Hierarchical Bayesian Nonparametric Models

HDP, HPYP, Sequence Memoizer

Jan Gasthaus & Yee Whye Teh

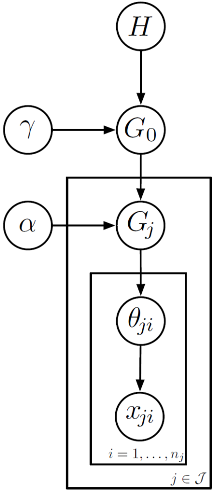
Bayesian Nonparametrics Course

Apr 13th, 2012

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

$$
\begin{aligned}
G_0 &\sim \mathcal{DP}(\gamma, H) \\
G \mid G_0 &\sim \mathcal{DP}(\alpha, G_0)
\end{aligned}
$$

$$G_0 \sim \mathcal{DP}(\gamma, H)$$
$$G_j \mid G_0 \sim \mathcal{DP}(\alpha_j, G_0) \quad j = 1, \dots, J$$
$$\theta_{ij} \mid G_j \sim G_j \qquad i = 1, \dots, N_J$$
$$x_{ij} \mid \theta_j \sim F(\theta_{ij})$$



$$
\begin{aligned}
G_0 &\sim \mathcal{PY}(\gamma, H) \\
G_1 \mid G_0 &\sim \mathcal{PY}(\alpha_1, G_0) \\
G_2 \mid G_1 &\sim \mathcal{PY}(\alpha_2, G_1) \\
&\vdots
\end{aligned}
$$

Outline

1. Hierarchical Dirichlet Processes
   - Representations: Stick-breaking and Chinese Restaurant Franchise
   - Prominent Models
     - HDP-LDA
     - Infinite HMM
2. Hierarchical Pitman-Yor Processes
   - Representations
3. Sequence Memoizer
   - Model
   - Coagulation-Fragmentation Properties
4. Inference

Hierarchical Dirichlet Processes

- Main idea: make the base measure of a DP a draw from another DP:

$$
\begin{aligned}
G_0 &\sim \mathcal{DP}(\gamma, H) \\
G_j \mid G_0 &\sim \mathcal{DP}(\alpha_j, G_0) \qquad\qquad j = 1, \ldots, J
\end{aligned}
$$

Hierarchical Dirichlet Processes

- Main idea: make the base measure of a DP a draw from another DP:

$$G_0 \sim \mathcal{DP}(\gamma, H)$$
$$G_j \mid G_0 \sim \mathcal{DP}(\alpha_j, G_0) \qquad\qquad j = 1, \ldots, J$$

- Induces sharing of atoms among the G_j
  - Atoms are are inherited from G_0
  - Each G_j has a distinct set of weights associated with the atoms

HDP: Stick-breaking Representation

- Stick-breaking representation of the DP $G_0 \sim \mathcal{DP}(\gamma, H)$:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}$$

where for $k = 1, 2 \ldots$

$$\nu_k \sim \text{Beta}(1, \gamma) \qquad \beta_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l) \qquad \theta_k^{**} \sim H$$

HDP: Stick-breaking Representation

- Stick-breaking representation of the DP $G_0 \sim \mathcal{DP}(\gamma, H)$:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}$$

- The support of each G_j is contained within the support of G_0, so that for each $j = 1, \ldots, J$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^{**}}$$

- What is the relationship between β and π_j?

HDP: Stick-breaking Representation

- Stick-breaking representation

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}} \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^{**}}$$

- Interpreting $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_j$ as discrete probability measures on $\{1, 2, \ldots\}$ we have

$$\boldsymbol{\pi}_j \mid \boldsymbol{\beta} \sim \mathcal{DP}(\alpha_j, \boldsymbol{\beta})$$

- Using the defining property of the DP, we can explicitly construct π_{jk} given β_k as follows:

$$\nu_{jk} \sim \text{Beta}\left(\alpha\beta_k, \alpha\left(1 - \sum_{l=1}^{k} \beta_l\right)\right) \qquad \pi_{jk} = \nu_k \prod_{l=1}^{k-1} (1 - \nu_{jl})$$

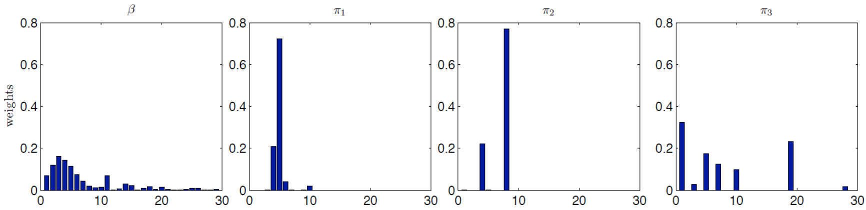
HDP: Stick-breaking Representation

- The weights are equal to the base distribution in expectation

$$E[\pi_{jk}] = E[\beta_k] = \gamma^{k-1}(1+\gamma)^{-k}$$

- However, the variance of the weight is higher, typically leading to “sparser” $\boldsymbol{\pi}_j$

$$\mathrm{Var}[\pi_{jk}] = E\left[\frac{\beta_k(1-\beta_k)}{1+\alpha}\right] + \mathrm{Var}[\beta_k] > \mathrm{Var}[\beta_k]$$



HDP: Chinese Restaurant Franchise

- The CRP describes the marginal distribution of draws $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$ with G integrated out

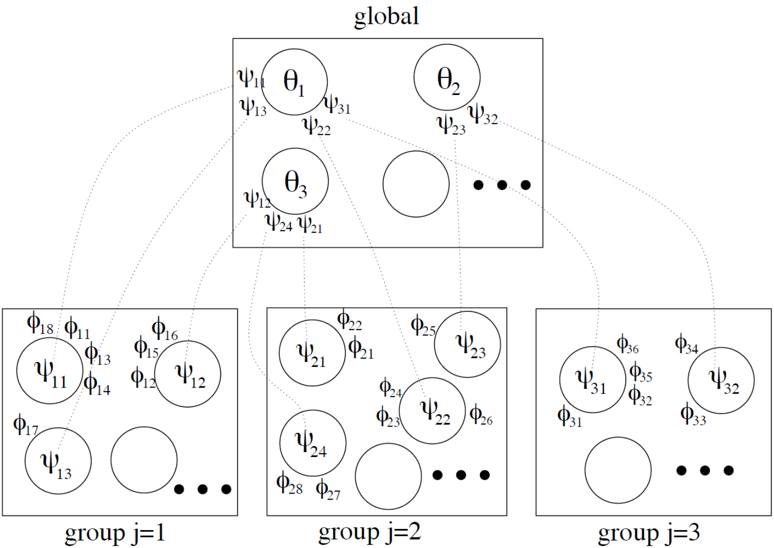
HDP: Chinese Restaurant Franchise

- The CRP describes the marginal distribution of draws $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$ with G integrated out
- The CRF extends the Chinese Restaurant metaphor for draws from a hierarchical model $G_0 \sim \mathcal{DP}(\gamma, H)$ and $G_j \mid G_0 \sim \mathcal{DP}(\alpha_j, G_0)$

HDP: Chinese Restaurant Franchise

- The CRP describes the marginal distribution of draws $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$ with G integrated out
- The CRF extends the Chinese Restaurant metaphor for draws from a hierarchical model $G_0 \sim \mathcal{DP}(\gamma, H)$ and $G_j \mid G_0 \sim \mathcal{DP}(\alpha_j, G_0)$
- The idea is to have a "franchise" with a shared menu of dishes
- In each restaurant, dishes are chose with probability proportional to the total number of tables serving them (in the entire franchise)

HDP: Chinese Restaurant Franchise



HDP: Chinese Restaurant Franchise

- Some notation
  - i-th customer in j-th restaurant $\theta_{ji} \sim G_j$
  - t-th table in j-th restaurant $\theta^*_{jt} \sim G_0$
  - k-th dish $\theta^{**}_k \sim H$
  - Customer i in restaurant j sits at table t_{ji} and table t serves dish k_{jt}
  - $\theta_{ji} = \theta^*_{jt_{ji}} = \theta^{**}_{k_{jt_{ji}}}$
  - n_{jtk} number of customers in restaurant j around table t serving dish k
  - m_{jk} number of tables in restaurant j serving dish k

HDP: Chinese Restaurant Franchise

- Recall the CRP for the DP $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$:

$$\theta_i | \theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha}{\alpha + n_\cdot} H + \sum_{t=1}^{T} \frac{n_t}{\alpha + n_\cdot} \delta^*_{\theta_t}$$

HDP: Chinese Restaurant Franchise

- Recall the CRP for the DP $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$:

$$\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha}{\alpha + n_\cdot} H + \sum_{t=1}^{T} \frac{n_t}{\alpha + n_\cdot} \delta^*_{\theta_t}$$

- In the HDP, integrating out the G_j we have similarily:

$$\theta_{ji} \mid \theta_{j1}, \ldots, \theta_{ji-1}, G_0 \sim \frac{\alpha_j}{\alpha_j + n_{j\cdot\cdot}} G_0 + \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha_j + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}}$$

HDP: Chinese Restaurant Franchise

- Recall the CRP for the DP $\theta_i \sim G$, $G \sim \mathcal{DP}(\alpha, H)$:

$$\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha}{\alpha + n_{\cdot}} H + \sum_{t=1}^{T} \frac{n_t}{\alpha + n_{\cdot}} \delta^*_{\theta_t}$$

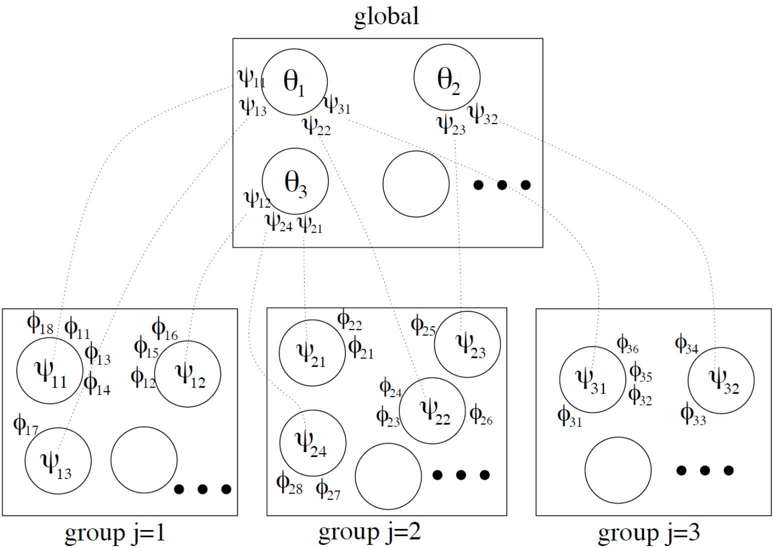
- In the HDP, integrating out the G_j we have similarily:

$$\theta_{ji} \mid \theta_{j1}, \ldots, \theta_{ji-1}, G_0 \sim \frac{\alpha_j}{\alpha_j + n_{j\cdot\cdot}} G_0 + \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha_j + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}}$$

- And for the customers in the higher-level restaurant

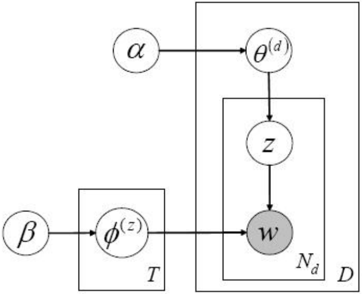
$$\theta^*_{jt} \mid \boldsymbol{\theta}^* \sim \frac{\gamma}{\gamma + m_{\cdot\cdot}} H + \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma + m_{\cdot\cdot}} \delta_{\theta^{**}_k}$$

HDP: Chinese Restaurant Franchise



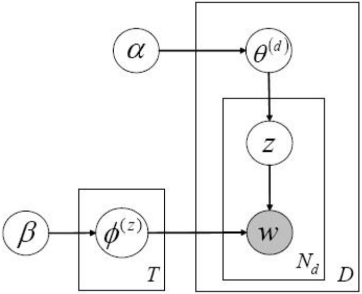
HDP-LDA

- Recall the standard LDA model



HDP-LDA

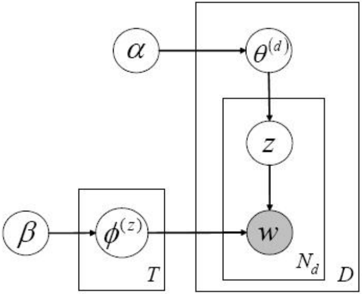
- Recall the standard LDA model



- Within each document, each word is drawn from a finite mixture model, where each mixture component is a distribution over words (a “topic”)
- The mixture components are shared between documents, but their weights differ.

HDP-LDA

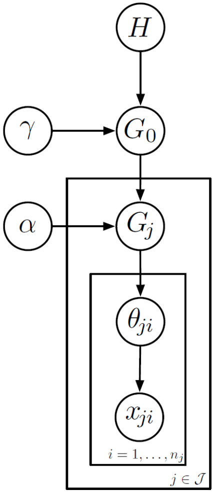
- Recall the standard LDA model



- Within each document, each word is drawn from a finite mixture model, where each mixture component is a distribution over words (a “topic”)
- The mixture components are shared between documents, but their weights differ.
- Can we take $T \to \infty$?

HDP-LDA

$$
\begin{aligned}
G_0 &\sim \mathcal{DP}(\gamma, H) & \\
G_j \mid G_0 &\sim \mathcal{DP}(\alpha_j, G_0) & j = 1, \ldots, J \\
\theta_{ij} \mid G_j &\sim G_j & i = 1, \ldots, N_J \\
x_{ij} \mid \theta_j &\sim F(\theta_{ij}) &
\end{aligned}
$$



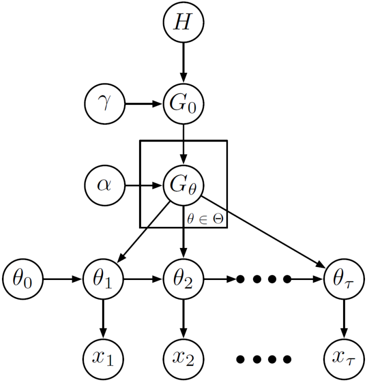
The Infinite HMM

- A traditional Hidden Markov Model is described by a set of states $\theta_1, \ldots, \theta_K$, a transition distribution $\pi(\theta_t|\theta_{t-1})$ and an emission distribution $f(x_t|\theta_t)$
- Note that this defines a set of mixture distributions – one for each state – with shared mixture components

The Infinite HMM

- We can define the iHMM as an infinite collection of DP draws G_θ with a common base measure G_0, representating the transition distributions.
- However, the description becomes clearer in the stick-breaking representation:

$$\theta_k^{**} \sim H$$
$$\boldsymbol{\beta} \sim \text{GEM}(\gamma)$$
$$\pi_{\theta_k^{**}} \sim \mathcal{DP}(\alpha, \boldsymbol{\beta})$$



Hierarchical Pitman-Yor Processes

- Same idea as with the HDP, but with a PYP:

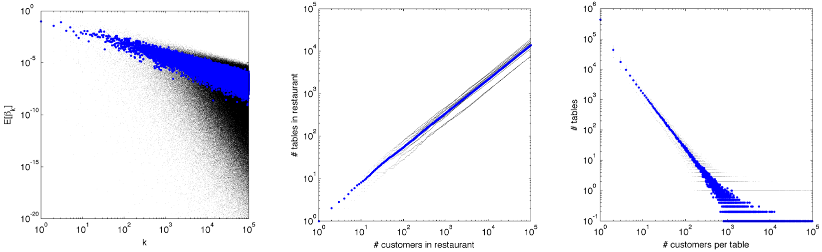
$$
\begin{aligned}
G_0 &\sim \mathcal{PY}(d_0, \alpha_0, H) \\
G_j \mid G_0 &\sim \mathcal{PY}(d_j, \alpha_j, G_0) \qquad\qquad j = 1, \ldots, J
\end{aligned}
$$

Hierarchical Pitman-Yor Processes

- Same idea as with the HDP, but with a PYP:

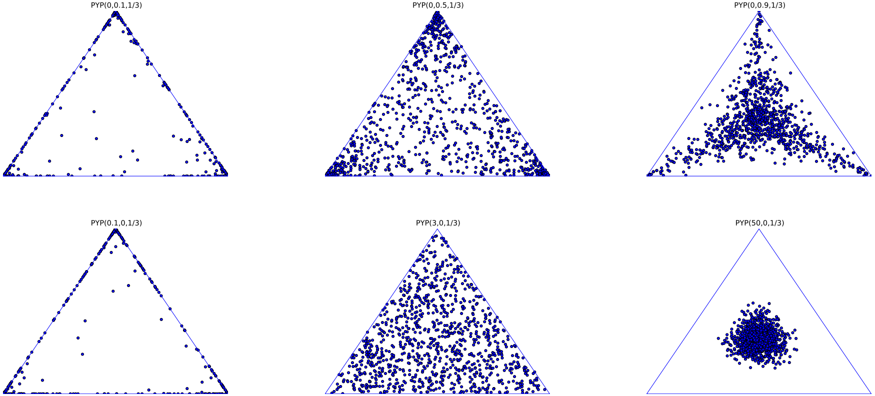
$$G_0 \sim \mathcal{PY}(d_0, \alpha_0, H)$$
$$G_j \mid G_0 \sim \mathcal{PY}(d_j, \alpha_j, G_0) \qquad j = 1, \ldots, J$$

- Useful if distributions have known power-law properties



Pitman-Yor Process

- What does $\text{PY}(G|\alpha, d, H)$ look like?
- No closed form expression, but can draw $G \sim \text{PY}(\alpha, d, H)$



HPYP: Stick-breaking Representation

- Stick-breaking representation of the PYP $G_0 \sim \mathcal{DP}(d, \alpha, H)$:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}$$

where for $k = 1, 2 \ldots$

$$\nu_k \sim \text{Beta}(1 - d, \alpha + kd) \qquad \beta_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l) \qquad \theta_k^{**} \sim H$$

HPYP: Chinese Restaurant Process

- Customers labeled $\{1, \ldots, c\}$ enter restaurant sequentially
- Customer i either joins other customers or sits at a new table

$$P(\text{join table } a) \propto |a| - d \qquad P(\text{new table}) \propto \alpha + |A_{i-1}|d$$

where $A_{i-1} \in \mathcal{A}_{i-1}$ is the current arrangement and $a \in A$

- Induces $\text{CRP}_c(\alpha, d)$, a distribution over $\mathcal{A}_c$
- Let $G \sim \text{PY}(\alpha, d, H)$ and $x_{1:c}|G \overset{\text{iid}}{\sim} G$; equivalently draw

$$A \sim \text{CRP}_c(\alpha, d) \qquad \theta_a \sim H \quad \text{for all } a \in A$$

and set $x_i = \theta_a$ for all $i \in a$.

HPYP: Chinese Restaurant Process

- CRP seating arrangement with c customers around t tables; $A \sim \text{CRP}_c(\alpha, d)$:

$$P(A) = \frac{[\alpha + d]_d^{|A|-1}}{[\alpha + 1]_1^{c-1}} \prod_{a \in A} [1 - d]_1^{|a|-1} \quad \text{for each } A \in \mathcal{A}_c, \qquad (1)$$

- CRP with fixed # of tables t; $A \sim \text{CRP}_{ct}(\alpha, d)$

$$P(A) = \frac{\prod_{a \in A} [1 - d]_1^{|a|-1}}{S_d(c, t)} \quad \text{for each } A \in \mathcal{A}_{ct},$$

- Normalization constant is a generalized Stirling number of type $(-1, -d, 0)$

Joint & Predictive Distribution

UCL

- Joint distribution of all seating arrangements

$$P(\{c_{\mathbf{u}s}, t_{\mathbf{u}s}, A_{\mathbf{u}s}\}, x_{1:T}) = \left(\prod_{s\in\Sigma} H(s)^{t_{\varepsilon s}}\right) \prod_{\mathbf{u}\in\Sigma^*} \left(\frac{[\alpha_{\mathbf{u}} + d_{\mathbf{u}}]_{d_{\mathbf{u}}}^{t_{\mathbf{u}\cdot}-1}}{[\alpha_{\mathbf{u}} + 1]_1^{c_{\mathbf{u}\cdot}-1}} \prod_{s\in\Sigma} \prod_{a\in A_{\mathbf{u}s}} [1 - d_{\mathbf{u}}]_1^{|a|-1}\right).$$

- Predictive distribution

$$P^*_{\mathbf{v}}(s) = \frac{c_{\mathbf{v}s} - t_{\mathbf{v}s} d_{\mathbf{v}}}{\alpha_{\mathbf{v}} + c_{\mathbf{v}\cdot}} + \frac{\alpha_{\mathbf{v}} + t_{\mathbf{v}\cdot} d_{\mathbf{v}}}{\alpha_{\mathbf{v}} + c_{\mathbf{v}\cdot}} P^*_{\sigma(\mathbf{v})}(s).$$

- The numbers of customers and tables have to satisfy the constraints

$$c_{\mathbf{u}s} = c^x_{\mathbf{u}s} + \sum_{\mathbf{v}:\sigma(\mathbf{v})=\mathbf{u}} t_{\mathbf{v}s}, \tag{2}$$

where $c^x_{\mathbf{u}s} = 1$ if $s = x_i$ and $\mathbf{u} = x_{1:i-1}$ for some i, and 0 otherwise.

HPYP Sequence Model

- Model for discrete sequences with power law properties
  - $\mathrm{P}(x_{1:N}) = \mathrm{P}(x_1) \prod_{i=2}^{N} \mathrm{P}(x_i | x_{1:i-1})$

HPYP Sequence Model

- Model for discrete sequences with power law properties
  - $\mathrm{P}(x_{1:N}) = \mathrm{P}(x_1) \prod_{i=2}^{N} \mathrm{P}(x_i | x_{1:i-1})$
- Directly estimate the set $\{\mathrm{P}(\cdot | x_{1:i-1})\}_{i=1,\ldots,N}$
- Treat distributions $\mathrm{P}(\cdot | x_{1:i-1})$ as random variables; call them $G_{[x_{1:i-1}]}(\cdot)$
  - $G_{[\mathbf{u}]}(t)$ = probability of observing symbol t in context $\mathbf{u}$

HPYP Sequence Model

- Model for discrete sequences with power law properties
  - $\mathrm{P}(x_{1:N}) = \mathrm{P}(x_1) \prod_{i=2}^{N} \mathrm{P}(x_i | x_{1:i-1})$
- Directly estimate the set $\{\mathrm{P}(\cdot | x_{1:i-1})\}_{i=1,\ldots,N}$
- Treat distributions $\mathrm{P}(\cdot | x_{1:i-1})$ as random variables; call them $G_{[x_{1:i-1}]}(\cdot)$
  - $G_{[\mathbf{u}]}(t)$ = probability of observing symbol t in context $\mathbf{u}$

1. Make prior assumptions about each individual G
   - Pitman-Yor process prior: $G \sim \mathrm{PY}(\alpha, d, H)$

HPYP Sequence Model

- Model for discrete sequences with power law properties
  - $\mathrm{P}(x_{1:N}) = \mathrm{P}(x_1) \prod_{i=2}^{N} \mathrm{P}(x_i | x_{1:i-1})$
- Directly estimate the set $\{\mathrm{P}(\cdot | x_{1:i-1})\}_{i=1,\dots,N}$
- Treat distributions $\mathrm{P}(\cdot | x_{1:i-1})$ as random variables; call them $G_{[x_{1:i-1}]}(\cdot)$
  - $G_{[\mathbf{u}]}(t)$ = probability of observing symbol t in context $\mathbf{u}$

1. Make prior assumptions about each individual G
   - Pitman-Yor process prior: $G \sim \mathrm{PY}(\alpha, d, H)$
2. Make use of hierarchical structure

HPYP Language Model

$$G_{[]} \mid d_0, \alpha_0, H \quad \sim \quad \mathrm{PY}(d_0, \alpha_0, H)$$

HPYP Language Model

$$
\begin{aligned}
G_{[]} \mid d_0, \alpha_0, H &\sim \mathrm{PY}(d_0, \alpha_0, H) \\
G_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]}) \quad \forall \mathbf{u} \in \bigcup_{k \leq m} \Sigma^k
\end{aligned}
$$

HPYP Language Model

$$
\begin{aligned}
G_{[]} \mid d_0, \alpha_0, H &\sim \mathrm{PY}(d_0, \alpha_0, H) \\
G_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]}) \quad \forall \mathbf{u} \in \bigcup_{k \leq m} \Sigma^k \\
x_i \mid \mathbf{x}_{i-m:i-1} = \mathbf{u} &\sim G_{[\mathbf{u}]} \qquad i = 1, \ldots, T
\end{aligned}
$$

HPYP Language Model

$$
\begin{aligned}
G_{[]} \mid d_0, \alpha_0, H &\sim \mathrm{PY}(d_0, \alpha_0, H) \\
G_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]}) \quad \forall \mathbf{u} \in \bigcup_{k \le m} \Sigma^k \\
x_i \mid \mathbf{x}_{i-m:i-1} = \mathbf{u} &\sim G_{[\mathbf{u}]} \qquad i = 1, \ldots, T
\end{aligned}
$$

$x_{1:5} = (\mathtt{o}, \mathtt{a}, \mathtt{c}, \mathtt{a}, \mathtt{c})$

```
[]o
[o]a
[oa]c
[oac]a
[aca]c
[cac]
```

HPYP Language Model

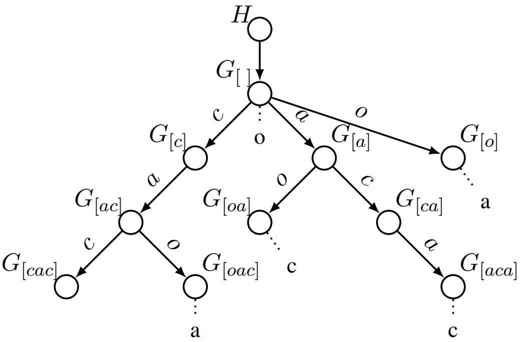
$$G_{[]} \mid d_0, \alpha_0, H \sim \mathrm{PY}(d_0, \alpha_0, H)$$

$$G_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]} \sim \mathrm{PY}(d_{|\mathbf{u}|}, \alpha_{|\mathbf{u}|}, G_{[\sigma(\mathbf{u})]}) \quad \forall \mathbf{u} \in \bigcup_{k \leq m} \Sigma^k$$

$$x_i \mid \mathbf{x}_{i-m:i-1} = \mathbf{u} \sim G_{[\mathbf{u}]} \quad i = 1, \ldots, T$$

$x_{1:5} = (\texttt{o}, \texttt{a}, \texttt{c}, \texttt{a}, \texttt{c})$

```
[]o
[o]a
[oa]c
[oac]a
[aca]c
[cac]
```



Sequence Memoizer: Details

1. At the root:

$$G_\varepsilon \mid \alpha_\varepsilon, d_\varepsilon, H \quad \sim \quad \mathrm{PY}(\alpha_\varepsilon, d_\varepsilon, H)$$

Sequence Memoizer: Details

1. At the root:

$$G_\varepsilon \mid \alpha_\varepsilon, d_\varepsilon, H \quad \sim \quad \mathrm{PY}(\alpha_\varepsilon, d_\varepsilon, H)$$

2. For all possible contexts $\mathbf{u} \in \Sigma^+$:

$$G_{[\mathbf{u}]} \mid \alpha_\mathbf{u}, d_\mathbf{u}, G_{[\sigma(\mathbf{u})]} \quad \sim \quad \mathrm{PY}(\alpha_\mathbf{u}, d_\mathbf{u}, G_{[\sigma(\mathbf{u})]})$$

Sequence Memoizer: Details

1. At the root:

$$G_\varepsilon \mid \alpha_\varepsilon, d_\varepsilon, H \quad \sim \quad \mathrm{PY}(\alpha_\varepsilon, d_\varepsilon, H)$$

2. For all possible contexts $\mathbf{u} \in \Sigma^+$:

$$G_{[\mathbf{u}]} \mid \alpha_{\mathbf{u}}, d_{\mathbf{u}}, G_{[\sigma(\mathbf{u})]} \quad \sim \quad \mathrm{PY}(\alpha_{\mathbf{u}}, d_{\mathbf{u}}, G_{[\sigma(\mathbf{u})]})$$

3. Draw observations from context-dependent distributions:

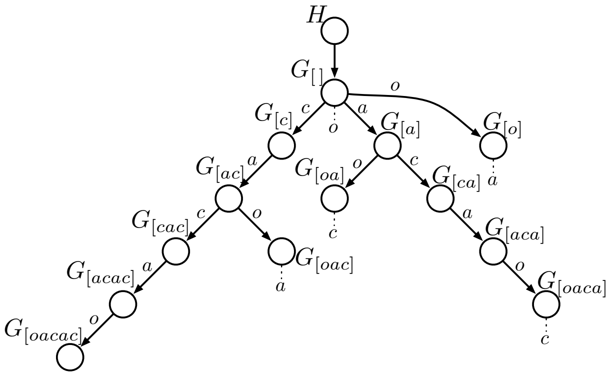
$$x_i \mid \mathbf{x}_{1:i-1} = \mathbf{u} \quad \sim \quad G_{[\mathbf{u}]} \qquad\qquad i = 1, \ldots, T$$

Sequence Memoizer: Illustration

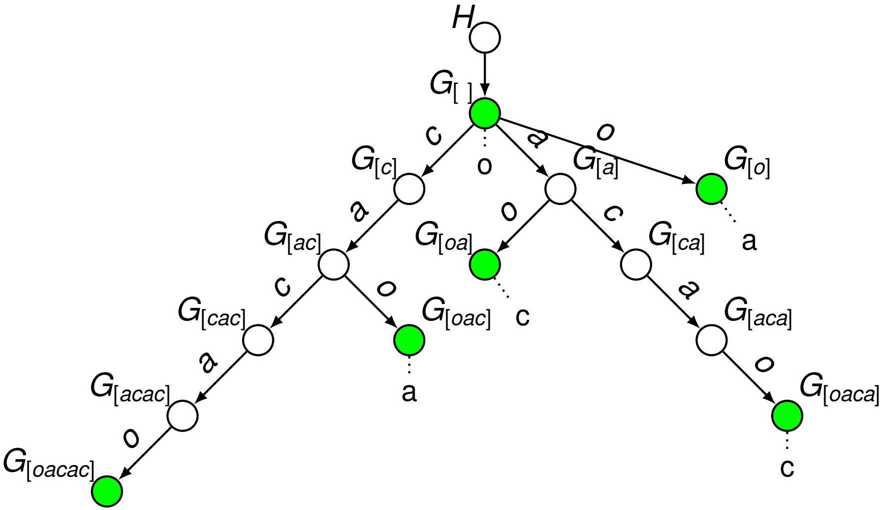
- Hierarchical prior over distributions arranged in a context tree
- Prior assumption $\mathrm{E}[G_{[\mathbf{u}]}(\cdot)|G_{[\sigma(\mathbf{u})]}] = G_{[\sigma(\mathbf{u})]}(\cdot)$

$x_{1:5} = (\mathtt{o}, \mathtt{a}, \mathtt{c}, \mathtt{a}, \mathtt{c})$

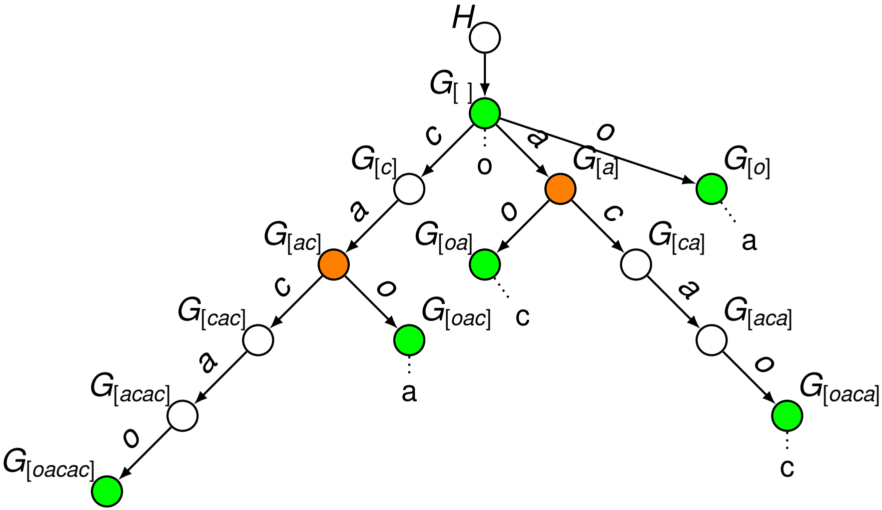
[ ] o
[o] a
[oa] c
[oac] a
[oaca] c
[oacac]



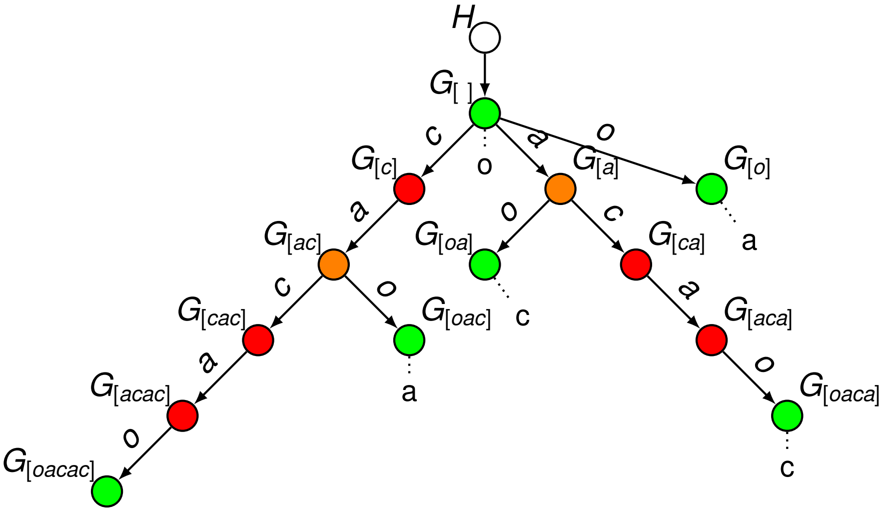
Marginalization: $\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$



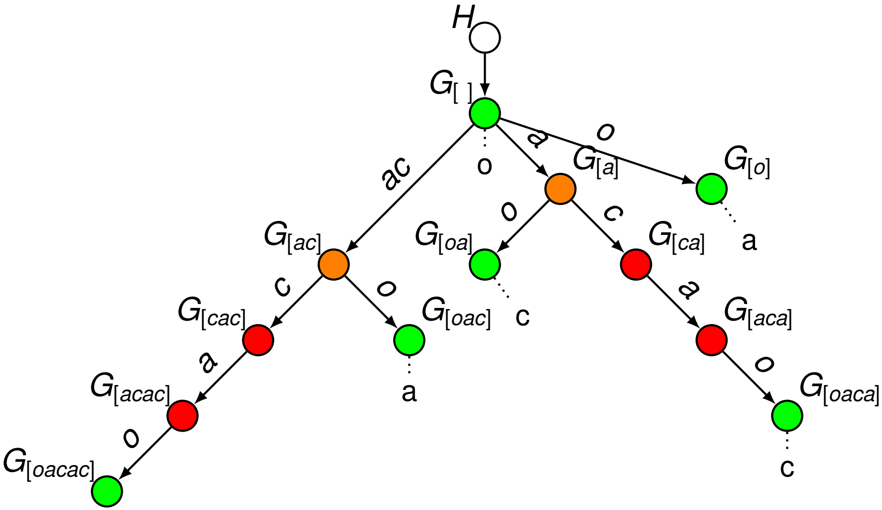
Marginalization: $\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$



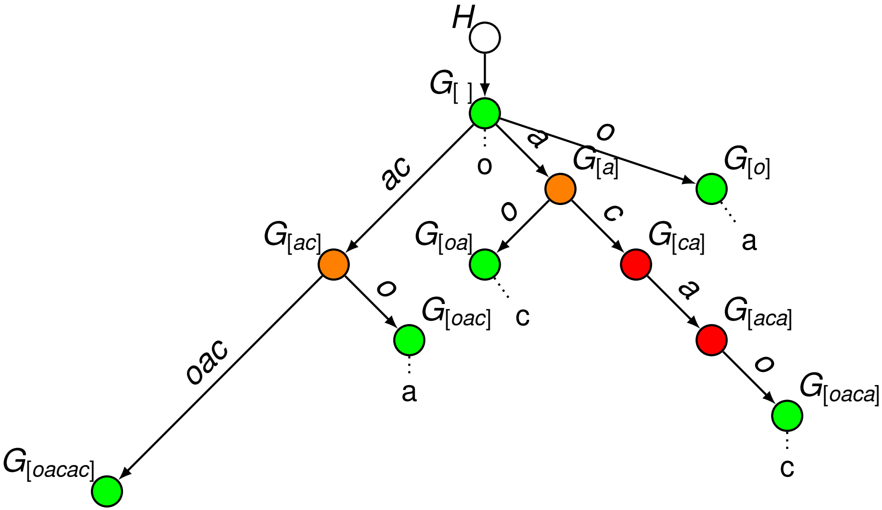
Marginalization: $\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$



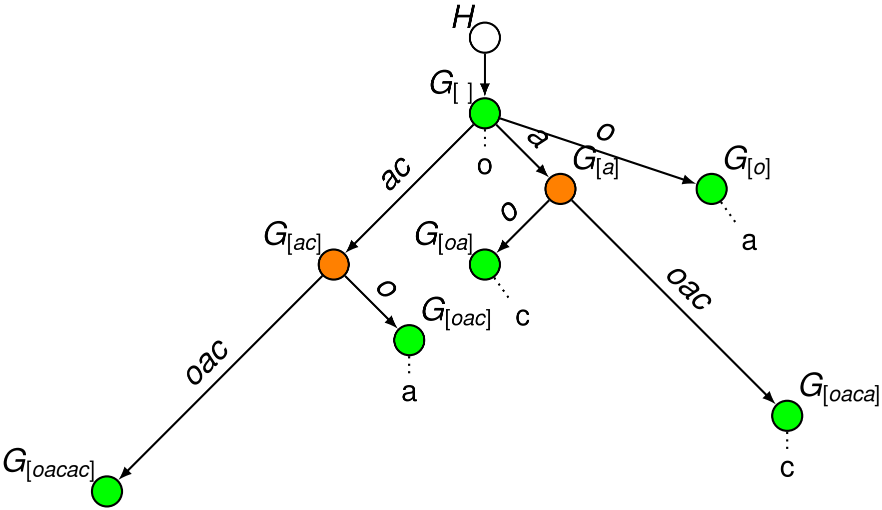
Marginalization: $\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$



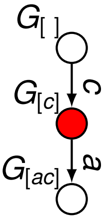
Marginalization: $\mathcal{O}(n^2) \to \mathcal{O}(n)$



Marginalization: $\mathcal{O}(n^2) \rightarrow \mathcal{O}(n)$



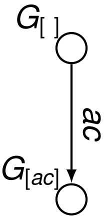
Marginalization



Theorem (Pitman, 1999; Ho et al., 2006):

If

$$G_{[c]} | G_{[\,]} \sim \mathrm{PY}(\alpha d_1, d_1, G_{[\,]})$$
$$G_{[ac]} | G_{[c]} \sim \mathrm{PY}(\alpha d_1 d_2, d_2, G_{[c]})$$



then

$$G_{[ac]} | G_{[\,]} \sim \mathrm{PY}(\alpha d_1 d_2, d_1 d_2, G_{[\,]})$$

with $G_{[c]}$ marginalized out.

I.e. we set $\alpha_{\mathbf{u}} = \alpha_{\sigma(\mathbf{u})} d_{\mathbf{u}}$.

HPYP: Coagulation & Fragmentation

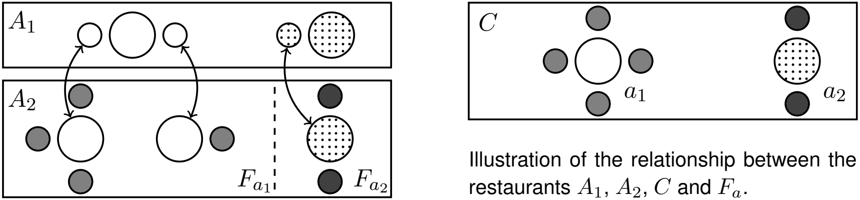
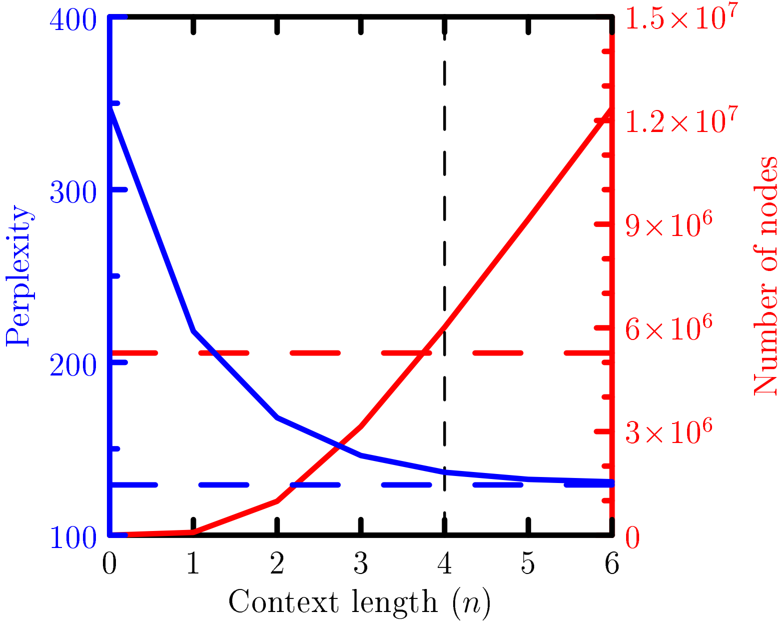


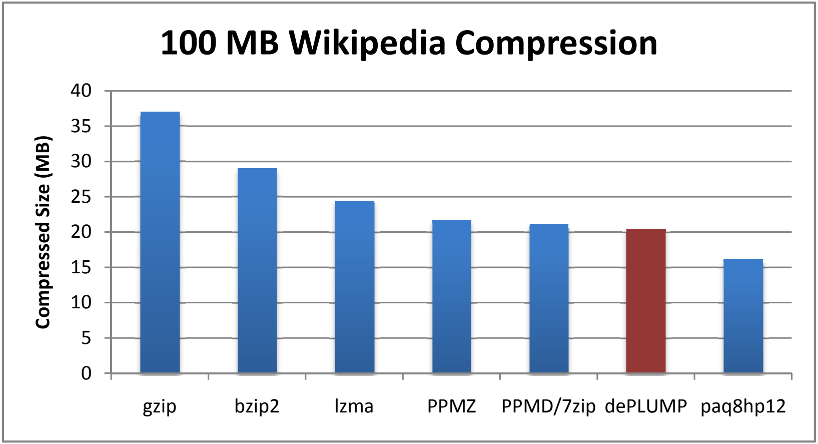
Illustration of the relationship between the restaurants A_1, A_2, C and F_a.

- **Theorem**: Suppose $A_2 \in \mathcal{A}_c$, $A_1 \in \mathcal{A}_{|A_2|}$, $C \in \mathcal{A}_c$ and $F_a \in \mathcal{A}_{|a|}$ for each $a \in C$ are related as above. Then the following describe equivalent distributions:
  (I) $A_2 \sim \text{CRP}_c(\alpha d_2, d_2)$ and $A_1 | A_2 \sim \text{CRP}_{|A_2|}(\alpha, d_1)$
  (II) $C \sim \text{CRP}_c(\alpha d_2, d_1 d_2)$ and $F_a | C \sim \text{CRP}_{|a|}(-d_1 d_2, d_2)$
  for each $a \in C$

Language Modeling Results



Results: Text Compression



INFERENCE

CRF Gibbs Sampler for Conjugate HDP

- Basically the hierarchical extension of the conjugate sampler for DP mixture models

$$
\begin{cases}
t_{ji} = t & \text{with probability} \propto \frac{n_{jt\cdot}^{\neg ji}}{n_{j\cdot\cdot}^{\neg ji}+\alpha} f_{k_{jt}}(\{x_{ji}\}) \\
t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{m_{\cdot k}^{\neg ji}}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_k(\{x_{ji}\}) \\
t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k^{\text{new}} & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{\gamma}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_{k^{\text{new}}}(\{x_{ji}\})
\end{cases}
$$

CRF Gibbs Sampler for Conjugate HDP

- Basically the hierarchical extension of the conjugate sampler for DP mixture models

$$\begin{cases} t_{ji} = t & \text{with probability} \propto \frac{n_{jt\cdot}^{\neg ji}}{n_{j\cdot\cdot}^{\neg ji}+\alpha} f_{k_{jt}}(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{m_{\cdot k}^{\neg ji}}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_k(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k^{\text{new}} & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{\gamma}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_{k^{\text{new}}}(\{x_{ji}\}) \end{cases}$$

$$k_{jt} = \begin{cases} k & \text{with probability} \propto \frac{m_{\cdot k}^{\neg jt}}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_k(\{x_{ji} : t_{ji} = t\}) \\ k^{\text{new}} & \text{with probability} \propto \frac{\gamma}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_{k^{\text{new}}}(\{x_{ji} : t_{ji} = t\}) \end{cases}$$

CRF Gibbs Sampler for Conjugate HDP

- Basically the hierarchical extension of the conjugate sampler for DP mixture models

$$\begin{cases} t_{ji} = t & \text{with probability} \propto \frac{n_{jt\cdot}^{\neg ji}}{n_{j\cdot\cdot}^{\neg ji}+\alpha} f_{k_{jt}}(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{m_{\cdot k}^{\neg ji}}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_k(\{x_{ji}\}) \\ t_{ji} = t^{\text{new}}, k_{jt^{\text{new}}} = k^{\text{new}} & \text{with probability} \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ji}+\alpha} \frac{\gamma}{m_{\cdot\cdot}^{\neg ji}+\gamma} f_{k^{\text{new}}}(\{x_{ji}\}) \end{cases}$$

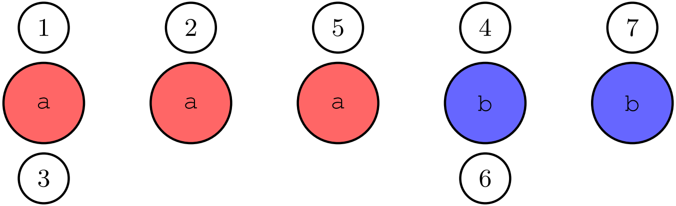
$$k_{jt} = \begin{cases} k & \text{with probability} \propto \frac{m_{\cdot k}^{\neg jt}}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_k(\{x_{ji} : t_{ji} = t\}) \\ k^{\text{new}} & \text{with probability} \propto \frac{\gamma}{m_{\cdot\cdot}^{\neg jt}+\gamma} f_{k^{\text{new}}}(\{x_{ji} : t_{ji} = t\}) \end{cases}$$

- In the non-conjugate case, extensions similar to the ones developed for the non-nojugate DP mixture model can be used
- In many models for discrete data (especially HPYP models), the observed data are direct draws from the random distributions *G*

Inference using Stick-breaking

- Variational inference can be performed in the stick-breaking representation
- Usually the number of stick pieces is fixed to some finite number

CRP Representations



Name	**Representation**	**Size**
PARTITIONS	$\{\{1,3\},\{2\},\{5\},\{4,6\},\{7\}\}$	$O(c)$
COUNTS	$[2,1,1]$ $[2,1]$	$O(t)$
COMPACT	$(4,3)$ $(3,2)$	$O(1)$
HISTOGRAM	$[1:2,2:1]$ $[1:1,2:1]$	$O(t)$

The new **COMPACT** representation only stores the **# of customers** and the **# of tables** (per type).

Gibbs Samplers

- Re-seating sampler
  - Iterate through all contexts/restaurants $\mathbf{u}$ and symbols $s \in \Sigma$
  - Sequentially remove and re-insert all $c_{\mathbf{u}s}$ customers
  - If removing/inserting a customer leads to removal/creation of a table, update the parent restaurant by removing/inserting a customer
  - In all but the PARTITIONS representation, there is no explicit customer-table assignment $\Longrightarrow$ sample table to remove from

Non-Compact Gibbs

- Pick table k to remove customer from with probability $\propto c_{\mathbf{u}sk}$
- Remove customer from selected table (recursively)
- Insert customer again (recursively)
- Time complexity: $O(c_{\mathbf{u}s} \times t_{\mathbf{u}s})$

Compact Original Gibbs Sampler

- Compute probability that a randomly chosen customer sits alone

$$P(\text{decrement } t_{\mathbf{u}s}) = \frac{S_{d_{\mathbf{u}}}(c_{\mathbf{u}s} - 1, t_{\mathbf{u}s} - 1)}{S_{d_{\mathbf{u}}}(c_{\mathbf{u}s}, t_{\mathbf{u}s})}$$

- Flip coin; if $t_{\mathbf{u}s}$ decremented, remove customer from parent
- Insert customer again (recursively)

$$P(\text{increment } t_{\mathbf{u}s}) = \frac{(\alpha_{\mathbf{u}} + d_{\mathbf{u}} t_{\mathbf{u}\cdot}) P^*_{\sigma(\mathbf{u})}(s)}{(\alpha_{\mathbf{u}} + d_{\mathbf{u}} t_{\mathbf{u}\cdot}) P^*_{\sigma(\mathbf{u})}(s) + c_{\mathbf{u}s} - t_{\mathbf{u}s} d_{\mathbf{u}}}$$

- Time complexity: $O(c_{\mathbf{u}s} \times t_{\mathbf{u}s})$; large constant because of log/exp

Re-Instantiating Gibbs Sampler

- Re-instantiate table sizes for restaurants along the path to **u**
- Apply original Gibbs sampler
- Discard sizes of individual tables
- Time complexity: $O(c_{\mathbf{u}s} \times t_{\mathbf{u}s})$; no log/exp necessary
- Preferred choice for compact representation

Direct Gibbs Sampler

- Instead of removing/inserting individual customers, sample $t_{\mathbf{u}s} \in \{1, \ldots, c_{\mathbf{u}s}\}$ directly from

$$P(t_{\mathbf{u}s}|\text{rest}) \propto \frac{[\alpha_{\mathbf{u}} + d_{\mathbf{u}}]_{d_{\mathbf{u}}}^{t_{\mathbf{u}\cdot}-1}}{[\alpha_{\sigma(\mathbf{u})} + 1]_1^{c_{\sigma(\mathbf{u})\cdot}-1}} S_{d_{\mathbf{u}}}(c_{\mathbf{u}s}, t_{\mathbf{u}s}) S_{d_{\sigma(\mathbf{u})}}(c_{\sigma(\mathbf{u})s}, t_{\sigma(\mathbf{u})s})$$

- Time complexity: $O(c_{\mathbf{u}s}^2)$; slow (need log/exp operations)

References

Y.W. Teh and M.I. Jordan. (2010). Hierarchical Bayesian Nonparametric Models with Applications. *Bayesian Nonparametrics*, Cambridge University Press.

Ho, M. W., James, L. F., & Lau, J. W. (2006). Coagulation fragmentation laws induced by general coagulations of two-parameter Poisson-Dirichlet processes.

Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, *27*, 1870–1902.