Approximate Mean Field for Dirichlet-Based Models

Arthur U. Asuncion ASUNCION@UCI.EDU
Department of Computer Science, University of California, Irvine

Abstract

Variational inference is an important class of approximate inference techniques that has been applied to many graphical models, including topic models. We propose to improve the efficiency of mean field inference for Dirichlet-based models by introducing an approximative framework that converts weighted geometric means in the updates into weighted arithmetic means. This paper also discusses a close resemblance between our approach and other methods, such as the factorized neighbors algorithm and belief propagation. Empirically, we find that our approach is accurate and efficient compared to standard mean field.

1. INTRODUCTION

Exact probabilistic inference is usually intractable for complicated graph-structured models. For such models, one must resort to approximate inference techniques such as Markov chain Monte Carlo or variational inference. In particular, mean field is a widely-used variational technique.

We propose a framework that improves the computational efficiency of mean field inference for models such as directed Bayesian networks with Dirichlet priors. Furthermore, we posit that our method can potentially be as accurate as standard mean field (or even more accurate), since it tries to counteract the error introduced by Jensen's inequality. While our approach is approximative and does not maintain the bound on the marginal loglikelihood, we find that our technique is very accurate for various models. Interestingly, the proposed approach has connections to loopy belief propagation (BP), the factorized neighbors algorithm (FNA), Gibbs sampling, and MAP estimation.

In the next section, we detail our AMF framework. We then show how AMF is related to FNA and loopy BP. We apply AMF to Dirichlet-based models and empirically find that AMF is both accurate and efficient.

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

2. APPROXIMATE MEAN FIELD (AMF)

Mean field (MF) has been applied to many different models (Opper & Saad, 2001; Blei et al., 2003; Beal & Ghahramani, 2006). In MF, a factorized variational distribution $q(x)$ over hidden variables is introduced, and the goal is to minimize KL divergence $\mathrm{KL}[q(x)||p(x|y)]$. To make this method explicit, we derive "free energies" by starting at the negative marginal loglikelihood ("energy"), denoting observed variables as y and hidden variables as x, and keeping implicit the dependence on parameters θ,

$$-\log p(y) = -\log \sum_x q(x) \frac{p(x,y)}{q(x)}$$

$$\leq -\sum_x q(x) \log p(x,y) - \mathcal{H}_x \quad (1)$$

$$\mathcal{F}_1 = -\sum_x \prod_i q_i(x_i) \log p(x|y) - \mathcal{H}_x + c_1 \quad (2)$$

$$\mathcal{F}_2 = -\sum_x \prod_i q_i(x_i) \log p(x_i|x_{N(i)}, y) - \mathcal{H}_x + c_2. \quad (3)$$

In Eq (1), Jensen's inequality is used to break up the "average" energy (1st term) from the "entropy" (2nd term), where entropy is $H_x = -\sum_x q(x) \log q(x)$. It is the application of Jensen's inequality, in tandem with the use of a factorized variational posterior, where error creeps into MF. In Eq (2), we arrive at the free energy $\mathcal{F}_1$ by introducing a fully factorized $q(x) = \prod_i q_i(x_i)$, and we also break off the "$\log p(y|\theta)$" term and move it into constant c_1. Then $\mathcal{F}_2$ is created in Eq (3) by breaking off "$\log p(x_{N(i)}|y,\theta)$" and moving it into c_2, which is is a constant with respect to $q_i(x_i)$. Here the focus is on $q_i(x_i)$, since we will take a gradient with respect to $q_i(x_i)$ to obtain a variational update. Note that $N(i)$ denotes the neighborhood of x_i.

One can optimize $\mathcal{F}_2$ by taking the gradient, setting it equal to zero, and solving for $q_i(x_i)$[1], yielding the following MF update (which uses a weighted *geometric* mean),

$$q_i(x_i) \propto \exp\{E_{q(x_{N(i)})}\left[\log p(x_i|x_{N(i)}, y, \theta)\right]\}$$

$$= \prod_{x_{N(i)}} \left[p(x_i|x_{N(i)}, y, \theta)\right]^{q(x_{N(i)})}. \quad (4)$$

[1]We assume that the variational distribution $q_i(x_i = k)$ is discrete, so there is a different update for each k.

Now we describe our method. Starting from $\mathcal{F}_2$, we move the $\log$ out of $N-1$ expectations in the average energy (where N is the number of variables, and where the exception is the expectation over $q_i(x_i)$). By Jensen's inequality, this transformation provides a lower bound $\tilde{\mathcal{F}}_2^i \leq \mathcal{F}_2$. Recall that $\mathcal{F}_2$ is itself an upper bound on the true energy. This modified "free energy" $\tilde{\mathcal{F}}_2^i$ is specific to index i,

$$\tilde{\mathcal{F}}_2^i = -\sum_{x_i} q_i(x_i) \log E_{q(x_{N(i)})}\left[p(x_i|x_{N(i)}, y, \theta)\right] - \mathcal{H}_x.$$

Taking the gradient of $\tilde{\mathcal{F}}_2^i$ and solving for $q_i(x_i)$ yields an update with a weighted *arithmetic* mean,

$$q_i(x_i) \propto E_{q(x_{N(i)})}\left[p(x_i|x_{N(i)}, y, \theta)\right]. \tag{5}$$

Thus, the AMF transformation, which moves $\log$ functions out of expectations, produces a weighted arithmetic mean in the update rather than the weighted geometric mean that was present in the MF update in Eq (4). For topic models with Dirichlet-multinomial conjugacy, a weighted arithmetic mean is more efficient to compute. Furthermore, we argue that moving the logarithm back out of the expectations can potentially "reverse" the error of Jensen's inequality incurred in Eq (1), since the AMF transformation is applying Jensen's inequality in the opposite direction from the original application of Jensen's inequality in Eq (1).

AMF can also be used to separate terms within $p(x_i|x_{N(i)}, y, \theta)$, prior to moving the $\log$ out of the expectations – this technique is further discussed in Section 5.

3. CONNECTION TO FNA

The Dobrushin-Lanford-Ruelle (DLR) equations represent all possible marginalizations of a distribution,

$$P(x_R) = \sum_{x_{N(R)}} p(x_R|x_{N(R)})P(x_{N(R)}) \qquad \forall R \in \Lambda,$$

where Λ is the power set over variables, R is an arbitrary subset, and $N(R)$ is the neighborhood of R. This system of equations ensures that the joint, conditionals, and marginals are consistent with each other, and solving these equations is tantamount to performing exact inference.

Consider a "reduced" system of DLR equations, where now $\Lambda_1 = \{i\}$ (singleton variables). Let $b_i(x_i) \equiv P(x_i)$. The reduced DLR equations are the following,

$$b_i(x_i) = \sum_{x_{N(i)}} p(x_i|x_{N(i)}) \prod_{j \in N(i)} b_j(x_j) \quad \forall i \in \Lambda_1. \tag{6}$$

The factorized neighbors algorithm (FNA) inspired by these reduced DLR equations (Rosen-Zvi et al., 2005) uses the following update for each x_i,

$$b_i(x_i) \leftarrow \sum_{x_{N(i)}} p(x_i|x_{N(i)}) \prod_{j \in N(i)} b_j(x_j). \tag{7}$$

Surprisingly, the FNA updates are precisely the updates in Eq (5), obtained by performing our AMF transformations. The fixed points of these updates are solutions to the reduced system of equations in Eq (6). FNA has been shown to empirically outperform MF on spin-glass models in terms of accuracy (Rosen-Zvi et al., 2005). The fact that FNA is a special case of our AMF framework suggests that our framework can produce accurate algorithms.

4. CONNECTION TO LOOPY BP

The AMF transformations can produce a modified version of loopy belief propagation (BP). We apply AMF to $\mathcal{F}_1$ in Eq (2), and introduce factors $p(x|y,\theta) = \prod_\alpha \Psi_\alpha(x_\alpha)$,

$$\mathcal{F}_1 = -\sum_\alpha \sum_{x_\alpha} \prod_{i \in N(\alpha)} q_i(x_i) \log \Psi_\alpha(x_\alpha) - \mathcal{H}_x.$$

We move the $\log$ out of $N-1$ expectations in the average energy as before, producing a lower bound $\tilde{\mathcal{F}}_1^i \leq \mathcal{F}_1$. Note that $\tilde{\mathcal{F}}_1^i$ focuses on terms specific to variable i,

$$\tilde{\mathcal{F}}_1^i = -\sum_{\alpha \in N(i)} \sum_{x_i} q_i(x_i) \log E_{q(x_\alpha^{\neg i})}[\Psi_\alpha(x_\alpha)] - \mathcal{H}_x.$$

Taking the gradient and solving for $q_i(x_i)$ produces an update which features weighted *arithmetic* means,

$$q_i(x_i) \propto \prod_{\alpha \in N(i)} \sum_{x_\alpha \setminus x_i} \left[\prod_{j \in N(\alpha) \setminus i} q_j(x_j)\right] \Psi_\alpha(x_\alpha). \tag{8}$$

Consider the standard BP updates for a factor graph,

$$m_{i\alpha}(x_i) \leftarrow \prod_{\beta \in N(i) \setminus \alpha} m_{\beta i}(x_i) \tag{9}$$

$$m_{\alpha i}(x_i) \leftarrow \sum_{x_\alpha \setminus x_i} \Psi_\alpha(x_\alpha) \prod_{j \in N_\alpha \setminus i} m_{j\alpha}(x_j) \tag{10}$$

$$q_i(x_i) \leftarrow \prod_{\beta \in N(i)} m_{\beta i}(x_i). \tag{11}$$

Now consider a slightly modified message update between variable i and factor α that includes the back-message $m_{\alpha i}(x_i)$ and thus becomes the belief $q_i(x_i)$:

$$m_{i\alpha}(x_i) \leftarrow \prod_{\beta \in N(i)} m_{\beta i}(x_i) \equiv q_i(x_i). \tag{12}$$

Surprisingly, by substituting this modified message into the update for $m_{\alpha i}(x_i)$ in Eq (10) and then substituting $m_{\alpha i}(x_i)$ into the expression for $q_i(x_i)$ in Eq (11), the exact AMF update in Eq (8) is obtained. Thus, BP and AMF are closely connected, with the difference being the inclusion of back-messages from the factor to the node. Typically, for BP, one strives to *prevent* this back-flow, since

back-messages can cause the algorithm to destabilize and collapse to a mode which may be undesirable in certain situations. A well-known property of MF is its ability to break symmetry and gravitate towards a mode (Jaakkola, 2000), since MF minimizes $\mathrm{KL}(q||p)$; thus, it is unsurprising that AMF corresponds to BP with back-messages. In some models, like topic models (with multiple symmetric modes), this symmetry-breaking property is desirable.

5. APPLICATION TO VARIOUS MODELS

In this section, we apply the AMF transformations to topic models and hidden Markov models, which lead to efficient algorithms. As we develop these algorithms, we also highlight connections to Gibbs sampling and MAP estimation.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003), known as a "topic model," is widely used in machine learning. The generative process is summarized below:

$$\theta_{k,j} \sim \mathcal{D}[\alpha] \quad \phi_{w,k} \sim \mathcal{D}[\beta] \quad x_{ij} \sim \theta_{k,j} \quad y_{ij} \sim \phi_{w,x_{ij}}.$$

Collapsed Gibbs sampling (in which ϕ and θ are integrated out) is a common method for LDA (Griffiths & Steyvers, 2004), where the conditional is used to sample each x_{ij},

$$p(x_{ij} = k|x_{\neg ij}, y) \propto (N_{kj}^{\neg ij} + \alpha)\frac{(N_{wk}^{\neg ij} + \beta)}{(N_{k}^{\neg ij} + W\beta)} \quad (13)$$

where $N_{wkj} = \sum_i \mathbb{I}[y_{ij} = w, x_{ij} = k]$, with the convention that missing indices are summed out, and $\neg ij$ denotes that the corresponding word is excluded from the counts.

The conditional in Eq (13) can be used to create an AMF algorithm for LDA. We begin at the free energy $\mathcal{F}_2$ in Eq (3):

$$\begin{aligned}
\mathcal{F}_2 &= -\sum_x \prod_{ij} q_{ij}(x_{ij}) \log p(x_{ij}|x_{\neg ij}, y) - \mathcal{H}_x \\
&= -\sum_x \prod_{ij} q_{ij}(x_{ij}) \log \left[(N_{kj}^{\neg ij} + \alpha)\frac{(N_{wk}^{\neg ij} + \beta)}{(N_{k}^{\neg ij} + W\beta)} \right] - \mathcal{H}_x \\
&= -\sum_x \prod_{ij} q_{ij}(x_{ij}) \log(N_{kj}^{\neg ij} + \alpha) \quad (14) \\
&\quad -\sum_x \prod_{ij} q_{ij}(x_{ij}) \log(N_{wk}^{\neg ij} + \beta) \\
&\quad +\sum_x \prod_{ij} q_{ij}(x_{ij}) \log(N_{k}^{\neg ij} + W\beta) - \mathcal{H}_x \\
&\approx -\sum_{x_{ij}} q_{ij}(x_{ij}) \log E_{q(x^{\neg ij})}[N_{kj}^{\neg ij} + \alpha] \quad (15) \\
&\quad -\sum_{x_{ij}} q_{ij}(x_{ij}) \log E_{q(x^{\neg ij})}[N_{wk}^{\neg ij} + \beta] \\
&\quad +\sum_{x_{ij}} q_{ij}(x_{ij}) \log E_{q(x^{\neg ij})}[N_{k}^{\neg ij} + W\beta] - \mathcal{H}_x
\end{aligned}$$

In Eq (14), the log breaks up the conditional into three terms. AMF transformations are applied in Eq (15), where the log is moved out of $N-1$ expectations for each of the three terms. Since one of the terms is positive while the other two are negative, the "reversal of error" discussed in Section 2 is mitigated when applying Jensen's inequality on the positive term[2]. Nonetheless, an efficient update is obtained when using the approximate free energy in Eq (15),

$$q_{ij}(x_{ij} = k) \propto E_{q(x^{\neg ij})}[N_{kj}^{\neg ij} + \alpha]\frac{E_{q(x^{\neg ij})}[N_{wk}^{\neg ij} + \beta]}{E_{q(x^{\neg ij})}[N_{k}^{\neg ij} + W\beta]} \quad (16)$$

Incidentally, this AMF update for LDA is equivalent to the "CVB0" algorithm proposed by Asuncion et al. (2009), which was originally derived as an approximation to the update in Eq (18). For comparison, consider the MF update which can be derived from the free energy in Eq (14),

$$\begin{aligned}
q_{ij}(x_{ij} = k) &\propto \exp\{E_{q(x^{\neg ij})}[\log(N_{kj}^{\neg ij} + \alpha)]\} * \quad (17) \\
&\left(\frac{\exp\{E_{q(x^{\neg ij})}[\log(N_{wk}^{\neg ij} + \beta)]\}}{\exp\{E_{q(x^{\neg ij})}[\log(N_{k}^{\neg ij} + W\beta)]\}} \right)
\end{aligned}$$

This MF update consists of weighted geometric means and is costly to compute. To address this problem, Teh et al. (2007) proposed a Gaussian approximation of MF which uses a second-order Taylor expansion to obtain the update,

$$\begin{aligned}
q_{ij}(x_{ij} = k) &\propto E_{q(x^{\neg ij})}[N_{kj}^{\neg ij} + \alpha]\frac{E_{q(x^{\neg ij})}[N_{wk}^{\neg ij} + \beta]}{E_{q(x^{\neg ij})}[N_{k}^{\neg ij} + W\beta]} \\
&\exp\left(-\frac{Var[N_{kj}^{\neg ij}]}{2(E_{q^{\neg ij}}[N_{kj}^{\neg ij} + \alpha])^2} - \frac{Var[N_{wk}^{\neg ij}]}{2(E_{q^{\neg ij}}[N_{wk}^{\neg ij} + \beta])^2} \right. \\
&\left. + \frac{Var[N_{k}^{\neg ij}]}{2(E_{q^{\neg ij}}[N_{k}^{\neg ij} + W\beta])^2} \right). \quad (18)
\end{aligned}$$

This update contains several variance terms. Eq (18) is similar to the AMF update in Eq (16), with the difference being the presence of these second-order terms which significantly add to the overhead in comparison to AMF. For LDA, Asuncion et al. (2009) compared MF in Eq (18) with AMF ("CVB0") and found that while both methods achieve the same accuracy, AMF is more than twice as fast than MF, suggesting that AMF should be preferred over MF.

AMF is closely connected to other methods. AMF in Eq (16) is similar to the Gibbs sampler in Eq (13), with the main difference being that the AMF update is deterministic and uses expected counts. Asuncion et al. (2009) also draw links between AMF and MAP estimation; furthermore, Beal & Ghahramani (2006) study MAP estimation with a softmax basis, which is very similar to AMF (with the difference being the exclusion of counts, i.e. $\neg ij$).

[2] It is possible to only apply our AMF transformations to certain terms; however, we are generally interested in efficiency and usually it is beneficial to apply the transformations to all the terms.

Our AMF approach can be applied more generally to other models. As a testbed, consider the HMM (Rabiner, 1990); in particular, consider an HMM with time-varying transition parameters. In this model, the observed sequences, $\{y_i\}, 1 \leq i \leq N$, are each of length T, where each y_{it} is discrete, taking one of M values. Each observed sequence has a hidden sequence x_i, and each x_{it} takes one of S state values. The transition matrices (of size $S \times S$) are denoted by θ_t (with initial distribution being a vector θ_0), and the emission probabilities (of size $S \times M$) are denoted by ϕ:

$$\theta_0[\cdot] \sim \mathcal{D}[\alpha], \quad \theta_t[\cdot|s] \sim \mathcal{D}[\alpha], \quad \phi[\cdot|s] \sim \mathcal{D}[\beta]$$
$$x_{i,1} \sim \theta_0[\cdot], \quad x_{i,t} \sim \theta_{t-1}[\cdot|x_{i,t-1}], \quad y_{i,t} \sim \phi[\cdot|x_{i,t}].$$

One way to perform direct collapsed Gibbs sampling over the hidden variables (with θ and ϕ integrated out) is to make use the following conditional distribution (Scott, 2002),

$$\begin{aligned} &p(x_{it} = k|x_{\neg it}, y) \qquad (19)\\ &\propto p(y_{it}|x_{it} = k)p(x_{it} = k|x_{i,t-1})p(x_{i,t+1}|x_{it} = k) \\ &\propto \left(\frac{N^{\neg it}_{k,y_{it}} + \beta}{N^{\neg it}_{k} + M\beta}\right)\left(N^{\neg it}_{t-1,x_{i,t-1},k} + \alpha\right)\left(\frac{N^{\neg it}_{t,k,x_{i,t+1}} + \alpha}{N^{\neg it}_{t,k} + S\alpha}\right) \end{aligned}$$

where $N_{k,m} = \sum_{i,t} \mathbb{I}[x_{it} = k, y_{it} = m]$ and $N_{t,k,l} = \sum_i \mathbb{I}[x_{it} = k, x_{i,t+1} = l]$.

On can develop MF for this HMM, but the updates would be inefficient (similar to Eq (17)). Moreover, a Gaussian version of MF can be developed as in Eq (18). AMF would follow the same type of derivation as in Eq (15), yielding,

$$\begin{aligned} q_{it}(x_{it} = k) \propto &\left(\frac{E_{q(x^{\neg it})}[N^{\neg it}_{k,y_{it}} + \beta]}{E_{q(x^{\neg it})}[N^{\neg it}_{k} + M\beta]}\right) * \qquad (20)\\ &\left(E_{q(x^{\neg it})}[N^{\neg it}_{t-1,x_{i,t-1},k} + \alpha]\right)\left(\frac{E_{q(x^{\neg it})}[N^{\neg it}_{t,k,x_{i,t+1}} + \alpha]}{E_{q(x^{\neg it})}[N^{\neg it}_{t,k} + S\alpha]}\right) \end{aligned}$$

6. EXPERIMENTS

We perform experiments on topic models and HMMs, and we find that AMF is accurate and efficient.

Asuncion et al. (2009) showed that AMF ("CVB0") is computationally faster than MF for LDA. Here we tried the AMF approach on Hierarchical Dirichlet Processes (HDP). We obtained code for collapsed MF for HDP (which uses a Gaussian approximation similar to Eq (18)) from Teh et al. (2008), and we removed the second-order terms to obtain AMF. In Figure 1(a), we conducted an experiment on Reuters data, and we find that performing our technique does not negatively affect the test loglikelihood. In fact, the accuracy of AMF appears to better than MF. Note that this improvement may be an artifact of possibly suboptimal hyperparameters (e.g. see Asuncion et al. (2009)). Nevertheless, we can be reasonably confident that AMF is learning an accurate solution on this nonparametric topic model.

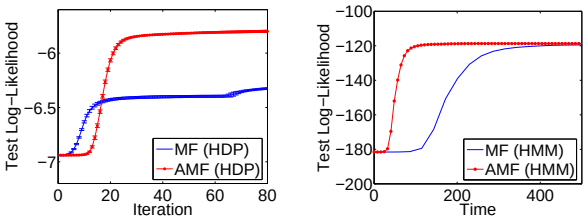

Figure 1. Test loglikelihoods achieved by AMF and MF on (a) HDP, (b) HMM. AMF is accurate and efficient compared to MF.

We also performed an experiment on the HMM in Section 5, with $T = 100$ slices, $M = 4$ hidden states, and $R = 9$ observed states, and we set $\alpha = 0.1$ and $\beta = 0.1$. We simulated ground truth θ_{true} and ϕ_{true} from $\mathcal{D}[0.1]$, and from these distributions, we simulated synthetic training data y_{train} with 1,000 sequences and test data y_{test} with 200 sequences. We ran both AMF and MF (with a Gaussian approximation) for 200 iterations on the training data.

Figure 1(b) shows test loglikelihoods achieved by MF and AMF as a function of time. While both algorithms achieve the same accuracy, AMF is significantly faster than MF due to AMF's efficient arithmetic updates. MF takes 401 seconds to reach a loglikelihood of -120, while AMF takes 105 seconds to reach the same loglikelihood; thus, our AMF transformations provide a substantial 4x speedup over the Gaussian version of MF for this model. Had we used standard MF (which is even less efficient than the Gaussian version of MF), this speedup would be even higher.

7. CONCLUSIONS

We have introduced an approximative framework that uses weighted arithmetic means in the MF updates. Furthermore, we have uncovered connections between AMF and techniques such as FNA, loopy BP, Gibbs sampling, and MAP estimation. Our experimental results on Dirichlet-based models suggest that AMF is as accurate as standard MF, while being more efficient than MF.

Acknowledgements

Thanks to Padhraic Smyth, Max Welling, Alex Ihler, and Drew Frank for useful discussions. This work was supported by an NSF graduate fellowship.

References

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. On smoothing and inference for topic models. In *UAI*, 2009.

Beal, M.J. and Ghahramani, Z. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 2006.

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet allocation. *JMLR*, 2003.

Griffiths, T.L. and Steyvers, M. Finding scientific topics. *PNAS*, 2004.

Jaakkola, T.S. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*. MIT Press, 2000.

Opper, M. and Saad, D. *Advanced mean field methods*. MIT Press, 2001.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in speech recognition*, 1990.

Rosen-Zvi, M., Jordan, M.I., and Yuille, A.L. The DLR hierarchy of approximate inference. In *UAI*, 2005.

Scott, S.L. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *JASA*, 2002.

Teh, Y.W., Newman, D., and Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS 19*. 2007.

Teh, Y.W., Kurihara, K., and Welling, M. Collapsed variational inference for HDP. In *NIPS 20*. 2008.