To appear: M. I. Jordan, (Ed.), *Learning in Graphical Models*, Kluwer Academic Publishers.

IMPROVING THE MEAN FIELD APPROXIMATION VIA THE USE OF MIXTURE DISTRIBUTIONS

TOMMI S. JAAKKOLA
University of California
Santa Cruz, CA

AND

MICHAEL I. JORDAN
Massachusetts Institute of Technology
Cambridge, MA

Abstract. Mean field methods provide computationally efficient approximations to posterior probability distributions for graphical models. Simple mean field methods make a completely factorized approximation to the posterior, which is unlikely to be accurate when the posterior is multimodal. Indeed, if the posterior is multi-modal, only one of the modes can be captured. To improve the mean field approximation in such cases, we employ mixture models as posterior approximations, where each mixture component is a factorized distribution. We describe efficient methods for optimizing the parameters in these models.

1. Introduction

Graphical models provide a convenient formalism in which to express and manipulate conditional independence statements. Inference algorithms for graphical models exploit these independence statements, using them to compute conditional probabilities while avoiding brute force marginalization over the joint probability table. Many inference algorithms, in particular the algorithms that construct a junction tree, make explicit their usage of conditional independence by constructing a data structure that captures the essential Markov properties underlying the graph. That is, the algorithm groups interacting variables into clusters, such that the hypergraph of clusters has Markov properties that allow simple local algorithms to be employed for inference. In the best case, in which the original graph is sparse and without long cycles, the clusters are small and inference is efficient. In the worst case, such as the case of a dense graph, the clusters are large and inference is inefficient (complexity scales exponentially in the size of the largest cluster). Architectures for which the complexity is prohibitive include the QMR database (Shwe, et al., 1991), the layered sigmoid belief network (Neal, 1992), and the factorial hidden Markov model (Ghahramani & Jordan, 1996).

Mean field theory (Parisi, 1988) provides an alternative perspective on the inference problem. The intuition behind mean field methods is that in dense graphs each node is subject to influences from many other nodes; thus, to the extent that each influence is weak and the total influence is roughly additive (on an appropriate scale), each node should roughly be characterized by its mean value. In particular this characterization is valid in situations in which the law of large numbers can be applied. The mean value is unknown, but it is related to the mean values of the other nodes. Mean field theory consists of finding consistency relations between the mean values for each of the nodes and solving the resulting equations (generally by iteration). For graphical models these equations generally have simple graphical interpretations; for example, Peterson and Anderson (1987) and Saul, Jaakkola, and Jordan (1996) found, in the cases of Markov networks and Bayesian networks, respectively, that the mean value of a given node is obtained additively (on a logit scale) from the mean values of the nodes in the Markov blanket of the node.

Exact methods and mean field methods might be said to be complementary in the sense that exact methods are best suited for sparse graphs and mean field methods are best suited for dense graphs. Both classes of methods have significant limitations, however, and the gap between their respective domains of application is large. In particular, the naive mean field methods that we have referred to above are based on the approximation that each node fluctuates independently about its mean value. If there are a strong interactions in the network, i.e. if higher-order moments are important, then mean field methods will generally fail. One way in which such higher-order interactions may manifest themselves is in the presence of multiple modes in the distribution; naive mean field theory, by assuming independent fluctuations, effectively assumes a unimodal distribution and will generally fail if there are multiple modes.

One approach to narrowing the gap between exact methods and mean field methods is the "structured mean field" methodology proposed by Saul and Jordan (1996). This approach involves deleting links in the graph and identifying a core graphical substructure that can be treated efficiently via exact methods (e.g., a forest of trees). Probabilistic dependencies not accounted for by the probability distribution on this core substructure are treated via a naive mean field approximation. For architectures with an obvious chain-like or tree-like substructure, such as the factorial hidden Markov model, this method is natural and successful. For architectures without such a readily identified substructure, however, such as the QMR database and the layered sigmoid belief network, it is not clear how to develop a useful structured mean field approximation.

The current paper extends the basic mean field methodology in a different direction. Rather than basing the approximation on the assumption of a unimodal approximating distribution,[1] we build in multimodality by allowing the approximating distribution to take the form of a mixture distribution. The components of the mixture are assumed to be simple factorized distributions, for reasons of computational efficiency. Thus, within a mode we assume independent fluctuations, and multiple modes are used to capture higher-order interactions.

In the following sections we describe our mixture-based approach that extends the basic mean field method. Section 2 describes the basics of mean field approximation, providing enough detail so as to make the paper self-contained. Section 3 then describes how to extend the mean field approximation by allowing mixture distributions in the posterior approximation. The following sections develop the machinery whereby this approximation can be carried out in practice.

2. Mean field approximation

Assume we have a probability model $P(S, S^*)$, where S^* is the set of instantiated or observed variables (the "evidence") and the variables S are hidden or unobserved. We wish to find a tractable approximation to the posterior probability $P(S|S^*)$.

In its simplest form mean field theory assumes that nodes fluctuate independently around their mean values. We make this assumption explicit by expressing the approximation in terms of a factorized distribution $Q_{mf}(S|S^*)$:

$$Q_{mf}(S|S^*) = \prod_i Q_i(S_i|S^*), \tag{1}$$

[1]By "unimodal distributions" we mean distributions that are log-concave. Mean field distributions that are products of exponential family distributions are unimodal in this sense.

where the parameters $Q_i(S_i|S^*)$ depend on each other and on the evidence S^* (i.e., we must refit the mean field approximation for each configuration of the evidence).

The mean field approximation can be developed from several points of view, but a particularly useful perspective for applications to graphical models is the "variational" point of view. Within this approach we use the KL divergence as a measure of the goodness of the approximation and choose values of the parameters $Q_i(S_i|S^*)$ that minimize the KL divergence:

$$\mathrm{KL}(\,Q_{mf}(S|S^*)\;||\;P(S|S^*)\,) = \sum_S Q_{mf}(S|S^*) \log \frac{Q_{mf}(S|S^*)}{P(S|S^*)}. \qquad (2)$$

Why do we use $\mathrm{KL}(Q_{mf}||P)$ rather than $\mathrm{KL}(P||Q_{mf})$? A pragmatic answer is that the use of $\mathrm{KL}(Q_{mf}||P)$ implies that averages are taken using the tractable Q_{mf} distribution rather than the intractable P distribution (cf. Eq. (2)); only in the former case do we have any hope of finding the best approximation in practice. A more satisfying answer arises from considering the likelihood of the evidence, $P(S^*)$. Clearly calculation of the likelihood permits calculation of the conditional probability $P(S|S^*)$. Using Jensen's inequality we can lower bound the likelihood (cf. for example Saul, et al., 1996):

$$\begin{aligned}
\log P(S^*) &= \log \sum_S P(S, S^*) \\
&= \log \sum_S Q(S|S^*) \frac{P(S, S^*)}{Q(S|S^*)} \\
&\geq \sum_S Q(S|S^*) \log \frac{P(S, S^*)}{Q(S|S^*)},
\end{aligned}$$

for arbitrary $Q(S|S^*)$, in particular for the mean field distribution $Q_{mf}(S|S^*)$. It is easily verified that the difference between the left hand side and the right hand side of this inequality is the KL divergence $\mathrm{KL}(Q||P)$; thus, minimizing this divergence is equivalent to maximizing a lower bound on the likelihood.

For graphical models, minimization of $\mathrm{KL}(Q_{mf}||P)$ with respect to $Q_i(S_i|S^*)$ leads directly to an equation expressing $Q_i(S_i|S^*)$ in terms of the $Q_j(S_j|S^*)$ associated with the nodes in the Markov blanket of node i (Peterson & Anderson, 1987; Saul, et al., 1996). These equations are coupled nonlinear equations and are generally solved iteratively.

3. Mixture approximations

In this section we describe how to extend the mean field framework by utilizing mixture distributions rather than factorized distributions as the approximating distributions. For notational simplicity, we will drop the dependence of $Q(S|S^*)$ on S^* in this section and in the remainder of the paper, but it must be borne in mind that the approximation is optimized separately for each configuration of the evidence S^*. We also utilize the notation $\mathcal{F}(Q)$ to denote the lower bound on the right hand side of Eq. (3):

$$\mathcal{F}(Q) \equiv \sum_S Q(S) \log \frac{P(S,S^*)}{Q(S)} \tag{3}$$

When $Q(S)$ takes the factorized form $Q_{mf}(S)$ in Eq. (1), we will refer to the approximation as a "factorized mean field" approximation and will denote the corresponding bound as $\mathcal{F}(Q_{mf})$.

Our proposal is to utilize a mixture distribution as the approximating distribution:

$$Q_{mix}(S) = \sum_m \alpha_m Q_{mf}(S|m), \tag{4}$$

where each of the component distributions $Q_{mf}(S|m)$ are factorized distributions. The mixing proportions α_m, which are constrained to be positive and sum to one, are additional parameters to be fit via the minimization of the KL divergence.

In the remainder of this section we focus on the relationship between $\mathcal{F}(Q_{mix})$ – the mixture bound – and the factorized mean field bounds $\mathcal{F}(Q_{mf}|m)$ corresponding to the mixture components. To this end, consider:

$$\begin{aligned}
\mathcal{F}(Q_{mix}) &= \sum_S Q_{mix}(S) \log \frac{P(S,S^*)}{Q_{mix}(S)} \\
&= \sum_{m,S} \alpha_m \left[Q_{mf}(S|m) \log \frac{P(S,S^*)}{Q_{mix}(S)} \right] \\
&= \sum_{m,S} \alpha_m \left[Q_{mf}(S|m) \log \frac{P(S,S^*)}{Q_{mf}(S|m)} + Q_{mf}(S|m) \log \frac{Q_{mf}(S|m)}{Q_{mix}(S)} \right] \\
&= \sum_m \alpha_m \mathcal{F}(Q_{mf}|m) + \sum_{m,S} \alpha_m Q_{mf}(S|m) \log \frac{Q_{mf}(S|m)}{Q_{mix}(S)} \\
&= \sum_m \alpha_m \mathcal{F}(Q_{mf}|m) + I(m;S)
\end{aligned} \tag{5}$$

where $I(m;S)$ is the mutual information between the mixture index m and the variables S in the mixture distribution[2]. The first term in Eq. (5)

[2]Recall that the mutual information between any two random variables x and y is I(x;y) $= \sum_{x,y} P(x,y) \log P(y|x)/P(y)$.

is a convex combination of factorized mean field bounds and therefore it yields no gain over factorized mean field as far as the KL divergence is concerned. It is the second term, i.e., the mutual information, that characterizes the gain of using the mixture approximation.[3] As a non-negative quantity, $I(m;S)$ increases the likelihood bound and therefore improves the approximation. We note, however, that since $I(m;S) \leq \log M$, where M is the number of mixture components, the KL divergence between the mixture approximation and the true posterior can decrease at most logarithmically in the number of mixture components.

4. Computing the mutual information

To be able to use the mixture bound of Eq. (5) we need to be able to compute the mutual information term. The difficulty with the expression as it stands is the need to average the log of the mixture distribution over all configurations of the hidden variables. Given that there are an exponential number of such configurations, this computation is in general intractable. We will make use of a refined form of Jensen's inequality to make headway.

Let us first rewrite the mutual information as

$$
\begin{aligned}
I(m;S) &= \sum_{m,S} \alpha_m Q_{mf}(S|m) \log \frac{Q_{mf}(S|m)}{Q_{mix}(S)} && (6) \\
&= \sum_{m,S} \alpha_m Q_{mf}(S|m) \left[-\log \frac{Q_{mix}(S)}{Q_{mf}(S|m)} \right] && (7) \\
&= \sum_{m,S} \alpha_m Q_{mf}(S|m) \left[-\log \frac{\alpha_m R(S|m)}{\alpha_m R(S|m)} \frac{Q_{mix}(S)}{Q_{mf}(S|m)} \right] && (8) \\
&= \sum_{m,S} \alpha_m Q_{mf}(S|m) \log R(S|m) - \sum_m \alpha_m \log \alpha_m \\
&\quad + \sum_{m,S} \alpha_m Q_{mf}(S|m) \left[-\log \frac{R(S|m)}{\alpha_m} \frac{Q_{mix}(S)}{Q_{mf}(S|m)} \right] && (9)
\end{aligned}
$$

where we have introduced additional smoothing distributions $R(S|m)$ whose role is to improve the convexity bounds that we describe below.[4] With some foresight we have also introduced α_m into the expression; as will become

[3]Of course, this is only the potential gain to be realized from the mixture approximation. We cannot rule out the possibility that it will be harder to search for the best approximating distribution in the mixture approximation than in the factorized case. In particular the mixture approximation may introduce additional local minima into the KL divergence.

[4]In fact, if we use Jensen's inequality directly in Eq. (6), we obtain a vacuous bound of zero, as is easily verified.

apparent in the following section the forced appearance of α_m turns out to be convenient in updating the mixture coefficients.

To avoid the need to average over the logarithm of the mixture distribution, we make use of the following convexity bound:

$$-\log(x) \geq -\lambda\, x + \log(\lambda) + 1, \tag{10}$$

replacing the $-\log(\cdot)$ terms in Eq. (9) with linear functions, separately for each m. These substitutions imply a lower bound on the mutual information given by:

$$\begin{aligned} I(m;S) \ \geq\ & \sum_{m,S} \alpha_m Q_{mf}(S|m) \log R(S|m) - \sum_m \alpha_m \log \alpha_m \\ & - \sum_m \lambda_m \sum_S R(S|m) Q_{mix}(S) + \sum_m \alpha_m \log \lambda_m + 1 \qquad (11) \\ \equiv\ & I_\lambda(m;S). \qquad (12) \end{aligned}$$

The tightness of the convexity bound is controlled by the smoothing functions $R(S|m)$, which can be viewed as "flattening" the logarithm function such that Jensen's inequality is tightened. In particular, if we set $R(S|m) \propto Q_{mf}(S|m)/Q_{mix}(S)$ and maximize over the new variational parameters λ_m we recover the mutual information exactly. Such a choice would not reduce the complexity of our calculations, however. To obtain an efficient algorithm we will instead assume that the distributions R have simple factorized forms, and optimize the ensuing lower bound with respect to the parameters in these distributions. It will turn out that this assumption will permit all of the terms in our bound to be evaluated efficiently. We note finally that this bound is never vacuous. To see why, simply set $R(S|m)$ equal to a constant for all $\{S, m\}$ in which case $\max_\lambda I_\lambda(m;S) = 0$; optimization of R can only improve the bound.

5. Finding the mixture parameters

In order to find a reasonable mixture approximation we need to be able to optimize the bound in Eq. (5) with respect to the mixture model. Clearly a necessary condition for the feasibility of this optimization process is that the component factorized mean field distributions can be fit to the model tractably. That is, we must assume that the following mean field equations:

$$\frac{\partial \mathcal{F}(Q_{mf})}{\partial Q_j(S_j)} = constant \tag{13}$$

can be solved efficiently (not necessarily in closed form) for any of the marginals $Q_j(S_j)$ in the factorized mean field distribution Q_{mf}.[5] Examples of such mean field equations are provided by Peterson and Anderson (1987) and Saul, et al. (1996).

Let us first consider the updates for the factorized mean field components of the mixture model. We need to take derivatives of the lower bound:

$$\mathcal{F}(Q_{mix}) \geq \sum_m \alpha_m \mathcal{F}(Q_{mf}|m) + I_\lambda(m;S) \equiv \mathcal{F}_\lambda(Q_{mix}) \tag{14}$$

with respect to the parameters of the component factorized distributions. The important property of $I_\lambda(m;S)$ is that it is linear in any of the marginals $Q_j(S_j|k)$, where k indicates the mixture component (see Appendix A). As a result, we have:

$$\frac{\partial I_\lambda(m;S)}{\partial Q_j(S_j|k)} = constant \tag{15}$$

where the constant is independent of the marginal $Q_j(S_j|k)$ (and is generally dependent on the other marginals in the particular component mean field distribution). It follows from our assumption that the equations:

$$\frac{\partial \mathcal{F}_\lambda(Q_{mix})}{\partial Q_j(S_j|k)} = \alpha_k \frac{\partial \mathcal{F}_\lambda(Q_{mf}|k)}{\partial Q_j(S_j|k)} + \frac{\partial I_\lambda(m;S)}{\partial Q_j(S_j|k)} = 0 \tag{16}$$

can be solved efficiently for any specific marginal $Q_j(S_j|k)$. We can thus find the best fitting mixture components by iteratively optimizing each of the marginals in the model.

Let us now turn to the mixture coefficients α_k. We note first that, apart from the entropy term $-\sum_k \alpha_k \log \alpha_k$, these coefficients appear linearly in $I_\lambda(m;S)$. The same is therefore true for $\mathcal{F}_\lambda(Q_{mix})$ (see Eq. (14)), and we can write:

$$\mathcal{F}_\lambda(Q_{mix}) \;=\; \sum_k \alpha_k(-E(k)) - \sum_k \alpha_k \log \alpha_k + 1 \tag{17}$$

where $-E(k)$ is the collection of terms linear in α_k:

$$\begin{aligned} -E(k) \;\equiv\; & \mathcal{F}(Q_{mf}|k) + \sum_S Q_{mf}(S|k)\log R(S|k) \\ & + \sum_m \lambda_m \sum_S R(S|m) Q_{mf}(S|k) + \log \lambda_k \end{aligned} \tag{18}$$

[5]In the ordinary mean field equations the constant is zero, but for many probability models of interest this slight difference poses no additional difficulties.

Now Eq. (17) has the form of a free energy, and the mixture coefficients that maximize the lower bound $\mathcal{F}_\lambda(Q_{mix})$ therefore come from the Boltzmann distribution:

$$\alpha_k = \frac{e^{-E(k)}}{\sum_{k'} e^{-E(k')}}. \tag{19}$$

(This fact is easily verified using Lagrange multipliers to optimize Eq. (17) with respect to α_m).

Finally, we must optimize the bound with respect to the parameters λ_m and the parameters of the smoothing distribution. We discuss the latter updates in Appendix A. Differentiating the bound with respect to the parameters λ_m, we readily obtain:

$$\lambda_m = \frac{\alpha_m}{\sum_k \alpha_k \sum_S R(S|m) Q_{mf}(S|k)}, \tag{20}$$

where a simplified form for the expression in the denominator is presented in the Appendix.

6. Discussion

We have presented a general methodology for using mixture distributions in approximate probabilistic calculations. We have shown that the bound on the likelihood resulting from mixture-based approximations is composed of two terms, one a convex combination of simple factorized mean field bounds and the other the mutual information between the mixture labels and the hidden variables S. It is the latter term which represents the improvement available from the use of mixture models.

The basic mixture approximation utilizes parameters α_m (the mixing proportions), $Q_i(S_i|m)$ (the parameters of the factorized component distributions). These parameters are fit via Eq. (19) and Eq. (16), respectively. We also introduced variational parameters λ_m and $R_i(S_i|m)$ to provide tight approximations to the mutual information. Updates for these parameters were presented in Eq. (20) and Eq. (29).

Bishop, et al. (1998) have presented empirical results using the mixture approximation. In experiments with random sigmoid belief networks, they showed that the mixture approximation provides a better approximation than simple factorized mean field bounds, and moreover that the approximation improves monotonically with the number of components in the mixture.

An interesting direction for further research would involve the study of hybrids that combine the use of the mixture approximation with exact methods. Much as in the case of the structured mean field approximation of

Saul and Jordan (1996), it should be possible to utilize approximations that identify core substructures that are treated exactly, while the interactions between these substructures are treated via the mixture approach. This would provide a flexible model for dealing with a variety of forms of high-order interactions, while remaining tractable computationally.

References

Bishop, C., Lawrence, N., Jaakkola, T. S., & Jordan, M. I. Approximating posterior distributions in belief networks using mixtures. In M. I. Jordan, M. J. Kearns, & S. A. Solla, *Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge MA (1998).

Ghahramani, Z. & Jordan, M. I. Factorial Hidden Markov models. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge MA (1996).

Neal, R. Connectionist learning of belief networks, *Artificial Intelligence*, *56*: 71-113 (1992).

Parisi, G. *Statistical Field Theory.* Addison-Wesley: Redwood City (1988).

Peterson, C. & Anderson, J. R. A mean field theory learning algorithm for neural networks. *Complex Systems* **1**:995–1019 (1987).

Saul, L. K., Jaakkola, T. S., & Jordan, M. I. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, *4*:61–76 (1996).

Saul, L. K. & Jordan, M. I. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge MA (1996).

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., & Cooper, G. F. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Meth. Inform. Med. 30*: 241-255 (1991).

A. Optimization of the smoothing distribution

Let us first adopt the following notation:

$$\pi_{R,Q}(m,m') \equiv \sum_S R(S|m) Q_{mf}(S|m') \tag{21}$$

$$= \prod_i \left[\sum_{S_i} R_i(S_i|m) Q_i(S_i|m') \right] \tag{22}$$

$$H(Q \parallel R\,|m) \equiv \sum_S Q_{mf}(S|m) \log R(S|m) \tag{23}$$

$$= \sum_i \sum_{S_i} Q_i(S_i|m) \log R_i(S_i|m) \tag{24}$$

$$H(m) \equiv -\sum_m \alpha_m \log \alpha_m \tag{25}$$

where we have evaluated the sums over S by making use of the assumption that $R(S|m)$ factorizes. We denote the marginals of these factorized forms by $R_i(S_i|m)$. Eq. (11) now reads as:

$$\begin{aligned} I_\lambda(m;S) &= \sum_m \alpha_m H(Q \parallel R\,|m) + H(m) \\ &\quad - \sum_m \lambda_m \left[\sum_{m'} Q(m')\, \pi_{R,Q}(m,m')\right] + \sum_m \alpha_m \log \lambda_m + 1 \end{aligned} \tag{26}$$

To optimize the bound $I_\lambda(m;S)$ with respect to R, we need the following derivatives:

$$\frac{\partial\, H(Q \parallel R\,|m)}{\partial\, R_j(S_j|k)} = \delta_{k,m}\, \frac{Q_j(S_j|k)}{R_j(S_j|k)} \tag{27}$$

$$\frac{\partial\, \pi_{R,Q}(m,m')}{\partial\, R_j(S_j|k)} = \delta_{k,m}\, Q_j(S_j|m') \prod_{i\neq j} \left[\sum_{S_i} R_i(S_i|m) Q_i(S_i|m')\right]. \tag{28}$$

Denoting the product in Eq. (28) by $\pi^j_{R,Q}(m,m')$, we can now characterize the optimal R via the consistency equations:

$$\frac{\partial\, I_\lambda(m;S)}{\partial\, R_j(S_j|k)} = \alpha_k \frac{Q_j(S_j|k)}{R_j(S_j|k)} - \lambda_k \left[\sum_{m'} Q(m')\, \pi^j_{R,Q}(k,m') Q_j(S_j|m')\right] = 0 \tag{29}$$

Note that the second term does not depend on the smoothing marginal $R_j(S_j|k)$ and therefore these consistency equations are easily solved for any specific marginal. The best fitting R is found by iteratively optimizing each of its marginals. We note finally that the iterative solutions are not necessarily normalized. This is of no consequence, however, since the information bound of Eq. (11)—once maximized over the λ parameters—is invariant to the normalization of R.