Stochastic Collapsed Variational Inference for Nonparametric Hidden Markov Models

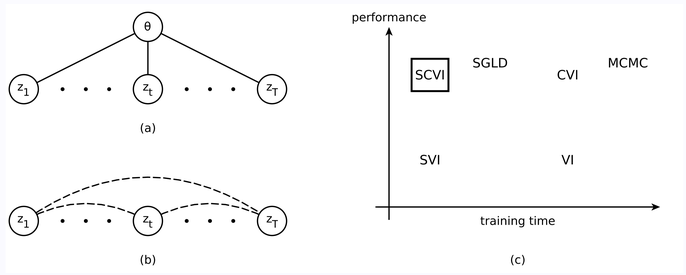
Pengyu Wang, Phil Blunsom

Department of Computer Science, University of Oxford, UK
email: firstname.lastname@cs.ox.ac.uk

Introduction

Stochastic Collapsed Variational Inference (SCVI)

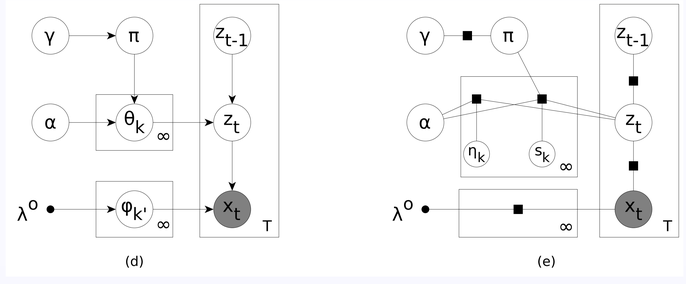
- In (a), Variational Inference (VI) breaks the strong dependencies between the parameters θ and the hidden variables $\mathbf{z} = z_1, ..., z_t, ..., z_T$.
- In (b), Collapsed Variational Inference (CVI) breaks the weak dependencies between the hidden variables in the collapsed space.
- (c) summarizes the efficiencies and performances of VI, CVI, MCMC and their minibatch-based inferential counterparts, namely stochastic VI (SVI), stochastic CVI (SCVI), stochastic gradient Langevin dynamics (SGLD).



- There has been little research on whether and how we can apply the recent advanced approximate inference algorithms in a time depedent data setting.

Hierarchical Dirichlet Process Hidden Markov Models (HDP-HMMs)

- The graphical model of a HDP-HMM is shown in (d).
- After introducing the auxiliary variables, the collapsed representation of a HDP-HMM is shown in (e).



- Goal: 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,

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

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

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

- The primary challenge is the sequential dependencies, which stand in the way of updating both the subchains and the HDP posteriors.

Our Contribution One

Inference of Subchain Posteriors

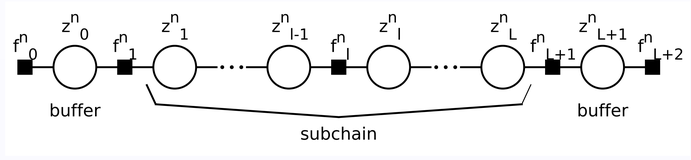
- Idea: break a long Markov chain into a set of subchains, $q(\mathbf{z}) = \prod_{n=1}^{N} q(\mathbf{z}^n)$.

$$q(\mathbf{z}^n) \approx\propto \hat{\theta}_{\cdot,z_1^n}\left(\prod_{l=2}^{L}\hat{\theta}_{z_{l-1}^n,z_l^n}\right)\hat{\theta}_{z_L^n,\cdot}\left(\prod_{l=1}^{L}\hat{\phi}_{z_l^n,x_l^n}\right)$$

- Challenge: the boundary transitions $\hat{\theta}_{\cdot,z_1^n}$ and $\hat{\theta}_{z_L^n,\cdot}$ prevent us from running the standard forward backward algorithm.
- Our Solution: expand $q(\mathbf{z}^n)$ to $q(\mathbf{z}^{(n)}, z_0^n, z_{L+1}^n)$,

$$q(\mathbf{z}^{(n)}, z_0^n, z_{L+1}^n) \propto f_0^n(z_0^n)\left(\prod_{l=1}^{L+1} f_l^n(z_{l-1}^n, z_l^n)\right) f_{L+2}^n(z_{L+1}^n)$$

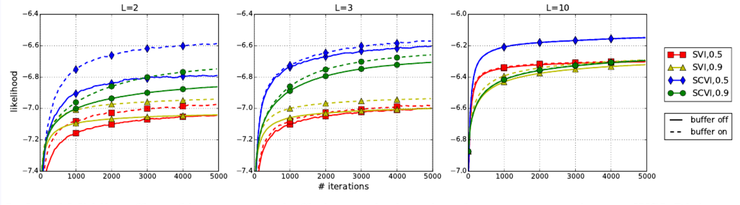
which can be represented by a factor graph,



define the potential functions to make sure $\sum_{z_0^n, z_{L+1}^n} q(\mathbf{z}^{(n)}, z_0^n, z_{L+1}^n) = q(\mathbf{z}^n)$, and develop a novel sum product algorithm.

Experiment One

- We evaluated the utility of our buffering method and compared the performances of our SCVI algorithm against the SVI algorithm on two datasets, the Wall Street Journal (WSJ) and New York Times (NYT).
- We created two very long synthetic time series by joining sentences. As the evaluation metrics, we used predictive log likelihoods by holding out small subsets.
- We fixed $L \times M = 1000$, where L is the subchain length and M is the minibatch size. We varied L and the forgetting rates, $\kappa = 0.5, 0.9$.



Left and Middle: effect of incorporating buffering methods and performance comparison on WSJ. Right: performance comparison on NYT.

- In most settings our SCVI algorithm outperformed the SVI algorithm by large margins. When L is small, there are noticeable improvements using respective buffering methods in both algorithms. For SCVI, we attribute the improvement to the inter subchain communication through the buffering variables.

Our Contribution Two

Inference of HDP Posteriors

- Basics: the HDP posteriors are governed by the variational parameters, which we aim to update after a minibatch. For example, $q(\tilde{\pi}_{k'}) = \text{Beta}(u_{k'}, v_{k'})$.
- Idea: take a weighted average of the intermediate variational parameters on the N replicates of (x^n, z^n) with their old estimates. For example, the update equation for $u_{k'}$ is $u_{k'} := (1-\rho_n)u_{k'} + \rho_n(1 + \mathbb{E}[s_{\cdot k'}^{(N)}])$.
- Challenge: compute $E[s_{kk'}^N]$ (the expected number of tables in the metaphor of the Chinese Restaurant Process) on the N replicates of (x^n, z^n). $E[s_{kk'}^N]$ is not a linear function of N. That is $E[s_{kk'}^N] \neq NE[s_{kk'}^n]$.
- Our Solution: use the approximation technique by Teh et al. (2008),

$$\mathbb{E}[s_{kk'}^{(N)}] \approx \mathbb{G}[\alpha\pi_{k'}]q(C_{kk'}^{(N)} > 0)(\psi(\mathbb{G}[\alpha\pi_{k'}] + \mathbb{E}_+[C_{kk'}^{(N)}]) - \psi(\mathbb{G}[\alpha\pi_{k'}])),$$

linearly and exponentially scale other local statistics, respectively,

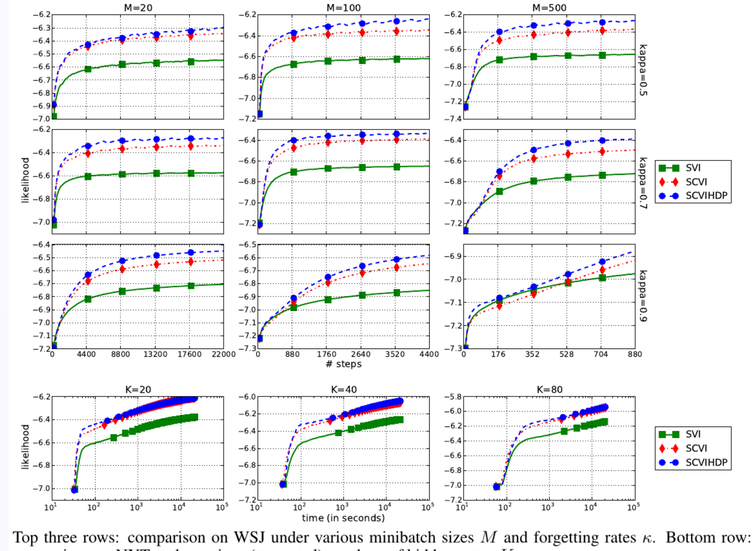
$$\mathbb{E}_+[C_{kk'}^{(N)}] = N\mathbb{E}[C_{kk'}^n]/q(C_{kk'}^{(N)} > 0)$$
$$q(C_{kk'}^{(N)} > 0) = 1 - q(C_{kk'}^{(N)} = 0) = 1 - \exp\{N \log q(C_{kk'}^n = 0)\},$$

and develop a fast approximation method,

$$q(C_{kk'}^n = 0) \approx \exp\{\textstyle\sum_l \log(1 - q((z_{l-1}^n, z_l^n) = (k, k')))\}.$$

Experiment Two

- The data and metric are the same as in the experiment one except that we consider each sentence as an independent subchain.



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.

- In all cases our SCVI with HDP inference performed the best.