Bayesian Nonparametrics: Models Based on the Dirichlet Process

Alessandro Panella

Department of Computer Science
University of Illinois at Chicago

Machine Learning Seminar Series
February 18, 2013
Sources and Inspirations

Tutorials (slides)

Articles etc.
- ...
Outline

1. Introduction and background
   - Bayesian learning
   - Nonparametric models

2. Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3. Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4. A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5. The hierarchical Dirichlet process
Introduction and background

1. Introduction and background
   - Bayesian learning
   - Nonparametric models

2. Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3. Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4. A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5. The hierarchical Dirichlet process
The meaning of it all

BAYESIAN NONPARAMETRICS
The meaning of it all
The meaning of it all
Bayesian statistics

Estimate a parameter $\theta \in \Theta$ after observing data $x$.

Frequentist
Maximum Likelihood (ML): $\hat{\theta}_{MLE} = \arg\max_{\theta} p(x|\theta) = \arg\max_{\theta} \mathcal{L}(\theta : x)$

Bayesian

- Bayes Rule: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
- Bayesian prediction (using the whole posterior, not just one estimator)

$$p(x_{new}|x) = \int_{\Theta} p(x_{new}|\theta)p(\theta|x) \, d\theta$$

- Maximum A Posteriori (MAP)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(x|\theta)p(\theta)$$
Bayesian statistics

Estimate a parameter $\theta \in \Theta$ after observing data $x$.

**Frequentist**

Maximum Likelihood (ML): $\hat{\theta}_{MLE} = \arg\max_{\theta} p(x|\theta) = \arg\max_{\theta} \mathcal{L}(\theta : x)$

**Bayesian**

- Bayes Rule: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
- Bayesian prediction (using the whole posterior, not just one estimator)
  \[
p(x_{\text{new}}|x) = \int_{\Theta} p(x_{\text{new}}|\theta)p(\theta|x) \, d\theta
\]
- Maximum A Posteriori (MAP)
  \[
  \hat{\theta}_{MAP} = \arg\max_{\theta} p(x|\theta)p(\theta)
  \]
Bayesian statistics

Estimate a parameter $\theta \in \Theta$ after observing data $x$.

**Frequentist**

Maximum Likelihood (ML): $\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} p(x|\theta) = \operatorname{argmax}_{\theta} L(\theta : x)$

**Bayesian**

- Bayes Rule: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$

- Bayesian prediction (using the whole posterior, not just one estimator)

$$p(x_{\text{new}}|x) = \int_{\Theta} p(x_{\text{new}}|\theta)p(\theta|x) \, d\theta$$

- Maximum A Posteriori (MAP)

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} p(x|\theta)p(\theta)$$
De Finetti’s theorem

A premise:

**Definition**
An infinite sequence random variables \((x_1, x_2, \ldots)\) is said to be (infinitely) exchangeable if, for every \(N\) and every possible permutation \(\pi\) on \((1, \ldots, N)\),

\[
p(x_1, x_2, \ldots, x_N) = p(x_{\pi(1)}, x_{\pi(2)} \ldots, x_{\pi(N)})
\]

Note: exchangeability *not* equal i.i.d!

**Example (Polya Urn)**
An urn contains some red balls and some black balls; an infinite sequence of colors is drawn recursively as follows: draw a ball, mark down its color, then put the ball back in the urn along with an additional ball of the same color.
De Finetti’s theorem

A premise:

**Definition**
An infinite sequence random variables \((x_1, x_2, \ldots)\) is said to be (infinitely) exchangeable if, for every \(N\) and every possible permutation \(\pi\) on \((1, \ldots, N)\),

\[
p(x_1, x_2, \ldots, x_N) = p(x_{\pi(1)}, x_{\pi(2)} \ldots, x_{\pi(N)})
\]

Note: exchangeability not equal i.i.d!

**Example (Polya Urn)**
An urn contains some red balls and some black balls; an infinite sequence of colors is drawn recursively as follows: draw a ball, mark down its color, then put the ball back in the urn along with an additional ball of the same color.
De Finetti’s theorem (cont’d)

Theorem (De Finetti, 1935. Aka Representation Theorem)

A sequence of random variables \((x_1, x_2, \ldots)\) is infinitely exchangeable if for all \(N\), there exists a random variable \(\theta\) and a probability measure \(p\) on it such that

\[
p(x_1, x_2, \ldots, x_N) = \int_{\Theta} p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \, d\theta
\]

i.e., there exists a parameter space and a measure on it that makes the variables iid!

The representation theorem motivates (and encourages!) the use of Bayesian statistics.
De Finetti’s theorem (cont’d)

**Theorem (De Finetti, 1935. Aka Representation Theorem)**

A sequence of random variables \((x_1, x_2, \ldots)\) is infinitely exchangeable if for all \(N\), there exists a random variable \(\theta\) and a probability measure \(p\) on it such that

\[
p(x_1, x_2, \ldots, x_N) = \int_\Theta p(\theta) \prod_{i=1}^{N} p(x_i | \theta) \, d\theta
\]

i.e., there exists a parameter space and a measure on it that makes the variables iid!

The representation theorem motivates (and encourages!) the use of Bayesian statistics.
Bayesian learning

- Hypothesis space $\mathcal{H}$
- Given data $D$, compute

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)}$$

Then, we probably want to predict some future data $D'$, by either:

- Average over $\mathcal{H}$, i.e. $p(D'|D) = \int_\mathcal{H} p(D'|h)p(h|D)p(h) \, dh$
- Choose the MAP $h$ (or compute it directly), i.e. $p(D'|D) = p(D'|h_{MAP})$
- Sample from the posterior
- ...

- $\mathcal{H}$ can be anything! Bayesian learning as a general learning framework
- We will consider the case in which $h$ is a probabilistic model itself, i.e. a parameter vector $\theta$. 

Alessandro Panella (CS Dept. - UIC)
A simple example

Infer the bias $\theta \in [0, 1]$ of a coin after observing $N$ tosses.

- $H = 1, T = 0, p(H) = \theta$
- $h = \theta$, hence $H = [0, 1]$
- Sequence of Bernoulli trials:

\[
p(x_1, \ldots, x_n|\theta) = \theta^{n_H} (1 - \theta)^{N-n_H}
\]

where $n_H = \# \text{ heads}$.

- Unknown $\theta$:

\[
p(x_1, \ldots, x_N) = \int_0^1 \theta^{n_H} (1 - \theta)^{n_H-k} p(\theta) \, d\theta
\]

- Need to find a “good” prior $p(\theta)$
  Beta distribution!
A simple example (cont’d)

Beta distribution: $\theta \sim \text{Beta}(a, b)$
- $p(\theta|a, b) = \frac{1}{B(a,b)} \theta^{a-1}(1 - \theta)^{b-1}$
- Bayesian learning: $p(h|D) \propto p(D|h)p(h)$; for us:

$$p(\theta|x_1, \ldots, x_N) \propto p(x_1, \ldots, x_n|\theta)p(\theta)$$

$$= \theta^{n_H}(1 - \theta)^{n_T} \frac{1}{B(a,b)} \theta^{a-1}(1 - \theta)^{b-1}$$

$$\propto \theta^{n_H+a-1}(1 - \theta)^{n_T+b-1}$$

i.e. $\theta|x_1, \ldots, x_N \sim \text{Beta}(a + N_H, b + N_T)$

- We’re lucky! The Beta distribution is a conjugate prior to the binomial distribution.
A simple example (cont’d)

Beta distribution: $\theta \sim \text{Beta}(a, b)$

- $p(\theta|a, b) = \frac{1}{B(a,b)} \theta^{a-1} (1 - \theta)^{b-1}$

- Bayesian learning: $p(h|D) \propto p(D|h)p(h)$; for us:

$$p(\theta|x_1, \ldots, x_N) \propto p(x_1, \ldots, x_n|\theta)p(\theta)$$

$$= \theta^{n_H} (1 - \theta)^{n_T} \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

$$\propto \theta^{n_H + a - 1} (1 - \theta)^{n_T + b - 1}$$

i.e. $\theta|x_1, \ldots, x_N \sim \text{Beta}(a + N_H, b + N_T)$

- We’re lucky! The Beta distribution is a conjugate prior to the binomial distribution.
A simple example (cont’d)

Three sequences of four tosses:
Nonparametric models

“Nonparametric” doesn’t mean “no parameters”! Rather,

- The number of parameters grows as more data are observed.
- $\infty$-dimensional parameter space.
  - Finite data $\Rightarrow$ Bounded number of parameters

Definition

A nonparametric model is a Bayesian model on an $\infty$-dimensional parameter space.

Example

(from Orbanz and Teh, NIPS 2011)
Nonparametric models

“Nonparametric” doesn’t mean “no parameters”! Rather,
- The number of parameters grows as more data are observed.
- $\infty$-dimensional parameter space.
- Finite data $\Rightarrow$ Bounded number of parameters

Definition
A nonparametric model is a Bayesian model on an $\infty$-dimensional parameter space.

Example

(from Orbanz and Teh, NIPS 2011)
Outline

1. Introduction and background
   - Bayesian learning
   - Nonparametric models

2. Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3. Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4. A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5. The hierarchical Dirichlet process
Model

- **Generative** process.
- Expresses how we think the data is generated.
- Contains *hidden variables* (the subject of learning.)
- Specifies relations between variables.
- E.g. **graphical models**.

Posterior inference

- Knowing $p(D|M, \theta)$... $\leftarrow$ “how data is generated”
- ... compute $p(\theta|M, D)$
- Akin to “reversing” the generative process.
Models in Bayesian data analysis

Model

- **Generative** process.
- Expresses how we think the data is generated.
- Contains *hidden variables* (the subject of learning.)
- Specifies relations between variables.
- E.g. graphical models.

Posterior inference

- Knowing $p(D|M, \theta)$... $\leftarrow$ “how data is generated”
- ... compute $p(\theta|M, D)$
- Akin to “reversing” the generative process.
Finite mixture models (FMMs)

Bayesian approach to clustering. Each data point is assumed to belong to one of $K$ clusters.

**General form**
A sequence of data points $\mathbf{x} = (x_1, \ldots, x_N)$ each with probability

$$p(x_i | \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x_i | \theta_k) \quad \pi \in \Pi_{K-1}$$

**Generative process**
For each $i$:
- Draw a cluster assignment $z_i \sim \pi$
- Draw a data point $x_i \sim F(\theta_{z_i})$. 

Finite mixture models

Clustering with FMMs

FMMs (example)

Mixture of univariate Gaussians

- $\theta_k = (\mu_k, \sigma_k)$
- $x_i \sim \mathcal{N}(\mu_k, \sigma_k)$

$$p(x_i | \pi, \mu, \sigma) = \sum_{k=1}^{K} \pi_k f_N(x_i; \mu_k, \sigma_k)$$

\[ \pi = (0.15, 0.25, 0.6) \]

\[ \mathcal{N}(1, 1) \quad \mathcal{N}(4, 0.5) \quad \mathcal{N}(6, 0.7) \]
Clustering with FFMs

- Need priors for $\pi$, $\theta$
- Usually, $\pi$ is given a (symmetric) Dirichlet distribution prior.
- $\theta_k$’s are given a suitable prior $H$, depending on the data.

\[
\begin{align*}
\pi & \sim \text{Dir}(\alpha/K, \ldots, \alpha/K) \\
\theta_k | H & \sim H \\
\pi & \sim \pi \\
x_i | \theta, z_i & \sim F(\theta_{z_i}) \\
\end{align*}
\]
Dirichlet distribution

Multivariate generalization of Beta.

Dirichlet Distributions (from Teh, MLSC 2008)

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

Dir(5, 5, 2)  Dir(5, 2, 2)  Dir(0.7, 0.7, 0.7)

(from Teh, MLSC 2008)
Dirichlet distribution (cont’d)

\[ \pi \sim \text{Dir}(\alpha/K, \ldots, \alpha/K) \quad \text{iff} \quad p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha/K)} \prod_{k=1}^{K} \pi_k^{\alpha/K - 1} \]

Conjugate prior to categorical/multinomial, i.e.

\[ \pi \sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \]
\[ z_i \sim \pi \quad i = 1 \ldots N \]

implies

\[ \pi | z_1, \ldots, z_N \sim \text{Dir}\left(\frac{\alpha}{K} + n_1, \frac{\alpha}{K} + n_2, \ldots, \frac{\alpha}{K} + n_K\right) \]

Moreover,

\[ p(z_1, \ldots, z_N | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)} \]

and

\[ p(z_i = k | z^{(-i)}, \alpha) = \frac{n_k^{(-i)} + \alpha/K}{\alpha + N - 1} \]
Inference in FMMs

Clustering: infer $z$ (marginalize over $\pi, \theta$)

$$p(z|x, \alpha, H) = \frac{p(x|z, H)p(z|\alpha)}{\sum_z p(x|z, H)p(z|\alpha)}$$

where

$$p(z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}$$

$$p(x|z, H) = \int_{\Theta} \left[ \prod_{i=1}^{N} p(x_i|\theta_{z_i}) \prod_{k=1}^{K} H(\theta_k) \right] d\theta$$

Parameter estimation: infer $\pi, \theta$

$$p(\pi, \theta|x, \alpha, H) = \sum_z \left[ p(\pi|z, \alpha) \prod_{k=1}^{K} p(\theta_k|x, H) \right] p(z|x, \alpha, H)$$

$\Rightarrow$ No analytic procedure.
Inference in FMMs

**Clustering:** infer $z$ (marginalize over $\pi, \theta$)

\[
p(z|x, \alpha, H) = \frac{p(x|z, H)p(z|\alpha)}{\sum_z p(x|z, H)p(z|\alpha)}, \text{ where}
\]

\[
p(z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}
\]

\[
p(x|z, H) = \int_{\Theta} \left[ \prod_{i=1}^{N} p(x_i|\theta_{z_i}) \prod_{k=1}^{K} H(\theta_k) \right] d\theta
\]

**Parameter estimation:** infer $\pi, \theta$

\[
p(\pi, \theta|x, \alpha, H) = \sum_z \left[ p(\pi|z, \alpha) \prod_{k=1}^{K} p(\theta_k|x, H) \right] p(z|x, \alpha, H)
\]

$\Rightarrow$ No analytic procedure.
Inference in FMMs

Clustering: infer \( z \) (marginalize over \( \pi, \theta \))

\[
p(z|x, \alpha, H) = \frac{p(x|z, H)p(z|\alpha)}{\sum_z p(x|z, H)p(z|\alpha)}, \quad \text{where}
\]

\[
p(z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}
\]

\[
p(x|z, H) = \int_\Theta \left[ \prod_{i=1}^{N} p(x_i|\theta_{z_i}) \prod_{k=1}^{K} H(\theta_k) \right] d\theta
\]

Parameter estimation: infer \( \pi, \theta \)

\[
p(\pi, \theta|x, \alpha, H) = \sum_z \left[ p(\pi|z, \alpha) \prod_{k=1}^{K} p(\theta_k|x, H) \right] p(z|x, \alpha, H)
\]

\( \Rightarrow \) No analytic procedure.
Inference in FMMs

Clustering: infer $z$ (marginalize over $\pi$, $\theta$)

\[
p(z|x, \alpha, H) = \frac{p(x|z, H)p(z|\alpha)}{\sum_z p(x|z, H)p(z|\alpha)},
\]

where

\[
p(z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}
\]

\[
p(x|z, H) = \int_{\Theta} \left[ \prod_{i=1}^{N} p(x_i|\theta_{z_i}) \prod_{k=1}^{K} H(\theta_k) \right] d\theta
\]

Parameter estimation: infer $\pi$, $\theta$

\[
p(\pi, \theta|x, \alpha, H) = \sum_z \left[ p(\pi|z, \alpha) \prod_{k=1}^{K} p(\theta_k|x, H) \right] p(z|x, \alpha, H)
\]

⇒ No analytic procedure.
Inference in FMMs

Clustering: infer \( z \) (marginalize over \( \pi, \theta \))

\[
p(z|x, \alpha, H) = \frac{p(x|z, H)p(z|\alpha)}{\sum_z p(x|z, H)p(z|\alpha)},
\]

where

\[
p(z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}
\]

\[
p(x|z, H) = \int_{\Theta} \left[ \prod_{i=1}^{N} p(x_i|\theta_{z_i}) \prod_{k=1}^{K} H(\theta_k) \right] d\theta
\]

Parameter estimation: infer \( \pi, \theta \)

\[
p(\pi, \theta|x, \alpha, H) = \sum_z \left[ p(\pi|z, \alpha) \prod_{k=1}^{K} p(\theta_k|x, H) \right] p(z|x, \alpha, H)
\]

\( \Rightarrow \) No analytic procedure.
Approximate inference for FMMs

No exact inference because of the unknown clusters identifiers $z$

**Expectation-Maximization (EM)**

Widely used, but we will focus on MCMC because of the connection with Dirichlet Process.

**Gibbs sampling**

Markov chain Monte Carlo (MCMC) integration method

- Set of random variables $\mathbf{v} = \{v_1, v_2, \ldots, v_M\}$.
- We want to compute $p(\mathbf{v})$.
- Randomly initialize their values.
- At each iteration, sample a variable $v_i$ and hold the rest constant:

$$
v_i^{(t)} \sim p(v_i|v_j^{(t-1)}, j \neq i)
\quad \leftarrow \text{usually tractable}
$$

$$
v_j^{(t)} = v_j^{(t-1)}
$$

- This creates a Markov chain with $p(\mathbf{v})$ as equilibrium distribution.
Gibbs sampling for FMMs

State variables: $z_1, \ldots, z_N, \theta_1, \ldots, \theta_K, \pi$.

Conditional distributions:

$$p(\pi|z, \theta) = \text{Dir}\left(\alpha K + n_1, \ldots, \alpha K + n_k\right)$$

$$p(\theta_k|x, z) \propto p(\theta_k) \prod_{i:z_i=k} p(x_i|\theta_k)$$

$$= H(\theta_k) \prod_{i:z_i=k} F_{\theta_k}(x_i)$$

$$p(z_i = k|\pi, \theta, x) \propto p(z_i = k|\pi_k)p(x_i|z_i = k, \theta_k)$$

$$= \pi_k F_{\theta_k}(x_i)$$

We can avoid sampling $\pi$:

$$p(z_i = k|z_{-i}, \theta, x) \propto p(x_i|\theta_k)p(z_i = k|z_{-i})$$

$$\propto F_{\theta_k}(x_i)(n_k^{(-i)} + \alpha/K)$$
Gibbs sampling for FMMs (example)

Mixture of 4 bivariate Gaussians
Normal-inverse Wishart prior on $\theta_k = (\mu_k, \Sigma_k)$, conjugate to normal distribution.

$$\Sigma_k \sim \mathcal{W}(\nu, \Delta) \quad \mu_k \sim \mathcal{N}(\vartheta, \Sigma_k/\kappa)$$

Figure 2.18. Learning a mixture of $K=4$ Gaussians using the Gibbs sampler of Alg. 2.1. Columns show the current parameters after $T=2$ (top), $T=10$ (middle), and $T=40$ (bottom) iterations from two random initializations. Each plot is labeled by the current data log-likelihood.
Finite mixture models

FMMs: alternative representation

\[ \alpha \quad \pi \quad \theta_k \quad z_i \quad x_i \quad H \]

\[ \pi \sim \text{Dir}(\alpha) \]
\[ \theta_k \sim H \]
\[ z_i \sim \pi \]
\[ x_i \sim F(\theta_{z_i}) \]

\[ G(\theta) = \sum_{k=1}^{K} \pi_k \delta(\theta, \theta_k) \]
\[ \theta_k \sim H \]
\[ \pi \sim \text{Dir}(\alpha) \]
\[ \bar{\theta}_i \sim G \]
\[ x_i \sim F(\bar{\theta}_i) \]
Outline

1 Introduction and background
   - Bayesian learning
   - Nonparametric models

2 Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3 Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4 A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5 The hierarchical Dirichlet process
Going nonparametric!

The problem with finite FMMs

- What if \( K \) is unknown?
- How many parameters?

Idea

- Let’s use \( \infty \) parameters!
- We want something of the kind:

\[
p(x_i | \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k p(x_i | \theta_k)
\]

How to define such a measure?

- We’d like the nice conjugancy properties of Dirichlet to carry on…
- Is there such a thing, the \( \infty \) limit of a Dirichlet?
Going nonparametric!

The problem with finite FMMs

- What if $K$ is unknown?
- How many parameters?

Idea

- Let's use $\infty$ parameters!
- We want something of the kind:

$$p(x_i | \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k p(x_i | \theta_k)$$

How to define such a measure?

- We'd like the nice conjugancy properties of Dirichlet to carry on... 
- Is there such a thing, the $\infty$ limit of a Dirichlet?
Going nonparametric!

The problem with finite FMMs

- What if $K$ is unknown?
- How many parameters?

Idea

- Let’s use $\infty$ parameters!
- We want something of the kind:

$$p(x_i | \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k p(x_i | \theta_k)$$

How to define such a measure?

- We’d like the nice conjugancy properties of Dirichlet to carry on. . .
- Is there such a thing, the $\infty$ limit of a Dirichlet?
The (practical) Dirichlet process

The Dirichlet process is a distribution over probability measures over $\Theta$.

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

$H(\theta)$ is the base (mean) measure.

- Think $\mu$ for a Gaussian...
- ...but in the space of probability measures.

$\alpha$ is the concentration parameter.

- Controls the dispersion around the mean $H$. 
The Dirichlet process (cont’d)

A draw $G \sim \text{DP}(\alpha, H)$ is an infinite discrete probability measure:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k), \quad \text{where}$$

$\theta_k \sim H$, and $\pi$ is sampled from a “stick-breaking prior.”

Break a stick

Imagine a stick of length one. For $k = 1 \ldots \infty$, do the following:

- Break the stick at a point drawn from $\text{Beta}(1, \alpha)$.
- Let $\pi_k$ be such value and keep the remainder of the stick.

Following standard convention, we write $\pi \sim \text{GEM}(\alpha)$.

(Details in second part of talk)
The Dirichlet process (cont’d)

A draw $G \sim \text{DP}(\alpha, H)$ is an infinite discrete probability measure:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k), \quad \text{where}$$

$\theta_k \sim H$, and $\pi$ is sampled from a “stick-breaking prior.”

Break a stick
Imagine a stick of length one. For $k = 1 \ldots \infty$, do the following:

- Break the stick at a point drawn from $\text{Beta}(1, \alpha)$.
- Let $\pi_k$ be such value and keep the remainder of the stick.

Following standard convention, we write $\pi \sim \text{GEM}(\alpha)$.
(Details in second part of talk)
Stick-breaking, intuitively

- Small $\alpha \Rightarrow$ lots of weight assigned to few $\theta_k$'s.
  - $\Rightarrow$ $G$ will be very different from base measure $H$.
- Large $\alpha \Rightarrow$ weights equally distributed on $\theta_k$'s.
  - $\Rightarrow$ $G$ will resemble the base measure $H$.
Dirichlet process mixture models

The Dirichlet process

(from Navarro et al., 2005)
Let's use $G \sim DP(\alpha, H)$ to build an infinite mixture model.

$$G \sim DP(\alpha, H)$$

$$\bar{\theta}_i \sim G$$

$$x_i \sim F_{\bar{\theta}_i}$$
Using explicit clusters indicators $\mathbf{z} = (z_1, z_2, \ldots, z_N)$.

- $\pi \sim \text{GEM}(\alpha)$
- $\theta_k \sim H$ for $k = 1, \ldots, \infty$
- $z_i \sim \pi$
- $x_i \sim F_{\theta_{z_i}}$ for $i = 1, \ldots, N$
Chinese restaurant process

- So far, we only have a generative model.
- Is there a “nice” conjugacy property to use during inference?
- It turns out (details in part 2) that, if

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

the distribution \( p(z|\alpha) = \int p(z|\pi)p(\pi) \, d\pi \) is easily tractable, and is known as the Chinese restaurant process (CRP).
Restaurant with $\infty$ tables with $\infty$ capacity.

- $z_i =$ table at which customer $i$ sits upon entering.
  - Customer 1 sits at table 1
  - Customer 2 sits:
    - at table 1 w. prob $\propto 1$
    - at table 2 w. prob. $\propto \alpha$
  - Customer $i$ sits:
    - at table $k$ w. prob. $\propto n_k$ (# ppl at $k$)
    - at new table w. prob. $\propto \alpha$

\[
p(z_i = k) = \frac{n_k}{\alpha + i - 1}
\]
\[
p(z_i = k_{\text{new}}) = \frac{\alpha}{\alpha + i - 1}
\]
Chinese restaurant process (cont’d)

Restaurant with $\infty$ tables with $\infty$ capacity.

- $z_i = \text{table at which customer } i \text{ sits upon entering.}$

- **Customer 1 sits at table 1**
  - **Customer 2 sits:**
    - at table 1 w. prob $\propto 1$
    - at table 2 w. prob. $\propto \alpha$
  - **Customer } i \text{ sits:}
    - at table } k \text{ w. prob. } \propto n_k \text{ (# ppl at } k\text{)}
    - at new table w. prob. $\propto \alpha$

\[
p(z_i = k) = \frac{n_k}{\alpha + i - 1}
\]
\[
p(z_i = k_{\text{new}}) = \frac{\alpha}{\alpha + i - 1}
\]
Chinese restaurant process (cont’d)

Restaurant with ∞ tables with ∞ capacity.

- \( z_i \) = table at which customer \( i \) sits upon entering.
- Customer 1 sits at table 1
- Customer 2 sits:
  - at table 1 w. prob \( \propto 1 \)
  - at table 2 w. prob. \( \propto \alpha \)
- Customer \( i \) sits:
  - at table \( k \) w. prob. \( \propto n_k \) (# ppl at \( k \))
  - at new table w. prob. \( \propto \alpha \)

\[
p(z_i = k) = \frac{n_k}{\alpha + i - 1}
\]

\[
p(z_i = k_{\text{new}}) = \frac{\alpha}{\alpha + i - 1}
\]
Chinese restaurant process (cont’d)

Restaurant with $\infty$ tables with $\infty$ capacity.

- $z_i =$ table at which customer $i$ sits upon entering.
- Customer 1 sits at table 1
- Customer 2 sits:
  - at table 1 w. prob $\propto 1$
  - at table 2 w. prob. $\propto \alpha$
- Customer $i$ sits:
  - at table $k$ w. prob. $\propto n_k$ (# ppl at $k$)
  - at new table w. prob. $\propto \alpha$

\[
p(z_i = k) = \frac{n_k}{\alpha + i - 1}
\]
\[
p(z_i = k_{new}) = \frac{\alpha}{\alpha + i - 1}
\]
Chinese restaurant process (cont’d)

Restaurant with $\infty$ tables with $\infty$ capacity.

- $z_i$ = table at which customer $i$ sits upon entering.
- Customer 1 sits at table 1
- Customer 2 sits:
  - at table 1 w. prob $\propto 1$
  - at table 2 w. prob. $\propto \alpha$
- Customer $i$ sits:
  - at table $k$ w. prob. $\propto n_k$ (# ppl at $k$)
  - at new table w. prob. $\propto \alpha$

\[
p(z_i = k) = \frac{n_k}{\alpha + i - 1} \\
p(z_i = k_{\text{new}}) = \frac{\alpha}{\alpha + i - 1}
\]
Gibbs sampling for DPMMs

- Via the CRP, we can find the conditional distributions for Gibbs sampling.
- State: $\theta_1, \ldots, \theta_k, z$.

\[
p(\theta_k|x, z) \propto p(\theta_k) \prod_{i: z_i = k} p(x_i|\theta_k) = h(\theta_k)f(x_i|\theta_k)
\]

\[
p(z_i = k|z_{-i}, \theta, x) \propto p(x_i|\theta_k)p(z_i = k|z_{-i}) \propto \begin{cases} n_k^{(-i)}f(x_i|\theta_k) & \text{existing } k \\ \alpha f(x_i|\theta_k) & \text{new } k \end{cases}
\]

$K$ grows as more data are observed, asymptotically as $\alpha \log n$. 
Gibbs sampling for DPMMs (example)

Mixture of bivariate Gaussians

T=2

T=10

T=40

(from Sudderth, 2008)
END OF FIRST PART.
Outline

1. Introduction and background
   - Bayesian learning
   - Nonparametric models

2. Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3. Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4. A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5. The hierarchical Dirichlet process
De Finetti’s REDUX

Theorem (De Finetti, 1935. Aka Representation Theorem)

A sequence of random variables \((x_1, x_2, \ldots)\) is infinitely exchangeable if for all \(N\), there exists a random variable \(\theta\) and a probability measure \(p\) on it such that

\[
p(x_1, x_2, \ldots, x_N) = \int_{\Theta} p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \, d\theta
\]

- The theorem wouldn’t be true if \(\theta\)’s range is limited to Euclidean’s vector spaces.
- We need to allow \(\theta\) to range over measures.
- \(\Rightarrow p(\theta)\) is a distribution on measures, like the DP.
De Finetti’s REDUX

Theorem (De Finetti, 1935. Aka Representation Theorem)

A sequence of random variables \((x_1, x_2, \ldots)\) is infinitely exchangeable if for all \(N\), there exists a random variable \(\theta\) and a probability measure \(p\) on it such that

\[
p(x_1, x_2, \ldots, x_N) = \int_\Theta p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \, d\theta
\]

- The theorem wouldn’t be true if \(\theta\)’s range is limited to Euclidean’s vector spaces.
- We need to allow \(\theta\) to range over measures.
- \(\Rightarrow p(\theta)\) is a distribution on measures, like the DP.
Dirichlet Process REDUX

Definition
Let $\Theta$ be a measurable space (of parameters), $H$ be a probability distribution on $\Theta$, and $\alpha$ a positive scalar. A Dirichlet process is the distribution of a random probability measure $G$ over $\Theta$, such that for any finite partition $(T_1, \ldots, T_k)$ of $\Theta$, we have

$$(G(T_1), \ldots, G(T_K)) \sim \text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)).$$

$\Theta$

$T_1$

$T_2$

$T_3$

$\tilde{T}_1$

$\tilde{T}_2$

$\tilde{T}_3$

$\tilde{T}_4$

$\tilde{T}_5$

(From Sudderth, 2008)

$\mathbb{E}[G(T_k)] = H(T_k)$
Posterior conjugancy

Via the conjugancy of the Dirichlet distribution, we know that:

\[ p(G(T_1), \ldots, G(T_K) | \bar{\theta} \in T_k) = \text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_k) + 1, \ldots, \alpha H(T_K)) \]

Formalizing this analysis, we obtain that if

\[ G \sim \text{DP}(\alpha, H) \]
\[ \bar{\theta}_i \sim G \quad i = 1, \ldots, N, \]

the posterior measure also follows a Dirichlet process:

\[ p(G | \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \text{DP} \left( \alpha + N, \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i} \right) \right) \]

The DP defines a conjugate prior for distributions on arbitrary measure spaces.
Generating samples: stick breaking


\[ G(\theta) \sim \text{DP}(\alpha, H) \quad \text{iff} \quad G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k), \]

where \( \theta \sim H \), and

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

(from Sudderth, 2008)
Stick-breaking (derivation) [Teh 2007]

- We know that (posterior):

\[
G \sim \text{DP}(\alpha, H)\quad \theta \mid G \sim G \quad \iff \quad \theta \sim H \quad G \mid \theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right)
\]

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

\[
(G(\Theta), G(\Theta \setminus \theta)) \sim \text{Dir} \left( (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\theta), (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\Theta \setminus \theta) \right)
\]

\[
= \text{Dir}(1, \alpha) = \text{Beta}(1, \alpha)
\]

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

\[
G = \beta \delta_\theta + (1 - \beta)G'
\]

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

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

- What is \(G'\)?
Stick-breaking (derivation) [Teh 2007]

- We know that (posterior):
  \[ G \sim \text{DP}(\alpha, H) \]
  \[ \theta | G \sim G \iff \theta \sim H \]
  \[ G | \theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right) \]

- Consider the partition \((\Theta, \Theta \setminus \theta)\) of \(\Theta\). We have:
  \[
  (G(\Theta), G(\Theta \setminus \theta)) \sim \text{Dir} \left( (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\theta), (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1} (\Theta \setminus \theta) \right)
  = \text{Dir}(1, \alpha) = \text{Beta}(1, \alpha)
  
- \(G\) has point mass located at \(\theta\):
  \[ G = \beta \delta_\theta + (1 - \beta) G' \]
  \[ \beta \sim \text{Beta}(1, \alpha) \]
  and \(G'\) is the renormalized probability measure with the point mass removed.

What is \(G'\)?
Stick-breaking (derivation) [Teh 2007]

- We know that (posterior):

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

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

\[
(G(\Theta), G(\Theta \setminus \theta)) \sim \text{Dir}\left((\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1}(\theta), (\alpha + 1) \frac{\alpha H + \delta_\theta}{\alpha + 1}(\Theta \setminus \theta)\right)
\]

\[
= \text{Dir}(1, \alpha) = \text{Beta}(1, \alpha)
\]

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

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

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

- What is \(G'\)?
Stick-breaking (derivation) [Teh 2007]

- We know that (posterior):
  \[ G \sim \text{DP}(\alpha, H) \]
  \[ \theta | G \sim G \iff G|\theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_{\theta}}{\alpha + 1} \right) \]

- Consider the partition \((\Theta, \Theta \setminus \theta)\) of \(\Theta\). We have:
  \[ (G(\Theta), G(\Theta \setminus \theta)) \sim \text{Dir} \left( (\alpha + 1) \frac{\alpha H + \delta_{\theta}}{\alpha + 1} (\theta), (\alpha + 1) \frac{\alpha H + \delta_{\theta}}{\alpha + 1} (\Theta \setminus \theta) \right) \]
  \[ = \text{Dir}(1, \alpha) = \text{Beta}(1, \alpha) \]

- \(G\) has point mass located at \(\theta\):
  \[ G = \beta \delta_{\theta} + (1 - \beta)G' \quad \beta \sim \text{Beta}(1, \alpha) \]

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

- What is \(G'\)?
Stick-breaking (derivation) [Teh 2007]

- We have:

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

\[
\Rightarrow \\
\theta \sim H \\
G | \theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right) \\
G = \beta \delta_\theta + (1 - \beta)G' \\
\beta \sim \text{Beta}(1, \alpha)
\]

- Consider a further partition \(\theta, T_1, \ldots, T_K\) of \(\Theta\):

\[
(G(\theta), G(T_1), \ldots, G(T_K)) = (\beta, (1 - \beta)G'(T_1), \ldots, (1 - \beta)G'(T_K)) \\
\sim \text{Dir}(1, \alpha H(T_1), \ldots, \alpha H(T_K))
\]

- Using the agglomerative/decimative property of Dirichlet, we get

\[
(G'(T_1), \ldots, G'(T_K)) \sim \text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)) \\
G' \sim \text{DP}(\alpha, H)
\]
Stick-breaking (derivation) [Teh 2007]

- We have:

\[ G \sim \text{DP}(\alpha, H) \]
\[ \theta | G \sim G \]
\[ \Rightarrow \]
\[ \theta \sim H \]
\[ G | \theta \sim \text{DP} \left( \alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1} \right) \]
\[ G = \beta \delta_\theta + (1 - \beta) G' \]
\[ \beta \sim \text{Beta}(1, \alpha) \]

- Consider a further partition \( \theta, T_1, \ldots, T_K \) of \( \Theta \):

\[(G(\theta), G(T_1), \ldots, G(T_K)) = (\beta, (1 - \beta) G'(T_1), \ldots, (1 - \beta) G'(T_K)) \]
\[\sim \text{Dir}(1, \alpha H(T_1), \ldots, \alpha H(T_K)) \]

- Using the agglomerative/decimative property of Dirichlet, we get

\[(G'(T_1), \ldots, G'(T_K)) \sim \text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)) \]
\[G' \sim \text{DP}(\alpha, H) \]
A little more theory... Dirichlet process REDUX

Stick-breaking (derivation) [Teh 2007]

- We have:

\[
G \sim \text{DP}(\alpha, H) \\
\theta | G \sim G \Rightarrow \\
\theta \sim H \\
G | \theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta}{\alpha + 1}\right) \\
G = \beta \delta_{\theta} + (1 - \beta)G'
\]

- \(\beta \sim \text{Beta}(1, \alpha)\)

- Consider a further partition \(\theta, T_1, \ldots, T_K\) of \(\Theta\):

\[
(G(\theta), G(T_1), \ldots, G(T_K)) = (\beta, (1 - \beta)G'(T_1), \ldots, (1 - \beta)G'(T_K)) \\
\sim \text{Dir}(1, \alpha H(T_1), \ldots, \alpha H(T_K))
\]

- Using the agglomerative/decimative property of Dirichlet, we get

\[
(G'(T_1), \ldots, G'(T_K)) \sim \text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)) \\
G' \sim \text{DP}(\alpha, H)
\]
Therefore,

\[ G \sim \text{DP}(\alpha, H) \]
\[ G = \beta_1 \delta_{\theta_1} + (1 - \beta_1)G_1 \]
\[ G = \beta_1 \delta_{\theta_1} + (1 - \beta_1)(\beta_2 \delta_{\theta_2} + (1 - \beta_2)G_2) \]
\[ \vdots \]
\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \]

where

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

which is the stick-breaking construction.
Chinese restaurant (derivation)

Once again, we start from the posterior:

$$p(G|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \text{DP} \left( \alpha + N, \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i} \right) \right)$$

The expected measure of any subset $T \subset \Theta$ is:

$$\mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}(T) \right)$$

Since $G$ is discrete, some of the $\{\bar{\theta}_i\}_{i=1}^{N} \sim G$ take identical values. Assume $K \leq N$ unique values:

$$\mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{K} N_k \delta_{\bar{\theta}_i}(T) \right)$$
Chinese restaurant (derivation)

Once again, we start from the posterior:

\[ p(G|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \text{DP} \left( \alpha + N, \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i} \right) \right) \]

The expected measure of any subset \( T \subset \Theta \) is:

\[ \mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}(T) \right) \]

Since \( G \) is discrete, some of the \( \{\bar{\theta}_i\}_{i=1}^{N} \sim G \) take identical values. Assume \( K \leq N \) unique values:

\[ \mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{K} N_k \delta_{\bar{\theta}_i}(T) \right) \]
Chinese restaurant (derivation)

Once again, we start from the posterior:

\[ p(G|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \text{DP} \left( \alpha + N, \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i} \right) \right) \]

The expected measure of any subset \( T \subset \Theta \) is:

\[ \mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}(T) \right) \]

Since \( G \) is discrete, some of the \( \{\bar{\theta}_i\}_{i=1}^{N} \sim G \) take identical values. Assume \( K \leq N \) unique values:

\[ \mathbb{E}[G(T)|\bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H] = \frac{1}{\alpha + N} \left( \alpha H + \sum_{i=1}^{K} N_k \delta_{\bar{\theta}_i}(T) \right) \]
Chinese restaurant (derivation)

- A bit informally...
- Let $T_k$ contain $\theta_k$ and shrink it arbitrarily. To the limit, we have that

$$p(\tilde{\theta}_{N+1} = \theta | \tilde{\theta}_1, \ldots, \tilde{\theta}_N, \alpha, H) = \frac{1}{\alpha + N} \left( \alpha h(\theta) + \sum_{i=1}^{K} N_k \delta_{\tilde{\theta}_i}(\theta) \right)$$

This is the generalized Polya urn scheme
An urn contains one ball for each preceding observation, with a different color for each distinct $\theta_k$. For each ball drawn from the urn, we replace that ball and add one more ball of the same color. There is a special “weighted” ball which is drawn with probability proportional to $\alpha$ normal balls, and has a new, previously unseen color $\tilde{\theta}_k$. [This description is from Sudderth, 2008]

- This allows to sample from a Dirichlet process without explicitly constructing the underlying $G \sim \text{DP}(\alpha, H)$. 
Chinese restaurant (derivation)

- A bit informally...
- Let $T_k$ contain $\theta_k$ and shrink it arbitrarily. To the limit, we have that

$$p(\bar{\theta}_{N+1} = \theta | \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N} \left( \alpha h(\theta) + \sum_{i=1}^{K} N_k \delta_{\bar{\theta}_i}(\theta) \right)$$

This is the generalized Polya urn scheme

An urn contains one ball for each preceding observation, with a different color for each distinct $\theta_k$. For each ball drawn from the urn, we replace that ball and add one more ball of the same color. There is a special “weighted” ball which is drawn with probability proportional to $\alpha$ normal balls, and has a new, previously unseen color $\theta_{\bar{k}}$. [This description is from Sudderth, 2008]

- This allows to sample from a Dirichlet process without explicitly constructing the underlying $G \sim DP(\alpha, H)$. 

Alessandro Panella (CS Dept. - UIC)  Bayesian Nonparametrics  February 18, 2013  48 / 57
The Dirichlet process implicitly partitions the data.

Let $z_i$ indicate the subset (cluster) associated with the $i$th observation, i.e. $\bar{\theta}_i = \theta_{z_i}$.

From the previous slide, we get:

$$p(z_{N+1} = z | z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left( \alpha \delta(z, \bar{k}) + \sum_{i=1}^{K} N_k \delta(z, k) \right)$$

This is the Chinese restaurant process (CRP).

It induces an exchangeable distribution on partitions.

The joint distribution is invariant to the order the observations are assigned to clusters.
Chinese restaurant (derivation)

- The Dirichlet process implicitly *partitions* the data.
- Let $z_i$ indicate the subset (cluster) associated with the *i*th observation, i.e. $\tilde{\theta}_i = \theta_{z_i}$.
- From the previous slide, we get:

$$p(z_{N+1} = z|z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left( \alpha \delta(z, \bar{k}) + \sum_{i=1}^{K} N_k \delta(z, k) \right)$$

This is the **Chinese restaurant process (CRP).**

It induces an *exchangeable* distribution on partitions.

- The joint distribution is invariant to the order the observations are assigned to clusters.
Take away message

These representations are all equivalent!

- **Posterior DP:**
  \[
  G \sim \text{DP}(\alpha, H) \\
  \theta | G \sim G \\
  \theta \sim H \\
  G | \theta \sim \text{DP}(\alpha + 1, \frac{\alpha H + \delta \theta}{\alpha + 1})
  \]

- **Stick-breaking construction:**
  \[
  G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \\
  \theta_k \sim H \\
  \pi \sim \text{GEM}(\alpha)
  \]

- **Generalized Polya urn**
  \[
  p(\bar{\theta}_{N+1} = \theta | \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N} \left( \alpha h(\theta) + \sum_{i=1}^{K} N_k \delta_{\bar{\theta}_i}(\theta) \right)
  \]

- **Chinese restaurant process**
  \[
  p(z_{N+1} = z | z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left( \alpha \delta(z, \bar{k}) + \sum_{i=1}^{K} N_k \delta(z, k) \right)
  \]
Outline

1. Introduction and background
   - Bayesian learning
   - Nonparametric models

2. Finite mixture models
   - Bayesian models
   - Clustering with FMMs
   - Inference

3. Dirichlet process mixture models
   - Going nonparametric!
   - The Dirichlet process
   - DP mixture models
   - Inference

4. A little more theory...
   - De Finetti’s REDUX
   - Dirichlet process REDUX

5. The hierarchical Dirichlet process
The DP mixture model (DPMM)

Let's use $G \sim DP(\alpha, H)$ to build an infinite mixture model.

$G \sim DP(\alpha, H)$

$\bar{\theta}_i \sim G$

$x_i \sim F_{\bar{\theta}_i}$
Related subgroups of data

- Dataset with $J$ related groups $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$.
- $\mathbf{x}_j = (x_{j1}, \ldots, x_{jN_j})$ contains $N_j$ observations.
- We want these groups to share clusters (transfer knowledge.)

(from Jordan, 2005)
Hierarchical Dirichlet process (HDP)

- **Global probability measure** \( G_0 \sim \text{DP}(\gamma, \mathcal{H}) \)
  - Defines a set of shared clusters.
  \[
  G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \quad \theta_k \sim \mathcal{H} \\
  \beta \sim \text{GEM}(\gamma)
  \]

- **Group specific distributions** \( G_j \sim \text{DP}(\alpha, G_0) \)
  \[
  G_j(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_k \delta(\theta, \tilde{\theta}_k) \\
  \tilde{\theta}_t \sim G_0 \\
  \tilde{\pi} \sim \text{GEM}(\gamma)
  \]

- Note \( G_0 \) as base measure!
- Each local cluster has parameter \( \tilde{\theta}_k \) copied from some global cluster
- For each group, data points are generated according to:
  \[
  \bar{\theta}_{ji} \sim G_j \\
  x_{ji} \sim F(\bar{\theta}_{ji})
  \]
Hierarchical Dirichlet process (HDP)

- **Global probability measure** $G_0 \sim \text{DP}(\gamma, H)$
  - Defines a set of shared clusters.
  
  $$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \quad \theta_k \sim H$$

  $$\beta \sim \text{GEM}(\gamma)$$

- **Group specific distributions** $G_j \sim \text{DP}(\alpha, G_0)$
  
  $$G_j(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_k \delta(\theta, \tilde{\theta}_k) \quad \tilde{\theta}_t \sim G_0$$

  $$\tilde{\pi} \sim \text{GEM}(\gamma)$$

- **Note** $G_0$ as base measure!
- **Each local cluster has parameter** $\tilde{\theta}_k$ *copied* from some global cluster

- For each group, data points are generated according to:

  $$\tilde{\theta}_{ji} \sim G_j$$

  $$x_{ji} \sim F(\tilde{\theta}_{ji})$$
Hierarchical Dirichlet process (HDP)

- **Global probability measure** \( G_0 \sim \text{DP}(\gamma, H) \)
  - Defines a set of shared clusters.
  
  \[
  G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \quad \theta_k \sim H \\
  \beta \sim \text{GEM}(\gamma)
  \]

- **Group specific distributions** \( G_j \sim \text{DP}(\alpha, G_0) \)
  
  \[
  G_j(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_k \delta(\theta, \tilde{\theta}_k) \\
  \tilde{\theta}_t \sim G_0 \\
  \tilde{\pi} \sim \text{GEM}(\gamma)
  \]

- Note \( G_0 \) as base measure!
- Each local cluster has parameter \( \tilde{\theta}_k \) copied from some global cluster

- For each group, data points are generated according to:
  
  \[
  \tilde{\theta}_{ji} \sim G_j \\
  x_{ji} \sim F(\tilde{\theta}_{ji})
  \]
The HDP mixture model (DPMM)

\[ G_0 \sim \text{DP}(\gamma, H) \]
\[ G_j \sim \text{DP}(\alpha, G_0) \]
\[ \bar{\theta}_{ji} \sim G_j \]
\[ x_{ji} \sim F(\bar{\theta}_{ji}) \]
The hierarchical Dirichlet process

The HDP mixture model (DPMM)

\[ G_j(\theta) = \sum_{t=1}^{\infty} \tilde{\pi}_k \delta(\theta, \tilde{\theta}_k) \]

\[ \tilde{\theta}_t \sim G_0 \]

\[ \tilde{\pi} \sim \text{GEM}(\gamma) \]

- \( G_0 \) is discrete.
- Each group might create several copies of the same global cluster.
- Aggregating the probabilities:

\[ G_j(\theta) = \sum_{t=1}^{\infty} \pi_k \delta(\theta, \tilde{\theta}_k) \]

\[ \pi_{jk} = \sum_{t:kjt=k} \tilde{\pi}_{jt} \]

It can be shown that \( \pi \sim \text{DP}(\alpha, \beta) \).

- \( \beta = (\beta_1, \beta_2, \ldots) \): average weight of local clusters.
- \( \pi = (\pi_1, \pi_2, \ldots) \) group-specific weights.
- \( \alpha \) controls the variability of clusters weight across groups.
THANK YOU. QUESTIONS?