Averaged Collapsed Variational Bayes Inference and its application to Infinite Relational Model

Katsuhiko Ishiguro
NTT Communication Science Labs
Kyoto, Japan
ishiguro.katsuhiko@lab.ntt.co.jp

Issei Sato
The University of Tokyo
Tokyo, Japan
sato@r.dl.itc.u-tokyo.ac.jp

Naonori Ueda
NTT Communication Science Labs
Kyoto, Japan
ueda.naonori@lab.ntt.co.jp

Abstract

A newly introduced Collapsed variational Bayes (CVB) inference is known for better inference performance then the original VB. However, CVB has no guarantee of convergence unlike VB and maximum likelihood EM algorithm. We overcome this problem by proposing a simple and easy-to-use CVB algorithm, called Averaged CVB (ACVB). ACVB has two advantages. First, ACVB posterior update assures convergence thanks to its use of a simple annealing technique. Second, the stationary point of the CVB lower bound is equivalent to the converged solution of ACVB, if the lower bound has a stationary point (which has yet to be nailed down in the literature). ACVB is applicable for general probabilistic models with discrete hidden variables such as LDA, and is equally valid for CVB and CVB0. We apply ACVB for Infinite Relational Model (IRM), which is a basic probabilistic model for relational data clustering, and conduct experiments to validate the performance of ACVB.

1 Introduction

Recently, collapsed variational Bayes (CVB) solutions have been intensively studied as a new inference algorithm for probabilistic models, especially for topic models [16, 2, 14, 13]. The original paper [16] examined a 2nd-order Taylor approximation of the variational expectation. A simpler 0th order-approximated CVB (CVB0) has been also developed, and its property was studied in [14] by using α-divergence. These papers report that CVB and CVB0 yield better inference results than VB or collapsed Gibbs, in data modeling [11, 16, 2], link predictions, and neighborhood search [13].

However, CVB inference has one drawback: its lack of convergence guarantee makes it differ from naive VB and maximum-likelihood EM algorithm. This problem, interestingly, has not been much discussed in the literature. However, this is a tricky and problematic issue for practitioners who are not familiar with but want to try state-of-the-art machine learning techniques. Users are required to determine the convergence of CVB inference manually: this is not an easy task for non-expert users. In that sense, CVB is not as favorable as naive VB and EM algorithms.

In this paper, we propose a simple and easy-to-use CVB algorithm, called Averaged CVB (ACVB), that overcomes this problem. ACVB has two advantages. First, ACVB posterior update offers assured convergence thanks to its simple annealing mechanism. Second, the stationary point of the CVB lower bound is equivalent to the converged solution of ACVB, if the lower bound has a stationary point (an issue unresolved in the literature). Our formulation is applicable to any model,

and is equally valid for CVB and CVB0. Convergence-guaranteed ACVB is the preferred choice for practitioners who want to apply state-of-the-art inference to their problems.

We apply ACVB to the Infinite Relational Model (IRM) [9], which is a relatively simple nonparametric Bayes relational data clustering model. Experiments show that ACVB0, which is a combination of CVB0 and ACVB, is better than VB and ACVB in terms of test data log likelihood and computation time to convergence.

2 CVB inference

Let us denote observed data as $\boldsymbol{X}$, hidden variables as $\boldsymbol{Z}$, and parameters as $\boldsymbol{\Theta}$. For the time being, we assume the hidden variables are discrete: like as cluster (topic) indicators in LDA and GMM models. CVB inference maximizes the following lower bound w.r.t. $q(Z)$:

$$\mathscr{L}(\mathbf{Z}) = \int q(\mathbf{Z}) \log \frac{p(\boldsymbol{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}. \tag{1}$$

The above lower bound collapses out the parameter Θ. This lower bound is always tighter than that of naive VB [16], indicating CVB inference is better than VB.

To derive the variational posterior q of hidden variable $z_i \in \mathbf{Z}$, we take a partial derivative of Eq. (1) w.r.t. $q(z_i)$ to obtain the stationary point. In general, we have the following update rule [3]:

$$\log q(z_i) = \mathbb{E}_{\mathbf{Z}^{\setminus i}} \left[\log p(\boldsymbol{X}, \mathbf{Z}) \right] + (\text{const.}), \tag{2}$$

where $\mathbb{E}_y \left[f(x) \right]$ indicates the expectation of $f(x)$ over the variational posterior $q(y)$, and $\mathbf{Z}^{\setminus i}$ indicates that z_i is excluded from $\mathbf{Z}$. Specific forms of Eq. (2) depend on the models solved. Unfortunately, we cannot exactly evaluate the expectation over $q(\mathbf{Z})$ for CVB. Thus we Taylor-approximate Eq. (1) and derive the update rules of q. Originally, [16] employed a 2nd-order approximation (**CVB**). Many papers (e.g. [2]) claim that the 0th-order approximation (**CVB0**) performs better in topic models.

Due to this Taylor-approximation, the CVB inference update rules do not correctly maximize the lower bound Eq. (1). So there is no guarantee that CVB updates monotonically increase the lower bound. To the best of our knowledge, no researchers have proposed a solution for this problem

3 Proposed: Averaged CVB (ACVB)

To make the CVB inference more useful for practitioners, assured and easy convergence detection is essential. An easy convergence detection algorithm for CVB would allow non-expert users to use CVB inference more easily, which has been reported more precise than naive VB and comparable with collapsed Gibbs.

The theoretical analysis of CVB convergence remains an important but difficult problem. Instead of tackling the problem theoretically, we propose a simple and easy-to-use technique that offers assured automatic CVB inference convergence. Our proposal, called **Averaged CVB (ACVB)**, provides a simple annealing technique. We would like to emphasize that ACVB supports CVB inference on any model (for discrete hidden variables $\mathbf{Z}$). Moreover, it is equally valid for CVB and CVB0. After a certain number of iterations for "burn-in", ACVB gradually decreases the portion of variational posterior changes using the following equation:

$$\bar{q}^{(s+1)} = \left(1 - \frac{1}{s+1}\right) \bar{q}^{(s)} + \frac{1}{s+1} q^{(s+1)}, \quad \text{or} \quad \bar{q}^{(S)} = \frac{1}{S} \sum_{s=1}^{S} q^{(s)}, \tag{3}$$

where s denotes the iterations after completion of the "burn-in" period, $\bar{q}^{(s)}$ denotes the "averaged" (or annealed) variational posterior in the sth iteration, $q^{(s)}$ denotes the variational posterior by CVB inference in the sth iteration, and S is the total number of iterations. After the "burn-in" period, we monitor the ratio of changes of $\bar{q}$ and detect convergence when the ratio falls below a predefined threshold. As the final result, we use $\bar{q}^{(s)}$ instead of $q^{(s)}$. During the burn-in period, we may monitor the changes of q.

ACVB has the following two merits. The first point is rather evident but makes ACVB useful for practical CVB inference. ACVB offers assured convergence, and we can easily detect convergence by taking the difference of $\bar{q}$ in successive iterations.

Theorem 1. *The averaged variational posterior $\bar{q}^{(s)}$ is convergence-assured:* $\forall \epsilon > 0$, $\exists S_0$, *s.t.* $\forall S > S_0 \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left| \bar{q}_i^{(S)} - \bar{q}_i^{(S-1)} \right| < \epsilon$.

Thus, we can automatically stop ACVB inference by creating a halt rule based on the difference of ACVB posteriors.

The next point is noteworthy and validates the use of ACVB in Bayesian inference.

Theorem 2. *If the variational posterior $q^{(s)}$ converges to a stationary point in the CVB lower bound, then the averaged variational posterior $\bar{q}^{(s)}$ also converges to a stationary point in the CVB lower bound.*

We want to stress that it remains unknown in the literature as to whether CVB inference has a stationary point. However, we can still safely use ACVB because it assures convergence of the inference process and ACVB will find the "true" solution if CVB has a stationary point. Such solutions for the convergence of CVB have been never studied, to the best of our knowledge. Please consult [7] for more technical details.

Hereafter, we denote the (naive) CVB solution and the CVB0 solution, both with ACVB, as the **ACVB** solution and the **ACVB0** solution, respectively.

4 Experiments on IRM-ACVB

In this section, we apply ACVB inference to the Infinite Relational Model (IRM) [9], which is a Bayesian nonparametric model for relational data that realizes simultaneous clustering on the row and column dimensions of a given pairwise relational data matrix. As an extended abstract, we skip technical details. For readers who are interested, we refer to [7].

We generated two synthetic dense relation datasets. The size and true numbers of clusters of these datasets were: $N_1 = 100, N_2 = 200, K_1 = 4, K_2 = 5$ (**synth 1**), and $N_1 = 1000, N_2 = 1500, K_1 = 7, K_2 = 6$ (**synth 2**).

The first real-world relational dataset is the **Enron** e-mail dataset [10]. This is a famous relational dataset used in many studies [15, 5, 6, 8]. We extracted monthly e-mails transactions for 2001. The dataset contained $N = N_1 = N_2 = 151$ company members of Enron. $x_{i,j} = 1(0)$ if there is (not) an e-mail sent from member i to member j. Out of twelve months, we selected the transactions of June (**Enron Jun.**), August (**Enron Aug.**), October (**Enron Oct.**), and December (**Enron Dec.**).

The second real-world relational dataset is the **Lastfm** dataset.[1] This dataset contains several records for the Last.fm music service, including lists of users' most listened-to musicians, tag assignments for artists, and friend relations between users. We employ the friend relations between $N = N_1 = N_2 = 1892$ users (**Lastfm UserXUser**). $x_{i,j} = 1(0)$ if there is (not) a friend relation from a user i to a user j. We also employ the artist-tag relations between 17632 artists and 11946 tags. The original observations are the co-occurrence counts of (artist name, tag) pairs. We binarize the observations by examining if the (artist name, tag) pair count is greater than 1 or not: i.e. we ignore one single occasional co-occurrences of (artist name, tag). If the counts are greater than 1, then the observation entries are set to 1: otherwise, set to 0. Then, all rows (artists) and columns (tags) that have no "1" entries are removed. The resulting binary matrix consists of $N_1 = 6099$ artists and $N_2 = 1088$ tags (**Lastfm ArtistXTag**). $x_{i,j} = 1(0)$ if artist i is (not) associated with the tag word j more than once.

The modeling performances of the inference solutions at $K = 20$ and $K = 60$ are presented in Table 1. They show the averages of test data marginal log likelihood after convergence. Results of the best hyperparameter setup are presented for each solution. In addition, we conducted t-tests to examine the statistical significance.

These results reveal the characteristics of the solutions in a few aspects. First, ACVB inferences are significantly better than those of VB for larger datasets: synth2, and two Lastfm datasets. In particular, ACVB0 performed better than the others in several cases. This indicates that, as expected, ACVB inferences are potentially superior to naive VB inferences.

[1] provided by HetRec2011. http://ir.ii.uam.es/hetrec2011/

Table 1: 20-run averages of marginal test data log likelihood per test data entry (10% test data). Larger values are better. Boldface indicates the best method, which is significantly better than the method(s) marked with $*$ (by t-test, $p = 0.05$).

Dataset	$K = 20$			$K = 60$		
	VB	ACVB	ACVB0	VB	ACVB	ACVB0
Synth1	**-0.3260**	-0.3337	-0.3372*	**-0.3281**	-0.3379*	-0.3452*
Synth2	-0.3737*	-0.3348*	**-0.3261**	-0.3736*	-0.3261	**-0.3258**
Enron Jun.	-0.0547	-0.0559	**-0.0540**	-0.0569	-0.0572	**-0.0563**
Enron Aug.	-0.0789	-0.0766	**-0.0763**	-0.0772	-0.0781	**-0.0754**
Enron Oct.	-0.1164*	-0.1098	**-0.1099**	-0.1162	-0.1145	**-0.1139**
Enron Dec.	-0.0693	-0.0686	**-0.0685**	**-0.0682**	-0.0686	-0.0690
Lastfm (UserXUser)	-0.0287*	-0.0271*	**-0.0267**	-0.0287*	-0.0272*	**-0.0267**
Lastfm (ArtistXTag)	-0.0165*	-0.0161*	**-0.0158**	-0.0167*	**-0.0163**	**-0.0163**

Second, we found no advantage to ACVB inferences over VB for smaller datasets: synth1 and Enron datasets. Specifically, VB performed significantly better than ACVB on synth1 data. However the data is artificial, dense and small cross-domain relationships. In general, we don't face such data in actual data analysis applications so the results on larger and sparser data cases are much more informative.

5 Conclusion

We proposed Averaged CVB (ACVB): a simple and easy-to-use solution for convergence guaranteed Collapsed Variational Bayes (CVB) inference. ACVB assures inference convergence, a goal hither to missing from CVB studies. Moreover, we proved that the converged posterior of ACVB may be equivalent to the stationary point of the CVB lower bound. We demonstrated the usefulness of ACVB by applying it to IRM [9].

As a future work, we will further enhance its inference speed. One possible solution is to stochastically approximate the sample size as in SGD. It is also important to examine the possibility of ACVB for more complicated relational data including time-series relational data [6, 5, 4] and mixed-membership models [1, 12]. In theoretical aspect, the convergence property of the original CVB algorithm for general models (not limited to topic models [14]) is difficult but very important question to ask. Finally, we assume that the hidden variables $\boldsymbol{Z}$ are discrete as many generative models assume so. The current approach of ACVB, however, is not necessarily valid for continuous hidden variables. For more broader use of ACVB, we need to solve this problem.

Appendix

In this appendix, we present the proof for theorems (though this is rather evident).

Proof for Theorem 1. The averaged variational posterior $\bar{q}^{(s)}$ is convergence-assured: $\forall \epsilon > 0,\ \exists S_0,$ s.t. $\forall S > S_0 \Rightarrow \frac{1}{N}\sum_{i=1}^{N}\left|\bar{q}_i^{(S)} - \bar{q}_i^{(S-1)}\right| < \epsilon$.

Proof. Since

$$\frac{1}{S}\sum_{s=1}^{S} q^{(s)} = \left(1 - \frac{1}{S}\right)\frac{1}{S-1}\sum_{s=1}^{S-1} q^{(s)} + \frac{1}{S}q^{(S)},$$

we have

$$\begin{aligned}\left|\frac{1}{S}\sum_{s=1}^{S} q^{(s)} - \frac{1}{S-1}\sum_{s=1}^{S-1} q^{(s)}\right| &= \left|-\frac{1}{S}\frac{1}{S-1}\sum_{s=1}^{S-1} q^{(s)} + \frac{1}{S}q^{(S)}\right| \\ &\leq \frac{1}{S}\frac{1}{S-1}\sum_{s=1}^{S-1}|q^{(s)}| + \frac{1}{S}|q^{(S)}| \\ &\leq \frac{1}{S}\frac{1}{S-1}(S-1) + \frac{1}{S} = \frac{2}{S}.\end{aligned}$$

Thus,

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{S}\sum_{s=1}^{S} q_i^{(s)} - \frac{1}{S-1}\sum_{s=1}^{S-1} q_i^{(s)}\right| \leq \frac{2}{S}.$$

If we set $S_0 = \frac{2}{\epsilon}$, then $\forall S > S_0$,

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{S}\sum_{s=1}^{S} q_i^{(s)} - \frac{1}{S-1}\sum_{s=1}^{S-1} q_i^{(s)}\right| \leq \frac{2}{S} < \frac{2}{S_0} = \epsilon.$$

This means

$$\frac{1}{N}\sum_{i=1}^{N}\left|\bar{q}_i^{(S)} - \bar{q}_i^{(S-1)}\right| < \epsilon.$$

□

Proof for Thorem 2. If the variational posterior $q^{(s)}$ converges to a stationary point in the CVB lower bound, then the averaged variational posterior $\bar{q}^{(s)}$ also converges to a stationary point in the CVB lower bound.

Proof. Let q^* be a stationary point in the CVB lower bound. From this assumption,

$$\lim_{s\to\infty} q^{(s)} = q^* \Leftrightarrow \forall \epsilon > 0, \exists s_0 \text{ s.t. } \forall s > s_0 \Rightarrow |q^{(s)} - q^*| < \epsilon/2.$$

Here, we define

$$\left|\sum_{s=1}^{s_0}(q^{(s)} - q^*)\right| = M > 0,$$

and thus,

$$\lim_{s\to\infty}\frac{M}{s} = 0 \Leftrightarrow \forall \epsilon > 0, \exists s_0' \text{ s.t. } \forall s > s_0' \Rightarrow \frac{M}{s} < \epsilon/2.$$

When $S_0 = \max\{s_0, s_0'\}$, we have

$$\begin{aligned}\forall S > S_0, |\bar{q}^{(S)} - q^*| &= \left|\sum_{s=1}^{S} \frac{1}{S}(q^{(s)} - q^*)\right| \\ &< \frac{M}{S} + \sum_{s=S_0+1}^{S} \left|\frac{1}{S}(q^{(s)} - q^*)\right| \\ &\leq \epsilon/2 + \left|\frac{S - S_0}{S}\right| \epsilon/2 \leq \epsilon/2 + \epsilon/2 = \epsilon.\end{aligned}$$

Therefore,

$$\lim_{s\to\infty} \bar{q}^{(s)} = q^*.$$

□

References

[1] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed Membership Stochastic Blockmodels. *JMLR*, 9:1981–2014, 2008.

[2] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On Smoothing and Inference for Topic Models. In *Proc. UAI*, 2009.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[4] James Foulds, Christopher Dubois, Arthur U. Asuncion, Carter T. Butts, and Padhraic Smyth. A Dynamic Relational Infinite Feature Model for Longitudinal Social Networks. In *Proc. AISTATS*, pages 287–295, 2011.

[5] Wenjie Fu, Le Song, and Eric P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proc. ICML*, 2009.

[6] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua Tenenbaum. Dynamic Infinite Relational Model for Time-varying Relational Data Analysis. In *Proc. NIPS*, 2010.

[7] Katsuhiko Ishiguro, Issei Sato, and Naonori Ueda. Collapsed Variational Bayes Inference of Infinite Relational Model. *arXiv*, page arXiv:1409.4757 [cs.LG], September 2014.

[8] Katushiko Ishiguro, Naonori Ueda, and Hiroshi Sawada. Subset Infinite Relational Models. In *Proc. AISTATS*, pages 547–555, 2012.

[9] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning Systems of Concepts with an Infinite Relational Model. In *Proc. AAAI*, 2006.

[10] Bryan Klimt and Yiming Yang. The Enron Corpus : A New Dataset for Email Classification Research. In *Proc. ECML*, 2004.

[11] Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed Variational Dirichlet Process Mixture Models. In *Proc. IJCAI*, 2007.

[12] Konstantina Palla, David A Knowles, and Zoubin Ghahramani. An Infinite Latent Attribute Model for Network Data. In *Proc. ICML*, 2012.

[13] Issei Sato, Kenichi Kurihara, and Hiroshi Nakagawa. Practical collapsed variational bayes inference for hierarchical dirichlet process. In *Proc. KDD*, 2012.

[14] Issei Sato and Hiroshi Nakagawa. Rethinking Collapsed Variational Bayes Inference for LDA. In *Proc. ICML*, 2012.

[15] Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *Proc. KDD*, pages 677–685. ACM Press, 2008.

[16] Yee Whye Teh, David Newman, Max Welling, and D Neaman. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 19 (Proceedings of NIPS)*, 2007.