Conditional Simultaneous Draws from Hierarchical Chinese Restaurant Processes

Takaki Makino*
Division of Project Coordination, University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa, 277-8568 Japan `mak@scint.dpc.u-tokyo.ac.jp`

Shunsuke Takei
Dept. of Computational Biology, Univ. of Tokyo

Daichi Mochihashi
NTT Communication Science Laboratories

Issei Sato
Dept. of Mathematical Informatics, Univ. of Tokyo

Toshihisa Takagi
Database Center for Life Science, ROIS

1 Introduction

Chinese Restaurant Process (CRP) [1] is one of the most widely-used nonparametric bayesian prior that represents the Dirichlet process (DP). One limitation of CRP is that only *single* draws can be correctly handled; often we want to consider *simultaneous* multiple draws from a single DP or correlated DPs. For example, step-wise Gibbs sampling for Hierarchical DP-Hidden Markov Model (HDP-HMM) requires two simultaneous samples from distributions over hidden states, $s_t \sim G_{s_{t-1}}$ and $s_{t+1} \sim G_{s_t}$ given a condition that the resulting s_{t+1} is consistent with the following sequences. Even this simplest case is considered infeasible if the "parent" distribution is also represented as CRP [2], because the draws in the parent distribution must be considered too. Recent approaches on HDP-HMM [2, 3] avoid this problem by using stick breaking representation for the parent distribution, but this solution is not scalable over more than two-level hierarchical DPs.

In this paper, we point out that the root of this problem is that we only consider distribution over *samples* (in the example above, states s). In CRPs, we are actually sampling *seating arrangements*, which is a kind of latent variables associated to each draw. When we consider conditional distributions on simultaneous draws, we cannot neglect seating arrangements, since the conditional distribution of seating arrangements cannot be represented by a single-draw operation for CRPs.

Here we propose a comprehensive method for simultaneous sampling of seating arrangements from hierarchical CRPs: a Markov chain Monte Carlo sampler based on Metropolis-Hastings algorithm [4]. Since our method does not depend on a particular structure of hierarchy, it is applicable to any combination or hierarchical structure of CRPs. Moreover, our new procedure provides complete manipulations for seating arrangements, thus modellers can use our method to infer a conditional distributions of multiple draws without worrying about details. We believe that this opens a correct way for a family of complex models based on hierarchical CRPs.

In this abstract, we show our algorithm for sampling from conditional simultaneous draws. Presentation includes experiments of Gibbs sampler for HDP-HMM implemented on hierarchical CRP.

2 Hierarchical Chinese Restaurant Processes

Consider a hierarchical DP, in which each DP is indexed by a sequence $\boldsymbol{u} = u_1 u_2 \cdots$ of symbols (say, integers):

$$G_\varepsilon \sim \mathrm{DP}(\alpha_\varepsilon, H) \qquad\qquad G_{\boldsymbol{u}w} \sim \mathrm{DP}(\alpha_{\boldsymbol{u}}, G_{\boldsymbol{u}})$$

*This work is supported by JSPS (KAKENHI 20700126).

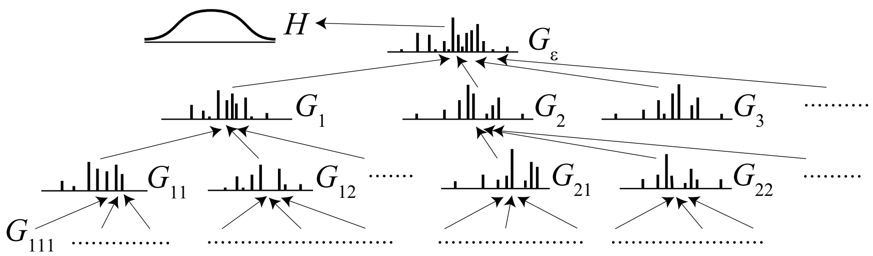

Figure 1: An example structure of a hierarchical Dirichlet Process we deal with in this paper. Each arrow indicates the relationship between a DP and its base measure.

where ε is a null sequence, H is the probability measure over measurable space $\mathcal{B}$, and $\boldsymbol{u}w$ is the concatenation of a symbol w to the sequence $\boldsymbol{u}$. Figure 1 illustrates this relation. We also define the prefix operator $'$ that strips the last symbol from the sequence: $(\boldsymbol{u}w)' = \boldsymbol{u}$.

We introduce a CRP representation for this hierarchical DP. Each DP is represented by a CRP indexed by $\boldsymbol{u}$: $\mathrm{CRP}(\boldsymbol{u})$ represents $G_{\boldsymbol{u}}$. Imagine that there have been N draws $x_1, \ldots, x_N$ at $\mathrm{CRP}(\boldsymbol{u})$; then a seating arrangement of $\mathrm{CRP}(\boldsymbol{u})$ is denoted as $S_{\boldsymbol{u}} = \langle \boldsymbol{\phi}_{\boldsymbol{u}}, \boldsymbol{j}_{\boldsymbol{u}}, \boldsymbol{t}_{\boldsymbol{u}} \rangle$. Meanings of $\boldsymbol{\phi}_{\boldsymbol{u}}$, $\boldsymbol{j}_{\boldsymbol{u}}$, and $\boldsymbol{t}_{\boldsymbol{u}}$ are as follows: $\phi_{\boldsymbol{u},t} \in \mathcal{B}$ is the dish on the t-th table in $\mathrm{CRP}(\boldsymbol{u})$, $j_{\boldsymbol{u},t}$ is the index of the table serving dish $\phi_{\boldsymbol{u},t}$ in the parent, and $t_{\boldsymbol{u},i}$ is the index of the table that the customer x_i is sitting. We denote the number of tables in $S_{\boldsymbol{u}}$ as $T_{\boldsymbol{u}}$, and a collection of seatings over all restaurants as $\boldsymbol{S}$.

In the following, we use number $n_{\boldsymbol{u},t}$ of customers on t-th table, number $n_{\boldsymbol{u}|\phi}$ of all customers eating dish ϕ, and number $n_{\boldsymbol{u}}$ of all customers in $\mathrm{CRP}(\boldsymbol{u})$, which can be calculated as follows:

$$n_{\boldsymbol{u},t} = \sum_{i=1}^{N} \mathbb{I}[\, t_{\boldsymbol{u},i} = t \,] + \sum_{w} \sum_{t'} \mathbb{I}[\, j_{\boldsymbol{u}w,t'} = t \,] \qquad n_{\boldsymbol{u}|\phi} = \sum_{t} \mathbb{I}[\, \phi_{\boldsymbol{u},t} = \phi \,]\, n_{\boldsymbol{u},t} \qquad n_{\boldsymbol{u}} = \sum_{t} n_{\boldsymbol{u},t},$$

where $\mathbb{I}$ denotes the indicator function, whose value is 1 if the condition is true and 0 otherwise. [1]

3 Gibbs Sampling on the Conditional Simultaneous Draws

A Gibbs sampler for the hierarchical CRP interleaves randomly removing a customer and adding him again. Let we denote $\boldsymbol{S}\backslash S(x)$ as the seating after removing the customer $S(x)$ corresponding to the draw x from $\boldsymbol{S}^{old}$. Then, re-drawing from CRP with seating $\boldsymbol{S}\backslash S(x)$ will give the new sample x^{new} and a new seating $\boldsymbol{S}^{new}$. Often we want samples from posterior distribution given observed data, which requires samples from a conditional distribution of hierarchical CRP: that is, x_i is restricted to one of the values that satisfies a condition (surrounding states in HMM, for example), which we denote as C. in such a case, we can choose x^{new} in proportinal to the expected probability $p(x|\boldsymbol{S}\backslash S(x))$, and update seating by $\mathsf{addCustomer}(x^{new}, \boldsymbol{u}, \boldsymbol{S})$ operation, which samples the seating of the new customer for x^{new}.

However, this procedure is not applicable to sampling from conditional distributions on simultaneous draws. Consider a condition C on a set of k draws, $\boldsymbol{x} = x_1, \ldots, x_k$, from $\mathrm{CRP}(\boldsymbol{u}_1), \ldots, \mathrm{CRP}(\boldsymbol{u}_k)$, respectively.[2] We want to sample from the following conditional distribution:

$$p(\boldsymbol{x}|C, \boldsymbol{S}\backslash S(\boldsymbol{x})) = \frac{1}{Z_C}\, \mathbb{I}[\, \boldsymbol{x} \in C \,]\, p(\boldsymbol{x}|\boldsymbol{S}\backslash S(\boldsymbol{x})) \quad . \tag{1}$$

There are two problems for conditional distributions on simultaneous draws. First, the exact distribution of $\boldsymbol{x}$ is complicated than that of a single draw, because the seating arrangements are correlated. Even if the CRP indices $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ are different, some draw may change the seating arrangement of the common parent CRP, which will cause a change on the probability of other draws. Second, even if the correct distribution of $\boldsymbol{x}$ is determined, we have to compute an exact distribution of seating arrangements $\boldsymbol{S}^{new}$ under that condition, because in hierarchical CRPs, $\boldsymbol{S}$ is a part of latent variables

[1] In single draw cases, $n_{\boldsymbol{u},t}$ can represent the seating of $\mathrm{CRP}(\boldsymbol{u})$ without $\boldsymbol{t}$ and $\boldsymbol{j}$ if $\mathsf{removeCustomer}$ operation samples the customer to be removed. This approximation is also applicable to our sampling method.

[2] For simplicity, we give fixed indices $\boldsymbol{u}_i$ for the CRPs to be drawn. However, our method is applicable to more general settings, where $\boldsymbol{u}_i$ depends on the preceding draws $x_1, \ldots, x_{k-1}$. For example, in the step-wise Gibbs sampling on HDP-HMM, the index of the second draw is determined by the result of the first draw.

that is being sampled [5]. The common operations for hierarchical CRPs such as addCustomer are suitable only for *single* draw and cannot be combined for sampling seating arrangements from the conditional *simultaneous* draws.

4 Metropolis-Hastings Sampler on Seating Arrangements

Our idea is to use a Metropolis-Hastings sampler [4] on the seating arrangements, which uses a proposal distribution of seating arrangements, and accepts the proposal probabilistically. Although it is hard to give a proposal distribution $q(\boldsymbol{S})$ of seating arrangements directly, our method constructs $q(\boldsymbol{S})$ from a proposal distribution of draws, $q(\boldsymbol{x})$, and use addCustomer to create a proposal distribution of seatings from the proposed draws. The algorithm to obtain new samples, $\boldsymbol{x}^{new}$ and $\boldsymbol{S}^{new}$ given the previous sample $\boldsymbol{x}^{old}$ and $\boldsymbol{S}^{old}$ is as follows:

1. Let $\boldsymbol{S}^{(0)} = \boldsymbol{S}^{old} \backslash S(\boldsymbol{x}^{old})$ be a seating that removes $\boldsymbol{x}^{old}$ from $\boldsymbol{S}^{old}$.
2. Draw $\boldsymbol{x}^*$ from proposal distribution of draws $q(\boldsymbol{x})$.
3. Create proposal seating $\boldsymbol{S}^*$ by sequentially adding customers for $x_1^*, \ldots, x_k^*$ to $\boldsymbol{S}^{(0)}$: $\boldsymbol{S}^{(i)} = \mathsf{addCustomer}(x_i^*, \boldsymbol{u}_i, \boldsymbol{S}^{(i-1)})$, and $\boldsymbol{S}^* = \boldsymbol{S}^{(k)}$.
4. Calculate an acceptance probability $r = \min\left\{ 1,\ \frac{p(\boldsymbol{S}^*)}{p(\boldsymbol{S}^{old})} \frac{q(\boldsymbol{S}^{old}|\boldsymbol{x}^{old}, \boldsymbol{S}^{(0)})}{q(\boldsymbol{S}^*|\boldsymbol{x}^*, \boldsymbol{S}^{(0)})} \frac{q(\boldsymbol{x}^{old})}{q(\boldsymbol{x}^*)} \right\}$.
5. $\langle \boldsymbol{x}^{new}, \boldsymbol{S}^{new} \rangle = \begin{cases} \langle \boldsymbol{x}^*, \boldsymbol{S}^* \rangle & \text{with probability } r \\ \langle \boldsymbol{x}^{old}, \boldsymbol{S}^{old} \rangle & \text{otherwise.} \end{cases}$

Here, each factor in r can be calculated as follows:

True probability of all seatings $p(\boldsymbol{S})$ is the product of seating probability for each CRP:

$$p(\boldsymbol{S}) = p(\boldsymbol{\phi}_\varepsilon|H) \cdot \prod_{\boldsymbol{u}} p(S_{\boldsymbol{u}}) = \prod_t p(\phi_{\varepsilon,t}|H) \cdot \prod_{\boldsymbol{u}} \frac{(\alpha_{\boldsymbol{u}})^{T_{\boldsymbol{u}}} \prod_t \Gamma(n_{\boldsymbol{u},t})}{\Gamma(\alpha_{\boldsymbol{u}} + n_{\boldsymbol{u}})} \ . \tag{2}$$

Proposal distribution of draws $q(\boldsymbol{x})$ can be anything, but we suggest to use the product of expected probability for each draw:

$$q(\boldsymbol{x}) = \prod_k p(x_k|\boldsymbol{S}^{(0)}) \ . \tag{3}$$

Proposal distribution of seatings $q(\boldsymbol{S}^*|\boldsymbol{x}, \boldsymbol{S}^{(0)})$ is the product of probability for each operation for adding a customer:

$$q(\boldsymbol{S}^*|\boldsymbol{x}, \boldsymbol{S}^{(0)}) = q(\boldsymbol{S}^{(1)}|x_1, \boldsymbol{S}^{(0)})\, q(\boldsymbol{S}^{(2)}|x_2, \boldsymbol{S}^{(1)}) \cdots q(\boldsymbol{S}^*|x_k, \boldsymbol{S}^{(k-1)}) \ , \tag{4}$$

where $q(\boldsymbol{S}^{(i)}|x_i, \boldsymbol{S}^{(i-1)})$ is the probability $p(t_{\boldsymbol{u},i}|x_i, \boldsymbol{u}_i, \boldsymbol{S})$ of obtaining the seating arrangement $\boldsymbol{S}^{(i)}$ as a result of $\mathsf{addCustomer}(x_i, \boldsymbol{u}_i, \boldsymbol{S}^{(i-1)})$ operation.

$$p(x|\mathrm{CRP}(\boldsymbol{u}), \boldsymbol{S}) = \frac{n_{\boldsymbol{u}|x} + \alpha_{\boldsymbol{u}} \cdot p(x|\mathrm{CRP}(\boldsymbol{u}'), \boldsymbol{S})}{n_{\boldsymbol{u}} + \alpha_{\boldsymbol{u}}} \tag{5}$$

$$p(t|x, \boldsymbol{u}, \boldsymbol{S}) \propto \begin{cases} n_{\boldsymbol{u},t} & n_{\boldsymbol{u},t} \geq 1 \\ \alpha_{\boldsymbol{u}} \cdot p(x|H) & n_{\boldsymbol{u},t} = 0, \boldsymbol{u} = \varepsilon \\ \alpha_{\boldsymbol{u}} \cdot p(x|\mathrm{CRP}(\boldsymbol{u}'), \boldsymbol{S}) \cdot p(j_{\boldsymbol{u},t}|x, \boldsymbol{u}', \boldsymbol{S}) & n_{\boldsymbol{u},t} = 0, \boldsymbol{u} \neq \varepsilon \end{cases} \tag{6}$$

References

[1] D. Aldous. Exchangeability and related topics. In *École d'été de Probabilités de Saint-Flour XIII—1983, Lecture notes in mathematics 1117*, pages 1–198. Springer Verlag, 1985.

[2] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort et al., editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2009.

[3] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. In *Proc. of the 25th Intl. Conf. on Machine Learning (ICML 2008)*, pages 312–319, 2008.

[4] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.