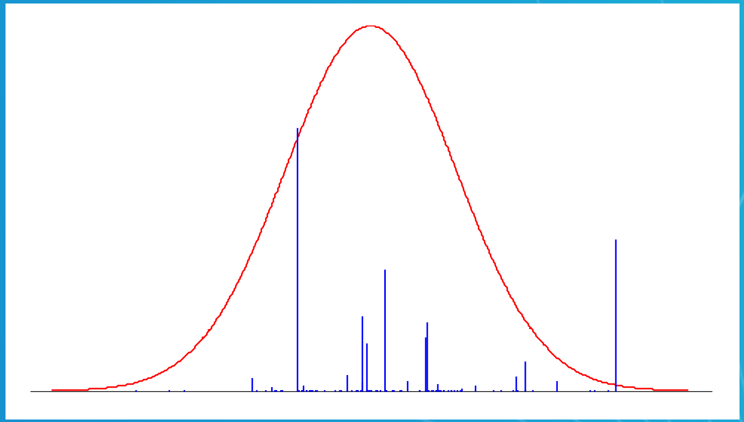
Probabilistic Graphical Models

Bayesian nonparametrics:
Dirchlet Process, Indian Buffet Process

Eric Xing

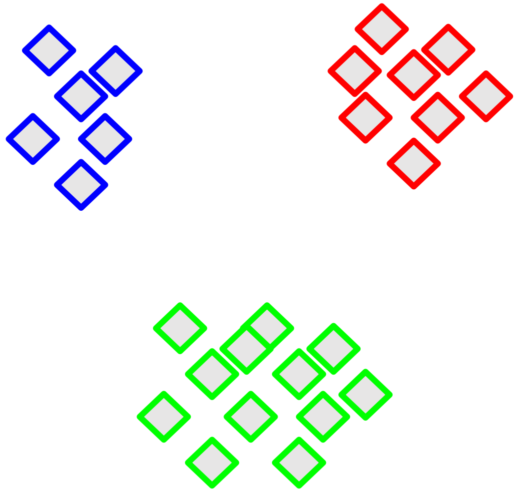
Lecture 22, April 8, 2019



Reading: see class homepage

Motivation via Clustering

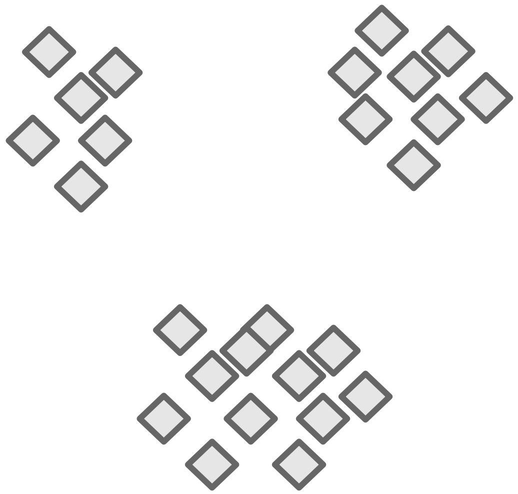
How to pick number of cluster?



Points on a 2D plane with color as attribute

Motivation via Clustering

How to pick number of cluster?



Points on a 2D plane
without color as attribute

Motivation via Clustering

How to pick number of cluster?

T=1

Streaming Data

T=2

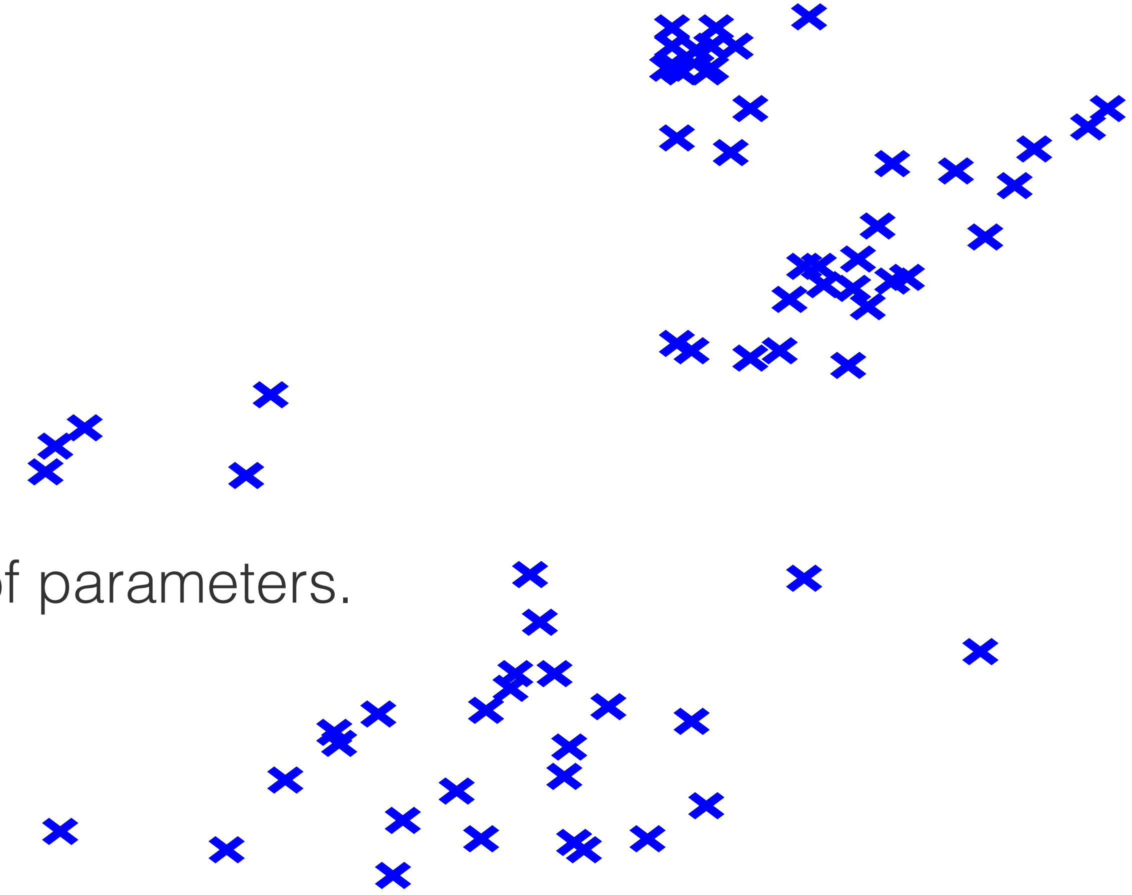
Clustered data

- How to model this data?

- Mixture of Gaussians:

$$p(x_1, \ldots, x_N | \pi, \{\mu_k\}, \{\Sigma_k\}) \\ = \underline{\prod_{n=1}^{\infty}} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_k | \mu_k, \Sigma_k) \right)$$

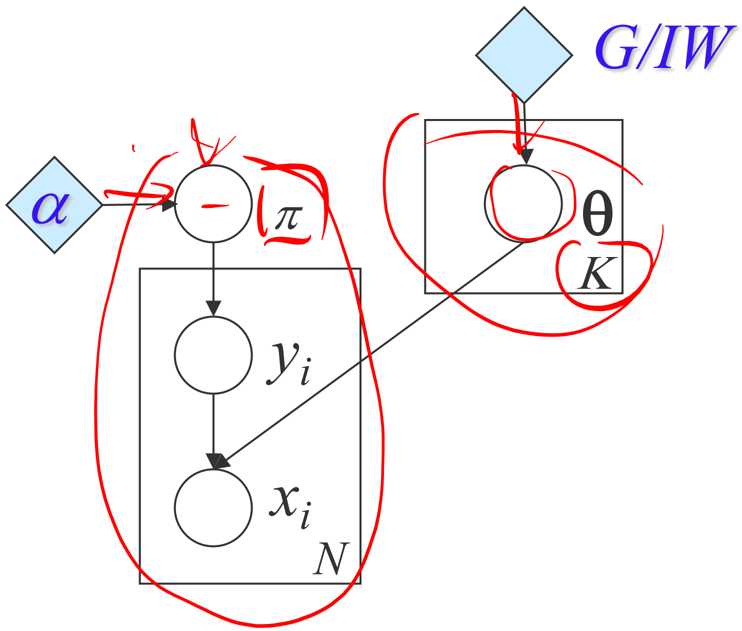
- Parametric model: Fixed finite number of parameters.



Bayesian finite mixture model

- How to choose the mixing weights and mixture parameters?
- Bayesian choice: Put a prior on them and integrate out:

$$
\begin{aligned}
&p(x_1, \ldots, x_N) \\
&= \int\int\int \left(\prod_{n=1}^{\infty} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_k | \mu_k, \Sigma_k) \right) \\
&\quad p(\pi) p(\mu_{1:K}) p(\Sigma_{1:K}) d\pi d\mu_{1:K} d\Sigma_{1:K}
\end{aligned}
$$



- Where possible, use conjugate priors
  - Gaussian/inverse Wishart for mixture parameters
  - What to choose for mixture weights?

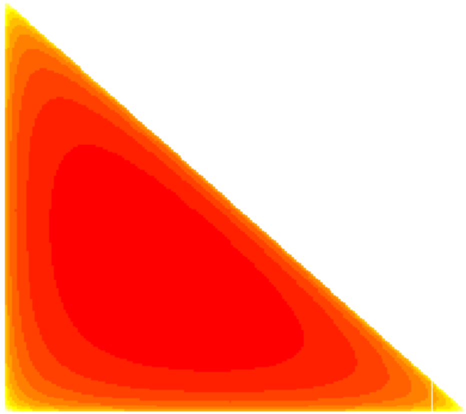
The Dirichlet distribution

- The Dirichlet distribution is a distribution over the (*K*-*1*)-dimensional simplex.
- It is parametrized by a *K*-dimensional vector $(\alpha_1, \ldots, \alpha_K)$ such that $\alpha_k \geq 0, k = 1, \ldots, K$ and $\sum_k \alpha_k > 0$
- Its distribution is given by

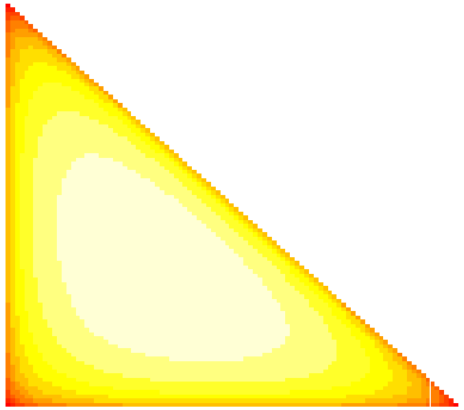
$$\frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Samples from the Dirichlet distribution

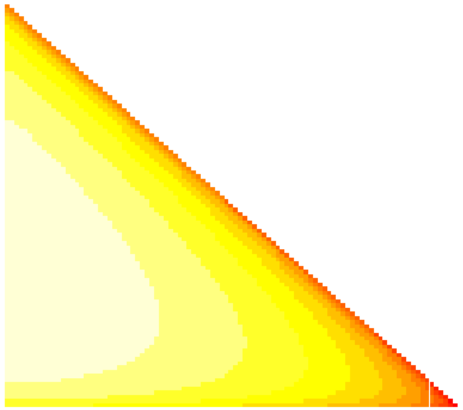
- If $\pi \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ then $\pi_k \geq 0$ for all *k*, and $\sum_{k=1}^{K} \pi_k = 1.$
- Expectation: $\mathbb{E}\Big[(\pi_1, \ldots, \pi_K)\Big] = \frac{(\alpha_1, \ldots, \alpha_K)}{\sum_k \alpha_k}$



$\alpha = (0.01, 0.01, 0.01)$



$\alpha = (100, 100, 100)$



$\alpha = (5, 50, 100)$

Conjugacy to the multinomial

- If $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ and $x_n \overset{iid}{\sim} \theta$

$$
\begin{aligned}
p(\pi | x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n | \pi) p(\pi) \\
&= \left(\frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \left(\frac{n!}{m_1! \ldots m_K!} \pi_1^{m_1} \ldots \pi_K^{m_K} \right) \\
&\propto \frac{\prod_{k=1}^K \Gamma(\alpha_k + m_k)}{\Gamma(\sum_{k=1}^K \alpha_k + m_k)} \prod_{k=1}^K \pi_k^{\alpha_k + m_k - 1} \\
&= \text{Dirichlet}(\pi | \alpha_1 + m_1, \ldots, \alpha_K + m_K)
\end{aligned}
$$

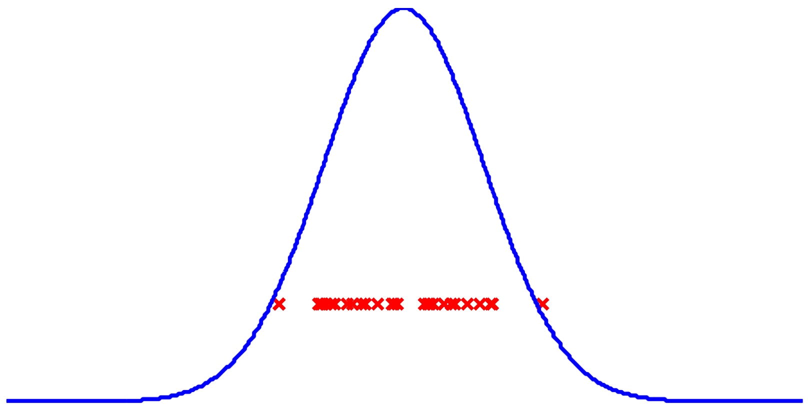
Distributions over distributions

$\pi \quad \theta = [\mu \mid \Sigma]$

- The Dirichlet distribution is a distribution over positive vectors that sum to one.
- We can associate each entry with a set of parameters
  - e.g. finite mixture model: each entry associated with a mean and covariance.
- In a Bayesian setting, we want these parameters to be *random.*
- We can combine the distribution over probability vectors with a distribution over parameters to get a **distribution over distributions over parameters.**

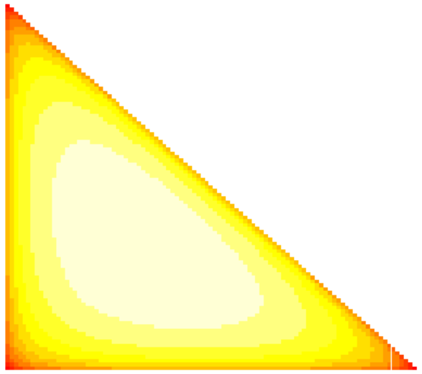
Example: finite mixture model

- Gaussian distribution: distribution over means.
  - Sample from a Gaussian is a real-valued number/vector.



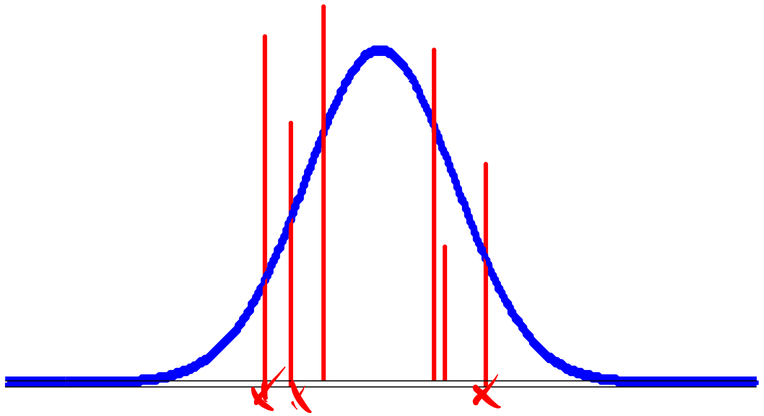
Example: finite mixture model

- Gaussian distribution: distribution over means.
  - Sample from a Gaussian is a real-valued number/vector.
- Dirichlet distribution:
  - Sample from a Dirichlet distribution is a probability vector.



Example: finite mixture model

- Dirichlet prior
  - Each element of a Dirichlet-distributed vector is associated with a parameter value drawn from some distribution.
  - Sample from a Dirichlet prior is a probability distribution over parameters.



Properties of the Dirichlet distribution (collapsing)

- Relationship to gamma distribution: If $\eta_k \sim \text{Gamma}(\alpha_k, 1)$,

$$\frac{(\eta_1, \ldots, \eta_K)}{\sum_k \eta_k} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

- If $\eta_1 \sim \text{Gamma}(\alpha_1, 1)$ and $\eta_2 \sim \text{Gamma}(\alpha_2, 1)$ then

$$\eta_1 + \eta_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)$$

- Therefore, if $(\pi_1 \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ then

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

Properties of the Dirichlet distribution (splitting)

- The beta distribution is a Dirichlet distribution on the 1-simplex.
- Let $(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ and $\theta \sim \text{Beta}(\alpha_1 b, \alpha_1(1-b)), 0 < b < 1.$
- Then

$$(\pi_1\theta, \pi_1(1-\theta), \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1(1-b_1), \alpha_2, \ldots, \alpha_K)$$

- More generally, if $\theta \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_N), \sum_i b_i = 1.$ then

$$(\pi_1\theta_1, \ldots, \pi_1\theta_N, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \ldots, \alpha_1 b_N, \alpha_2, \ldots, \alpha_K)$$

Properties of the Dirichlet distribution

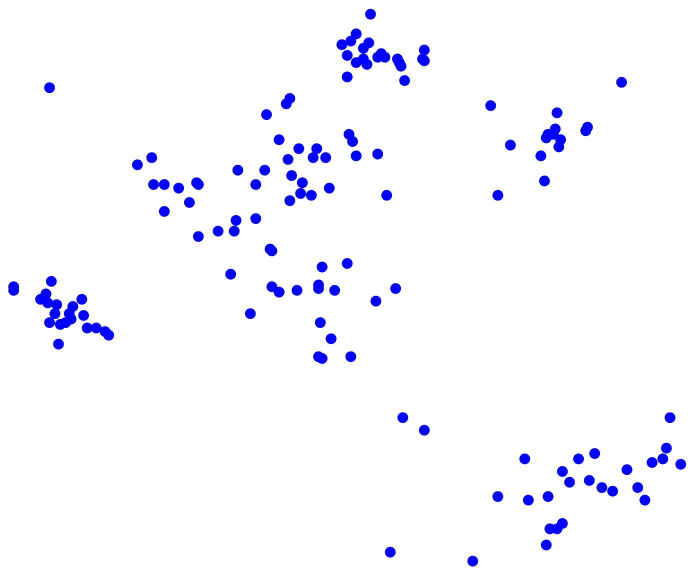
- Renormalization:

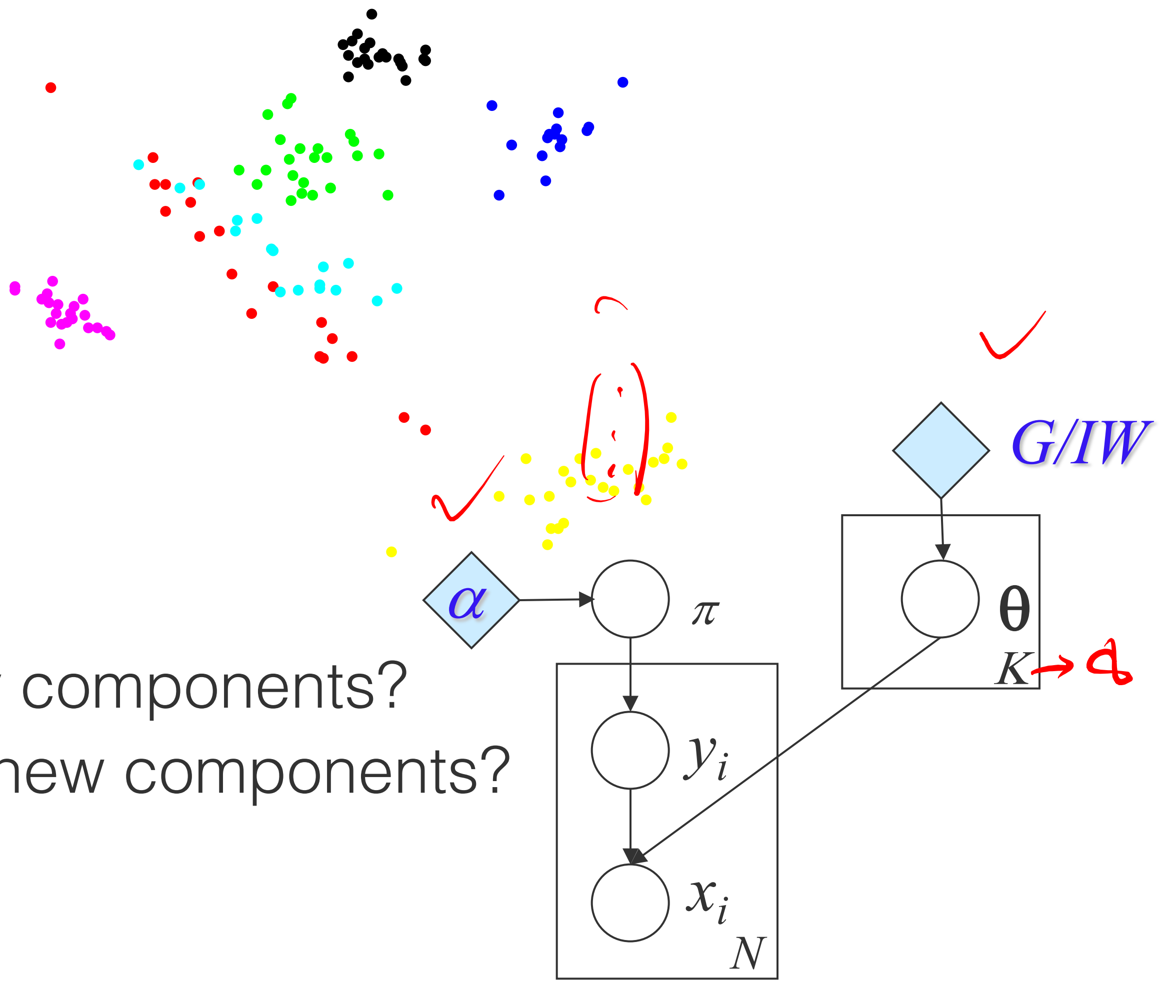
  If $$(\pi_1 \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

  then $$\frac{(\pi_2, \ldots, \pi_K)}{\sum_{k=1}^K \pi_k} \sim ?$$

$$\frac{(\pi_2, \ldots, \pi_K)}{\sum_{k=1}^K \pi_k} \sim \text{Dirichlet}(\alpha_2, \ldots, \alpha_K)$$

Choosing the number of clusters





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

Parametric vs nonparametric

Parametric model:

- Assumes all data can be represented using a fixed, finite number of parameters.
  - Mixture of K Gaussians, polynomial regression.

Nonparametric model:

- Number of parameters can grow with sample size.
- Number of parameters may be random.
  - Kernel density estimation.

Bayesian nonparametrics:

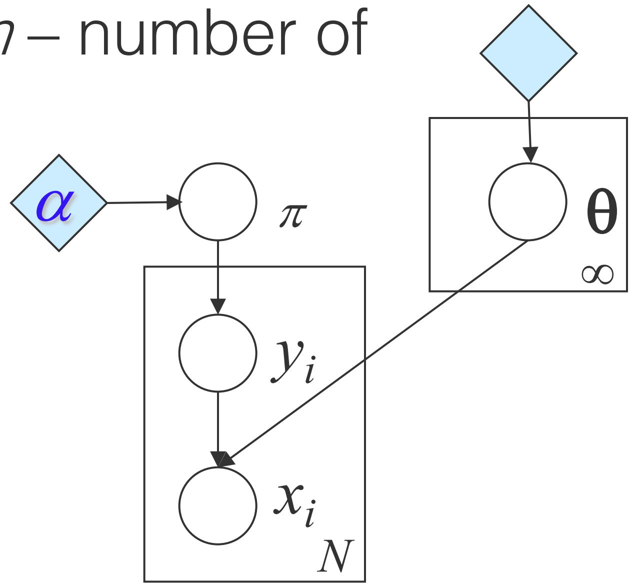
- Allow an *infinite* number of parameters *a priori.*
- A finite data set will only use a finite number of parameters.
- Other parameters are integrated out.

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.



Constructing an appropriate prior

- Start off with $$\pi^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$$
- Split each component according to the splitting rule:

$$\theta_1^{(2)}, \theta_2^{(2)} \overset{iid}{\sim} \text{Beta}\left(\frac{\alpha}{2} \cdot \frac{1}{2}, \frac{\alpha}{2} \cdot \frac{1}{2}\right)$$

$$\pi^{(4)} = (\theta_1^{(2)}\pi_1^{(2)}, (1-\theta_1^{(2)})\pi_1^{(2)}, \theta_2^{(2)}\pi_2^{(2)}, (1-\theta_2^{(2)})\pi_2^{(2)})$$

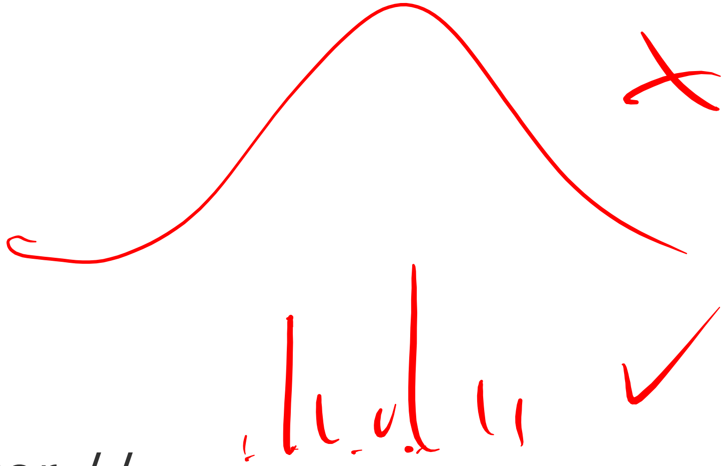
$$\sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

α

- Repeat to get $$\pi^{(K)} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$
- As $K \to \infty$, we get a vector with infinitely many components

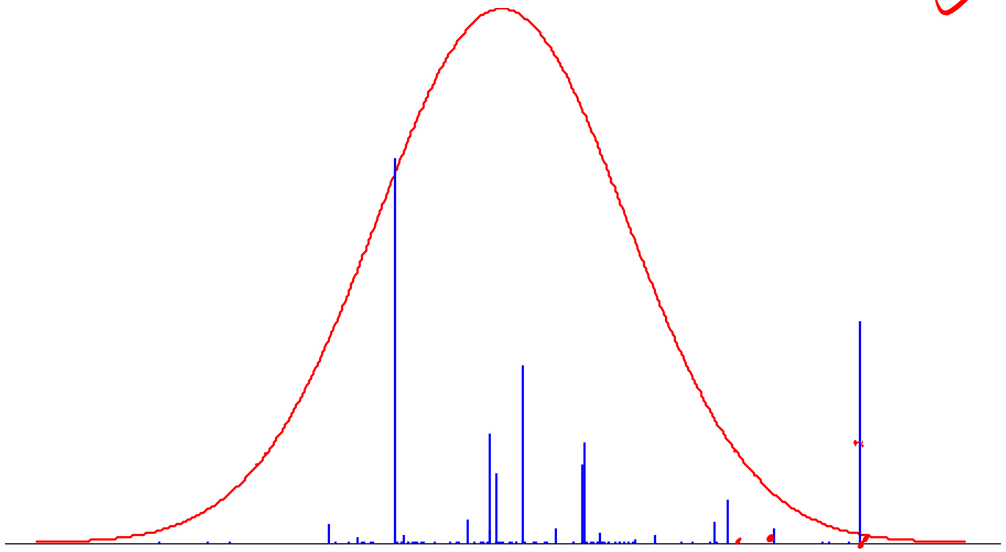
The Dirichlet process

- Let H be a distribution on some space Ω – e.g. a Gaussian distribution on the real line.
- Let $\boxed{\pi} \sim \lim_{K\to\infty} \mathrm{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$
- For $k = 1, \dots, \infty$ let $\theta_k \sim H$.
- Then $G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ is an infinite distribution over H.
- We write $G \sim \mathrm{DP}(\alpha, H)$



Samples from the Dirichlet process

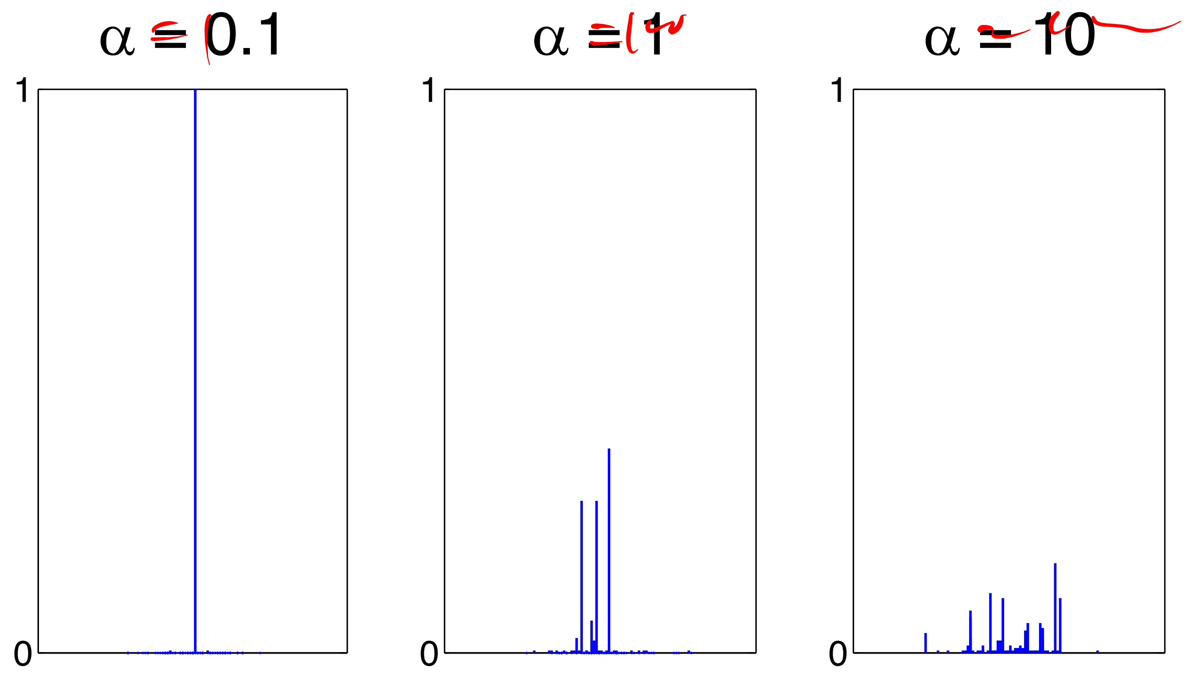
- Samples from the Dirichlet process are *discrete*.
- We call the point masses in the resulting distribution, *atoms*.



- The *base measure* H determines the *locations* of the atoms.

Samples from the Dirichlet process

- The *concentration parameter* α determines the distribution over atom sizes.
- Small values of α give *sparse* distributions.

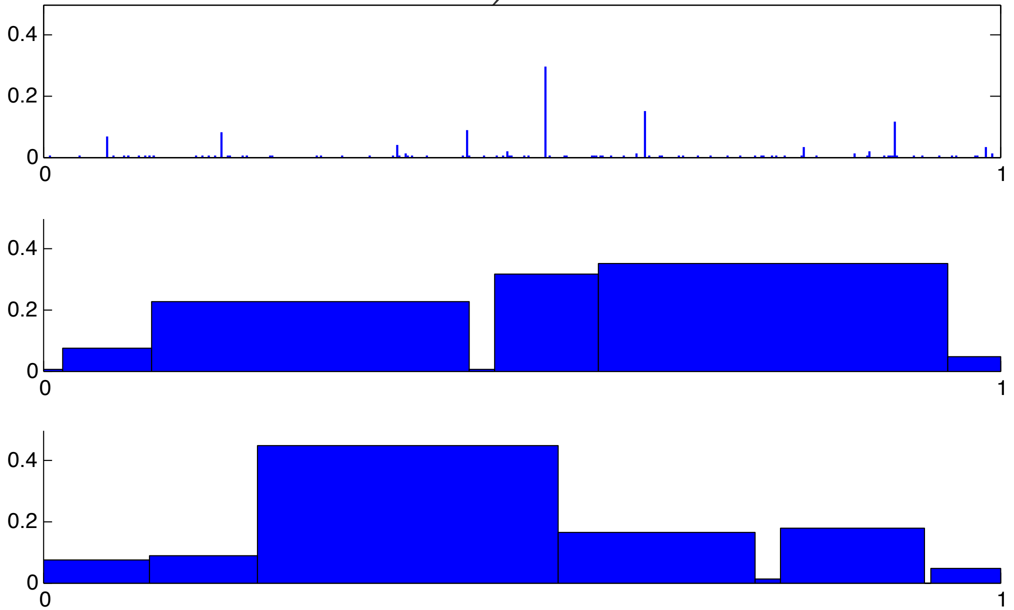


Properties of the Dirichlet process

- For any partition $A_1,\ldots,A_K$ of Ω (imagine different colors from a possibly infinite color library), the total mass assigned to each partition is distributed according to

$$\sim Dir(\alpha H(A_1),\ldots,\alpha H(A_K))$$

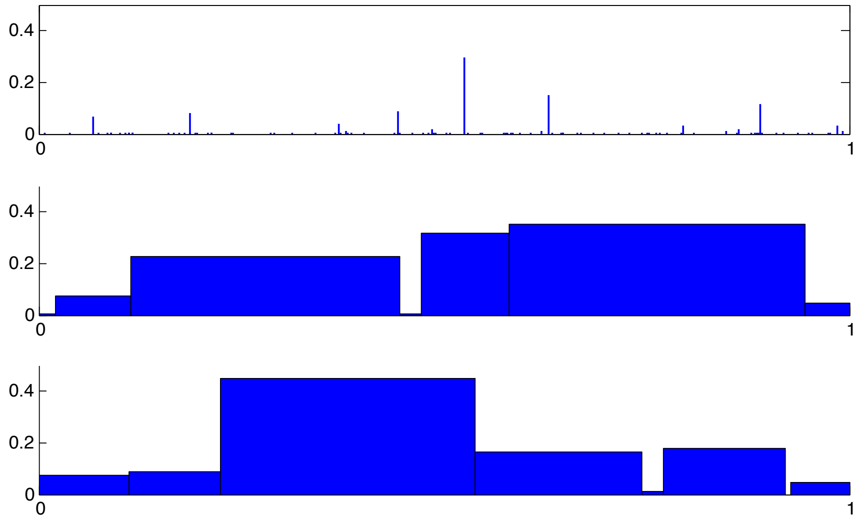
$\alpha H(A_K)$



Definition: Finite marginals

- A Dirichlet process is the unique distribution over probability distributions on some space Ω, such that for any finite partition $A_1, \ldots, A_K$ of Ω,

$$(P(A_1), \ldots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)).$$



[Ferguson, 1973]

Conjugacy of the Dirichlet process

- Let $A_1,\ldots,A_K$ be a partition of Ω, and let H be a measure on Ω. Let $P(A_k)$ be the mass assigned by $G \sim \mathrm{DP}(\alpha, H)$ to partition A_k. Then

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

- If we see an observation in the J^{th} segment, then

$$(P(A_1),\ldots,P(A_j),\ldots,P(A_K)|X_1 \in A_j)$$
$$\sim \mathrm{Dirichlet}(\alpha H(A_1),\ldots,\alpha H(A_j)+1,\ldots,\alpha H(A_K)).$$

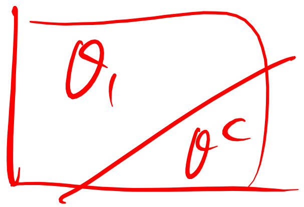
- This must be true for *all possible partitions of* Ω.
- This is only possible if the posterior of G, given an observation x, is given by

$$G|X_1 = x \sim \mathrm{DP}\left(\alpha+1, \frac{\alpha H + \delta_x}{\alpha+1}\right)$$

Predictive distribution

- The Dirichlet process clusters observations.
- A new data point can either join an existing cluster, or start a new cluster.
- Question: What is the predictive distribution for a new data point?
- Assume H is a continuous distribution on Ω. This means for every point θ in Ω, $P_H(\theta) = 0$.
- First data point:
  - Start a new cluster.
  - Sample a parameter θ_1 for that cluster.

Predictive distribution



X_1

- We have now split our parameter space in two: the singleton θ_1, and everything else.
- Let π_1 be the size of atom at θ_1.
- The combined mass of all the other atoms is $\pi_* = 1 - \pi_1$.
- *A priori*, $(\pi_1, \pi_*) \sim \text{Dirichlet}(0, \alpha)$

$\alpha H(\theta_1)$ $\alpha H(\theta^c)$

- A posteriori, $(\pi_1, \pi_*)|X_1 = \theta_1 \sim \text{Dirichlet}(1, \alpha)$

Predictive distribution

$P(X_2 = \theta_k \mid \pi_1, \pi_*, \theta_1)$

- If we integrate out π_1 we get

$$\begin{aligned}
P(X_2 = \theta_k | X_1 = \theta_1) &= \int P(X_2 = \theta_k | (\pi_1, \pi_*)) P((\pi_1, \pi_* | X_1 = \theta_1) d\pi_1 \\
&= \int \pi_k \text{Dirichlet}((\pi_1, 1 - \pi_1) | 1, \alpha) d\pi_1 \\
&= \mathbb{E}_{\text{Dirichlet}(1,\alpha)}[\pi_k] \\
&= \begin{cases} \frac{1}{1+\alpha} & \text{if } k = 1 \\ \frac{\alpha}{1+\alpha} & \text{for new } k. \end{cases}
\end{aligned}$$

$\theta_3 \mid X_1\ X_2$?

$(\pi_1\ \pi_2\ \pi_*)$

Predictive distribution

- Lets say we choose to start a new cluster, and sample a new parameter $\theta_2 \sim H$. Let π_2 be the size of the atom at θ_2.
- A posteriori, $(\pi_1, \pi_2, \pi_*)|X_1 = \theta_1, X_2 = \theta_2 \sim \text{Dirichlet}(1, \alpha).$
- If we integrate out $\pi = (\pi_1, \pi_2, \pi_*)$, we get

$$
\begin{aligned}
P(X_3 = \theta_k | X_1 = \theta_1, X_2 = \theta_2) \\
&= \int P(X_3 = \theta_k|\pi)P(\pi|X_1 = \theta_1, X_2 = \theta_2)d\pi \\
&= \mathbb{E}_{\text{Dirichlet}(1,1,\alpha)}[\pi_k] \\
&= \begin{cases} \frac{1}{2+\alpha} & \text{if } k = 1 \\ \frac{1}{2+\alpha} & \text{if } k = 2 \\ \frac{\alpha}{2+\alpha} & \text{for new } k. \end{cases}
\end{aligned}
$$

Predictive distribution

π_k

- In general, if m_k is the number of times we have seen X_i=k, and K is the total number of observed values,

$$P(X_{n+1} = \theta_k | X_1, \ldots, X_n) = \int P(X_{n+1} = \theta_k | \pi) P(\pi | X_1, \ldots, X_n) d\pi$$

$$= \mathbb{E}_{\text{Dirichlet}(m_1, \ldots, m_K, \alpha)}[\pi_k]$$

$$= \begin{cases} \frac{m_k}{n+\alpha} & \text{if } k \leq K \\ \frac{\alpha}{n+\alpha} & \text{for new cluster.} \end{cases}$$

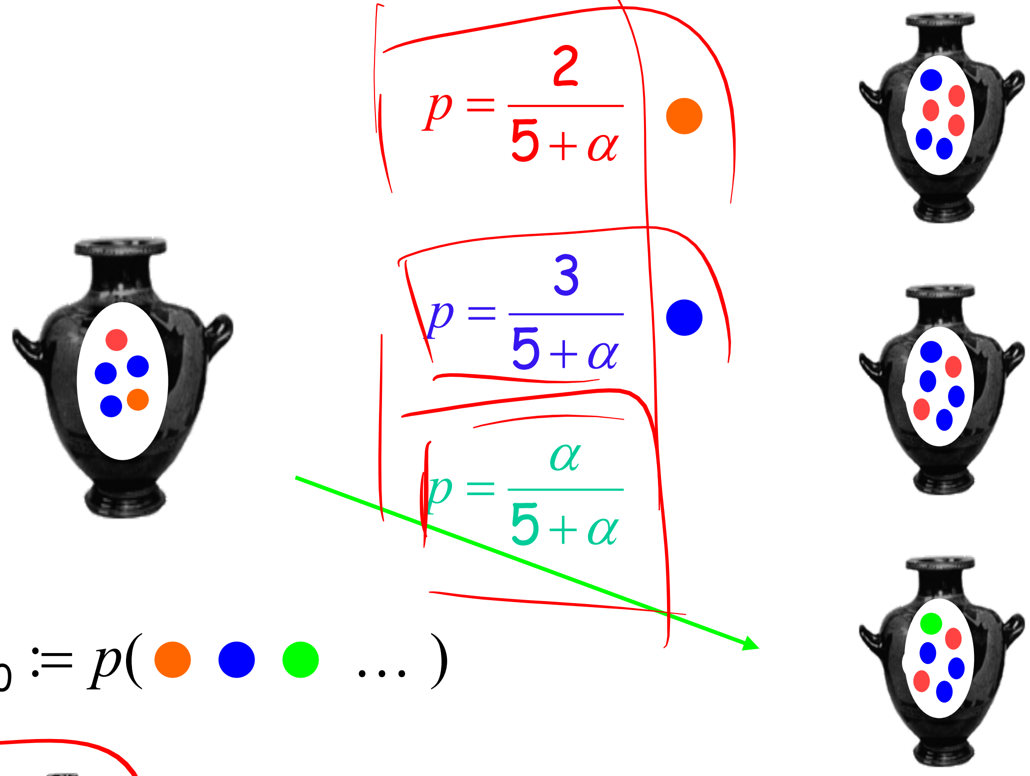
$\Downarrow$

$n_1 \; m_1 \cdots m_k$

α

- We tend to see observations that we have seen before – *rich-get-richer property.*
- We can always add new features – *nonparametric.*

DP – a Pólya urn Process



$$p = \frac{2}{5+\alpha}$$

$$p = \frac{3}{5+\alpha}$$

$$p = \frac{\alpha}{5+\alpha}$$

$$G_0 := p(\bullet \; \bullet \; \bullet \; \ldots)$$

Joint: $G(\ ) \sim DP(\alpha G_0)$

Marginal: $\phi_i \mid \phi_{-i}, \alpha, G_0 \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha} \delta_{\phi_k} + \frac{\alpha}{i-1+\alpha} G_0 .$

- Self-reinforcing property
- exchangeable partition of samples

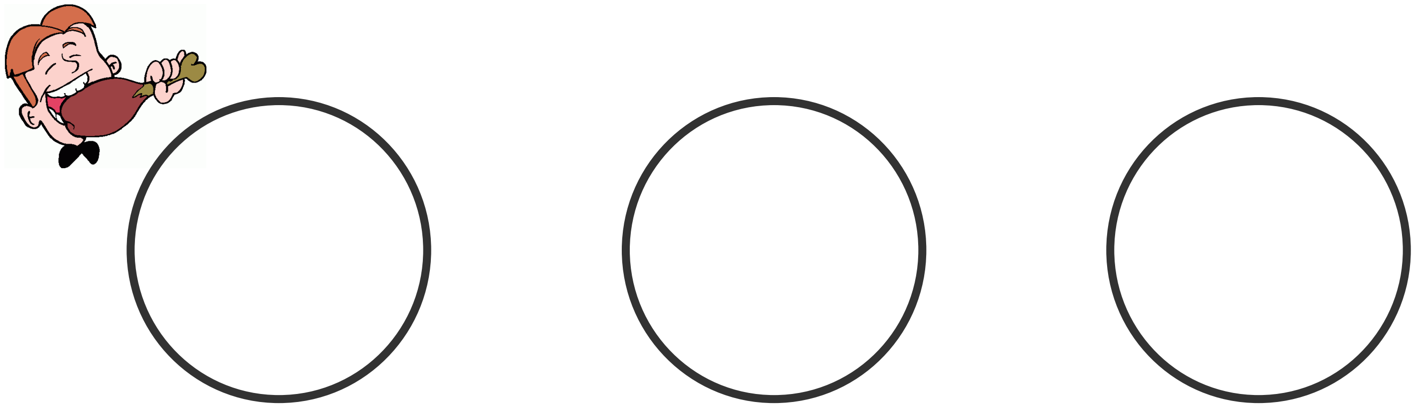
Polya urn scheme

- The resulting distribution over data points can be thought of using the following urn scheme.
- An urn initially contains a black ball of mass α.
- For n=1,2,… sample a ball from the urn with probability proportional to its mass.
- If the ball is black, choose a previously unseen color, record that color, and return the black ball plus a unit-mass ball of the new color to the urn.
- If the ball is not black, record it's color and return it, plus another unit-mass ball of the same color, to the urn

[Blackwell and MacQueen,1973]

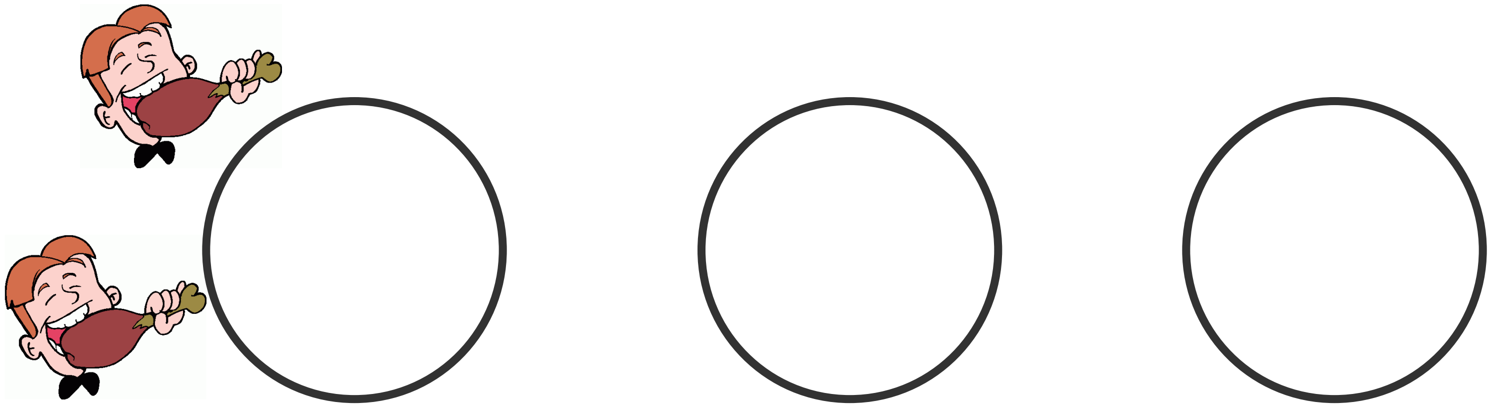
Chinese restaurant process

- The distribution over partitions can be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.



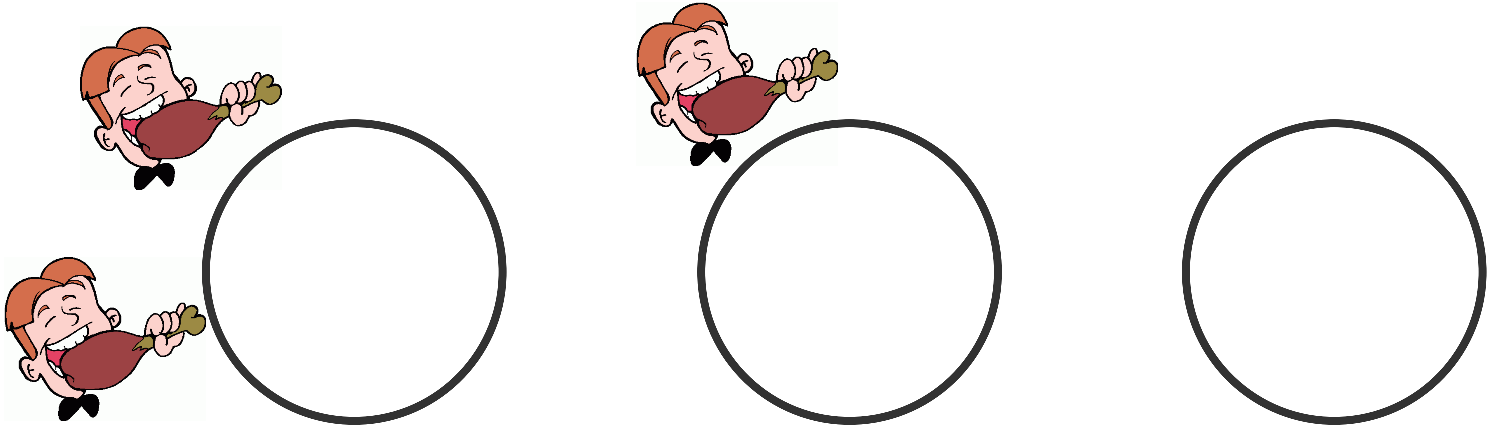
Chinese restaurant process

- 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. 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 $\alpha/(n-1+\alpha)$.



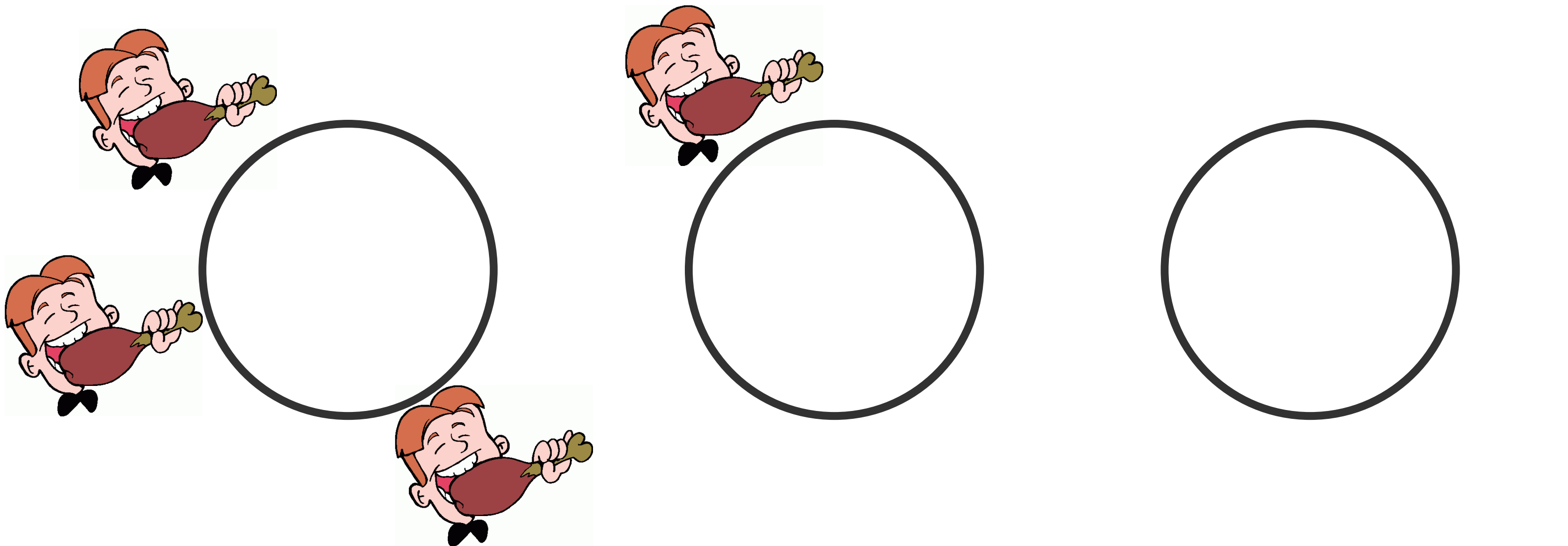
Chinese restaurant process

- 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. 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 $\alpha/(n-1+\alpha)$.



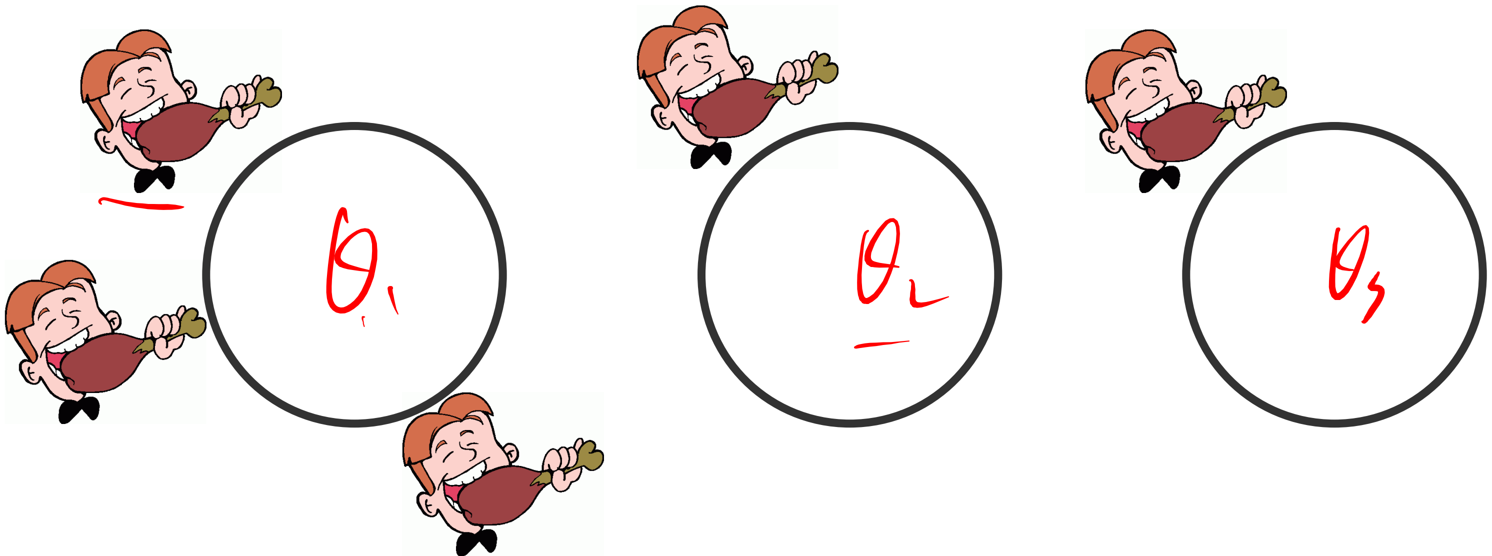
Chinese restaurant process

- 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. 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 process

- 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. 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 $\alpha/(n-1+\alpha)$.



Exchangeability

- An interesting fact: the distribution over the clustering of the first N customers *does not depend on the order in which they arrived*.
- Homework: Prove to yourself that this is true.
- However, the customers are not independent – they tend to sit at popular tables.
- We say that distributions like this are *exchangeable*.
- De Finetti's theorem: If a sequence of observations is exchangeable, there must exist a distribution given which they are iid.
- The customers in the CRP are iid given the underlying Dirichlet process – by integrating out the DP, they become dependent.

The Stick-breaking Process

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

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

$$\theta_k \sim G_0$$

Location

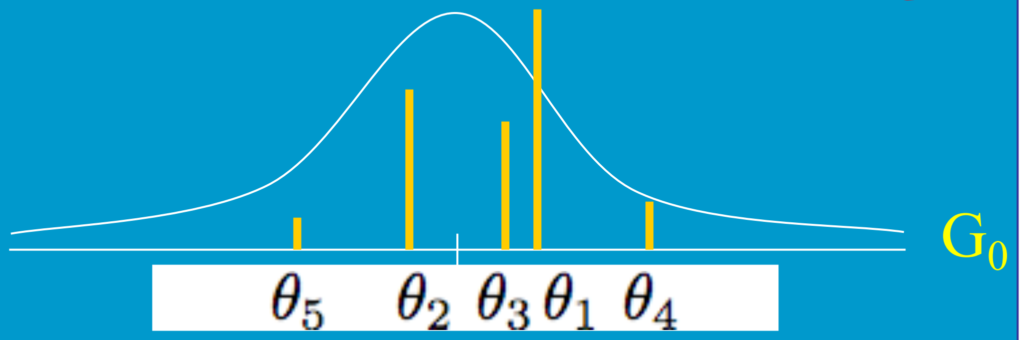
$$\sum_{k=1}^{\infty} \pi_k = 1$$

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

Mass

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

$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24



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,…
  - Sample a beta(1,α) random variable b_k.
  - Break off a fraction b_k of the stick. This is the k^{th} atom size
  - Sample a random location for this atom.
  - Recurse on the remaining stick.

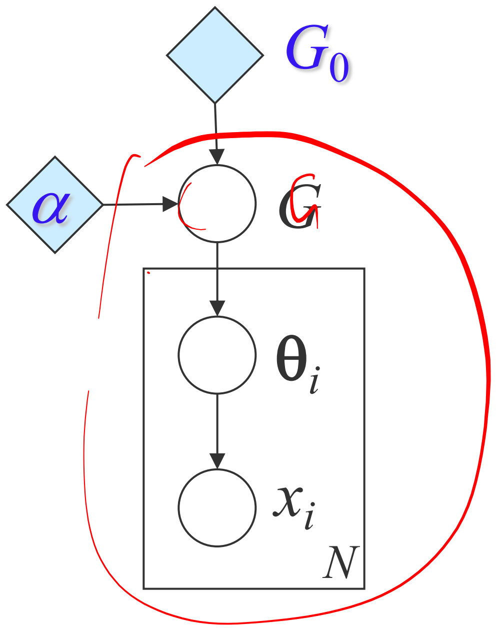
$$G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

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

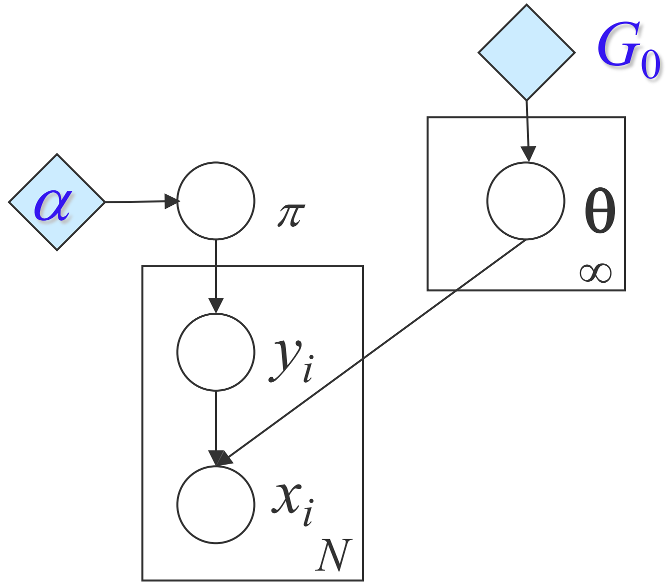
$$b_k \sim \mathrm{Beta}(1, \alpha)$$

[Sethuraman, 1994]

Graphical Model Representations of DP



The Pólya urn construction



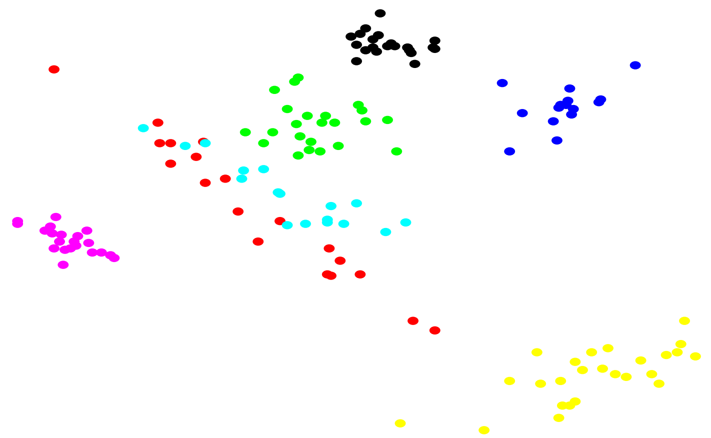
The Stick-breaking construction

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 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 nth 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 sampler

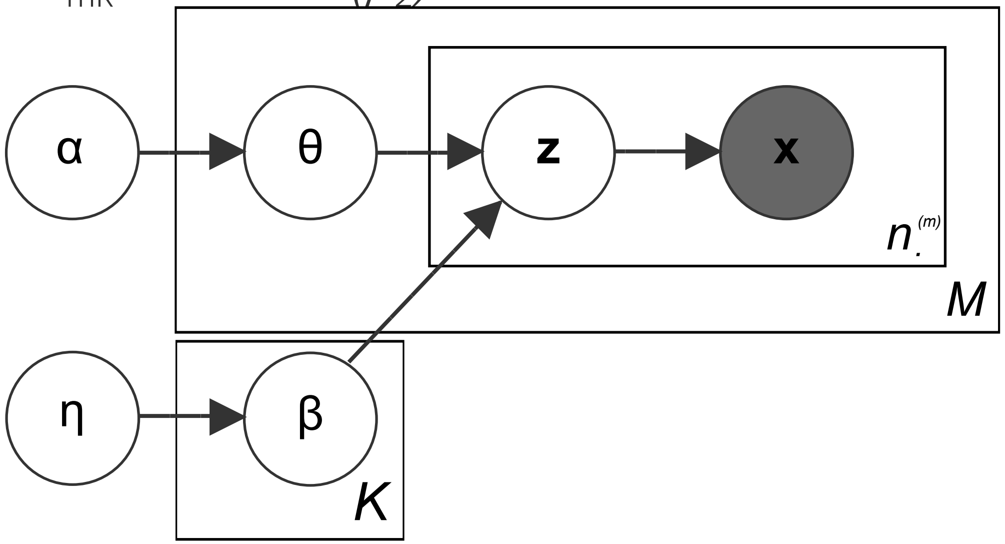
- We are only updating one data point at a time.
- Imagine two "true" clusters are merged into a single cluster – a single data point is unlikely to "break away".
- Getting to the true distribution involves going through low probability states ➔ mixing can be slow.
- If the likelihood is not conjugate,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

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



Blei et al, 2002

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.

Sharing topics

- Because the base measure is *continuous*, we have zero probability of picking the same topic twice.
- 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.

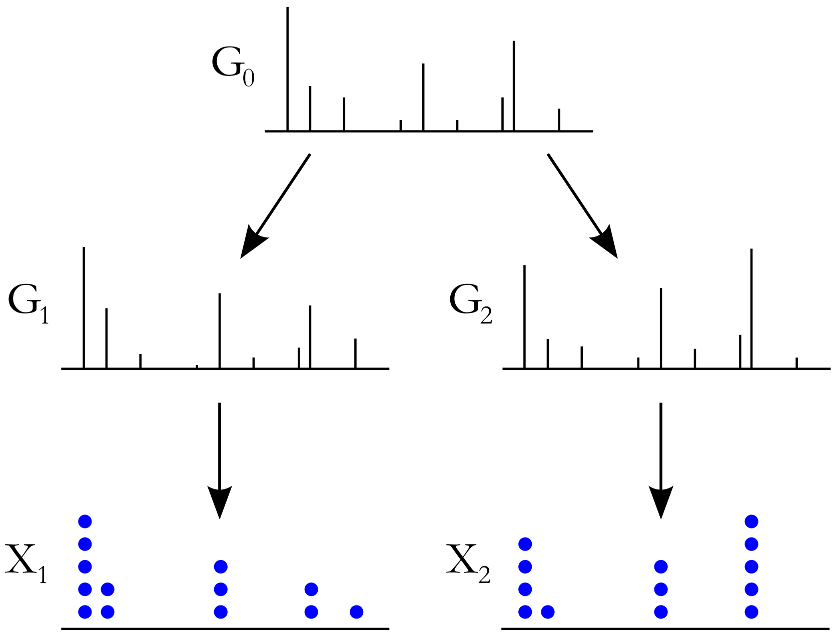
- We want there to be an infinite number of topics, so we want an *infinite, discrete* base measure.
- We want the location of the topics to be random, so we want an *infinite, discrete, random* base measure.

Hierarchical Dirichlet Process (Teh et al, 2006)

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

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

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

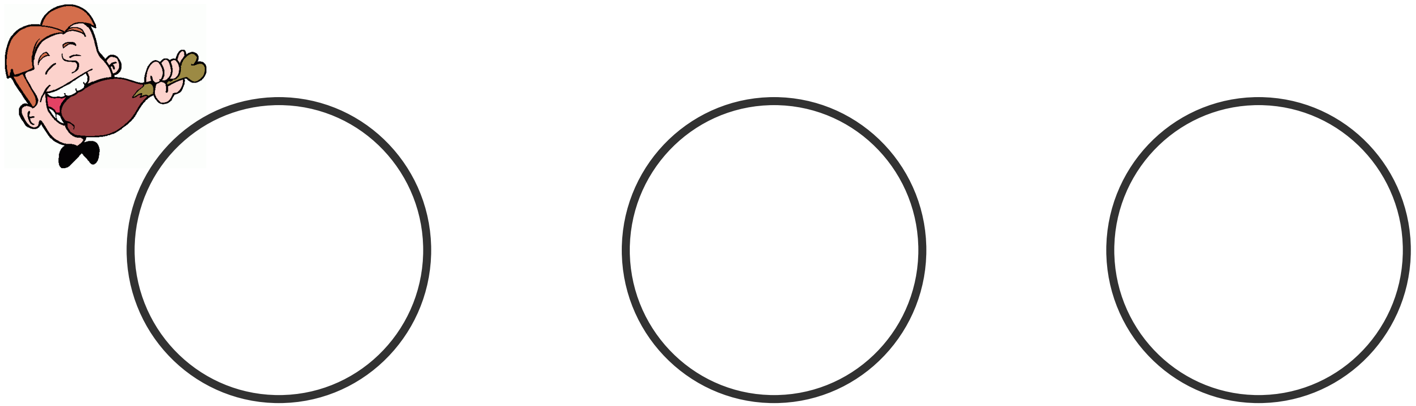


Chinese restaurant franchise

- Imagine a *franchise* of restaurants, serving 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.

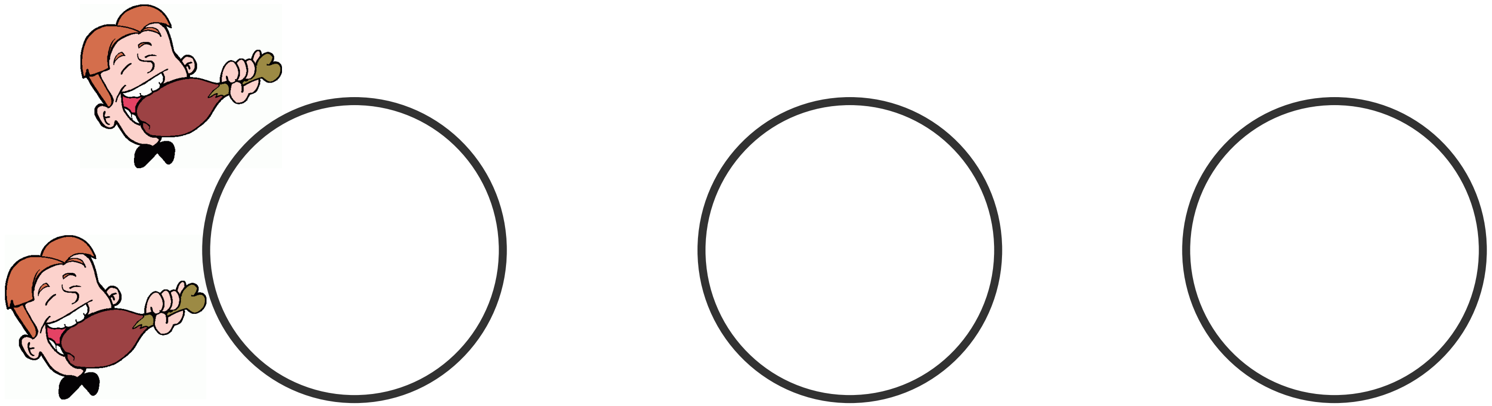
Chinese restaurant franchise

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



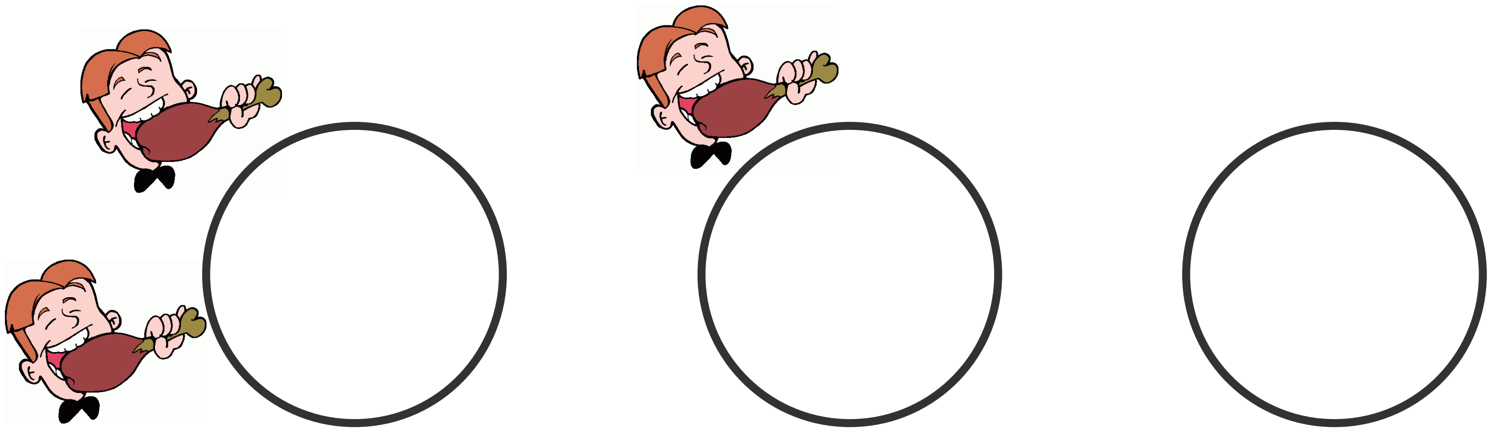
Chinese restaurant franchise

- 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 $\alpha/(n-1+\alpha)$.



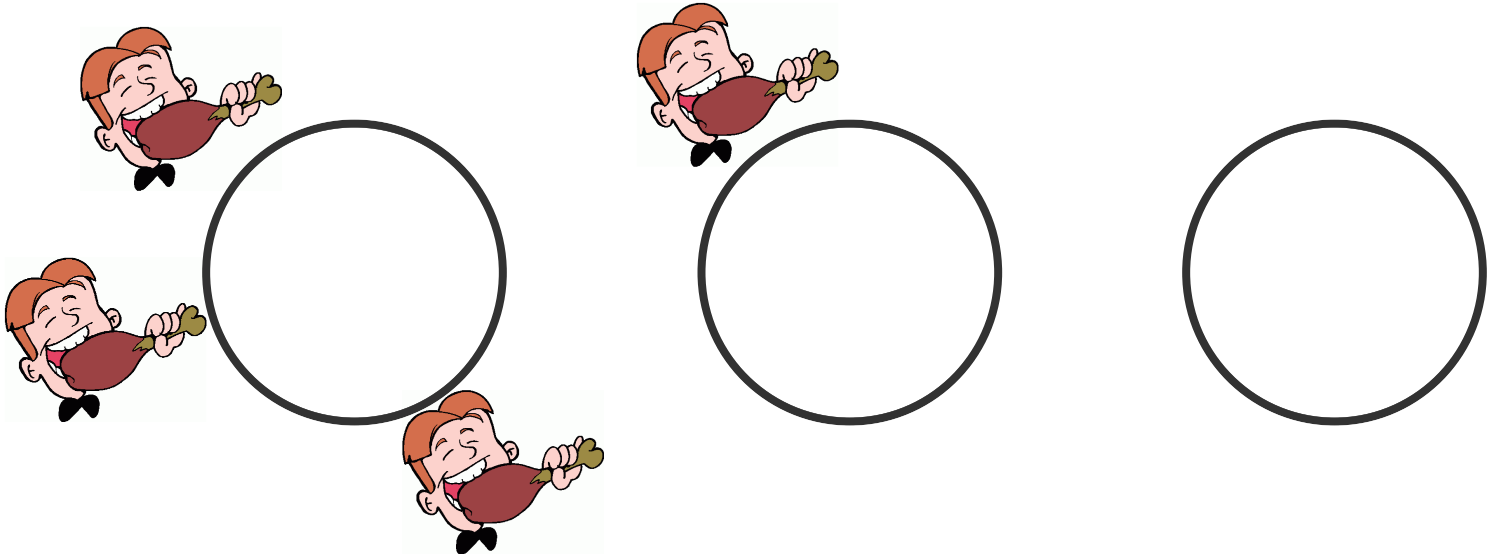
Chinese restaurant franchise

- 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 $\alpha/(n-1+\alpha)$.



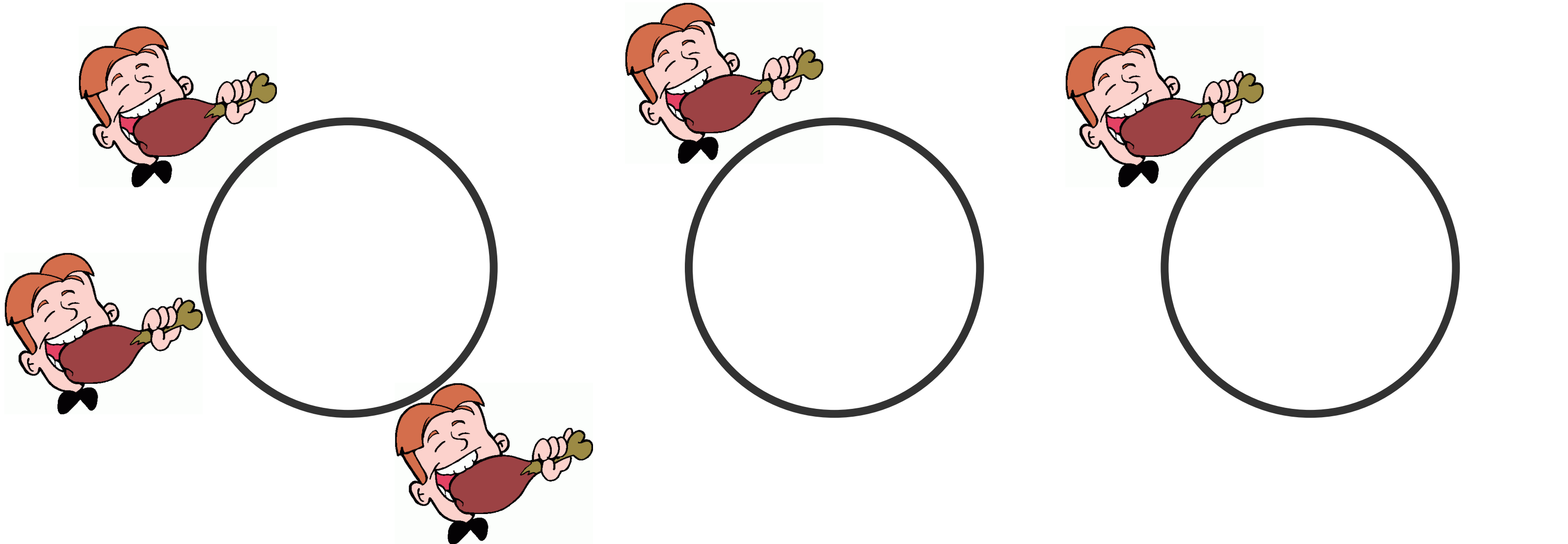
Chinese restaurant franchise

- 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

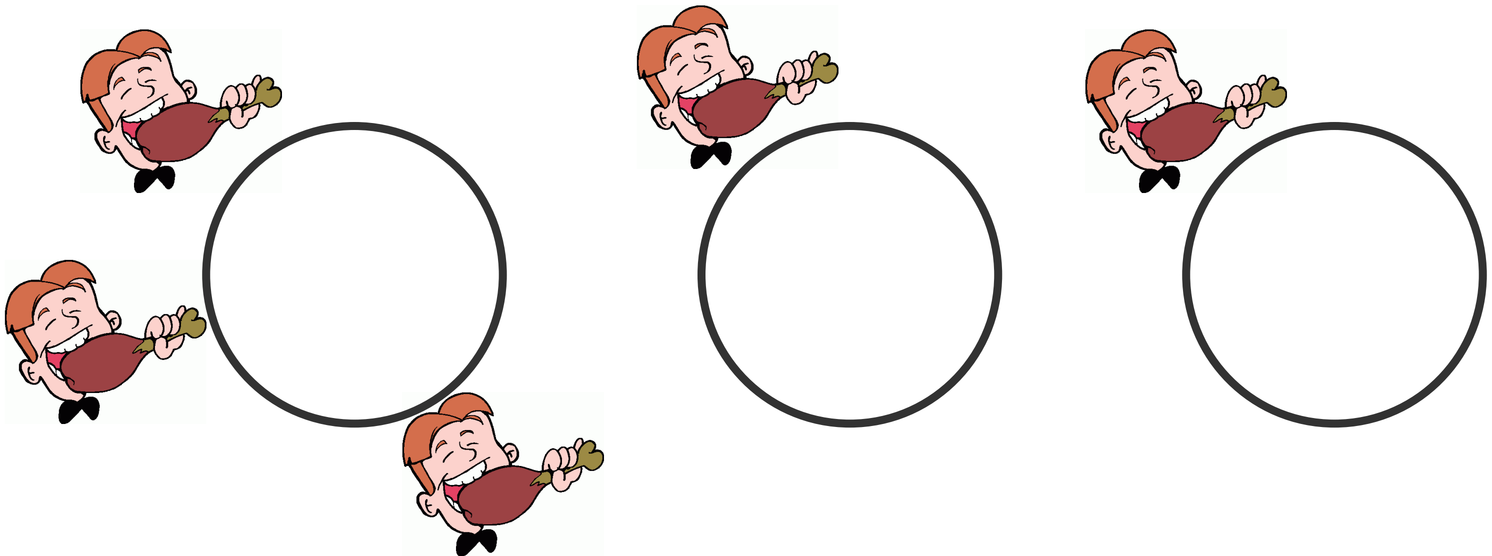
- Customers enter the restaurants, and sit at tables according to the Chinese restaurant process
  - The first customer enters a restaurant, and picks a table.
  - The n^{th} customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where m_k is the number of people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

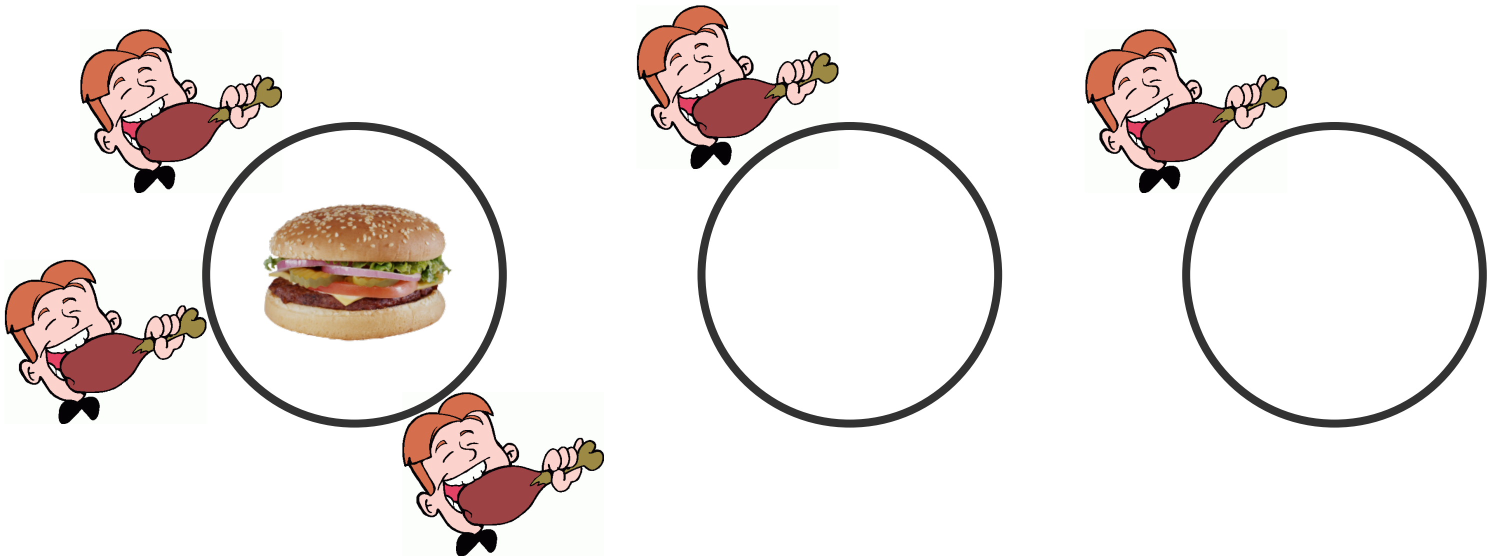
$$p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



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.

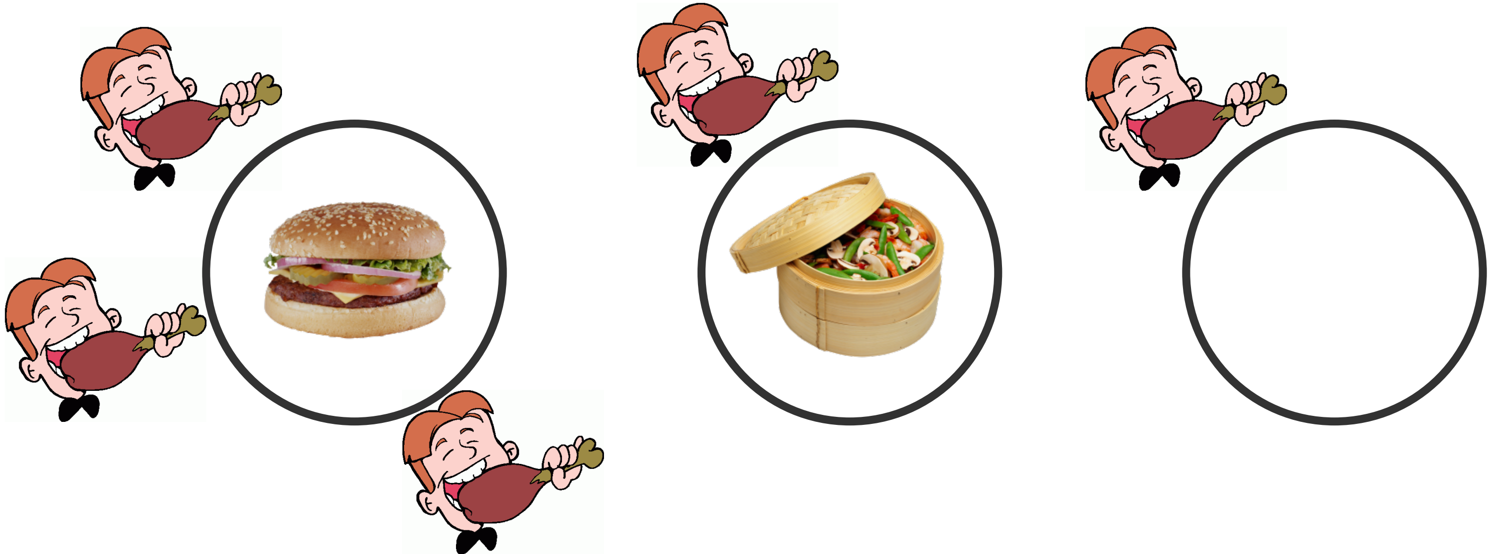
$$p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



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.

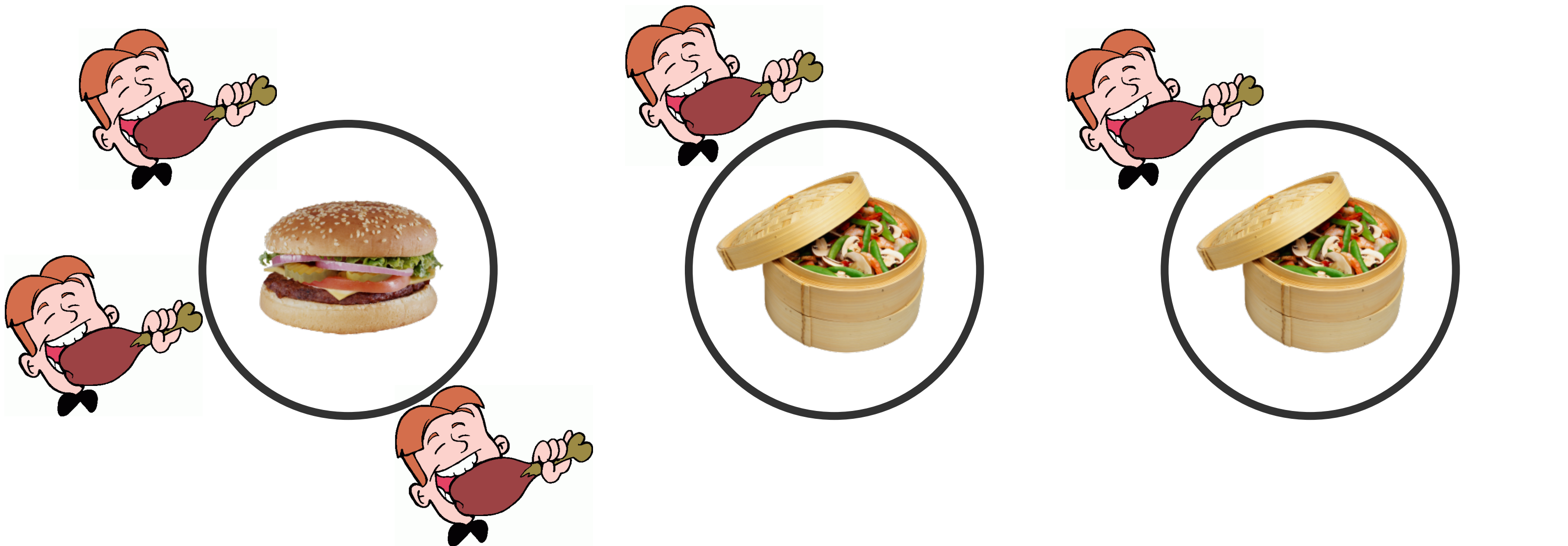
$$p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



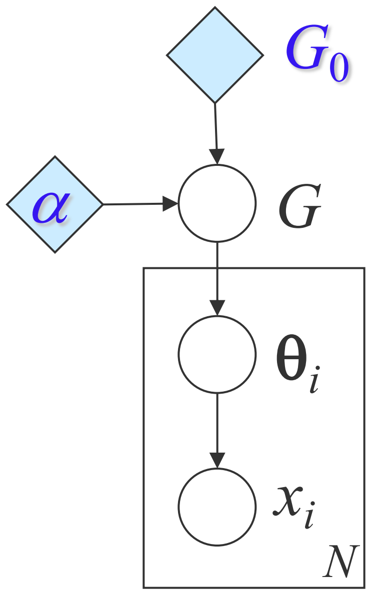
Chinese restaurant franchise

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

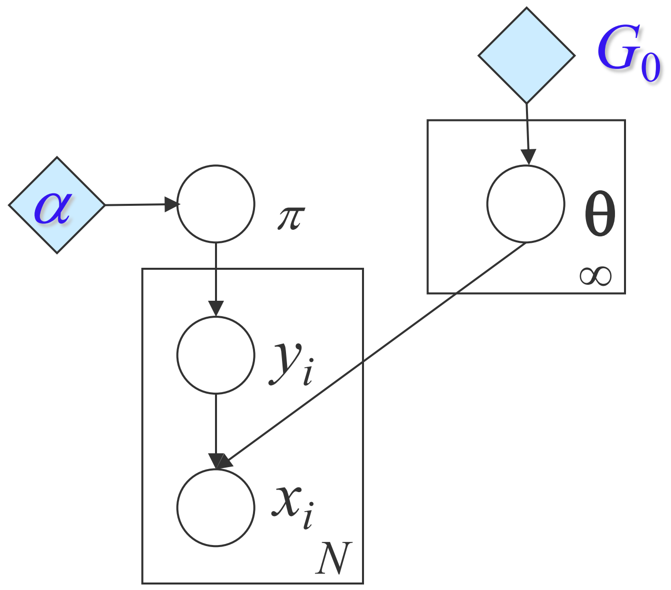
$$p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}$$



Recall: Graphical Model Representations of DP

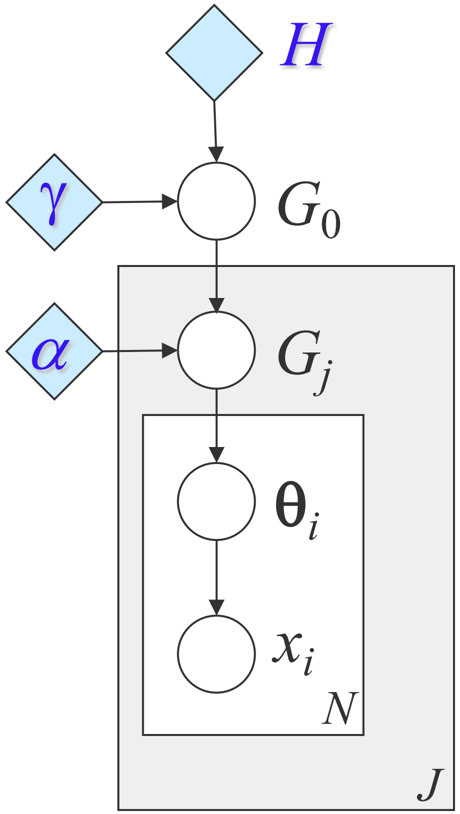


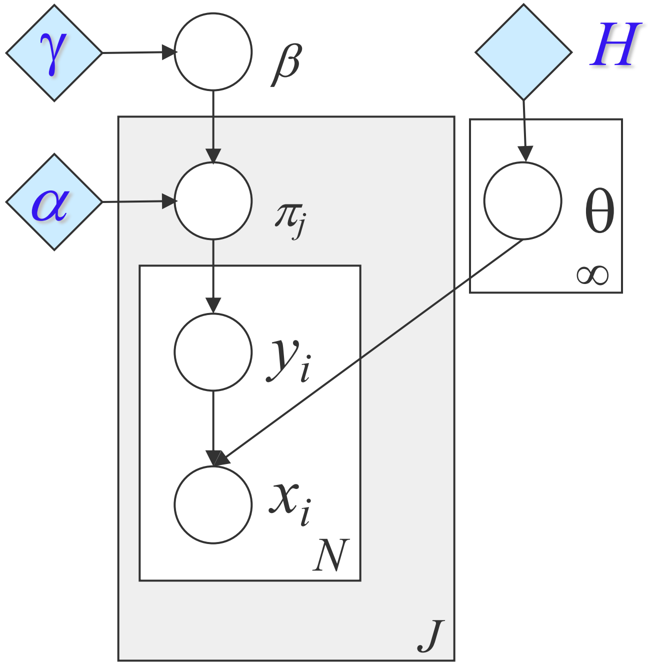
The Pólya urn construction



The Stick-breaking construction

Hierarchical DP Mixture





$\theta_k \sim H$

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

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

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

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

An infinite topic model

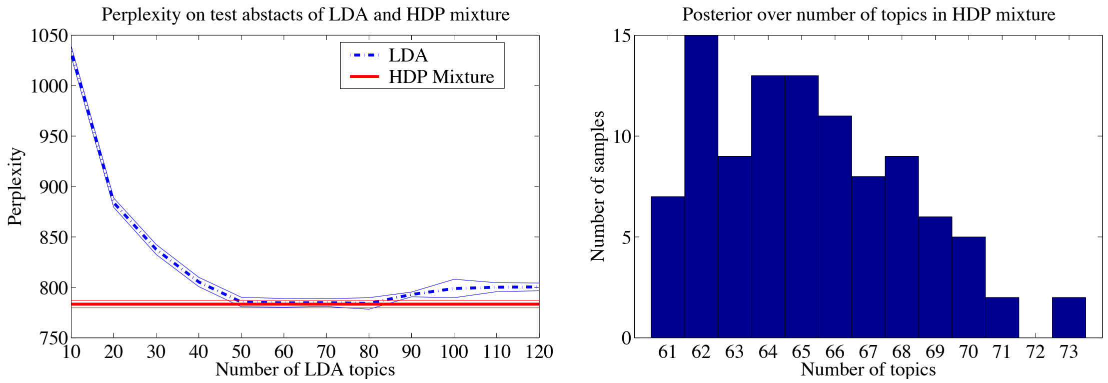
- Restaurants = documents; dishes = topics.

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

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: Factor analysis

  $X = WA^T + \varepsilon$

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

Infinite latent feature models

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

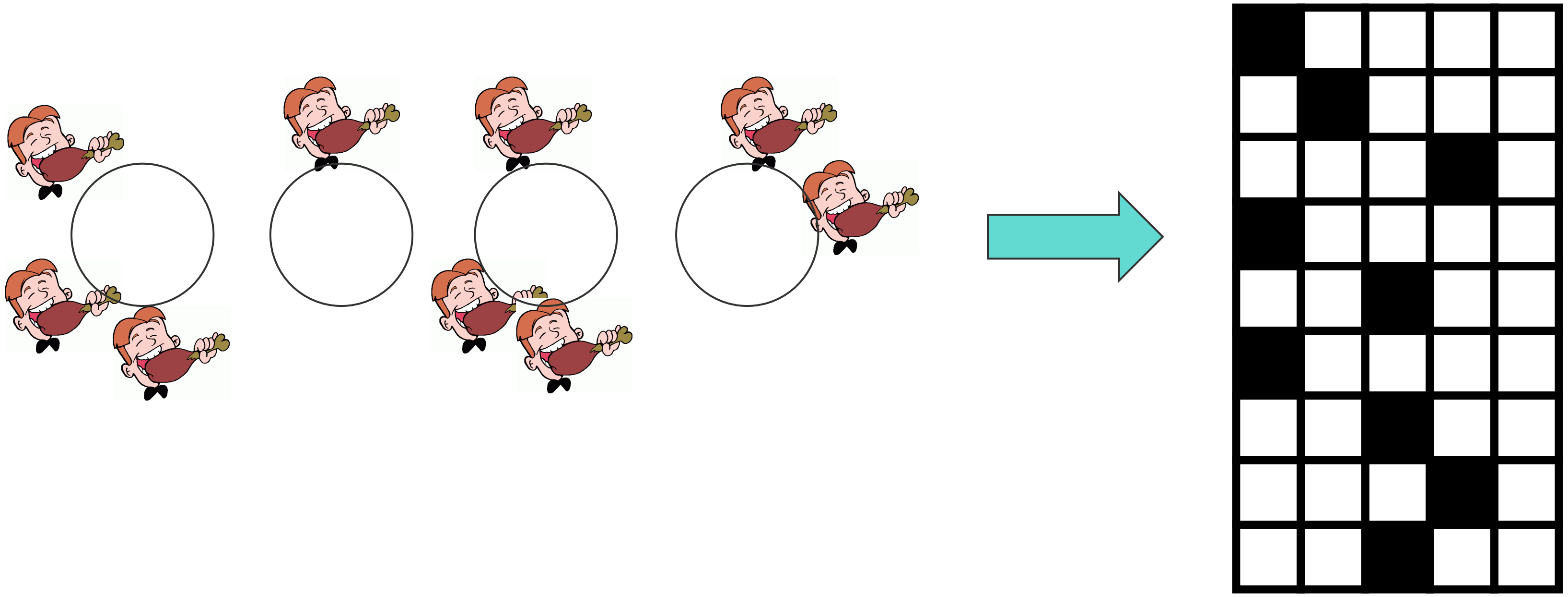
  $$X = WA^T + \varepsilon$$

- 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 – ie 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

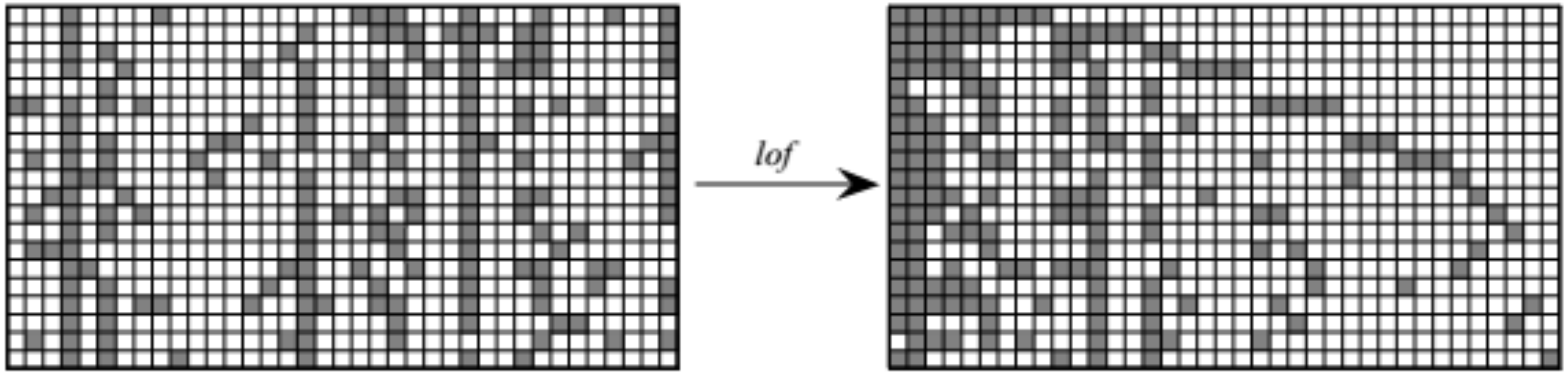


Figure 5: Binary matrices and the left-ordered form. The binary matrix on the left is transformed into the left-ordered binary matrix on the right by the function $lof(\cdot)$. This left-ordered matrix was generated from the exchangeable Indian buffet process with $\alpha = 10$. Empty columns are omitted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

$$p([\mathbf{Z}]) = \sum_{\mathbf{Z}\in[\mathbf{Z}]} p(\mathbf{Z})$$

$$= \frac{K!}{\prod_{n=0}^{2^N-1} K_n!} \prod_{k=1}^{K} \frac{\alpha}{K} \frac{\Gamma(m_k + \alpha/K)\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \alpha/K)}$$

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

$$\cdot \prod_{k=1}^{K_+} \frac{(N - m_k)! \prod_{j=1}^{m_k-1} (j + \alpha/K)}{N!}$$

Taking the infinite limit

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

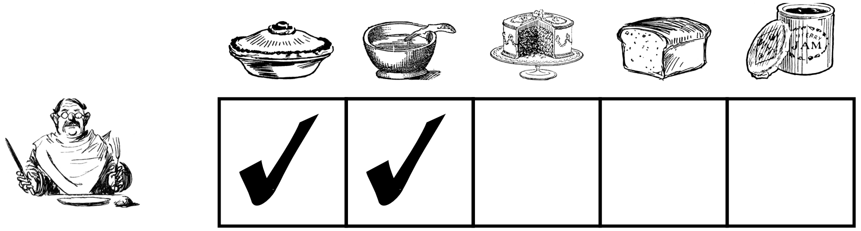
$$\frac{\alpha^{K_+}}{\prod_{n=1}^{2^N-1} K_n!} \frac{K!}{K_0! K^{K_+}} \left(\frac{N!}{\prod_{j=1}^{N} j + \frac{\alpha}{K}} \right)^K \prod_{k=1}^{K_+} \frac{(N-m_k)! \prod_{j=1}^{m_k-1}(j+\frac{\alpha}{K})}{N!}$$

$$\downarrow K \rightarrow \infty$$

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

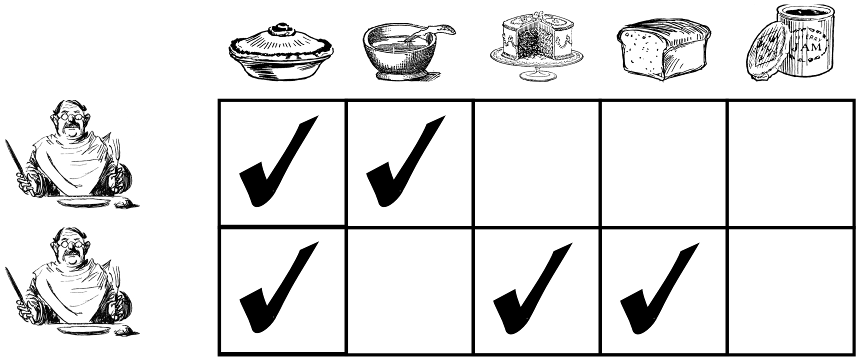
Predictive distribution: The Indian buffet process

- 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

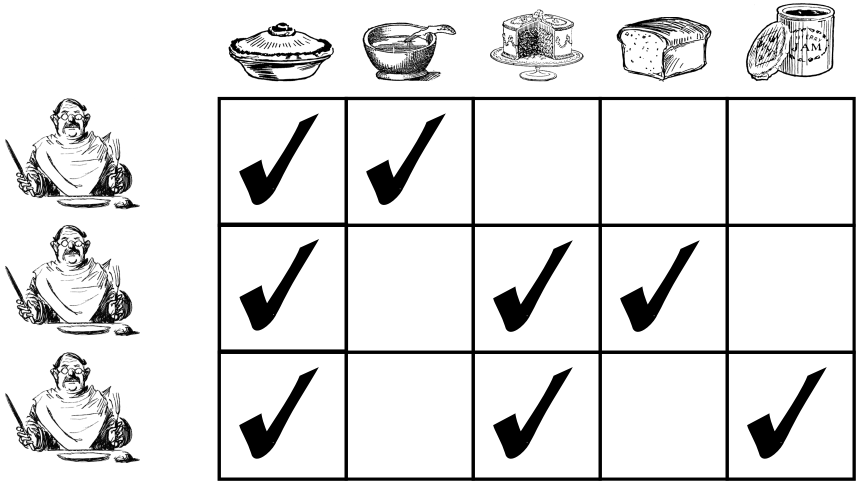
- We can describe this model in terms of the following restaurant analogy.
  - A customer enters a restaurant with an infinitely large buffet
  - He helps himself to Poisson(α) dishes.
  - The n^{th} customer enters the restaurant
  - He helps himself to each dish with probability m_k/n
  - He then tries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



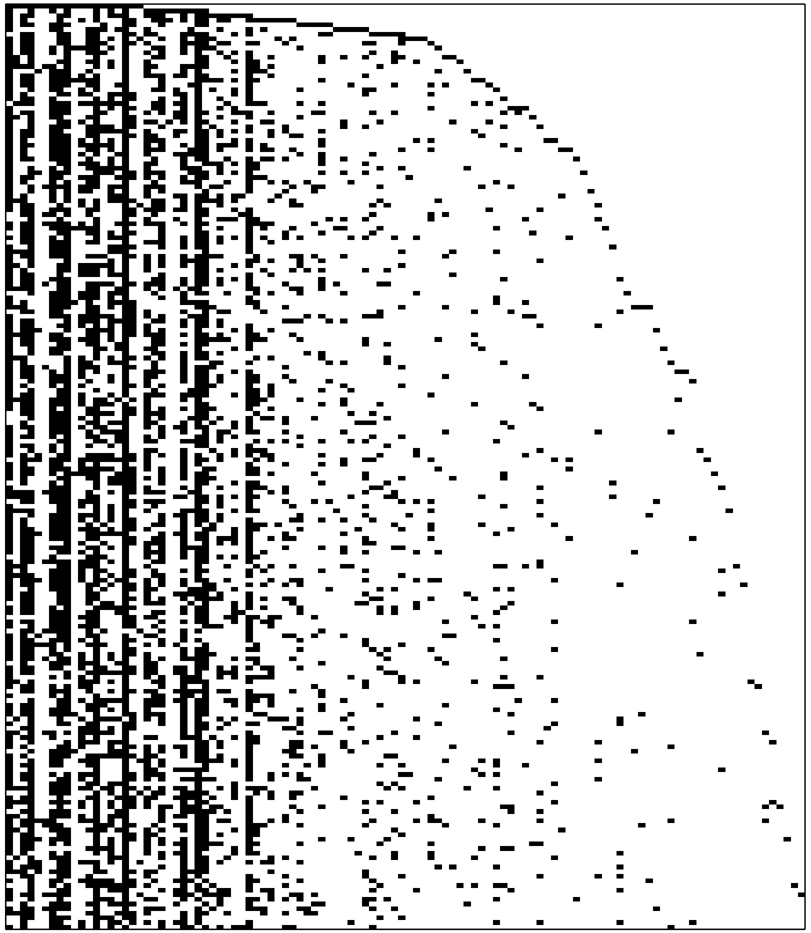
✓	✓			
✓		✓	✓	

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.

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

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

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

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

- 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~Poisson(α_1) and x_2~Poisson(α_2), then (x_1+x_2)~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.