

---

# Collapsed Variational Inference for Time-varying Dirichlet Process Mixture Models

---

Amr Ahmed      Eric Xing  
School of Computer Science  
Carnegie Mellon University

## 1 Introduction

Dirichlet process mixture models provide a flexible Bayesian framework for estimating a distribution as an infinite mixture of simpler distributions that could identify latent classes in the data. However the full exchangeability assumption they employ makes them an unappealing choice for modeling longitudinal data such as text, audio and video streams that can arrive or accumulate as epochs, where data points inside the same epoch can be assumed to be fully exchangeable, whereas across the epochs both the structure (i.e., the number of mixture components) and the parameterizations of the data distributions can evolve and therefore unexchangeable. Several approaches have been proposed to introduce temporal dependencies between distinct DPs [1, 2, 3, 4] by utilizing different representations of DP. Notwithstanding these developments, inference is mainly carried via Gibbs sampling which poses computational challenges when confronted with large datasets. In this paper we focus on the Temporal DPM, a framework for introducing temporal dependencies between DPs that we introduced earlier in [1], and show how to carry out collapsed variational inference in this model.

## 2 Temporal Dirichlet Process Mixture Model

The Dirichlet process (DP) is a distribution over distributions [5]. A DP denoted by  $DP(G_0, \alpha)$  is parameterized by a base measure  $G_0$  and a concentration parameter  $\alpha$ . We write  $G \sim DP(G_0, \alpha)$  for a draw of a distribution  $G$  from the Dirichlet process.  $G$  itself is a distribution over a given parameter space  $\theta$ , therefore we can draw parameters  $\theta_{1:N}$  from  $G$ . Integrating out  $G$ , the parameters  $\theta$  follow a Polya urn distribution [6], also known as a Chinese restaurant process (CRP), in which the previously drawn values of  $\theta$  have strictly positive probability of being redrawn again, thus making the underlying probability measure  $G$  discrete with probability one. More formally,

$$\theta_i | \theta_{1:i-1}, G_0, \alpha \sim \sum_k \frac{n_k}{i-1+\alpha} \delta(\phi_k) + \frac{\alpha}{i-1+\alpha} G_0. \quad (1)$$

where,  $\phi_{1:k}$  denotes the distinct values among the parameters  $\theta$ , and  $n_k$  is the number of parameter  $\theta$  having value  $\phi_k$ . By using the DP at the top of a hierarchical model, one obtains the Dirichlet process mixture model, DPM [7]. The generative process thus proceeds as follows:

$$G | \alpha, G_0 \sim DP(\alpha, G_0), \quad \theta_n | G \sim G, \quad x_n | \theta_n \sim F(\cdot | \theta_n), \quad (2)$$

where  $F$  is a given likelihood function parameterized by  $\theta$ .

Several approaches have been proposed to introduce temporal dependencies in DPs [1, 2, 3, 4], to name a few. Here we focus on the temporal DPM introduced in [1]. The temporal Dirichlet process mixture model (TDPM) is a framework for modeling complex longitudinal data, in which the number of mixture components at each time point is unbounded; the components themselves can retain, die out or emerge over time; and the actual parameterization of each component can also evolve over time in a Markovian fashion. In TDPM, the random measure  $G$  is time-varying, and the process stipulates that:

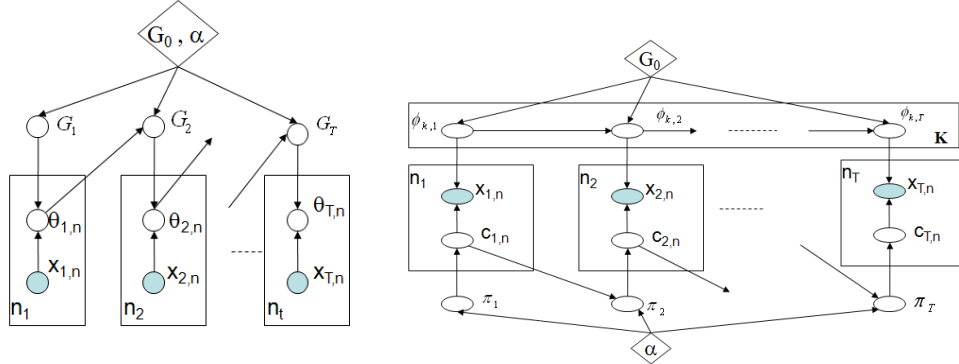


Figure 1: Left, the TDPM and right, the finite fixed-dimensional model construction. The figure shows a first-order TDPM to avoid cluttering the display but the text assumes a generic  $\Delta$ -order process.

$$G_t | \phi_{1:k}, G_0, \alpha \sim DP \left( \alpha + \sum_k m'_{kt}, \sum_k \frac{m'_{k,t}}{\sum_l m'_{lt} + \alpha} \delta(\phi_k) + \frac{\alpha}{\sum_l m'_{lt} + \alpha} G_0 \right) \quad (3)$$

where  $\{\phi_{1:k}\}$  are the collection of unique values of  $\theta_{1:t-\Delta}$ , and  $m'_{kt} = \sum_{\delta=1}^{\Delta} \exp \frac{-\delta}{\lambda} m_{k,t-\delta}$ , where  $m_{kt}$  is the number of parameters in  $\theta_{t,\cdot}$  associated with component  $k$  and  $\Delta, \lambda$  define the width and decay factor of the time-decaying kernel. Moreover, the parameterization  $\phi_k$  of each component changes over time in a markovian fashion, i.e.:  $\phi_{kt} | \phi_{k,t-1} \sim P(\cdot | \phi_{k,t-1})$ . Integrating out the random measures  $G_{1:T}$ , the parameters  $\theta_{1:t}$  follows a poly-urn distribution with time-decay, or the recurrent Chinese restaurant process. More formally:

$$\theta_{ti} | \theta_{t-1:t-\Delta}, \theta_{t,1:i-1}, G_0, \alpha \propto \sum_k \left( m'_{kt} + m_{kt} \right) \delta(\phi_{kt}) + \alpha G_0 \quad (4)$$

Finally, [1] gives a third construction of the same process using the limit of a finite dynamic mixture models with  $K$  mixtures as shown in Figure 1, whose generative process proceeds as follows: for each  $t$  do,

1.  $\forall k$ : Draw  $\phi_{kt} \sim P(\cdot | \phi_{k,t-1})$
2. Draw  $\pi_t \sim \text{Dir}(m'_{1t} + \alpha/K, \dots, m'_{Kt} + \alpha/K)$
3.  $\forall i \in N_t$  Draw  $c_{ti} \sim \text{Multi}(\pi_t)$ ,  $x_{ti} \sim F(\cdot | \phi_{c_{ti},t})$

where  $c_{ti}$  is the mixture component associated with data  $x_{ti}$ . By integrating over the mixing proportion  $\pi_t$ , It is quite easy to show the equivalence with the TDPM as  $k \rightarrow \infty$  (see [1] for more details).

### 3 Collapsed Variational Inference

The inference problem is to compute  $p(c, \phi | \mathbf{x})$ . This problem was addressed using a Gibbs sampling algorithm in [1], here we provide a collapsed VI algorithm based on a truncation level  $K$  of the finite model in Figure 1. We collapse the  $\pi$  variables from the model, and maintain the dependency between the atoms' locations over time, thus we have the following variational distribution:

$$q(c, \phi) = \prod_t \prod_i q(c_{ti}) \prod_k q(\phi_k) \quad (5)$$

Using the method in [8, 9, 10], the variational update can be found as follows. The form for  $q(\phi_k)$  depends on the emission and dynamic models. For instance, for a linear dynamics and gaussian emission, this distribution can be calculated using the RTS smoother using the expected sufficient statistics under  $q(\cdot)$ , while for logistic normal emission, an approximate variational analogous is available either via Laplace [11] or Taylor approximation [12]. Here we focus on the form for  $q(c_{ti})$ , it can be shown that:

$$q(c_{ti} = k) \propto \exp\left(\mathbb{E}_{q(c)}\left[\log p(c_{ti} = k | \mathbf{c}_{t:t-\Delta})\right] + \mathbb{E}_{q(\phi_k)}\left[\log p(x_{t,i} | \phi_{k,t})\right]\right) \times \exp\left(\sum_{\delta=1}^{\Delta} \mathbb{E}_{q(c)}\left[\log p(\mathbf{c}_{t+\delta} | \mathbf{c}_{t:t+\delta-\Delta}, c_{ti} = k)\right]\right) \quad (6)$$

where

$$P(c_{ti} = k | \dots) \propto \frac{m'_{kt} + m_{kt}^{-ti} + \alpha}{\sum_t m'_{it} + N_t - 1 + \alpha} \quad (7)$$

$$P(\mathbf{c}_{t+\delta} | \mathbf{c}_{t:t+\delta-\Delta}, c_{ti} = k) = \frac{\Gamma\left(\sum_s m'_{s,t+\delta}{}^{c_{ti}=k} + \alpha/K\right) \prod_s \Gamma\left(m'_{s,t+\delta}{}^{c_{ti}=k} + m_{s,t+\delta} + \alpha/K\right)}{\prod_s \Gamma\left(m'_{s,t+\delta}{}^{c_{ti}=k} + \alpha/K\right) \Gamma\left(\sum_s m'_{s,t+\delta}{}^{c_{ti}=k} + m_{s,t+\delta} + \alpha/K\right)} \quad (8)$$

Note that the terms of the form  $\Gamma(\sum \dots)$  in (8) and the denominator in (7) are fixed regardless of  $k$  and thus will be factored out when normalizing (6). Plugging (7,8) into (6) we get,

$$q(c_{ti} = k) \propto \exp\left(\mathbb{E}_{q(c)}\left[\log(m'_{kt} + m_{kt}^{-ti} + \alpha)\right] + \mathbb{E}_{q(\phi_k)}\left[\log p(x_{t,i} | \phi_{k,t})\right]\right) \times \exp\left(\sum_{\delta=1}^{\Delta} \sum_{s=1}^K \mathbb{E}_{q(c)}\left[\log \Gamma\left(m'_{s,t+\delta}{}^{c_{ti}=k} + m_{s,t+\delta} + \alpha/K\right)\right] - \mathbb{E}_{q(c)}\left[\log \Gamma\left(m'_{s,t+\delta}{}^{c_{ti}=k} + \alpha/K\right)\right]\right) \quad (9)$$

Now we focus on calculating expectations of the terms in (9). First note that:  $m_{kt} + m'_{kt} = \sum_{i=1}^{N_t} \mathbb{I}(c_{ti} = k) + \sum_{\delta=1}^{\Delta} \sum_{i=1}^{N_t-\delta} \mathbb{I}(c_{t-\delta,i} = k) \exp^{-\frac{\delta}{\lambda}}$ . Under the factorized variational distribution, the distribution of this quantity can be calculated using convolution as noted in [13], however this is still intractable for large datasets. Following the insight in [13], and using the central limit theorem, this quantity can be approximated by a Gaussian distribution with mean and variance given by:

$$\begin{aligned} \mathbb{E}_{q(c)}\left[m_{kt} + m'_{kt}\right] &= \sum_{i=1}^{N_t} q(c_{ti} = k) + \sum_{\delta=1}^{\Delta} \sum_{i=1}^{N_t-\delta} q(c_{t-\delta,i} = k) \exp^{-\frac{\delta}{\lambda}} \\ \text{Var}_{q(c)}\left[m_{kt} + m'_{kt}\right] &= \sum_{i=1}^{N_t} q(c_{ti} = k) \left(1 - q(c_{ti} = k)\right) \\ &\quad + \sum_{\delta=1}^{\Delta} \sum_{i=1}^{N_t-\delta} q(c_{t-\delta,i} = k) \left(1 - q(c_{t-\delta,i} = k)\right) \left(\exp^{-\frac{\delta}{\lambda}}\right)^2 \end{aligned} \quad (10)$$

Using the above approximation, the expectation of the terms in (7,8) can be approximated using a second-order Taylor expansion as follows:  $\mathbb{E}\left[f(m)\right] = f\left(\mathbb{E}[m]\right) + \frac{1}{2} f''\left(\mathbb{E}[m]\right) \text{Var}[m]$ . Note the second derivative of  $\log \Gamma(\cdot)$  is given by the  $\psi_1(\cdot)$ , which is the trigamma function. As noted in [13], this Gaussian approximation works well in practice in the context of the static DPM. Moreover, the expectation of the second term in 6 depends on the emission model. Finally, smart caching can be utilized when evaluating (9,10) across different values of  $k, \delta$ .

We are still in the process of implementing this scheme and we expect to provide preliminary results over simulation data in the workshop and contrast this inference algorithm with the Gibbs sampling algorithm in [1].

## References

- [1] A. Ahmed and E.P. Xing. Non-parametric mixture models and the recurrent chinese restaurant process with application to evolutionary clustering. In *SDM*, 2008.

- [2] F. Caron, M. Davy, and A. Doucet. Generalized polya urn for time-varying dirichlet processes. In *UAI*, 2007.
- [3] X. Zhu Z. Ghahramani and J. Lafferty. Time-sensitive dirichlet process mixture models. In *Technical Report CMU-CALD-05-104*, 2005.
- [4] J.E. Griffin and M.F.J. Steel. Order-based dependent dirichlet processes. In *Journal of the American Statistical Association*, volume 101, pages 1566–1581, 2006.
- [5] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.
- [6] D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.
- [7] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- [8] E.P. Xing, M.I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *UAI*, 2003.
- [9] C. Bishop, D. SPIEGELHALTER, and J. Winn. Vibes: A variational inference engine for bayesian networks. In *NIPS 15*, 2002.
- [10] M. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. In *Technical Report 649, Dept. Statistics, U.C. Berkeley*, 2003.
- [11] A. Ahmed and E.P. Xing. Modeling topic evolution in text streams. In *Technical Report, CMU-ML-07-117*, 2006.
- [12] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, 2006.
- [13] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational dirichlet process mixture models. In *IJCAI*, 2007.