School of Computer Science
Carnegie Mellon

10-708 Probabilistic Graphical Models

Bayesian Nonparametrics: Indian Buffet Process

Readings:
Griffiths & Ghahramani (2011)

Matt Gormley
Lecture 19
March 23, 2016

Reminders

- Typo fixed: lecture 16, slide 29, detailed balance

- Today: wrap up DPMM

Outline

- **Motivation:** *Infinite* Latent Feature Models
- **Finite Feature Model**
  - Beta-Bernoulli Model
  - Marginalized Beta-Bernoulli Model
  - Expected # of non-zeros
  - Taking the **Infinite** Limit
  - Left-ordered form (equivalence classes)
- **The Indian Buffet Process (IBP)**
  - Nonexchangeable IBP
  - Exchangeable IBP
  - Gibbs Sampling with Exchangeable IBP
- **IBP properties**
- **Applications**
- **Summary**

Motivation

- **Latent Feature Models**
  - Examples:
    - factor analysis
    - probabilistic PCA
    - cooperative vector quantization
    - sparse PCA
- **Applications**
  - choice behavior (i.e. option A over option B)
  - proteomics: modeling the functional interactions of proteins – which can belong to multiple complexes at the same time
  - collaborative filtering: modeling features of movie preferences (a la. Netflix challenge)
  - structure learning for graphical models (i.e. bipartite graphs)

Latent Feature Models

Let $\mathbf{x}_i$ be the ith data instance

$\mathbf{f}_i$ be its features

Define $\mathbf{X} = [\mathbf{x}_1^T, x_2^T, \ldots, x_N^T]$

$\mathbf{F} = [\mathbf{f}_1^T, f_2^T, \ldots, f_N^T]$

Model: $p(\mathbf{X}, \mathbf{F}) = p(\mathbf{X}|\mathbf{F})p(\mathbf{F})$

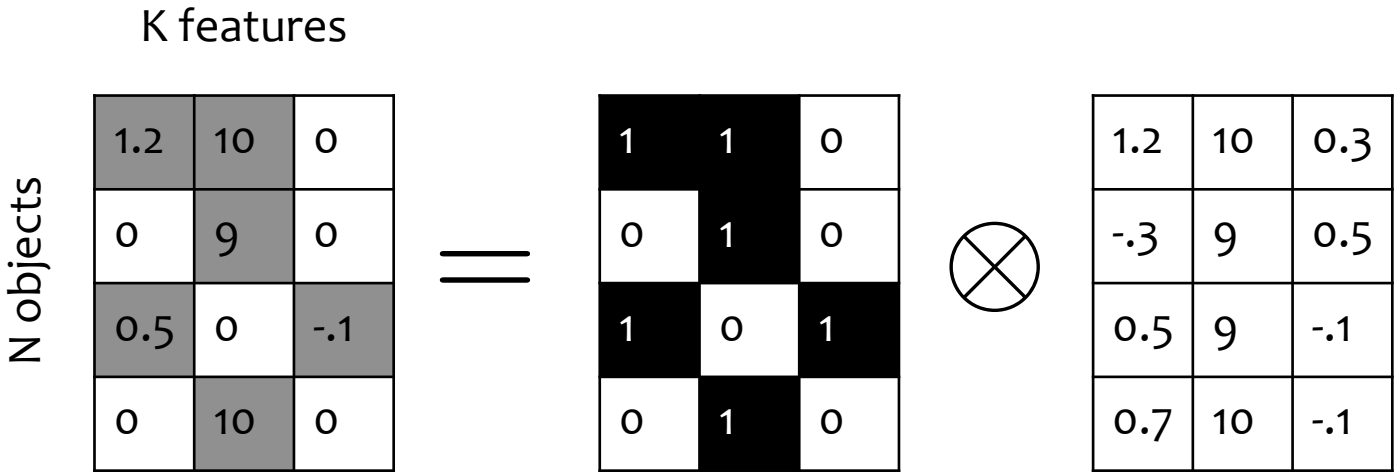
Latent Feature Models

Decompose the feature matrix, F, into a sparse binary matrix, Z, and a value matrix, V.

$\mathbf{F} = \mathbf{Z} \otimes \mathbf{V}$ where $\otimes$ is the elementwise product

$$z_{ij} \in \{0, 1\}$$

$$v_{ij} \in \mathcal{R}$$



Latent Feature Models

Decompose the feature matrix, F, into a sparse binary matrix, Z, and a value matrix, V.

$\mathbf{F} = \mathbf{Z} \otimes \mathbf{V}$ where $\otimes$ is the elementwise product

$z_{ij} \in \{0, 1\}$

$v_{ij} \in \mathcal{R}$

Model: $p(\mathbf{X}, \mathbf{F}) = p(\mathbf{X}|\mathbf{F})p(\mathbf{F})$

$= p(\mathbf{X}|\mathbf{F})p(\mathbf{Z})p(\mathbf{V})$

The IBP will provide p(Z) for the case of infinite K!

Figure from Griffiths & Ghahramani (2011)

Finite Feature Model

Beta-Bernoulli Model

Generative Story:

- for each feature $k \in \{1, \ldots, K\}$:
  - $\pi_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$ where $\alpha > 0$
  - for each object $i \in \{1, \ldots, N\}$:
    - $z_{ik} \sim \text{Bernoulli}(\pi_k)$

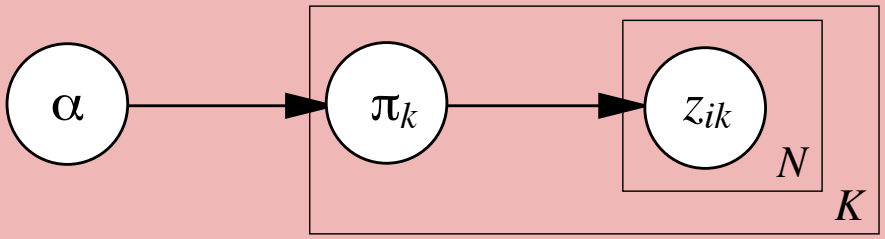
[row]

[prob. of feat. k]

[column]

[is feat. ON/OFF]

$p(\mathbf{Z}, \boldsymbol{\pi} \mid \alpha)$



Finite Feature Model

Marginalized Beta-Bernoulli Model

Because of the **conjugacy** of the **Beta** and **Bernoulli**, we can analytically **marginalize out** the feature prevalence parameters, π_k.

$$P(\mathbf{Z}) = \prod_{k=1}^{K} \int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k) \right) p(\pi_k)\, d\pi_k$$

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

where $m_k = \sum_{i=1}^{N} z_{ik}$ is # features ON in column k,

Γ is the Gamma function

Finite Feature Model

Expected # of non-zeroes

Generative Story:

- for each feature $k \in \{1, \ldots, K\}$:
  - $\pi_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$ where $\alpha > 0$
  - for each object $i \in \{1, \ldots, N\}$:
    - $z_{ik} \sim \text{Bernoulli}(\pi_k)$

[row]

[prob. of feat. k]

[column]

[is feat. ON/OFF]

Recall: if $X \sim \text{Beta}(r, s)$, then $\mathbb{E}[X] = \frac{r}{r+s}$

if $Y \sim \text{Bernoulli}(p)$, then $\mathbb{E}[Y] = p$

$$\mathbb{E}[z_{ik}] = \frac{\frac{\alpha}{K}}{1 + \frac{\alpha}{K}}$$

$$\Rightarrow \mathbb{E}[\mathbf{1}^T \mathbf{Z} \mathbf{1}] = \mathbb{E}\left[\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik}\right] = \frac{N\alpha}{1 + \frac{\alpha}{K}}$$

So the expected number of non-zero entries in Z is ≤ $N\alpha$

What happens as $K \to \infty$?

Finite Feature Model

*Taking the **Infinite** Limit*

$$\lim_{K\to\infty} p(\mathbf{Z}) = \lim_{K\to\infty} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(N+1+\frac{\alpha}{K})}.$$

$$= 0$$

Problem: Every **matrix** has **zero** probability!

Finite Feature Model

Left-Ordered Form (lof)

Topic Modeling:

- Consider many samples of the k[th] topic from the Markov chain: $\phi_k^{(1)}, \phi_k^{(2)}, \ldots, \phi_k^{(T)}$

  This topic will "drift" over time (e.g. from {politics} at time (t) to {geology} at time (t+m))
- Instead of averaging, it's common to use a MAP estimate of the topics
- The **order** of the topics is **not important** to the model (i.e. the topics are not identifiable)

Finite Feature Model

Left-Ordered Form (lof)

Back to our model:

- Q: In a **latent** feature model, what's the difference between feature k=13 and k=27?
- A: Nothing!

The use of left-ordered form **capitalizes** on the fact that **features are not identifiable** (i.e. order of features doesn't matter to the model)

Finite Feature Model

Left-Ordered Form (lof)

Define the history of feature k to be the magnitude of the binary value given by the column:

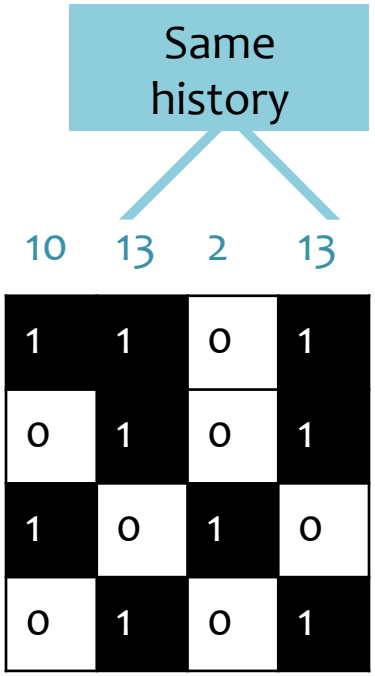
$$h_k = \sum_{i=1}^{N} 2^{(N-i)} z_{ik}$$

K_h = # of features with history h

K_0 = # of features with $m_k = 0$ (i.e. $h = 0$)

$$K_+ = \sum_{h=1}^{2^N - 1} K_h, \text{ \# of features with non-zero history}$$

$$\Rightarrow K = K_0 + K_+$$



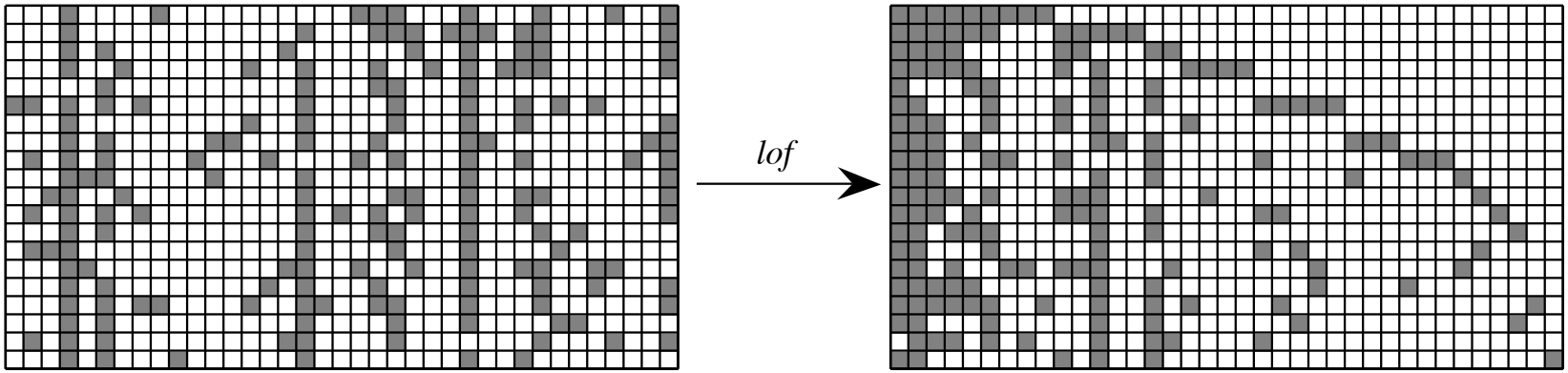
Define lof(Z) to be sorted left-to-right by the history of each feature.

Figure from Griffiths & Ghahramani (2005)

Finite Feature Model

Left-Ordered Form (lof)

Define *lof(Z)* to be sorted left-to-right
by the history of each feature.



Define equivalence class $[Z] = \{Z' : lof(Z') = lof(Z)\}$

Cardinality of $[Z] = \dfrac{K!}{\prod_{h=0}^{2^N-1} K_h!}$

Finite Feature Model

*Taking the **Infinite** Limit*

$$\lim_{K\to\infty} p(\mathbf{Z}) = \lim_{K\to\infty} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}.$$

$$= 0$$

Problem: Every **matrix** has **zero** probability!

$$\lim_{K\to\infty} p([\mathbf{Z}]) = \lim_{K\to\infty} \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} p(\mathbf{Z})$$

$$= \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \cdot \exp\{-\alpha H_N\} \cdot \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!},$$

where $H_N = \sum_{j=1}^{N} \frac{1}{j}$ is the Nth harmonic number

Solution: Every **equivalence class** has **non-zero** probability!

Figure from Griffiths & Ghahramani (2005)

The Indian Buffet Process

Non-exchangeable

- Imagine an Indian restaurant with a buffet containing an **infinite** # of dishes.
- N customers make a plate by selecting dishes from the buffet:
  - **1st customer:**
    Starts at the left and selects a Poisson(α) number of dishes
  - **ith customer:**
    1. Samples *previously sampled* dishes according to their popularity:
       (i.e. with prob. m_k/i where m_k is the # of previous customers who tried dish k)
    2. Then selects a Poisson(α/i) number of new dishes

Problem: the process is **not exchangeable** – dishes sampled as "new" **depend on the customer order.**

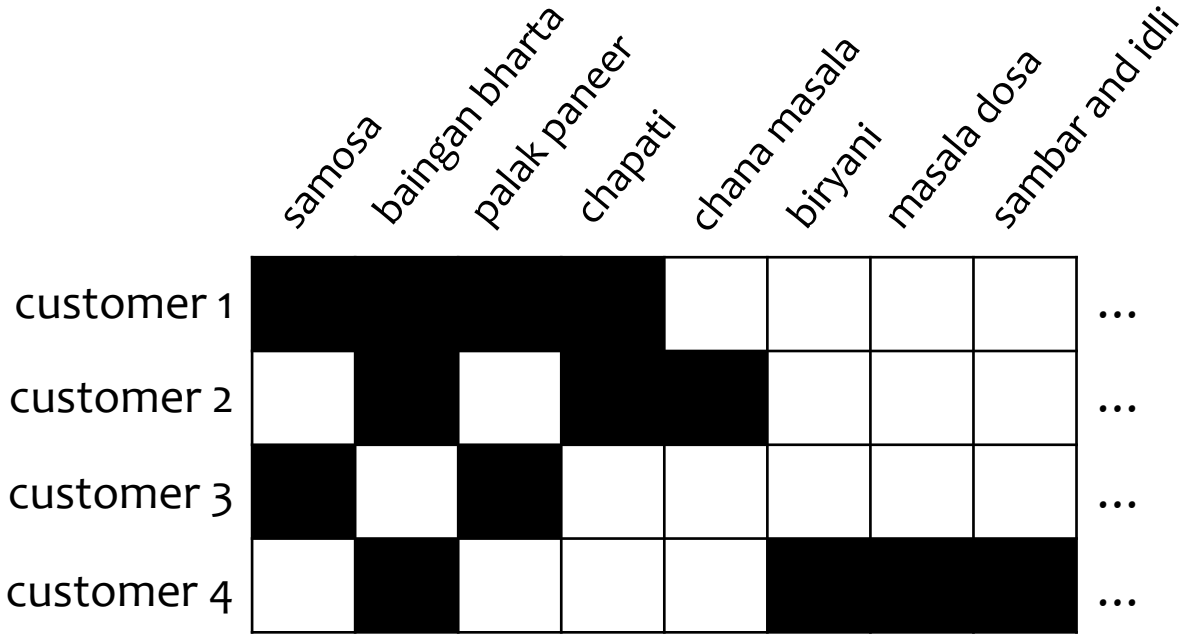


Figure from Griffiths & Ghahramani (2005)

The Indian Buffet Process

^Exchangeable

- Imagine an Indian restaurant with a buffet containing an **infinite** # of dishes.
- N customers make a plate by selecting dishes from the buffet:
  - **1st customer:**
    Starts at the left and selects a Poisson(α) number of dishes
  - **ith customer:**
    1. Makes a single decision for dishes with same history, h:
       (i.e. If there are K_h dishes w/history h sampled by m_h customers, then she samples a Binomial(m_h/I, K_h) number starting at the left)
    2. Then selects a Poisson(α/i) number of new dishes

This yields a *lof* matrix, Z.

Does so with probability p([Z])!

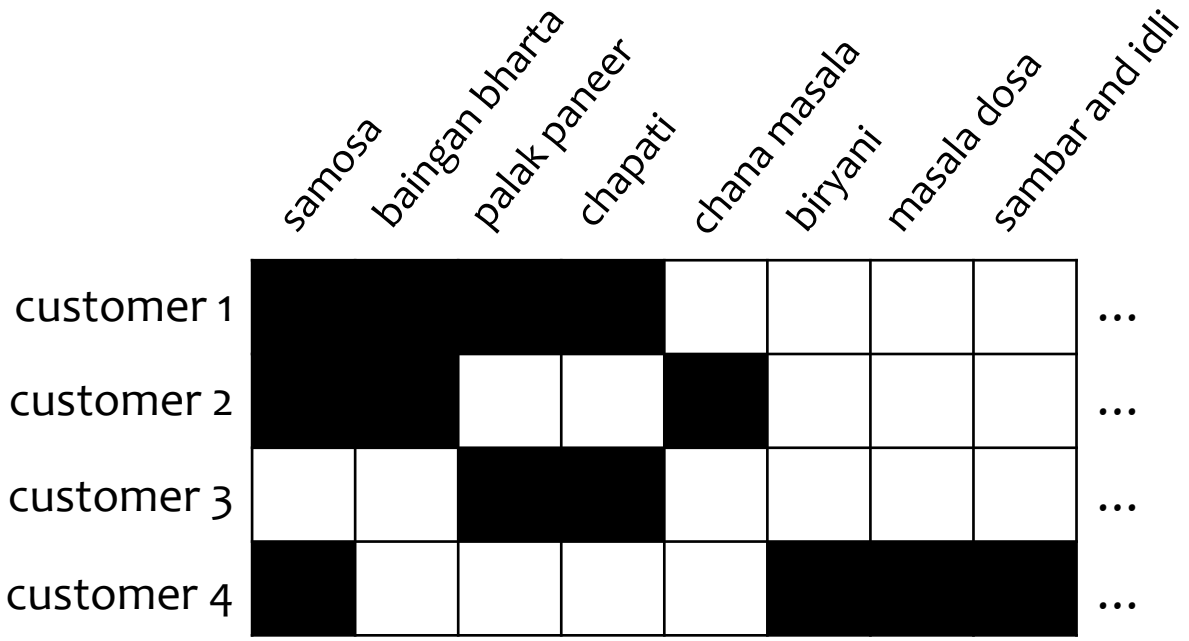
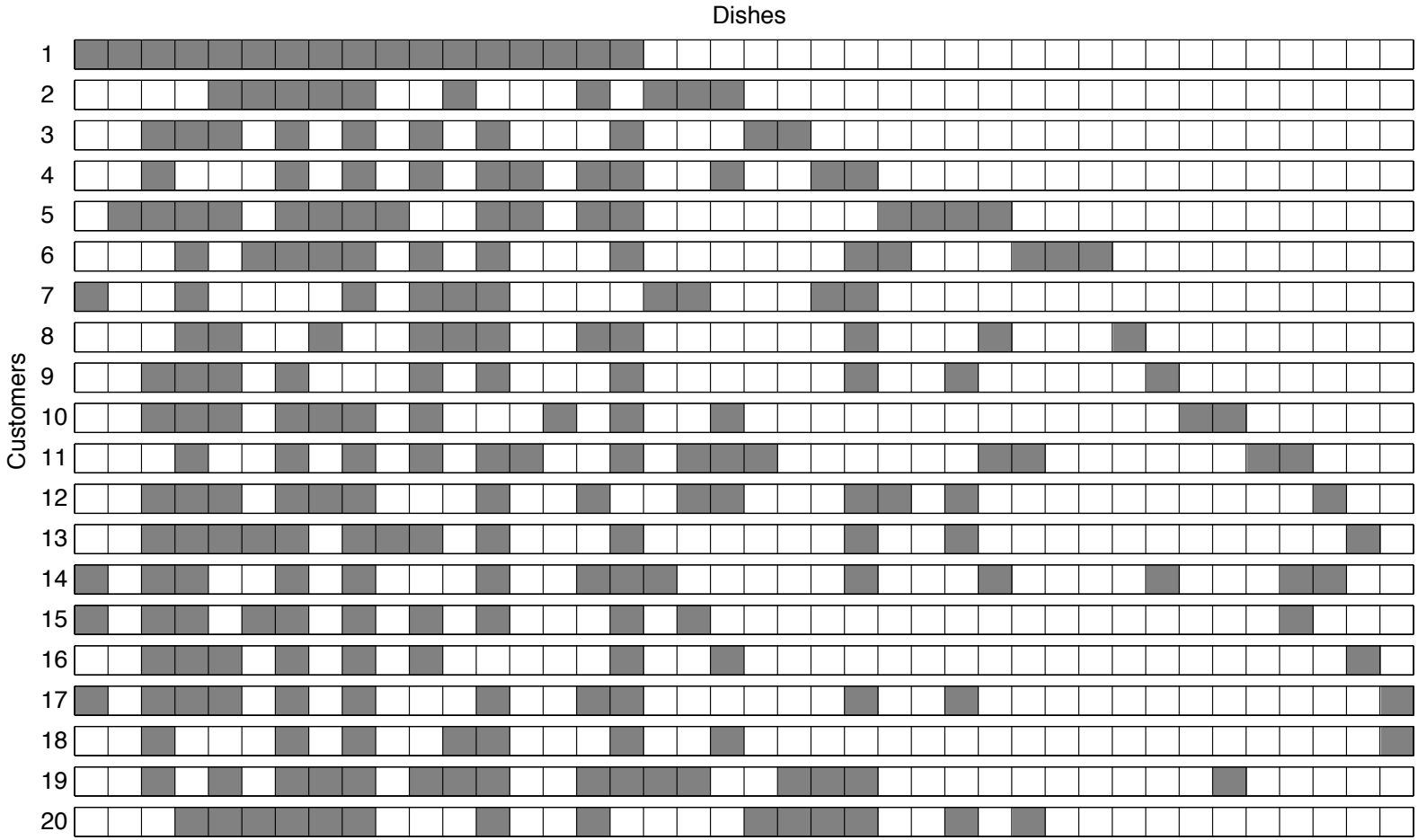


Figure from Griffiths & Ghahramani (2005)

The Indian Buffet Process

Example:



Gibbs Sampler for IBP

Consider a "prior only" sampler of p(Z | α)

- For finite K:

$$P(z_{ik} = 1|\mathbf{z}_{-i,k}) = \int_0^1 P(z_{ik}|\pi_k)p(\pi_k|\mathbf{z}_{-i,k})\,d\pi_k$$
$$= \frac{m_{-i,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}},$$

where $z_{-i,k}$ is the kth column except row i,

$m_{-i,k}$ is the # of rows w/feat. k except i

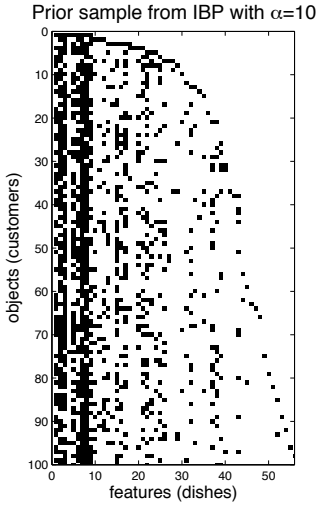
- For infinite K:
  - The "Exchangeable IBP" is *exchangeable*!
  - Choose an order s.t. the i^{th} customer was the last to enter (just like CRP sampler)
  - For any k s.t. $m_{-i,k} > 0$, resample:

$$P(z_{ik} = 1|\mathbf{z}_{-i,k}) = \frac{m_{-i,k}}{N},$$

  - Then draw a Poisson(α/i) # of new dishes.

Properties of the Indian buffet process

$$P([\mathbf{Z}]|\alpha) = \exp\{-\alpha H_N\} \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$



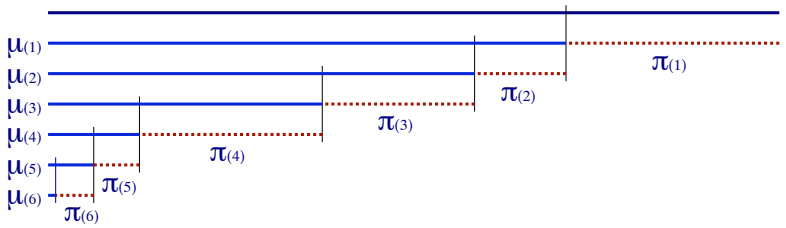


Figure 1: Stick-breaking construction for the DP and IBP. The black stick at top has length 1. At each iteration the vertical black line represents the break point. The brown dotted stick on the right is the weight obtained for the DP, while the blue stick on the left is the weight obtained for the IBP.

Shown in (Griffiths and Ghahramani, 2005):

- It is infinitely exchangeable.
- The number of ones in each row is Poisson(α)
- The expected total number of ones is αN.
- The number of nonzero columns grows as $O(\alpha \log N)$.

Additional properties:

- Has a stick-breaking representation (Teh, Görür, Ghahramani, 2007)
- Can be interpreted using a Beta-Bernoulli process (Thibaux and Jordan, 2007)

Posterior Inference in IBPs

$$P(\mathbf{Z}, \alpha|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha)P(\alpha)$$

Gibbs sampling: $P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \mathbf{X}, \alpha) \propto P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \alpha)P(\mathbf{X}|\mathbf{Z})$

- If $m_{-n,k} > 0$, $P(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$
- For infinitely many k such that $m_{-n,k} = 0$: Metropolis steps with truncation* to sample from the number of new features for each object.
- If α has a Gamma prior then the posterior is also Gamma $\rightarrow$ Gibbs sample.

Conjugate sampler: assumes that $P(\mathbf{X}|\mathbf{Z})$ can be computed.

Non-conjugate sampler: $P(\mathbf{X}|\mathbf{Z}) = \int P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta)d\theta$ cannot be computed, requires sampling latent θ as well (c.f. (Neal 2000) non-conjugate DPM samplers).

***Slice sampler:** non-conjugate case, is not approximate, and has an adaptive truncation level using a stick-breaking construction of the IBP (Teh, et al, 2007).

Particle Filter: (Wood & Griffiths, 2007).

Accelerated Gibbs Sampling: maintaining a probability distribution over some of the variables (Doshi-Velez & Ghahramani, 2009).

Variational inference: (Doshi-Velez, Miller, van Gael, & Teh, 2009).

Modelling Data

Latent variable model: let $\mathbf{X}$ be the $N \times D$ matrix of observed data, and $\mathbf{Z}$ be the $N \times K$ matrix of binary latent features

$$P(\mathbf{X}, \mathbf{Z}|\alpha) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha)$$

By combining the IBP with different likelihood functions we can get different kinds of models:

- Models for graph structures (w/ Wood, Griffiths, 2006)
- Models for protein complexes (w/ Chu, Wild, 2006)
- Models for overlapping clusters (w/ Heller, 2007)
- Models for choice behaviour (Görür, Jäkel & Rasmussen, 2006)
- Models for users in collaborative filtering (w/ Meeds, Roweis, Neal, 2006)
- Sparse latent factor models (w/ Knowles, 2007)

Summary

- Beta-Bernoulli model is a **simple** *finite* feature model
- Can treat features as **latent**
- **Infinite limit** of Beta-Bernoulli yields the Indian Buffet Process (IBP)
- Many properties of the IBP are similar to the CRP