For any \(n\), the joint distribution of \(\theta_1, \ldots, \theta_n\) is:

\[
p(\theta_1, \ldots, \theta_n) = \frac{\alpha^K \prod_{k=1}^K h(\theta^*_k)(m_{nk} - 1)!}{\prod_{i=1}^n i - 1 + \alpha}
\]

where \(h(\theta)\) is density of \(\theta\) under \(H\), \(\theta^*_1, \ldots, \theta^*_K\) are the unique values, and \(\theta^*_k\) occurred \(m_{nk}\) times among \(\theta_1, \ldots, \theta_n\).

- The joint distribution is exchangeable wrt permutations of \(\theta_1, \ldots, \theta_n\).

- *De Finetti’s Theorem* says that there must be a random probability measure \(G\) making \(\theta_1, \theta_2, \ldots\) iid. This is the DP.
De Finetti’s Theorem

Let $\theta_1, \theta_2, \ldots$ be an infinite sequence of random variables with joint distribution $p$. If for all $n \geq 1$, and all permutations $\sigma \in \Sigma_n$ on $n$ objects,

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

That is, the sequence is infinitely exchangeable. Then there exists a (unique) latent random parameter $G$ such that:

$$p(\theta_1, \ldots, \theta_n) = \int p(G) \prod_{i=1}^{n} p(\theta_i | G) dG$$

where $p$ is a joint distribution over $G$ and $\theta_i$’s.

- $\theta_i$’s are independent given $G$.
- Sufficient to define $G$ through the conditionals $p(\theta_n | \theta_1, \ldots, \theta_{n-1})$.
- $G$ can be infinite dimensional (indeed it is often a random measure).
- The set of infinitely exchangeable sequences is convex and it is an important theoretical topic to study the set of extremal points.
- Partial exchangeability: Markov, group, arrays,...
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Binary Latent Variable Models

Consider a latent variable model with binary sources/features,

\[ Z_{ik} = \begin{cases} 
1 & \text{with probability } \mu_k; \\
0 & \text{with probability } 1 - \mu_k.
\end{cases} \]

Example: Data items could be movies like “Terminator 2”, “Shrek” and “Lord of the Rings”, and features could be “science fiction”, “fantasy”, “action” and “Arnold Schwarzenegger”.

Place beta prior over the probabilities of features:

\[ \mu_k \sim \text{Beta}(\frac{\alpha}{K}, 1) \]

We will again take \( K \to \infty \).
Indian Buffet Processes

- The *Indian Buffet Process* (IBP) is akin to the Chinese restaurant process but describes each customer with a binary vector instead of cluster.

- Generating from an IBP:
  - Parameter $\alpha$.
  - First customer picks $\text{Poisson}(\alpha)$ dishes to eat.
  - Subsequent customer $i$ picks dish $k$ with probability $\frac{m_k}{i}$; and picks $\text{Poisson}(\frac{\alpha}{i})$ new dishes.
Indian Buffet Processes and Exchangeability

- The IBP is infinitely exchangeable. For this to make sense, we need to “forget” the ordering of the dishes.
  - “Name” each dish $k$ with a $\Lambda_k^*$ drawn iid from $H$.
  - Each customer now eats a set of dishes: $\Psi_i = \{\Lambda_k^* : z_{ik} = 1\}$.
  - The joint probability of $\Psi_1, \ldots, \Psi_n$ can be calculated:

$$p(\Psi_1, \ldots, \Psi_n) = \exp \left( -\alpha \sum_{i=1}^{n} \frac{1}{i} \right) \alpha^K \prod_{k=1}^{K} \frac{(m_k - 1)! (n - m_k)!}{n!} h(\Lambda_k^*)$$

$K$: total number of dishes tried by $n$ customers.
$\Lambda_k^*$: Name of $k$th dish tried.
$m_k$: number of customers who tried dish $\Lambda_k^*$.

- De Finetti’s Theorem again states that there is some random measure underlying the IBP.
- This random measure is the beta process.

[Griffiths and Ghahramani 2006, Thibaux and Jordan 2007]
A beta process $B \sim \text{BP}(c, \alpha H)$ is a random discrete measure with form:

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

where the points $P = \{(\theta_1^*, \mu_1), (\theta_2^*, \mu_2), \ldots\}$ are spikes in a 2D Poisson process with rate measure:

$$c\mu^{-1}(1 - \mu)^{c-1} d\mu \alpha H(d\theta)$$

The beta process with $c = 1$ is the de Finetti measure for the IBP. When $c \neq 1$ we have a two parameter generalization of the IBP.

This is an example of a completely random measure.

A beta process does not have Beta distributed marginals.

[Hjort 1990, Ghahramani et al. 2007]
When $c = 1$ it was shown that the following generates a draw of $B$:

$$v_k \sim \text{Beta}(1, \alpha) \quad \mu_k = (1 - v_k) \prod_{i=1}^{k-1} (1 - v_i) \quad \theta^*_k \sim H$$

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta^*_k}$$

The above is the complement of the stick-breaking construction for DPs!

[Teh et al. 2007]
Applications of Indian Buffet Processes

- The IBP can be used in concert with different likelihood models in a variety of applications.

\[ Z \sim \text{IBP}(\alpha) \quad X \sim F(Z, Y) \]

\[ Y \sim H \quad p(Z, Y|X) = \frac{p(Z, Y)p(X|Z, Y)}{p(X)} \]

- Latent factor models for distributed representation [Griffiths and Ghahramani 2005].
- Matrix factorization for collaborative filtering [Meeds et al 2007].
- Protein complex discovery [Chu et al 2006].
- Psychological choice behaviour [Görür and Rasmussen 2006].
- Independent Components Analysis [Knowles and Ghahramani 2007].
Infinite Independent Components Analysis

- Each image $X_i$ is a linear combination of sparse features:

$$X_i = \sum_k \Lambda_k^* y_{ik}$$

where $y_{ik}$ is activity of feature $k$ with sparse prior. One possibility is a mixture of a Gaussian and a point mass at 0:

$$y_{ik} = z_{ik} a_{ik} \quad a_{ik} \sim \mathcal{N}(0, 1) \quad Z \sim \text{IBP}(\alpha)$$

- An ICA model with infinite number of features.

[Knowles and Ghahramani 2007]
Outline

Some Examples of Parametric Models
Bayesian Nonparametric Modelling
Infinite Mixture Models
Some Measure Theory
Dirichlet Processes
Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Infer topics from a document corpus, topics being sets of words that tend to co-occur together.

Using (Bayesian) latent Dirichlet allocation:

\[ \pi_j \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \]
\[ \theta_k \sim \text{Dirichlet}(\frac{\beta}{W}, \ldots, \frac{\beta}{W}) \]
\[ z_{ji} | \pi_j \sim \text{Multinomial}(\pi_j) \]
\[ x_{ji} | z_{ji}, \theta_{z_{ji}} \sim \text{Multinomial}(\theta_{z_{ji}}) \]

Can we take \( K \to \infty \)?
Hierarchical Dirichlet Processes
Hierarchical Dirichlet Processes

- Use a DP mixture for each group.

- Unfortunately there is no sharing of clusters across different groups because $H$ is smooth.

- Solution: make the base distribution $H$ discrete.

- Put a DP prior on the common base distribution.

[Teh et al. 2006]
Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

  \[ G_0 \sim \text{DP}(\alpha_0, H) \]
  \[ G_1, G_2 | G_0 \sim \text{DP}(\alpha, G_0) \text{ iid} \]

- Extension to larger hierarchies is straightforward.
Hierarchical Dirichlet Processes

- Making $G_0$ discrete forces shared cluster between $G_1$ and $G_2$. 
Hierarchical Dirichlet Processes

- Document topic modelling:
  - Allows documents to be modelled with DP mixtures of topics, with topics shared across corpora.

- Infinite hidden Markov modelling:
  - Allows HMMs with an infinite number of states, with transitions from each allowable state to every other allowable state.

- Learning discrete structures from data:
  - Determining number of objects, nonterminals, states etc.
Infinite Hidden Markov Models

- Hidden Markov models with an infinite number of states.
- Hierarchical DPs used to share information among transition probability vectors prevents “run-away” states.

[Beal et al. 2002, Teh et al. 2006]
Hierarchical Modelling

- Better estimation of parameters.
- Multitask learning, learning to learn: generalizing across related tasks.
Hierarchical Modelling

- Better estimation of parameters.
- Multitask learning, learning to learn: generalizing across related tasks.
Outline

Some Examples of Parametric Models

Bayesian Nonparametric Modelling

Infinite Mixture Models

Some Measure Theory

Dirichlet Processes

Indian Buffet and Beta Processes

Hierarchical Dirichlet Processes

Pitman-Yor Processes

Summary
Pitman-Yor Processes

▶ Two-parameter generalization of the Chinese restaurant process:

\[
p(\text{customer } n \text{ sat at table } k|\text{past}) = \begin{cases} 
\frac{n_k-\beta}{n-1+\alpha} & \text{if occupied table} \\
\frac{\alpha+\beta K}{n-1+\alpha} & \text{if new table}
\end{cases}
\]

▶ Associating each cluster \( k \) with a unique draw \( \theta_k^* \sim H \), the corresponding Pólya urn scheme is also exchangeable.

▶ De Finetti’s Theorem states that there is a random measure underlying this two-parameter generalization.

▶ This is the Pitman-Yor process.

▶ The Pitman-Yor process also has a stick-breaking construction:

\[
\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i) \quad \beta_k \sim \text{Beta}(1 - \beta, \alpha + \beta k) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
\]

[Pitman and Yor 1997, Perman et al. 1992]
Pitman-Yor Processes

- Two salient features of the Pitman-Yor process:
  - With more occupied tables, the chance of even more tables becomes higher.
  - Tables with smaller occupancy numbers tend to have lower chance of getting new customers.

- The above means that Pitman-Yor processes produce Zipf’s Law type behaviour, with $K = O(\alpha n^\beta)$. 

![Graphs showing the relationship between number of customers and tables with 1 customer and proportion of tables with 1 customer.](image)
Pitman-Yor Processes

Draw from a Pitman-Yor process

- \( \alpha = 1, \ d = 0.5 \)
- \( \alpha = 30, \ d = 0 \)

Draw from a Dirichlet process

- \( \alpha = 1, \ d = 0.5 \)
- \( \alpha = 30, \ d = 0 \)
Hierarchical Pitman-Yor Language Models

- Pitman-Yor processes can be suitable models for many natural phenomena with power-law statistics.

- Language modelling with Markov assumption:

\[
p(Mary \text{ has a little lamb}) \\
\approx p(Mary)p(has|Mary)p(a|Mary \text{ has})p(little|has a)p(lamb|a \text{ little})
\]

- Parameterize with \( p(w_3|w_1, w_2) = G_{w_1,w_2}[w_3] \) and use a hierarchical Pitman-Yor process prior:

\[
G_{w_1,w_2}|G_{w_2} \sim PY(\alpha_2, \beta_2, G_{w_2}) \\
G_{w_2}|G_{\emptyset} \sim PY(\alpha_1, \beta_1, G_{\emptyset}) \\
G_{\emptyset}|U \sim PY(\alpha_0, \beta_0, U)
\]

- State-of-the-art results, connection to Kneser-Ney smoothing.

[Goldwater et al. 2006a, Teh 2006b, Wood et al. 2009]
Human segmentations of images also seem to follow power-law.

An unsupervised image segmentation model based on dependent hierarchical Pitman-Yor processes achieves state-of-the-art results.

[Sudderth and Jordan 2009]
Outline

Some Examples of Parametric Models
Bayesian Nonparametric Modelling
Infinite Mixture Models
Some Measure Theory
Dirichlet Processes
Indian Buffet and Beta Processes
Hierarchical Dirichlet Processes
Pitman-Yor Processes

Summary
Summary

- Motivation for Bayesian nonparametrics:
  - Allows practitioners to define and work with models with large support, sidesteps model selection.
  - New models with useful properties.
  - Large variety of applications.

- Various standard Bayesian nonparametric models:
  - Dirichlet processes
  - Hierarchical Dirichlet processes
  - Infinite hidden Markov models
  - Indian buffet and beta processes
  - Pitman-Yor processes

- Touched upon two important theoretical tools:
  - Consistency and Kolmogorov’s Consistency Theorem
  - Exchangeability and de Finetti’s Theorem

- Described a number of applications of Bayesian nonparametrics.

- Missing: Inference methods based on MCMC, variational etc, consistency and convergence.
Other Introductions to Bayesian Nonparametrics

- Volker Tresp, ICML nonparametric Bayes workshop 2006.
- Peter Orbanz, Foundations of Nonparametric Bayesian Methods, 2009.
- I have given a number myself (check webpage).
- I have an introduction to Dirichlet processes [Teh 2007], and another to hierarchical Bayesian nonparametric models [Teh and Jordan 2010].
Bayesian Nonparametric Software

- Hierarchical Bayesian Compiler (HBC). Hal Daume III. http://www.cs.utah.edu/hal/HBC/
- Others...
AISTATS 2010

May 13-15, 2010. Chia Laguna Resort, Sardinia, Italy

http://www.aistats.org

Deadline: November 6, 2009
Outline

Relating Different Representations of Dirichlet Processes

Representations of Hierarchical Dirichlet Processes

Extended Bibliography
Representations of Dirichlet Processes

- **Posterior Dirichlet process**: 
  \[ G \sim \text{DP}(\alpha, H) \]
  \[ \theta | G \sim G \]
  \[ \theta \sim H \]
  \[ G | \theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_{\theta}}{\alpha + 1} \right) \]

- **Pólya urn scheme**: 
  \[ \theta_n | \theta_1:n-1 \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1} \]

- **Chinese restaurant process**: 
  \[ p(\text{customer } n \text{ sat at table } k | \text{past}) = \begin{cases} 
  \frac{n_k}{n-1+\alpha} & \text{if occupied table} \\
  \frac{\alpha}{n-1+\alpha} & \text{if new table} 
\end{cases} \]

- **Stick-breaking construction**: 
  \[ \pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \]
  \[ \beta_k \sim \text{Beta}(1, \alpha) \]
  \[ \theta^*_k \sim H \]
  \[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_k} \]
Posterior Dirichlet Processes

- Suppose $G$ is DP distributed, and $\theta$ is $G$ distributed:

  \[ G \sim \text{DP}(\alpha, H) \]
  \[ \theta | G \sim G \]

- We are interested in:
  - The marginal distribution of $\theta$ with $G$ integrated out.
  - The posterior distribution of $G$ conditioning on $\theta$. 
Posterior Dirichlet Processes

Conjugacy between Dirichlet Distribution and Multinomial.

- Consider:

\[
(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)
\]

\[
z | (\pi_1, \ldots, \pi_K) \sim \text{Discrete}(\pi_1, \ldots, \pi_K)
\]

\(z\) is a multinomial variate, taking on value \(i \in \{1, \ldots, n\}\) with probability \(\pi_j\).

- Then:

\[
z \sim \text{Discrete}\left(\frac{\alpha_1}{\sum_i \alpha_i}, \ldots, \frac{\alpha_K}{\sum_i \alpha_i}\right)
\]

\[(\pi_1, \ldots, \pi_K) | z \sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \ldots, \alpha_K + \delta_K(z))\]

where \(\delta_i(z) = 1\) if \(z\) takes on value \(i\), 0 otherwise.

- Converse also true.
Posterior Dirichlet Processes

- Fix a partition \((A_1, \ldots, A_K)\) of \(\Theta\). Then

\[
(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))
\]

\[
P(\theta \in A_i | G) = G(A_i)
\]

- Using Dirichlet-multinomial conjugacy,

\[
P(\theta \in A_i) = H(A_i)
\]

\[
(G(A_1), \ldots, G(A_K)) | \theta \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_K) + \delta_\theta(A_K))
\]

- The above is true for every finite partition of \(\Theta\). In particular, taking a really fine partition,

\[
p(d\theta) = H(d\theta)
\]

i.e. \(\theta \sim H\) with \(G\) integrated out.

- Also, the posterior \(G|\theta\) is also a Dirichlet process:

\[
G|\theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right)
\]
Posterior Dirichlet Processes

\[
\begin{align*}
G &\sim \text{DP}(\alpha, H) \\
\theta | G &\sim G \\
\iff \\
\theta &\sim H \\
G | \theta &\sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)
\end{align*}
\]
Pólya Urn Scheme

- **First sample:**
  \[
  \begin{align*}
  \theta_1 | G & \sim G \\
  G & \sim \text{DP}(\alpha, H) \\
  \iff \theta_1 & \sim H \\
  G|\theta_1 & \sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha+1})
  \end{align*}
  \]

- **Second sample:**
  \[
  \begin{align*}
  \theta_2 | \theta_1, G & \sim G \\
  G|\theta_1 & \sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha+1}) \\
  \iff \theta_2 | \theta_1 & \sim \frac{\alpha H + \delta_{\theta_1}}{\alpha+1} \\
  G|\theta_1, \theta_2 & \sim \text{DP}(\alpha + 2, \frac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+2})
  \end{align*}
  \]

- **n\textsuperscript{th} sample**
  \[
  \begin{align*}
  \theta_n | \theta_1:n-1, G & \sim G \\
  G|\theta_1:n-1 & \sim \text{DP}(\alpha + n - 1, \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1}) \\
  \iff \theta_n | \theta_1:n-1 & \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} \\
  G|\theta_1:n & \sim \text{DP}(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n})
  \end{align*}
  \]
Stick-breaking Construction

- Returning to the posterior process:

\[ G \sim \text{DP}(\alpha, H) \quad \Leftrightarrow \quad \theta \sim H \]

\[ \theta | G \sim G \quad \Leftrightarrow \quad G | \theta \sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}) \]

- Consider a partition \((\theta, \Theta \setminus \theta)\) of \(\Theta\). We have:

\[
(G(\theta), G(\Theta \setminus \theta)) | \theta \sim \text{Dirichlet}((\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\theta), (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\Theta \setminus \theta))
= \text{Dirichlet}(1, \alpha)
\]

- \(G\) has a point mass located at \(\theta\):

\[ G = \beta \delta_\theta + (1 - \beta) G' \quad \text{with} \quad \beta \sim \text{Beta}(1, \alpha) \]

and \(G'\) is the (renormalized) probability measure with the point mass removed.

- What is \(G'\)?
Stick-breaking Construction

Currently, we have:

\[
\begin{align*}
G &\sim \text{DP}(\alpha, H) \\
\theta &\sim G \\
\Rightarrow & \\
G | \theta &\sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta \theta}{\alpha + 1}) \\
\beta &\sim \text{Beta}(1, \alpha)
\end{align*}
\]

Consider a further partition \((\theta, A_1, \ldots, A_K)\) of \(\Theta\):

\[
\begin{align*}
(G(\theta), G(A_1), \ldots, G(A_K)) &= (\beta, (1 - \beta)G'(A_1), \ldots, (1 - \beta)G'(A_K)) \\
&\sim \text{Dirichlet}(1, \alpha H(A_1), \ldots, \alpha H(A_K))
\end{align*}
\]

The agglomerative/decimative property of Dirichlet implies:

\[
\begin{align*}
(G'(A_1), \ldots, G'(A_K)) | \theta &\sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)) \\
G' &\sim \text{DP}(\alpha, H)
\end{align*}
\]
Stick-breaking Construction

- We have:

\[ G \sim \text{DP}(\alpha, H) \]
\[ G = \beta_1 \delta_{\theta^*_1} + (1 - \beta_1)G_1 \]
\[ G = \beta_1 \delta_{\theta^*_1} + (1 - \beta_1)(\beta_2 \delta_{\theta^*_2} + (1 - \beta_2)G_2) \]
\[ \vdots \]
\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_k} \]

where

\[ \pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \quad \beta_k \sim \text{Beta}(1, \alpha) \quad \theta^*_k \sim H \]
Outline

Relating Different Representations of Dirichlet Processes

Representations of Hierarchical Dirichlet Processes

Extended Bibliography
Stick-breaking Construction

- We shall assume the following HDP hierarchy:

\[
G_0 \sim \text{DP}(\gamma, H) \\
G_j | G_0 \sim \text{DP}(\alpha, G_0) \quad \text{for } j = 1, \ldots, J
\]

- The stick-breaking construction for the HDP is:

\[
G_0 = \sum_{k=1}^{\infty} \pi_{0k} \delta_{\theta^*_k} \quad \theta^*_k \sim H \\
\pi_{0k} = \beta_{0k} \prod_{l=1}^{k-1} (1 - \beta_{0l}) \quad \beta_{0k} \sim \text{Beta}(1, \gamma) \\
G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta^*_k} \\
\pi_{jk} = \beta_{jk} \prod_{l=1}^{k-1} (1 - \beta_{jl}) \quad \beta_{jk} \sim \text{Beta}(\alpha \beta_{0k}, \alpha(1 - \sum_{l=1}^{k} \beta_{0l}))
\]
Hierarchical Pòlya Urn Scheme

- Let $G \sim DP(\alpha, H)$.
- We can visualize the Pòlya urn scheme as follows:

\[
\begin{array}{ccccccc}
\theta_1^* & \theta_2^* & \theta_3^* & \theta_4^* & \theta_5^* & \theta_6^* & \cdots \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
\theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \theta_6 & \theta_7 & \cdots
\end{array}
\]

where the arrows denote to which $\theta_k^*$ each $\theta_i$ was assigned and

\[
\begin{align*}
\theta_1, \theta_2, \ldots & \sim G \text{ i.i.d.} \\
\theta_1^*, \theta_2^*, \ldots & \sim H \text{ i.i.d.}
\end{align*}
\]

(but $\theta_1, \theta_2, \ldots$ are not independent of $\theta_1^*, \theta_2^*, \ldots$).
Hierarchical Pòlya Urn Scheme

- Let \( G_0 \sim \text{DP}(\gamma, H) \) and \( G_1, G_2 | G_0 \sim \text{DP}(\alpha, G_0) \).
- The hierarchical Pòlya urn scheme to generate draws from \( G_1, G_2 \):
Let $G_0 \sim \text{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \text{DP}(\alpha, G_0)$.

The Chinese restaurant franchise describes the clustering of data items in the hierarchy:
Outline

Relating Different Representations of Dirichlet Processes

Representations of Hierarchical Dirichlet Processes

Extended Bibliography
Dirichlet Processes were first introduced by [Ferguson 1973], while [Antoniak 1974] further developed DPs as well as introduce the mixture of DPs. [Blackwell and MacQueen 1973] showed that the Pólya urn scheme is exchangeable with the DP being its de Finetti measure. Further information on the Chinese restaurant process can be obtained at [Aldous 1985, Pitman 2002]. The DP is also related to Ewens’ Sampling Formula [Ewens 1972]. [Sethuraman 1994] gave a constructive definition of the DP via a stick-breaking construction. DPs were rediscovered in the machine learning community by [Neal 1992, Rasmussen 2000].

Hierarchical Dirichlet Processes (HDPs) were first developed by [Teh et al. 2006], although an aspect of the model was first discussed in the context of infinite hidden Markov models [Beal et al. 2002]. HDPs and generalizations have been applied across a wide variety of fields.

Dependent Dirichlet Processes are sets of coupled distributions over probability measures, each of which is marginally DP [MacEachern et al. 2001]. A variety of dependent DPs have been proposed in the literature since then [Srebro and Roweis 2005, Griffin 2007, Caron et al. 2007]. The infinite mixture of Gaussian processes of [Rasmussen and Ghahramani 2002] can also be interpreted as a dependent DP.

Indian Buffet Processes (IBPs) were first proposed in [Griffiths and Ghahramani 2006], and extended to a two-parameter family in [Ghahramani et al. 2007]. [Thibaux and Jordan 2007] showed that the de Finetti measure for the IBP is the beta process of [Hjort 1990], while [Teh et al. 2007] gave a stick-breaking construction and developed efficient slice sampling inference algorithms for the IBP.

Nonparametric Tree Models are models that use distributions over trees that are consistent and exchangeable. [Blei et al. 2004] used a nested CRP to define distributions over trees with a finite number of levels. [Neal 2001, Neal 2003] defined Dirichlet diffusion trees, which are binary trees produced by a fragmentation process. [Teh et al. 2008] used Kingman’s coalescent [Kingman 1982b, Kingman 1982a] to produce random binary trees using a coalescent process. [Roy et al. 2007] proposed annotated hierarchies, using tree-consistent partitions first defined in [Heller and Ghahramani 2005] to model both relational and featural data.

Markov chain Monte Carlo Inference algorithms are the dominant approaches to inference in DP mixtures. [Neal 2000] is a good review of algorithms based on Gibbs sampling in the CRP representation. Algorithm 8 in [Neal 2000] is still one of the best algorithms based on simple local moves. [Ishwaran and James 2001] proposed blocked Gibbs sampling in the stick-breaking representation instead due to the simplicity in implementation. This has been further explored in [Porteous et al. 2006]. Since then there has been proposals for better MCMC samplers based on proposing larger moves in a Metropolis-Hastings framework [Jain and Neal 2004, Liang et al. 2007a], as well as sequential Monte Carlo [Fearnhead 2004, Mansingkha et al. 2007].

Other Approximate Inference Methods have also been proposed for DP mixture models. [Blei and Jordan 2006] is the first variational Bayesian approximation, and is based on a truncated stick-breaking representation. [Kurihara et al. 2007] proposed an
improved VB approximation based on a better truncation technique, and using KD-trees for extremely efficient inference in large scale applications. [Kurihara et al. 2007] studied improved VB approximations based on integrating out the stick-breaking weights. [Minka and Ghahramani 2003] derived an expectation propagation based algorithm. [Heller and Ghahramani 2005] derived tree-based approximation which can be seen as a Bayesian hierarchical clustering algorithm. [Daume III 2007] developed admissible search heuristics to find MAP clusterings in a DP mixture model.

**Computer Vision and Image Processing.** HDPs have been used in object tracking [Fox et al. 2006, Fox et al. 2007b, Fox et al. 2007a]. An extension called the transformed Dirichlet process has been used in scene analysis [Sudderth et al. 2006b, Sudderth et al. 2006a, Sudderth et al. 2008], a related extension has been used in fMRI image analysis [Kim and Smyth 2007, Kim 2007]. An extension of the infinite hidden Markov model called the nonparametric hidden Markov tree has been introduced and applied to image denoising [Kivinen et al. 2007a, Kivinen et al. 2007b].

**Natural Language Processing.** HDPs are essential ingredients in defining nonparametric context free grammars [Liang et al. 2007b, Finkel et al. 2007]. [Johnson et al. 2007] defined adaptor grammars, which is a framework generalizing both probabilistic context free grammars as well as a variety of nonparametric models including DPs and HDPs. DPs and HDPs have been used in information retrieval [Cowans 2004], word segmentation [Goldwater et al. 2006b], word morphology modelling [Goldwater et al. 2006a], coreference resolution [Haghighi and Klein 2007], topic modelling [Blei et al. 2004, Teh et al. 2006, Li et al. 2007]. An extension of the HDP called the hierarchical Pitman-Yor process has been applied to language modelling [Teh 2006a, Teh 2006b, Goldwater et al. 2006a].[Savova et al. 2007] used annotated hierarchies to construct syntactic hierarchies. Theses on nonparametric methods in NLP include [Cowans 2006, Goldwater 2006].

**Other Applications.** Applications of DPs, HDPs and infinite HMMs in bioinformatics include [Xing et al. 2004, Xing et al. 2007, Xing et al. 2006, Xing and Sohn 2007a, Xing and Sohn 2007b]. DPs have been applied in relational learning [Shafto et al. 2006, Kemp et al. 2006, Xu et al. 2006], spike sorting [Wood et al. 2006a, Görür 2007]. The HDP has been used in a cognitive model of categorization [Griffiths et al. 2007]. IBPs have been applied to infer hidden causes [Wood et al. 2006b], in a choice model [Görür et al. 2006], to modelling dyadic data [Meeds et al. 2007], to overlapping clustering [Heller and Ghahramani 2007], and to matrix factorization [Wood and Griffiths 2006].


Information retrieval using hierarchical Dirichlet processes.

Probabilistic Document Modelling.

Fast search for Dirichlet process mixture models.

Ewens, W. J. (1972).
The sampling theory of selectively neutral alleles.
Theoretical Population Biology, 3:87–112.

Particle filters for mixture models with an unknown number of components.
Statistics and Computing, 14:11–21.

A Bayesian analysis of some nonparametric problems.

The infinite tree.
Nonparametric Bayesian methods for large scale multi-target tracking.

Tracking a non-cooperative maneuvering target using hierarchical Dirichlet processes.
In *Proceedings of the Adaptive Sensor Array Processing Conference*.

Hierarchical Dirichlet processes for tracking maneuvering targets.
In *Proceedings of the International Conference on Information Fusion*.

Bayesian nonparametric latent feature models (with discussion and rejoinder).
In *Bayesian Statistics*, volume 8.

*Nonparametric Bayesian Models of Lexical Acquisition*.

Interpolating between types and tokens by estimating power-law generators.

Contextual dependencies in unsupervised word segmentation.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

A choice model with infinitely many latent features.

The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference.
Technical report, Department of Statistics, University of Warwick.

Unifying rational models of categorization via the hierarchical Dirichlet process.

Infinite latent feature models and the Indian buffet process.

Unsupervised coreference resolution in a nonparametric Bayesian model.

Bayesian hierarchical clustering.
References V

A nonparametric Bayesian approach to modeling overlapping clusters.

Nonparametric Bayes estimators based on beta processes in models for life history data.

Gibbs sampling methods for stick-breaking priors.

A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.
Technical report, Department of Statistics, University of Toronto.

Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models.

Learning systems of concepts with an infinite relational model.

Learning Hierarchical Probabilistic Models with Random Effects with Applications to Time-series and Image Data.
PhD thesis, Information and Computer Science, University of California at Irvine.
Hierarchical dirichlet processes with random effects.

The coalescent.

On the genealogy of large populations.
Essays in Statistical Science.

Image denoising with nonparametric hidden Markov trees.
In IEEE International Conference on Image Processing (ICIP), San Antonio, TX.

Learning multiscale representations of natural scenes using Dirichlet processes.
In IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil.

Infinite sparse factor analysis and infinite independent components analysis.

Accelerated variational DP mixture models.
Nonparametric Bayes pachinko allocation.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

A permutation-augmented sampler for Dirichlet process mixture models.
In *Proceedings of the International Conference on Machine Learning*.

The infinite PCFG using hierarchical Dirichlet processes.

Spatial nonparametric Bayesian models.

AClass: An online algorithm for generative classification.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Modeling dyadic data with binary latent factors.

Expectation propagation for infinite mixtures.
Presented at NIPS2003 Workshop on Nonparametric Bayesian Methods and Infinite Models.
References VIII

Bayesian mixture modeling.
In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211.

Markov chain sampling methods for Dirichlet process mixture models.

Defining priors for distributions using Dirichlet diffusion trees.
Technical Report 0104, Department of Statistics, University of Toronto.

Density modeling and clustering using Dirichlet diffusion trees.

Size-biased sampling of Poisson point processes and excursions.

Combinatorial stochastic processes.
Technical Report 621, Department of Statistics, University of California at Berkeley.
Lecture notes for St. Flour Summer School.

The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.
Gibbs sampling for (Coupled) infinite mixture models in the stick-breaking representation.

The infinite Gaussian mixture model.

Occam’s razor.

Infinite mixtures of Gaussian process experts.

Gaussian Processes for Machine Learning.
MIT Press.

Learning annotated hierarchies from relational data.

Discovering syntactic hierarchies.
A constructive definition of Dirichlet priors.

Learning cross-cutting systems of categories.

Time-varying topic models using dependent Dirichlet processes.
Technical Report UTML-TR-2005-003, Department of Computer Science, University of Toronto.

Shared segmentation of natural scenes using dependent Pitman-Yor processes.

Depth from familiar objects: A hierarchical model for 3D scenes.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Describing visual scenes using transformed Dirichlet processes.

Describing visual scenes using transformed objects and parts.
*International Journal of Computer Vision*, 77.


Hierarchical beta processes and the Indian buffet process.

A stochastic memoizer for sequence data.

A non-parametric Bayesian approach to spike sorting.

Particle filtering for nonparametric Bayesian matrix factorization.

A non-parametric Bayesian method for inferring hidden causes.

Bayesian haplotype inference via the dirichlet process.

Bayesian haplotype inference via the Dirichlet process.
Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space.
*Bayesian Analysis, 2*(2).

A nonparametric Bayesian approach for haplotype reconstruction from single and multi-population data.

Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture.

Infinite hidden relational models.