Sampling Table Configulations for the Hierarchical Poisson-Dirichlet Process

Changyou Chen[1,2], Lan Du[1,2], Wray Buntine[2,1]

[1]ANU College of Engineering and Computer Science
The Australian National University
[2]National ICT, Australia

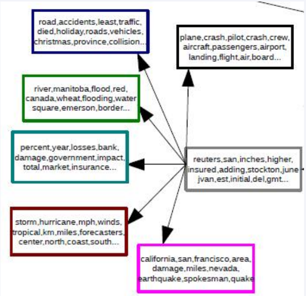
September 7, 2011

Outline

1. The Poisson-Dirichlet Process
2. Table Indicator Representation of the HPDP
3. Experiments: Topic Modeling using the HPDP
4. Conclusion

What can We Do with the Poisson-Dirichlet Process?

- Applications of the Poisson-Dirichlet process:
  - **Topic modeling**: Finding meaningful topics discussed in large scale documents. Beneficial to automatic document analysis and understanding.
  - **Computational linguistic**: *e.g.*, the n-gram model, adaptor grammar.
  - **Computer vision**. Using PDP/HPDP for image annotation, image segmentation, scene learning, and *etc*.
  - **Others**: Data compression, relational modeling, *etc*.

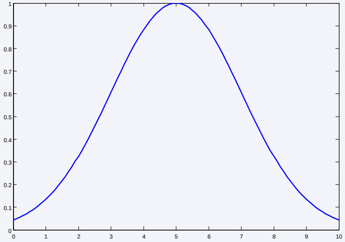


Cows, grass, hill, sky, tree...

...

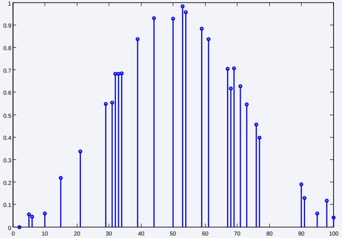
What is the Poisson-Dirichlet Process?

- The Poisson-Dirichlet process takes as input a base distribution, and yields as output a discreet distribution which is somewhat similar to the base distribution.



$$y = f(x)$$

$\longmapsto$



$$y = \sum_{i=0}^{K} p_i \delta_{x_i}$$

What is the Poisson-Dirichlet Process?

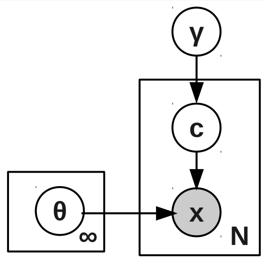
- The *Poisson-Dirichlet process* (PDP) is a random probability measure, or a distribution over distributions.
- The basic form of the PDP is:

$$\sum_{k=1}^{\infty} p_k \delta_{X_k^*}(\cdot) \tag{1}$$

  where $\vec{p} = (p_1, p_2, ...)$ is a probability vector so $0 \leq p_k \leq 1$ and $\sum_{k=1}^{\infty} p_k = 1$. Also, $\delta_{X_k^*}(\cdot)$ is a discrete measure concentrated at X_k^*. The values $X_k^* \in \mathscr{X}$ are *i.i.d.* drawn from the base measure $H(\cdot)$.
- The one parameter version of the PDP is the *Dirichlet process* (DP), and the two parameter version is the *Pitman-Yor process*.

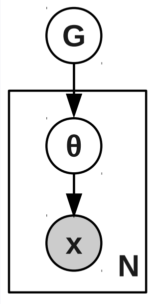
Why Poisson-Dirichlet Processes?

- It allows us to extend finite mixture models to infinite mixture models, by putting a PDP prior on the mixture components.



$$c \sim P(\gamma|\lambda), \quad x \sim P(x|\theta_c)$$

$P(\gamma|\lambda)$ is a infinite discreet distribution with parameter λ, or a Poisson-Dirichlet distribution.



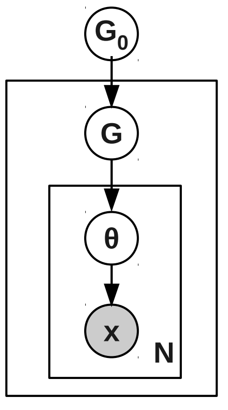
$$\theta \sim G, \quad x \sim P(x|\theta)$$

G is the distribution over mixture components θ, or a Poisson-Dirichlet Process.

Why Poisson-Dirichlet Processes?

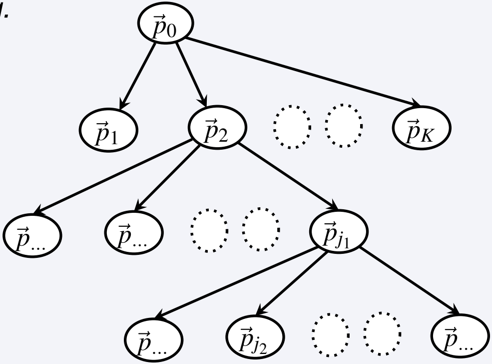
- Easily extend to model hierarchical discreet distributions, *e.g.*, hierarchical Dirichlet processes (HDP) or hierarchical Poisson-Dirichlet process (HPDP).

e.g.



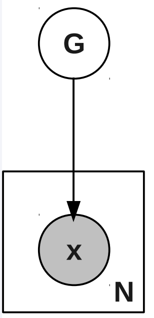
(a) HDP

e.g.

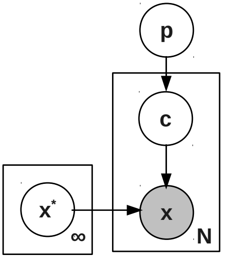


(b) Probability vector hierarchy

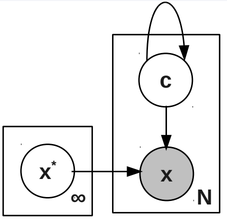
The Chinese Restaurant Process



(a) PDP



(b) PDP

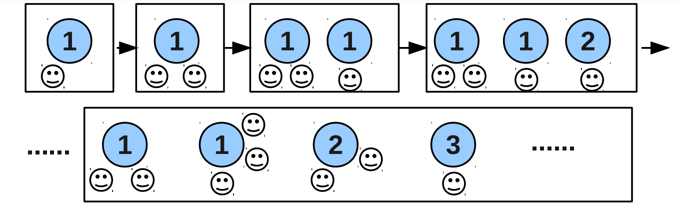


(c) CRP

- The Chinese restaurant process (c) (CRP) is the marginalized version of the PDP (b).
- Sampling the PDP can be done following the CRP's metaphor.

The Chinese Restaurant Process

- The CRP assumes a Chinese restaurant has an infinite number of tables, each with infinite capacity, the customers go in and sit at the restaurant as:
  - The first customer sits at the first unoccupied table with probability 1.
  - For the subsequent $(n+1)$-th customer:
    - He can choose to sit at an occupied table to share the dish with other seated customers with probability proportional to the number of customers that have already sit at that table.
    - or to sit at an unoccupied table with probability proportional to the sum of strength parameter and the total table count.



Outline

1 The Poisson-Dirichlet Process

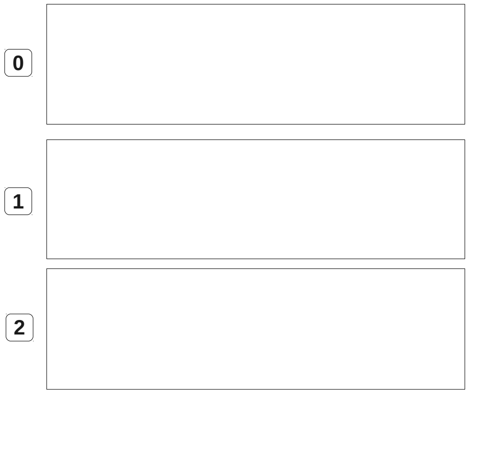
2 Table Indicator Representation of the HPDP

3 Experiments: Topic Modeling using the HPDP

4 Conclusion

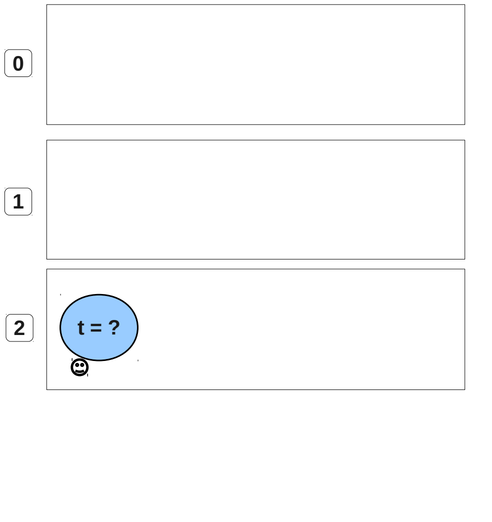
HPDP in the Chinese Restaurant Process Metaphor

- This shows a simple linear hierarchy of 3 Chinese restaurants. We'll bring customers in and watch the seating.



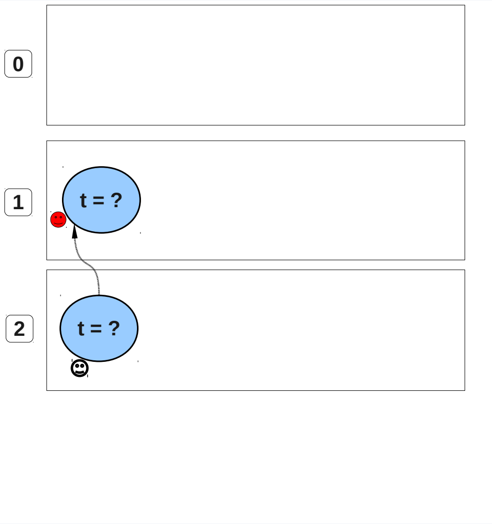
z: *dish*
u: *table indicator*

HPDP in the Chinese Restaurant Process Metaphor



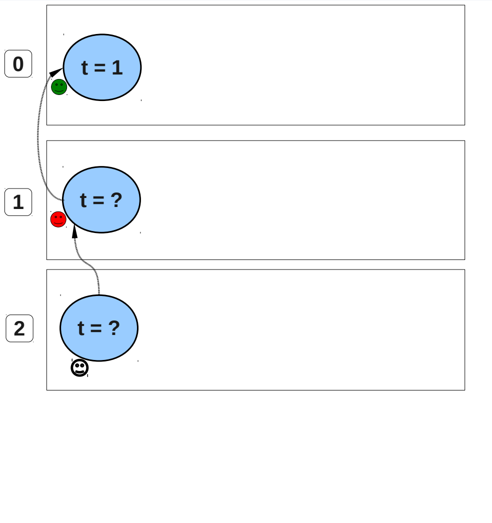
$z = ?$
$u = 2$

HPDP in the Chinese Restaurant Process Metaphor



$z = ?$
$u = 1$

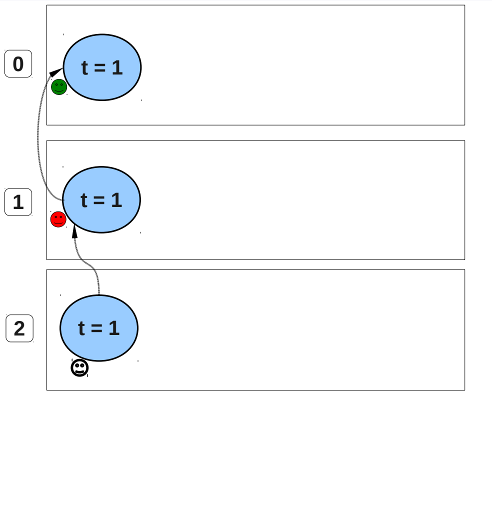
HPDP in the Chinese Restaurant Process Metaphor



$z = ?$

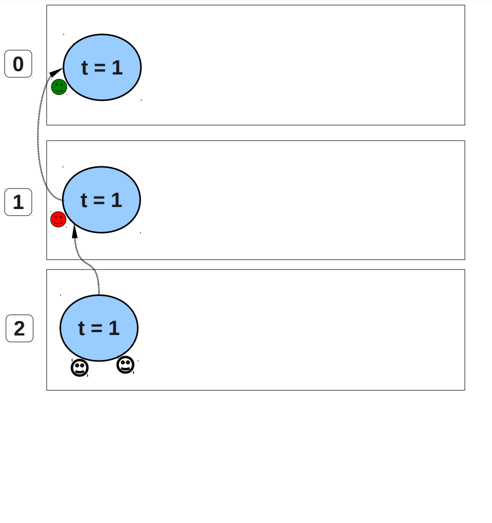
$u = 0$

HPDP in the Chinese Restaurant Process Metaphor



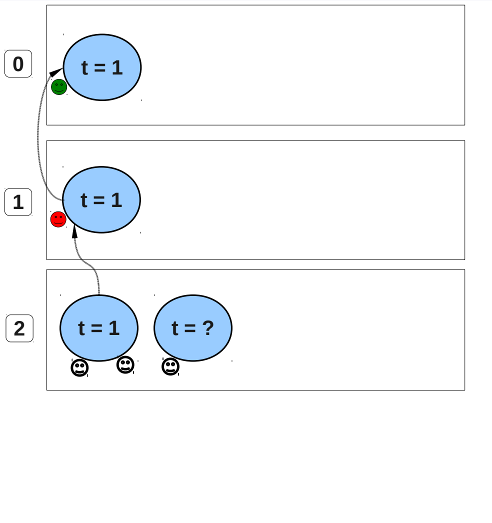
$$z = 1$$
$$u = 0$$

HPDP in the Chinese Restaurant Process Metaphor



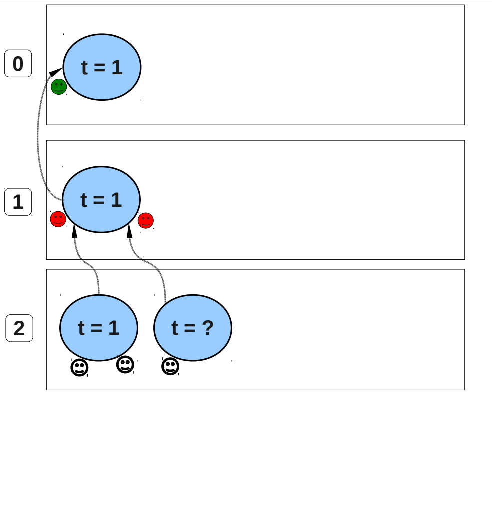
$$z = 1$$
$$u = NA$$

HPDP in the Chinese Restaurant Process Metaphor



$z = ?$
$u = 2$

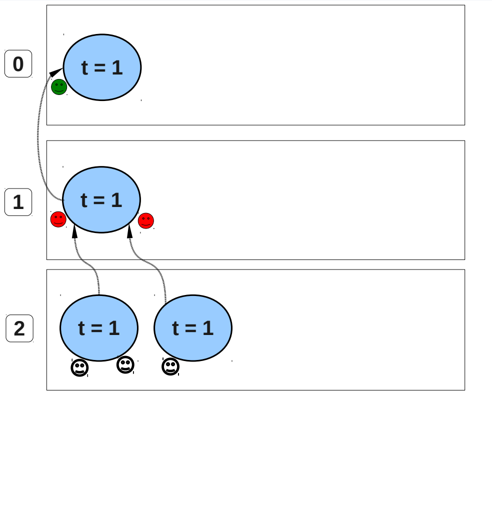
HPDP in the Chinese Restaurant Process Metaphor



$z = ?$

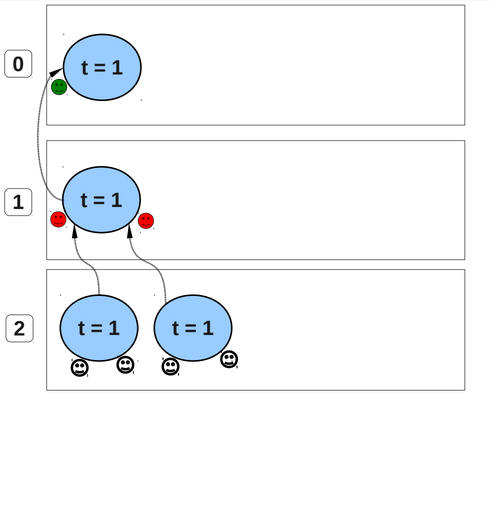
$u = 2$

HPDP in the Chinese Restaurant Process Metaphor



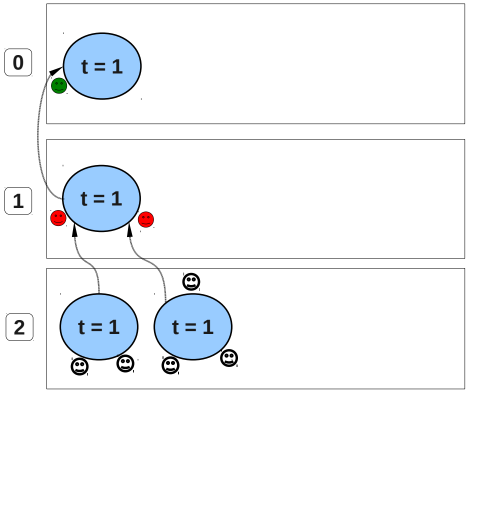
$z = 1$
$u = 2$

HPDP in the Chinese Restaurant Process Metaphor



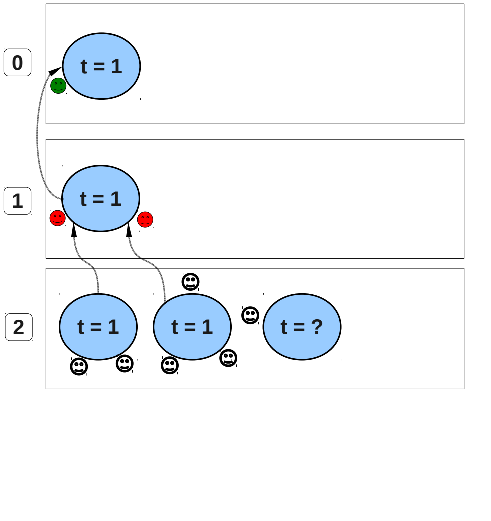
$$z = 1$$
$$u = NA$$

HPDP in the Chinese Restaurant Process Metaphor



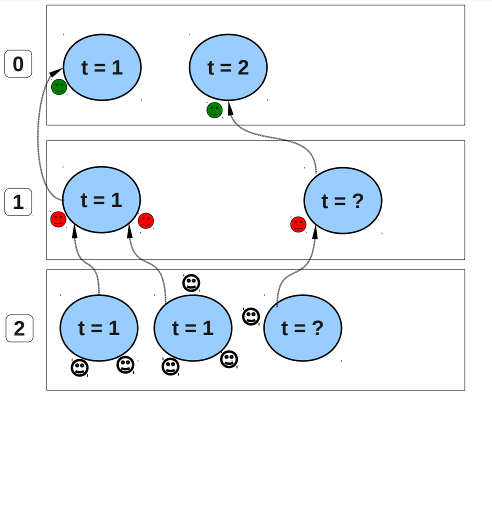
$$z = 1$$
$$u = NA$$

HPDP in the Chinese Restaurant Process Metaphor



$z = ?$
$u = 2$

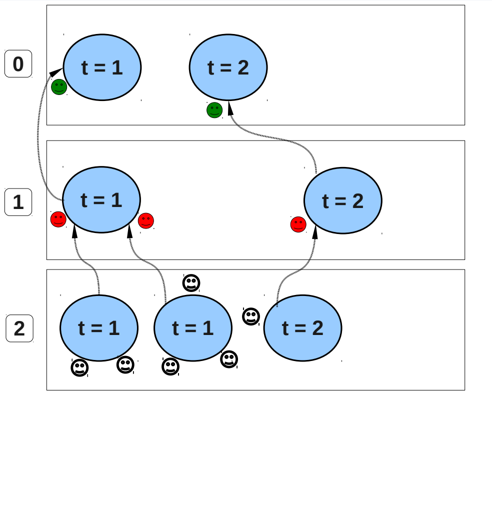
HPDP in the Chinese Restaurant Process Metaphor



$z = ?$

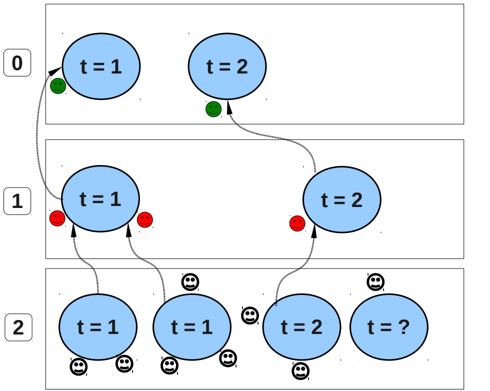
$u = 0$

HPDP in the Chinese Restaurant Process Metaphor



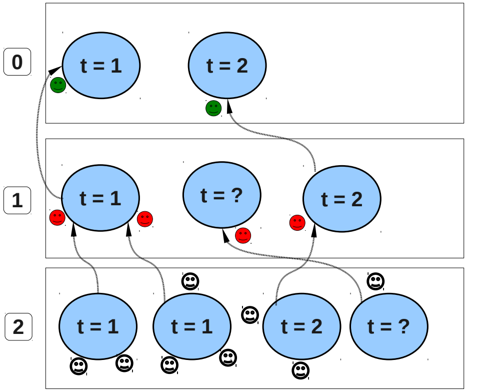
$$z = 2$$
$$u = 0$$

HPDP in the Chinese Restaurant Process Metaphor



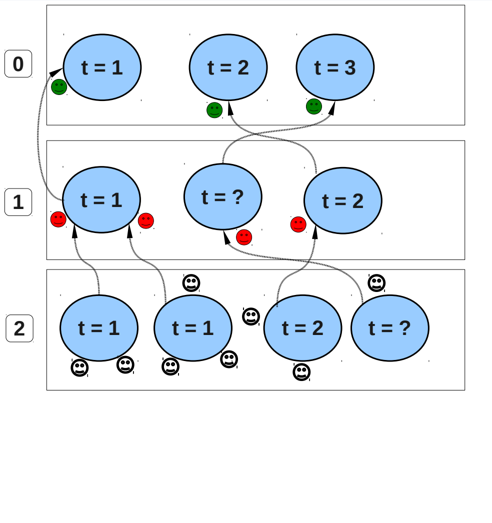
$z = ?$
$u = 2$

HPDP in the Chinese Restaurant Process Metaphor



$z = ?$
$u = 1$

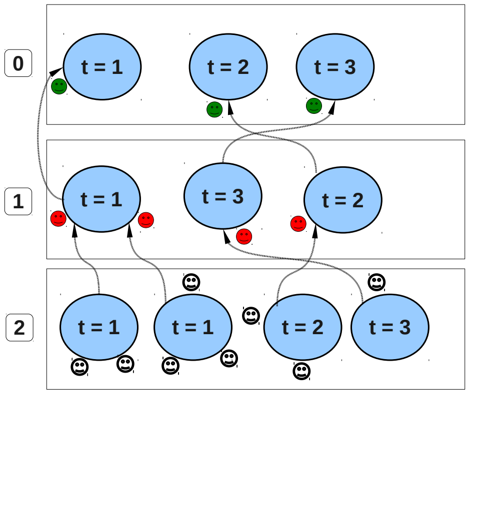
HPDP in the Chinese Restaurant Process Metaphor



$z = ?$

$u = 0$

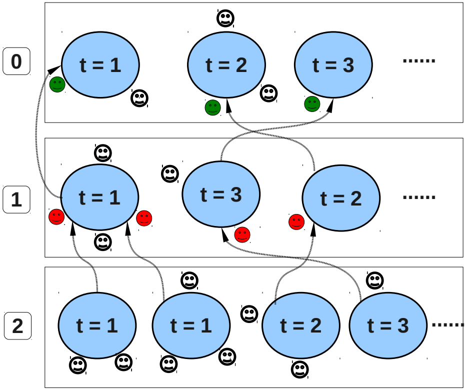
HPDP in the Chinese Restaurant Process Metaphor



$z = 3$
$u = 0$

HPDP in the Chinese Restaurant Process Metaphor

- We can also add customers at any restaurant.



HPDP Statistics used in Sampling

Seating arrangements represenation

$t_{0,1} = 1, m_{0,1,1} = 2$ $t_{0,2} = 1, m_{0,2,1} = 3$ $t_{0,3} = 1, m_{0,3,1} = 1$
$t_{1,1} = 1, m_{1,1,1} = 4$ $t_{1,2} = 1, m_{1,2,1} = 1$ $t_{1,3} = 1, m_{1,3,1} = 2$
$t_{2,1} = 2, m_{2,1,1} = 2,$ $m_{2,1,2} = 3$ $t_{2,2} = 1, m_{2,2,1} = 2$ $t_{2,3} = 1, m_{2,3,1} = 1$

$\vec{m}$ = counts at each table

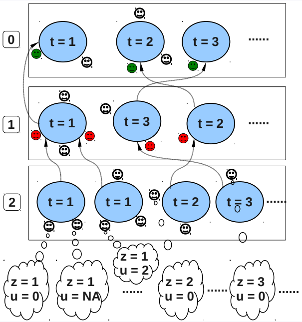


Table indicators representation

$n_{0,1} = 2, t_{0,1} = 1$ $n_{0,2} = 3, t_{0,2} = 1$ $n_{0,3} = 1, t_{0,3} = 1$
$n_{1,1} = 4, t_{1,1} = 1$ $n_{1,2} = 1, t_{1,2} = 1$ $n_{1,3} = 2, t_{1,3} = 1$
$n_{2,1} = 5, t_{2,1} = 2$ $n_{2,2} = 2, t_{2,2} = 1$ $n_{2,3} = 1, t_{2,3} = 1$

$\vec{t}$ = number of tables

Table Indicator Representation of the HPDP

- The above u is called the table indicator, defined as:

Definition (Table indicator u_l)

The table indicator u_l for each customer l is an auxiliary latent variable which indicates up to which level in the restaurant hierarchy l has contributed a table count (*i.e.* activated a new table).

- It is enough to use this variable to represent the HPDP, *e.g.*, other statistics (#customers and #tables in each restaurant) can be reconstructed from the table indicators easily.
- A block Gibbs sampler can be derived using this representation.

Table Counts Representation of the HPDP

Theorem: (Teh *et. al.*, 2006 [1]; Buntine and Hutter, 2010 [2])

Posterior of the HPDP with table count representation:

$$P_r(\vec{z}_{1:J}, \vec{t}_{1:J} \mid H_0) \quad = \quad \prod_{j \geq 0} \left(\frac{(b_j|a_j)_{T_j}}{(b_j)_{N_j}} \prod_k S^{n_{jk}}_{t_{jk}, a_j} \right)$$

$$\begin{aligned} n_{jk} &: \quad \text{\#customers in the } j\text{-th restaurant eating dish } k \\ t_{jk} &: \quad \text{\#tables in the } j\text{-th restaurant serving dish } k \\ N_j &: \quad = \sum_k n_{jk}, \qquad T_j := \sum_k t_{jk} \\ S^N_{M,a} &: \quad \text{the generalized Stirling number} \\ (x|y)_N &: \quad \text{the Pochhammer symbol with increment } y \\ H_0 &: \quad \text{base distribution} \end{aligned} \tag{2}$$

Table Indicator Representation of the HPDP

Theorem

Posterior of the HPDP in our representation:

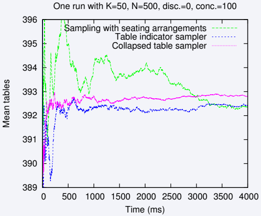
$$P_r(\vec{z}_{1:J}, \vec{u}_{1:J} \mid H_0) = \prod_{j \geq 0} \left(\frac{(b_j|a_j)_{T_j}}{(b_j)_{N_j}} \prod_k S^{n_{jk}}_{t_{jk}, a_j} \frac{t_{jk}!(n_{jk} - t_{jk})!}{n_{jk}!} \right)$$

the symbols are the same with the previous ones except that t_{jk} can be constructed from the table indicators:

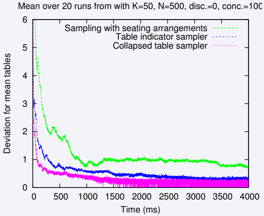
$$t_{jk} = \sum_{j' \in T(j)} \sum_{l \in D(j')} \delta_{z_l = k} \delta_{u_l \leq d(j)}$$

How does our Sampler Work?

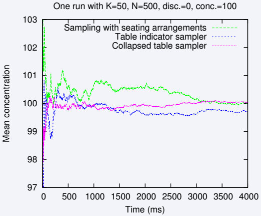
- Sampling a DP with $dim = 50$, data=500 points, true conc.=100:



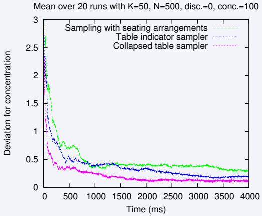
(a) Estimate of total table counts



(b) Std.Dev. of total table counts



(c) Estimate of concentration par. b



(d) Std.Dev. of concentration par. b

Outline

1 The Poisson-Dirichlet Process

2 Table Indicator Representation of the HPDP

3 Experiments: Topic Modeling using the HPDP

4 Conclusion

Datasets, Compared Algorithms and Evaluations

- We use five datasets (from Blogs, Reuters, NIPS, UCI):

	Health	Person	Obama	NIPS	Enron
# words	1,119,678	1,656,574	1,382,667	1,932,365	6,412,172
# documents	1,655	8,616	9,295	1,500	39,861
vocabulary size	12,863	32,946	18,138	12,419	28,102

- Compared algorithms:
  - Teh *et.al.* 's sampling direct assignment `SDA` [1]
  - Buntine and Hutter's collapsed table sampler `CTS` [2]
  - Our proposed block Gibbs sampler `STC`
  - `STC` initialized with `SDA`, denoted as `STC + SDA`
- We used the unbiased `left-to-right` algorithm [3] to calculate the testing perplexities for the topic models, which is a standard evaluation.

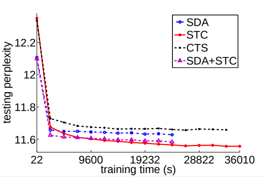
Perplexities

Table: Test log_2(perplexities) on the five datasets

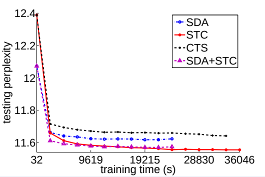
Dataset	Health		Person		Obama	
	$I = 100$	$I = 200$	$I = 1000$	$I = 2000$	$I = 1000$	$I = 2000$
SDA	11.628281	11.619546	11.930657	11.904425	11.144188	11.134732
CTS	11.655493	11.636743	11.940532	11.947740	11.191377	11.174327
SDA+STC	**11.582969**	**11.573457**	**11.844319**	**11.829628**	**11.094079**	**11.090389**
STC	**11.547999**	**11.551453**	**11.858719**	**11.852253**	11.210295	11.201241
Dataset	Enron		NIPS			
	$I = 500$	$I = 1000$	$I = 1000$	$I = 2000$		
SDA	10.847454	10.768568	10.564221	10.558330		
SDA+STC	**10.768568**	**10.659724**	**10.534148**	**10.518792**		
STC	10.853034[1]	10.810127	**10.474467**	**10.425393**		

[1] updated result

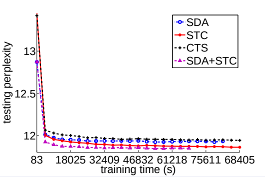
Convergence Speed



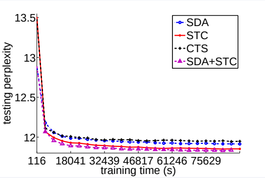
(a) Health with I = 100



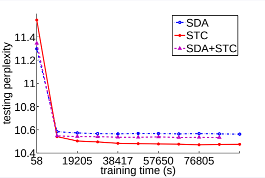
(b) Health with I = 200



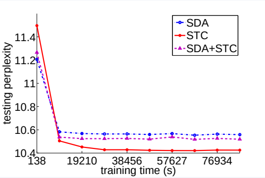
(c) Person with I = 1000



(d) Person with I = 2000



(e) NIPS with I = 1000



(f) NIPS with I = 2000

Outline

1 The Poisson-Dirichlet Process

2 Table Indicator Representation of the HPDP

3 Experiments: Topic Modeling using the HPDP

4 Conclusion

Conclusion

- Proposed a new representation for the HPDP based on the CRP metaphor.
- All useful statistics of the CRP can be reconstructed from the table indicator.
- No dynamic memory allocations for table counts.
- A blocked Gibbs sampler can be easily derived, *e.g.*, we do not have to sample the table counts separately.
- Experimental results on topic modeling indicate fast mixing of the proposed algorithm.
- All other PDP related applications can be adapted to this representation.

Reference

- Teh, Y.:
  A Bayesian interpretation of interpolated Kneser-Ney.
  Technical Report TRA2/06, School of Computing, National University of Singapore (2006)
- Buntine, W., Hutter, M.:
  A Bayesian review of the Poisson-Dirichlet process.
  Technical Report arXiv:1007.0296, NICTA and ANU, Australia (2010)
- Buntine, W.:
  Estimating likelihoods for topic models.
  In: ACML '09. (2009) 51–64

Thanks for your attention!!!