Bayesian Nonparametrics: Dirichlet Process

Yee Whye Teh
Gatsby Computational Neuroscience Unit, UCL

http://www.gatsby.ucl.ac.uk/~ywteh/teaching/npbayes2012

Dirichlet Process

- Cornerstone of modern Bayesian nonparametrics.
- Rediscovered many times as the infinite limit of finite mixture models.
- Formally defined by [Ferguson 1973] as a distribution over measures.
- Can be derived in different ways, and as special cases of different processes.

- Random partition view:
    - Chinese restaurant process, Blackwell-mcQueen urn scheme
- Random measure view:
    - stick-breaking construction, Poisson-Dirichlet, gamma process

The Infinite Limit of Finite Mixture Models

Finite Mixture Models

- Model for data from heterogeneous unknown sources.
- Each cluster (source) modelled using a parametric model (e.g. Gaussian).
- Data item i:

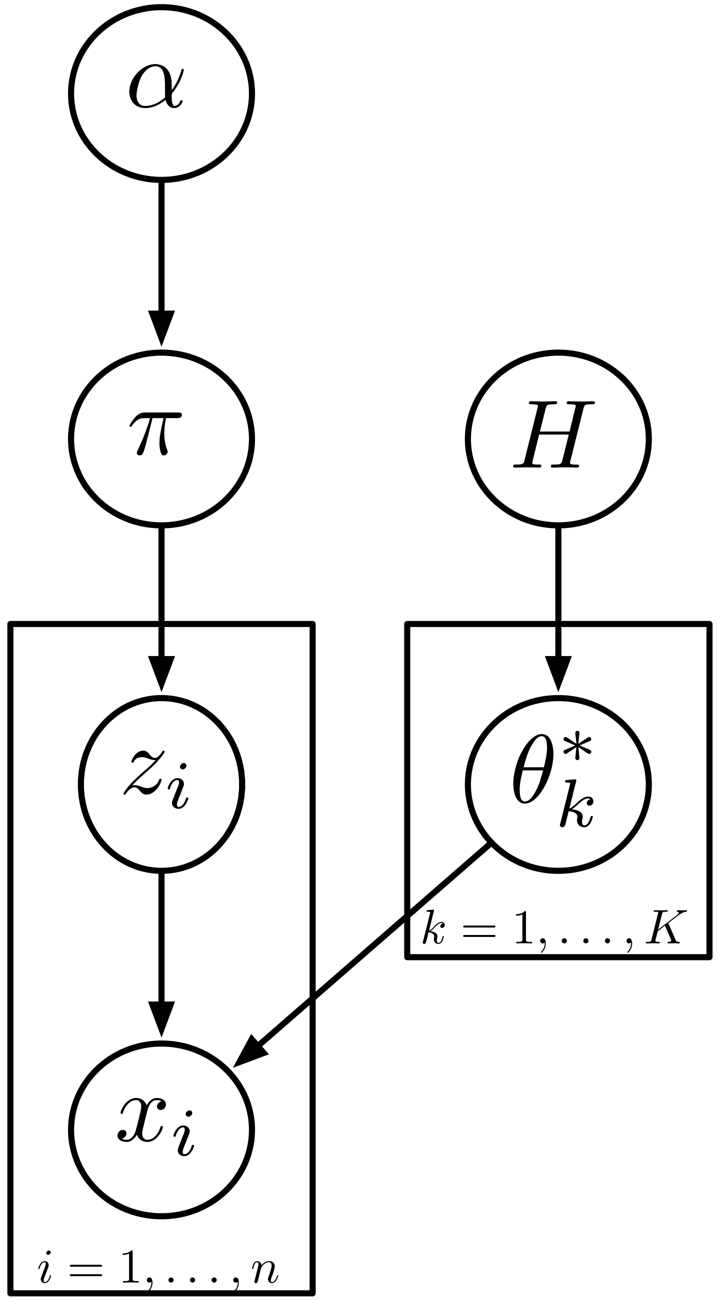
$$z_i | \pi \sim \mathrm{Discrete}(\pi)$$
$$x_i | z_i, \theta_k^* \sim F(\theta_{z_i}^*)$$

- **Mixing proportions**:

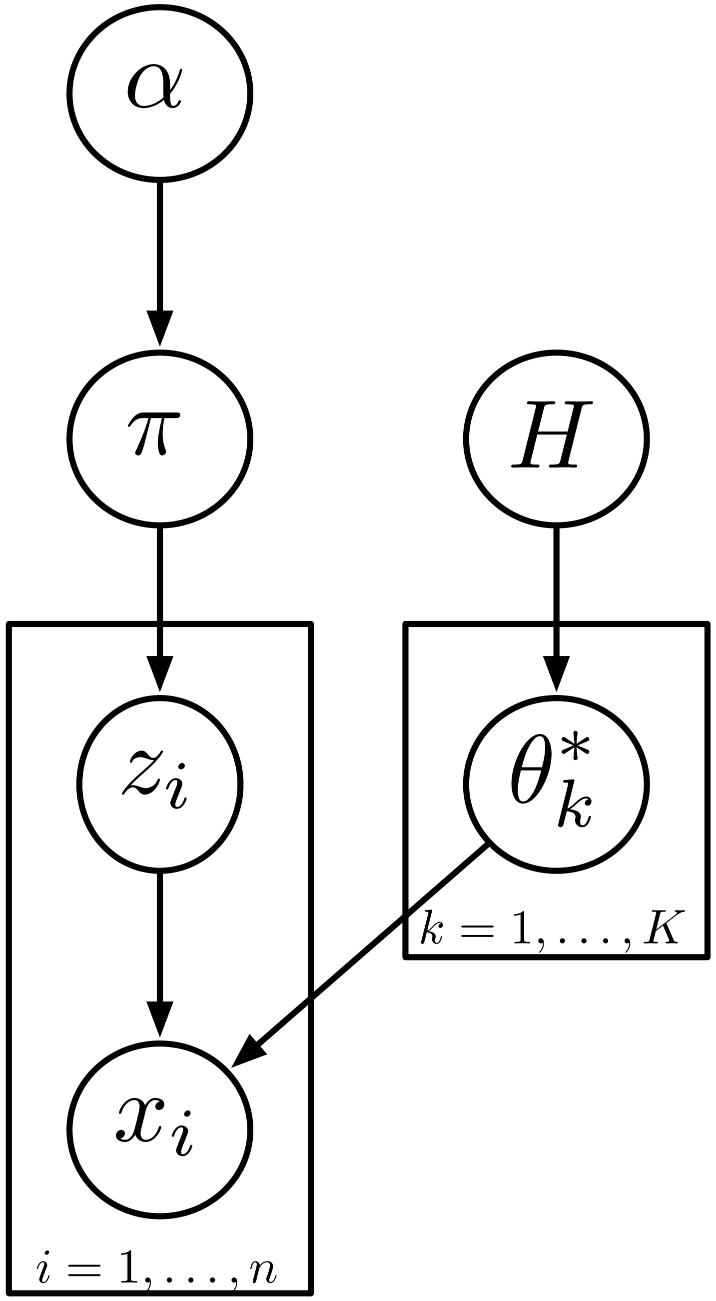
$$\pi = (\pi_1, \ldots, \pi_K) | \alpha \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

- Cluster k:

$$\theta_k^* | H \sim H$$



Finite Mixture Models



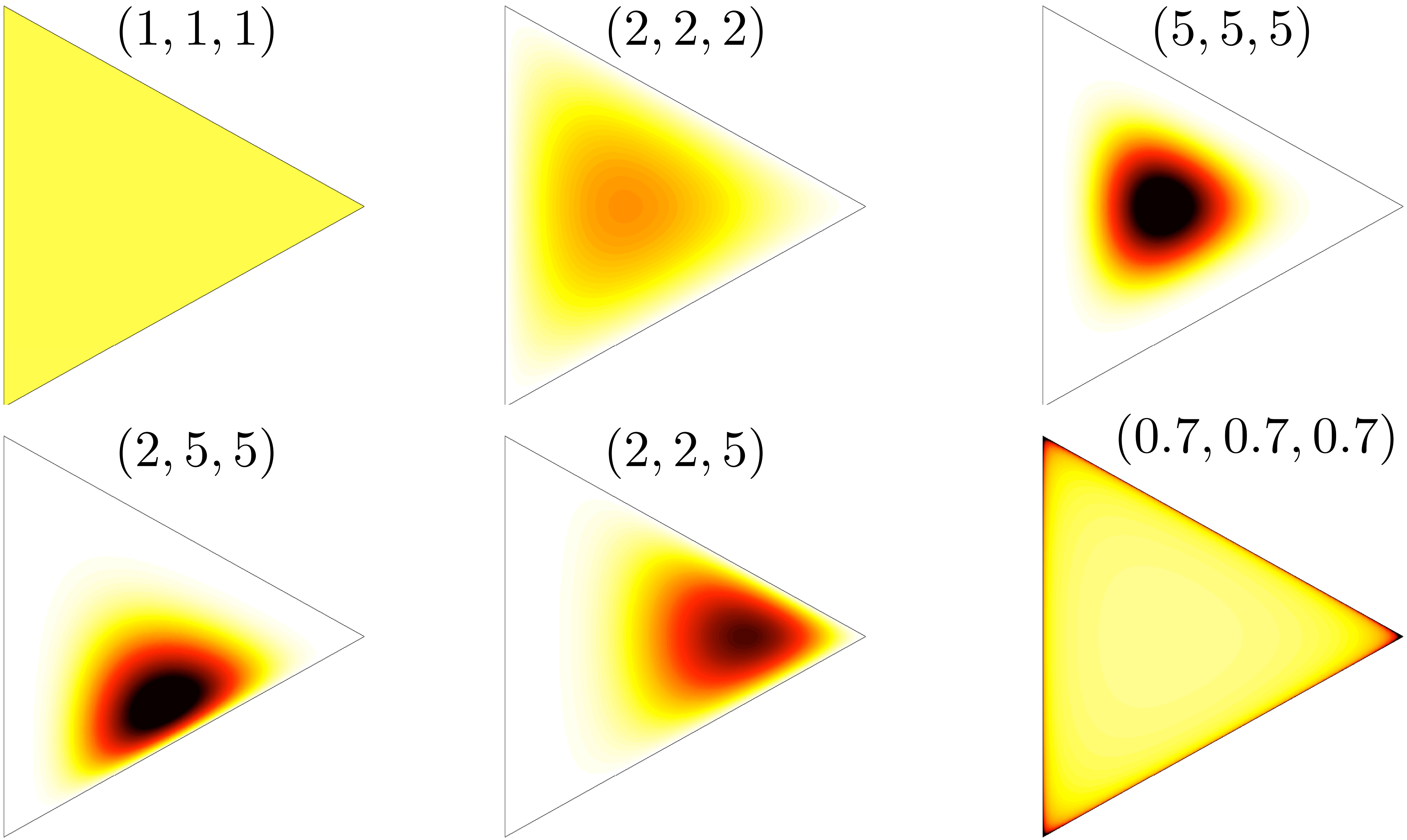
- Dirichlet distribution on the K-dimensional probability simplex $\{ \pi \mid \Sigma_k \pi_k = 1 \}$:

$$P(\pi|\alpha) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha/K)} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}$$

with $\Gamma(a) = \int_0^\infty x^{a-1} e^x dx.$

- Standard distribution on probability vectors, due to **conjugacy** with multinomial.

Dirichlet Distribution



$$P(\pi|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Dirichlet-Multinomial Conjugacy

- Joint distribution over z_i and π:

$$P(\pi|\alpha) \times \prod_{i=1}^{n} P(z_i|\pi) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha/K)} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} \times \prod_{k=1}^{K} \pi_k^{n_k}$$

where $n_c = \#\{\, z_i = c \,\}$.

- Posterior distribution:

$$P(\pi|\mathbf{z},\alpha) = \frac{\Gamma(n+\alpha)}{\prod_{k=1}^{K} \Gamma(n_k+\alpha/K)} \prod_{k=1}^{K} \pi_k^{n_k+\alpha/K-1}$$

- Marginal distribution:

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

Gibbs Sampling

- All conditional distributions are simple to compute:

$$p(z_i = k|\text{others}) \propto \pi_k f(x_i|\theta^*_k)$$

$$\pi|\text{others} \sim \text{Dirichlet}(\tfrac{\alpha}{K} + n_1, \ldots, \tfrac{\alpha}{K} + n_K)$$

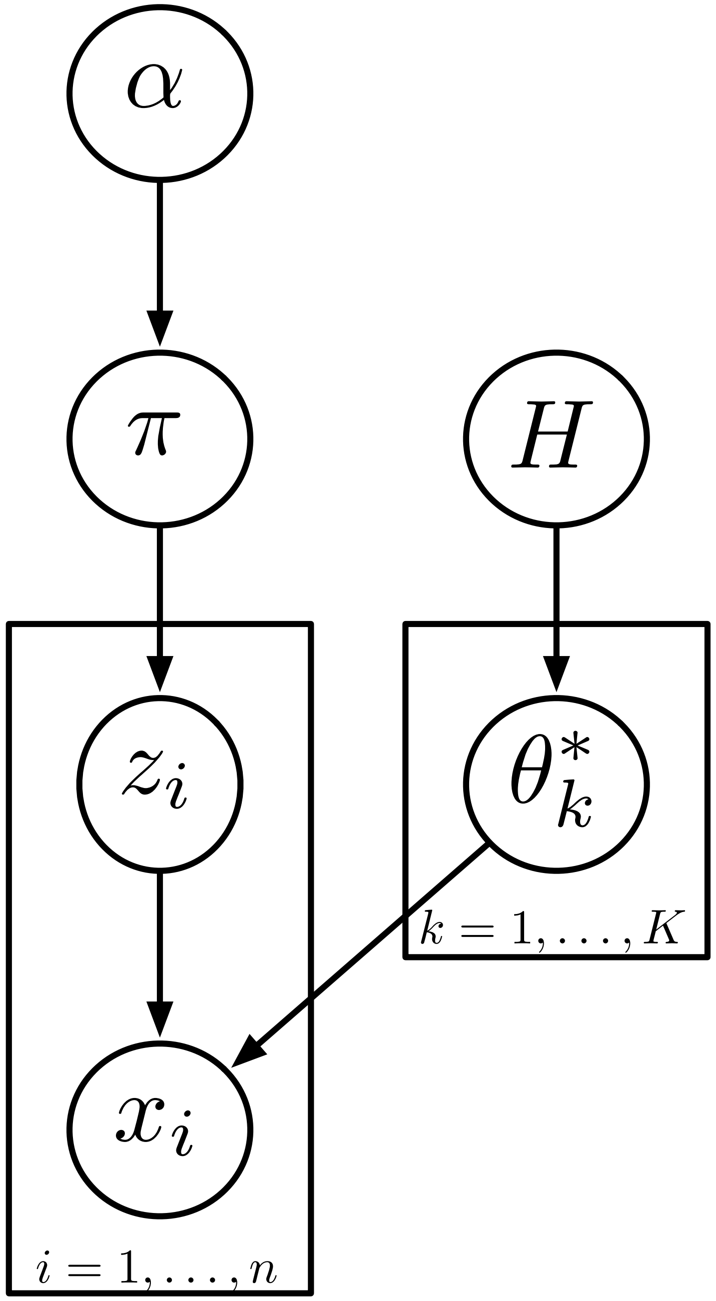
$$p(\theta^*_k = \theta|\text{others}) \propto h(\theta) \prod_{j:z_j=k} f(x_j|\theta)$$

- Not as efficient as collapsed Gibbs sampling, which integrates out π, θ*'s:

$$p(z_i = k|\text{others}) \propto \frac{\frac{\alpha}{K} + n_k^{\neg i}}{\alpha + n - 1} f(x_i|\{x_j : j \neq i, z_j = k\})$$

$$f(x_i|\{x_j : j \neq i, z_j = k\}) \propto \int h(\theta) f(x_i|\theta) \prod_{j \neq i: z_j = k} f(x_j|\theta) d\theta$$

- Conditional distributions can be efficiently computed if *F* is conjugate to *H*.

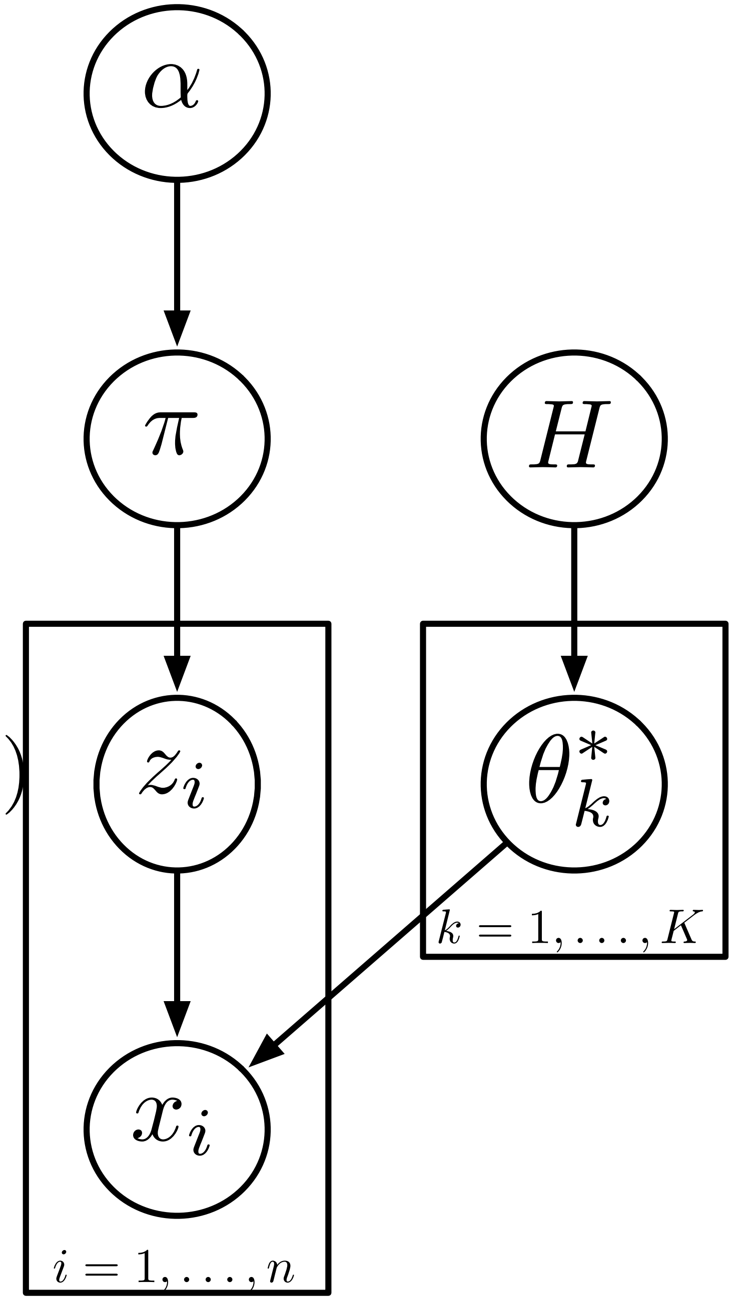


Infinite Limit of Collapsed Gibbs Sampler

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

$$p(z_i = k|\text{others}) = \frac{n_k^{\neg i} + \frac{\alpha}{K}}{n - 1 + \alpha} f(x_i|\{x_j : j \neq i, z_j = k\})$$

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



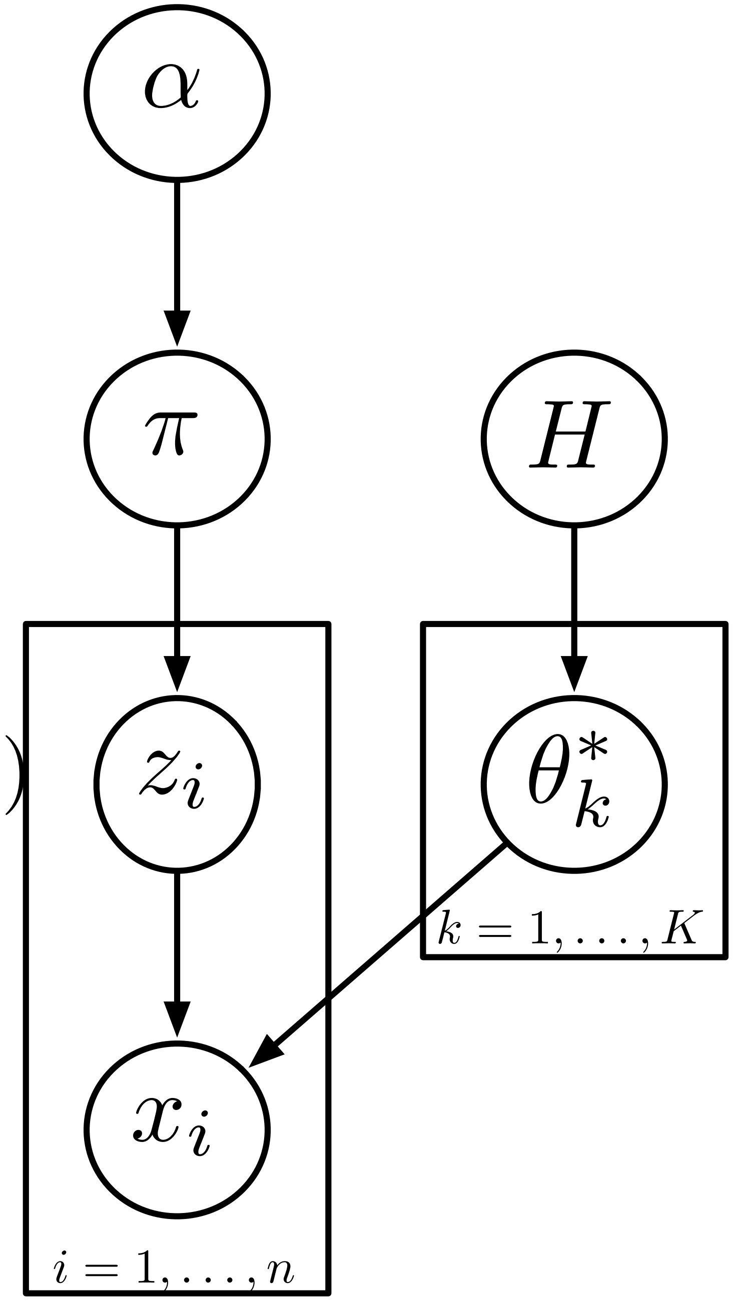
[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Infinite Limit of Collapsed Gibbs Sampler

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

$$p(z_i = k|\text{others}) = \frac{n_k^{\neg i}}{n - 1 + \alpha} f(x_i | \{x_j : j \neq i, z_j = k\})$$

$$p(z_i = k_{\text{empty}}|\text{others}) = \frac{\alpha}{n - 1 + \alpha} f(x_i | \{\})$$



[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Infinite Limit

- The actual infinite limit of the finite mixture model does not make sense:
    - any particular cluster will get a mixing proportion of 0.
- Better ways of making this infinite limit precise:
    - Chinese restaurant process.
    - 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.

Ferguson's Definition of the Dirichlet Process

Ferguson's Definition of Dirichlet Processes

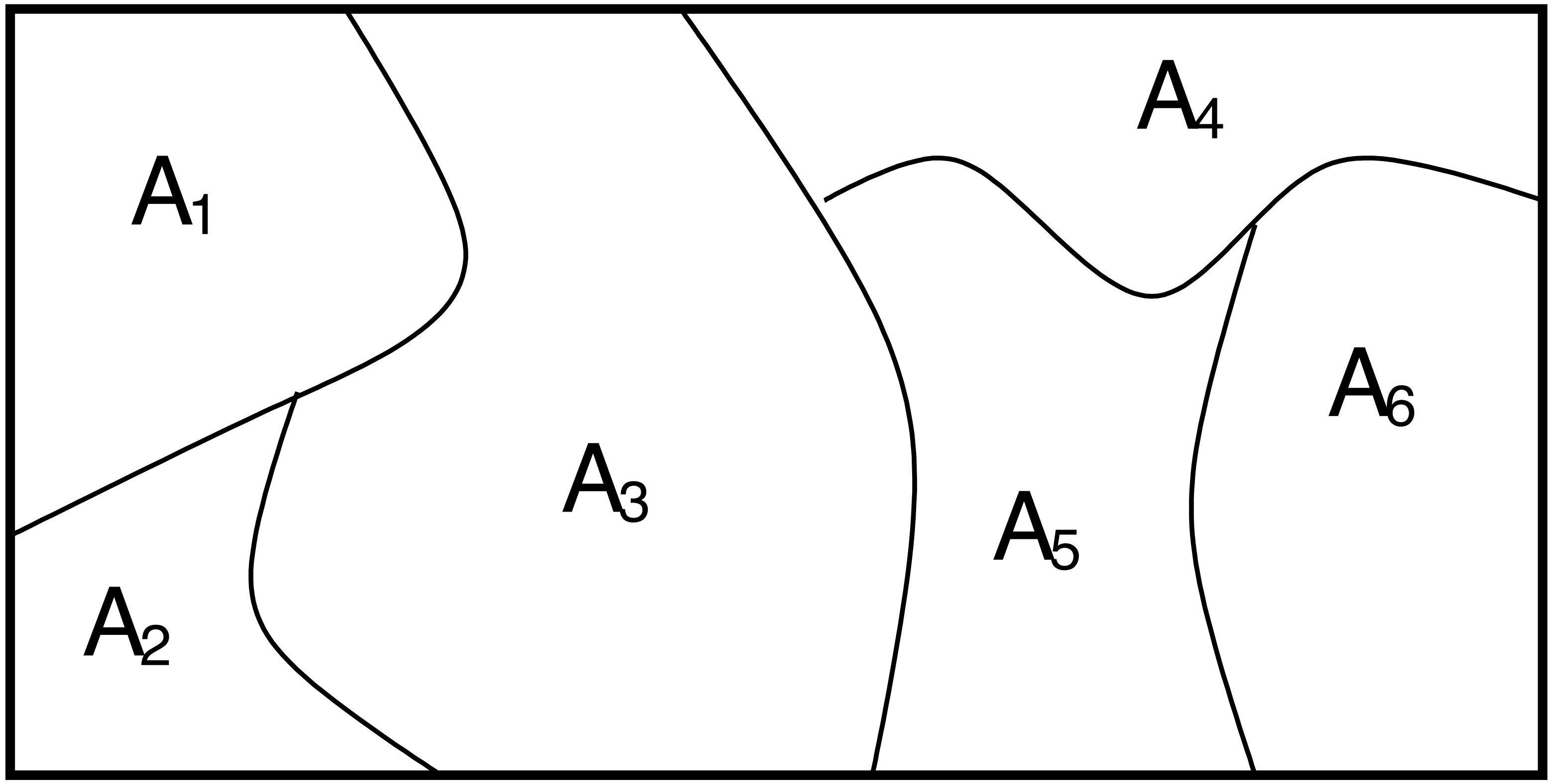
- A **Dirichlet process** (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of measurable sets $A_1,\ldots A_K \in \Sigma$ partitioning Θ, i.e.

$$A_1 \dot\cup \cdots \dot\cup A_K = \Theta$$

we have

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

where α and H are parameters of the DP.



[Ferguson 1973]

Parameters of the Dirichlet Process

- α is called the **strength, mass** or **concentration parameter**.
- H is called the **base distribution**.
- Mean and variance:

$$\mathbb{E}[G(A)] = H(A)$$
$$\mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

  where A is a measurable subset of Θ.
- H is the mean of G, and α is an inverse variance.

Posterior Dirichlet Process

- Suppose

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

- We can define random variables that are G distributed:

$$\theta_i | G \sim G \qquad \text{for } i = 1, \ldots, n$$

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

$$G | \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha + n, \tfrac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n})$$

Pólya Urn Scheme

$$G \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i | G \sim G \qquad \text{for } i = 1, 2, \ldots$$

- Marginalizing out G, we get:

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$$

- This is called the **Pólya**, **Hoppe** or **Blackwell-MacQueen urn scheme**.
  - Start with an urn with α balls of a special colour.
  - Pick a ball randomly from urn:
    - If it is a special colour, make a new ball with colour sampled from H, note the colour, and return both balls to urn.
    - If not, note its colour and return two balls of that colour to urn.

[Blackwell & MacQueen 1973, Hoppe 1984]

Clustering Property

$$G \sim \mathrm{DP}(\alpha, H)$$
$$\theta_i | G \sim G \quad \text{for } i = 1, 2, \ldots$$

- The n variables $\theta_1, \theta_2, \ldots, \theta_n$ can take on $K \leq n$ distinct values.
- Let the distinct values be $\theta_1^*, \ldots, \theta_K^*$. This defines a partition of $\{1, \ldots, n\}$ such that i is in cluster k if and only if $\theta_i = \theta_k^*$.
- The induced distribution over partitions is the **Chinese restaurant process**.

Discreteness of the Dirichlet Process

- Suppose

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

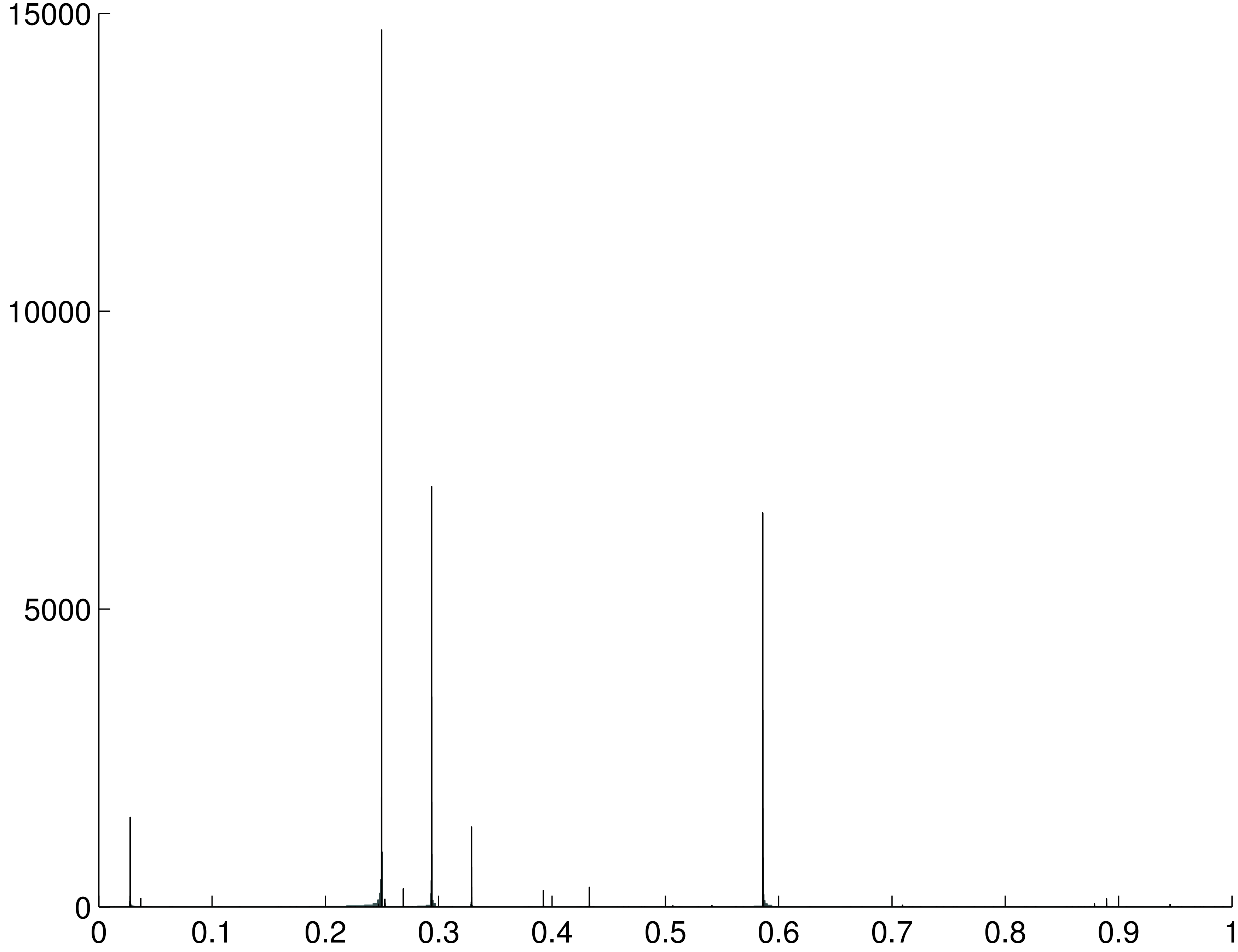
- G is discrete if

$$\mathbb{P}(G(\{\theta\}) > 0) = 1$$

- Above holds, since joint distribution is equivalent to:

$$\theta \sim H$$
$$G|\theta \sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1})$$

A draw from a Dirichlet Process



Atomic Distributions

- Draws from Dirichlet processes will always be atomic:

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

  where $\Sigma_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

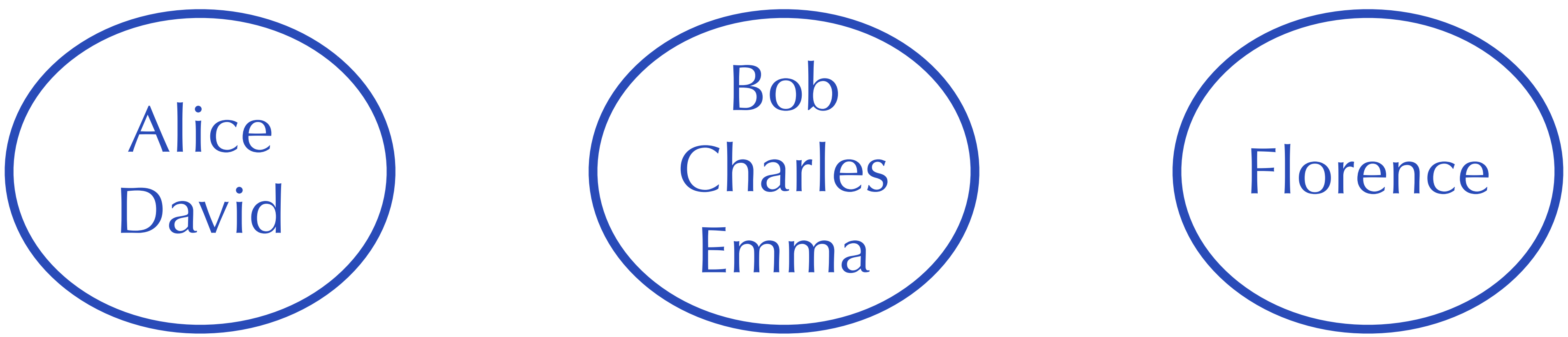
- A number of ways to specify the joint distribution of $\{\pi_k, \theta_k^*\}$.
  - **Stick-breaking construction**;
  - **Poisson-Dirichlet distribution**.

Random Partitions

[Aldous 1985, Pitman 2006]

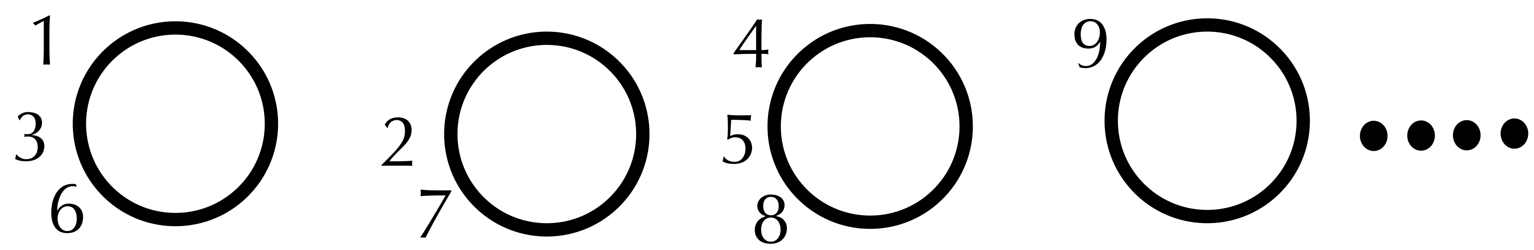
Partitions

- A **partition** ϱ of a set S is:
  - A disjoint family of non-empty subsets of S whose union in S.
  - $S = \{$Alice, Bob, Charles, David, Emma, Florence$\}$.
  - $\varrho = \{\ \{$Alice, David$\}$, $\{$Bob, Charles, Emma$\}$, $\{$Florence$\}\ \}$.



- Denote the set of all partitions of S as $\mathcal{P}_S$.
- **Random partitions** are random variables taking values in $\mathcal{P}_S$.
- We will work with partitions of $S = [n] = \{1,2,\ldots n\}$.

Chinese Restaurant Process



- Each customer comes into restaurant and sits at a table:

$$p(\text{sit at table } c) = \frac{n_c}{\alpha + \sum_{c \in \varrho} n_c} \qquad p(\text{sit at new table}) = \frac{\alpha}{\alpha + \sum_{c \in \varrho} n_c}$$

- Customers correspond to elements of S, and tables to clusters in ϱ.
- **Rich-gets-richer**: large clusters more likely to attract more customers.
- Multiplying conditional probabilities together, the overall probability of ϱ, called the **exchangeable partition probability function** (EPPF), is:

$$P(\varrho|\alpha) = \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c \in \varrho} \Gamma(|c|)$$

[Aldous 1985, Pitman 2006]

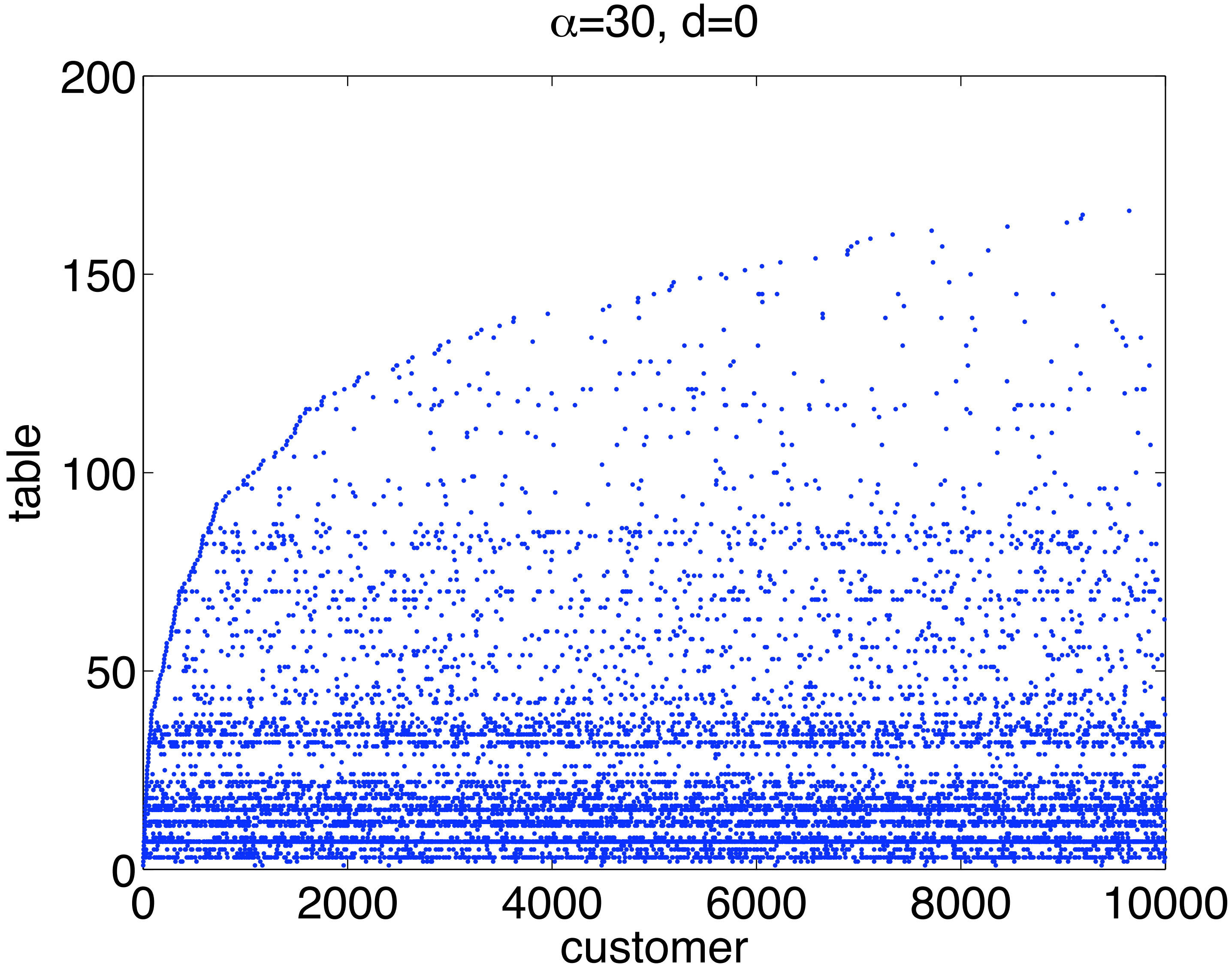
Number of Clusters

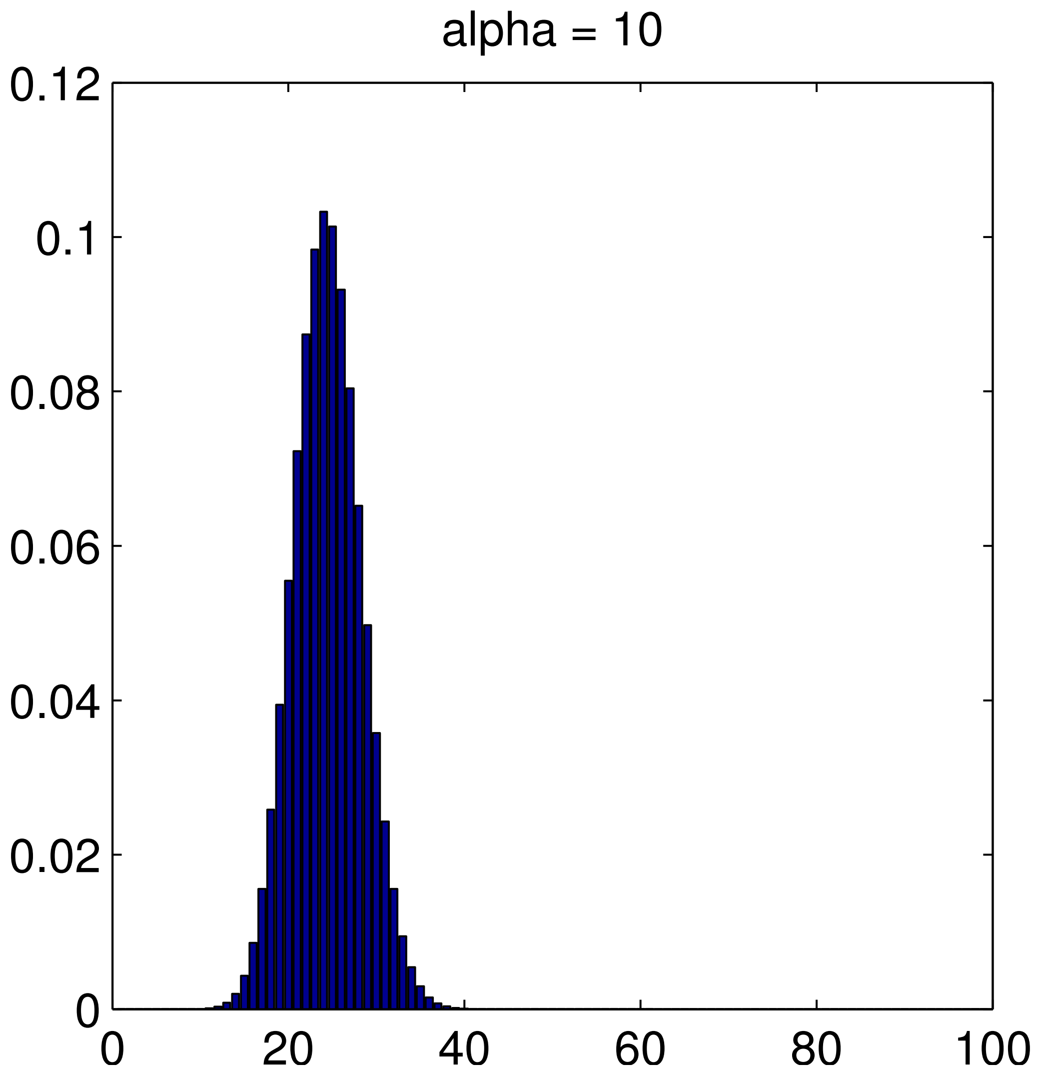
- The prior mean and variance of K are:

$$\mathbb{E}[|\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$

$$\mathbb{V}[|\rho||\alpha, n] = \alpha(\psi(\alpha + n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha + n) - \psi'(\alpha)) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$$

$$\psi(\alpha) = \frac{\partial}{\partial \alpha} \log \Gamma(\alpha)$$





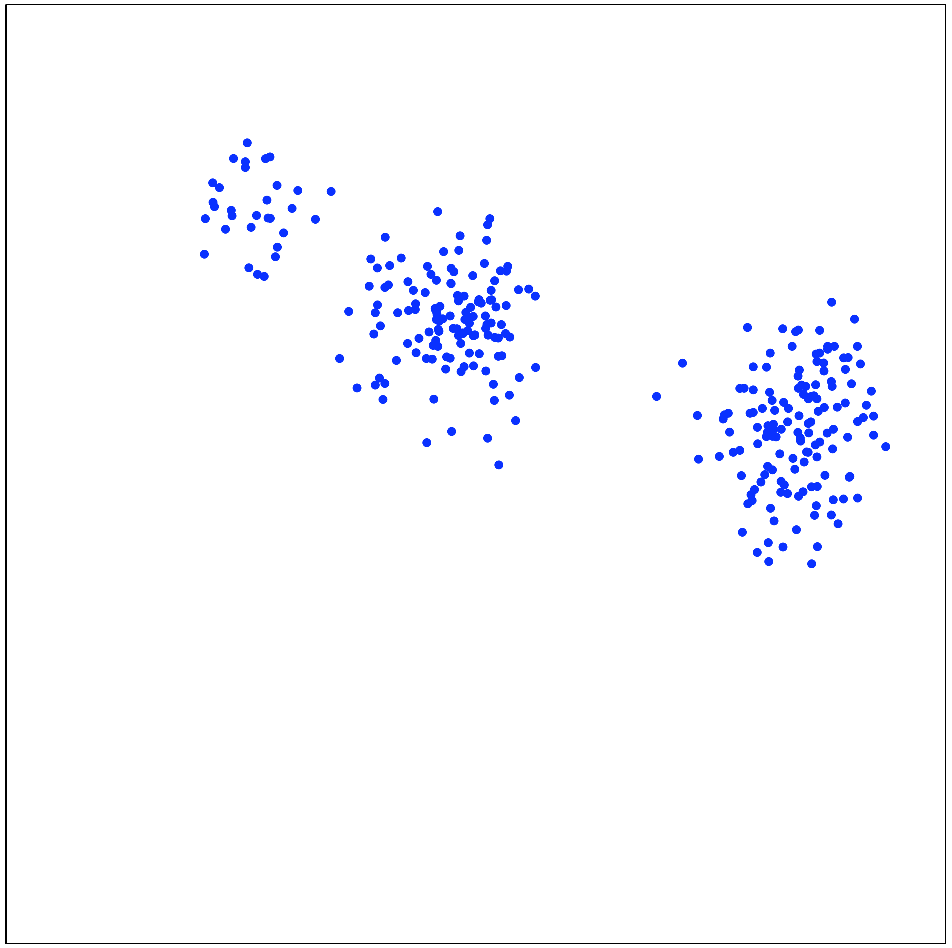
Model-based Clustering with Chinese Restaurant Process

Partitions in Model-based Clustering

- Partitions are the natural latent objects of inference in clustering.
  - Given a dataset S, partition it into clusters of similar items.
- Cluster c $\in \varrho$ described by a model

$$F(\theta^*_c)$$

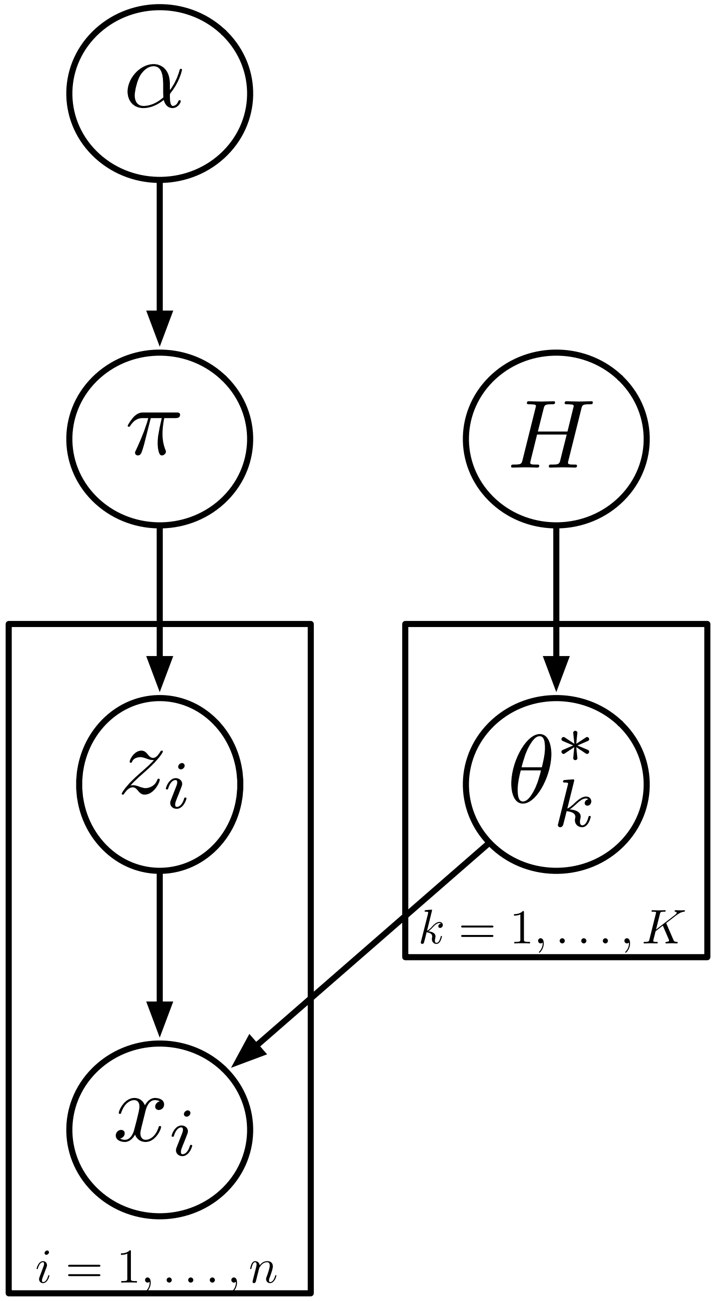
  parameterized by $\theta_c{}^*$.
- Bayesian approach: introduce prior over ϱ and $\theta_c{}^*$; compute posterior over both.



Finite Mixture Model

- Explicitly allow only K clusters in partition:
  - Each cluster k has parameter θ_k.
  - Each data item i assigned to k with **mixing probability** π_k.
  - Gives a random partition with at most K clusters.
- Priors on the other parameters:

$$\pi|\alpha \sim \mathrm{Dirichlet}(\alpha/K, \dots, \alpha/K)$$
$$\theta_k^*|H \sim H$$



Induced Distribution over Partitions

$$P(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha/K)} \frac{\prod_k \Gamma(n_k + \alpha/K)}{\Gamma(n + \alpha)}$$

- P($\mathbf{z}|\alpha$) describes a partition of the data set into clusters, *and a labelling of each cluster with a mixture component index.*

- Induces a distribution over partitions ϱ (without labelling) of the data set:

$$P(\varrho|\alpha) = [K]_{-1}^{k} \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in \varrho} \frac{\Gamma(|c| + \alpha/K)}{\Gamma(\alpha/K)}$$

  where $[x]_b^a = x(x+b)\cdots(x+(a-1)b)$.

- Taking $K \to \infty$, we get a proper distribution over partitions without a limit on the number of clusters:

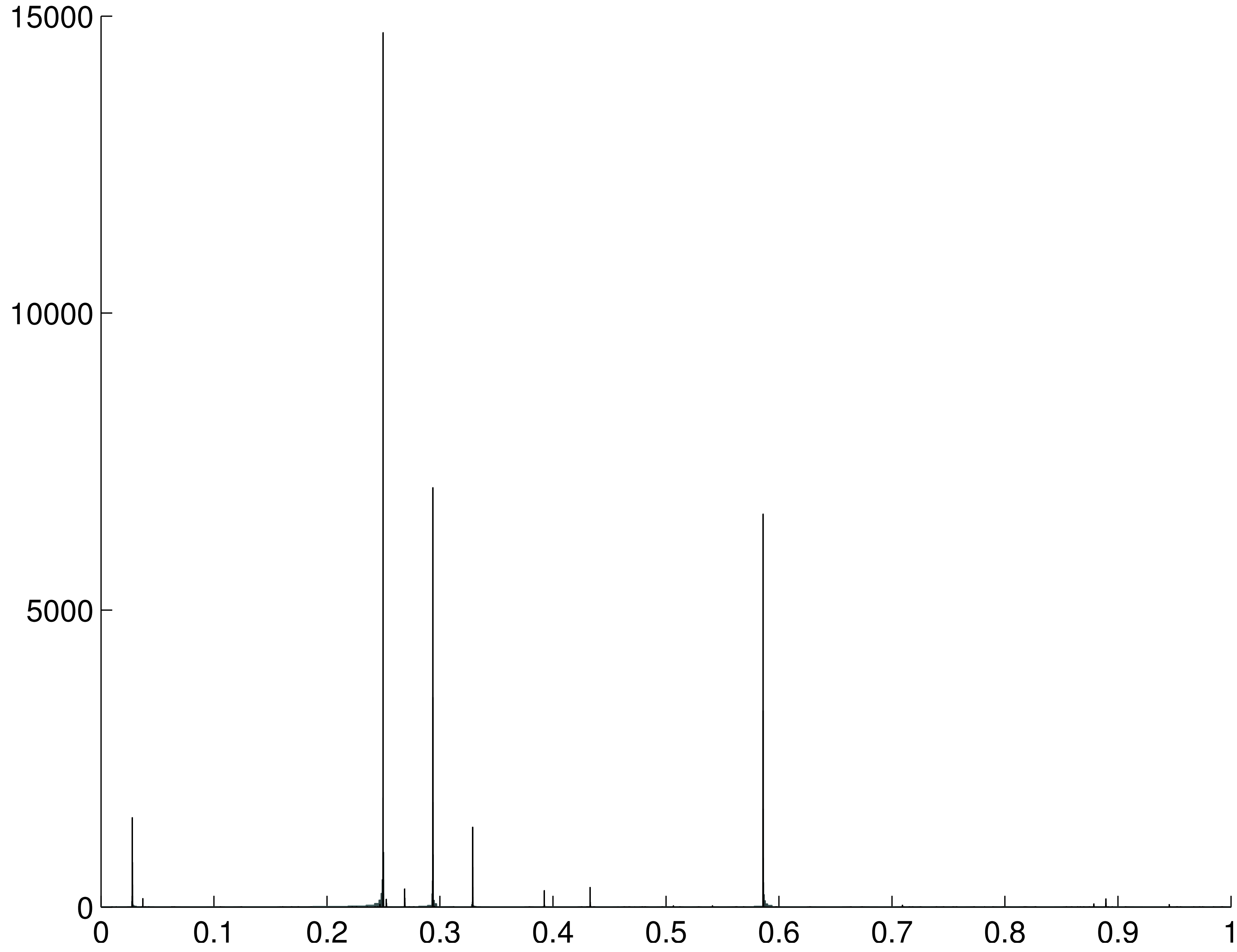
$$P(\varrho|\alpha) \to \frac{\alpha^{|\varrho|}\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in \varrho} \Gamma(|c|)$$

Chinese Restaurant Process

- An important representation of the Dirichlet process
- An important object of study in its own right.
- Predates the Dirichlet process and originated in genetics (related to Ewen's sampling formula there).
- Large number of MCMC samplers using CRP representation.
- Random partitions are useful concepts for clustering problems in machine learning
  - CRP mixture models for nonparametric model-based clustering.
  - hierarchical clustering using concepts of fragmentations and coagulations.
  - clustering nodes in graphs, e.g. for community discovery in social nets.
  - Other combinatorial structures can be built from partitions.

Random Probability Measures

A draw from a Dirichlet Process



Atomic Distributions

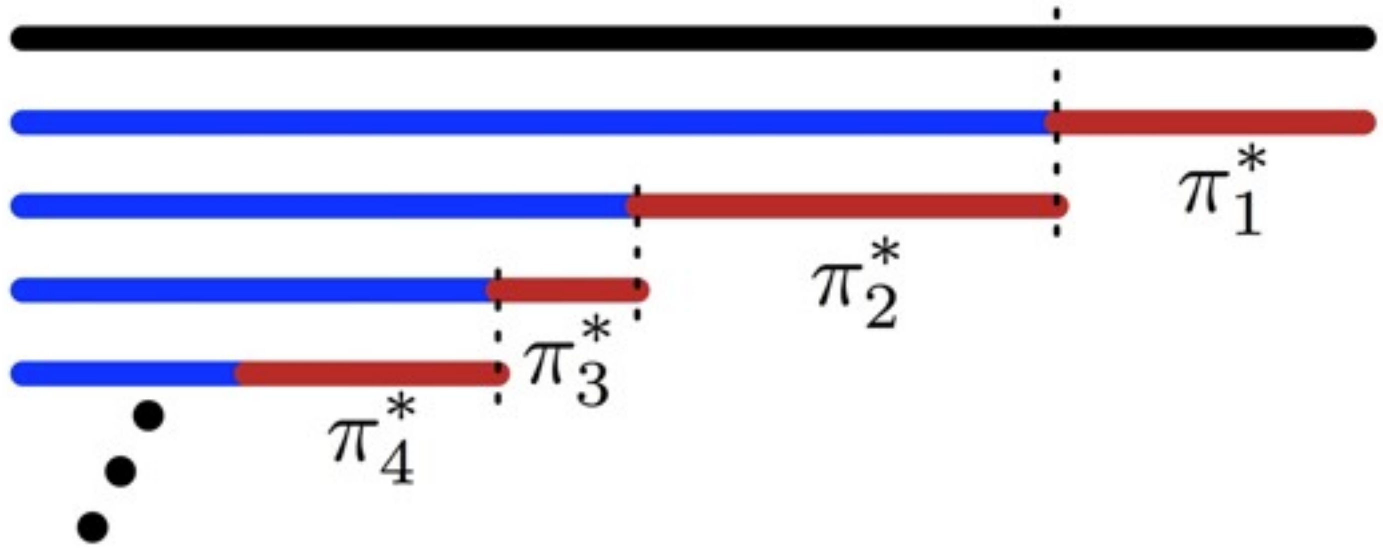
- Draws from Dirichlet processes will always be atomic:

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

where $\Sigma_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

- A number of ways to specify the joint distribution of $\{\pi_k, \theta_k^*\}$.
    - Stick-breaking construction;
    - Poisson-Dirichlet distribution.

Stick-breaking Construction



- **Stick-breaking construction** for the joint distribution:

$$\theta_k^* \sim H \qquad v_k \sim \text{Beta}(1, \alpha) \qquad \text{for } k = 1, 2, \ldots.$$

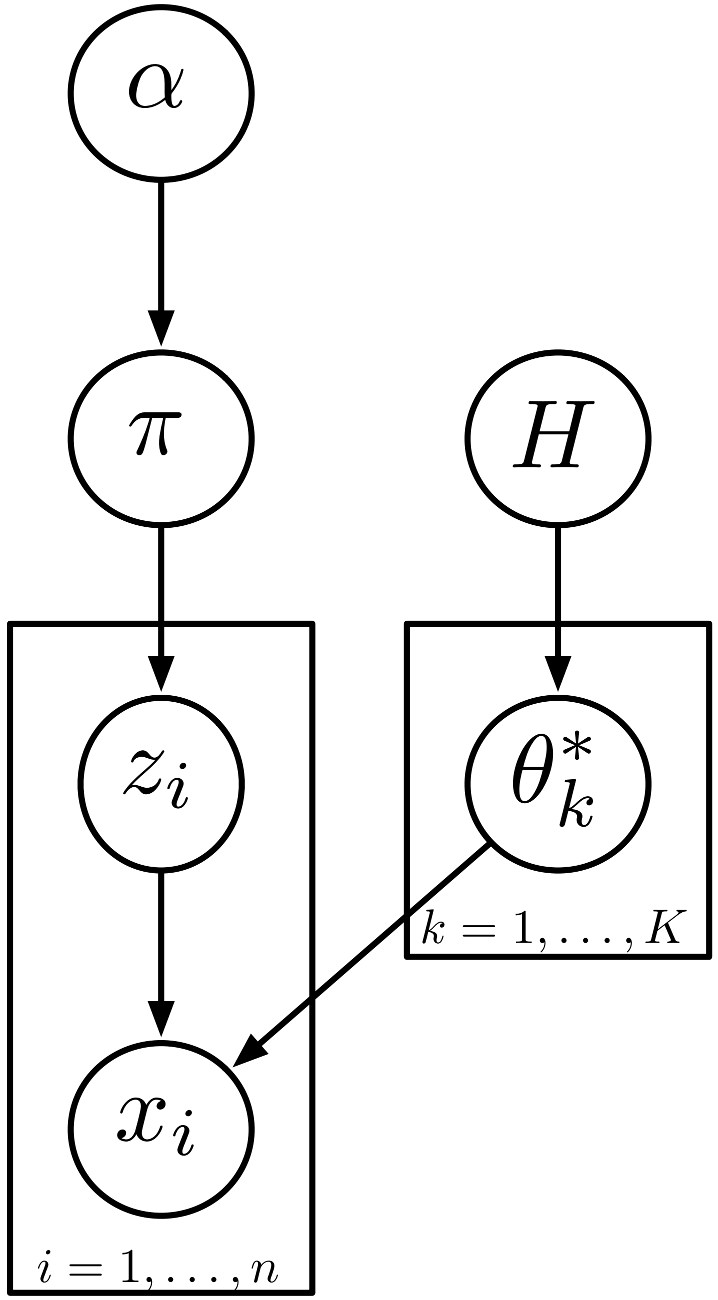
$$\pi_k^* = v_k \prod_{j=1}^{k-1} (1 - v_j) \qquad G = \sum_{k=1}^{\infty} \pi_k^* \delta_{\theta_k^*}$$

- π_k's are decreasing on average but not strictly.
- Distribution of $\{\pi_k\}$ is the **Griffiths-Engen-McCloskey** (GEM) distribution.
- **Poisson-Dirichlet distribution** [Kingman 1975] gives a strictly decreasing ordering (but is not computationally tractable).

Finite Mixture Model

- Explicitly allow only K clusters in partition:
  - Each cluster k has parameter θ_k.
  - Each data item i assigned to k with mixing probability π_k.
  - Gives a random partition with at most K clusters.
- Priors on the other parameters:

$$\pi|\alpha \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$$
$$\theta^*_k|H \sim H$$



Size-biased Permutation

- Reordering clusters do not change the marginal distribution on partitions or data items.
- By strictly decreasing π_k: Poisson-Dirichlet distribution.
- Reorder stochastically as follows gives stick-breaking construction:
  - Pick cluster k to be first cluster with probability π_k .
  - Remove cluster k and renormalize rest of $\{ \pi_k : j \neq k \}$; repeat.
- Stochastic reordering is called a **size-biased permutation**.
- After reordering, taking $K \to \infty$ gives the corresponding DP representations.

Stick-breaking Construction

- Easy to generalize stick-breaking construction:
  - to other random measures;
  - to random measures that depend on covariates or vary spatially.
- Easy to work with different algorithms:
  - MCMC samplers;
  - variational inference;
  - parallelized algorithms.

[Ishwaran & James 2001, Dunson 2010 and many others]

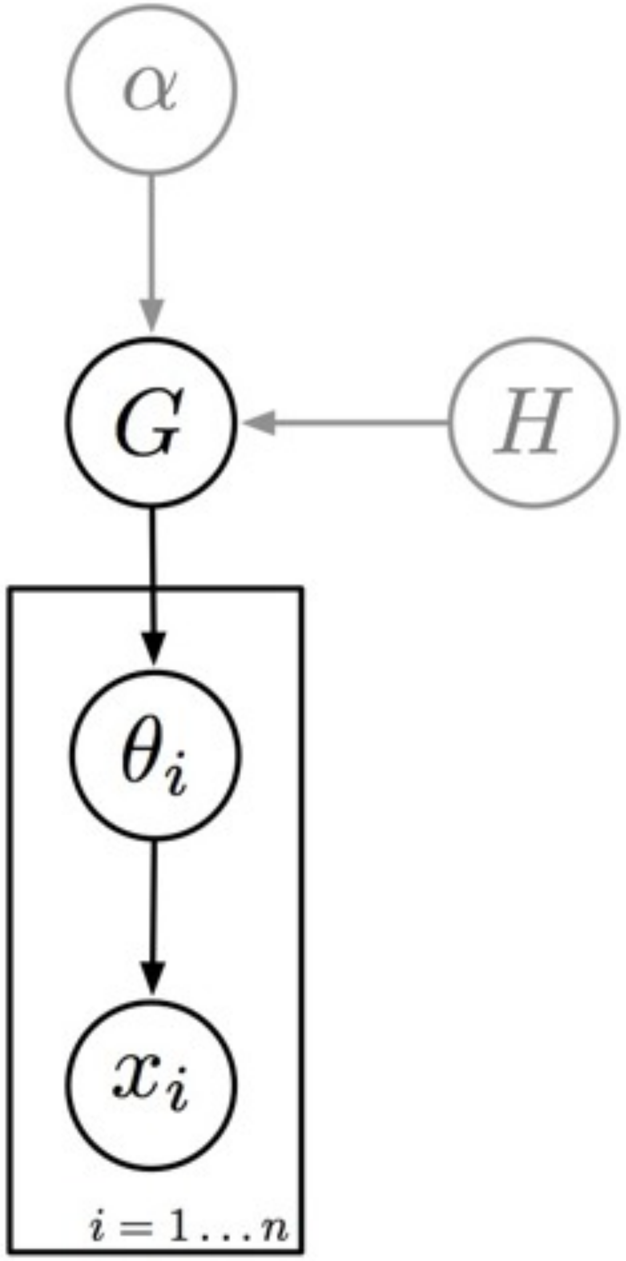
DP Mixture Model: Representations and Inference

DP Mixture Model

- A **DP mixture model**:

$$
\begin{aligned}
G|\alpha, H &\sim \mathrm{DP}(\alpha, H) \\
\theta_i|G &\sim G \\
x_i|\theta_i &\sim F(\theta_i)
\end{aligned}
$$

- Different representations:
  - $\theta_1, \theta_2, \ldots, \theta_n$ are clustered according to Pólya urn scheme, with induced partition given by a CRP.
  - G is atomic with weights and atoms described by stick-breaking construction.



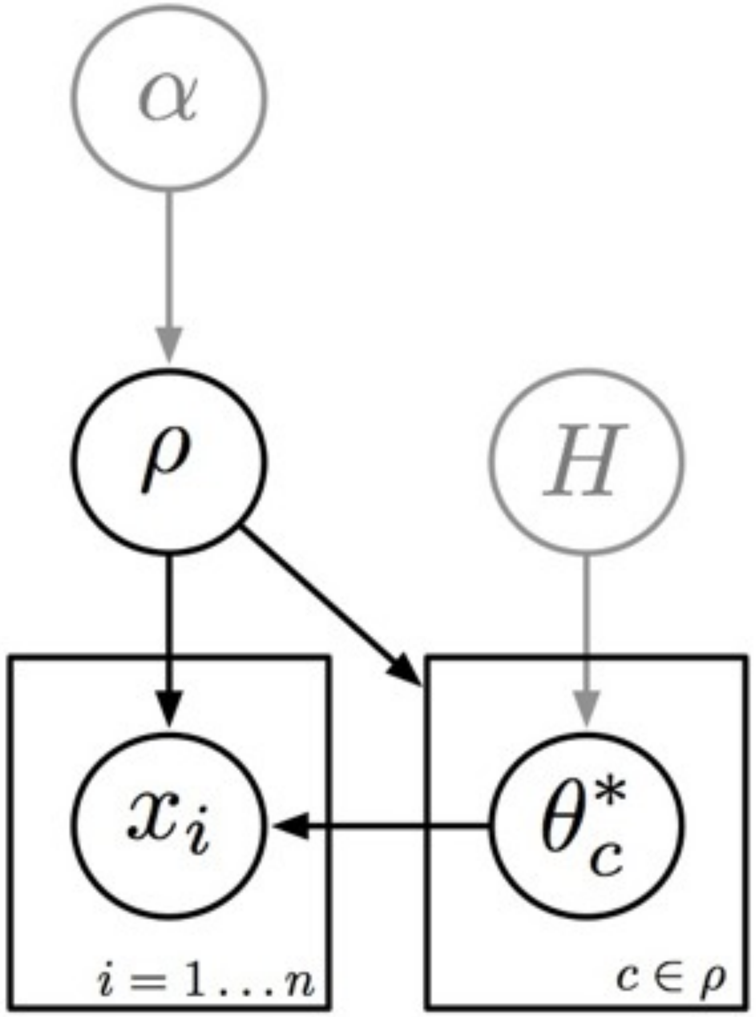
[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

CRP Representation

- Representing the partition structure explicitly with a CRP:

$$
\begin{aligned}
\rho|\alpha &\sim \mathrm{CRP}([n], \alpha) \\
\theta^*_c|H &\sim H \text{ for } c \in \rho \\
x_i|\theta^*_c &\sim F(\theta^*_c) \text{ for } c \ni i
\end{aligned}
$$

- Makes explicit that this is a clustering model.
- Using a CRP prior for ϱ obviates need to limit number of clusters as in finite mixture models.



[Neal 2000, Rasmussen 2000, Ishwaran & Zarepour 2002]

Marginal Sampler

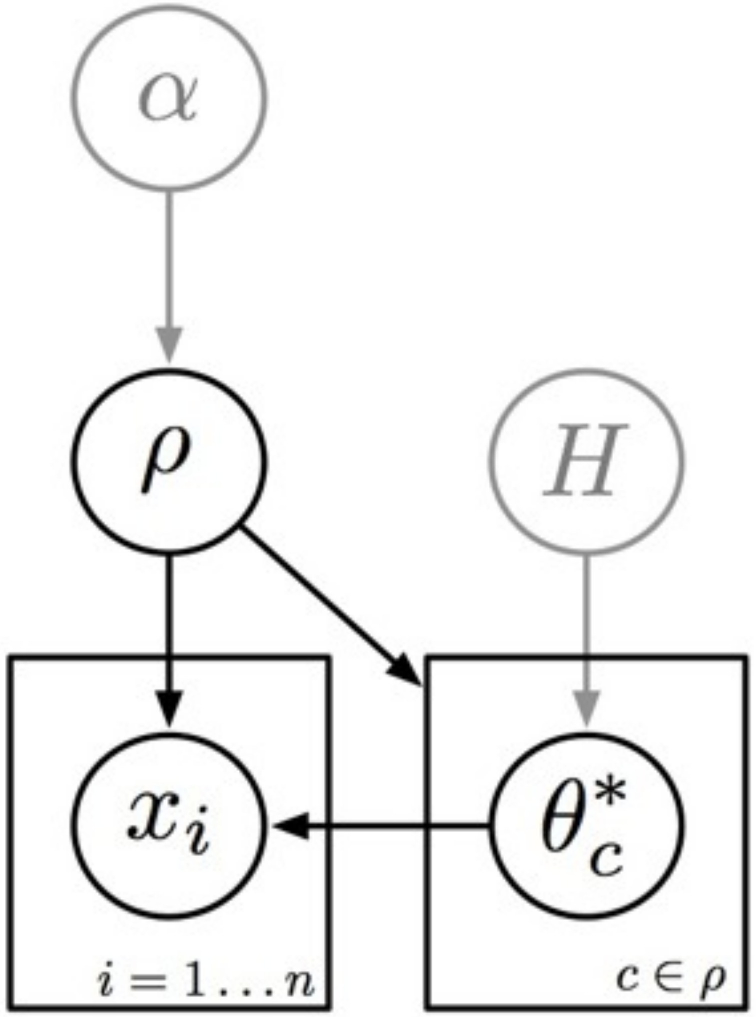
$$\rho|\alpha \sim \mathrm{CRP}([n], \alpha)$$
$$\theta_c^*|H \sim H \text{ for } c \in \rho$$
$$x_i|\theta_c^* \sim F(\theta_c^*) \text{ for } c \ni i$$

- "Marginal" MCMC sampler.
  - Marginalize out G, and Gibbs sample partition.
- Conditional probability of cluster of data item i:

$$P(\rho_i|\rho_{\backslash i}, \mathbf{x}, \boldsymbol{\theta}) = P(\rho_i|\rho_{\backslash i})P(x_i|\rho_i, \mathbf{x}_{\backslash i}, \boldsymbol{\theta})$$

$$P(\rho_i|\rho_{\backslash i}) = \begin{cases} \frac{|c|}{n-1+\alpha} & \text{if } \rho_i = c \in \rho_{\backslash i} \\ \frac{\alpha}{n-1+\alpha} & \text{if } \rho_i = \text{new} \end{cases}$$

$$P(x_i|\rho_i, \mathbf{x}_{\backslash i}, \boldsymbol{\theta}) = \begin{cases} f(x_i|\theta_{\rho_i}) & \text{if } \rho_i = c \in \rho_{\backslash i} \\ \int f(x_i|\theta)h(\theta)d\theta & \text{if } \rho_i = \text{new} \end{cases}$$



- A variety of methods to deal with new clusters.
- Difficulty lies in dealing with new clusters, especially when prior h is not conjugate to f.

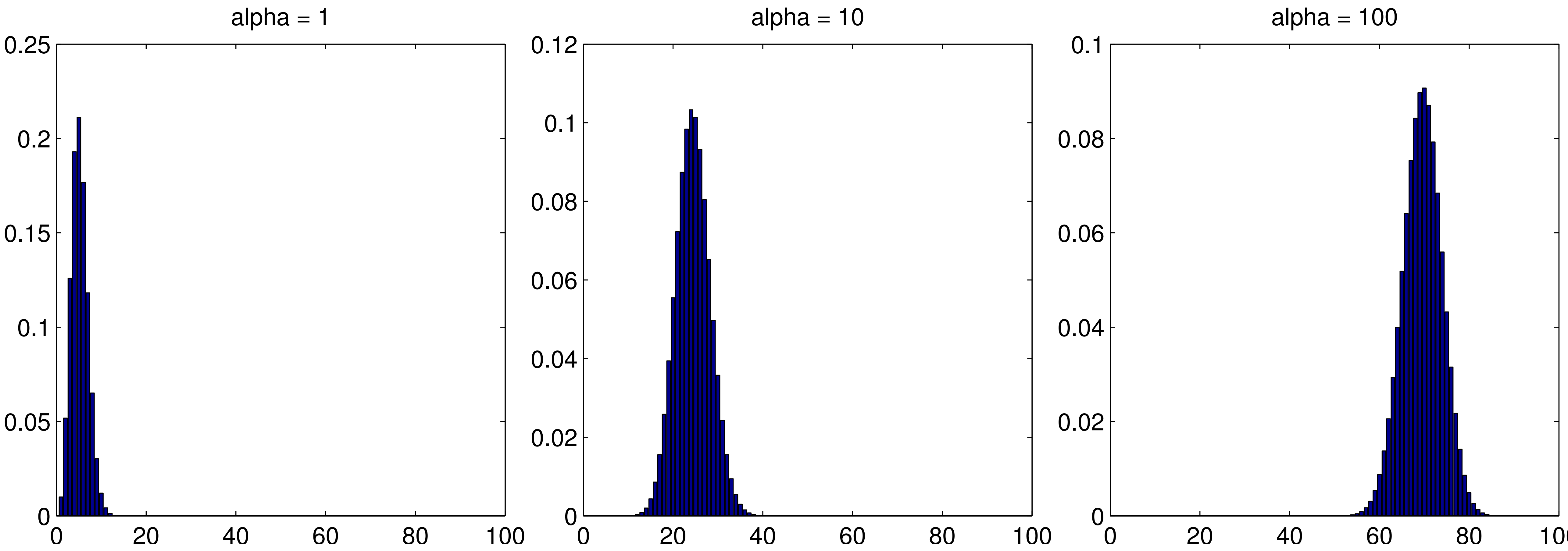
[Neal 2000]

Induced Prior on the Number of Clusters

- The prior expectation and variance of $|\varrho|$ are:

$$\mathbb{E}[|\rho||\alpha,n] = \alpha(\psi(\alpha+n)-\psi(\alpha)) \approx \alpha\log\left(1+\frac{n}{\alpha}\right)$$

$$\mathbb{V}[|\rho||\alpha,n] = \alpha(\psi(\alpha+n)-\psi(\alpha)) + \alpha^2(\psi'(\alpha+n)-\psi'(\alpha)) \approx \alpha\log\left(1+\frac{n}{\alpha}\right)$$



Marginal Gibbs Sampler Pseudocode

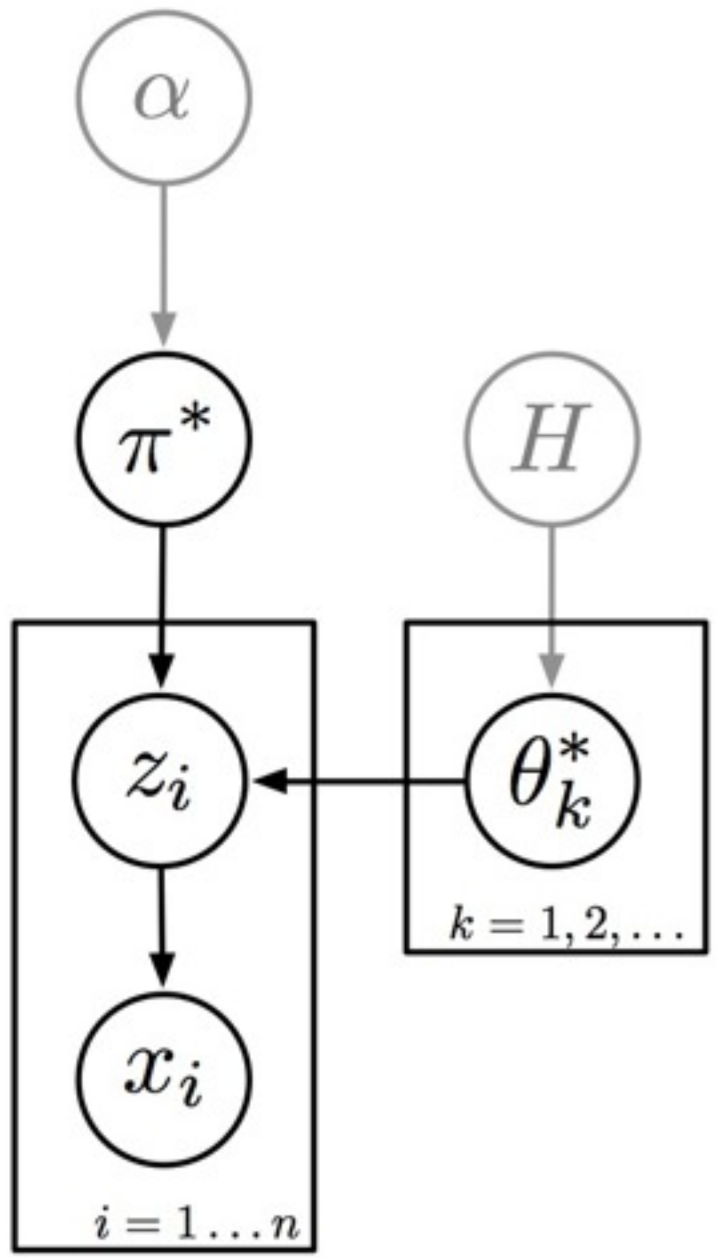
- Initialize: randomly assign each data item to some cluster.
- K := the number of clusters used.
- For each cluster $k = 1 \ldots K$:
  - Compute sufficient statistics $s_k := \Sigma \{ s(x_i) : z_i = k \}$.
  - Compute cluster sizes $n_k := \# \{ i : z_i = k \}$.
- Iterate until convergence:
  - For each data item $i = 1 \ldots n$:
    - Let $k := z_i$ be the current cluster data item is assigned to.
    - Remove data item: $s_k -= s(x_i)$, $n_k -= 1$.
    - If $n_k = 0$ then remove cluster k ($K -= 1$ and relabel rest of clusters).
    - Compute conditional probabilities p(zi=c|others) for $c = 1 \ldots K$, $k_{\text{empty}} := K+1$.
    - Sample new cluster for data item from conditional probabilities.
    - If $c = k_{\text{empty}}$ then create new cluster: $K+=1$, $s_c := 0$, $n_c = 0$.
    - Add data item: $z_i := c$, $s_c += s(x_i)$, $n_c += 1$.

Stick-breaking Representation

- Dissecting stick-breaking representation for G:

$$
\begin{aligned}
\pi^*|\alpha &\sim \mathrm{GEM}(\alpha) \\
\theta_k^*|H &\sim H \\
z_i|\pi^* &\sim \mathrm{Discrete}(\pi^*) \\
x_i|z_i, \theta_{z_i}^* &\sim F(\theta_{z_i}^*)
\end{aligned}
$$

- Makes explicit that this is a mixture model with an infinite number of components.
- Conditional sampler:
  - Standard Gibbs sampler, except need to truncate the number of clusters.
  - Easy to work with non-conjugate priors.
  - For sampler to mix well need to introduce moves for permuting the order of clusters.



[Ishwaran & James 2001, Walker 2007, Papaspiliopoulos & Roberts 2008]

Explicit G Sampler

- Represent G explicitly, alternately sampling $\{\theta_i\}|G$ (simple) and $G|\{\theta_i\}$:.

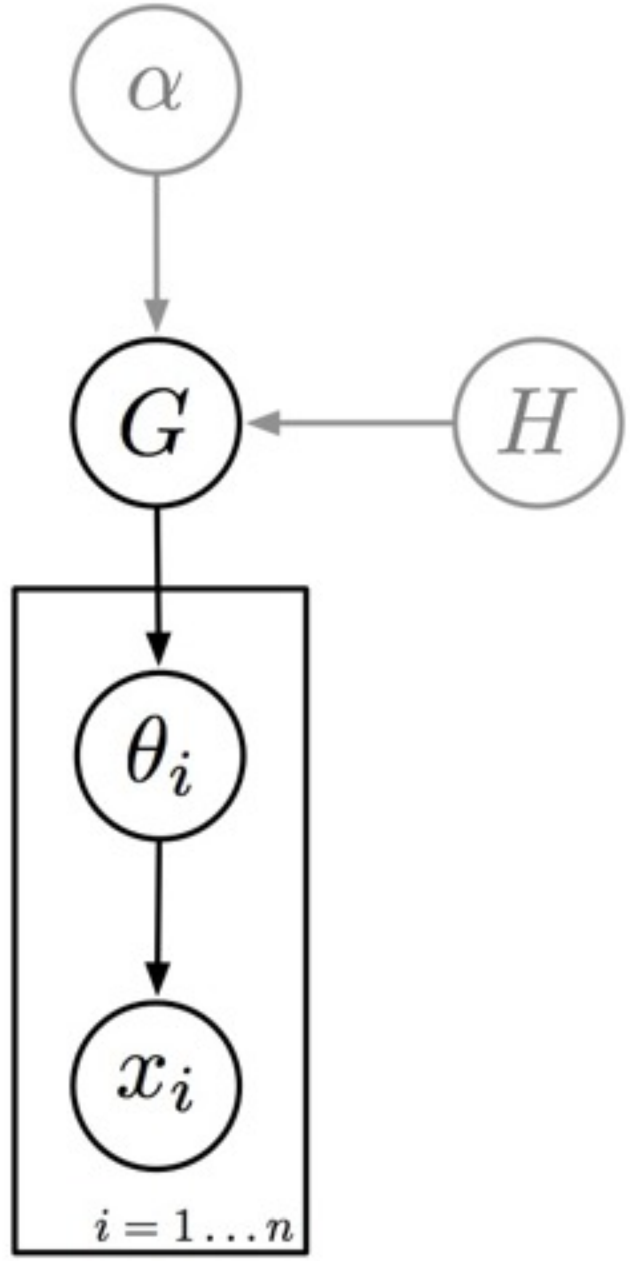
$$G|\theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha+n, \tfrac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha+n})$$

$$G = \pi_0^* G' + \sum_{k=1}^K \pi_k^* \delta_{\theta_k^*}$$

$$(\pi_0^*, \pi_1^*, \ldots, \pi_K^*) \sim \mathrm{Dirichlet}(\alpha, n_1, \ldots, n_K)$$

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

- Use a stick-breaking representation for G' and truncate as before.
- No explicit ordering of the non-empty clusters makes for better mixing.
- Explicit representation of G allows for posterior estimates of functionals of G.



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

$$\theta_i|G \sim G$$

$$x_i|\theta_i \sim F(\theta_i)$$

Other Inference Algorithms

- Split-merge algorithms [Jain & Neal 2004].
  - Close in spirit to reversible-jump MCMC methods [Green & richardson 2001].

- Sequential Monte Carlo methods [Liu 1996, Ishwaran & James 2003, Fearnhead 2004, Mansingkha et al 2007].

- Variational algorithms [Blei & Jordan 2006, Kurihara et al 2007, Teh et al 2008].

- Expectation propagation [Minka & Ghahramani 2003, Tarlow et al 2008].