PETUUM

Carnegie Mellon University
School of Computer Science

Probabilistic Graphical Models

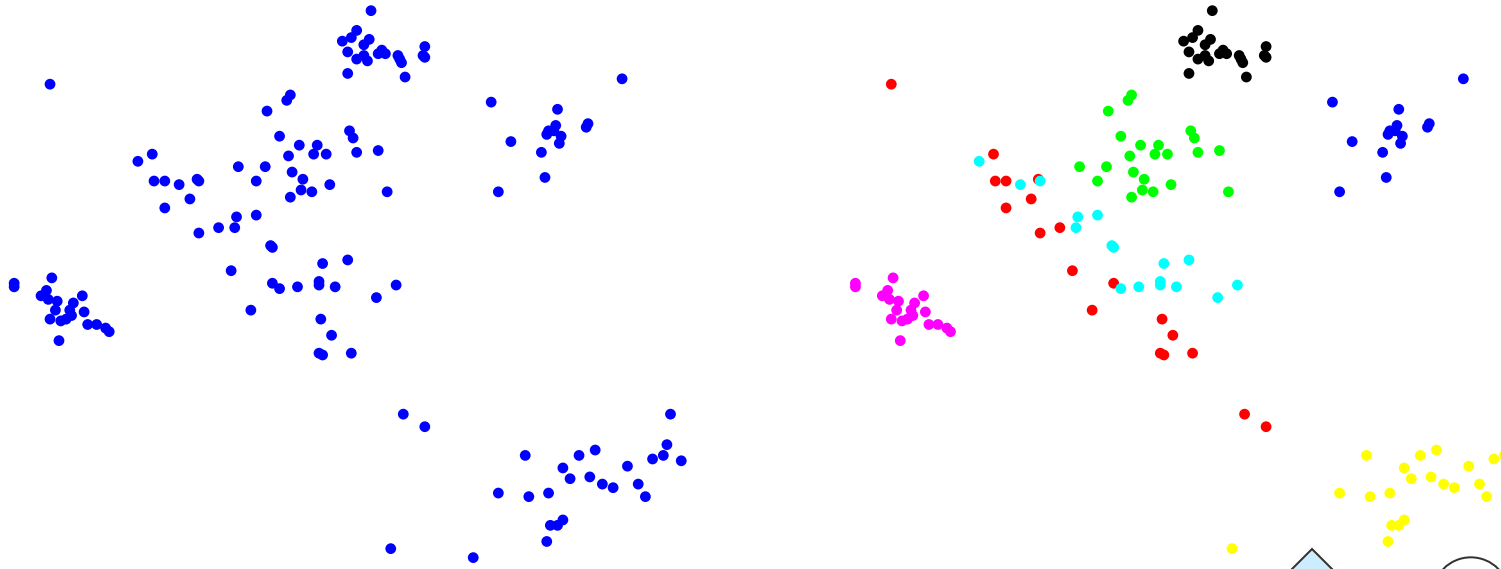
Bayesian Nonparametrics (continued): Hierarchical Dirchlet Process, Indian Buffet Process

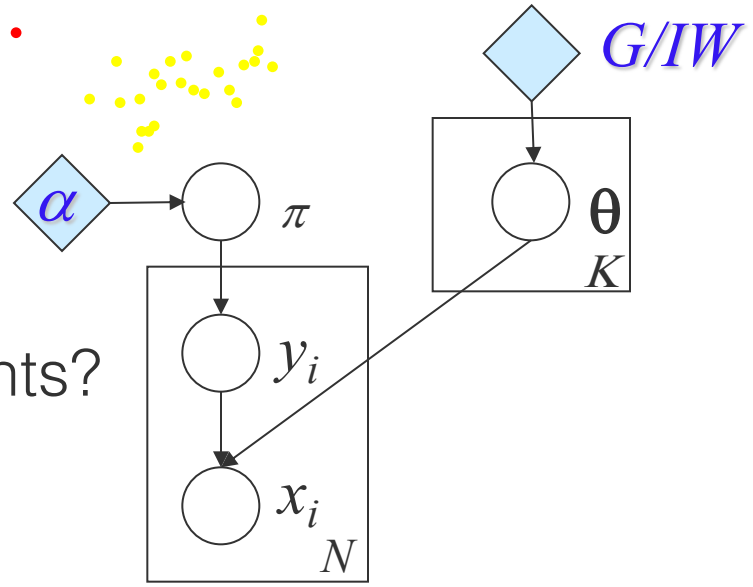
Maruan Al-Shedivat

Lecture 23, April 10, 2019

Reading: see class homepage

Recap: Distribution over Distributions





- Mixture of Gaussians – but how many components?
- What if we see more data – may find new components?

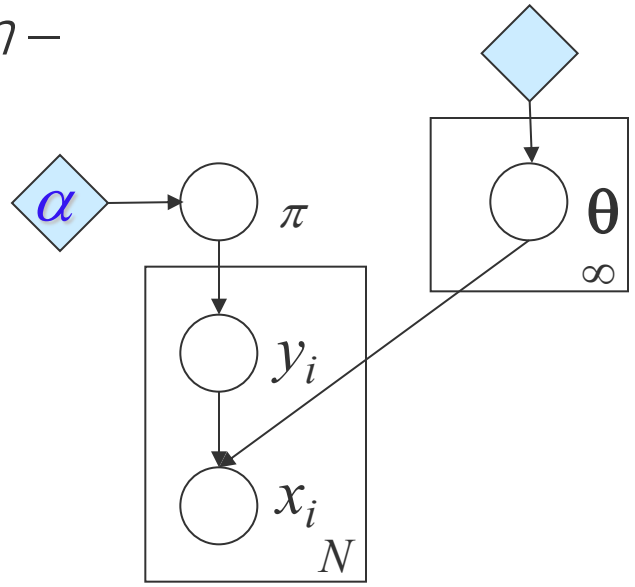
Recap: Bayesian Nonparametric Mixture Models

- Make sure we always have more clusters than we need.
- Solution – infinite clusters *a priori*!

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

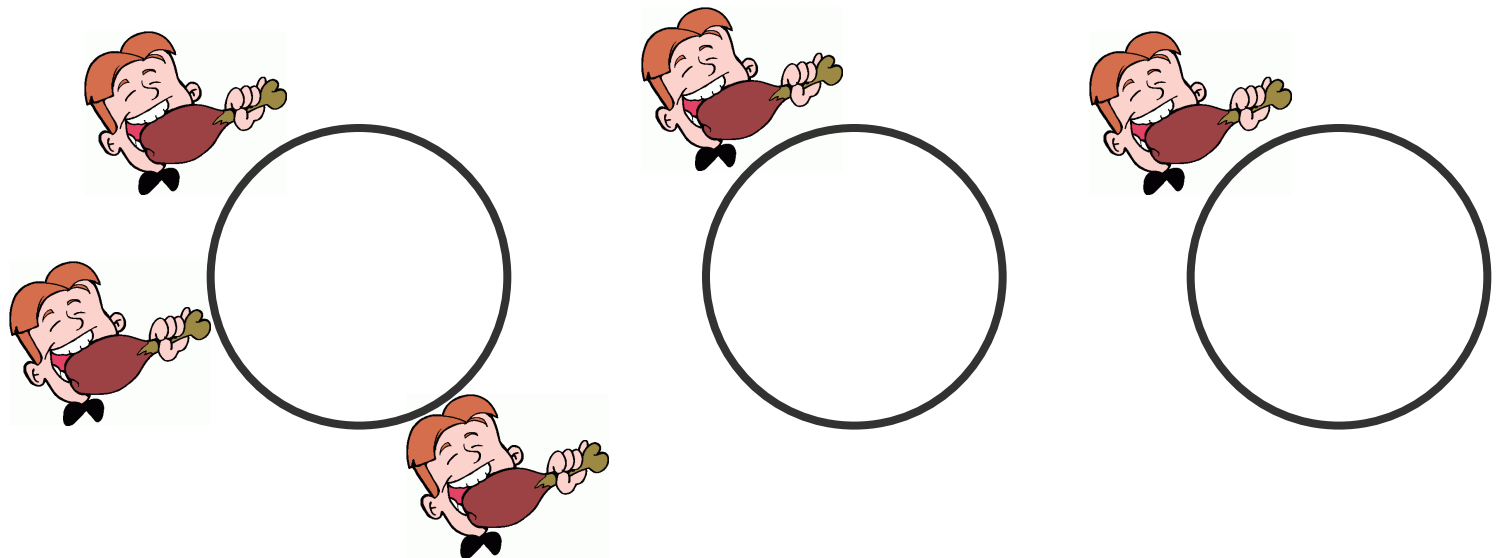
- A finite data set will always use a finite – but *random* – number of clusters.

- How to choose the prior?
- We want something *like* a Dirichlet prior – but with an infinite number of components.



Recap: Chinese Restaurant Process (CRP)

- The distribution over partitions can be described in terms of the following restaurant metaphor:
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant and:
    - Sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people sat at table k.
    - Starts a new table with probability $\alpha/(n-1+\alpha)$.



Recap: Stick-breaking Construction of DP

- We can represent samples from the Dirichlet process exactly.
- Imagine a stick of length 1, representing total probability.
- For k=1,2,…
  1. Sample a beta(1,α) random variable b_k.
  2. Break off a fraction b_k of the stick. This is the k^{th} atom size
  3. Sample a random location for this atom.
  4. Recurse on the remaining stick.

$$G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$
$$\pi_k := b_k \prod_{j=1}^{k-1} (1 - b_k)$$
$$b_k \sim \text{Beta}(1, \alpha)$$

[Sethuraman, 1994]

Recap: Stick-breaking Construction of DP

$$G \sim \mathrm{DP}(\alpha, G_0)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

$$\theta_k \sim G_0$$

Location

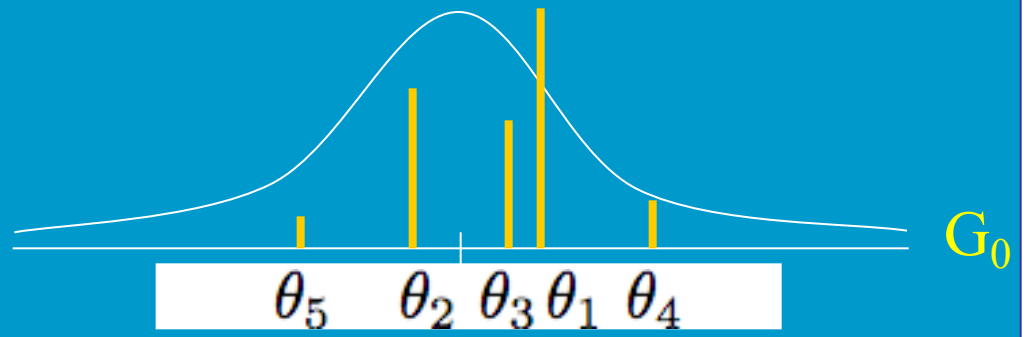
$$\sum_{k=1}^{\infty} \pi_k = 1$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

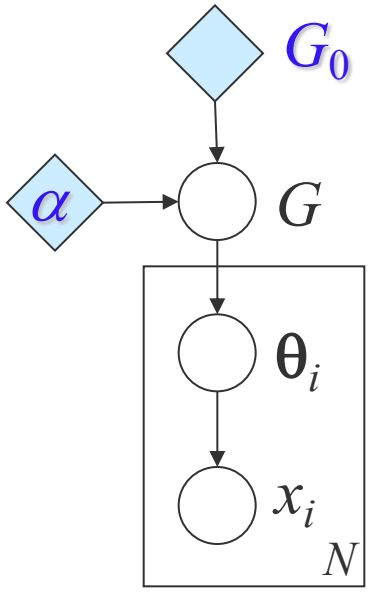
Mass

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

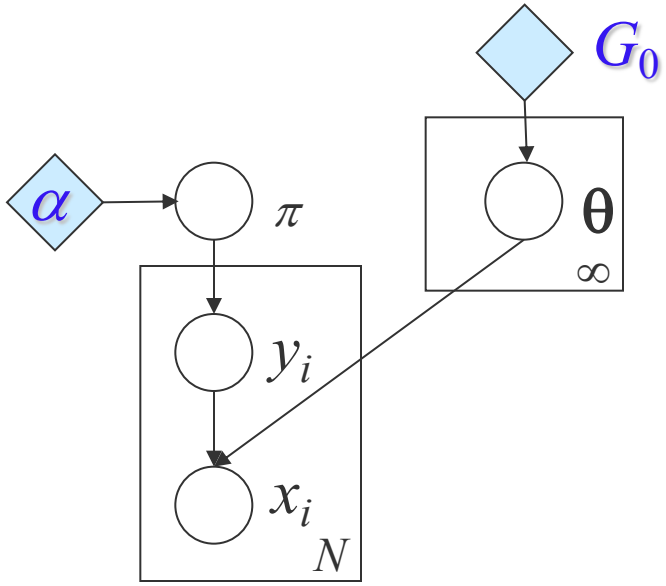
$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
1.0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24



Recap: Graphical Model Representations of DP



The Pólya urn construction



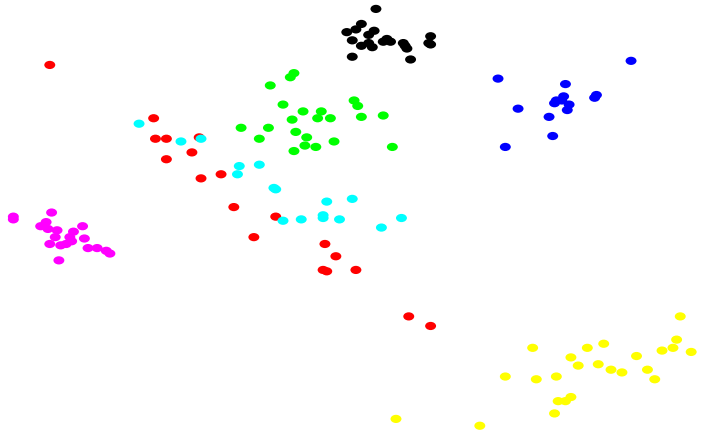
The Stick-breaking construction

Inference in the DP mixture model

$$G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \mathrm{DP}(\alpha, H)$$

$$\phi_n \sim G$$

$$x_n \sim f(\phi_n)$$



Inference: Collapsed Gibbs Sampler

- We can integrate out G to get the CRP.
- **Reminder:** Observations in the CRP are exchangeable.
- **Corollary:** When sampling any data point, we can always rearrange the ordering so that it is the last data point.
- Let z_n be the cluster allocation of the n-th data point.
- Let K be the total number of instantiated clusters.
- Then:

$$p(z_n = k | x_n, z_{-n}, \phi_{1:K}) \propto \begin{cases} m_k f(x_n|\phi_k) & k \leq K \\ \alpha \int_\Omega f(x_n|\phi) H(d\phi) & k = K+1 \end{cases}$$

- If we use a conjugate prior for the likelihood, we can often integrate out the cluster parameters.

Problems with the Collapsed Gibbs Sampler

- We are only updating one data point at a time.
- Imagine two “true” clusters are merged into a single cluster – a single data point is unlikely to “break away”.
- Getting to the true distribution involves going through low probability states ➔ mixing can be slow.
- If the likelihood is not conjugate, integrating out parameter values for new features can be difficult.
- Neal [2000] offers a variety of algorithms.
- **Alternative:** instantiate the latent measure.

Inference: Blocked Gibbs Sampler

- Rather than integrate out *G*, we can instantiate it.
- **Problem:** *G* is infinite-dimensional.
- **Solution:** Approximate it with a truncated stick-breaking process:

$$G^K := \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

$$\pi_k = b_k \prod_{j=1}^{k-1} (1 - b_j)$$

$$b_k \sim \text{Beta}(1, \alpha), k = 1, \ldots, K-1$$

$$b_K = 1$$

Inference: Blocked Gibbs sampler

- Sampling the cluster indicators:

$$p(z_n = k|\text{rest}) \propto \pi_k f(x_n|\theta_k)$$

- Sampling the stick breaking variables:
  - We can think of the stick breaking process as a sequence of binary decisions.
  - Choose $z_n = 1$ with probability b_1.
  - If $z_n \neq 1$, choose $z_n = 2$ with probability b_2.
  - *etc.*

$$b_k|\text{rest} \sim \text{Beta}\left(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j\right)$$

Inference: Slice sampler

- **Problem with batch sampler:** Fixed truncation introduces error.

- Idea:
  - Introduce *random truncation*.
  - If we marginalize over the random truncation, we recover the full model.
- Introduce a uniform random variable u_n for each data point.
- Sample indicator z_n according to

$$p(z_n = k|\text{rest}) = I(\pi_k > u_n) f(x_n|\theta_k)$$

- Only a **finite** number of possible values.

Inference: Slice sampler

- The conditional distribution for u_n is just:

$$u_n|\text{rest} \sim \text{Uniform}[0, \pi_{z_n}]$$

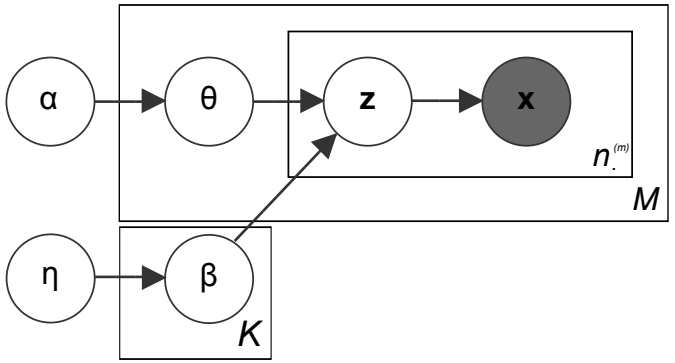
- Conditioned on the u_n and the z_n, the π_k can be sampled according to the block Gibbs sampler.

$$b_k|\text{rest} \sim \text{Beta}\left(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j\right)$$

- Only need to represent a finite number K of components such that

$$1 - \sum_{k=1}^{K} \pi_k < min(u_n)$$

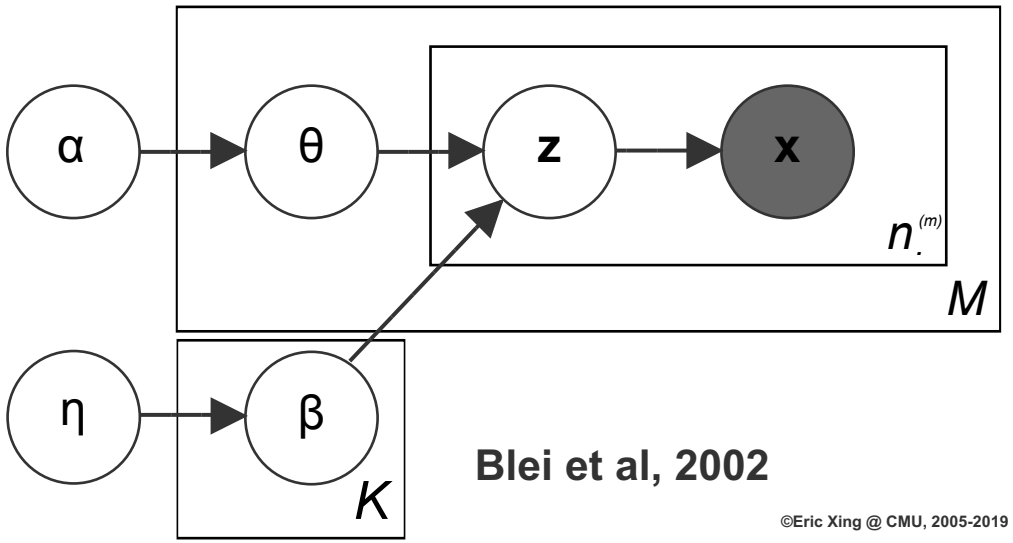
Topic models



- Topic models describe documents using a distribution over features.
- Each feature is a distribution over words
- Each document is represented as a collection of words (usually unordered – “bag of words” assumption).
- The words within a document are distributed according to a document-specific mixture model
- Each word in a document is associated with a feature.
- The features are shared between documents.
- The features learned tend to give high probability to semantically related words – “topics”

Latent Dirichlet Allocation

- For each topic $k=1,\ldots,K$
  - Sample a distribution over words, $\beta \sim \text{Dir}(\eta_1,\ldots,\eta_V)$
- For each document $m=1,\ldots,M$
  - Sample a distribution over topics, $\theta_m \sim \text{Dir}(\alpha_1,\ldots,\alpha_K)$
  - For each word $n=1,\ldots,N_m$
    - Sample a topic $z_{mn} \sim \text{Discrete}(\theta_m)$
    - Sample a word $w_{mk} \sim \text{Discrete}(\beta_z)$



Blei et al, 2002

How do we know the number of topics a priori?

Constructing a Topic Model with Infinitely Many Topics

- LDA: Each distribution is associated with a distribution over *K* topics.
- Problem: How to choose the number of topics?

- Solution:
  - Infinitely many topics!
  - Replace the Dirichlet distribution over topics with a Dirichlet process!

- Problem: We want to make sure the topics are *shared* between documents

Sharing Topics

- In LDA, we have *M* independent samples from a Dirichlet distribution.
- The weights are different, but the topics are fixed to be the same.

- If we replace the Dirichlet distributions with Dirichlet processes, each atom of each Dirichlet process will pick a topic *independently* of the other topics.
- Because the base measure is *continuous*, we have zero probability of picking the same topic twice.

Sharing Topics

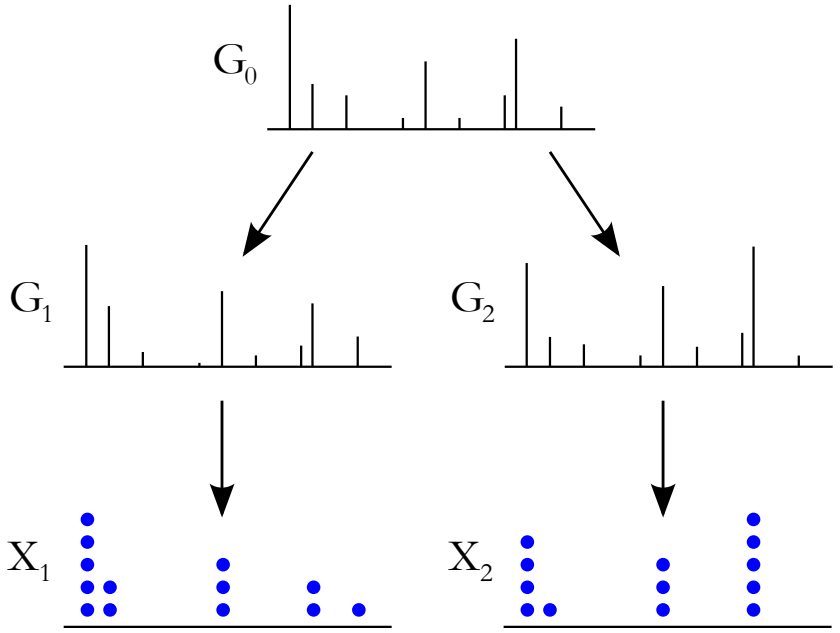
- If we want to pick the same topic twice, we need to use a *discrete* base measure.
- For example, if we chose the base measure to be $H = \sum_{k=1}^{K} \alpha_k \delta_{\beta_k}$, then we would have LDA again.

- We want there to be an infinite number of topics, so we want an *infinite, discrete* base measure.
- We want the location of the topics to be random, so we want an *infinite, discrete, random* base measure.

Hierarchical Dirichlet Process (Teh et al, 2006)

- Solution: Sample the base measure from a Dirichlet process!

$$G_0 \sim \mathrm{DP}(\gamma, H)$$
$$G_m \sim \mathrm{DP}(\alpha, G_0)$$

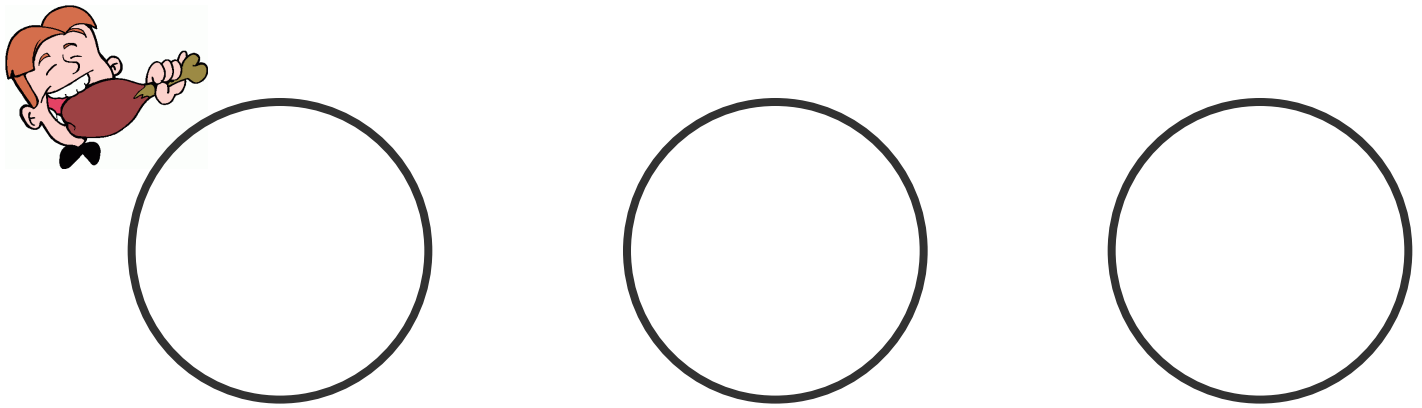


Chinese Restaurant Franchise

- Imagine a *franchise* of restaurants that serve an infinitely large, global menu.
- Each table in each restaurant orders a single dish.
- Let n_{rt} be the number of customers in restaurant r sitting at table t.
- Let m_{rd} be the number of tables in restaurant r serving dish d.
- Let $m_{.d}$ be the number of tables, across *all* restaurants, serving dish d.

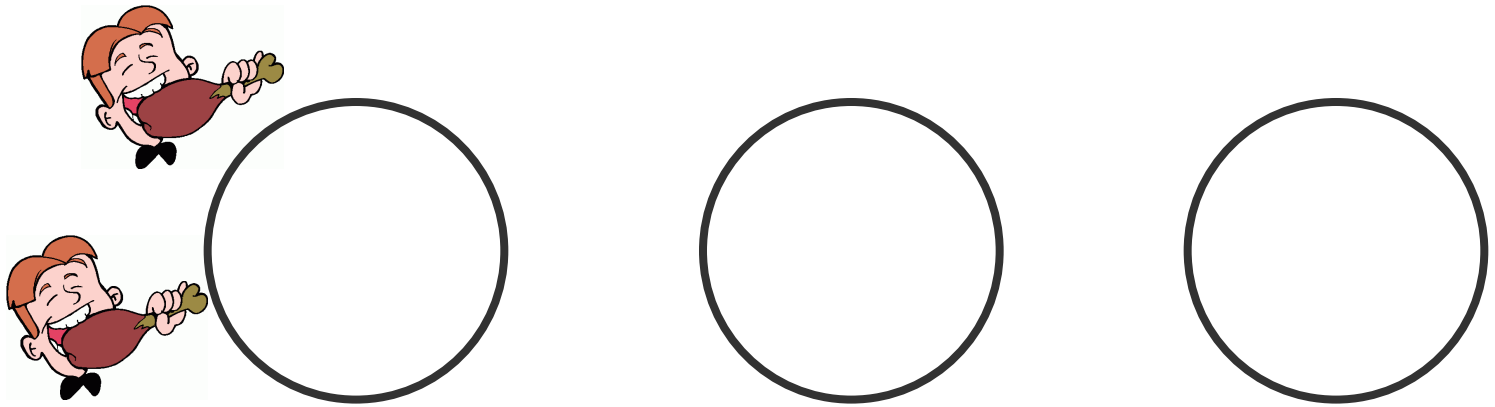
Chinese Restaurant Franchise

- Customers enter the restaurants, and sit at tables according to the Chinese Restaurant Process
  - The first customer enters a restaurant, and picks a table.



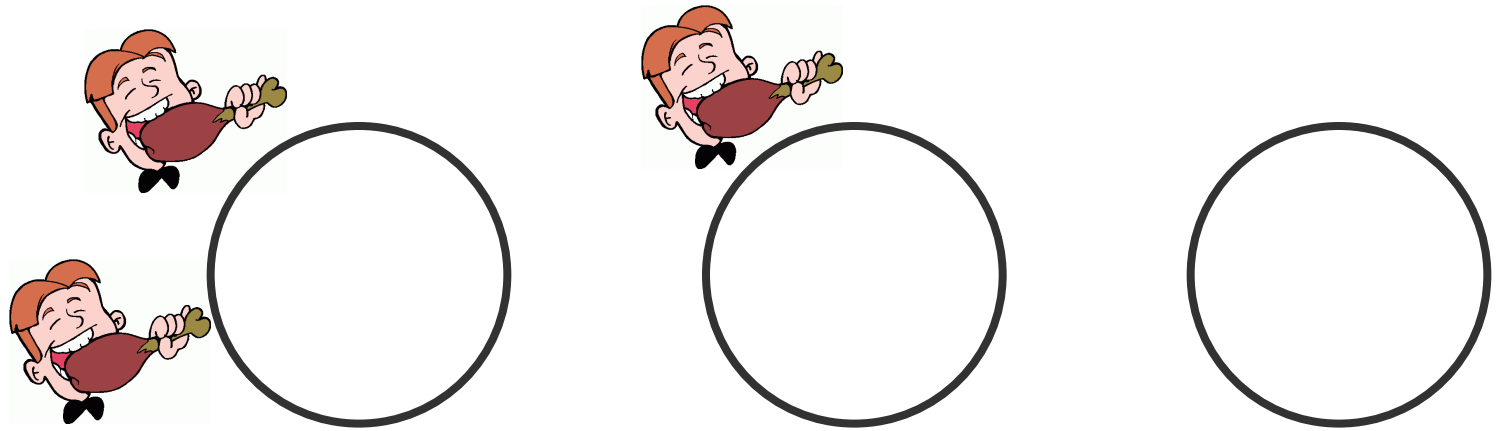
Chinese Restaurant Franchise

- Customers enter the restaurants, and sit at tables according to the Chinese Restaurant Process
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people sat at table k. He starts a new table with probability $\alpha/(n-1+\alpha)$.



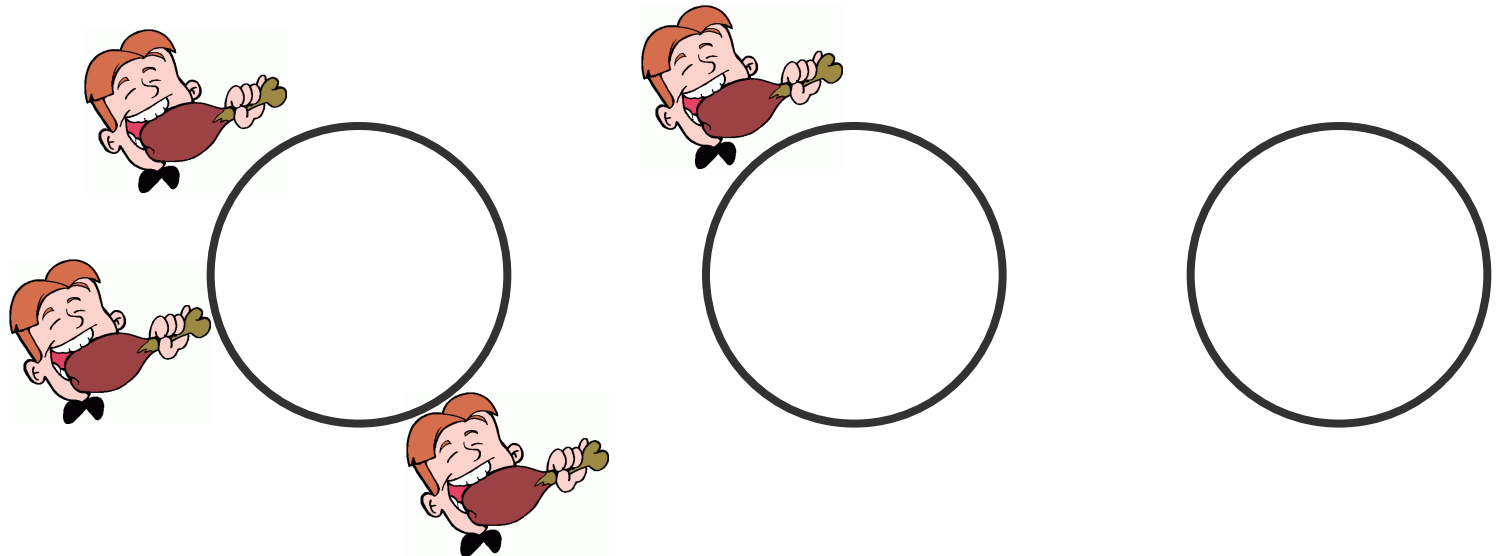
Chinese Restaurant Franchise

- Customers enter the restaurants, and sit at tables according to the Chinese Restaurant Process
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people sat at table k. He starts a new table with probability $\alpha/(n-1+\alpha)$.



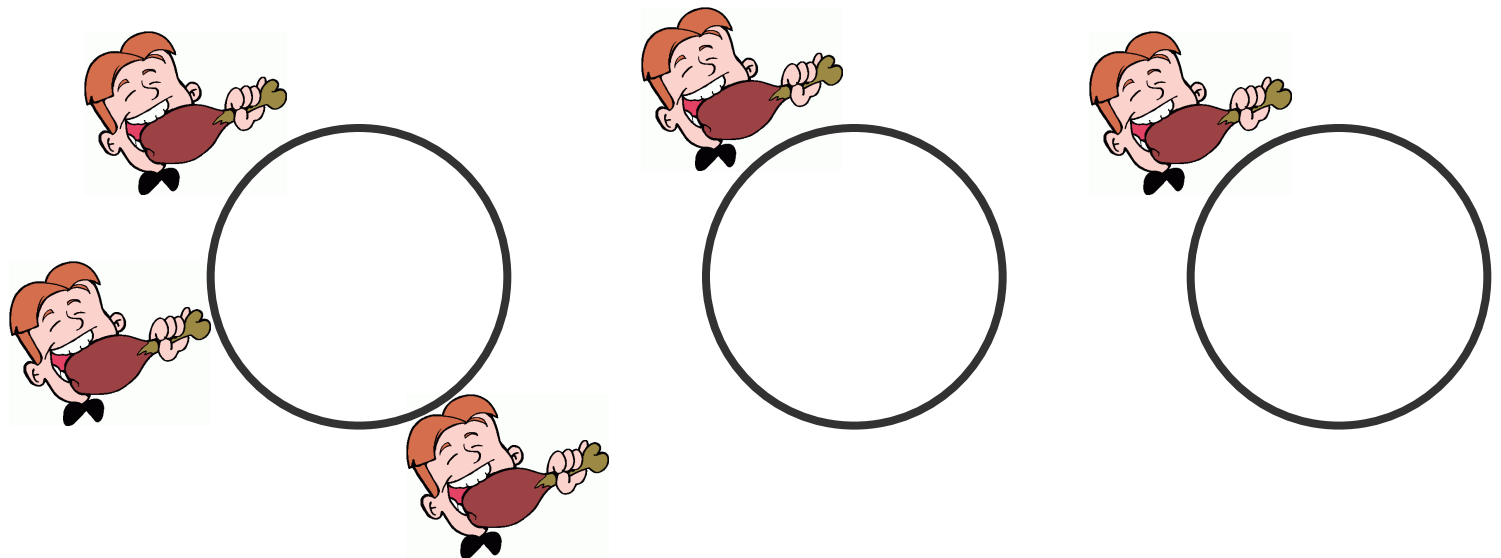
Chinese Restaurant Franchise

- Customers enter the restaurants, and sit at tables according to the Chinese Restaurant Process
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people sat at table k. He starts a new table with probability $\alpha/(n-1+\alpha)$.



Chinese Restaurant Franchise

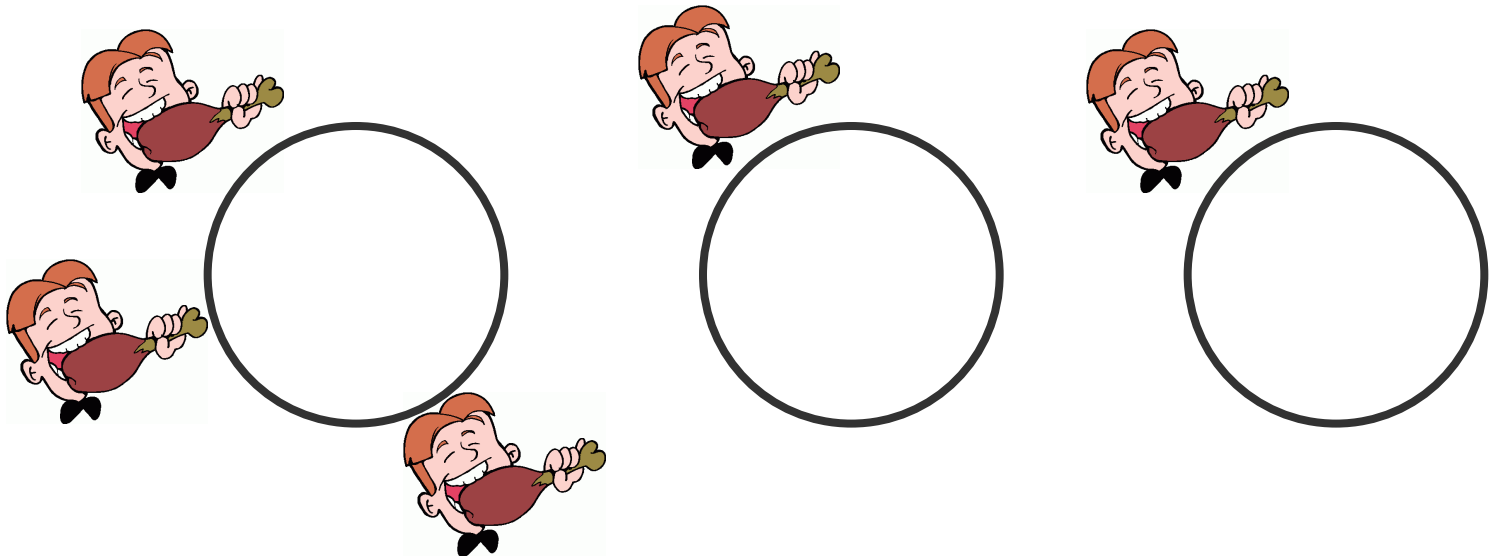
- Customers enter the restaurants, and sit at tables according to the Chinese Restaurant Process
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people sat at table k. He starts a new table with probability $\alpha/(n-1+\alpha)$.



Chinese Restaurant Franchise

- Each *table* in each restaurant picks a *dish*, with probability proportional to the number of times it has been served across *all* restaurants.

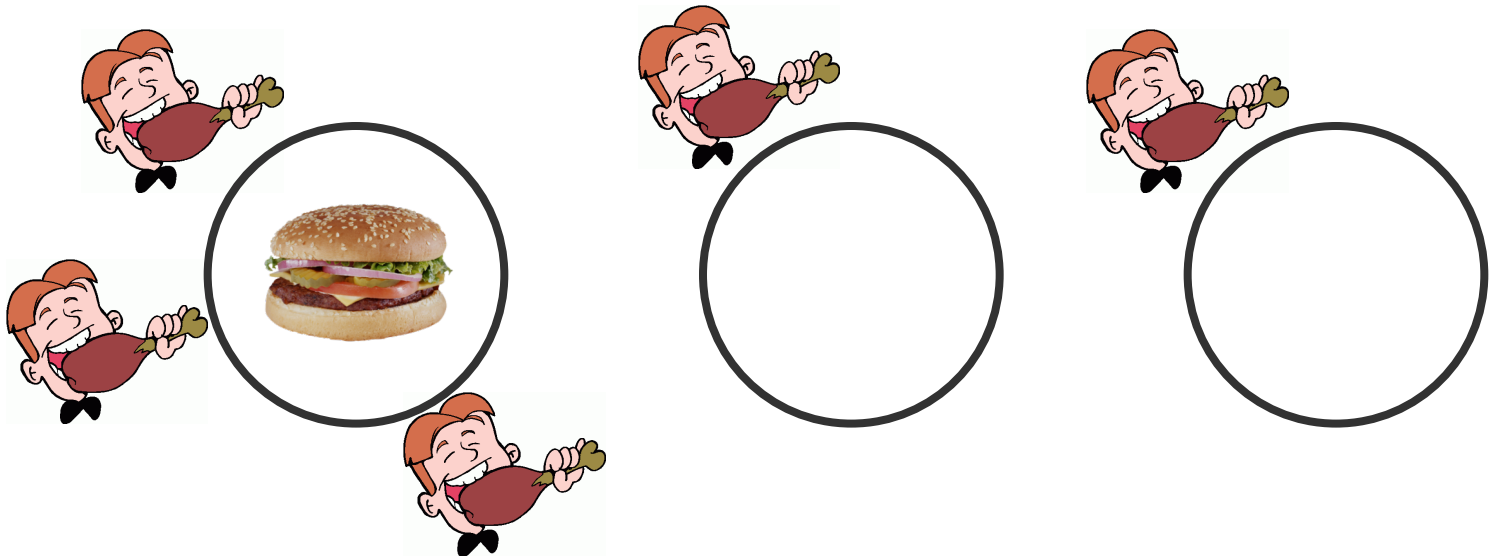
$$p(\text{table } t \text{ chooses dish } d|\text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



Chinese Restaurant Franchise

- Each *table* in each restaurant picks a *dish*, with probability proportional to the number of times it has been served across *all* restaurants.

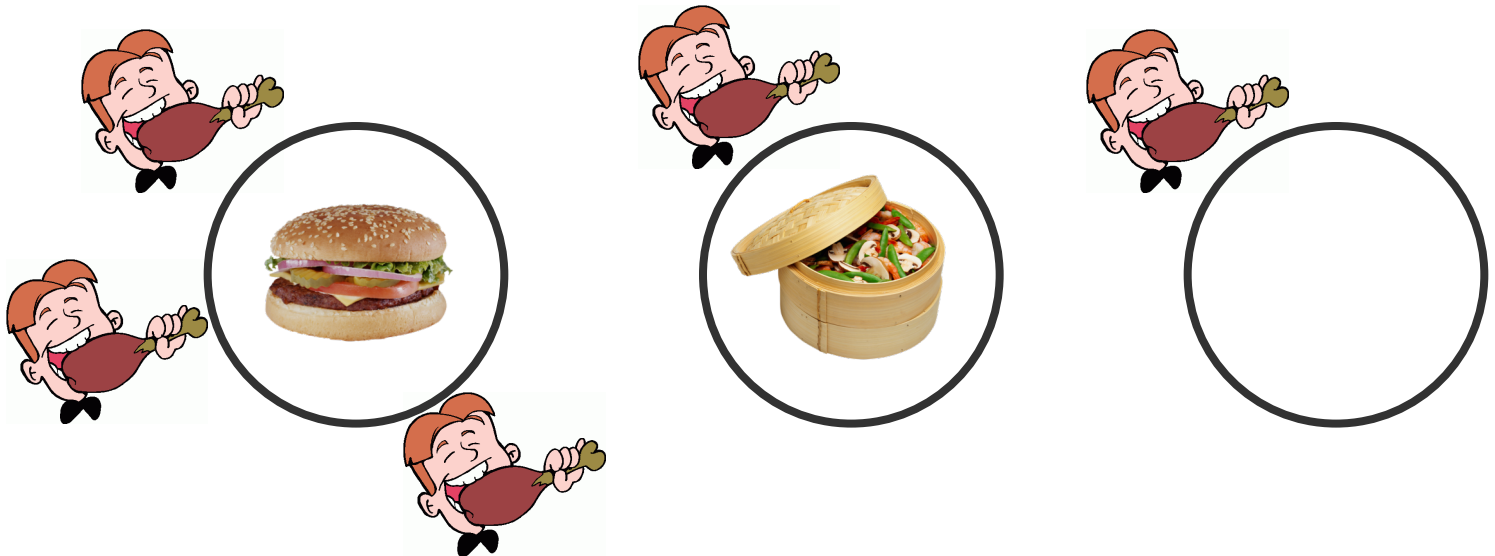
$$p(\text{table } t \text{ chooses dish } d|\text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



Chinese Restaurant Franchise

- Each *table* in each restaurant picks a *dish*, with probability proportional to the number of times it has been served across *all* restaurants.

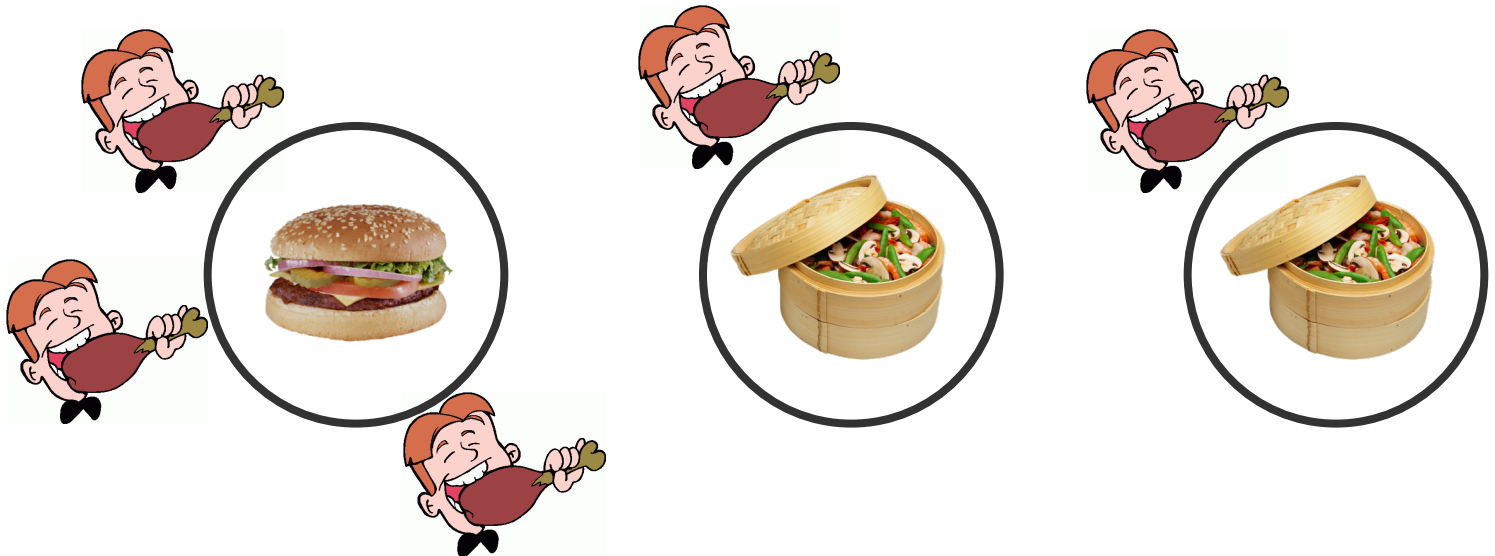
$$p(\text{table } t \text{ chooses dish } d|\text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



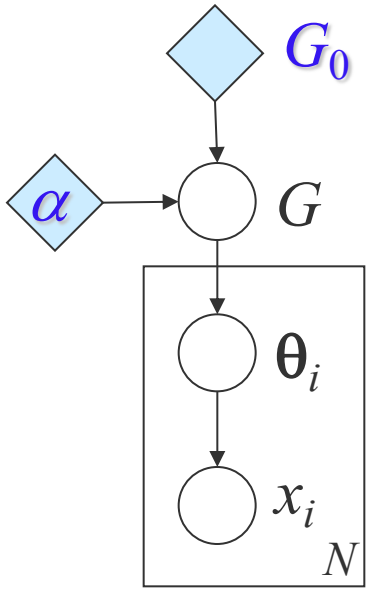
Chinese Restaurant Franchise

- Each *table* in each restaurant picks a *dish*, with probability proportional to the number of times it has been served across *all* restaurants.

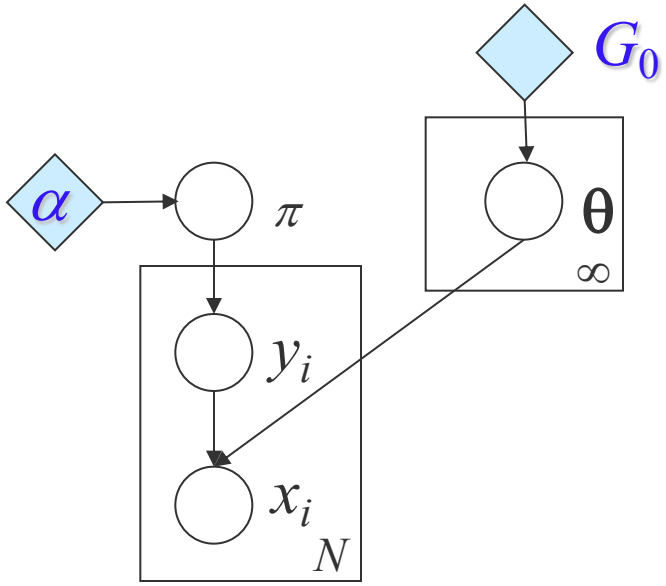
$$p(\text{table } t \text{ chooses dish } d|\text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



Recall: Graphical Model Representations of DP

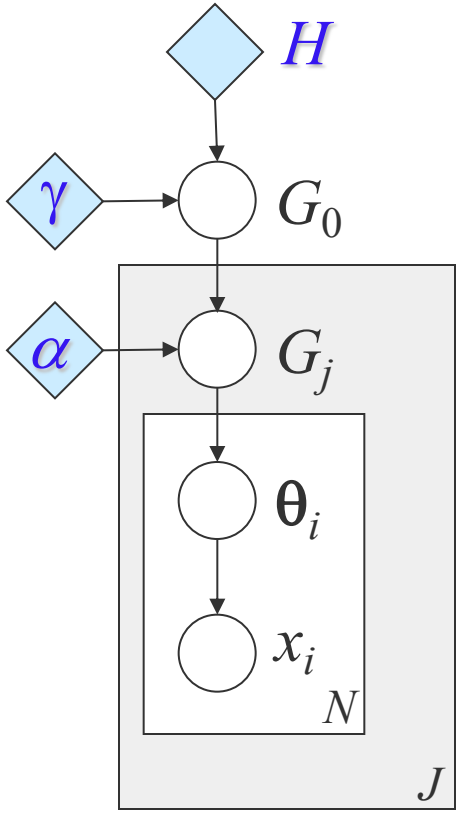


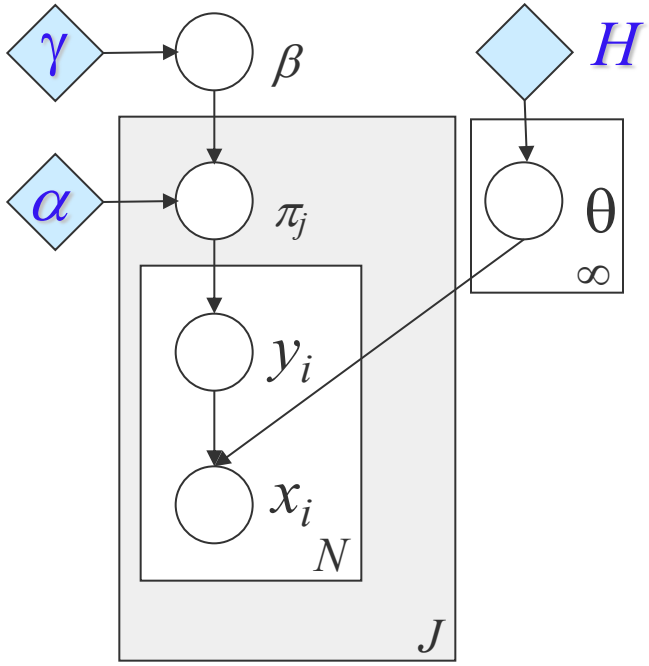
The Pólya urn construction



The Stick-breaking construction

Hierarchical DP Mixture





$\text{Stick}(\alpha,\beta)$:

$$\pi'_{jk} \sim \text{Beta}\left(\alpha\beta_k, \alpha\left(1-\sum_{l=1}^{k}\beta_l\right)\right), \quad \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}\left(1-\pi'_{jl}\right).$$

$\theta_k \sim H$

$$\beta = \text{Stick}(\gamma), G_0 = \sum_{k=1}^{\infty}\beta_k\delta(\theta_k)$$

$$\pi_j = \text{Stick}(\alpha,\beta), G_j = \sum_{k=1}^{\infty}\pi_k\delta(\theta_k)$$

An Infinite Topic Model

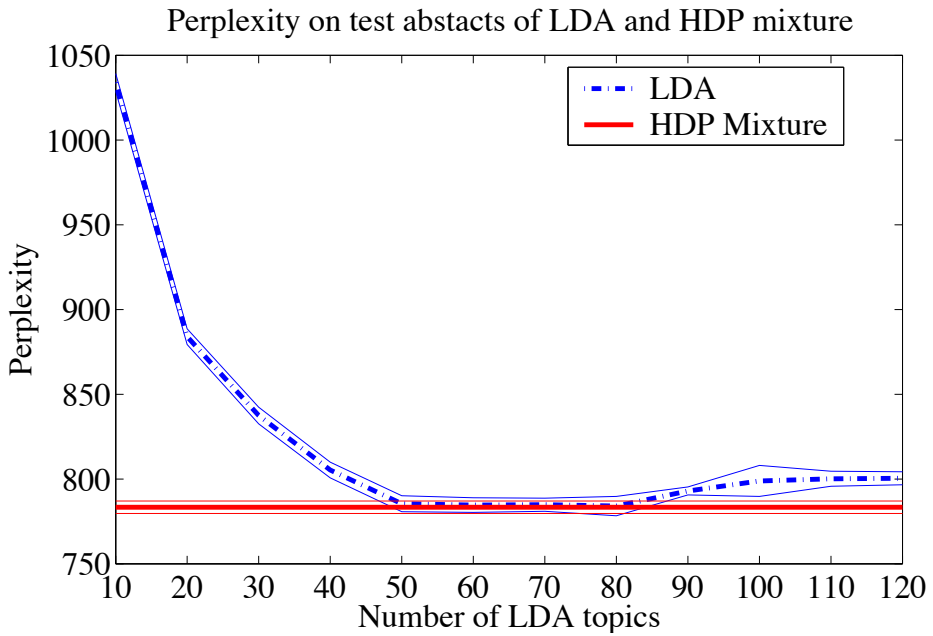
- Restaurants = documents; dishes = topics.

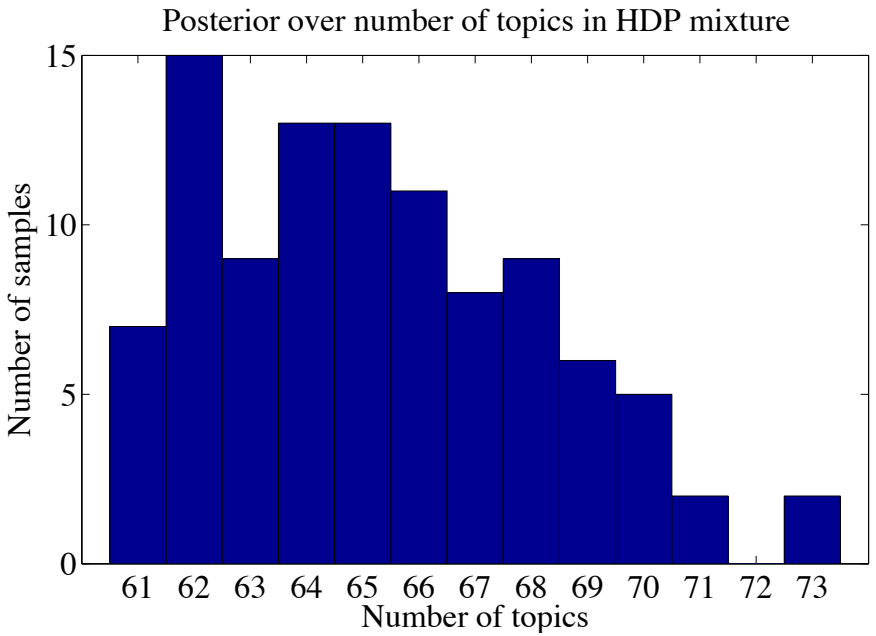
- Let H be a V-dimensional Dirichlet distribution, so a sample from H is a distribution over a vocabulary of V words.
- Sample a global distribution over topics:

$$G_0 := \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k} \sim \mathrm{DP}(\alpha, H)$$

- For each document $m=1,\ldots,M$
  - Sample a distribution over topics, $G_m \sim \mathrm{DP}(\gamma, G_0)$.
  - For each word $n=1,\ldots,N_m$
    - Sample a topic $\phi_{mn} \sim \mathrm{Discrete}(G_0)$.
    - Sample a word $w_{mk} \sim \mathrm{Discrete}(\phi_{mn})$.

The “right” number of topics





Limitations of a Simple Mixture Model

- The Dirichlet distribution and the Dirichlet process are great if we want to cluster data into non-overlapping clusters.
- However, DP/Dirichlet mixture models cannot share features between clusters.
- In many applications, data points exhibit properties of multiple latent features
  - Images contain multiple objects.
  - Actors in social networks belong to multiple social groups.
  - Movies contain aspects of multiple genres.

Latent Variable Models

- Latent variable models allow each data point to exhibit *multiple* features, to *varying degrees*.
- **Example 1:** Factor analysis

$$X = WA^T + \varepsilon$$

  - Rows of **A** = latent features
  - Rows of **W** = data-point-specific weights for these features
  - ε = Gaussian noise
- **Example 2:** LDA
  - Each document represented by a *mixture* of features

Infinite Latent Feature Models

- **Problem:** How to choose the number of features?
- **Example:** Factor analysis

$$X = WA^T + \varepsilon$$

- Each column of **W** (and row of **A**) corresponds to a feature.

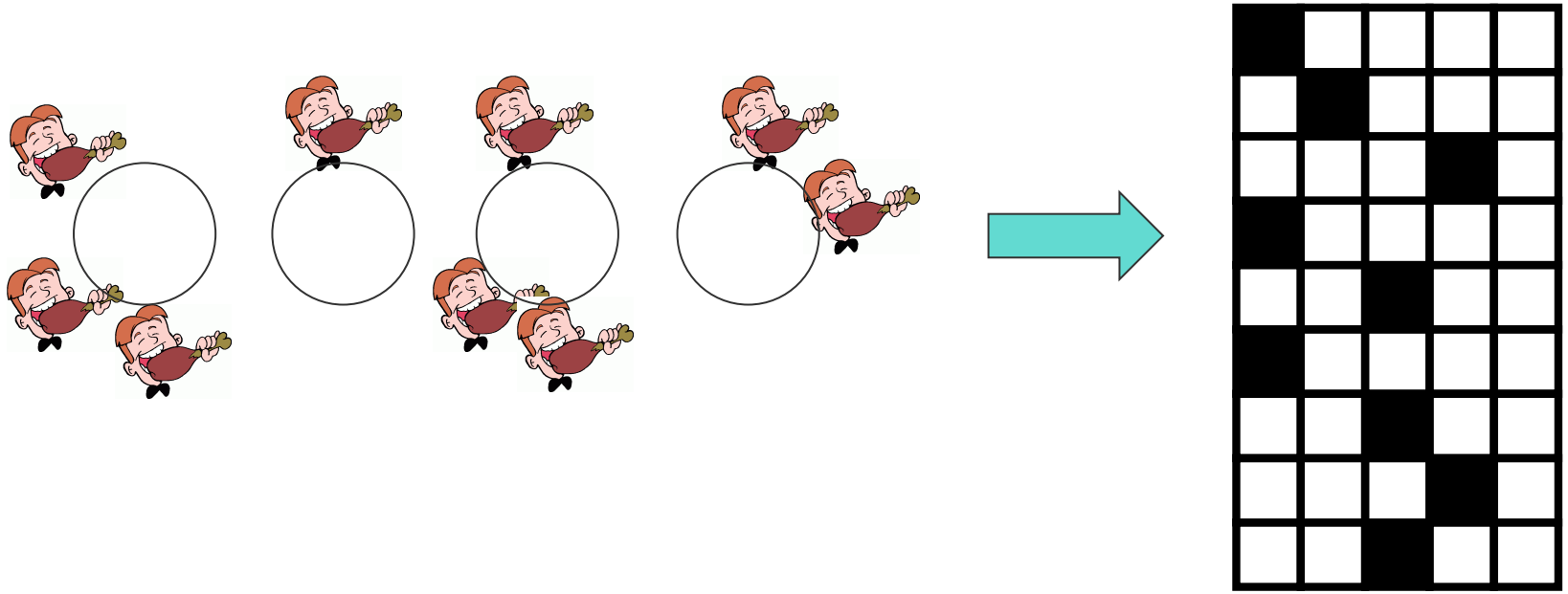
- **Question:** Can we make the number of features *unbounded* a posteriori, as we did with the DP?
- **Solution:** Allow *infinitely many* features a priori – i.e. let **W** (or **A**) have infinitely many columns (rows).

- **Problem:** We can't represent infinitely many features!
- **Solution:** Make our infinitely large matrix *sparse*.

Griffiths and Ghaharamani, 2006

The CRP: A Distribution over Binary Matrices

- Recall that the CRP gives us a distribution over *partitions* of our data.
- We can represent this as a distribution over *binary matrices*, where each row corresponds to a data point, and each column to a cluster.



A sparse, finite latent variable model

- We want a *sparse* model – so let

$$\mathbf{X} = \mathbf{W}\mathbf{A}^T + \epsilon$$
$$\mathbf{W} = \mathbf{Z} \odot \mathbf{V}$$

  for some sparse matrix Z.
- Place a *beta-Bernoulli prior* on Z:

$$\pi_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right), k = 1, \ldots, K$$
$$z_{nk} \sim \text{Bernoulli}(\pi_k), n = 1, \ldots, N.$$

A sparse, finite latent variable model

- If we integrate out the π_k, the marginal probability of a matrix Z is:

$$p(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{n=1}^{N} p(z_{nk}|\pi_k) \right) p(\pi_k) d\pi_k$$

$$= \prod_{k=1}^{K} \frac{B(m_k + \alpha/K, N - m_k + 1)}{B(\alpha/K, 1)}$$

$$= \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$

- This is *exchangeable* (doesn't depend on the order of the rows or columns)

A sparse, finite latent variable model

- If we integrate out the π_k, the marginal probability of a matrix Z is:

$$p(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{n=1}^{N} p(z_{nk}|\pi_k) \right) p(\pi_k) d\pi_k$$

$$= \prod_{k=1}^{K} \frac{B(m_k + \alpha/K, N - m_k + 1)}{B(\alpha/K, 1)}$$

$$= \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$

- How is this sparse?

An equivalence class of matrices

- We can naively take the infinite limit by taking *K* to infinity
- Because all the columns are equal in expectation, as *K* grows we are going to have more and more empty columns.
- We do not want to have to represent infinitely many empty columns!
- Define an *equivalence class* [Z] of matrices where the non-zero columns are all to the left of the empty columns.
- Let *lof(.)* be a function that maps binary matrices to *left-ordered* binary matrices – matrices ordered by the binary number made by their rows.

Left-ordered matrices

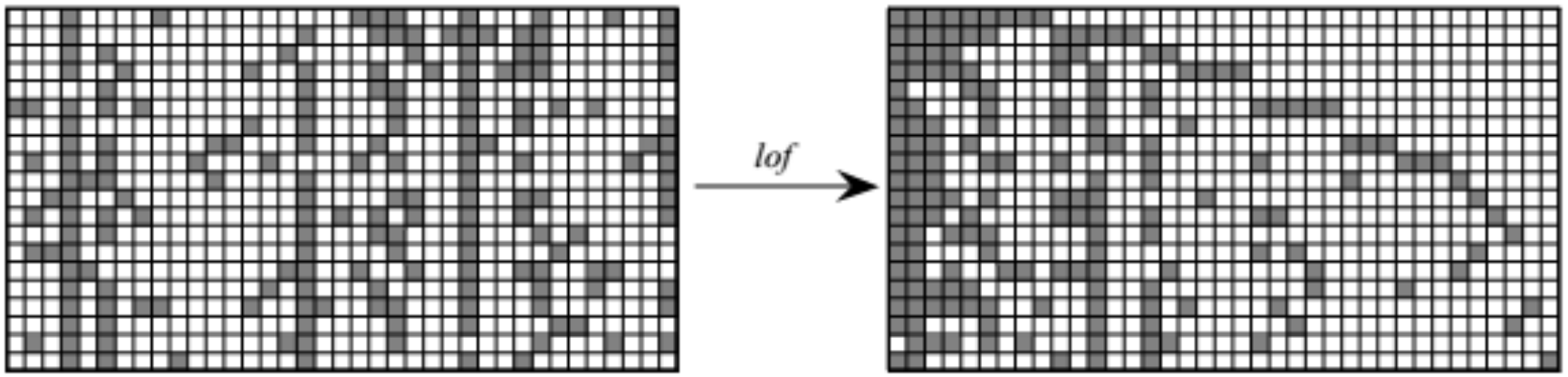


Figure 5: Binary matrices and the left-ordered form. The binary matrix on the left is transformed into the left-ordered binary matrix on the right by the function $lof(\cdot)$. This left-ordered matrix was generated from the exchangeable Indian buffet process with $\alpha = 10$. Empty columns are omitted from both matrices.

Image from Griffiths and Ghahramani, 2011

How big is the equivalence set?

- All matrices in the equivalence set [Z] are equiprobable (by exchangeability of the columns), so if we know the size of the equivalence set, we know its probability.
- Call the vector $(z_{1k}, z_{2,k}, \ldots, z_{(n-1)k})$ the *history* of feature k at data point n (a number represented in binary form).
- Let K_h be the number of features possessing history h, and let K_+ be the total number of features with non-zero history.
- The total number of lof-equivalent matrices in [Z] is

$$\binom{K}{K_0 \cdots K_{2^N-1}} = \frac{K!}{\prod_{n=0}^{2^N-1} K_n!}$$

Probability of an equivalence class of finite binary matrices

- If we know the size of the equivalence class [Z], we can evaluate its probability:

$$
\begin{aligned}
p([\mathbf{Z}]) &= \sum_{\mathbf{Z}\in[\mathbf{Z}]} p(\mathbf{Z}) \\
&= \frac{K!}{\prod_{n=0}^{2^N-1} K_n!} \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k+\alpha/K)\Gamma(N-m_k+1)}{\Gamma(N+1+\alpha/K)} \\
&= \frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \frac{K!}{K_0! K^{K_+}} \left(\frac{N!}{\prod_{j=1}^{N} j+\alpha/K} \right)^K \\
&\quad \cdot \prod_{k=1}^{K_+} \frac{(N-m_k)! \prod_{j=1}^{m_k-1} (j+\alpha/K)}{N!}
\end{aligned}
$$

Taking the infinite limit

- We are now ready to take the limit of this finite model as K tends to infinity:

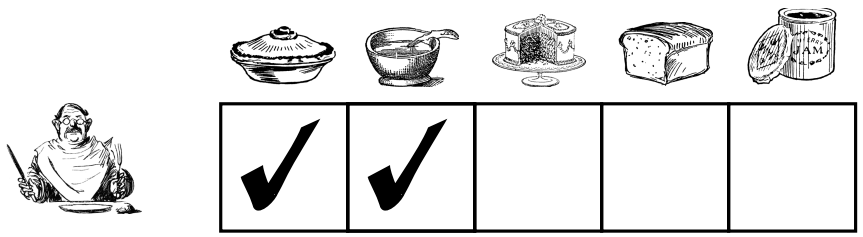
$$\frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \frac{K!}{K_0! K^{K_+}} \left(\frac{N!}{\prod_{j=1}^{N} j + \frac{\alpha}{K}} \right)^K \prod_{k=1}^{K_+} \frac{(N-m_k)! \prod_{j=1}^{m_k-1} (j + \frac{\alpha}{K})}{N!}$$

$$\downarrow K \to \infty$$

$$\frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \qquad 1 \qquad \exp\{-\alpha H_N\} \qquad \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

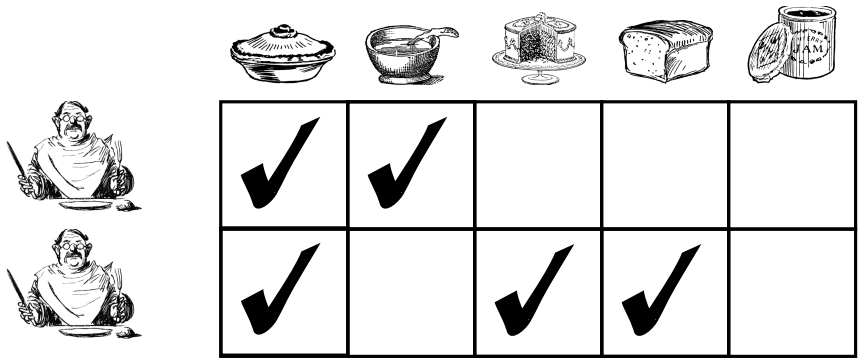
Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet.
  - He helps himself to Poisson(α) dishes.



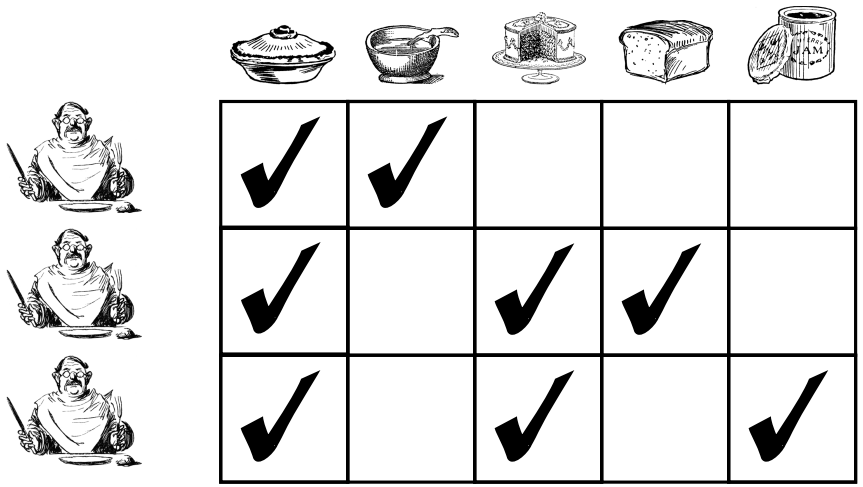
Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet.
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each dish with probability m_k/n
  - He then tries Poisson(α/n) new dishes

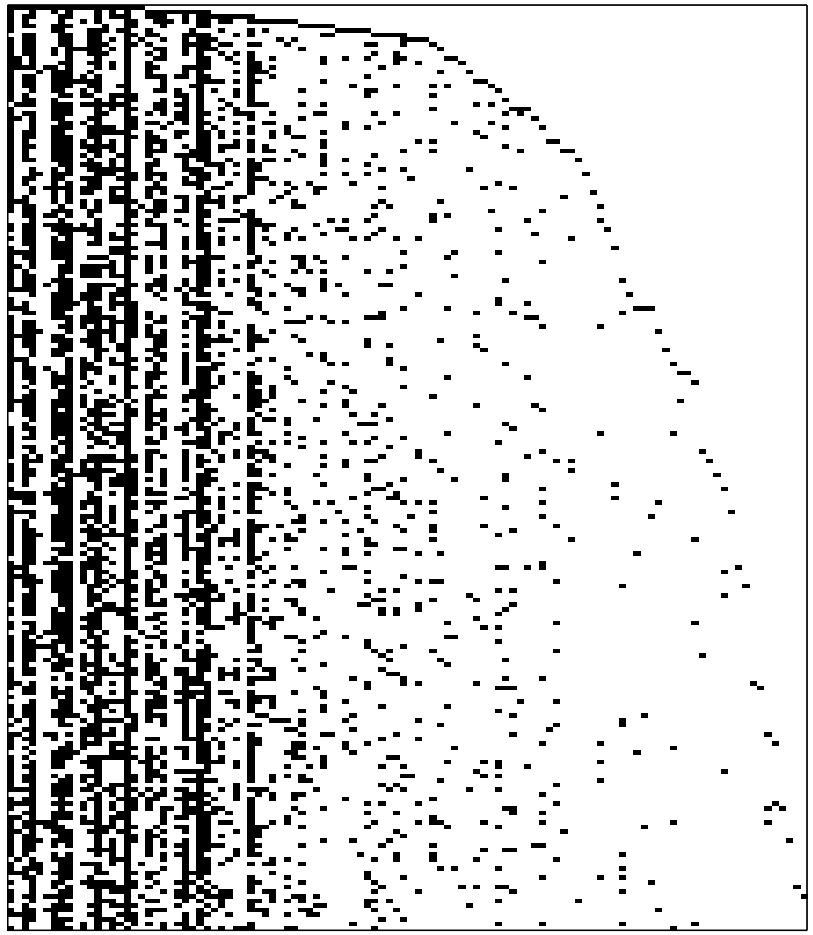


Predictive distribution: The Indian buffet process

- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each dish with probability m_k/n
  - He then tries Poisson(α/n) new dishes



Example



Proof that the IBP is lof-equivalent to the infinite beta-Bernoulli model

- What is the probability of a matrix Z?
- Let $K_1^{(n)}$ be the number of new features in the n^{th} row.

$$p(\mathbf{Z}) = \prod_{n=1}^{N} p(\mathbf{z}_n | \mathbf{z}_{1:(n-1)})$$

$$= \prod_{n=1}^{N} \text{Poisson}\left(K_1^{(n)} \middle| \frac{\alpha}{n}\right) \prod_{k=1}^{K_+} \left(\frac{\sum_{i=1}^{n-1} z_{ik}}{n}\right)^{z_{nk}} \left(\frac{n - \sum_{i=1}^{n-1} z_{ik}}{n}\right)^{1-z_{nk}}$$

$$= \prod_{n=1}^{N} \left(\frac{\alpha}{n}\right)^{K_1^{(n)}} \frac{1}{K_1^{(n)}!} e^{-\alpha/n} \prod_{k=1}^{K_+} \left(\frac{\sum_{i=1}^{n-1} z_{ik}}{n}\right)^{z_{nk}} \left(\frac{n - \sum_{i=1}^{n-1} z_{ik}}{n}\right)^{1-z_{nk}}$$

$$= \frac{\alpha^{K_+}}{\prod_{n=1}^{N} K_1^{(n)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{N - m_k)!(m_k - 1)!}{N!}$$

- If we include the cardinality of [Z], this is the same as before

Properties of the IBP

- "Rich get richer" property – "popular" dishes become more popular.
- The number of nonzero entries for each row is distributed according to Poisson(α) – due to exchangeability.
- Recall that if $x_1 \sim \text{Poisson}(\alpha_1)$ and $x_2 \sim \text{Poisson}(\alpha_2)$, then $(x_1+x_2) \sim \text{Poisson}(\alpha_1+\alpha_2)$
  - The number of nonzero entries for the whole matrix is distributed according to Poisson($N\alpha$).
  - The number of non-empty columns is distributed according to Poisson(αH_N)

Building latent feature models using the IBP

- We can use the IBP to build latent feature models with an unbounded number of features.
- Let each column of the IBP correspond to one of an *infinite* number of features.
- Each row of the IBP selects a *finite subset* of these features.
- The **rich-get-richer** property of the IBP ensures features are shared between data points.
- We must pick a *likelihood model* that determines **what the features look like** and **how they are combined**.