An Auxiliary Variational Method

Felix V. Agakov[1] and David Barber[2]

[1] University of Edinburgh, 5 Forrest Hill, EH1 2QL Edinburgh, UK
felixa@inf.ed.ac.uk, http://anc.ed.ac.uk
[2] IDIAP, Rue du Simplon 4, CH-1920 Martigny Switzerland,
david.barber@idiap.ch

Abstract. An attractive feature of variational methods used in the context of approximate inference in undirected graphical models is a rigorous lower bound on the normalization constants. Here we explore the idea of using augmented variable spaces to improve on the standard mean-field bounds. Our approach forms a more powerful class of approximations than any structured mean field technique. Moreover, the existing variational mixture models may be seen as computationally expensive special cases of our method. A byproduct of our work is an efficient way to calculate a set of mixture coefficients for any set of tractable distributions that principally improves on a flat combination.

1 Introduction

Probabilistic treatment of uncertainty provides a principled way of reasoning in stochastic domains. Unfortunately, mathematical consistency often comes at a price of inherent intractability of many interesting models, such as Boltzmann machines

$$p(\mathsf{x}) = \exp\{-E(\mathsf{x})\}/Z, \quad Z = \sum_{\mathsf{x}} \exp\{-E(\mathsf{x})\}. \tag{1}$$

In general the complexity of evaluating the partition function Z is exponential in the size of the largest clique in the associated junction tree. For dense models the exact evaluations are in general computationally infeasible, and approximations need to be considered. In this paper we focus on computation of lower bounds on Z, which may also be used to approximate formally intractable marginals.

Variational approximations have been widely used in physics and engineering and more recently applied to graphical modeling (e.g. [5]). In this context they are typically used to obtain rigorous (but relatively simple) bounds on the normalizing constant. A popular class of such methods is based on the KL-divergence

$$KL(q(\mathsf{x})\|p(\mathsf{x})) = \langle \log q(\mathsf{x}) \rangle_{q(\mathsf{x})} - \langle \log p(\mathsf{x}) \rangle_{q(\mathsf{x})} \geq 0, \tag{2}$$

where $\langle \ldots \rangle_{q(\mathsf{x})}$ denotes an average over $q(\mathsf{x})$, and the bound is saturated if and only if $q(\mathsf{x}) \equiv p(\mathsf{x})$. In the case of the Boltzmann distribution (1), non-negativity of (2) yields the well-known class of lower bounds

$$\log Z \geq -\langle \log q(\mathsf{x}) \rangle_{q(\mathsf{x})} - \langle E(\mathsf{x}) \rangle_{q(\mathsf{x})}, \tag{3}$$

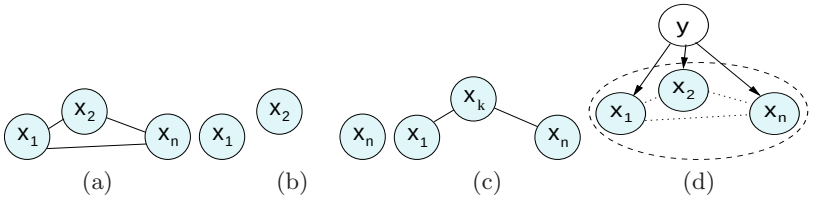


Fig. 1. (a) A fully connected network representing the intractable $p(\mathsf{x})$; (b) standard mean field model $q_{MF}(\mathsf{x})$; (c) structured mean field model $q_{SMF}(\mathsf{x})$; (d) a mixture of mean field models [all the variables x are coupled through the mixture label y.]

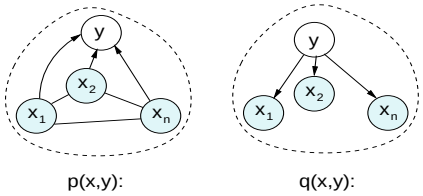


Fig. 2. An auxiliary MF model. The target $p(\mathsf{x}, y)$ is approximated by $q(\mathsf{x}, y)$, which is structured *in the augmented space*. [Note that the marginal $p(\mathsf{x})$ expressed from $p(\mathsf{x}, y)$ is identical to the original fully connected pairwise distribution shown on Figure 1 (a).

where $q(\mathsf{x})$ is typically restricted to lie in a tractable family and varied to obtain the tightest bound within the family. Coupled with an upper bound on Z, expression (3) may be used for bounding marginals of $p(\mathsf{x})$. This procedure may also be used to optimize a lower bound on the marginal likelihood in partially observable models, which is a natural generalization of the EM algorithm [7].

1.1 Existing Variational Approximations

The tractability of the bound (3) depends on the choice of the approximating distribution $q(\mathsf{x})$, which is in the simplest case given by the factorized *mean field* (MF) model $q_{MF}(\mathsf{x}) = \prod_i q(x_i)$ (see Fig. 1 (a), (b)). Factorized approximations may be simple, but inaccurate when $p(\mathsf{x})$ is strongly coupled; moreover, due to uni-modality they may miss a significant mass contributing to Z. One way to go beyond the factorized assumption for $q(\mathsf{x})$ is to consider a *structured mean field* approximation, which also introduces conditional independencies, but retains some of the structure of $p(\mathsf{x})$. Often it is assumed that $q(\mathsf{x})$ has a sparse graphical representation (e.g. it is a (poly)tree, see Figure 1 (c)), which typically leads to an improvement on the bound at a moderate increase in computational cost. Another approach examined recently [4], [6] uses mixtures of mean field type models (see Figure 1 (d)). This is a powerful extension of factorized approximations, since the resulting $q(\mathsf{x})$ is in general multi-modal and coupled in x (Fig. 1); however, in this case the bound (3) is itself intractable. Known techniques

[6], [4], [3] handle the intractability by effectively using the Jensen's bound on top of (3), which is computationally costly and numerically unstable in practice unless *all* the mixture components $q(\mathsf{x}|\mathsf{y})$ have the same structure.

2 Auxiliary Variational Method

Intuitively, evaluation of the bound (3) on $\log Z$ for variational mixture approximations requires minimization of the KL-divergence between two fully-connected distributions (see (2)). However, computationally it could be useful to retain a sparser structural form of $q(\mathsf{x},\mathsf{y})$ and use it as an approximation. To do this, we introduce *auxiliary variables* y to the target distribution in such a way that the marginal $\tilde{p}(\mathsf{x})$ of the augmented model $p(\mathsf{x},\mathsf{y})$ has the same graphical structure as the original target $p(\mathsf{x})$ (see Fig. 2). Then we minimize the KL-divergence between $q(\mathsf{x},\mathsf{y})$ and $p(\mathsf{x},\mathsf{y})$ in the joint variable spaces. This case is different from standard structured approximations, as all the variables x of the marginal $\langle q(\mathsf{x}|\mathsf{y})\rangle_{q(\mathsf{y})}$ remain fully connected. However, similarly to structured mean field methods, the approximation $q(\mathsf{x},\mathsf{y})$ in the joint space is constrained to be sparse.

Another motivation for this work is the reported success of auxiliary sampling techniques, such as *Hybrid Monte-Carlo* or the *Swendsen-Wang* [8] algorithms. It has been shown that by augmenting the original variable space with auxiliary variables and sampling from joint distributions in the augmented spaces, one can achieve a significant improvement over standard MCMC approaches. The purpose of the auxiliary variables in this context is to capture (structural) information about clusters of correlated variables. It may therefore be hoped that an *auxiliary variational* method performing approximations in the augmented space $\{\mathsf{x},\mathsf{y}\}$ may improve on simple approximations in $\{\mathsf{x}\}$.

2.1 Optimizing the Auxiliary Variational Bound

Let $p(\mathsf{x},\mathsf{y}) = p(\mathsf{x})p(\mathsf{y}|\mathsf{x})$ define the joint distribution of the original variables x and auxiliary variables y in the augmented $\{\mathsf{x},\mathsf{y}\}$ space. From the divergence $KL(q(\mathsf{x},\mathsf{y})\|p(\mathsf{x},\mathsf{y}))$ in the joint space it is easy to obtain an expression for the lower bound on the normalizing constant of (1), which is given by

$$\log Z \geq \sum_{\mathsf{y}} q(\mathsf{y}) \left[\langle -E(\mathsf{x}) - \log q(\mathsf{x}|\mathsf{y})\rangle_{q(\mathsf{x}|\mathsf{y})}\right] + \tilde{I}, \quad \tilde{I} = \sum_{\mathsf{x}}\sum_{\mathsf{y}} q(\mathsf{x},\mathsf{y}) \log \frac{p(\mathsf{y}|\mathsf{x})}{q(\mathsf{y})} \quad (4)$$

where $p(\mathsf{y}|\mathsf{x})$ is an arbitrary *auxiliary conditional* distribution. Clearly, (4) decomposes as a convex sum of the standard lower bounds with approximations $q(\mathsf{x}|\mathsf{y})$ and a lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ on the mutual information. This may be used to improve on a single best (tractable) approximation $q(\mathsf{x}|\mathsf{y})$, which is reconstructed by trivially setting $p(\mathsf{y}|\mathsf{x}) \equiv p(\mathsf{y})$. Note that (4) is tractable as long as $p(\mathsf{y}|\mathsf{x})$, $q(\mathsf{y})$, and $q(\mathsf{x}|\mathsf{y})$ are constrained to lie in tractable families – there is no need to use further variational relaxations in this case. One tractable choice for the auxiliary mapping $p(\mathsf{y}|\mathsf{x})$ is a Gaussian (in this case $q(\mathsf{x}|\mathsf{y})$ should also be

parameterized as y is real-valued). Another case leading to exact computations is obtained when each node y_i in $p(\mathsf{y}|\mathsf{x})$ has a small number of x-parents, and $q(\mathsf{x},\mathsf{y})$ is a tree. Some other parameterizations may not lead to exact bounds, but may nevertheless result in efficient and practically useful approximations (see [1] for discussions and derivations of the EM algorithms for some of these cases).

2.2 Specific Auxiliary Representations

Here we briefly look at two useful choices of $p(\mathsf{y}|\mathsf{x})$ for pairwise Markov networks with $E(\mathsf{x}) \stackrel{\text{def}}{=} \mathsf{x}^T\mathsf{W}\mathsf{x}$. As usual, we assume that $q(\mathsf{x},\mathsf{y})$ is tractable, i.e. the only problematic term in (4) is the auxiliary expectation $\langle \log p(\mathsf{y}|\mathsf{x})\rangle_{q(\mathsf{x},\mathsf{y})}$.

Parametric Constraints on the Auxiliary Distributions

If the auxiliary space is given by a single multinomial variable $y \in \{1,\ldots,M\}$, a natural choice for $p(y|\mathsf{x})$ is to use a *softmax* type representation

$$p(y_k|\mathsf{x}) \propto \exp\left\{ f(\mathsf{x}^T\mathsf{u}^{(k)} + b^{(k)})\right\}, \quad \mathsf{U} = \{\mathsf{u}^{(1)},\ldots,\mathsf{u}^{(M)}\} \in \mathbb{R}^{|\mathsf{x}|\times M}, \mathsf{b} \in \mathbb{R}^M \quad (5)$$

where $f(\mathsf{x};\mathsf{U},\mathsf{b})$ is some differentiable function and $p(y_k|\mathsf{x})$ is the probability of the auxiliary variable y being in state k. Unfortunately, if the weight vector $\mathsf{u}^{(k)}$ is dense, one may need to bound $\langle \log p(y_k|\mathsf{x})\rangle_{q(\mathsf{x},y)}$ (a cheaper alternative is to approximate such terms as $\log p(y_k|\langle\mathsf{x}\rangle_{q(\mathsf{x})})$ or use a multivariate factorial representation of the auxiliary variables $p(\mathsf{y}|\mathsf{x}) = \prod_i p(y_i|\mathsf{x})$ with the *Gaussian field* [2] approximation of $\langle \log p(y_i|\mathsf{x})\rangle_{q(\mathsf{x}|\mathsf{y})}$ – see [1] for details). For dense weights such approximations usually do not lead to significant deviations of the objective and are shown to be both accurate and efficient [1]; however, the optimized function is no longer a strict bound on $\log Z$. We now consider a tractable case when $p(\mathsf{y}|\mathsf{x})$ has constraints on the parental structure for each auxiliary variable.

Structural Constraints on Auxiliary Distributions

If $\boldsymbol{\pi}_x(y_i)$ and $\boldsymbol{\pi}_y(y_i)$ are x- and y-parents of y_i in the mapping $p(\mathsf{y}|\mathsf{x})$, we get

$$\langle \log p(\mathsf{y}|\mathsf{x})\rangle_{q(\mathsf{x},\mathsf{y})} = \sum_{i=1,\ldots|\mathsf{y}|} \langle \log p(y_i|\boldsymbol{\pi}_x(y_i),\boldsymbol{\pi}_y(y_i))\rangle_{q(y_i,\boldsymbol{\pi}_y(y_i),\boldsymbol{\pi}_x(y_i))}\,. \quad (6)$$

The representational complexity of each conditional in (6) in this case is of the order of $s^{|\boldsymbol{\pi}_x(y_i)|+|\boldsymbol{\pi}_y(y_i)|}$, where s is the number of states (for simplicity assumed to be equal for each variable). Since we are free to choose the form of the distribution $p(\mathsf{y}|\mathsf{x})$, we can limit its parental structure so that $|\boldsymbol{\pi}_x(y_i)| + |\boldsymbol{\pi}_y(y_i)|$ is small. Clearly, for discrete variables this allows an exact representation of the conditionals. It is also clear that the computational complexity of evaluating (6) is limited by the cost of marginalization of $|\mathsf{x}| + |\mathsf{y}| - |\boldsymbol{\pi}_x(y_i)| - |\boldsymbol{\pi}_y(y_i)| - 1$ variables from $q(\mathsf{x},\mathsf{y})$, which is tractable as long as $q(\mathsf{x},\mathsf{y})$ is in a tractable family. E.g., in the special case when $q(\mathsf{y},\mathsf{x}) = \prod_{l=1}^{|\mathsf{y}|} q(y_l) \prod_{j=1}^{|\mathsf{x}|} q(x_j|\boldsymbol{\rho}(x_j))$ is a polytree with a small number of y-parents $\boldsymbol{\rho}(x_j)$, the marginalization is exponential in $|\boldsymbol{\rho}(\boldsymbol{\pi}_x(y_i))\backslash\{\boldsymbol{\pi}_y(y_i)\cup y_i\}|$, which is acceptable if both $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$ are sparse.

3 Relation to Variational Mixture Models

The existing variational mixture approaches may be viewed as a special case of the *unconstrained* auxiliary formulation (where $p(\mathsf{y}|\mathsf{x}) \equiv q(\mathsf{y}|\mathsf{x})$). In this case (4) is intractable even for factorized mixtures with a few components (as $|\mathsf{x}|$ is large). This *requires* further factorized relaxations, such as the ones used by [4], [6], [3]. Also, unless all the components $q(\mathsf{x}|\mathsf{y})$ have identical structures, the optimization of the existing bounds may become numerically unstable (as it requires computing non-factorized summations of exponentially small terms [6]).

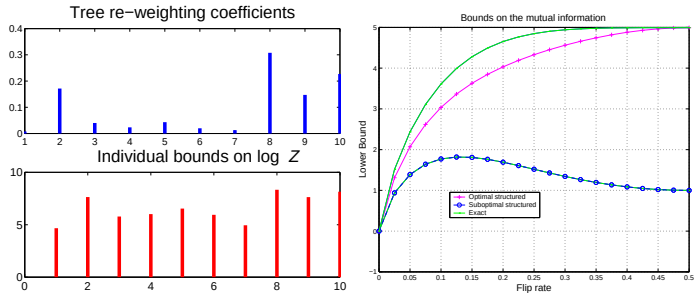
In our approach, we optimize the bound on $\log Z$ subject to constraints on the auxiliary conditional. By first constraining $p(\mathsf{y}|\mathsf{x})$ to be tractable and then optimizing the bound (4), we essentially incorporate the constrained Blahut-Arimoto algorithm into the variational inference framework. Arguably, our approach generalizes variational mixture approximations similarly to the way that the variational EM [7] generalizes the standard EM algorithm: by allowing a flexibility in the choice of $p(\mathsf{y}|\mathsf{x})$, it improves numerical stability and helps to significantly simplify the computations. Moreover, our method suggests a way to extend variational mixture approaches to structured auxiliary spaces, which may be useful for boosting the effective number of mixture states.

4 Experimental Results

Throughout the simulations, it was assumed that $p(\mathsf{x})$ is a pairwise Markov net with the energy $E(\mathsf{x}) = \mathsf{x}^T\mathsf{W}\mathsf{x} + \mathsf{x}^T\mathsf{b}$ and $\mathsf{x} \in \{-1, 1\}^{|\mathsf{x}|}$ (see [1] for details).

Reweighting Structured Representations: A by-product of our framework is a simple and fast way to re-weight any set of differently structured (tractable) distributions $q(\mathsf{x}|y)$, which principally improves on trivial combinations. For a uniformly chosen $\mathsf{W} \in \mathbb{I}^{10\times 10}$, we generated $K = 10$ spanning trees with the weights $\mathsf{W}^{(m)}$ such that $\mathsf{W}_{ij}^{(m)} = \mathsf{W}_{ij}$ for all i, j and $m = 1, \ldots, K$. Then we optimized (4) for $q(\mathsf{y})$ assuming that $q(\mathsf{x}|y_k)$ and $p(y_k|\mathsf{x})$ were fixed (see Fig. 3). In this case the bounds were $L_r \approx 8.38$, $L_u \approx 8.87$, $L_b \approx 8.33$, and $\mathbf{L_{av}} \approx \mathbf{9.52}$ for the random, uniform, best single, and auxiliary variational weightings respectively. As expected, the auxiliary method leads to the tightest bound.

Structured Auxiliary Mappings: To investigate an influence of a structure of $p(\mathsf{y}|\mathsf{x})$ on (4), we assumed a fixed $q(\mathsf{x}, \mathsf{y})$ with $\mathsf{x} \in \{-1, 1\}^{|\mathsf{x}|}$ and $\mathsf{y} \in \{-1, 1\}^{|\mathsf{y}|}$. The marginals $q(y_i)$ (flip rate) were fixed to be constant for all the nodes y_i. Fig. 3 (c) shows the improvement $\tilde{I}$ (over the convex combination in (4)) as a function of the flip rate for a model with $|\mathsf{x}| = 5$, $|\mathsf{y}| = 10$. The total number of x- and y-parents of each y_i was constrained to satisfy $|\boldsymbol{\pi}_x(y_i)| + |\boldsymbol{\pi}_y(y_i)| \leq 4$. We observed that some of the optimal auxiliary mappings were close to the theoretically optimal (but generally intractable) $q(\mathsf{x}|\mathsf{y})$, though the choice of the structure proved to be important (see [1] for details).



(a) (b)

Fig. 3. (a) *Top:* weights of the fixed structured approximations; *Bottom:* bounds on $\log Z$ for each approximation; (b) influence of the structure of $p(\mathsf{y}|\mathsf{x})$ on $\tilde{I}$.

5 Summary

We have presented an approach that generalizes the standard KL- variational procedure to the use of auxiliary variables, which provides a systematic improvement over standard structured approximations. We have also showed that the variational mixture approximations could be seen as special and more computationally expensive cases of our approach. Finally, we showed that our method can be easily generalized to factorial and structural state representations. One way to use it in practice is to find weightings for any set of tractable distributions.

References

1. Agakov, F. V. and Barber, D. (2004). An Auxiliary Variational Method. Technical report, EDI-INF-RR-0205, School of Informatics, University of Edinburgh.
2. Barber, D. and Sollich, P. (2000). Gaussian fields for approximate inference. In *Neural Information Processing Systems 12*. The MIT Press.
3. El-Hay, T. and Friedman, N. (2002). Incorporating Expressive Graphical Models in Variational Approximations: Chain-Graphs and Hidden Variables. In *UAI*.
4. Jaakkola, T. S. and Jordan, M. I. (1998). Improving the Mean Field Approximation via the Use of Mixture Distributions. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Publishers.
5. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An Introduction to Variational Methods for Graphical Models. In Jordan, M. I., editor, *Learning in Graphical Models*, chapter 1. Kluwer Academic Publishers.
6. Lawrence, N. D., Bishop, C. M., and Jordan, M. I. (1998). Mixture Representations for Inference and Learning in Boltzmann Machines. In *UAI: Proceedings of the 14th Conference*.
7. Neal, R. M. and Hinton, G. E. (1998). A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. In Jordan, M. I., editor, *Learning in Graphical Models*, chapter 1. Kluwer Academic Publishers.
8. Swendsen, R. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88.