Stochastic Collapsed Variational Inference for Sequential Data

Pengyu Wang[1] **Phil Blunsom**[1,2]
[1]Department of Computer Science, University of Oxford
[2]Google DeepMind
{pengyu.wang, phil.blunsom}@cs.ox.ac.uk

Abstract

Stochastic variational inference for collapsed models has recently been successfully applied to large scale topic modelling. In this paper, we propose a stochastic collapsed variational inference algorithm in the sequential data setting. Our algorithm is applicable to both finite hidden Markov models and hierarchical Dirichlet process hidden Markov models, and to any datasets generated by emission distributions in the exponential family. Our experiment results on two discrete datasets show that our inference is both more efficient and more accurate than its uncollapsed version, stochastic variational inference.

1 Background

A hidden Markov model (HMM) [1] consists of a hidden state sequence $\mathbf{z} = \{z_t\}_{t=0}^T$ and a corresponding observation sequence $\mathbf{x} = \{x_t\}_{t=1}^T$. Let there be K hidden states. For convenience, we let the start state be 0 and set $z_0 = 0$. Let θ be the transition matrix where $\theta_{k,k'} = p(z_t = k'|z_{t-1} = k)$, and θ_0 be the initial state distribution where $\theta_{0,k'} = p(z_1 = k')$. A hierarchical Dirichlet process HMM (HDP-HMM) [2, 3] allows to use an unbounded number of hidden states by constructing an infinite mean vector π from a stick breaking process and drawing transition vectors θ_k from the shared π. We have for $k = 1, 2, ...$ and for $k' = 1, 2, ...,$

$$\pi_{k'} = \tilde{\pi}_{k'} \prod_{l=1}^{k'-1} (1 - \tilde{\pi}_l) \qquad \tilde{\pi}_{k'} \sim \text{Beta}(1, \gamma) \qquad \theta_k \sim \text{DP}(\alpha, \pi). \tag{1}$$

A hidden sequence is generated by a first order Markov process, and each observation is generated conditioned on its hidden state. We have for $t = 1, ..., T$,

$$z_t|z_{t-1} = k \sim \text{Mult}(\theta_k) \qquad x_t|z_t = k' \sim p(\cdot|\phi_{k'}), \tag{2}$$

where $\phi_{k'}$ parametrizes the observation likelihoods for $k' = 1, 2, ...,$ with $\phi_{k',w} = p(x_t = w|z_t = k')$. We assume that the observation likelihoods and their conjugate prior take exponential forms:

$$p(w|\phi_{k'}) = h_l(w) \exp\{\phi_{k'}^T t(w) - a_l(\phi_{k'})\} \tag{3}$$

$$p(\phi_{k'}|\lambda^\circ) = h_g(\phi_{k'}) \exp\{(\lambda_1^\circ)^T \phi_{k'} + (\lambda_2^\circ)^T(-a_l(\phi_{k'})) - a_g(\lambda^\circ)\}. \tag{4}$$

The base measure h and log normalizer a are scalar functions; and the parameter $\phi_{k'}$ and sufficient statistics t are vector functions. The subscripts l and g represent the local hidden variables and global model parameters, respectively. The dimensionality of the prior hyperparameter $\lambda^\circ = (\lambda_1^\circ, \lambda_2^\circ)$ is equal to $\dim(\phi_{k'}) + 1$. For a complete Bayesian treatment, we place vague Gamma priors on α and γ, $\alpha \sim \text{Gamma}(a_\alpha^\circ, b_\alpha^\circ)$ and $\gamma \sim \text{Gamma}(a_\gamma^\circ, b_\gamma^\circ)$. The graphical model is shown in figure 1 (left).

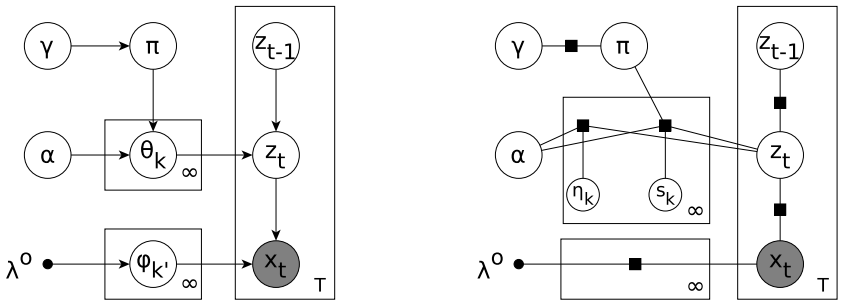

Figure 1: Left: an HDP-HMM. Right: a collapsed HDP-HMM with auxiliary variables. Here we suppress the first order Markov dependencies into one plate repeated T times to fit the page.

2 Stochastic Collapsed Variational Inference

HMMs and HDP-HMMs are popular probabilistic models for modelling sequential data. However, their traditional inference methods such as variational inference (VI) [4] and Markov chain Monte Carlo (MCMC) [3, 5] are not readily scalable to large datasets (e.g., one dataset in our experiment contains 5 million sequences with combined length over 100 million). In this paper, we follow the success of stochastic collapsed variational inference (SCVI) for latent Dirichlet allocation (LDA) [6], and we propose a scalable SCVI algorithm for HMMs and HDP-HMMs. Our algorithm achieved better predictive performances than the stochastic variational inference (SVI) [7] applied to HMMs [8] and to HDP-HMMs [9].

We present our derivation in the following three steps: 1. we marginalize out the model parameters (θ, ϕ); 2. we derive stochastic updates for each sequence; 3. we derive stochastic updates for the posteriors of the HDP parameters (α, β, γ) (for HDP-HMMs only). For notational simplicity, we consider a dataset of N sequences each of length T. That is $\mathbf{x} = \{\mathbf{x}^n\}_{n=1}^N$ and $\mathbf{x}^n = \{x_t^n\}_{t=1}^T$. Similarly, we write for hidden sequences $\mathbf{z} = \{\mathbf{z}^n\}_{n=1}^N$ and $\mathbf{z}^n = \{z_t^n\}_{t=1}^T$.

2.1 Collapsed HDP-HMMs

There is substantial empirical evidence [10, 11, 6] that marginalizing the model parameters is helpful for both accurate and efficient inference. The marginal data likelihood of an HDP-HMM is:

$$p(\mathbf{x}, \mathbf{z}) = \prod_{k=0}^{K} \frac{\Gamma(\alpha)}{\Gamma(\alpha + C_{k\cdot})} \prod_{k'=1}^{K} \frac{\Gamma(\alpha\pi_{k'} + C_{kk'})}{\Gamma(\alpha\pi_{k'})} \prod_{n=1}^{N} \prod_{t=1}^{T} h_l(x_t^n) \prod_{k'=1}^{K} \exp\{a_g(\lambda^{k'}) - \{a_g(\lambda^\circ)\}. \quad (5)$$

The gamma functions and log normalizers come from the marginalization. $C_{kk'}$ denotes the transition count from the hidden state k to k', $C_{kk'} = \#\{n, t : z_{t-1}^n = k, z_t^n = k'\}$. dot denotes the summed out column, e.g., $C_{\cdot k'} = \sum_k C_{kk'}$. $\lambda^{k'}$ denotes the posterior hyperparameter for the hidden state k', $\lambda_1^{k'} = \lambda_1^\circ + \sum_{n=1}^N \sum_{t=1}^T t(x_t^n)\delta(z_t^n = k')$ and $\lambda_2^{k'} = \lambda_2^\circ + C_{\cdot k'}$, where δ is the standard delta function.

The gamma functions are a nuisance to take derivatives of (5). Following [12], we replace them by integrals of some auxiliary variables η and $\mathbf{s}$ and the joint likelihood becomes:

$$\begin{aligned} & p(\mathbf{x}, \mathbf{z}, \eta, \mathbf{s}) \\ &= \prod_{k=0}^{K} \frac{\eta_k^{\alpha-1}(1-\eta_k)^{C_{k\cdot}-1}}{\Gamma(C_{k\cdot})} \prod_{k'=1}^{K} \begin{bmatrix} C_{kk'} \\ s_{kk'} \end{bmatrix} (\alpha\pi_{k'})^{s_{kk'}} \prod_{n=1}^{N} \prod_{t=1}^{T} h_l(x_t^n) \prod_{k'=1}^{K} \exp\{a_g(\lambda^{k'}) - \{a_g(\lambda^\circ)\} \quad (6) \end{aligned}$$

where $\eta_k \in [0, 1]$ is Beta distributed, $s_{kk'} \in \{0, 1, ..., C_{kk'}\}$ is the number of tables labelled with k' in the k^{th} Chinese restaurant in a Chinese restaurant franchise, and $\begin{bmatrix} C_{kk'} \\ s_{kk'} \end{bmatrix}$ is unsigned Stirling number of the first kind. The factor graph with the auxiliary variables is given in figure 1 (right).

We are interested in the posterior $p(\mathbf{z}, \eta, \mathbf{s}, \gamma, \alpha, \tilde{\pi}|\mathbf{x})$. As the exact computation is intractable, we introduce a variational distribution in a tractable family,

$$q(\mathbf{z}, \eta, \mathbf{s}, \gamma, \alpha, \tilde{\pi}) = q(\mathbf{z})q(\eta|\mathbf{z})q(\mathbf{s}|\mathbf{z})q(\gamma)q(\alpha)q(\tilde{\pi}), \quad (7)$$

and we maximize the evidence lower bound (ELBO) denoted by $\mathcal{L}(q)$,

$$\log p(\mathbf{x}) \geq \mathbb{E}[\log p(\mathbf{x}, \mathbf{z}, \eta, \mathbf{s}, \gamma, \alpha, \tilde{\pi})] - \mathbb{E}[\log q(\mathbf{z}, \eta, \mathbf{s}, \gamma, \alpha, \tilde{\pi})] \triangleq \mathcal{L}(q). \tag{8}$$

By the 'direct assignment truncation' [12, 9], we set the truncation level to be K. That is $q(\mathbf{z} = 0)$ if for any n and t such that $z_t^n > K$.

2.2 Inference for Sequences

To infer $q(\mathbf{z})$, we factorize it as a product of independent sequences, $q(\mathbf{z}) = \prod_{n=1}^{N} q(\mathbf{z}^n)$. Combining the work of SCVI for LDA [6] and CVI for HMM [11], we randomly sample $\mathbf{x}^n$ with $n \sim \mathcal{U}[1, N]$, and we derive the update for $q(\mathbf{z}^n)$ with a zeroth order Taylor approximation [13]:

$$q(\mathbf{z}^n) \propto \prod_{t=1}^{T} \hat{\theta}_{z_{t-1}^n, z_t^n} \prod_{t=1}^{T} \hat{\phi}_{z_t^n, x_t^n} \qquad \hat{\theta}_{k,k'} \propto \mathbb{G}[\alpha \pi_{k'}] + \mathbb{E}[C_{kk'}] \tag{9}$$

$$\hat{\phi}_{k',w} \propto h(w) \exp\{a_g(\lambda_1^\circ + t(w) + \mathbb{E}[t_{k'}(\mathbf{x}, \mathbf{z})], \lambda_2^\circ + 1 + \mathbb{E}[C_{\cdot k'}])\}, \tag{10}$$

in which $\mathbb{G}$ denotes the geometric expectation, $\mathbb{E}[C_{kk'}]$ denotes the expected transition count from state k to k', and $\mathbb{E}[t_{k'}(\mathbf{x}, \mathbf{z})] = \sum_{n=1}^{N} \sum_{t=1}^{T} q(z_t^n = k') t(x_t^n)$ denotes the emission statistics at the hidden state k'. The details on expectations that appear in the paper are in Appendix A.

As $q(\mathbf{z}^n)$ is proportional to a HMM parametrized by the surrogate parameters $\hat{\theta}$ and $\hat{\phi}$, we can use the forward backward algorithm [14]. After collecting the local transition counts $\mathbb{E}[C_{kk'}^n]$ and emission statistics $\mathbb{E}[t_{k'}(\mathbf{x}^n, \mathbf{z}^n)]$, we update the global statistics by taking a weighted average:

$$\mathbb{E}[C_{kk'}] = (1 - \rho_n)\mathbb{E}[C_{kk'}] + \rho_n N \mathbb{E}[C_{kk'}^n] \tag{11}$$

$$\mathbb{E}[t_{k'}(\mathbf{x}, \mathbf{z})] = (1 - \rho_n)\mathbb{E}[t_{k'}(\mathbf{x}, \mathbf{z})] + \rho_n N \mathbb{E}[t_{k'}(\mathbf{x}^n, \mathbf{z}^n)], \tag{12}$$

where ρ_n is the step size satisfying $\sum_n \rho_n^2 \leq \infty$ and $\sum_n \rho_n = \infty$.

Unlike CVI for HMM [11], our algorithm is memory efficient, since we update $q(\mathbf{z}^n)$ without subtracting the local statistics, as such they do not need to be explicitly stored.

2.3 Inference for HDP

For notational clarity, we write the variational posteriors of the HDP parameters to be governed by their variational parameters. We have,

$$q(\tilde{\pi}_{k'}) = \text{Beta}(u_{k'}, v_{k'}) \qquad q(\alpha) = \text{Gamma}(a_\alpha, b_\alpha) \qquad q(\gamma) = \text{Gamma}(a_\gamma, b_\gamma). \tag{13}$$

We derive stochastic updates for the HDP posteriors. For a randomly selected sequence $\mathbf{x}^n$, we form an artificial dataset $\{\mathbf{x}^{n^{(N)}}, \mathbf{z}^{n^{(N)}}\}$ consisting N replicates of the observed and hidden sequences $\{\mathbf{x}^n, \mathbf{z}^n\}$. Assuming we can compute $\mathbb{E}[s_{kk'}^{(N)}]$ and $\mathbb{E}[\log \eta_k^{(N)}]$ based on the artificial dataset, we derive the intermediate variational parameters and take a weighted average with their old estimates. Hence, we have the following updates ($\mathbb{E}[\log(1 - \tilde{\pi}_{k'})]$ in (16) is also a function of $\mathbb{E}[s_{kk'}^{(N)}]$):

$$u_{k'} = (1 - \rho_n)u_{k'} + \rho_n(1 + \mathbb{E}[s_{\cdot k'}^{(N)}]) \quad v_{k'} = (1 - \rho_n)v_{k'} + \rho_n(\mathbb{E}[\gamma] + \mathbb{E}[s_{\cdot > k'}^{(N)}]) \tag{14}$$

$$a_\alpha = (1 - \rho_n)a_\alpha + \rho_n(a_\alpha^\circ + \mathbb{E}[s_{\cdot\cdot}^{(N)}]) \quad b_\alpha = (1 - \rho_n)b_\alpha + \rho_n(b_\alpha^\circ - \textstyle\sum_k \mathbb{E}[\log \eta_k^{(N)}]) \tag{15}$$

$$a_\gamma = (1 - \rho_n)a_\gamma + \rho_n(a_\gamma^\circ + K) \quad b_\gamma = (1 - \rho_n)b_\gamma + \rho_n(b_\gamma^\circ - \textstyle\sum_{k'} \mathbb{E}[\log(1 - \tilde{\pi}_{k'})]) \tag{16}$$

where dot denotes the summed out column, and $> k'$ denotes summing over l for $l > k'$. The details on computing $\mathbb{E}[s_{kk'}^{(N)}]$ and $\mathbb{E}[\log \eta_k^{(N)}]$ are in Appendix B. Stochastic optimizations often benefit from the use of minibatches, to reduce the variance of noisy samples and the updating time of variational parameters. Thus we propose to update the global statistics after a minibatch is processed, and to update the HDP posteriors after a larger batch is processed. Altogether, our SCVI algorithm for HDP-HMMs is given in Appendix C, and it applies to HMMs by removing the outermost loop.

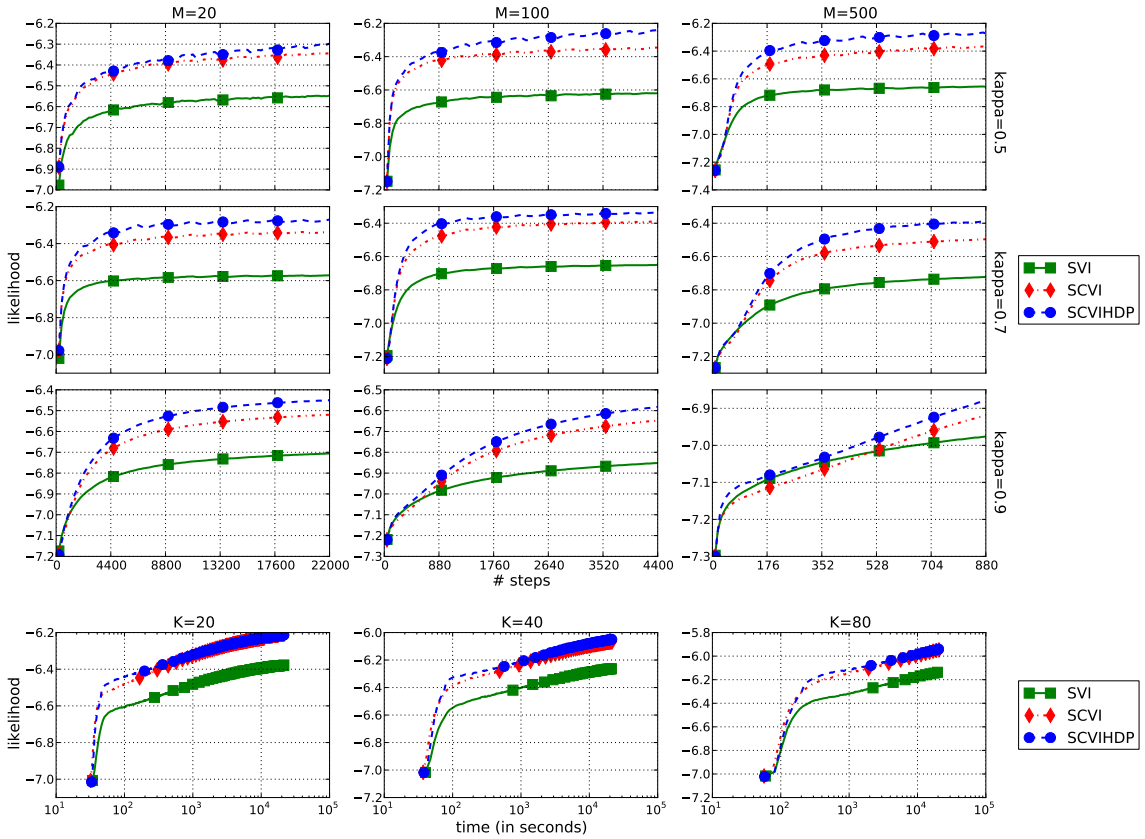

Figure 2: Top three rows: comparison on WSJ under various minibatch sizes M and forgetting rates κ. Bottom row: comparison on NYT under various (truncated) numbers of hidden states K.

3 Experiments

We evaluated our SCVI algorithm applied to HMMs (denoted by SCVI) and applied to HDP-HMMs (denoted by SCVIHDP) compared to the SVI algorithm applied to HMMs [8] (denoted by SVI). SVI applied to HDP-HMMs was omitted, since we were unable to make noticeable improvement over SVI using a point estimate of the top level stick π [9]. We used two discrete datasets, the Wall Street Journal (WSJ) and New York Times (NYT). Both datasets are made of sentences. For each sentence, the underlying sequence can be understood as a Markov chain of hidden part-of-speech (PoS) tags [15] and words are drawn conditioned on PoS tags, making (HDP)-HMMs natural models. We used the predictive log likelihoods as our evaluation metrics.

For SVI and SCVI, we set the transition priors to $\text{Dir}(0.1)$, to encourage sparsity. For SCVIHDP, we set the HDP priors to be vague, $\text{Gamma}(1, 0.1)$. We set $\mathbb{G}[\alpha\pi_{k'}] = 0.1$ for the first iteration such that all the algorithms started with the same transition prior counts. Finally, all the emission priors were set to $\text{Dir}(0.1)$; all the global statistics $\mathbb{E}[C_{k,k'}]$ and $\mathbb{E}[t_{k'}(\mathbf{x}, \mathbf{z})]$ were initialized using exponential distributions, as suggested by [7].

The first three rows in figure 2 presents the predictive log likelihood results of three inferences on WSJ ($49,000$ sentences, 90% for training and 10% for testing). We fixed the number of hidden states (or truncation level) $K = 45$[1] and varied the minibatch sizes M and forgetting rates κ, which parametrize the step sizes $\rho_n = (1+n)^{-\kappa}$. The large batch size was set to be $10,000$ for SCVIHDP. We let each inference run through the dataset 10 times and reported the per time step likelihoods. In all the settings, our SCVI outperformed SVI by large margins, extending the success of SCVI for

[1]One goal of our experiments is to show the improvement by sharing statistics using our HDP inference. Using a larger truncation level than the number of hidden states would put our SCVIHDP in an advantageous position and we would not be able to identify the source of improvement.

LDA [6] to time series data. Further, our collapsed HDP inference helped SCVIHDP to surpass our SCVI by noticeable margins.

The forth row in figure 2 presents the predictive log likelihood results of three inferences on NYT (5 million sentences, 99% for training and 1% for testing) using the complementary settings to WSJ. We fixed $\kappa = 0.5$ and $M = 1000$ and varied K. We ran all the algorithms (implemented in Cython) for 6 hours and reported the likelihood results versus wall-clock time. We see that given the same time, our SCVI converged much better than SVI. Our SCVIHDP overlapped with our SCVI towards the end, but it was always better prior to that, making better use of its time.

4 Conclusion

In this paper, we have presented a general stochastic collapsed variational inference algorithm that is scalable to very large time series datasets, memory efficient and significantly more accurate than the existing SVI algorithm. Our algorithm is also the first truly variational algorithm for HDP-HMMs, avoiding point estimates, and it comes with performance gains. For future work, we aim to derive the true nature gradients of the ELBO to prove and further speed up the convergence of our algorithm [16], although we never saw a nonconverging case in our experiments.

References

[1] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. pages 267–296, 1990.

[2] Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002.

[3] 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.

[4] Matthew Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.

[5] Jurgen Van Gael, Yunus Saatci, Yee W. Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden markov model. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1088–1095, New York, NY, USA, 2008. ACM.

[6] James R. Foulds, L. Boyles, C. DuBois, Padhraic Smyth, and Max Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *KDD*, 2013.

[7] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[8] Nicholas Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems 27*, pages 3599–3607. 2014.

[9] Matthew Johnson and Alan Willsky. Stochastic variational inference for Bayesian time series models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1854–1862. JMLR Workshop and Conference Proceedings, 2014.

[10] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States, 2009.

[11] Pengyu Wang and Phil Blunsom. Collapsed Variational Bayesian Inference for Hidden Markov Models. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Scottsdale, AZ, USA, 2013.

[12] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

[13] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *In Advances in Neural Information Processing Systems, volume 19*, 2007.

[14] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[15] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.

[16] Francisco J. R. Ruiz, Neil D. Lawrence, and James Hensman. True natural gradient of collapsed variational bayes. In *NIPS Workshop on Advances in Variational Inference*, Montreal, 2014.

Appendix A

In this section, we present the standard (geometric) expectations that appeared in the main paper.

If x is Beta distributed, $p(x|u,v) \propto x^{u-1}(1-x)^{v-1}$, (e.g., $\tilde{\pi}_{k'}$ in the main paper), we have,

$$\mathbb{E}[\log x] = \psi(u) - \psi(u+v) \qquad \mathbb{E}[\log(1-x)] = \psi(v) - \psi(u+v) \tag{17}$$

If x is Gamma distributed, $p(x|a,b) \propto x^{a-1}e^{-bx}$ (e.g., α in the main paper), we have,

$$\mathbb{E}[x] = a/b \qquad \mathbb{G}[x] = e^{\psi(a)}/b \tag{18}$$

If x and y are independent, (e.g., α and $\pi_{k'}$ in the main paper), we have $\mathbb{G}[xy] = \mathbb{G}[x]\mathbb{G}[y]$.

Appendix B

In this section, we present the details on computing $\mathbb{E}[s^{(N)}_{kk'}]$ and $\mathbb{E}[\log \eta^{(N)}_k]$. For $\mathbb{E}[s^{(N)}_{kk'}]$, we notice the inequality[2]: $\mathbb{E}[s^{(N)}_{kk'}] \neq N\mathbb{E}[s^n_{kk'}]$. Thus we compute it as follows:

$$\mathbb{E}[s^{(N)}_{kk'}] \approx \mathbb{G}[\alpha\pi_{k'}]q(C^{(N)}_{kk'} > 0)(\psi(\mathbb{G}[\alpha\pi_{k'}] + \mathbb{E}_+[C^{(N)}_{kk'}]) - \psi(\mathbb{G}[\alpha\pi_{k'}])) \tag{19}$$

$$q(C^{(N)}_{kk'} > 0) = 1 - q(C^{(N)}_{kk'} = 0) = 1 - \exp\{N \log q(C^n_{kk'} = 0)\} \tag{20}$$

$$q(C^n_{kk'} = 0) \approx \exp\{\textstyle\sum_t \log(1 - q((z^n_{t-1}, z^n_t) = (k,k')))\} \tag{21}$$

$$\mathbb{E}_+[C^{(N)}_{kk'}] \approx N\mathbb{E}[C^n_{kk'}]/q(C^{(N)}_{kk'} > 0) \tag{22}$$

The approximation in (19) comes from the technique proposed by Teh et al. detailed in [12]. In (20), $q(C^{(N)}_{kk'} > 0)$ denotes the probability of at least one transition from state k to state k'; and the second equality comes from the fact that $\mathbf{z}^n$ is repeated N times under exactly the same distribution. In (21) and (22), we propose a fast approximate method. We partition a hidden sequence $\mathbf{z}^n$ into a set of overlapping but independent clusters $\{(z^n_t, z^n_{t+1})\}_{t=1}^T$. Allowing to overlap is sufficient to preserve all the pairwise transition information, while making the independence assumption permits the above linear computations as in [12]. The same strategy applies to computing $\mathbb{E}[\log \eta^{(N)}_k]$.

$$\mathbb{E}[\log \eta^{(N)}_k] \approx q(C^{(N)}_{k\cdot} > 0)(\psi(\mathbb{E}[\alpha]) - \psi(\mathbb{E}[\alpha] + \mathbb{E}_+[C^{(N)}_{k\cdot}])) \tag{23}$$

$$q(C^{(N)}_{k\cdot} > 0) = 1 - q(C^{(N)}_{k\cdot} = 0) = 1 - \exp\{N \log q(C^n_{k\cdot} = 0)\} \tag{24}$$

$$q(C^n_{k\cdot} = 0) \approx \exp\{\textstyle\sum_t \log(1 - q((z^n_{t-1}) = (k)))\} \tag{25}$$

$$\mathbb{E}_+[C^{(N)}_{k\cdot}] \approx N\mathbb{E}[C^n_{k\cdot}]/q(C^{(N)}_{k\cdot} > 0) \tag{26}$$

Appendix C

In this section, we present our SCVI algorithm for HDP-HMMs.

Algorithm 1 SCVI for HDP-HMMs (and for HMMs by deleting the outermost loop)

```
Randomly initialize E[C_{k,k'}] and E[t_{k'}(x, z)]
for each large batch do
  for each mini batch do
    for each sequence x^n do
      update q(z^n) by Eqs. (9,10)
    end for
    update E[C_{k,k'}], E[t_{k'}(x, z)] by Eqs. (11,12)
  end for
  update q(π), q(α), q(γ) by Eqs. (14,15,16)
end for
```

[2] $\mathbb{E}[s^n_{kk'}]$ is the expected number of tables k' in the k^{th} restaurant. The inequality holds by the property of CRP: the expected number of tables grows not linearly but logarithmically with the number of customers.