Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process

Changyou Chen, Lan Du, Wray Buntine[*†]
{Changyou.Chen, Lan.Du, Wray. Buntine}@NICTA.com.au

1 Introduction

- Discrete hierarchies are ubiquitous in intelligent systems.
- The *Poisson-Dirichlet process* (*PDP*) [1] allow statistical inference and learning on discrete hierarchies, *e.g.*, hierarchy of Dirichlet distributions.
- Applications of the PDP/HPDP include but not limited to:
  - **Topic modeling**: Finding meaningful topics discussed in large set of documents. Beneficial to automatic document analysis and understanding.
  - **Computational linguistic**: For example, the n-gram model.
  - **Computer vision**. Using PDP/HPDP to do image annotation, image segmentation, scene learning, and *etc.*
  - **Others**: Data compression, relational modeling, *etc.*

1.1 What does our sampler do?

- It is a collapsed Gibbs sampler so is generally more efficient.
- It requires no dynamic storage for table counts.
- It can be used wherever HPDP's are used
- It improves existing performance a lot.

1.2 The Poisson-Dirichlet Process (PDP)

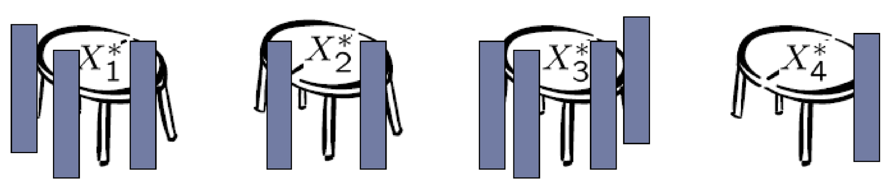
The Poisson-Dirichlet process [1] is a *random probability measure* defined as:

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

- $\vec{p} = (p_1, p_2, \cdots)$ is a probability vector satisfying $0 \le p_k \le 1$ and $\sum_{k=1}^{\infty} p_k = 1$, generated by a *stick-breaking process*.
- X_k's are drawn iid from a *base probability measure* $H(\cdot)$.

1.3 The Chinese Restaurant Process (CRP)

- It is the probability distribution of the partition of the integers.
- Explanation: a Chinese restaurant has an infinite number of circular tables, each with infinite capacity. Customer 1 is seated at an unoccupied table with probability 1. At time n + 1, a new customer chooses with probabilities to sit at one of the following n + 1 places: directly to the left of one of the n customers already sitting at an occupied table, or at a new, unoccupied circular table.



- Clearly, each table corresponds to a block of a random partition.
- The Poisson-Dirichlet process with probability vector marginalized out is equivalent to the Chinese Restaurant process, thus posterior sampling for the PDP can be done from the CRP's aspect.

1.4 The Hierarchical Poisson-Dirichlet Process (HPDP)

When using one PDP as the base measure for another PDP, we get a *hierarchical Poisson-Dirichlet process* [1].

*Research School of Computer Science, College of Engineering & Computer Science, Australian National University
†National ICT, Canberra, ACT, Australia

2 Table Indicator Representation of the HPDP

The *table indicator* u_l for each data item l (*i.e.*, a customer) is an auxiliary latent variable which indicates up to which level in the tree l has contributed a table count (*i.e.* activated a new table).See Figure 2.

3 Experiments

We applied the proposed algorithm for topic modeling (*HDP-LDA*) [2].

3.1 Datasets

- All three algorithms are implemented in C, and run on a desktop with Intel(R) Core(TM) Qaud CPU (2.4GHz).

- Data statistics:

n^0_{jk}: #customers (true customers) in the j-th restaurant eating dish k.
n_{jk}: #customers (including pseudo customers) in the j-th restaurant eating dish k.
t_{jk}: #tables in the j-th restaurant serving dish k.

- These statistics can be constructed using table indicator representation:

$$n^0_{jk} = \begin{cases} \sum_{l\in D(j)} \delta_{z_l=k}, & D(j) \neq \emptyset \\ 0, & \text{others} \end{cases} \qquad (2)$$

$$t_{jk} = \sum_{j'\in T(j)} \sum_{l\in D(j')} \delta_{z_l=k}\delta_{u_l\le d(j)} \qquad (3)$$

$$n_{jk} = n^0_{jk} + \sum_{j'\in C(j)} t_{j'k} \qquad (4)$$

$$T_j = \sum_k t_{jk}, \qquad (5)$$

$$N_j = \sum_k n_{jk} \qquad (6)$$

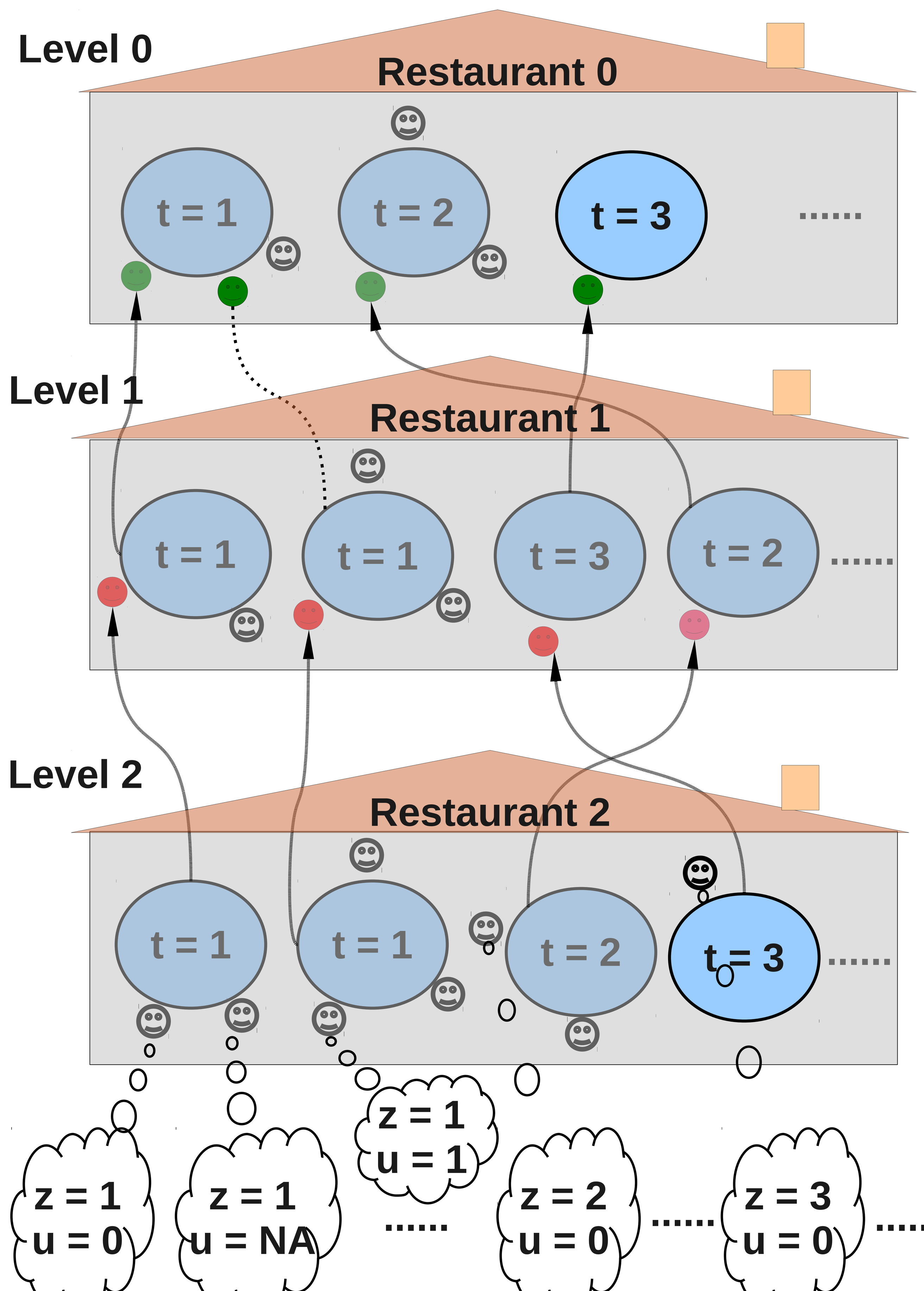


Figure 2: Table indicator representation of the HPDP.

Given the base distribution H_0 for the root node, the joint posterior distribution of the data $\vec{z}_{1:J}$ and their table indicators $\vec{u}_{1:J}$ for the HPDP in a tree structure is

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

where $S^N_{M,a}$ is the generalized Stirling number, $(x|y)_N$ denotes the Pochhammer symbol with increment y.

Table 1: Statistics of the five datasets

	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

- Five text datasets from Blogs, News articles, as well as the UCI repository. See Table 1 for details.

3.2 Testing Perplexities

We use the "*left-to-right*" algorithm [3] to calculate the testing perplexities, which is unbiased. See Table 2 for the results, the lower, the better .

Table 2: Test log_2(perplexities) on the five datasets. SDA means Sampling by Direct Assignment by Teh *et.al.*, CTS means Collapsed Table Sampler by Buntine *et.al.*, STC is our sampler.

Dataset	*Health*	*Person*	*Obama*	*Enron*	*NIPS*
SDA	11.628281	11.930657	11.144188	10.847454	10.564221
CTS	11.655493	11.940532	11.191377	–	10.595912
SDA+STC	**11.573457**	**11.829628**	**11.090389**	**10.659724**	**10.518792**
STC	**11.547999**	**11.852253**	11.201241	10.810127	**10.425393**

3.3 Convergence Speed

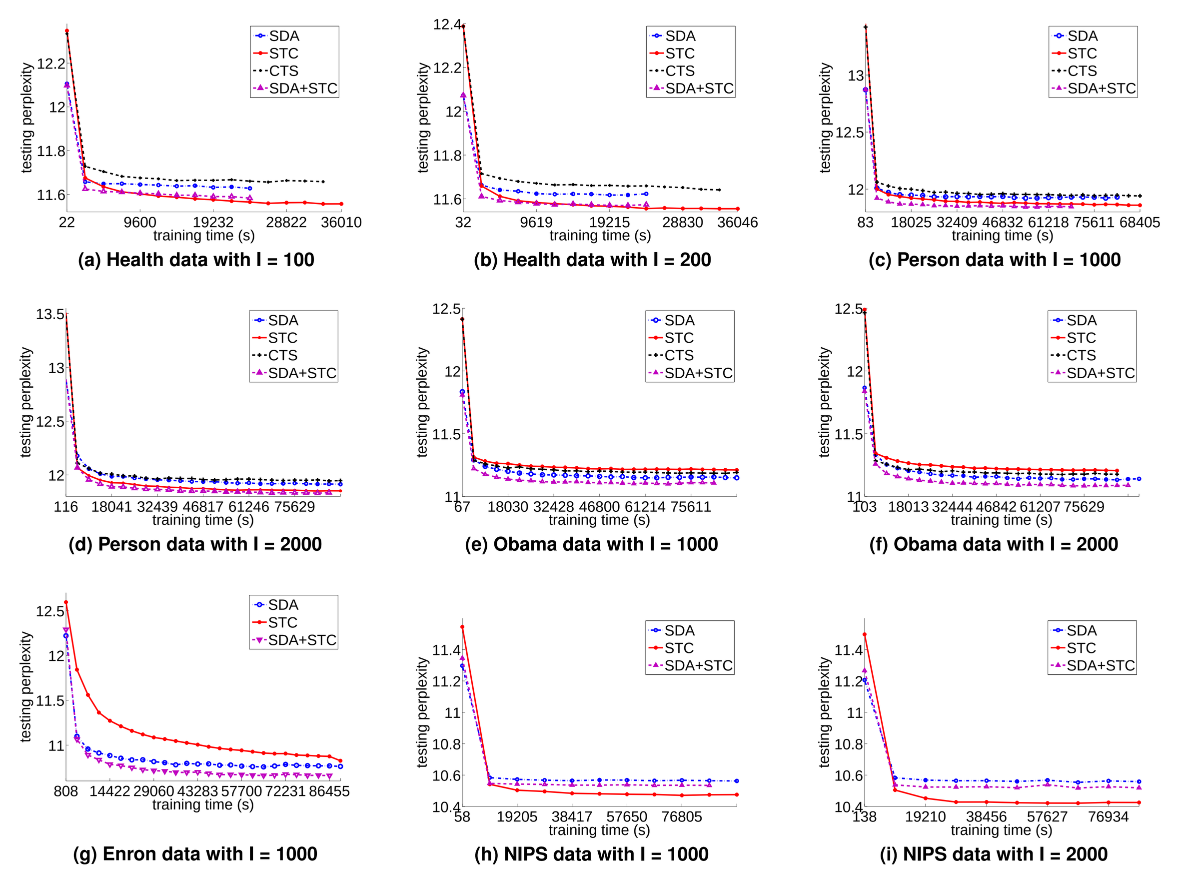


Figure 1: Test $\log_2$(perplexities) evolved with training time, I means initial number of topics

4 Conclusion

- Proposed a new representation for the HPDP.
- Useful statistics can be reconstructed from the table indicator.
- 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.

References

[1] Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: Bayesian Nonparametrics: Principles and Practice. (2010)
[2] Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the ASA **101** (2006) 1566–1581
[3] Buntine, W.: Estimating likelihoods for topic models. In: ACML '09. (2009) 51–64

Australian National University
ANU College of
Engineering & Computer Science
Statistical Machine Learning (SML) Group
NICTA