IDIAP RESEARCH REPORT

An Auxiliary Variational Method

Felix Agakov [a] David Barber [b]

IDIAP–RR 04-86

June 2004

[a] University of Edinburgh, EH12QL, Edinburgh, UK
[b] IDIAP, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-86

An Auxiliary Variational Method

Felix Agakov David Barber

June 2004

Abstract. Variational methods have proved popular and effective for inference and learning in intractable graphical models. An attractive feature of the approaches based on the Kullback-Leibler divergence is a rigorous lower bound on the normalization constants in undirected models. In the suggested work we explore the idea of using auxiliary variables to improve on the lower bound of standard mean field methods. Our approach forms a more powerful class of approximations than any structured mean field technique. Furthermore, the existing lower bounds of the variational mixture models could 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 graphical models provide a convenient framework for the graphical representation of joint probability distributions via local constraints, and facilitate computation of many quantities of interest required for both inference and learning. Such a probabilistic treatment of uncertainty offers a consistent and principled approach to inference in complex domains. However, many distributions used to model practical domains are inherently intractable. For example, for distributions $p(\mathsf{x})$ of the (Boltzmann) form

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

the complexity of evaluating the partition function (normalization constant) Z is exponential in the size of the largest clique in the associated junction tree Lauritzen and Spiegelhalter (1988). For dense graphs the exact evaluations are in general computationally infeasible, and approximations need to be considered. In this paper we focus on approximate inference, specifically the computation of lower bounds on normalization constants, which can also be used to approximate marginals of a formally intractable distribution.

Variational approximations have been extensively used in physics and engineering and more recently applied to approximate inference and learning in intractable graphical models (see e.g. Jordan et al. (1998); Barber and Wiegerinck (1998); Wainwright et al. (2002)). In this context they were shown to result in relatively simple dual representations of the induced optimization problems. Their other advantage for graphical models is availability of rigorous bounds on the normalizing constant (Jaakkola and Jordan, 1996).

A popular class of such approximations is based on the Kullback-Leibler divergence

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

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

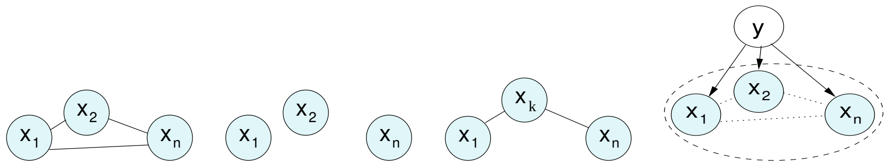
$$\log Z \geq \langle -\log q(\mathsf{x})\rangle_{q(\mathsf{x})} + \langle E(\mathsf{x})\rangle_{q(\mathsf{x})}, \tag{3}$$

where $q(\mathsf{x})$ is typically restricted to a class of tractable distributions and varied to obtain the tightest bound within the tractable class. Coupled with an upper bound on the normalizing constant (Wainwright et al., 2002), expression (3) may be used for bounding expectations in the original model. A further use for this procedure is to provide a lower bound on the marginal likelihood in situations of observed and unobserved variables, which is a natural derivation and extension of the EM procedure (Neal and Hinton, 1998).

1.1 Existing Variational Approximations

The computational tractability of the bound (3) depends on the choice of the approximating distribution $q(\mathsf{x})$. The simplest choice is given by the factorized *mean field* model (see Figure 1 (a), (b)) with $q_{MF}(\mathsf{x}) = \prod_i q(x_i)$, which discards all the edges from the original graph. This results in the simplest form of the bound (3). However, the bound may be inaccurate when the variables in the true distribution are strongly correlated. Moreover, factorized approximations of distributions of the exponential family are unimodal. This implies a fundamental limitation of the mean field approximation in the case when $p(\mathsf{x})$ is multi-modal, since significant mass contributing to the partition function may be missed.

One way to go beyond the factorized assumption for $q(\mathsf{x})$ is to consider a *structured mean field* approximation (Ghahramani and Jordan, 1995; Barber and Wiegerinck, 1998) which 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) (b) (c) (d)

Figure 1: (a) A fully connected pairwise Markov network representing the intractable distribution $p(\mathsf{x})$; (b) standard mean field approximation $q_{MF}(\mathsf{x})$; (c) structured mean field approximation $q_{SMF}(\mathsf{x})$; (d) mixture of mean field models. [All the variables x are coupled through the mixture label y. The dotted lines serve to indicate that the marginal $q_{MMF}(\mathsf{x})$ is in general fully connected.]

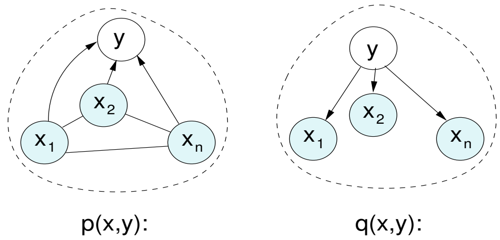


Figure 2: A special case of the auxiliary variational framework – an auxiliary mean field approximation. Here the target distribution $p(\mathsf{x}, y)$ is approximated by $q(\mathsf{x}, y)$, which is structured *in the augmented space*. Note that the marginal $p(\mathsf{x})$ expressed from $p(\mathsf{x}, y)$ is identical to the original fully connected pairwise target shown on Figure 1 (a).

a (poly)tree, see Figure 1 (c)), which typically leads to an improvement on the bound at a moderate increase in computational cost. Note, however, that discarded edges in $q(\mathsf{x})$ introduce conditional independencies which may not exist in the original distribution $p(\mathsf{x})$.

Another approach examined recently (Lawrence et al., 1998; Jaakkola and Jordan, 1998) uses mixtures of mean field type models (see Figure 1 (d)). This is a powerful extension of the standard factorized approximation, since the resulting approximating distribution $q(\mathsf{x})$ is in general multi-modal and not factorized in x. However, in general optimization of the bound (3) in this case requires minimization of the KL divergence between two fully connected distributions $p(\mathsf{x})$ and $q(\mathsf{x})$, which requires a formally computationally intractable evaluation of the entropy of the mixture $H(\mathsf{x})$. Further approximations suggested in Lawrence et al. (1998); Jaakkola and Jordan (1998); El-Hay and Friedman (2002) circumvent the computational intractability of variational mixture approximations by relaxing (3) at the cost of introducing additional variational parameters. Unfortunately, unless *all* the mixture components $q(\mathsf{x}|\mathsf{y})$ have the same tractable structure, optimization of the resulting bound on $\log Z$ is numerically unstable, which makes it difficult to generalize the existing results to variational mixtures of *arbitrarily structured* tractable experts. Moreover, additional variational parameters increase the computational cost of optimization.

2 Auxiliary Variational Method

The primary goal of this work is to extend the idea of using augmented variable spaces in a variational context. Note that the computational problems of the variational mixture approximations arise from the fact that marginalization of the mixture labels y from the joint distribution $q(\mathsf{x}, \mathsf{y})$ results in a fully connected marginal $q(\mathsf{x})$ (see Figure 1 (d)). Indeed, computation of the bound on $\log Z$ in this case requires minimization of the KL divergence between two fully-connected distributions (see expression (2)). It is intuitive that from the computational viewpoint it would be significantly more beneficial to retain the structural form of the joint distribution and use $q(\mathsf{x},\mathsf{y})$ as an approximation. In this case in order for the bound (3) to be well-defined, we introduce *auxiliary variables* y to the target distribution. This can be readily done in such a way that the marginal $\tilde{p}(\mathsf{x})$ of the joint $p(\mathsf{x},\mathsf{y})$ has the same graphical structure as the original target $p(\mathsf{x})$ (see Figure 2). Then we minimize the KL divergence between $q(\mathsf{x},\mathsf{y})$ and $p(\mathsf{x},\mathsf{y})$ in the joint variable spaces. Note that in contrast to standard structured approximations all the variables x of the marginal $q(\mathsf{x})$ remain fully connected. However, similarly to structured mean field techniques, our *auxiliary variational* method does not require evaluations of the computationally intractable entropy of the mixture $H(\mathsf{x})$. As we later show, this also applies to the cases when $q(\mathsf{x},\mathsf{y})$ has loops (as long as each $q(\mathsf{x}|\mathsf{y})$ remains structured).

Another motivation for our work on variational auxiliary approximations is the reported success of auxiliary sampling techniques, such as the *Swendsen-Wang* (Swendsen and Wang, 1987), *partial decoupling* (Higdon, 1998), and *Hybrid Monte-Carlo* (e.g. Neal (1993)) algorithms. It has been shown that by extending the original variable space with auxiliary variables and drawing samples from the joint distribution $p(\mathsf{x},\mathsf{y})$ in the augmented space, 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 is therefore hoped that an auxiliary variational method performing approximations in the augmented space may lead to a natural improvement over standard approaches.

Indeed, we demonstrate that the auxiliary variational technique forms a more powerful class of approximations than any structured mean field approach. Moreover, the method offers an efficient way of calculating a set of mixture coefficients for any choice of tractable approximators (for example, trees with different structures). These coefficients may be used to form a mixture which is principally better than a single best tractable approximator or their flat combination. Finally, we show that our approach provides a computationally simple extension of the existing variational mixture methods.

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 log partition function of $p(\mathsf{x}) = \exp\{E(\mathsf{x})\}/Z$, given by

$$\log Z \geq -\langle \log q(\mathsf{x},\mathsf{y})\rangle_{q(\mathsf{x},\mathsf{y})} + \langle E(\mathsf{x})\rangle_{q(\mathsf{x})} + \langle \log p(\mathsf{y}|\mathsf{x})\rangle_{q(\mathsf{x},\mathsf{y})}, \tag{4}$$

where $p(\mathsf{y}|\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{y}|\mathsf{x};\mathbf{\Psi})$ is an *auxiliary conditional* distribution parameterized by $\mathbf{\Psi}$. Equivalently, (4) may be written as

$$\log Z \geq \sum_{\mathsf{y}} q(\mathsf{y}) \left[\langle E(\mathsf{x}) - \log q(\mathsf{x}|\mathsf{y})\rangle_{q(\mathsf{x}|\mathsf{y})}\right] + \tilde{I}, \quad \tilde{I} = \sum_{\mathsf{x}}\sum_{\mathsf{y}} q(\mathsf{x},\mathsf{y}) \log \frac{p(\mathsf{y}|\mathsf{x})}{q(\mathsf{y})}. \tag{5}$$

The first of the two terms in expression (5) is a convex summation of the standard lower bounds (3) on $\log Z$ for the $\{\mathsf{x}\}$-variables, which cannot improve on the tightest individual bound in the set. The auxiliary variational lower bound on $\log Z$ may improve on the standard bounds only if $\tilde{I} > 0$. In the case that the auxiliary space y contains no information about x, i.e. $p(\mathsf{y}|\mathsf{x}) = p(\mathsf{y})$, it is straightforward to show that the method reproduces the standard variational bound which uses the single best approximation $q(\mathsf{x}|\mathsf{y})$. If $p(\mathsf{y}|\mathsf{x}) \equiv q(\mathsf{y}|\mathsf{x})$ then $\tilde{I}$ defines the non-negative mutual information between the original variables x and the auxiliary variables y. However, this specification leads to computational difficulties of evaluating or bounding the intractable entropy of the mixture $H(\mathsf{y})$.

We can attempt to avoid computational difficulties by imposing appropriate parametric constraints on $p(\mathsf{y}|\mathsf{x};\mathbf{\Psi})$ and maximizing the bound (5) with respect to the parameters or clique potentials $\mathbf{\Psi}$, the marginal $q(\mathsf{y})$, and the conditionals $q(\mathsf{x}|\mathsf{y})$. The general optimization algorithm in this case is given as follows:

1. Choose the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$. For the remainder, we choose

$$p(\mathsf{y}|\mathsf{x}) = \exp\{\mathbf{\Psi}(\mathsf{y};\mathsf{x})\}/Z_{\mathsf{y}|\mathsf{x}}, \quad Z_{\mathsf{y}|\mathsf{x}} = \sum_{\mathsf{y}} \exp\{\mathbf{\Psi}(\mathsf{y};\mathsf{x})\}, \tag{6}$$

   though more general distributions may potentially be considered.

2. Initialize $q(\mathsf{x}|\mathsf{y})$, $q(\mathsf{y})$, and parameters $\mathbf{\Psi}$ of $p(\mathsf{y}|\mathsf{x})$.

3. For the fixed $q(\mathsf{y},\mathsf{x})$, obtain $\mathbf{\Psi}^{new}$ by solving for zeros of

$$\partial \log Z/\partial \mathbf{\Psi} = \langle \partial \log p(\mathsf{y}|\mathsf{x})/\partial \mathbf{\Psi} \rangle_{q(\mathsf{x},\mathsf{y})}, \tag{7}$$

   or performing numerical ascent on $\log Z$ for $\mathbf{\Psi}$ [see the bound (5)].

4. For the fixed $p^{new}(\mathsf{y},\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{x})p(\mathsf{y}|\mathsf{x};\mathbf{\Psi}^{new})$ and $q(\mathsf{y})$ set

$$q^{new}(\mathsf{x}|\mathsf{y}) \propto p^{new}(\mathsf{y},\mathsf{x}) \tag{8}$$

   for all instances y.

5. For the fixed $p^{new}(\mathsf{y},\mathsf{x})$ and $q^{new}(\mathsf{x}|\mathsf{y})$ set

$$q^{new}(\mathsf{y}) \propto \exp\left\{ -\sum_{\mathsf{x}} q^{new}(\mathsf{x}|\mathsf{y}) \log \frac{q^{new}(\mathsf{x}|\mathsf{y})}{p(\mathsf{x})p(\mathsf{y}|\mathsf{x};\mathbf{\Psi}^{new})} \right\}. \tag{9}$$

6. Iterate steps 3–5 untill a termination criterion is met.

Note that for a parametric auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ step 3 is analogous to the M-step of the generalized EM algorithm (Neal and Hinton, 1998), while steps 4 and 5 are analogous to the E-step. An update rule for each term was obtained from (5) by taking functional derivatives while keeping other terms fixed.

Up to this point, the results are completely general. However, for computational reasons it is **fundamentally important** to impose constraints on the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ and the approximate joint $q(\mathsf{x},\mathsf{y})$. If both distributions are in a tractable family (for example, if each node y_i has a small number of x-parents), the bound (5) may be computed and optimized exactly. Another case leading to exact computations is when $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathbf{\Psi}\mathsf{x}, \mathbf{\Sigma})$, though in this case $q(\mathsf{x}|\mathsf{y})$ should also be parameterized.

If one wishes to have a large number of parental variables x influencing y, approximations need to be employed. For the special case of distributions of the form (6), we can utilize the standard linear upper bound $\log x \leq mx - \log m - 1$. This transforms the objective (5) to

$$\log Z \geq 1 + \sum_{\mathsf{y}} q(\mathsf{y}) \langle E(\mathsf{x}) - \log q(\mathsf{x},\mathsf{y}) \rangle_{q(\mathsf{x}|\mathsf{y})} + \left\langle \mathbf{\Psi}(\mathsf{y};\mathsf{x}) + \mu(\mathsf{x};\mathsf{y}) - e^{\mu(\mathsf{x};\mathsf{y})} Z_{\mathsf{y}|\mathsf{x}} \right\rangle_{q(\mathsf{x},\mathsf{y})} \tag{10}$$

where $e^{\mu(\mathsf{x};\mathsf{y})}$ is an additional variational functional of the exponential form.

In general, optimization of the bound (10) is numerically unstable and computationally expensive. However, as we show below, by imposing parametric or structural constraints on $p(\mathsf{y}|\mathsf{x})$ we may obtain a number of efficient approximations which obviate a recourse to (10).

2.2 Specific Auxiliary Representations

Here we briefly outline variational auxiliary representations for jointly Gaussian distributions, binary MRFs, and *structured* auxiliary models.

Jointly Gaussian Spaces

In order to test applicability of auxiliary variational approximations we may consider the simplest case of normally distributed data $p(\mathsf{x}) \sim \mathcal{N}(0, \mathbf{\Sigma}_x)$ approximated by the auxiliary variational model with a one-dimensional Gaussian auxiliary variable.

Assume that the approximating distributions are parameterized as

$$p(y|\mathsf{x}) \sim \mathcal{N}(\mathsf{u}^T\mathsf{x} + b, s^2), \quad q(y) \sim \mathcal{N}(\mu_y, \sigma_y^2), \quad q(x_i|y) \sim \mathcal{N}(\mu_i(y), \sigma_i^2). \tag{11}$$

By self-consistently solving (8) for parameters of $q(x_i|y)$ it is possible to show that the optimal variational conditional means are given by linear functions of the auxiliary variable y, which can be written explicitly as $\mu_i(y) = \Theta_i y + c_i$. By performing some algebraic manipulations we can transform the fixed point equations (7) – (9) to a system of non-linear equations for the parameters of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ and the approximate joint $q(\mathsf{x}, \mathsf{y})$ as

1. $q(x_k|y) \sim \mathcal{N}(\Theta_k y + c_k, \sigma_k^2)$:

$$\begin{cases} \sigma_k^2 = \left[w_{kk} + u_k^2/s^2\right]^{-1} \\ \Theta_k = \sigma_k^2 \left(u_k - \sum_{i \neq k} \Theta_i \left[u_i u_k/s^2 + w_{ik}\right]\right) \\ c_k = \sigma_k^2 \left(-\sum_{i \neq k} c_i \left[u_i u_k/s^2 + w_{ik}\right] - b u_k/s^2\right) \end{cases} \tag{12}$$

2. $q(y) \sim \mathcal{N}(\mu_y, \sigma_y^2)$:

$$\begin{cases} \sigma_y^2 = \left\{(1 - \mathsf{u}^T\mathbf{\Theta})^2/s^2 + \mathbf{\Theta}^T\mathsf{W}\mathbf{\Theta}\right\}^{-1} \\ \mu_y = \sigma_y^2 \left[(b + \mathsf{c}^T\mathsf{u})(1 - \mathsf{u}^T\mathbf{\Theta})/s^2 - \mathbf{\Theta}^T\mathsf{W}\mathsf{c}\right] \end{cases} \tag{13}$$

3. $p(y|\mathsf{x}) \sim \mathcal{N}(\mathsf{u}^T\mathsf{x} + b, s^2)$:

$$s^2 = \left[\sum_i \Theta_i^2/\sigma_i^2 + 1/\sigma_y^2\right]^{-1}, \quad u_i = \Theta_i s^2/\sigma_i^2, \quad b = -s^2 \sum_i c_i\Theta_i/\sigma_i^2 + \mu_y s^2/\sigma_y^2, \tag{14}$$

where $\mathsf{W} = \{w_{ij}\} = \mathbf{\Sigma}_x^{-1}$ is the inverse covariance matrix of the data. This may be viewed as an alternative formulation of the 1-factor factor analysis model.

Discrete Spaces with Parametric 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

$$p(y_k|\mathsf{x}) \propto \exp\left\{f(\mathsf{x}^T\mathsf{u}^{(k)} + b^{(k)})\right\}, \quad \mathsf{U} = \{\mathsf{u}^{(1)}, \ldots, \mathsf{u}^{(M)}\} \in \mathbb{R}^{|\mathsf{x}| \times M}, \mathsf{b} \in \mathbb{R}^M. \tag{15}$$

Here $f(\mathsf{x}; \mathsf{U}, \mathsf{b})$ is some differentiable function and $p(y_k|\mathsf{x})$ is the probability of the auxiliary variable y being in state k. Although the expectation $\langle \log p(y_k|\mathsf{x})\rangle_{q(\mathsf{x},y)}$ is in general intractable, we can make a use of the transformed bound (10). A significantly cheaper alternative is to perform optimization of (5) by approximating the term at $\log p(y_k|\langle\mathsf{x}\rangle_{q(\mathsf{x},y)})$. In practice this is often reasonably accurate, since $p(y|\mathsf{x})$ is typically uni-modal.

A potential disadvantage of using multinomial univariate y-spaces with the softmax parameterization (15) of $p(\mathsf{y}|\mathsf{x})$ is the large number of parameters, namely $|\mathsf{x} + 1| \times M$. One way to decrease it is to use a multivariate factorial representation of the auxiliary variables. In the case when $|\mathsf{x}|$ is large, $p(\mathsf{y}|\mathsf{x}) = \prod_k p(y_k|\mathsf{x})$, and each factor uses a simple (generalized) linear type dependency

$$p(y_k|\mathsf{x}) = p(y_k|a_k), \quad a_k \stackrel{\text{def}}{=} f_k(\mathsf{x}^T\mathsf{u}^{(k)} + b^{(k)}), \tag{16}$$

the *Gaussian field* approximation Barber and Sollich (2000) can be employed. From the Central Limit Theorem we can assume approximate Gaussianity of the field a_k and approximate the expectation $\langle \log p(y_i|\mathsf{x}^i)\rangle_{q(\mathsf{x}|y_k)}$ by performing 1-D Gaussian integration of the general form $\int_a f(a)p(a)$, where $p(a) \sim \mathcal{N}(\mu_a, \sigma_a^2)$. The mean and variance of the fields are readily relatable to first and second order moments of $q(\mathsf{x}|y_k)$.

In practice, the described approximations do not lead to significant deviations and are shown to be both accurate and efficient (Agakov and Barber, 2003). Probably the greatest disadvantage of these relaxations is due to the fact that for any realistic limit of $|\mathsf{x}|$ the bound is no longer strict. One way to address this is to impose additional structural constraints on the conditionals by limiting the number of parental variables for each factor.

Discrete Spaces with Structured Auxiliary Distributions

The use of sparsely-connected structured distributions $q(\mathsf{x})$ in the standard variational framework can greatly improve the accuracy of the bound. Similar improvements in performance may be obtainable from using sparse $q(\mathsf{x}|\mathsf{y})$ and $p(\mathsf{y}|\mathsf{x})$, which generally result in a tractable objective (5).

Indeed, the potentially problematic term in the bound (5) is the average of the auxiliary conditional over the joint approximating distribution $\langle \log p(\mathsf{y}|\mathsf{x})\rangle_{q(\mathsf{x},\mathsf{y})}$. However, since we are free to choose the form of the distribution $p(\mathsf{y}|\mathsf{x})$, we can limit its parental structure so that the average $\langle \log p(\mathsf{y}|\mathsf{x})\rangle_{q(\mathsf{x},\mathsf{y})}$ is tractable. For example, we may assume the factorial form for $p(\mathsf{y}|\mathsf{x})$ and constrain the number of x-parents $\boldsymbol{\pi}_x(y_i)$ of each auxiliary variable y_i, so that

$$p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\boldsymbol{\pi}_x(y_i)). \tag{17}$$

If the approximating distribution $q(\mathsf{x}|\mathsf{y})$ is also sparse and factorized in x, then all the terms in the objective (5) may be computed exactly.

Finally, note that if all components $q(\mathsf{x}|\mathsf{y})$ have the same structure for all values of y then the joint $q(\mathsf{x},\mathsf{y})$ is also structured, and the steps (8), (9) may be combined. Then if $p(\mathsf{y}|\mathsf{x})$ is fixed, the auxiliary variational method will be performing a structured approximation in the augmented space, and the existing techniques (e.g. Ghahramani and Jordan (1995); Barber and Wiegerinck (1998)) may be used.

3 Relation to Variational Mixture Models

In Section 2 we performed optimization of the bound on the log partition function subject to parametric constraints on the auxiliary conditional. Fundamentally, it is exactly due to the choice of a constrained $p(\mathsf{y}|\mathsf{x})$ that computationally efficient exact and approximate evaluations of the bound (5) are possible. In cases when there are no utilizable parametric constraints on $p(\mathsf{y}|\mathsf{x})$, optimization of (5) *requires* significantly more expensive relaxations, such as the one given by (10).

Formally, (10) generalizes the term in the objective criterion

$$\tilde{I}(\mathsf{x}, y) \geq \left\langle \log \frac{\tilde{q}(\mathsf{x}|y)}{q(y)} \right\rangle_{q(\mathsf{x},y)} + \langle \log \lambda(y)\rangle_{q(y)} + 1 - \sum_y \lambda(y) \sum_{\mathsf{x}} \tilde{q}(\mathsf{x}|y) q_{mix}(\mathsf{x}) \tag{18}$$

used by all the (known to us) variational mixture approaches (Jaakkola and Jordan (1998); Bishop et al. (1998); Lawrence et al. (1998); El-Hay and Friedman (2002)). The resulting bound is further optimized with respect to the variational functionals $\lambda(y)$, "smoothing" conditionals $\tilde{q}(\mathsf{x}|y)$, and parameters of the mixture $q(y)$ and $q(\mathsf{x}|y)$. While, theoretically, one could use (18) to variationally fit a mixture of (hyper)trees of different structures (as opposed to the mixture of completely factorized Lawrence et al. (1998) or identically structured El-Hay and Friedman (2002) models), the smoothing distributions will generally still need to be factorized, since the final term in (18) would otherwise be

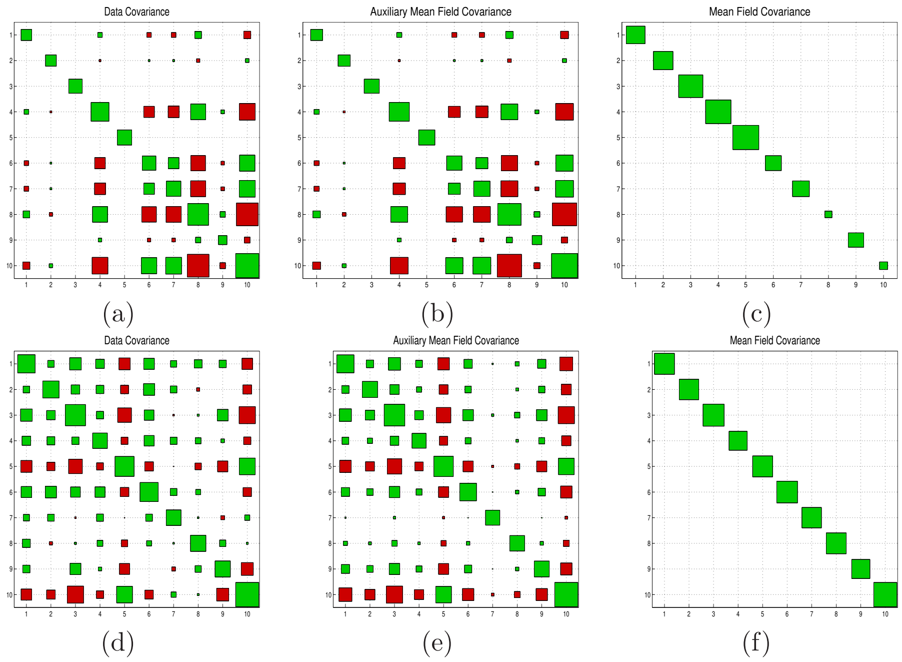


Figure 3: Approximation of the Gaussian data covariance [(a), (d)] with the auxiliary [(b), (e)] and traditional mean field [(c), (f)] approaches. *Top*: $\mathbf{\Sigma}_x$ corresponds to the data covariance of a 1-factor Factor Analysis model; *Bottom*: $\mathbf{\Sigma}_x$ is unstructured.

intractable. Also in practice, if some of the conditionals $q(\mathsf{x}|y)$ have different structures, the optimization of (18) in general becomes numerically unstable (e.g. Bishop et al. (1998) show that it involves non-factorized ratios of summations of exponentially small terms).

By permitting a tractable choice of the auxiliary conditional distribution $p(\mathsf{y}|\mathsf{x})$, the auxiliary variational method generalizes a variational mixture model in the same way as the variational EM algorithm (Neal and Hinton, 1998) generalizes the standard expectation maximization (Dempster et al., 1977). In addition to simplification of the theoretical framework, this can lead to a significant improvement in computational efficiency and numerical stability of the optimization. Moreover, it suggests fundamental extensions of the variational mixture approaches to factorial and structured auxiliary spaces. This is interesting, because it potentially allows us to boost the effective number of mixture states while remaining in the bounds of computational tractability (for example, when $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{y}|\mathsf{x})$ are constrained to be reasonably sparse). A further pleasing result from the auxiliary framework, demonstrated in Section 4, is a simple and fast way to re-weight any set of tractable distributions, which principally improves on any single structured approximation.

4 Experimental Results

Here we summarize some of the experimental results comparing performance of the auxiliary variational framework with the standard factorized approaches. Also, for discrete variable spaces we apply our method to computing reweightings of fixed approximations $q(\mathsf{x}|y)$ with different structures.

4.1 Approximating Covariance of a Gaussian

In order to verify efficiency of the auxiliary mean field approximation for normally distributed data we experimentally compared its performance with the standard mean field approximation. To quantify the performance, we computed the KL divergence between $q(\mathsf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_Q)$ and $p(\mathsf{x}) \sim \mathcal{N}(\mathbf{0}, \mathsf{W}^{-1})$, which is given by

$$KL(q\|p) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_Q \mathsf{W}| + \frac{1}{2}tr(\mathsf{W}[\boldsymbol{\Sigma}_Q + \boldsymbol{\mu}\boldsymbol{\mu}^T] - \mathsf{I}). \tag{19}$$

Figure 3 shows typical covariances obtained by auxiliary and standard MF models for structured 1-factor Factor Analysis (top) and unstructured (bottom) 10-dimensional data covariance matrices. After a few iterations of equations (12) – (14) the KL divergences are $KL(q\|p)_{MF} \approx 1.6077$, $KL(q\|p)_{AMF} \approx 0.0002$ for the structured and $KL(q\|p)_{MF} \approx 0.7484$, $KL(q\|p)_{AMF} \approx 0.2991$ for the unstructured case. As we see, the auxiliary mean field approximation results in a significantly higher accuracy than the standard mean field approach.

4.2 Inference in Discrete Markov Networks

Here we demonstrate systematic changes in the auxiliary variational estimates of the second-order moments for discrete variable spaces. Throughout the simulations, it was assumed that $p(\mathsf{x})$ is a pairwise Markov network with the energy $E(\mathsf{x}) = \mathsf{x}^T\mathsf{W}\mathsf{x} + \mathsf{x}^T\mathsf{b}$ and $\mathsf{x} \in \{-1, 1\}^{|\mathsf{x}|}$. The weight matrix W was constrained *not* to be positive-definite, so that $p(\mathsf{x})$ was inherently multi-modal. In the following experiments it is assumed that $y \in \{1, \ldots, M\}$ is multinomial and $p(y|\mathsf{x})$ has a softmax form (15) (though analogous experiments with the factorial auxiliary spaces led to principally similar results). In all cases the average $\langle \log p(y|\mathsf{x})\rangle_{q(\mathsf{x},y)}$ was approximated at the mean of $p(y|\mathsf{x})$.

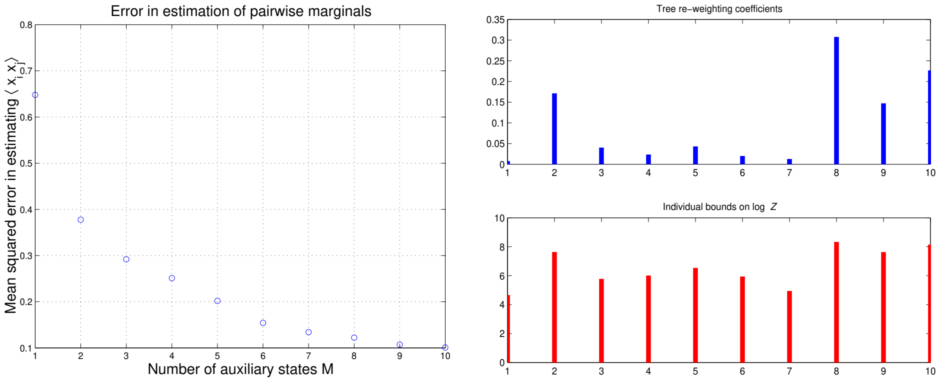
By analogy with Lawrence et al. (1998), we generated 100 biases $\mathsf{b} \in \mathbb{I}^{10}$ and matrices $\mathsf{W} \in \mathbb{I}^{10\times 10}$ of the 10-D pairwise Markov network $p(\mathsf{x})$, where $\mathbb{I}$ defines the uniform range $[-1, 1]$. Then we computed the squared errors $\epsilon(M)$ between exact and estimated second moments $\langle x_i x_j\rangle$, averaged for all networks and all $i \neq j$. Note that for $M = 1$ the auxiliary variational representation (5) is equivalent to the mean field model. As can be seen from Figure 4 (a), we obtain a systematic improvement in the accuracy with an increase in M. The scale of the changes in the accuracy is different from that reported by Lawrence et al. (1998), whose mean field error $\epsilon(1)$ was approximately 0.15, though we observe qualitatively similar improvements. This discrepancy is undoubtedly due to details of the optimization approaches, and does not detract from our conclusion that the auxiliary method conveys a systematic improvement.

4.3 Reweighting Structured Approximators

Finally, we applied our framework to computing reweightings of fixed structured approximations $q(\mathsf{x}|y)$. For a uniformly chosen $\mathsf{W} \in \mathbb{I}^{10\times 10}$, we generated $K = 10$ random spanning trees with the weights $\mathsf{W}^{(m)}$ such that $\mathsf{W}^{(m)}_{ij} = \mathsf{W}_{ij}$ for all i, j and $m = 1, \ldots, K$. Then we re-weighted the trees by recomputing $q(y)$ according to

$$q(y_k) \propto \exp\left\{\langle E(\mathsf{x}) - \log q(\mathsf{x}|y_k) + \log p(y_k|\mathsf{x})\rangle_{q(\mathsf{x}|y_k)}\right\}, \tag{20}$$

which follows directly from (9). Figure 4 (b) illustrates typical re-weightings of fixed structured approximators and the induced lower bounds on $\log Z$ for the case of the softmax parameterization of $p(y|\mathsf{x})$ and for fixed parameters of the auxiliary conditional (selected at uniform random on $\mathbb{I}$). The corresponding bounds in this case were $L_r \approx 8.38$, $L_u \approx 8.87$, $L_b \approx 8.33$, and $L_{av} \approx 9.52$ for the random, uniform, best single, and auxiliary variational weightings respectively. Note that the single-step optimization of (20) has a reasonable order of computational complexity $[\sim O(|\mathsf{x}|^2)]$ and can be easily extended to high dimensional approximations. Finally, we tried to optimize the standard bound (18) of the variational mixture models with the differently structured conditionals $q(\mathsf{x}|y)$ and fully factorized smoothing factors $\tilde{q}(\mathsf{x}|y)$. However, in this case even for $|\mathsf{x}| = 8$ we often encountered numerical problems.



(a) (b)

Figure 4: (a) Systematic changes in the mean squared error for estimates of the second-order moments with M; (b) *Top:* re-weightings of the fixed structured approximators; *Bottom:* the bounds on $\log Z$ for each individual tree.

5 Summary

We have presented an approach that generalizes the standard Kullback-Leibler variational procedure to the use of auxiliary variables, which provides a systematic improvement over the standard theory for any structured approximation. We also showed that the approximations commonly employed by variational mixture models could be seen as special and more computationally expensive cases of our approach. Furthermore, our method potentially avoids numerical difficulties present in other approaches. One way to use it in practice is to find weightings for any set of tractable distributions.

References

Agakov, F. V. and Barber, D. (2003). Temporal hidden hopfield models. In *ICANN*. Springer-Verlag.

Barber, D. and Sollich, P. (2000). Gaussian fields for approximate inference. In *Neural Information Processing Systems 12*. The MIT Press.

Barber, D. and Wiegerinck, W. (1998). Tractable variational structures for approximating graphical models. In *NIPS*. MIT Press.

Bishop, C. M., Lawrence, N., Jaakkola, T., and Jordan, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. In *Advances in Neural Information Processing Systems 10*. The MIT Press.

Dempster, A. P., Laird, M., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1).

El-Hay, T. and Friedman, N. (2002). Incorporating Expressive Graphical Models in Variational Approximations: Chain-Graphs and Hidden Variables. In *UAI: Proceedings of the 18th Conference*.

Ghahramani, Z. and Jordan, M. I. (1995). Factorial Hidden Markov Models. In *NIPS*. MIT Press.

Higdon, D. M. (1998). Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications. *Journal of the American Statistical Association*, 93(442):585–595.

Jaakkola, T. S. and Jordan, M. I. (1996). Computing upper and lower bounds on likelihoods in intractable networks. Technical Report AIM-1571.

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.

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.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of Royal Statistical Society B*, 50(2):157 – 224.

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*.

Neal, R. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

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.

Swendsen, R. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88.

Wainwright, M., Jaakkola, T., and Willsky, A. (2002). A new class of upper bounds on the log partition function. In *Uncertainty in Artificial Intelligence*.