Variational Inference for Sequential Distance Dependent Chinese Restaurant Process

Sergey Bartunov SBOS@SBOS.IN
Dorodnicyn Computing Centre of the Russian Academy of Sciences, 119333, Moscow, Vavilov st. 40 RUSSIA

Dmitry P. Vetrov VETROVD@YANDEX.RU
Lomonosov Moscow State University, 119992, Moscow, Leninskie Gory, 1, 2nd ed. bld., CMC department RUSSIA

Abstract

Recently proposed distance dependent Chinese Restaurant Process (ddCRP) generalizes extensively used Chinese Restaurant Process (CRP) by accounting for dependencies between data points. Its posterior is intractable and so far only MCMC methods were used for inference. Because of very different nature of ddCRP no prior developments in variational methods for Bayesian nonparametrics are appliable. In this paper we propose novel variational inference for important sequential case of ddCRP (seqddCRP) by revealing its connection with Laplacian of random graph constructed by the process. We develop efficient algorithm for optimizing variational lower bound and demonstrate its efficiency comparing to Gibbs sampler. We also apply our variational approximation to CRP-equivalent seqddCRP-mixture model, where it could be considered as alternative to one based on truncated stick-breaking representation. This allowed us to achieve significantly better variational lower bound than variational approximation based on truncated stick breaking for Dirichlet process.

1. INTRODUCTION

One of the most important problems in machine learning and statistics is model selection. Well-known examples of it are choosing the number of hidden layers of neural network or the number of clusters in the mixture model. Model selection almost always leads to a tradeoff between model complexity and fit accuracy. While one may fit several models by varying the number of structural components and choosing the best one by some criteria, Bayesian nonparametric methods provide an elegant alternative solution to this problem. Instead of comparing several models, nonparametric approach is to define a distribution on model structure which can adapt it's complexity to data. Furthermore, it may refine and even complicate its structure when new data is being observed.

For many nonparametric models such as widely used Dirichlet process (Ferguson, 1973) or Indian Buffet Process (Griffiths & Ghahramani, 2006), it is common to assume *exchangeability* of data, the property that every permutation of data points has the same probability under the model. While this assumption often holds, in many settings when data has temporal, spatial or any other internal dependencies, a proper non-exchangeable prior which takes such information into account could model data more adequately and thus fit it more accurately.

A number of such distributions were developed to date including dependent Dirichlet process (MacEachern, 1999) and other similar processes (Duan et al., 2005; Griffin & Steel, 2006; Xue et al., 2007), in which various constructions on top of Dirichlet process are considered. Besides that, a distance-dependent Chinese Restaurant Process (ddCRP) was proposed recently by (Blei & Frazier, 2011), which takes an alternative approach to modeling dependencies by considering a distribution over partitions. Given pairwise distances of any nature (generally not symmetric), each data point connects itself to other ones (including itself) with probability depending on the corresponding distance; after that data points that are reachable from each other form a partition.

Besides very general yet simple formulation which allows for many interesting applications including image segmentation (Ghosh et al., 2011) and natural language processing tasks (Haghighi & Klein, 2010), ddCRP has another advantage of not generally exhibiting *marginal invariance*, the property that a missing observation does not affect the joint distribution of data and parameters. This fact also distin-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

guishes ddCRP from other mentioned processes.

As for many interesting distributions, posterior of ddCRP is intractable, and so far only Markov Chain Monte Carlo (MCMC) inference techniques were developed for it. Although it is theoretically guaranteed that a properly constructed Markov chain will once converge to the true posterior, in practice it's often hard to determine convergence, and no general methods are provided to estimate the required number of samples to obtain a good approximation. Due to the fact that nature of ddCRP differs a lot from many other nonparametric distributions, prior developments of variational inference for DP, such as the one based on stick-breaking construction, are not applicable.

In this paper, we propose variational inference algorithm for the important sequential case of ddCRP called sequential distance-dependent Chinese Restaurant Process (seqddCRP), in which (a) data points arrive in sequential order, and (b) a data point can not be connected to those which arrived later (though the opposite is allowed). This assumption often holds in temporal data and is natural for many natural language processing tasks. Moreover, Chinese Restaurant Process (Aldous, 1985) could be formulated as a special case of seqddCRP.

The contributions of this paper are:

1. We introduce variational mean-field approximation for seqddCRP mixture model by revealing connection between partition assignments of data points and expected Laplacian of random graph modeled by the process.
2. We show that our approximation of posterior provides better lower bound to marginal likelihood for Dirichlet process mixture than the one based on truncated stick-breaking process and also converges in much fewer iterations.
3. We develop efficient coordinate-ascent inference algorithm and compare it with Gibbs sampler in terms of computational performance and inference quality.

The rest of the paper is organized as follows. We first review Dirichlet process and Chinese Restaurant process (section 2), then we briefly describe distance-dependent Chinese Restaurant Process (section 3). In section 4 we present our variational inference framework for seqddCRP mixture. We conclude with experiments on synthetic and real data (section 5).

2. DIRICHLET PROCESS AND RELATED PROCESSES

One of the most frequently used nonparametric models is Dirichlet process (DP) introduced by (Ferguson, 1973). It could be seen as infinite-dimensional generalization of Dirichlet distribution parametrized by base measure G_0 and concentration parameter α. A draw from DP is a random distribution over draws from G_0 with the property that some draws may contain repetitive elements. The frequency of repetitions is governed by α.

The infinite set of probabilities representing frequencies of unique draws from G_0 is distributed according to *stick-breaking process* (SBP) (Sethuraman, 1994) which divides total probability mass into diminishing probabilities π_k such that $\sum_k^\infty \pi_k = 1$. Each π_k is constructed by breaking the rest of the unit-length stick $1 - \sum_{p<k} \pi_p$ in a ratio of $v_k \sim \text{Beta}(1, \alpha)$. This allows for constructive definition of DP and is often used for variational inference.

Thus one may define nonparametric mixture model using DP by placing SBP prior over mixture component assignments z_i for data points $1, 2, \dots$ and drawing mixture parameters θ_k from appropriate base measure.

Another representation of DP is Chinese Restaurant Process (Aldous, 1985) which could be derived by integrating out base measure from DP. This makes mixture assignments dependent on each other, but leaves them *exchangeable* and transforms DP into distribution over the all possible partitions of natural numbers. It is defined as follows: consider a restaurant with infinite number of tables and no customers at the beginning. Customers enter the restaurant one by one, each drawing her *table assignment* z_i. By definition for the first customer $z_1 = 1$. All successive customers $i = 2, 3, \dots$ draw z_i according to the following distribution:

$$p(z_i = k | z_{\setminus i}) \propto \begin{cases} n_k, & k \leq K \\ \alpha, & k = K + 1 \end{cases} \tag{1}$$

where K is the number of ocuppied tables before i-th customers enters, and n_k is a total number of customers already chosen the table. After all the table assignments have been selected, they form clusters; after that cluster parameters θ_k are being drawn. Generative process is finished by drawing each data point x_i from the corresponing mixture component θ_{z_i}.

3. DISTANCE-DEPENDENT CHINESE RESTAURANT PROCESS

The generative process of table assignments in CRP described above may be reformulated using *customer assignments*. Let each customer in CRP choose not the table z_i, but rather exactly one customer c_i to share a table with:

$$p(c_i = j | \alpha) \propto \begin{cases} 1, & j < i \\ \alpha, & i = j \end{cases} \tag{2}$$

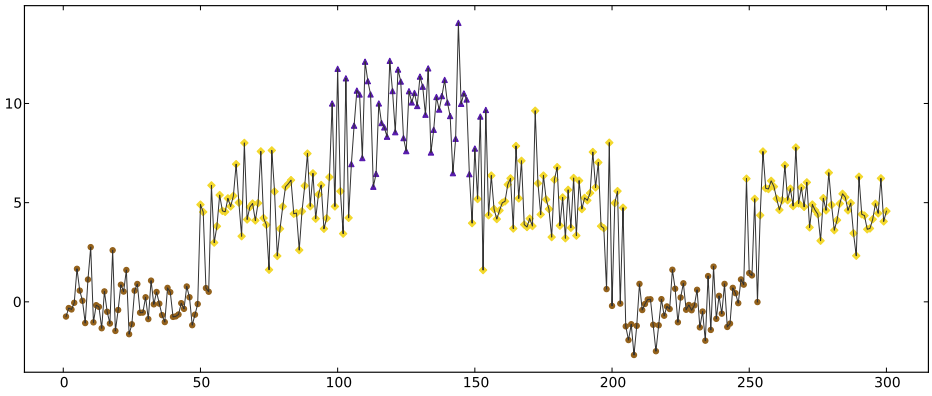

Figure 1. Time dependent mixture. Same-colored points were drawn from the same normal distribution

Thus the first customer always sits with herself, and successive customers sequentially choose whether they want to join existing table (occupied by some of the previous customers), or to initiate a new one. It is easy to show that if we define $z_i(\mathbf{c})$ as the minimal index of customers that are *reachable* from i through directed graph of customer assignments (i.e. those which are sitting at the same table), then the induced partitioning $z(c)$ based on such reachability is equivalent to the one constructed in CRP.

(Blei & Frazier, 2011) generalized (2) to make customer assignments *distance dependent*:

$$p(c_i = j|f, D, \alpha) \propto \begin{cases} f(d_{ij}), & j < i \\ \alpha, & i = j \end{cases} \tag{3}$$

where $f(d)$ is non-negative *decay function* such that $f(\infty) = 0$, D is *distance matrix* and positive α governs for table initiation. This distribution on customer assignments is called sequential distance-dependent Chinese Restaurant Process (seqddCRP). Traditional CRP appears as a special case of seqddCRP and the latter is generally a very different distribution. At first, it accounts for prior dependencies in data which may be of any nature and hence does not assume customers to be *exchangeable*. In addition, seqddCRP is not generally a distribution over partitions induced by random measures (see (Blei & Frazier, 2011) for details).

It is also possible to define general ddCRP without assuming sequential order of customers by allowing assignments $c_i > i$. This makes possible to start a new table not necessarily by drawing $c_i = i$ for some i. In that case, table assignments are constructed treating customer assignments as an undirected graph. Below we focus on sequential restaurants only.

Now using (3) as a prior over mixture component assignments, we may define the seqddCRP mixture model:

$$p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta}|\eta) = p(\mathbf{c}|\eta) \prod_{j=1}^{N} \left(p(\theta_j|\eta) \prod_{i, z_i(\mathbf{c})=j} p(x_i|\theta_j) \right)$$

$$p(\mathbf{c}|\eta) = \prod_{i=1}^{N} p(c_i|f, D, \alpha)$$

where $\eta = \{f, D, \alpha, G_0\}$ are process parameters, G_0 is a base measure generating mixture parameters $\mathbf{\Theta}$, and N is total number of data points.

Here we account for all N possible tables (since all customers have non-zero probability to start a new table) and identify each table with a customer that initiated it by choosing to sit with herself. Notice that while parameters of empty tables affect joint distribution, marginal distribution $p(\mathbf{x}, \mathbf{c})$ does not depend on empty tables, and so does the variational lower bound.

4. VARIATIONAL INFERENCE FOR SEQDD-CRP MIXTURE

Before we continue with the mean-field approximation of the posterior, let us first reformulate seqddCRP model. We denote as $r_{ij}(\mathbf{c})$ indicator function which is equal to 1 if and only if there exists a directed path from i-th customer to j-th customer in the graph induced by customer assignments $\mathbf{c}$. We also expand customer assignment $c_i = j$ to one-hot vector of size N such that $c_{ij} = 1$ and $c_{ik} = 0$ for $k \neq j$. This allows us to rewrite joint distribution of seqddCRP mixture in the following way:

$$p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta}|\eta) = p(\mathbf{c}|\eta) \prod_{j=1}^{N} p(\theta_j|G_0) \Big(\prod_{i=1}^{N} p(x_i|\theta_j)^{r_{ij}(\mathbf{c})} \Big)^{c_{jj}} \tag{4}$$

As we are interested in connected components $\mathbf{z}(\mathbf{c})$ one may observe that $z_i(\mathbf{c}) = j$ if and only if $c_{jj} = 1$ and $r_{ij}(\mathbf{c}) = 1$ which is still correct table assignment notation. Thus we enumerate tables not by abstract ordering $1, 2, \ldots, |\mathbf{z}(\mathbf{c})|$, but rather by customer numbers $1, 2, \ldots, N$. Below we denote $z_{ij}(\mathbf{c}) = c_{jj} r_{ij}(\mathbf{c})$.

To derive variational inference we need to define the factorized approximation of the posterior of seqddCRP parameters. Similarly to the variational algorithms for DP we choose the factorization that holds true when there are no data (i.e. for the priors). In DP such a property was true for the parameters responsible for the breaking ratio of the stick v_k (see section 2). In the case of ddCRP such parameters are c_i. Indeed, in the absence of any data the distribution $p(\mathbf{c}, \mathbf{\Theta}|\eta)$ is fully factorized. We consider such prior-induced factorization to be the most natural choice.

So we approximate posterior $p(\mathbf{c}, \mathbf{\Theta}|\mathbf{x}, \eta)$ as a fully factorized distribution $q(\mathbf{c}, \mathbf{\Theta}) = \prod_i q(c_i) q(\theta_i)$ by minimizing Kullback-Leibler divergence between the true posterior and its approximation, which is equivalent to maximization of marginal likelihood lower bound (Jordan et al., 1999):

$$\log p(\mathbf{x}) \geq \mathcal{L} = \mathbb{E}_q \Bigg[\sum_{i=1}^{N} \log p(c_i|f, D, \alpha) - \log q(c_i) + \sum_{j=1}^{N} \log p(\theta_j|G_0) - \log q(\theta_j) + \sum_{i=1}^{N} z_{ij}(\mathbf{c}) \log p(x_i|\theta_j) \Bigg]$$

The main difficulty, both conceptual and computational in deriving variational inference for ddCRP, is in expectation of table assignments w.r.t variational distribution $q(\mathbf{c})$, that is $\mathbb{E}_q z(\mathbf{c})$. Table assignments $z(\mathbf{c})$ are computed by deterministic function which maps independent customer assignments into table assignments. It is global in the sense that even if we are interested only whether customer i sits at the table started by customer j (that is $z_{ij}(\mathbf{c})$), in general case information about all customer assignments $\mathbf{c}$ is required to answer this question. The opposite is also true: changing just one customer assignment c_i may cause global change of table assignments. We now describe how to compute this expectation for seqddCRP and provide efficient variational inference algorithm.

4.1. Expected table assignments via inverse of the Laplacian

Consider directed graph modeled by seqddCRP. It consists of N vertices, one for each customer, all of them having exactly one link c_i pointing to another customer $j \leq i$ with probability $q(c_i = j)$. We define then the probabilistic adjacency matrix A such that $A_{ij} = q(c_i = j)$ for $i \neq j$ and $A_{ii} = 0$ for all i.

Further we denote $R_{ij} = \mathbb{E}_{q(\mathbf{c})} r_{ij}(\mathbf{c})$ which is the probability of existence of the directed path from customer i to customer j. One may observe that vertex i may be connected with j directly with probability A_{ij} or through other vertices. By noting that $R_{ii} = 1$ for all i[1] for $j > i$ we get:

$$R_{ij} = A_{ij} + \sum_{t<i} A_{it} R_{tj} = \sum_{t} A_{it} R_{tj}$$

We may vectorize this formula and obtain j-th column of R:

$$R_{\cdot j} = A \cdot R_{\cdot j} + e_j = (I - A)^{-1} e_j,$$

where e_j is j-th column of identity matrix. Finally, we obtain the whole matrix:

$$R = (I - A)^{-1} \tag{5}$$

Matrix $L = I - A$ is non-singular since it's triangular and has ones on main diagonal. It could be viewed as *expected Laplacian*[2] of random graph induced by the seqddCRP.

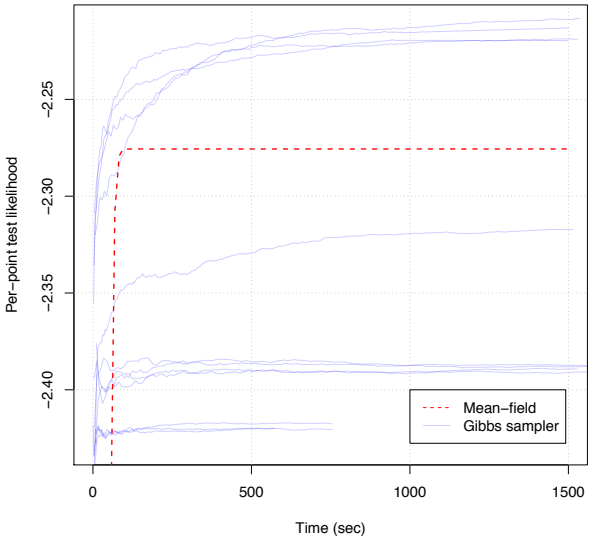

Figure 2. Per-point test data likelihood as function of time on time-dependent mixture modeling. For variational inference only best run is drawn.

[1] It may be confusing that $r_{ii}(\mathbf{c}) = 1$ regardless of the value $q(c_i = i)$. We emphase that former statement is correct as long as we are interested in *reachability* and not *table assignments*, while $q(c_i = i)$ 1) governs table initiation and 2) restrains probabilities of outcoming links since $\sum_{j \neq i} q(c_i = j) = 1 - q(c_i = i)$.

[2] By definition i-th diagonal element of Laplacian is either inner or outer degree of vertex i depending on the application (outer in our case). Thus strictly speaking our matrix L may not be valid Laplacian if $q(c_i = j) = 0$ for some i and all $j \neq i$, but this is insignificant for further reasoning as we do not rely on any of Laplacian properties.

Laplacian matrices are widely used in spectral graph theory for finding important graph properties, see e.g. (Chung, 1996).

To get expected table assignments, recall that $z_{ij}(\mathbf{c}) = c_{jj} r_{ij}(\mathbf{c})$ and use (5):

$$\mathbb{E}z(\mathbf{c}) = (I - A)^{-1}\text{diag}(\mathbb{E}\mathbf{c}) \tag{6}$$

Thus expected table assignments could be computed as column-weighted inverse of expected Laplacian induced by factorized distribution of customer assignments. This connection allows for deterministic inference algorithms and also provides lower bound for their computational complexity as matrix inverse operation is being involved.

Note that equation (5) holds true not only for approximated posterior $q(\mathbf{c})$ but also for prior $p(\mathbf{c}|D, \alpha, f, a)$ since it is also factorized. This allows us to analytically obtain various properties of prior table assignments, e.g. expected size of j-th table $\sum_{i \geq j} \mathbb{E}z_{ij}$ which wasn't possible before. Based on those properties one may select the values of hyper-parameters for the model such as decay function f or α.

4.2. Coordinate ascent algorithm

We iteratively perform fixed-point updates of each variational distribution $q(c_i)$ by setting derivative of KL divergence with respect to corresponding distribution to zero. Updates for $q(\theta_j)$ depend on a particular form of distributions and are quite straightforward, so we focus on updates for $q(\mathbf{c})$.

While we may naively use the result from the previous section to obtain the update equations, this would require a number of matrix inverse operations which is computationally expensive. Thus, we derived an efficient Algorithm 1 requiring only one implicit matrix inverse per iteration and relying on Woodbury (1950) matrix identity. We provide the details on derivation of efficient variational updates in the appendix.

The algorithm has complexity $O(N^3)$ per iteration which is equal to cost of a matrix inverse required to compute table assignments and variational lower bound and thus may be considered as quite efficient. It is possible to further improve it's computational performance by adding and removing clusters ad-hoc, i.e. clusters with probability of existence $q(c_k = k)$ below some threshold may not be taken into account.

Note that it is possible to update variational distributions $q(c_i)$ in any order and we have empirically observed that random order updates perform much better than sequential ones.

Algorithm 1 Variational inference for seqddCRP

Input: data $\mathbf{x}$, initial $q(\mathbf{c})$ and $q(\mathbf{\Theta})$, hyperparameters η
Compute reachability matrix R according to eq. (5)
repeat
  for $i = 1$ **to** N **do**
    Initialize zero vector $a_i \in \mathbb{R}^N$
    for $k = 1$ **to** i **do**
      $a_{ik} \leftarrow \sum_{i \geq i} R_{si} \mathbb{E}_q \log p(x_s|\theta_k)$
    end for
    Initialize $\gamma_{ij} = \log p(c_i = j|\eta)$ for all j
    for $j = 1$ **to** i **do**
      $\gamma_{ij} \leftarrow \gamma_{ij} + \sum_{k \leq i} a_{ik} R_{jk} q(c_k = k)$
    end for
    Exponentiate and normalize γ_i, update $q(c_i) \leftarrow \gamma_i$
    Perform rank-1 update to R by Woodbury formula
  end for
  for $k = 1$ **to** N **do**
    Update $q(\theta_k) \propto p(\theta_k) \prod_{i \geq k} p(x_i|\theta_k)^{R_{ik} q(c_k = k)}$
  end for
until convergence

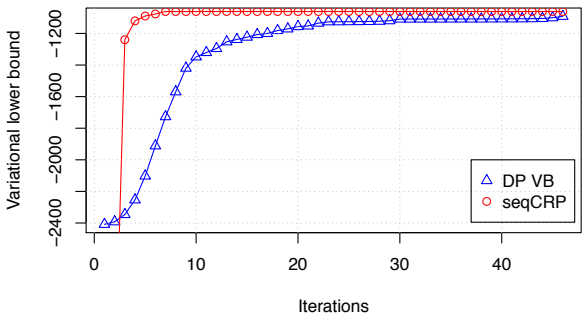

Figure 3. Variational lower bound over iterations for DP mixture with $R = 5$

5. EXPERIMENTS

In this section we empirically compare our variational algorithms with a set of baselines. We start from comparison with Gibbs sampler as it is currently the only available inference algorithm for ddCRP. Next we compare our variational inference for CRP-equivalent mixture model and variational inference for Dirichlet Process. Then we evaluate distance-dependent mixture model against CRP mixture. We release our software implementation used for the experiments[3].

[3] http://github.com/sbos/seqddcrp.jl

Table 1. Best results for predictive likelihood

algorithm/R	seqddCRP	seqCRP
1	-863.55	-965.76
2	-900.70	-905.75
3	-864.93	-888.10
4	-825.26	-836.58
5	-689.38	-691.96

5.1. Language modeling

Following (Blei & Frazier, 2011) we model natural text as fully observed seqddCRP where individual tokens $\mathbf{w}$ are organized in tables and each table then draws a word from discrete distribution over words V (which is just word frequencies). This is very simple "unigram" model, yet it is especially suitable for comparison with Gibbs sampler, because there are no hidden variables and thus nothing to collapse out. Note that derivation of collapsed variational inference is usually non-trivial task and comparison of collapsed and non-collapsed model would be flawed, hence it was decided to use this task for demonstration of computational efficiency of inference algorithms.

Our dataset consisted from 2246 news articles from the Associated Press, we performed word stemming, but did not remove stop-words. Distances between tokens was defined just as number of tokens between two given. Sigmoid decay function was used with hyper-parameters set as for best model in (Blei & Frazier, 2011).

We compared absolute convergence speed of our variational algorithm comparing to Gibbs sampler by visually assessing convergence of posterior estimates, in particular, expected number of tables was chosen as statistics of interest since it is important quantity in nonparametric analysis. We provide such a plot for a randomly selected document on figure 6. It was observed that our variational algorithm converged in just one iteration on all documents and estimates made with two considered algorithms are very close.

5.2. Mixture of Gaussians

We continue our experimental study with continuous mixture modeling. We generated 5 datasets each drawn from mixture of 5 two-dimensional Gaussian distributions with equal weights, spherical covariance and mean located in $(0, 0)$ and $(-R, -R)$, $(-R, R)$, $(R, -R)$, (R, R) respectively, varying R from 1 to 5. $R = 1$ means that clusters are almost undistinguishable and $R = 5$ makes them easely separable. Each dataset was split into 200 train data points and 200 test data points.

It is common to compare different models by *predictive likelihood* which is $p(\mathbf{x}_{\text{test}}|\mathbf{x})$ assuming that well-fitted model assigns high probability to test data which is obtained from the same context as train data, e.g. held-out part of the document. Unfortunately this quantity is intractable for ddCRP and thus we use the following estimate as test likelihood:

$$p(\mathbf{x}_{\text{test}}|\mathbf{x}, \hat{\Theta}) = \sum_{\mathbf{c}_{\text{test}}, \mathbf{c}} p(\mathbf{c}|\mathbf{x}, \hat{\Theta}) p(\mathbf{c}_{\text{test}}) p(\mathbf{x}_{\text{test}}|\mathbf{c}_{\text{test}}, \mathbf{c}, \hat{\Theta}) \quad (7)$$

where $\hat{\Theta}$ are cluster parameters estimated as posterior mean.

Since the purpose of test likelihood is to measure how well estimated cluster parameters explain unseen data, new clusters emerged in test data could flaw the results. It is impossible in traditional parametric models and in DP mixture models based on truncated stick-breaking, however ddCRP allows to assign $\mathbf{c}_{\text{test},i} = i$ for some i. Thus we restrict test data to start new tables by setting corresponding prior probabilities to zero and re-normalizing prior.

5.2.1. seqCRP and DP VB

We denote CRP-equivalent mixture model formulated as special case of seqddCRP as seqCRP below in the paper. We compare our variational inference algorithm for this model and one for DP based on truncated stick-breaking (Blei & Jordan, 2005), further we denote it as VB DP. We set parameter $\alpha = 0.1$ for both models to encourage small number of clusters and provided weak informative priors for covariance matrix to slightly suggest it's spherical form. Truncation level for DP VB was set to 50.

Since both algorithms are dependent of random initialization we performed 300 runs in each setting and selected best results. We observed that seqCRP achieved higher variational lower bound and since both models actually represent the same mixture model, it could be admitted that our variational approximation is tighter. Figure 4 contains histograms of variational lower bounds on datasets with $R = 5$ and $R = 3$. In the first case where the mixture is easily separable, best lower bound obtained by our approximation was -1881.59 comparing to -1886.29 from VB DP, and in the second case where it is harder to recover true number of clusters best results were -1891.73 and correspondingly -1892.66.

We also compared convergence rate for both algorithms (see fig. 3). Our variational algorithm is considerably slower because it involves matrix inverses extensively. However, it converges much faster than VB DP in number of iterations.

This empirically demonstrates potential of our variational algorithm which could be preferred over DP VB in situations when tight approximation is more important than run-

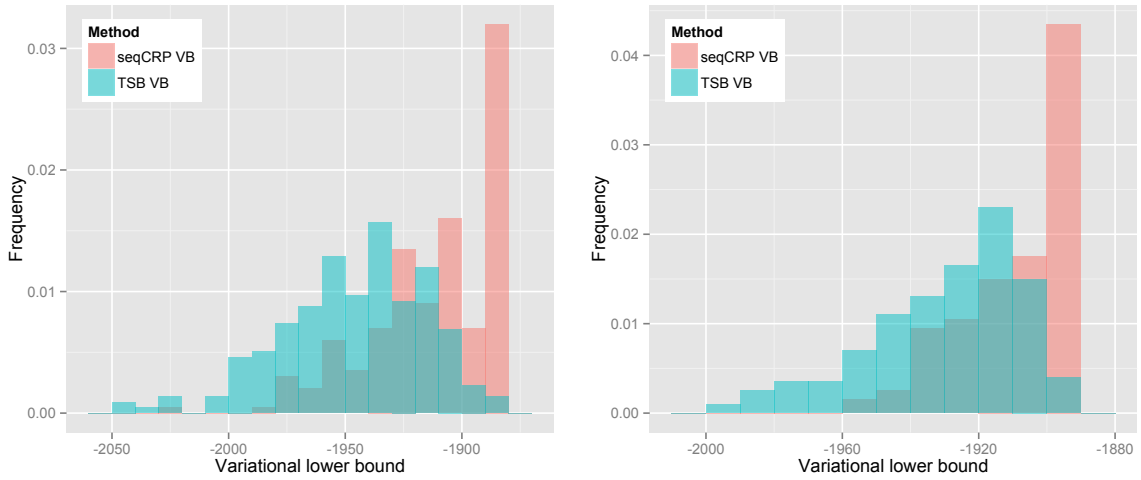

Figure 4. Histograms of converged variational bounds across random initializations on $R = 5$ (left) and $R = 3$ (right) for DP-equivalent mixture.

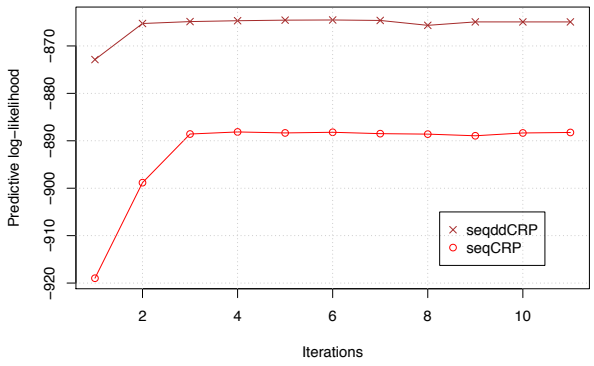

Figure 5. Predictive log-likelihood over iterations for informative and non-informative distance prior

ning time.

5.2.2. SEQCRP AND SEQDDCRP

Finally we evaluate seqddCRP with informative prior in the sense that data points generated from the same mixture component have lower distances than every pair of points which are not. In particular, we have used exponential decay function with parameter $a = 4$ meaning moderate decay and sparseness inducing $\alpha = 0.1$, as in language modeling task distances were plain and measured in number of data points between two given. Train data was generated sequentially from each Gaussian, while test data was randomly permuted and for them uninformative CRP distances were used.

Clearly, such information improved test likelihood comparing to plain seqCRP, see table 1. Also for both models test likelihood converged very fast, detailed graph is on fig. 5. Note that we didn't tune parameters of the process and thus even more performance increase could be achieved. This suggests seqddCRP as an alternative to various sequential mixture models such as Hidden Markov Model.

5.3. Time-dependent mixture

Motivated by previous experiment we evaluate non-collapsed Gibbs sampler and our variational algorithm on modeling of time-dependent normal mixtures. Data was generated from a number of states, each associated with its own normal distribution, by a random process visualized on figure 1. In each moment of time exactly one state is active and the the process draws a point from the corresponding distribution. It also may switch between states. Each normal distribution has mean $5 \cdot k$ where k is its number and random precision drawn from Gamma distribution with shape 0.8 and scale 1.1. This Gamma distribution was used as a prior for both algorithms. After generative process finished, we randomly permuted several points near state switches in order to simulate perturbations in transitive periods often observed in real data.

The setting was the same as in 5.2.2, except we used convenient average per-point test likelihood as evaluation metric (see e.g. (Blei & Jordan, 2005)). Besides it is tractable for both sample and analytic computation obtained from variational distribution, we empirically found that it is highly correlated with test likelihood estimate we used before (equation 7).

Results of the comparison are shown on figure 2. Clearly,

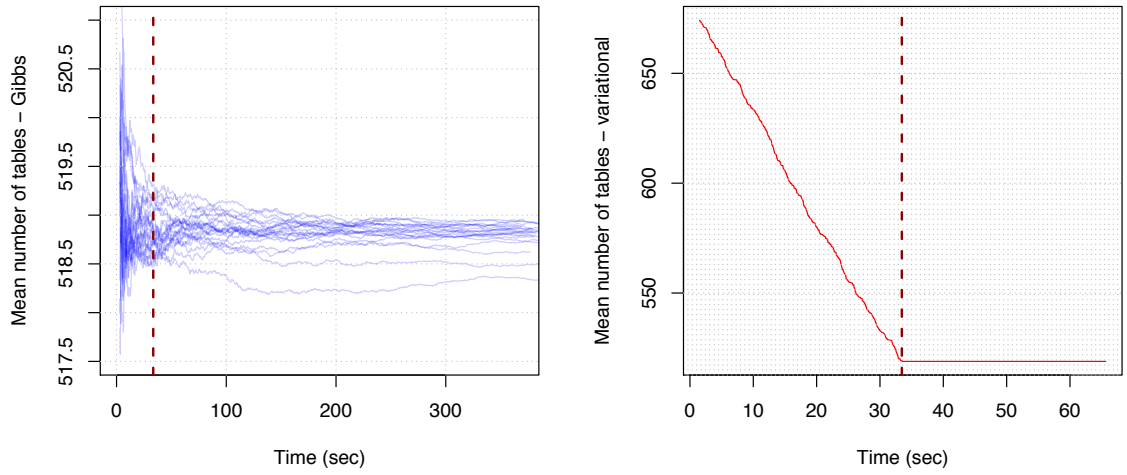

Figure 6. Mean number of tables over time for language modeling task over time

variational inference outperforms Gibbs sampler in convergence speed although the latter may achieve better test likelihood in time depending on the initialization.

6. CONCLUSION

In the paper we proposed mean-field approximation for sequential distance-dependent Chinese Restaurant process and developed efficient coordinate ascent algorithm. Our inference procedure is closely connected with Laplacian of random graph modeled by seqddCRP. We used special factorization w.r.t. customer assignments which is not only natural for seqddCRP but is also applicable for conventional CRP. For the latter case it can be regarded as an alternative to well-known stick-breaking variational approximation. We showed that it might yield better lower bounds of marginal probabilities for Dirichlet process mixture models. For the general case the only available inference framework is based Gibbs sampler. The proposed framework could serve as its faster deterministic analogue.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. Sergey Bartunov is supported by RFBR grant 14-01-31361, Dmitry P. Vetrov is supported by RFBR grants 12-01-00938 and 12-01-33085.

Appendix: Efficient variational update

The general form of the variational update for each $q(c_i)$ is (with hyper-parameters and explicit dependency on $\mathbf{c}$ omitted for clarity):

$$\log q(c_i) = \mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} \left[\log p(\mathbf{x}, \mathbf{c}, \boldsymbol{\Theta})\right] = \log p(c_i) + \sum_{s=1}^{N}\sum_{k=1}^{N} \log p(x_s|\theta_k)\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} z_{sk} + \text{const}$$

where we denote $\mathbf{c}_{\setminus \mathbf{i}}$ as a set of all customer assignments except for i-th customer.

Now consider how $\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} z_{sk} = q(c_k = k)\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{sk}$ depends on c_i. First note that if $s < i$ or $k > i$ then r_{sk} does not depend on assignment of i-th customer due to sequential property of the process. Denote $r^{\setminus i}_{sk} = 1$ if there exists a directed path from s to k that avoids i. Note that if $r_{sk} = 1$ then it immediately implies $r_{si} = 0$. Then we may write

$$r_{sk} = 1 \iff (r_{ik} = 1 \text{ and } r_{si} = 1) \text{ or } (r^{\setminus i}_{sk} = 1 \text{ and } r_{si} = 0)$$

Equivalently in algebraic form

$$1 - r_{sk} = (1 - r_{ik}r_{si})(1 - r^{\setminus i}_{sk}(1 - r_{si}) = 1 - r_{ik}r_{sk} - r^{\setminus i}_{sk}(1 - r_{si}) + r_{ik}r_{si}r^{\setminus i}_{sk}(1 - r_{si}) = (1 - r_{si}) = 1 - r_{ik}r_{sk} - r^{\setminus i}_{sk}(1 - r_{si})$$

In last equation only the second item depends on c_i. Let $c_i = t$. Then observing that $r_{ik} = r_{c_t k}$ we express the dependence on c_i by explicit formula:

$$\log q(c_i = t) = \log p(c_i = t) + + \sum_{s=1}^{N}\sum_{k=1}^{N} \mathbb{E}_q \log p(x_s|\theta_k)\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{tk}(\mathbf{c}) r_{si}(\mathbf{c}) c_{kk} + \text{const}.$$

First note that $r_{c_i k}$, r_{si}, and c_{kk} are independent of each other. Also note that $\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{si}$ does not depend on $q(c_j)$ for $j < i$. The value of $\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{tk}$ also does not depend on $q(c_j)$ for $j > t$, so we may compute it when updating $q(c_t)$ as follows:

$$\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{tk} = \sum_{j<t} q(c_t = j)\mathbb{E}_{q(\mathbf{c}_{\setminus \mathbf{i}})} r_{jk}.$$

References

Aldous, D.J. Exchangeability and related topics. pp. 1–198, 1985. Lecture Notes in Math. 1117.

Blei, David M. and Frazier, Peter I. Distance Dependent Chinese Restaurant Processes. *J. Mach. Learn. Res.*, 12: 2461–2488, November 2011. ISSN 1532-4435.

Blei, David M. and Jordan, Michael I. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

Chung, Fan R. K. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, December 1996. ISBN 0821803158.

Duan, A., Guindani, Michele, and Gelfand, Alan E. Generalized spatial dirichlet process models. In *Duke University*, pp. 05–23, 2005.

Ferguson, Thomas S. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364. doi: 10.2307/2958008. URL http://dx.doi.org/10.2307/2958008.

Ghosh, Soumya, Ungureanu, Andrei B., Sudderth, Erik B., and Blei, David M. Spatial distance dependent chinese restaurant processes for image segmentation. In Shawe-Taylor, John, Zemel, Richard S., Bartlett, Peter L., Pereira, Fernando C. N., and Weinberger, Kilian Q. (eds.), *NIPS*, pp. 1476–1484, 2011.

Griffin, Jim E. and Steel, Mark F. J. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

Griffiths, Tom L. and Ghahramani, Zoubin. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, 2006.

Haghighi, Aria and Klein, Dan. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 385–393, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125.

MacEachern, S. Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.

Sethuraman, Jayaram. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Woodbury, Max A. *Inverting Modified Matrices*. Number 42 in Statistical Research Group Memorandum Reports. Princeton University, Princeton, NJ, 1950.

Xue, Ya, Dunson, David, and Carin, Lawrence. The matrix stick-breaking process for flexible multi-task learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 1063–1070, Corvalis, Oregon, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273630.