# A Tutorial on Dirichlet Processes and Hierarchical Dirichlet Processes

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

Mar 1, 2007 / CUED

# Outline

1. Dirichlet Processes
   - Definitions, Existence, and Representations (recap)
   - Applications
   - Generalizations
   - Generalizations

2. Hierarchical Dirichlet Processes
   - Grouped Clustering Problems
   - Hierarchical Dirichlet Processes
   - Representations
   - Applications
   - Extensions and Related Models

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

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

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

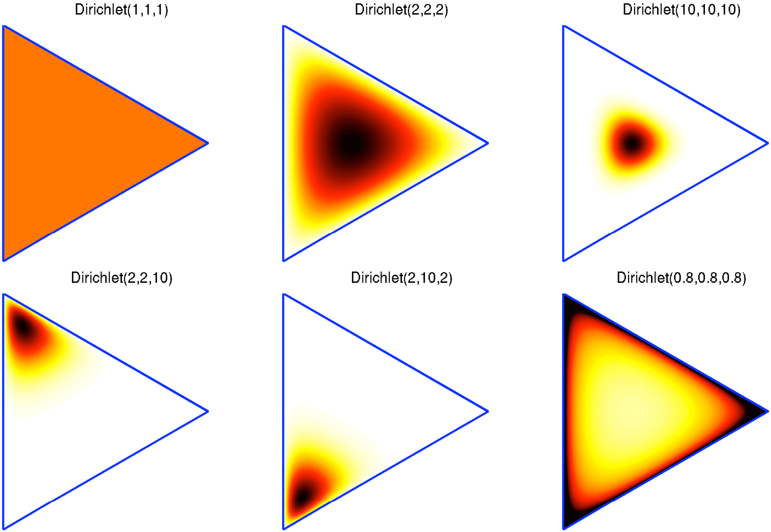
$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

  with parameters $(\alpha_1, \ldots, \alpha_K)$, if

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

# Dirichlet Processes

## Examples of Dirichlet distributions

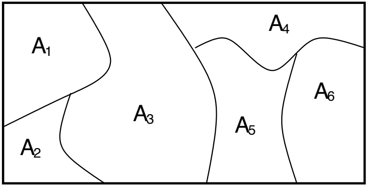


- A Dirichlet Process (DP) is a distribution over probability measures.
- A DP has two parameters:
  - Base distribution $H$, which is like the *mean* of the DP.
  - Strength parameter $\alpha$, which is like an *inverse-variance* of the DP.
- We write:

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

if for any partition $(A_1, \ldots, A_n)$ of $\mathbb{X}$:

$$(G(A_1), \ldots, G(A_n)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$



- A DP has two parameters:
  - Base distribution $H$, which is like the *mean* of the DP.
  - Strength parameter $\alpha$, which is like an *inverse-variance* of the DP.
- The first two cumulants of the DP:

$$\text{Expectation:} \qquad \mathbb{E}[G(A)] = H(A)$$

$$\text{Variance:} \qquad \mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

where $A$ is any measurable subset of $\mathbb{X}$.

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: if we can prescribe consistent finite dimensional distributions, then a distribution over functions exist.
- de Finetti's Theorem: if we have an infinite exchangeable sequence of random variables, then a distribution over measures exist making them independent. Pòlya's urn, Chinese restaurant process.
- Stick-breaking Construction: Just construct it.

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: if we can prescribe consistent finite dimensional distributions, then a distribution over functions exist.
- de Finetti's Theorem: if we have an infinite exchangeable sequence of random variables, then a distribution over measures exist making them independent. Pòlya's urn, Chinese restaurant process.
- Stick-breaking Construction: Just construct it.

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: if we can prescribe consistent finite dimensional distributions, then a distribution over functions exist.
- de Finetti's Theorem: if we have an infinite exchangeable sequence of random variables, then a distribution over measures exist making them independent. Pòlya's urn, Chinese restaurant process.
- Stick-breaking Construction: Just construct it.

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: if we can prescribe consistent finite dimensional distributions, then a distribution over functions exist.
- de Finetti's Theorem: if we have an infinite exchangeable sequence of random variables, then a distribution over measures exist making them independent. Pòlya's urn, Chinese restaurant process.
- Stick-breaking Construction: Just construct it.

- Distribution over probability measures.

$$(G(A_1), \ldots, G(A_n)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$

- Chinese restaurant process/Pòlya's urn scheme.

$$P(n^{\text{th}} \text{ customer sit at table } k) = \frac{n_k}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \frac{\alpha}{i-1+\alpha}$$

- Stick-breaking construction.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \qquad \beta_k \sim \text{Beta}(1, \alpha)$$

- Distribution over probability measures.

$$(G(A_1),\dots,G(A_n)) \sim \text{Dirichlet}(\alpha H(A_1),\dots,\alpha H(A_n))$$

- Chinese restaurant process/Pòlya's urn scheme.

$$P(n^{\text{th}} \text{ customer sit at table } k) = \frac{n_k}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \frac{\alpha}{i-1+\alpha}$$

- Stick-breaking construction.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1-\beta_k) \qquad \beta_k \sim \text{Beta}(1,\alpha)$$

- Distribution over probability measures.

$$(G(A_1), \ldots, G(A_n)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$

- Chinese restaurant process/Pòlya's urn scheme.

$$P(n^{\text{th}} \text{ customer sit at table } k) = \frac{n_k}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \frac{\alpha}{i-1+\alpha}$$

- Stick-breaking construction.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \qquad \beta_k \sim \text{Beta}(1, \alpha)$$

# Dirichlet Processes
## Representations

- Distribution over probability measures.

$$(G(A_1), \ldots, G(A_n)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$

- Chinese restaurant process/Pòlya's urn scheme.

$$P(n^{\text{th}} \text{ customer sit at table } k) = \tfrac{n_k}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \tfrac{\alpha}{i-1+\alpha}$$

- Stick-breaking construction.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \qquad \beta_k \sim \text{Beta}(1, \alpha)$$

# Pòlya's Urn Scheme

- A draw $G \sim \mathrm{DP}(\alpha, H)$ is a random probability measure.
- Treating $G$ as a distribution, consider i.i.d. draws from $G$:

$$\theta_i | G \sim G$$

- Marginalizing out $G$, marginally each $\theta_i \sim H$, while the conditional distributions are,

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

- This is the Pòlya urn scheme.

# Pòlya's Urn Scheme

- A draw $G \sim \mathrm{DP}(\alpha, H)$ is a random probability measure.
- Treating $G$ as a distribution, consider i.i.d. draws from $G$:

$$\theta_i | G \sim G$$

- Marginalizing out $G$, marginally each $\theta_i \sim H$, while the conditional distributions are,

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

- This is the Pòlya urn scheme.

# Pòlya's Urn Scheme

- Pòlya's urn scheme produces a sequence $\theta_1, \theta_2, \ldots$ with the following conditionals:

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

- Imagine 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 that color into the urn.
  - With probability $\propto n-1$, pick a ball at random from the urn, record $\theta_n$ to be its color, return the ball into the urn and place a second ball of same color into urn.

# Exchangeability and de Finetti's Theorem

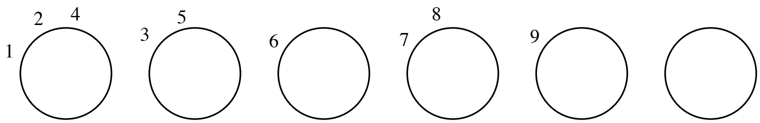
- Starting with a DP, we constructed Pòlya's urn scheme.
- The reverse is possible using de Finetti's Theorem.
- Since $\theta_i$ are i.i.d. $\sim G$, their joint distribution is invariant to permutations, thus $\theta_1, \theta_2, \ldots$ are exchangeable.
- Thus a distribution over measures must exist making them i.i.d..
- This is the DP.

# Chinese Restaurant Process

- Draw $\theta_1, \ldots, \theta_n$ from a Pòlya's urn scheme.
- They 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 $\theta_i = \theta_k^*$.
- Random draws $\theta_1, \ldots, \theta_n$ from a Pòlya's urn scheme induces a random partition of $1, \ldots, n$.
- The induced distribution over partitions is a Chinese restaurant process (CRP).

# Chinese Restaurant Process

- 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.



# Chinese Restaurant Process

- To get back from the CRP to Pòlya's urn scheme, simply draw

  $$\theta_k^* \sim H$$

  for $k = 1, \ldots, K$, then for $i = 1, \ldots, n$ set

  $$\theta_i = \theta_{k_i}^*$$

  where $k_i$ is the table that customer $i$ sat at.
- The CRP teases apart the clustering property of the DP, from the base distribution.

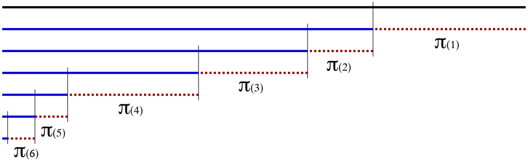
# Stick-breaking Construction

- But how do draws $G \sim \mathrm{DP}(\alpha, H)$ look like?
  - $G$ is discrete with probability one, so:

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

- The stick-breaking construction shows that $G \sim \mathrm{DP}(\alpha, H)$ if:

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$
$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$
$$\theta_k^* \sim H$$



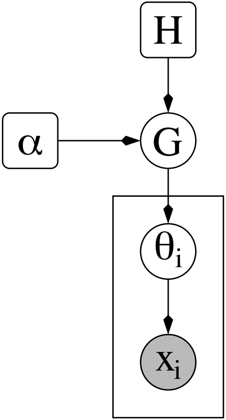

- We write $\pi \sim \mathrm{GEM}(\alpha)$ if $\pi = (\pi_1, \pi_2, \ldots)$ is distributed as above.

# Applications

- Mixture Modelling.
- Haplotype Inference.
- Nonparametric relaxation of parametric models.

- We model a data set $x_1, \ldots, x_n$ using the following model:

$$
\begin{aligned}
x_i &\sim F(\theta_i) \qquad \text{for } i = 1, \ldots, n \\
\theta_i &\sim G \\
G &\sim \mathrm{DP}(\alpha, H)
\end{aligned}
$$




- Each $\theta_i$ is a latent parameter modelling $x_i$, while $G$ is the unknown distribution over parameters modelled using a DP.
- This is the basic DP mixture model.

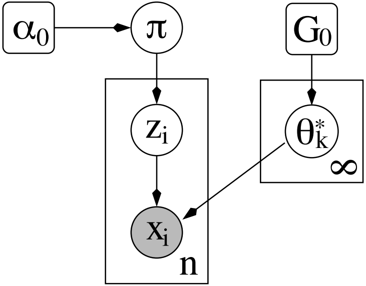
# Dirichlet Process Mixture Models

- Since $G$ is of the form
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$
we have $\theta_i = \theta_k^*$ with probability $\pi_k$.
- Let $k_i$ take on value $k$ with probability $\pi_k$. We can equivalently define $\theta_i = \theta_{k_i}^*$.
- An equivalent model is:

$$\begin{aligned}
x_i &\sim F(\theta_{k_i}^*) && \text{for } i = 1, \ldots, n \\
p(k_i = k) &= \pi_k && \text{for } k = 1, 2, \ldots \\
\pi_k &= \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \\
\beta_k &\sim \text{Beta}(1, \alpha) \\
\theta_k^* &\sim H
\end{aligned}$$

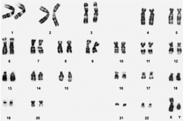
# Dirichlet Process Mixture Models

- So the DP mixture model is a mixture model with an infinite number of clusters.
- But only finitely clusters ever used.
- The DP mixture model can be used for clustering purposes.
  - The number of clusters is not known a priori.
  - Inference in model returns a posterior distribution over number of clusters used to represent data.
  - An alternative to model selection/averaging over finite mixture models.



# Haplotype Inference

- A bioinformatics problem relevant to the study of the evolutionary history of human populations.
- Consider a sequence of $M$ markers on a pair of chromosomes.
- Each marker marks the site where there is an observed variation in the DNA in across the human population.
- A sequence of marker states is called a haplotype.
- A genotype is a sequence of unordered pairs of marker states.

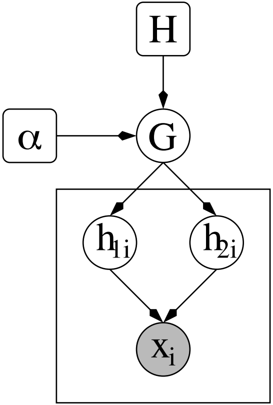


| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 1 | 1 |

$$\{0,0\}\ \{1,1\}\ \{0,1\}\ \{1,1\}\ \{0,1\}\ \{0,1\}\ \{1,1\}$$

# Haplotype Inference

- Biological assays allow us to read the genotype of an individual, not the two haplotypes.
- Problem: from the genotypes of a large number of individuals, can we reconstruct the haplotypes accurately?
- Observation: only a very small number of haplotypes are observed in human populations.
- Model the process as a mixture model.
- Because the actual number of haplotypes in the observed population is not known, we use a DP mixture model.



# Nonparametric Relaxation

- If $G \sim \mathrm{DP}(\alpha, H)$, then $G \to H$ as $\alpha \to \infty$, in the sense that for any function $f$,

$$\int f(\theta)G(\theta)d\theta \to \int f(\theta)H(\theta)d\theta$$

- We can use $G$ as a nonparametric relaxation of $H$.
- Example: generalized linear models.
    - Observed data $\{x_1, y_1, \ldots, x_n, y_n\}$ where , modelled as:

    $$x_i \sim H(f^{-1}(\lambda^\top y_i))$$

    where $H(\eta)$ is an exponential family distribution with parameter $\eta$ and $f$ is the link function.
    - If we do not believe that $H(f^{-1}(\lambda^\top y))$ is the true model, then we can relax our strong parametric assumption as:

    $$\begin{aligned} G(y_i) &\sim \mathrm{DP}(\alpha(w^\top y_i), H(f^{-1}(\lambda^\top y_i))) \\ x_i &\sim G(y_i) \end{aligned}$$

# Generalizations

- Pitman-Yor processes.
- General stick-breaking processes.
- Normalized inversed-Gaussian processes.

# Pitman-Yor Processes

- Pitman-Yor Processes are also known as Two-parameter Poisson-Dirichlet Processes.
- Chinese restaurant representation:

$$P(n^{\text{th}} \text{ customer sit at table } k, 1 \leq k \leq K) = \frac{n_k - d}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \frac{\alpha + dK}{i-1+\alpha}$$

  where $0 \leq d < 1$ and $\alpha > -d$.
- When $d = 0$ the Pitman-Yor process reduces to the DP.
- When $\alpha = 0$ the Pitman-Yor process reduces to a stable process.
- When $\alpha = 0$ and $d = \frac{1}{2}$ the stable process is a normalized inverse-gamma process.
- There is a stick-breaking construction for Pitman-Yor processes (later), but no known analytic expressions for its finite dimensional marginals, except for $d = 0$ and $d = \frac{1}{2}$.

# Pitman-Yor Processes

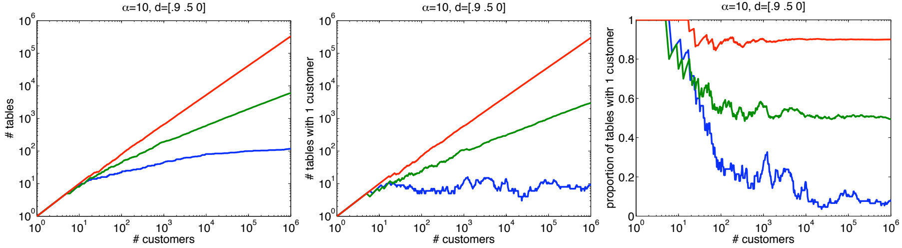
- Pitman-Yor Processes are also known as Two-parameter Poisson-Dirichlet Processes.
- Chinese restaurant representation:

$$P(n^{\text{th}} \text{ customer sit at table } k, 1 \leq k \leq K) = \frac{n_k - d}{n-1+\alpha}$$
$$P(n^{\text{th}} \text{ customer sit at new table}) = \frac{\alpha + dK}{i-1+\alpha}$$

  where $0 \leq d < 1$ and $\alpha > -d$.
- When $d = 0$ the Pitman-Yor process reduces to the DP.
- When $\alpha = 0$ the Pitman-Yor process reduces to a stable process.
- When $\alpha = 0$ and $d = \frac{1}{2}$ the stable process is a normalized inverse-gamma process.
- There is a stick-breaking construction for Pitman-Yor processes (later), but no known analytic expressions for its finite dimensional marginals, except for $d = 0$ and $d = \frac{1}{2}$.

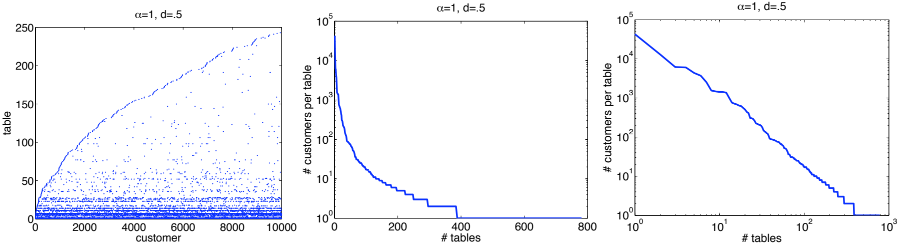
# Pitman-Yor Processes

- 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.

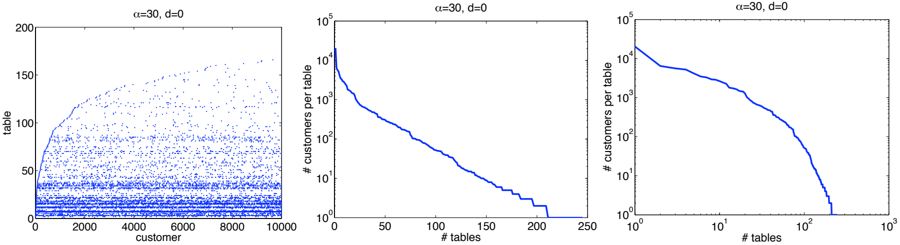


# Pitman-Yor Processes

## Draw from a Pitman-Yor process




## Draw from a Dirichlet process



# General Stick-breaking Processes

- We can relax the priors on $\beta_k$ in the stick-breaking construction:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \qquad \pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l)$$

$$\theta_k^* \sim H \qquad \beta_k \sim \text{Beta}(a_k, b_k)$$

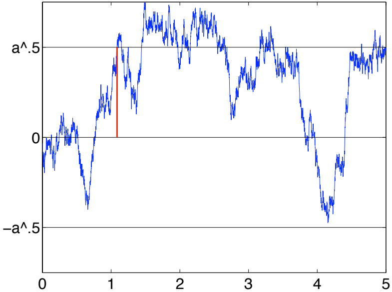
- We get the DP if $a_k = 1$, $b_k = \alpha$.
- We get the Pitman-Yor process if $a_k = 1 - d$, $b_k = \alpha + kd$.
- To ensure that $\sum_{k=1}^{\infty} \pi_k = 1$, we need $\beta_k$ to not go to 0 too quickly:

$$\sum_{k=1}^{\infty} \pi_k = 1 \quad \text{almost surely iff} \quad \sum_{k=1}^{\infty} \log(1 + a_k/b_k) = \infty$$

# Normalized Inverse-Gaussian Processes

- The inverse-Gaussian distribution with parameter $\alpha$ has density:

$$p(\nu) = \frac{\alpha}{\sqrt{2\pi}} \nu^{-3/2} \exp\left(-\frac{1}{2}\left(\frac{\alpha^2}{\nu} + \nu\right) + \alpha\right) \qquad \nu \geq 0$$




- Additive property of inverse-Gaussian variables: if $\nu_1 \sim \mathrm{IG}(\alpha_1)$ and $\nu_2 \sim \mathrm{IG}(\alpha_2)$ then $\nu_1 + \nu_2 \sim \mathrm{IG}(\alpha_1 + \alpha_2)$.

# Normalized Inverse-Gaussian Processes

- The normalized inverse-Gaussian is a distribution over the $m$-simplex obtained by normalizing $m$ inverse-Gaussian variables, and has density:

$$p(w_1,\ldots,w_m|\alpha_1,\ldots,\alpha_m)$$
$$=\frac{e^{\sum_{i=1}^m \alpha_i+\log\alpha_i}}{2^{m/2-1}\pi^{m/2}}K_{-m/2}\left(\sqrt{\sum_{i=1}^m \frac{\alpha_i^2}{w_i}}\right)\left(\sum_{i=1}^m \frac{\alpha_i^2}{w_i}\right)^{-m/4}\prod_{i=1}^m w_i^{-3/2}$$

- Agglomerative property: if $\{J_1,\ldots,J_{m'}\}$ is a partition of $\{1,\ldots,m\}$,

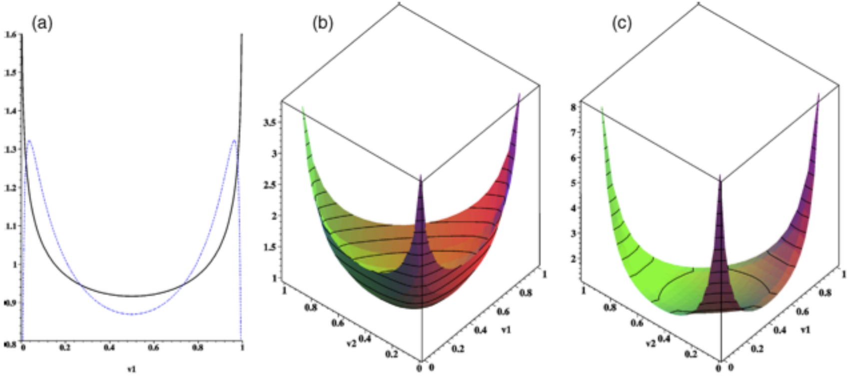
$$\left(\sum_{i\in J_1} w_i,\ldots,\sum_{i\in J_{m'}} w_i\right)\sim \text{NIG}\left(\sum_{i\in J_1}\alpha_i,\ldots,\sum_{i\in J_{m'}}\alpha_i\right)$$

- We can now define a normalized inverse-Gaussian process (NIGP) analogously to a Dirichlet process. $G\sim\text{NIGP}(\alpha,H)$ if for all partitions $(A_1,\ldots,A_m)$ of $\mathbb{X}$:

$$(G(A_1),\ldots,G(A_m))\sim\text{NIG}(\alpha H(A_1),\ldots,\alpha H(A_m))$$

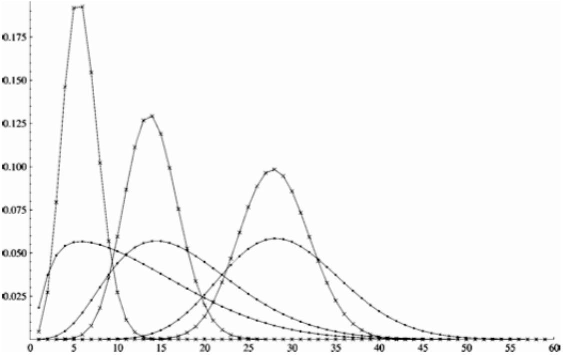
# Normalized Inverse-Gaussian Processes

- There is a tractable Pòlya urn scheme corresponding to the NIGP.
- The DP, the Pitman-Yor with $d = \frac{1}{2}$, and the NIG process are the only known normalized random measure with analytic finite dimensional marginals.
- The NIGP have wider support around its modes than does the DP:



# Normalized Inverse-Gaussian Processes

- There is a tractable Pòlya urn scheme corresponding to the NIGP.
- The DP, the Pitman-Yor with $d = \frac{1}{2}$, and the NIG process are the only known normalized random measure with analytic finite dimensional marginals.
- The NIGP have wider support around its modes than does the DP:

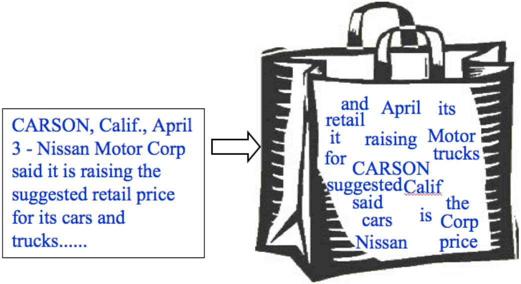


- Grouped Clustering Problems.
- Hierarchical Dirichlet Processes.
- Representations of Hierarchical Dirichlet Processes.
- Applications in Grouped Clustering.
- Extensions and Related Models.

# Grouped Clustering Problems

## Example: document topic modelling

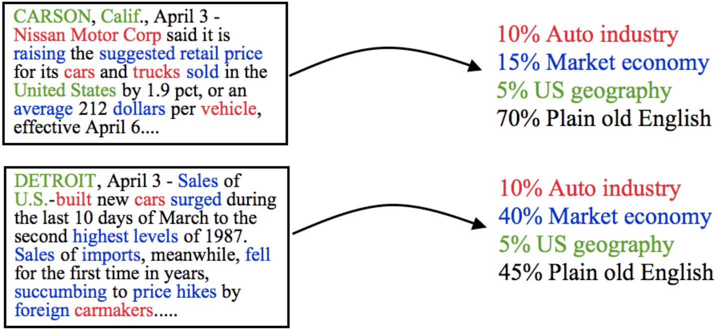
- Information retrieval: finding useful information from large collections of documents.
- Example: Google, CiteSeer, Amazon...
- Model documents as “bags of words”.



# Grouped Clustering Problems

## Example: document topic modelling

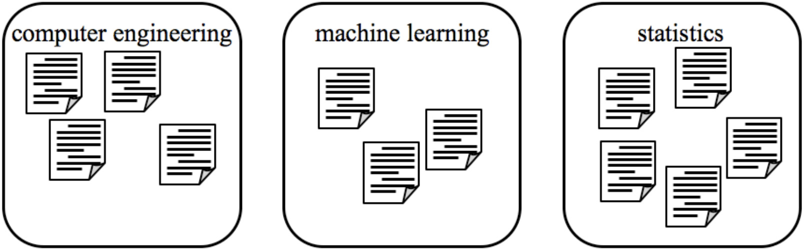
- We model documents as coming from an underlying set of topics.
  - Summarize documents.
  - Document/query comparisons.
  - Do not know the number of topics a priori—use DP mixtures somehow.
  - But: topics have to be shared across documents...



# Grouped Clustering Problems

## Example: document topic modelling

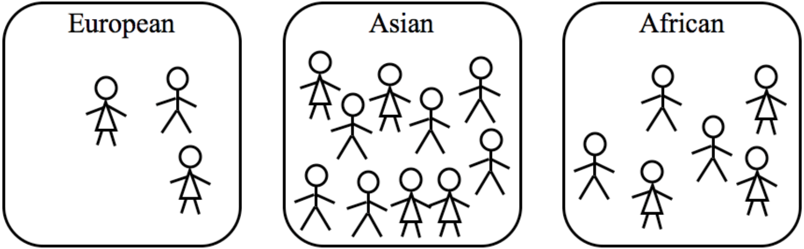
- Share topics across documents in a collection, and across different collections.
- More sharing within collections than across.
- Use DP mixture models as we do not know the number of topics a priori.



# Grouped Clustering Problems

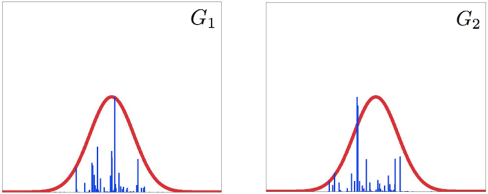
## Example: haplotype inference

- Individuals inherit both ancient haplotypes dispersed across multiple populations, as well as more recent population-specific haplotypes.
- Sharing of haplotypes among individuals in a population, and across different populations.
- More sharing within populations than across.
- Use DP mixture models as we do not know the number of haplotypes.

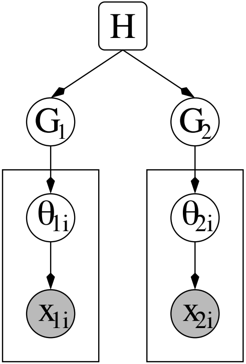


# Hierarchical Dirichlet Processes

- Use a DP mixture for each group.

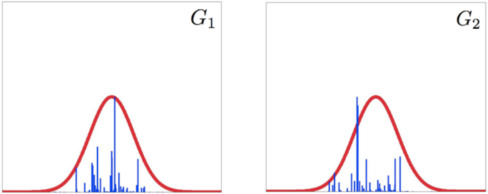



- Unfortunately there is no sharing of clusters across different groups because *H* is smooth.
- Solution: make the base distribution *H* discrete.
- Put a DP prior on the common base distribution.

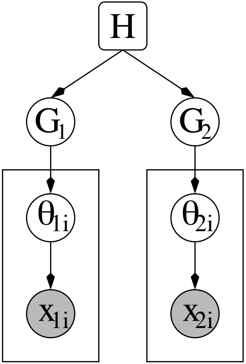


# Hierarchical Dirichlet Processes

- Use a DP mixture for each group.

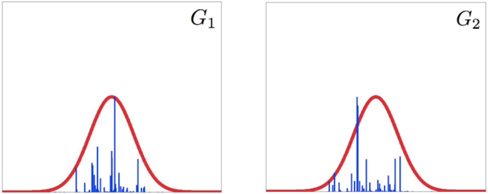



- Unfortunately there is no sharing of clusters across different groups because *H* is smooth.
- Solution: make the base distribution *H* discrete.
- Put a DP prior on the common base distribution.

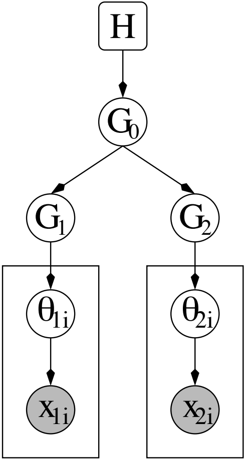


# Hierarchical Dirichlet Processes

- 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.

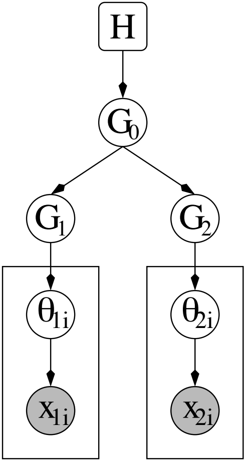


# Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

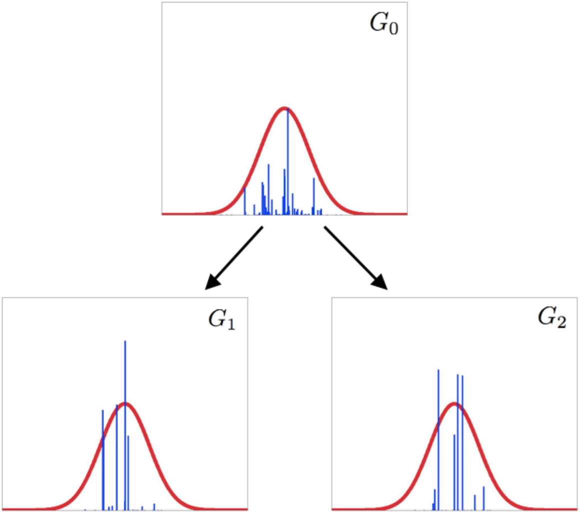
$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$$

- Extension to deeper hierarchies is straightforward.



# Hierarchical Dirichlet Processes

- Making $G_0$ discrete forces shared cluster between $G_1$ and $G_2$



- We shall assume the following HDP hierarchy:

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

- The stick-breaking construction for the HDP is:

$$G_0 = \sum_{k=1}^{\infty} \pi_{0k} \delta_{\theta_k^*} \qquad \theta_k^* \sim H$$
$$\pi_{0k} = \beta_{0k} \prod_{l=1}^{k-1} (1 - \beta_{0l}) \quad \beta_{0k} \sim \mathrm{Beta}\left(1, \gamma\right)$$
$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^*}$$
$$\pi_{jk} = \beta_{jk} \prod_{l=1}^{k-1} (1 - \beta_{jl}) \qquad \beta_{jk} \sim \mathrm{Beta}\left(\alpha\beta_{0k}, \alpha(1 - \sum_{l=1}^{k} \beta_{0l})\right)$$

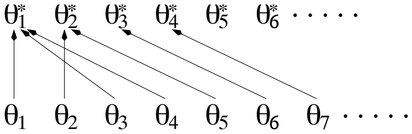
- We shall assume the following HDP hierarchy:

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_j | G_0 &\sim \mathrm{DP}(\alpha, G_0) \quad \text{for } j = 1, \ldots, J \end{aligned}$$

- The stick-breaking construction for the HDP is:

$$\begin{aligned} G_0 &= \textstyle\sum_{k=1}^\infty \pi_{0k} \delta_{\theta_k^*} & \theta_k^* &\sim H \\ \pi_{0k} &= \beta_{0k} \textstyle\prod_{l=1}^{k-1} (1 - \beta_{0l}) & \beta_{0k} &\sim \mathrm{Beta}\left(1, \gamma\right) \\ G_j &= \textstyle\sum_{k=1}^\infty \pi_{jk} \delta_{\theta_k^*} & & \\ \pi_{jk} &= \beta_{jk} \textstyle\prod_{l=1}^{k-1} (1 - \beta_{jl}) & \beta_{jk} &\sim \mathrm{Beta}\left(\alpha\beta_{0k}, \alpha(1 - \textstyle\sum_{l=1}^k \beta_{0l})\right) \end{aligned}$$

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

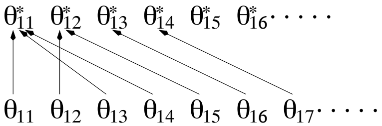



where the arrows denote to which $\theta_k^*$ each $\theta_i$ was assigned and

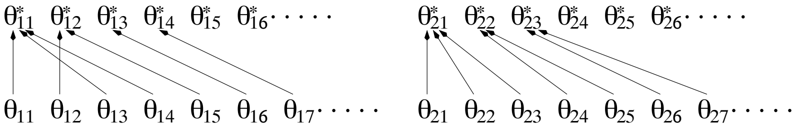
$$\theta_1, \theta_2, \ldots \sim G \text{ i.i.d.}$$
$$\theta_1^*, \theta_2^*, \ldots \sim H \text{ i.i.d.}$$

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

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



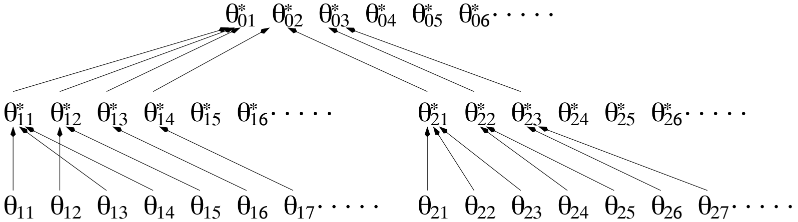
- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The hierarchical Pòlya urn scheme to generate draws from $G_1, G_2$:



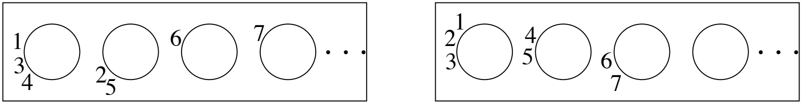
# Representations of Hierarchical Dirichlet Processes

## Hierarchical Pòlya urn scheme

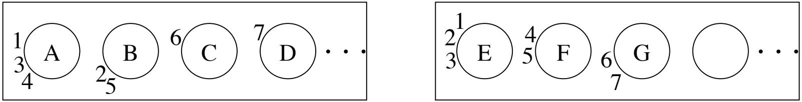
- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The hierarchical Pòlya urn scheme to generate draws from $G_1, G_2$:



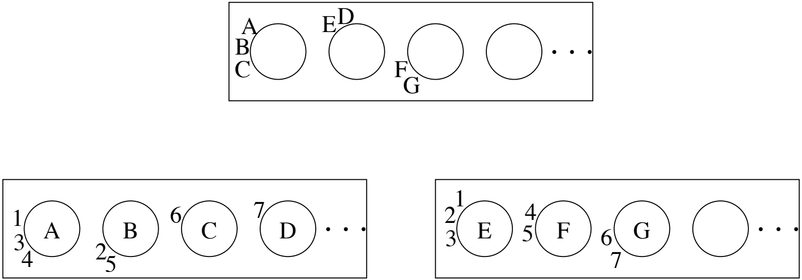
- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The Chinese restaurant franchise describes the clustering of data items in the hierarchy:



- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The Chinese restaurant franchise describes the clustering of data items in the hierarchy:

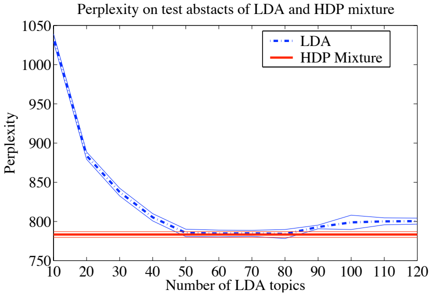


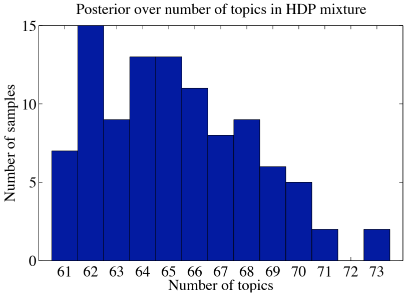
- Let $G_0 \sim \text{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \text{DP}(\alpha, G_0)$.
- The Chinese restaurant franchise describes the clustering of data items in the hierarchy:



# Application: Document Topic Modelling

- Compared against latent Dirichlet allocation, a parametric version of the HDP mixture for topic modelling.



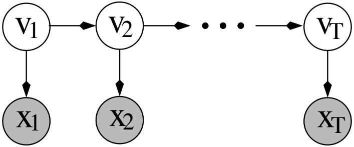
# Application: Document Topic Modelling

- Topics learned on the NIPS corpus.
  - Documents are separated into 9 subsections.
  - Model this with a 3 layer HDP mixture model.
- Shown are the topics shared between Vision Sciences and each other subsections.

| Cognitive Science | | Neuroscience | | Algorithms & Architecture | | Signal Processing | |
|---|---|---|---|---|---|---|---|
| task | examples | cells | visual | algorithms | distance | visual | signals |
| representation | concept | cell | cells | test | tangent | images | separation |
| pattern | similarity | activity | cortical | approach | image | video | signal |
| processing | Bayesian | response | orientation | methods | images | language | sources |
| trained | hypotheses | neuron | receptive | based | transformation | image | source |
| representations | generalization | visual | contrast | point | transformations | pixel | matrix |
| three | numbers | patterns | spatial | problems | pattern | acoustic | blind |
| process | positive | pattern | cortex | form | vectors | delta | mixing |
| unit | classes | single | stimulus | large | convolution | lowpass | gradient |
| patterns | hypothesis | fig | tuning | paper | simard | flow | eq |

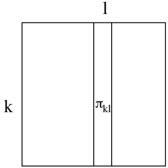
# Infinite Hidden Markov Models

- A hidden Markov model consists of a discrete latent state sequence $v_{1:T}$ and an observation sequence $x_{1:T}$.




- The transition and observation probabilities are:

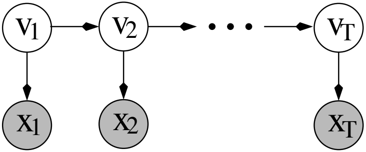
$$
\begin{aligned}
P(v_t = k | v_{t-1} = l) &= \pi_{kl} \\
p(x_t | v_t = k) &= f(x_t | \theta^*_k)
\end{aligned}
$$

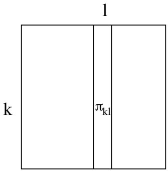



- In finite HMMs, we can place priors on the parameters easily:

$$
\begin{aligned}
(\pi_{1l}, \ldots, \pi_{Kl}) &\sim \mathrm{Dirichlet}(\alpha/K, \ldots, /\mathit{alpha}/K) \\
\theta^*_k &\sim H
\end{aligned}
$$

# Infinite Hidden Markov Models



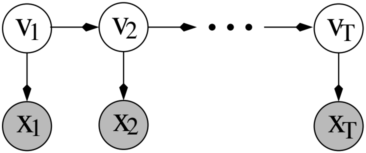

$$P(v_t = k | v_{t-1} = l) = \pi_{kl}$$
$$(\pi_{1l}, \ldots, \pi_{Kl}) \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

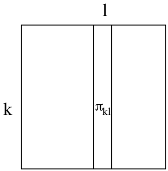
- Can we take $K \to \infty$?
- Probability of transitioning to a previously unseen state always 1...
- Say $v_{t_1} = l$ and this is first time we are in state $l$. Then

$$P(v_t = k | v_{t-1} = l) = 1/K \to 0$$

for all $k$.

# Infinite Hidden Markov Models




$$P(v_t = k | v_{t-1} = l) = \pi_{kl}$$
$$(\pi_{1l}, \dots, \pi_{Kl}) \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$$

- Can we take $K \to \infty$? Not just like that!
- Probability of transitioning to a previously unseen state always 1...
- Say $v_{t_1} = l$ and this is first time we are in state $l$. Then

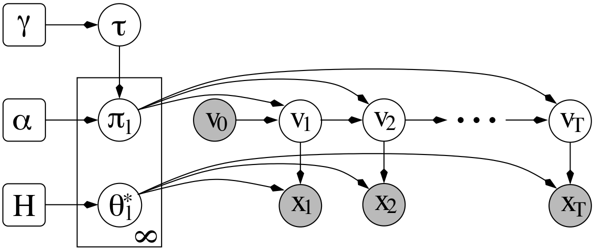
$$P(v_t = k | v_{t-1} = l) = 1/K \to 0$$

for all $k$.

# Infinite Hidden Markov Models

- Previous issue is that there is no sharing of possible next states across different current states.
- Implement sharing of next states using a HDP:

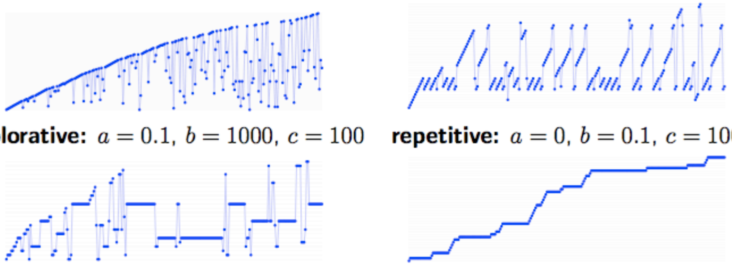
$$(\tau_1, \tau_2, \ldots) \sim \mathrm{GEM}(\gamma)$$
$$(\pi_{1l}, \pi_{2l}, \ldots)|\tau \sim \mathrm{DP}(\alpha, \tau)$$



# Infinite Hidden Markov Models

- A variety of trajectory characteristics can be modelled using different parameter regimes.

(modified to treat self-transitions specially)

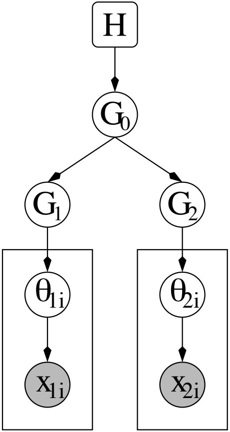


**explorative:** $a = 0.1,\ b = 1000,\ c = 100$ **repetitive:** $a = 0,\ b = 0.1,\ c = 100$

**self-transitioning:** $a = 2,\ b = 2,\ c = 20$ **ramping:** $a = 1,\ b = 1,\ c = 10000$

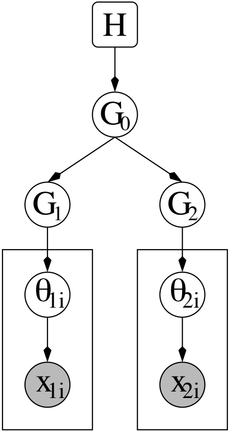
# Nested Dirichlet Processes

- The HDP assumes that data group structure is observed.
- The group structure may not be known in practice, even if there is prior belief in some group structure.
- Even if known, we may still believe that some groups are more similar to each other than to other groups.
- We can cluster groups using a second level of mixture models.
- Using a second DP mixture to model this leads to the nested Dirichlet process.



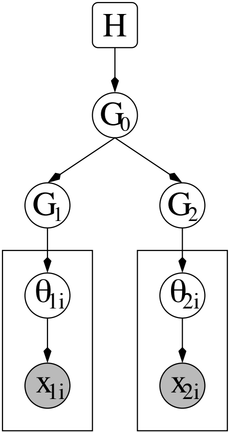
# Nested Dirichlet Processes

- The HDP assumes that data group structure is observed.
- The group structure may not be known in practice, even if there is prior belief in some group structure.
- Even if known, we may still believe that some groups are more similar to each other than to other groups.
- We can cluster groups using a second level of mixture models.
- Using a second DP mixture to model this leads to the nested Dirichlet process.



# Nested Dirichlet Processes

- The HDP assumes that data group structure is observed.
- The group structure may not be known in practice, even if there is prior belief in some group structure.
- Even if known, we may still believe that some groups are more similar to each other than to other groups.
- We can cluster groups using a second level of mixture models.
- Using a second DP mixture to model this leads to the nested Dirichlet process.



# Nested Dirichlet Processes

- Start with:

$$x_{ji} \sim F(\theta_{ji}) \qquad \theta_{ji} \sim G_j$$

- Cluster groups. Each group $j$ belongs to cluster $k_j$:

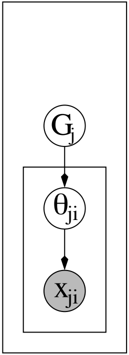
$$k_j \sim \pi \qquad \pi \sim \mathrm{GEM}(\alpha)$$

- Group $j$ inherits the DP from cluster $k_j$:

$$G_j = G^*_{k_j}$$

- Place a HDP prior on $\{G^*_k\}$:

$$G^*_k \sim \mathrm{DP}(\beta, G^*_0) \qquad G^*_0 \sim \mathrm{DP}(\gamma, H)$$



# Nested Dirichlet Processes

- Start with:

$$x_{ji} \sim F(\theta_{ji}) \qquad \theta_{ji} \sim G_j$$

- Cluster groups. Each group $j$ belongs to cluster $k_j$:

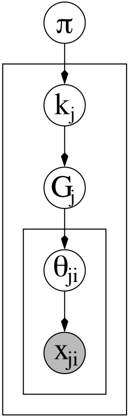
$$k_j \sim \pi \qquad \pi \sim \mathrm{GEM}(\alpha)$$

- Group $j$ inherits the DP from cluster $k_j$:

$$G_j = G^*_{k_j}$$

- Place a HDP prior on $\{G^*_k\}$:

$$G^*_k \sim \mathrm{DP}(\beta, G^*_0) \qquad G^*_0 \sim \mathrm{DP}(\gamma, H)$$



# Nested Dirichlet Processes

- Start with:

$$x_{ji} \sim F(\theta_{ji}) \qquad \theta_{ji} \sim G_j$$

- Cluster groups. Each group $j$ belongs to cluster $k_j$:

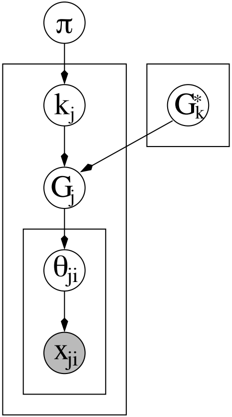
$$k_j \sim \pi \qquad \pi \sim \mathrm{GEM}(\alpha)$$

- Group $j$ inherits the DP from cluster $k_j$:

$$G_j = G^*_{k_j}$$

- Place a HDP prior on $\{G^*_k\}$:

$$G^*_k \sim \mathrm{DP}(\beta, G^*_0) \qquad G^*_0 \sim \mathrm{DP}(\gamma, H)$$



# Nested Dirichlet Processes

- Start with:

$$x_{ji} \sim F(\theta_{ji}) \qquad \theta_{ji} \sim G_j$$

- Cluster groups. Each group $j$ belongs to cluster $k_j$:

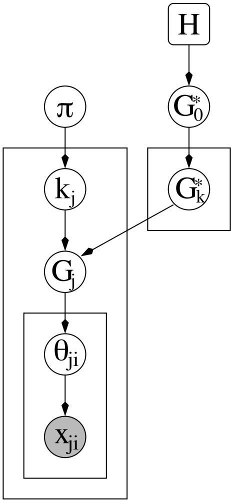
$$k_j \sim \pi \qquad \pi \sim \text{GEM}(\alpha)$$

- Group $j$ inherits the DP from cluster $k_j$:

$$G_j = G^*_{k_j}$$

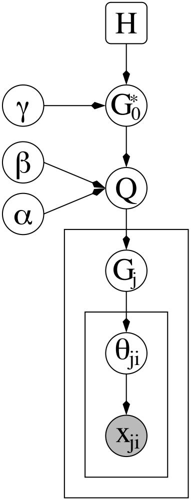
- Place a HDP prior on $\{G^*_k\}$:

$$G^*_k \sim \text{DP}(\beta, G^*_0) \qquad G^*_0 \sim \text{DP}(\gamma, H)$$



# Nested Dirichlet Processes

$$
\begin{aligned}
G_0^* &\sim \mathrm{DP}(\gamma, H) \\
Q &\sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, G_0^*)) \\
G_j &\sim Q \\
\theta_{ji} &\sim G_j \\
x_{ji} &\sim F(\theta_{ji})
\end{aligned}
$$



# Dependent Dirichlet Processes

- The HDP induces a straightforward dependency among groups.
- What if the data is smoothly varying across some spatial or temporal domain?
  - Topic modelling: topic popularity and composition can both change slowly as time passes.
  - Haplotype inference: haplotype occurrence can change smoothly as function of geography.
- a dependent Dirichlet process is a stochastic process $\{G_t\}$ indexed by $t$ (space or time), such that each $G_t \sim \mathrm{DP}(\alpha, H)$ and if $t, t'$ are neighbouring points, $G_t$ and $G_{t'}$ should be "similar" to each other.
- Simple example:

$$\pi \sim \mathrm{GEM}(\alpha) \qquad\qquad (\theta^*_{tk}) \sim \mathrm{GP}(\mu, \Sigma) \quad \text{for each } k$$

$$G_t = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_{tk}}$$

# Summary

- Dirichlet processes and hierarchical Dirichlet processes.
- Described different representations:
  distribution over distributions; Chinese restaurant process; Pòlya urn scheme; Stick-breaking construction.
- Described generalizations and extensions:
  Pitman-Yor processes; General stick-breaking processes; Normalized inverse-Gaussian processes; nested Dirichlet processes; Dependent Dirichlet processes.
- Described some applications:
  Document mixture models; Topic modelling; Haplotype inference; Infinite hidden Markov models.
- I have not described inference schemes.
- A rich and growing area, and much to be discovered and tried.