Stick-breaking Construction for the Indian Buffet Process

Yee Whye Teh   Dilan Görür   Zoubin Ghahramani

Presented by Hsin-Ta Wu
2011/11/15

Outlines

- Indian Buffet Process
- Indian Buffet Process with stick-breaking construction
  - Derivation
  - Related to DP
- Slice sampling
- Semi-ordered stick-breaking
- Experiments

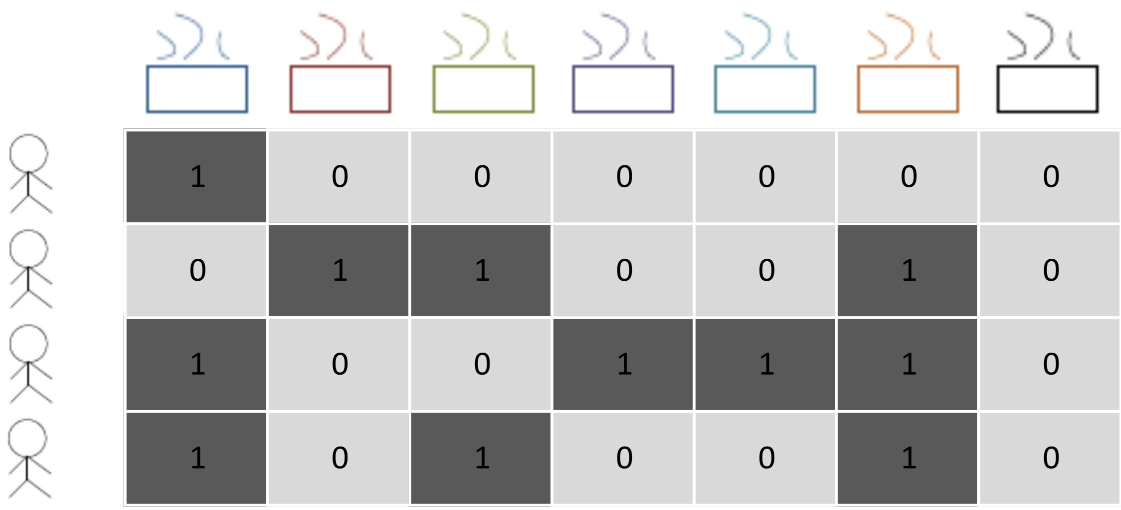
Introduction

- Indian Buffet Process (IBP)
  - A distribution over binary matrices consisting of N rows (objects) and an unbounded number of columns (features)
  - 1 and 0 in entry (*i*,*k*) indicates feature *k* present and absent from object *i*, respectively

	Action	Comedy	Animation	Brad Pitt	History	…	…
Terminator	1	0	0	0	0	0	0
Shrek	0	1	1	0	0	1	0
Troy	1	0	0	1	1	1	0
Avata	1	0	1	0	0	1	0

Introduction

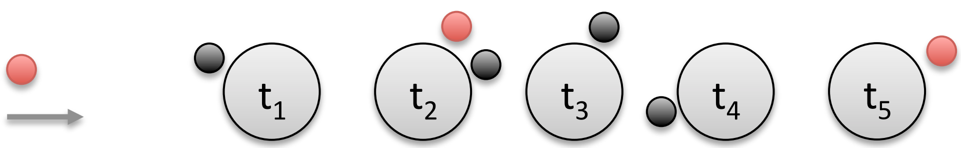
- Indian Buffet Process (IBP)
  - A distribution over binary matrices consisting of N rows (objects) and an unbounded number of columns (features)
  - 1 and 0 in entry (*i*,*k*) indicates feature *k* present and absent from object *i*, respectively



1	0	0	0	0	0	0
0	1	1	0	0	1	0
1	0	0	1	1	1	0
1	0	1	0	0	1	0

IBP vs. CRP

- Each object belongs to only one of infinitely many latent classes



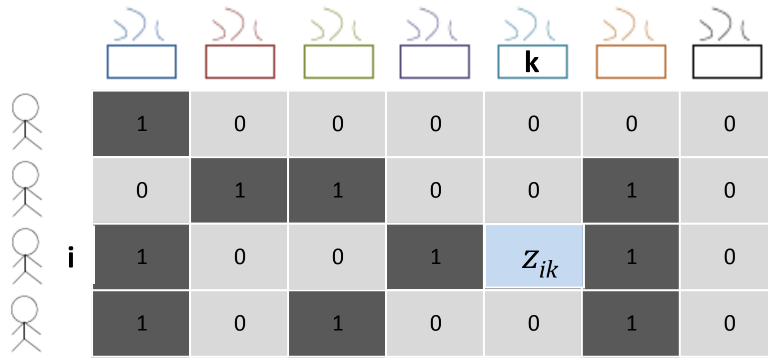
t_1	t_2	t_3	t_4	t_5	t_6	t_7
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	0	0	0	0

- Each object can possess potentially any combination of infinitely many latent features.

1	0	0	0	0	0	0
0	1	1	0	0	1	0
1	0	0	1	1	1	0
1	0	1	0	0	1	0

Indian Buffet Process (IBP)

Restaurant Construction



Z: a random binary NxK matrix
μ_k : prior probability that feature k presents in an object

$\mu_k \sim Beta(\alpha/K, 1)$
$z_{ik} | \mu_k \sim Bernoulli(\mu_k)$

$\theta_k \sim H$
$x_i \sim F(z_{i,:}, \theta_:)$

For the first customer, the distribution over the number of features it has is: (the number of dishes he tried)

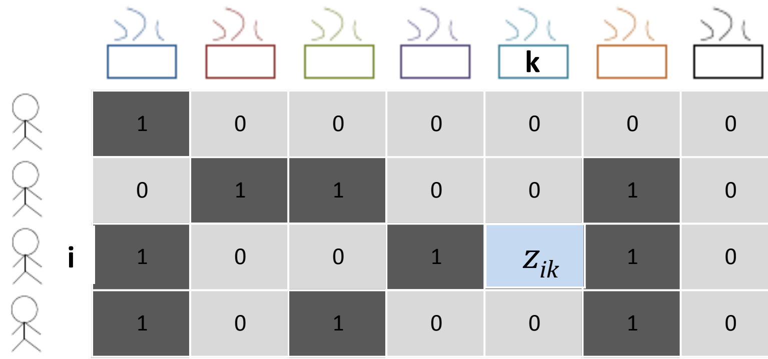
$Binomial(\alpha/K, K)$

when $K \longrightarrow \infty$

$Poisson(\alpha)$

Indian Buffet Process (IBP)

Restaurant Construction



Z: a random binary NxK matrix
μ_k : prior probability that feature k presents in an object

$\mu_k \sim Beta(\alpha/_K, 1)$ $\qquad$ $\theta_k \sim H$
$z_{ik} | \mu_k \sim Bernoulli(\mu_k)$ $\qquad$ $x_i \sim F(z_{i,:}, \theta_:)$

The i-th customer takes portions from previously sampled dishes with probability:

$$\frac{m_{<i\,k}}{i}$$

He can also tries $Poisson(\alpha/_i)$ new dishes.

Posterior Inference in IBP

- Gibbs Sampling:
  - Imagine that the object we are sampling as the last customer to the buffet.
- Iterate through i=1,…,N, for each object *i*,
  - Update the feature occurrences for the currently used features K^+

$$p\left(z_{ik}=1 \middle| z_{-(i,k)}, x_i, \theta_{1:K^+}\right) \propto p\left(z_{ik}=1 \middle| z_{-(i,k)}, \theta_{1:K^+}\right) p(x_i|Z, \theta_{1:K^+})$$

$$\propto \frac{m_{-i,k}}{N} p(x_i|z_{i,-k}, z_{ik}=1, \theta_{1:K^+})$$

  - Add L_i new features

$$p\left(L_i \middle| z_{i,1:K^+}, x_i, \theta_{1:K^+}\right)$$

$$\propto Poisson(L_i, \frac{\alpha}{N}) \times \int p\left(x_i \middle| z_{i,1:K^+}, z^\circ{}_{i,1:L_i}=1, \theta_{1:K^+}, \theta^\circ{}_{1:L_i}\right) \mathrm{dh}\left(\theta^\circ{}_{1:L_i}\right)$$

$$\propto \frac{(\frac{\alpha}{N})^{L_i} e^{-\frac{\alpha}{N}}}{L_i!} \times \int p\left(x_i \middle| z_{i,1:K^+}, z^\circ{}_{i,1:L_i}=1, \theta_{1:K^+}, \theta^\circ{}_{1:L_i}\right) \mathrm{dh}(\theta^\circ{}_{1:L_i})$$

Conjugacy on the IBP

- When new features being introduced:

$$p(L_i | z_{i,1:K^+}, x_i, \theta_{1:K^+})$$
$$\propto Poisson(L_i, \frac{\alpha}{N}) \times \int p(x_i | z_{i,1:K^+}, z^{\circ}_{i,1:L_i} = 1, \theta_{1:K^+}, \theta^{\circ}_{1:L_i}) \, \mathrm{dh}(\theta^{\circ}_{1:L_i})$$

What if *h* is not the conjugate prior for the data likelihood p(x|Z, θ) ?

The Integrals in equation will not be tractable.

Alternative representation of the IBP:

- the feature probabilities are not integrated out

Indian Buffet Process (IBP)

Stick-breaking Construction

- A decreasing ordering of $\mu_{1:K} = \{\mu_1, \dots, \mu_K\}$:

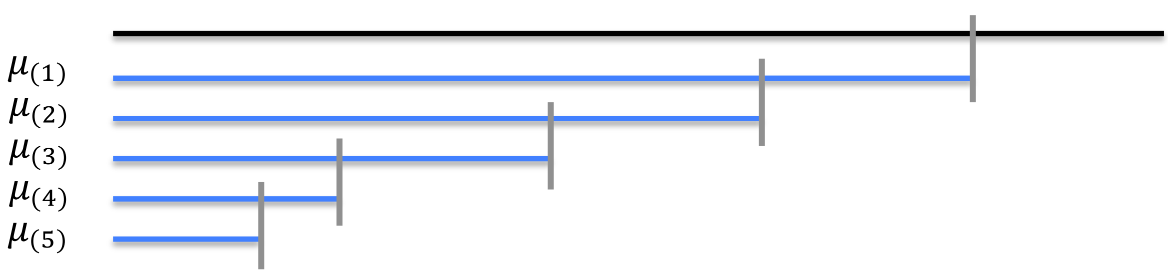
$$\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$$

where each $\mu_l \sim Beta(\alpha/_K, 1)$.

- K⟶∞, the $\mu_{(k)}{}'s$ obey the following law:

$$\nu_{(k)} \sim Beta(\alpha, 1) \qquad \mu_{(k)} = \nu_{(k)}\mu_{(k-1)} = \prod_{l=1}^{k} \nu_{(l)}$$

- Metaphorical representation:



Derivation

- Start from $\mu_{(1)} = \max_{l=1,\ldots,K} \mu_l$ where each μ_l is $\text{Beta}(\frac{\alpha}{K}, 1)$ and has density:

  $$p(\mu_l) = \frac{\alpha}{K}\mu_l^{\frac{\alpha}{K}-1}\mathbb{I}(0 \leq \mu_l \leq 1)$$

- cdf for μ_l

  $$F(\mu_l) = \int_{-\infty}^{\mu_l} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\mathbb{I}(0 \leq t \leq 1)\, dt$$
  $$= \mu_l^{\frac{\alpha}{K}}\mathbb{I}(0 \leq \mu_l \leq 1) + \mathbb{I}(1 < \mu_l)$$

- cdf for $\mu_{(1)}$

  $$F(\mu_{(1)}) = \left(\mu_{(1)}^{\frac{\alpha}{K}}\mathbb{I}(0 \leq \mu_{(1)} \leq 1) + \mathbb{I}(1 < \mu_{(1)} < \infty)\right)^K$$
  $$= \mu_{(1)}^{\alpha}\mathbb{I}(0 \leq \mu_{(1)} \leq 1) + \mathbb{I}(1 < \mu_{(1)}) \qquad (9)$$

- Differentiate $p(\mu_{(1)}) = \alpha\mu_{(1)}^{\alpha-1}\mathbb{I}(0 \leq \mu_{(1)} \leq 1)$

Derivation

- Considering $\mu_{(k)}$'s.

$$\mathbf{L}_k$$

$$\boxed{\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(k)}} > \boxed{\mu_{(k+1)} > \mu_{(k+2)} > \cdots > \mu_{(K)}}$$

for each $l \in \mathbf{L}_k$ $\quad \mu_l \leq \min_{k' \leq k} \mu_{(k')} = \mu_{(k)}$

- CDF for μ_l

where each μ_l is $\mathrm{Beta}(\frac{\alpha}{K}, 1)$ and has density:

$$p(\mu_l) = \frac{\alpha}{K}\mu_l^{\frac{\alpha}{K}-1}\mathbb{I}(0 \leq \mu_l \leq 1)$$

$$F(\mu_l|\mu_{(1:k)}) = \frac{\int_0^{\mu_l} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\, dt}{\int_0^{\mu_{(k)}} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\, dt}$$

$$= \mu_{(k)}^{-\frac{\alpha}{K}} \mu_l^{\frac{\alpha}{K}} \mathbb{I}(0 \leq \mu_l \leq \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_l)$$

Derivation

- CDF for μ_l

$$F(\mu_l|\mu_{(1:k)}) = \frac{\int_0^{\mu_l} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\,dt}{\int_0^{\mu_{(k)}} \frac{\alpha}{K} t^{\frac{\alpha}{K}-1}\,dt}$$

$$= \mu_{(k)}^{-\frac{\alpha}{K}} \mu_l^{\frac{\alpha}{K}} \mathbb{I}(0 \le \mu_l \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_l)$$

- CDF for $\mu_{(k+1)} = \max_{l \in \mathbf{L}_k} \mu_l$

$$F(\mu_{(k+1)}|\mu_{(1:k)}) \tag{13}$$

$$= \mu_{(k)}^{-\frac{K-k}{K}\alpha} \mu_{(k+1)}^{\frac{K-k}{K}\alpha} \mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_{(k+1)})$$

$$\rightarrow \mu_{(k)}^{-\alpha} \mu_{(k+1)}^{\alpha} \mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)}) + \mathbb{I}(\mu_{(k)} < \mu_{(k+1)})$$

- Differentiate the density of $\mu_{(k+1)}$

$$p(\mu_{(k+1)}|\mu_{(1:k)})$$

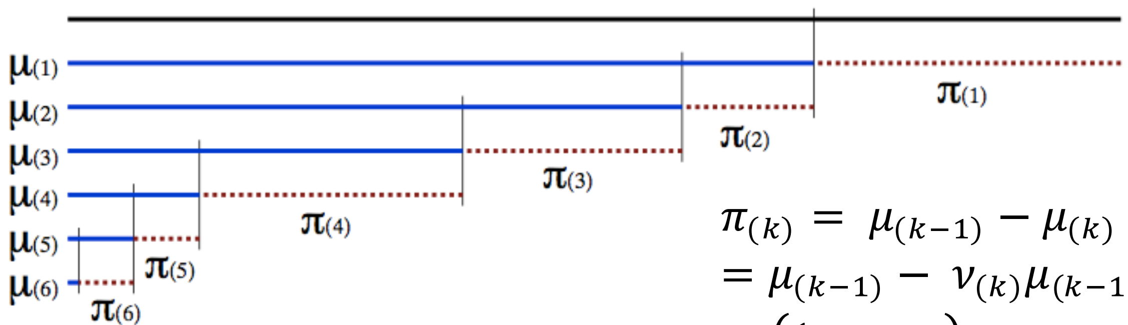
$$= \alpha \mu_{(k)}^{-\alpha} \mu_{(k+1)}^{\alpha-1} \mathbb{I}(0 \le \mu_{(k+1)} \le \mu_{(k)})$$

$$p(\nu_{(k)}|\mu_{(1:k-1)}) = \alpha \nu_{(k)}^{\alpha-1} \mathbb{I}(0 \le \nu_{(k)} \le 1)$$

Relation To Dirichlet Process (DP)

- Stick-breaking for IBP:

$$\nu_{(k)} \sim Beta(\alpha, 1) \qquad \mu_{(k)} = \nu_{(k)}\mu_{(k-1)} = \prod_{l=1}^{k} \nu_{(l)}$$



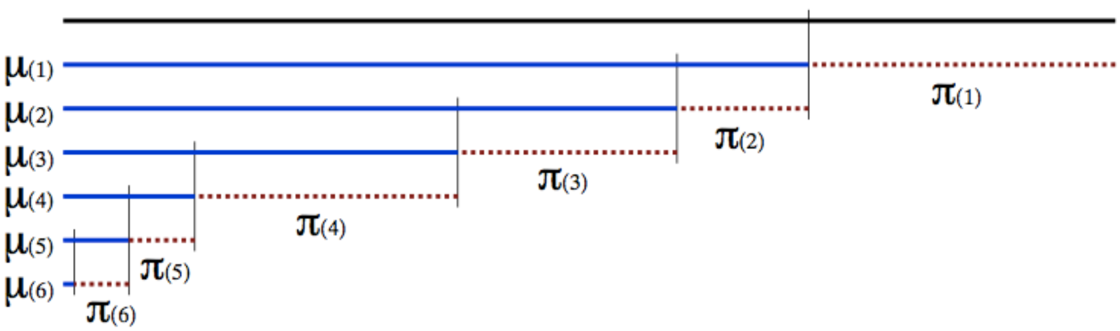
$$\pi_{(k)} = \mu_{(k-1)} - \mu_{(k)}$$
$$= \mu_{(k-1)} - \nu_{(k)}\mu_{(k-1)}$$
$$= \left(1 - \nu_{(k)}\right)\mu_{(k-1)}$$
$$= \left(1 - \nu_{(k)}\right)\prod_{l=1}^{k-1} \nu_{(l)}$$

- Stick-breaking for DP:

$$w_{(k)} = 1 - \nu_{(k)} \qquad w_{(k)} \sim Beta(1, \alpha) \qquad \pi_{(k)} = w_{(k)} \prod_{l=1}^{k-1}\left(1 - w_{(l)}\right)$$

Relation To Dirichlet Process (DP)

- Different properties:
  - DPs: stick lengths sum to a length of 1, and not decreasing
  - IBPs: stick lengths need not sum to 1, but decreasing



$$\mu_{(k)} = \nu_{(k)}\mu_{(k-1)} = \prod_{l=1}^{k} \nu_{(l)}$$

$$\pi_{(k)} = w_{(k)} \prod_{l=1}^{k-1} (1 - w_{(l)})$$

The correspondence to stick-breaking in DPs implies that a range of techniques for DP can be adapted for the IBP. E.g. Pitman-Yor of the IBP, truncated stick-breaking construction

Adapt truncated stick-breaking for the DP to the IBP

- Let K* be the truncation level.
- Set $\mu_{(k)} = 0$ for each k > K* while the joint density of $\mu_{(1:K^*)}$:

$$p(\mu_{(1:K^*)}) = \prod_{k=1}^{K^*} p(\mu_{(k)} | \mu_{(k-1)})$$

- The conditional distribution of Z given $\mu_{(1:K^*)}$:

$$p(Z | \mu_{(1:K^*)}) = \prod_{i=1}^{N} \prod_{k=1}^{K^*} \mu_{(k)}^{z_{ik}} (1 - \mu_{(k)})^{1 - z_{ik}}$$

- Gibbs sampling in the truncated stick-breaking construction is simple to implement, however…

Slice Sampler

- The truncated stick-breaking construction
  - Predetermined truncation level
  - Approximation scheme

- Proposing a non-approximate scheme based on SLICE Sampling.
  - Choosing the truncation level adaptively at each iteration

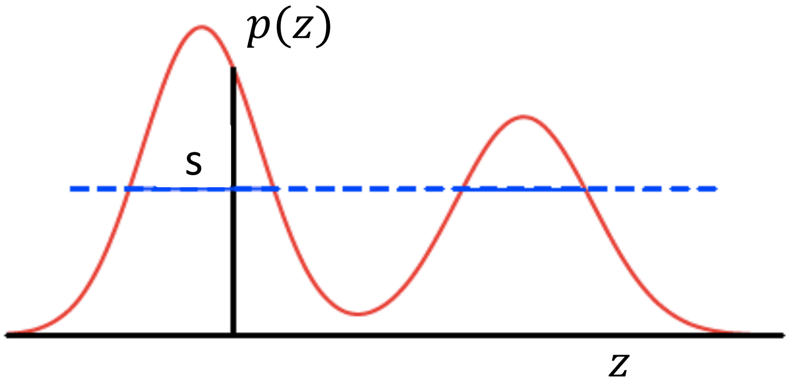
Slice Sampler

- Suppose we wish to sample a new value for the variable of interest z from some distribution p(z)

- The key concept is to introduce an auxiliary variable ***s*** does not change the underlying distribution, i.e.

$$\int_s p(s, z) ds = p(z)$$

- Alternatively sample z and s,
  - Given z, sample s uniformly from the range $0 \leq s \leq p(z)$
  - Given s, sample a new value for the variables of interest **z**, considering only **z** such that $p(z) > s$



Slice Sampler for IBP

- Draw ***s***

$$s | Z, \mu_{(1:\infty)} \sim \text{Uniform}[0, \mu^*] \qquad \mu^* = \min\left\{1, \min_{k:\, \exists i, z_{ik}=1} \mu_{(k)}\right\}$$

μ*: last active (used feature)

- Given s, the distribution of *Z* :

$$p(Z|\mathbf{x}, s, \mu_{(1:\infty)}) \propto p(Z|\mathbf{x}, \mu_{(1:\infty)})\, p(s|Z, \mu_{(1:\infty)})$$

$$\propto p(Z|\mathbf{x}, \mu_{(1:\infty)})\, \frac{1}{\mu^*} \mathbb{I}(0 \leq s \leq \mu^*)$$

We need only consider updating those features k, where $\mu_{(k)} > s$.

- $z_{ik} = 0$ where $\mu_{(k)} < s$

Slice Sampler for IBP

- Draw ***s*** (updating ***s***)

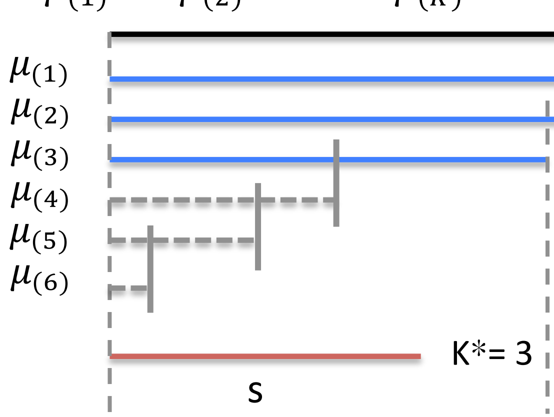
$$s | Z, \mu_{(1:\infty)} \sim \text{Uniform}[0, \mu^*] \qquad \mu^* = \min\left\{1, \min_{k:\, \exists i, z_{ik}=1} \mu_{(k)}\right\}$$

μ*: last active (used feature)

- In IBP stick-breaking:

Active (used) features (dishes)
Inactive (unused) features (dishes)

$$\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$$



K*: maximal feature index with where $\mu_{(K^*)} > s$.
Only consider updating those features $k \leq K^*$,

- $z_{ik} = 0$ for $k > K^*$, i.e. $\mu_{(k)} < s$

Slice Sampler for IBP

- Draw ***s*** (updating ***s***)

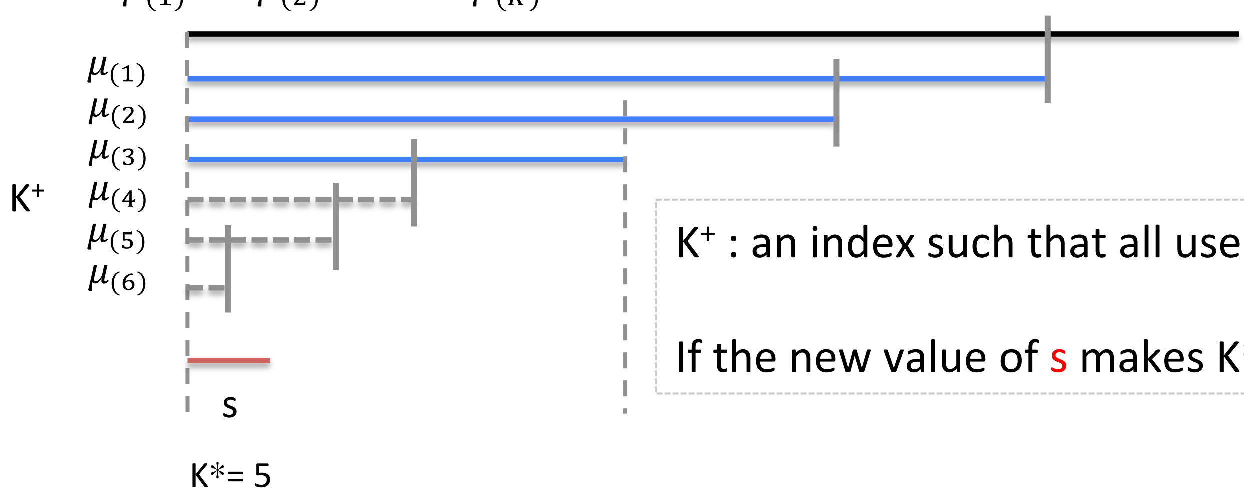
$$s|Z, \mu_{(1:\infty)} \sim \text{Uniform}[0, \mu^*] \qquad \mu^* = \min\left\{1, \min_{k:\, \exists i, z_{ik}=1} \mu_{(k)}\right\}$$

μ*: last active (used feature)

- In IBP stick-breaking:

Active (used) features (dishes)
Inactive (unused) features (dishes)

$$\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$$



K^+ : an index such that all used features have index $k < K^+$

If the new value of s makes $K^* \geq K^+$, i.e. $s < \mu_{(K+)}$ …

Slice Sampler for IBP

- Draw ***s*** (updating ***s***)

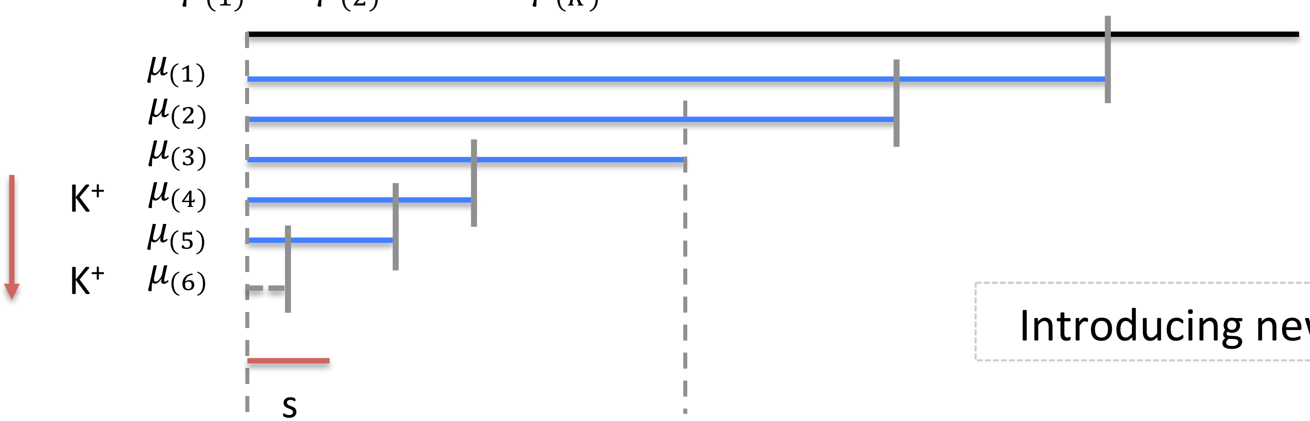
$$s|Z, \mu_{(1:\infty)} \sim \text{Uniform}[0, \mu^*] \qquad \mu^* = \min\left\{1, \min_{k:\ \exists i, z_{ik}=1} \mu_{(k)}\right\}$$

μ*: last active (used feature)

- In IBP stick-breaking:

Active (used) features (dishes)
Inactive (unused) features (dishes)

$$\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$$



K*= 5

Introducing new features!

Slice Sampler for IBP

- Draw ***s*** (updating ***s***)

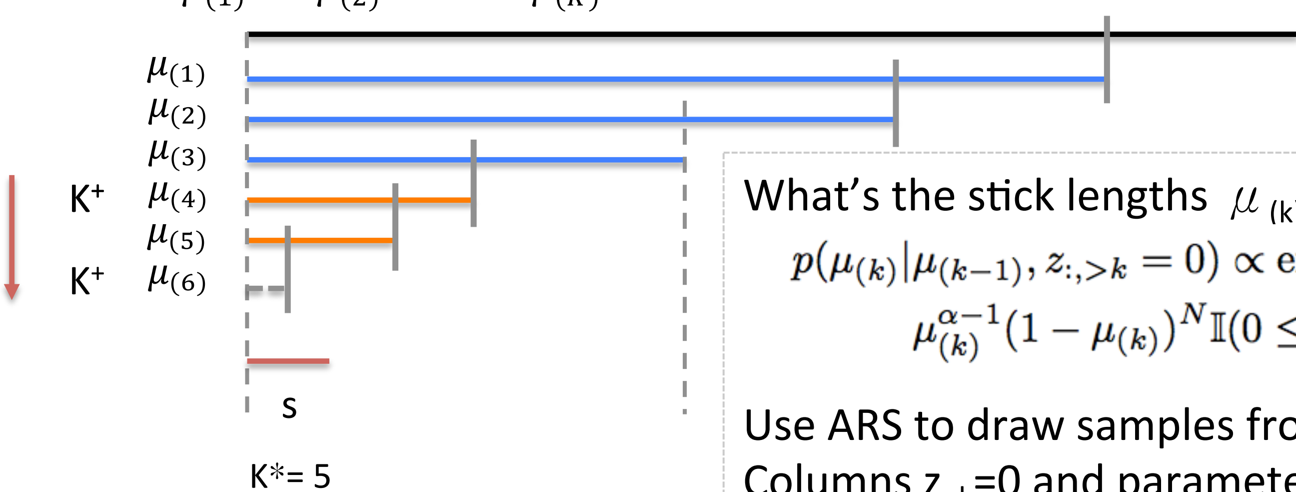
$$s|Z, \mu_{(1:\infty)} \sim \text{Uniform}[0, \mu^*] \qquad \mu^* = \min\left\{1, \min_{k:\, \exists i, z_{ik}=1} \mu_{(k)}\right\}$$

μ*: last active (used feature)

- In IBP stick-breaking:

Active (used) features (dishes)
Inactive (unused) features (dishes)

$$\mu_{(1)} > \mu_{(2)} > \cdots > \mu_{(K)}$$



What's the stick lengths $\mu_{(k)}$ for the new features k?

$$p(\mu_{(k)}|\mu_{(k-1)}, z_{:,>k}=0) \propto \exp(\alpha \textstyle\sum_{i=1}^{N} \frac{1}{i}(1-\mu_{(k)})^i)$$
$$\mu_{(k)}^{\alpha-1}(1-\mu_{(k)})^N \mathbb{I}(0 \le \mu_{(k)} \le \mu_{(k-1)})$$

Use ARS to draw samples from the distribution.
Columns $z_{:,k}$=0 and parameters $\theta_k \sim H$

Slice Sampler for IBP

Updating *Z*

- Given s, we only update z_{ik} for each *i* and $k \leq K^*$

$$p(z_{ik} = 1|\text{rest}) \propto \frac{\mu_{(k)}}{\mu^*} f(x_i|z_{i,\neg k}, z_{ik} = 1, \theta_{1:K^\dagger})$$

μ*: last active (used feature)

Updating θ_k

- for k=1,…,K^+

$$p(\theta_k|\text{rest}) \propto h(\theta_k) \prod_{i=1}^{N} f(x_i|z_{i,1:K^\dagger}, \theta_{\neg k}, \theta_k)$$

Slice Sampler for IBP

Updating $\mu_{(k)}$

- for k=1,…,K^+-1 (Active features)

$$p(\mu_{(k)}|\text{rest}) \propto \mu_{(k)}^{m_{\cdot k}-1}(1-\mu_{(k)})^{N-m_{\cdot k}}$$
$$\mathbb{I}(\mu_{(k+1)} \le \mu_{(k)} \le \mu_{(k-1)})$$

$$m_{\cdot k} = \sum_{i=1}^{N} z_{ik}$$

- For k=K^+ (Inactive features)

$$p(\mu_{(k)}|\mu_{(k-1)}, z_{:,>k}=0) \propto \exp(\alpha \sum_{i=1}^{N} \tfrac{1}{i}(1-\mu_{(k)})^i)$$
$$\mu_{(k)}^{\alpha-1}(1-\mu_{(k)})^N \mathbb{I}(0 \le \mu_{(k)} \le \mu_{(k-1)})$$

Change of Representations

- IBP – ignoring the ordering on features;
- Stick-breaking IBP – enforcing an ordering with decreasing weights.

- Stick-breaking IBP to IBP:
  - Drop the stick lengths and the inactive features,
  - leaving only the K^+ active feature columns along with the corresponding parameters.

- IBP to stick-breaking IBP:
  - Draw both the stick lengths and order the features in decreasing stick lengths,
  - Introducing inactive features K° into the representation

IBP to stick-breaking IBP

- We have K^+ active features in the IBP,
  - Feature occurrence matrix: $Z_{1:N,1:K+}$
  - Suppose we have $K >> K^+$ features
  - For the active features, the posterior for the lengths are

$$\mu_k^+ | z_{:,k} \sim \text{Beta}(m_{\cdot,k}, 1 + N - m_{\cdot,k})$$

  - For the rest of the $K-K^+$ inactive features
    - Consider only those inactive features with stick lengths larger than $\min_k \mu_k^+$
  - Reorder $\mu^+_{(1:K+)}$, $\mu^\circ_{(1:K^\circ)}$ in decreasing order

$\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(k)}$

$\mu_{(k+1)} > \mu_{(k+2)} > \cdots > \mu_{(K)}$

Semi-ordered Stick-breaking

- μ_k^+ on active features are **unordered** and draw from the following distribution:

$$\mu_k^+ | z_{:,k} \sim \text{Beta}(m_{\cdot,k}, 1 + N - m_{\cdot,k})$$

- The stick length on inactive feature is similar to the stick-breaking IBP:

$$p(\mu_{(k)}^\circ | \mu_{(k-1)}^\circ, z_{:,>k} = 0) \propto \exp\left(\sum_{i=1}^{N} \frac{1}{i}(1 - \mu_{(k)}^\circ)^i\right) (\mu_{(k)}^\circ)^{\alpha-1}(1 - \mu_{(k)}^\circ)^N \mathbb{I}(0 \leq \mu_{(k)}^\circ \leq \mu_{(k-1)}^\circ) \quad (31)$$

- The auxiliary variable ***s*** determines how many inactive features need to add

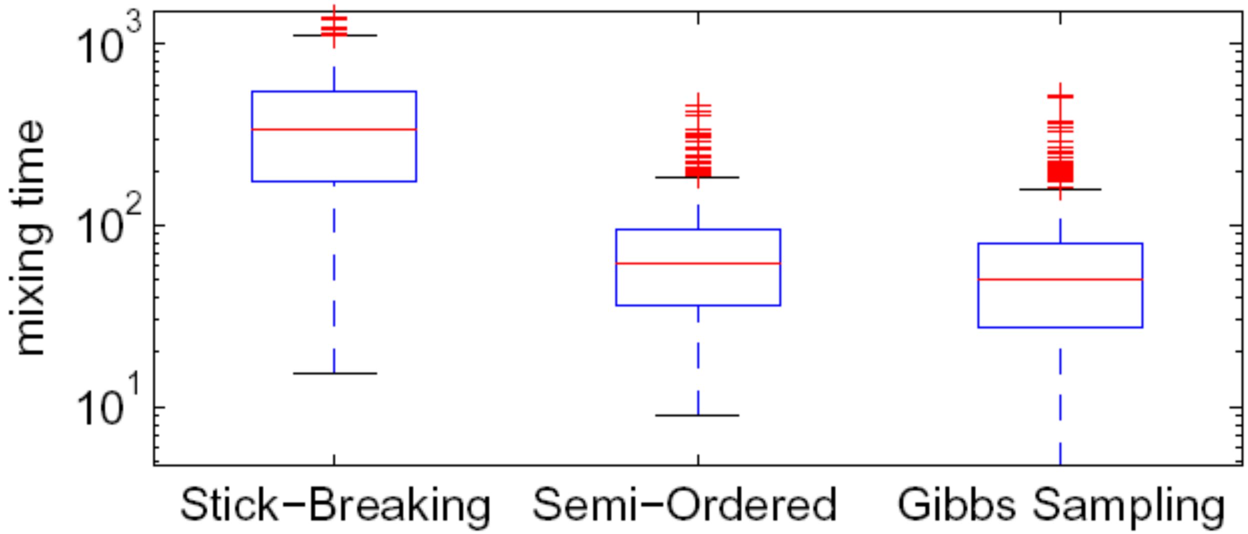
$$s \sim \text{Uniform}[0, \mu^*] \quad \mu^* = \min\left\{1, \min_{1 \leq k \leq K^+} \mu_k^+\right\} \quad (32)$$

- Drop from the list of active features any that become inactive and add to the list any inactive feature that became active
- New list of active features are drawn from

$$\mu_k^+ | z_{:,k} \sim \text{Beta}(m_{\cdot,k}, 1 + N - m_{\cdot,k})$$

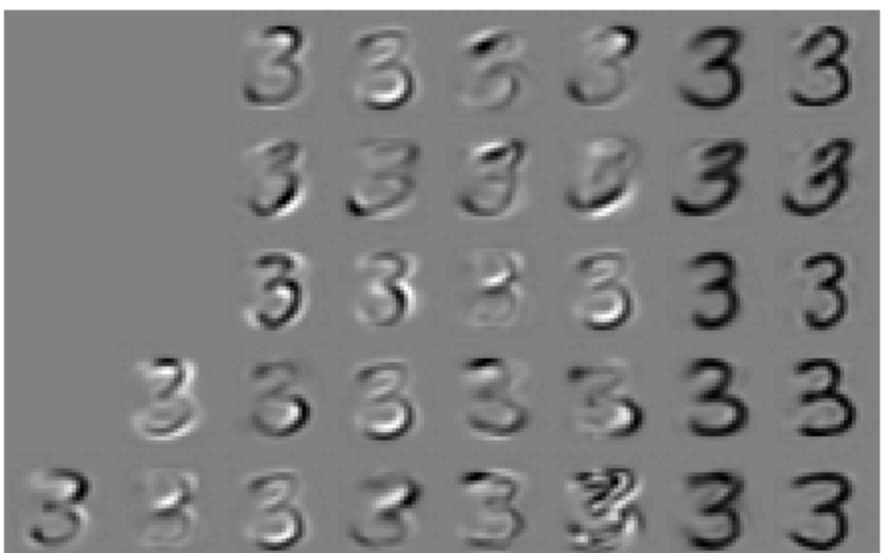
Results

- Use the conjugate linear-Gaussian binary latent feature model for comparing the performance of the different samplers.



Demonstration

- Apply semi-ordered slice sampler to 1000 examples of handwritten images of 3’s in the MNIST dataset.



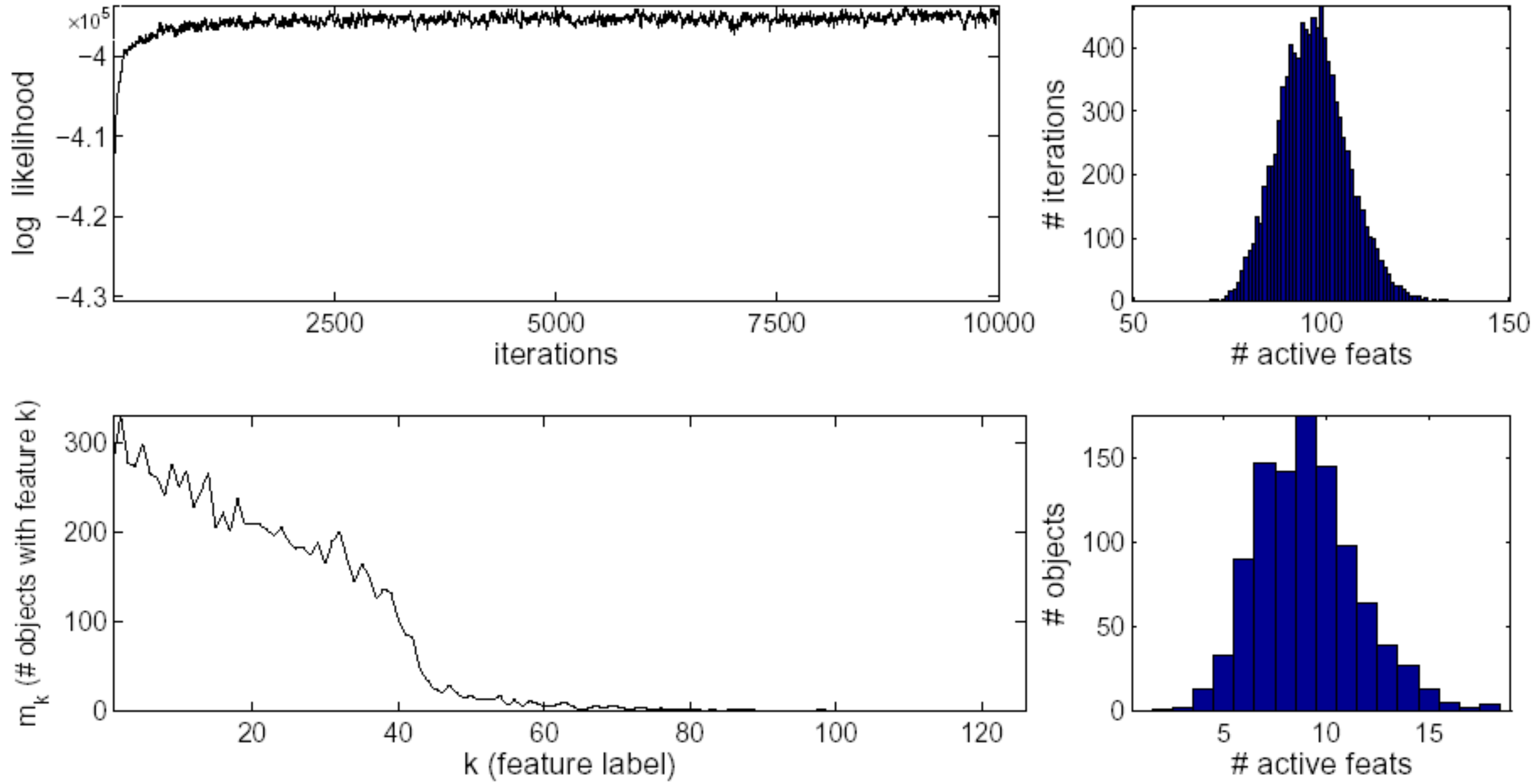


Figure 3: *Top-left*: the log likelihood trace plot. The sampler quickly finds a high likelihood region. *Top-right*: histogram of the number of active features over the 10000 iterations. *Bottom-left*: number of images sharing each feature during the last MCMC iteration. *Bottom-right*: histogram of the number of active features used by each input image. Note that about half of the features are used by only a few data points, and each data point is represented by a small subset of the active features.

Conclusions

- Derived novel stick-breaking representations of the IBP
- New MCMC samplers are proposed based on the new representations.
- The new samplers show as efficient as Gibbs without using conjugacy.