6.892 Machine learning seminar

http://www.ai.mit.edu/courses/6.892-ml

Prof. Tommi S. Jaakkola
tommi@ai.mit.edu

March 17, 1999

1 Maximum entropy estimation

- We want to induce a probability distribution from a set of observations (constraints) so that the resulting distribution makes fewest assumptions about the examples.
- The nature of these "assumptions" is very different from the independence assumptions considered earlier in the class.
- The key idea is that we shouldn't assign different probabilities to configurations that we have no reason to differentiate.

1.1 Maximum entropy principle

- Find the probability distribution $P(x)$ that has the maximum entropy subject to the "moments" constraints

$$\sum_x P(x) r_i(x) = \alpha_i \tag{1}$$

for $i = 1, \ldots, m$. We allow inequality constraints as well. The "moments" can be e.g.

$$r_i(x) = 1, \;\; r_i(x) = x, \;\; r_i(x) = x^2, \;\; r_i(x) = \log Q(x), \;\; \ldots \tag{2}$$

- The solution is obtained easily via Lagrange multipliers: we introduce Lagrange multipliers λ_i, one for each constraint[1], and maximize

$$J(P, \lambda) = \lambda_0 \left[\sum_x P(x) - 1 \right] + \sum_i \lambda_i \left[\sum_x P(x) r_i(x) - \alpha_i \right] + H(P) \tag{3}$$

[1] λ_i corresponding to an inequality constraint $\ldots \leq \alpha_i$ (or $\ldots \geq \alpha_i$) must remain non-negative (non-positive).

with respect to P. Here $H(P)$ is the entropy of distribution P:

$$H(P) = -\sum_x P(x) \log P(x) \tag{4}$$

The solution $P = P_\lambda$ has the following general form

$$P_\lambda(x) = \frac{1}{Z(\lambda)}\, e^{\lambda_1 r_1(x)+\ldots+\lambda_m r_m(x)} \tag{5}$$

where

$$Z(\lambda) = \sum_x e^{\lambda_1 r_1(x)+\ldots+\lambda_m r_m(x)} \tag{6}$$

is the normalization constant (partition function) and plays an important role.

- The coefficients $\{\lambda_i\}, i = 1, \ldots, m$, are chosen to satisfy the moments constraints. In the above solution we had already chosen λ_0 so that the distribution sums to one. An efficient method for finding the remaining coefficients is described below. We note, however, that it is not always possible to find such coefficients.

  Ex: Assume $-\infty \leq x \leq \infty$ and we have only one additional constraint $r_1(x) = x$. In this case, we cannot even normalize the distribution $P(x) \propto e^{-\lambda_1 x}$. The reason in this case is that the maximum entropy is not attainable (it is ∞ in this case).

- Some properties of the partitition function:

$$\begin{aligned}
\frac{\partial}{\partial \lambda_i} \log Z(\lambda) &= \sum_x P_\lambda(x) r_i(x) = E_{P_\lambda}\{\, r_i(x)\,\} && (7)\\
\frac{\partial^2}{\partial \lambda_i^2} \log Z(\lambda) &= Var_{P_\lambda}\{\, r_i(x)\,\} && (8)\\
\frac{\partial^2}{\partial \lambda_i \partial \lambda_j} \log Z(\lambda) &= Cov_{P_\lambda}\{\, r_i(x), r_j(x)\,\} && (9)
\end{aligned}$$

  Theorem 1 $\log Z(\lambda)$ *is a jointly convex function of the coefficients* $\{\lambda_i\}$.

  Proof: Covariance matrix is positive (semi-)definite. The result follows from the properties we have listed above.

- Assuming the coefficients $\lambda = \lambda^*$ have been set to satisfy the constraints we may ask: what is the entropy of the max entropy distribution $P_{\lambda^*}(x)$?

$$\begin{aligned}
-H(P^*; \alpha) &= \sum_x P_{\lambda^*}(x) \log P_{\lambda^*}(x) && (10)\\
&= \sum_x P_{\lambda^*}(x) \{\, \sum_i \lambda_i^* r_i(x) - \log Z(\lambda^*)\,\} && (11)\\
&= \sum_i \lambda_i^* \alpha_i - \log Z(\lambda^*) && (12)
\end{aligned}$$

Consider for a moment the resulting expression for the negative entropy as a function of λ, i.e.

$$\sum_i \lambda_i \alpha_i - \log Z(\lambda) \tag{13}$$

By our earlier results, this must be a concave function of $\{\lambda_i\}$. Its maximum is thus obtained at the point were the derivatives vanish, i.e., when

$$\alpha_i - \frac{\partial}{\partial \lambda_i} \log Z(\lambda) = \alpha_i - \sum_x P_\lambda(x) r_i(x) = 0 \tag{14}$$

for all i. It is clear that these fixed point equations correspond to satisfying the moments constraints. Thus $\{\lambda_i^*\}$ are the coefficients that maximize the above expression. By exploiting this observation we can write the negative entropy as

$$-H(P^*; \alpha) \quad = \quad \max_\lambda \{ \sum_i \lambda_i \alpha_i - \log Z(\lambda) \} \tag{15}$$

which provides us with an efficient and general way of setting the coefficients $\{\lambda_i\}$ to satisfy the constraints.

Note that this maximization step actually *minimizes* the entropy (maximizes the negative entropy). This has to do with the max–min nature of the Lagrange multiplier method.

1.2 Convex duality digression

Theorem 2 *Any sufficiently regular convex function can be written as*

$$f(\lambda) = \max_\alpha \{ \lambda^T \alpha - f^*(\alpha) \} \tag{16}$$

where $f^*(\alpha)$ is the dual or conjugate function of $f(\lambda)$, it is also convex and satisfies

$$f^*(\alpha) = \max_\lambda \{ \lambda^T \alpha - f(\lambda) \} \tag{17}$$

Lemma 1 $f^{**} = f$ *(dual of the dual is the original function).*

- The results obtained in the previous section suggest that the negative entropy is the dual of the log-partition function.

1.3 Minimum relative entropy principle

- The minimum relative entropy principle is a slight generalization of the maximum entropy principle. Here we introduce a "prior" distribution $P_0(x)$ and try to induce a new distribution $P(x)$ that is consistent with the observed constraints but nevertheless close to the prior distribution P_0.

- The objective is to find a distribution $P(x)$ which achieves the minimum relative entropy between $P(x)$ and $P_0(x)$, i.e.,

$$D(P\|P_0) = \sum_x P(x) \log \frac{P(x)}{P_0(x)} \tag{18}$$

  subject to the contraints

$$\sum_x P(x) r_i(x) = \alpha_i \tag{19}$$

  for $i = 1, \ldots, m$. Similarly to the max entropy case, we can accomodate equality as well as inequality constraints.

- By setting $P_0(x)$ to the uniform distribution (non-informative prior) we recover the maximum entropy principle. This is because the minimization of $D(P\|P_0)$ reduces to minimizing $-H(P)$ whenever P_0 is a uniform distribution. Minimizing the negative entropy is, of course, the same as maximizing the entropy.

- The solution to this variational problem can be obtained via Lagrange multipliers as before:

$$P_\lambda(x) = \frac{1}{Z(\lambda; P_0)} P_0(x)\, e^{\lambda_1 r_1(x) + \ldots + \lambda_m r_m(x)} \tag{20}$$

  where

$$Z(\lambda; P_0) = \sum_x P_0(x)\, e^{\lambda_1 r_1(x) + \ldots + \lambda_m r_m(x)} \tag{21}$$

  and the coefficients λ are set to satisfy the constraints.

- Note the "Bayesian" style combination of the *priori* distribution $P_0(x)$ and the distribution that depends on the constraints (evidence).

- The minimum relative entropy achieved by the solution is

$$D(P_{\lambda^*}\|P_0) = \max_\lambda \{ \sum_i \lambda_i \alpha_i - \log Z(\lambda; P_0) \} \tag{22}$$

  where, again, the right hand side provides us with an efficient way of setting the coefficients $\{\lambda_i\}$ (to satisfy the moments constraints).

  Note that relative entropy distance as a function of P_λ is the dual of the log partition function in this setting.

1.4 Relation to large deviation theory

- The minimum relative entropy principle is quite general and appears for example in large deviation theory. For later utility, we will elaborate this relation here a bit.

- Suppose we have a sequence $\{x_1, \ldots, x_n\}$ of i.i.d. discrete random variables x_l. The probability of any observed sequence X is in this case given by

$$P_0^{(n)}(X) = \prod_{l=1}^{n} P_0(x_l) \tag{23}$$

  where $P_0(x)$ is the generative distribution.

- A typical problem in large deviation theory is to bound the probability of rare events, i.e., events whose probability goes to zero as $n \to \infty$. For example, we may be interested in calculating the probability that the empirical mean of an observed sequence deviates from the true mean by more than ϵ. Since the empirical mean converges to the true mean in the limit, the probability of this event goes to zero for any fixed ϵ. In fact, it goes to zero exponentially fast.

- The rare event we are interested in here is defined by a set of k inequality constraints:

$$\sum_{l=1}^{n} r_i(x_l) \geq n\,\alpha_i \tag{24}$$

  for $i = 1, \ldots, k$. The log-probability that these constraints are satisfied for a sequence drawn at random from $P_0^{(n)}$ is given by

$$\begin{aligned}
&\log P\left(\forall i,\ \sum_{l=1}^{n} r_i(x_l) \geq n\alpha_i\right) && (25)\\
&= \log \sum_X P_0^{(n)}(X) \prod_i \text{step}\left(\sum_{l=1}^{n} r_i(x_l) - n\alpha_i\right) && (26)\\
&\leq \log \sum_X P_0^{(n)}(X) \prod_i e^{\lambda_i(\sum_{l=1}^{n} r_i(x_l) - n\alpha_i)} && (27)\\
&= \log \prod_{l=1}^{n} \sum_{x_l} P_0(x_l) e^{\sum_i \lambda_i r_i(x_l) - \alpha_i)} && (28)\\
&= -n\left[\sum_i \lambda_i \alpha_i - \log \sum_{x_l} P_0(x_l) e^{\sum_i \lambda_i r_i(x_l)}\right] && (29)
\end{aligned}$$

  where we have used a standard bound

$$\text{step}(z) \leq e^{\lambda z} \tag{30}$$

valid for $\lambda \geq 0$ and rearranged the terms a little.

The best large deviation bound is found by optimizing the above expression with respect to λ. In other words,

$$
\begin{aligned}
& \log P\left(\forall i,\ \sum_{l=1}^{n} r_i(x_l) \geq n\alpha_i\right) && (31)\\
\leq\ & -n\ \max_{\lambda}\left[\sum_i \lambda_i\alpha_i - \log\sum_{x_l} P_0(x_l)e^{\sum_i \lambda_i r_i(x_l)}\right] && (32)\\
=\ & -n\ \max_{\lambda}\left[\sum_i \lambda_i\alpha_i - \log Z(\lambda; P_0)\right] && (33)\\
=\ & -n\ D(P^*\|P_0) && (34)
\end{aligned}
$$

where $D(P^*\|P_0)$ is the minimum attainable relative entropy between the generative (prior) distribution P_0 and P subject to the constraints

$$\sum_x P(x)r_i(x) \geq \alpha_i \tag{35}$$

- Note a few things:

  1. The log-probability of this rare event decreases linearly as a function of n or, equivalently, the probability of the event goes down exponentially fast. The rate of decrease is governed by the minimum relative entropy $D(P^*\|P_0)$.
  2. When the generative distribution P_0 also satisfies the constraints, then necessarily $D(P^*\|P_0) = 0$ as we can choose $P^* = P_0$. In this case, we are dealing with typical events, i.e., events whose probability does not decrease as a function of n.
  3. The minimum relative entropy distribution P^* is unique. It is rather remarkable that the probability of the rare event is governed by $D(P^*\|P_0)$, the distance from this single point to the generative distribution (as opposed to some integral over the whole set of distributions consistent with the constraints).

1.5 Variational free energy

- Suppose we have a large physical system consisting of N identical subsystems. Each subsystem can be in configuration $x \in \mathcal{X}$ with the corresponding energy $E(x)$.
- The total energy of the system is conserved:

$$\frac{1}{N}E_{tot} = \sum_x \frac{N(x)}{N}\,E(x) = \sum_x P(x)\,E(x) \tag{36}$$

where $N(x)$ is the number of subsystems that happen to be in configuration x.

- We want to compute the probability of finding a randomly chosen subsystem in configuration x. For large N this probability is simply $N(x)/N$ but how do we find $N(x)$?

  The most likely setting of $\{N(x)\}, x \in \mathcal{X}$ is the one that can be realized in the largest number of ways. In other words, the ratios $N(x)/N$, $x \in \mathcal{X}$ for large N must settle into values that impose least restrictions on the underlying subsystems, leaving the configuration of each individual subsystem most uncertain.

- This is equivalent to finding the maximum entropy distribution $P(x) = N(x)/N$ subject to the energy constraint. Thus we find P by maximizing

$$(-\beta\Phi(P)) = -\beta \sum_x P(x)\, E(x) + H(P) \tag{37}$$

  The sign in front of the Lagrange multiplier β is simply a convention. $\Phi(P)$ is known as the variational free energy.

  Note: β and E_{tot} interchangeable specifications of the maximum entropy solution. We will stick with β.

  The maximum entropy distribution is the Boltzmann distribution

$$P^*(x) = \frac{1}{Z}\, e^{-\beta E(x)} \tag{38}$$

  Varying β in this distribution corresponds to satisfying different total energy constraints. Typically $\beta = 1/T$ where T is the temperature. Note the connection between temperature and the total energy of the system.

- The variational free energy evaluated at the Boltzmann distribution is given by

$$(-\beta\Phi(P^*)) = \log Z(P^*) \tag{39}$$

  Thus free energy corresponds to the log-partition function.

- For large systems, we typically cannot compute the log-partition function. An important relation for any approximate computation of free energy is

$$(-\beta\Phi(P)) \leq (-\beta\Phi(P^*)) \tag{40}$$

  where P can be any distribution. In other words, if we can evaluate the variational free energy for any simple distribution P, then we have a bound on the free energy.

2 Probabilistic inference

Here we establish a correspondence between variational free energy and probabilistic calculations. In essence, we'll transform probabilistic inference calculations into an optimization problem analogous to the calculation of free energy.

Assume we have a probability model $P_0(x)$ and we are interested in calculating marginal probabilities over some subset of variables. Let $x = x_v \cup x_h$ where x_v is the set of "visible" variables whose probability we are interested in computing and x_h is the set of remaining or "hidden" variables.

(Note that h is a set of nodes in the graph while x_h is the set of variables associated with these nodes)

- Consider the following correspondence:

  $\beta \quad \leftrightarrow \quad 1$

  $E(x_h) \quad \leftrightarrow \quad -\log P_0(x_v, x_h)$

  Setting $\beta = 1$ implies that we are looking at a system in constant temperature. Defining energy as negative log-probability makes sense from a coding perspective. Assuming x_v and x_h are drawn at random from P_0, then the length of the optimal code encoding (x_v, x_h) would be $-\log P_0(x_v, x_h)$. Thus the lower the energy, the shorter the code.

- In the maximum entropy formulation and given the above substitutions we want to maximize

$$L(x_v, P) = \sum_x P(x) \log P_0(x_v, x_h) + H(P) \tag{41}$$

  with respect to the variational distribution P. We use here the notation $L(x_v, P)$ instead of variational free energy $\Phi(P)$ as it is more convenient and allows us to phase out the physics terminology.

- The maximum entropy solution is the Boltzmann distribution:

$$P^*(x_h) \;=\; \frac{1}{Z} e^{-E(x_h)} = \frac{1}{Z} e^{\log P_0(x_v, x_h)} = \frac{1}{Z} P_0(x_v, x_h) \tag{42}$$

  where $Z = \sum_{x_h} P_0(x_v, x_h) = P_0(x_v)$. Thus $P^*(x_h)$ is the posterior distribution over the hidden variables x_h given x_v:

$$P^*(x) = \frac{1}{P_0(x_v)} P_0(x_v, x_h) = P_0(x_h | x_v) \tag{43}$$

We saw earlier that the variational free energy evaluated at the Boltzmann distribution was proportional to the log-partition function. Analogously, in terms of $L(x_v, P)$, we have

$$L(x_v, P^*) = \log P_0(x_v) \tag{44}$$

- This is an imporant result. It says that the maximum of $L(x_v, P)$ over all distributions P is the log-probability over the variables of interest, i.e., $\log P_0(x_v)$. This completes our transformation of inference problems in graphical models as optimization problems.

- It is important to note that the above variational calculation of $\log P_0(x_v)$ is no more efficient than calculating this marginal directly, i.e.,

$$\log P_0(x_v) = \log \sum_{x_h} P_0(x_v, x_h) \tag{45}$$

The utility of the variational formulation lies in our ability to approximate these calculations.

2.1 Approximate inference

Based on the previous results, we know that

$$\begin{aligned} \log P_0(x_v) &= \max_P L(x_v, P) && (46) \\ &\geq L(x_v, P) && (47) \\ &= \sum_{x_h} P(x_h) \log P_0(x_v, x_h) + H(P) && (48) \end{aligned}$$

which holds for any distribution P.

- Thus whenever we cannot compute $\log P_0(x_v)$ exactly, we can always obtain a lower bound by choosing *any* simple enough variational distribution P for which we are able to evaluate the bound efficiently. Optimizing this bound within the chosen class of simple distributions can improve the quality of the bound considerably.

2.2 MRF example and the mean field approximation

To illustrate the above framework for obtaining lower bounds on marginal probabilities consider any joint distribution that can be written in terms of pairwise marginals, i.e.,

$$P_0(x_v, x_h) = \prod_e \phi_e(x_e) \tag{49}$$

where the product is over the edges in the undirected graph. For notational simplicity, we have left implicit the division of variables into visible and hidden. An example of this kind of model is a 2-D square lattice:

```
   |   |
 --O--O--  ...
   |   |
 --O--O--  ...
   |   |
```

It is typically not feasible to evaluate $\log P_0(x_v)$ exactly in these models. We can, however, compute a lower bound on $\log P_0(x_v)$:

$$\log P_0(x_v) \geq \sum_{x_h} P(x_h) \log \prod_e \phi_e(x_e) + H(P) \tag{50}$$

$$= \sum_e \sum_{x_{h\cap e}} P(x_{h\cap e}) \log \phi_e(x_e) + H(P) \tag{51}$$

where $h \cap e$ corresponds to the nodes that are associated with a particular edge and are also identified as "hidden". Now, what variational distribution P should we use to be able to evaluate this bound efficiently?

The simplest choice is the completely factorized distribution:

$$P(x_h) = \prod_{i\in h} P_i(x_i) \tag{52}$$

This choice leads to what is known as *mean field* approximation. There are two components to this approximation: 1) evaluation of the bound for any fixed factorized distribution P and 2) the optimization or adjustment of P so as to obtain the tightest bound in this class of distributions.

2.2.1 Evaluation of the lower bound

Since P is assumed to be completely factorized, it is easy to compute the required marginals $P(x_{h\cap e})$:

$$P(x_{h\cap e}) = \prod_{i\in h\cap e} P_i(x_i) \tag{53}$$

where the product has at most two terms. Each of the averages over the edge potentials, i.e.,

$$\sum_{x_{h\cap e}} P(x_{h\cap e}) \log \phi_e(x_e) \tag{54}$$

can be computed in time at most $\mathcal{O}(m^2)$ where m the number of values the variables can take. The complexity of evaluating the first part of the lower bound is thus no more than $\mathcal{O}(n_e m^2)$ where n_e is the number of edges in the graph. Note that this does not depend on the clique size. For example, the graph could be fully connected.

The entropy of the mean field distribution is simply the sum of entropies of the component distributions

$$H(P) = \sum_i H(P_i) = \sum_i \left[-\sum_{x_i} P(x_i) \log P(x_i) \right] \tag{55}$$

and takes only $\mathcal{O}(n_h m)$ to compute, where n_h is the number of hidden variables.

2.2.2 Optimization of the lower bound

Here we want to update each of the component distributions $P_i(x_i)$ to obtain the tightest possible lower bound on the desired marginal probability $P_0(x_v)$.

It is important to note in this context that while the maximum entropy distribution is unique, the optimization problem in the space of completely factorized distributions has local maxima. This is because the set of completely factorized distributions is not convex:

$$P'(x_h) = p \prod_i P_i^{(1)}(x_i) + (1-p) \prod_i P_i^{(2)}(x_i) \tag{56}$$

is no longer completely factorized in general. Thus we cannot expect to find the best possible factorized distribution, one that would yield the tightest bound. We can nevertheless incrementally and monotonically optimize P starting from any initial guess.

To derive the update equations for the component distributions, we will make use of the following concise notation for averages with respect to the mean field distribution:

$$\langle f(x_h) \rangle = \sum_{x_h} \left[\prod_{i \in h} P_i(x_i) \right] f(x_h) \tag{57}$$

So, for example, we can write

$$\sum_{x_{h \cap e}} P(x_{h \cap e}) \log \phi_e(x_e) \tag{58}$$

simply as $\langle \log \phi_e(x_e) \rangle$. Note that this does not involve a summation over all the configurations of the hidden variables but only those that $\log \phi_e(x_e)$ depends on (the summation over the remaining variables simply gives 1). We assume that the summations are carried out efficiently in this way. The complexity of each average is thus governed by the number of hidden variables the quantity of interest (e.g. $\log \phi_e(x_e)$) depends on.

We also define a partial average

$$\langle f(x_h) \rangle_j = \sum_{x_h \backslash x_j} \left[\prod_{i \neq j} P_i(x_i) \right] f(x_h) \tag{59}$$

Note that since we are not averaging over a particular variable x_j, $\langle f(x_h) \rangle_j$ is in general a function of x_j. Moreover, by design it is not a function of the component distribution $P_j(x_j)$. A very convenient property of the partial average is that

$$\sum_{x_j} P_j(x_j) \langle f(x_h) \rangle_j = \langle f(x_h) \rangle \tag{60}$$

- In the new notation, we can write the variational lower bound as

$$\begin{aligned} \log P_0(x_v) &\geq \sum_e \langle \log \phi_e(x_e) \rangle + \sum_i H(P_i) && (61) \\ &= \sum_{x_j} P(x_j) \sum_e \langle \log \phi_e(x_e) \rangle_j + H(P_j) + \sum_{i \neq j} H(P_i) && (62) \end{aligned}$$

where we have explicated how the bound depends on a particular component distribution $P_j(x_j)$. Note that from the point of view of P_j this bound is in the "maximum entropy form". Thus we can maximize the lower bound with respect to each P_j while keeping the other component distributions fixed. This leads to setting P_j equal to the max entropy distribution:

$$P_j(x_j) \leftarrow \frac{1}{Z_j} e^{\sum_e \langle \log \phi_e(x_e) \rangle_j} \tag{63}$$

where Z_j normalizes the distribution over x_j.

Note a few things:

1. In evaluating the exponent $\sum_e \langle \log \phi_e(x_e) \rangle_j$ we have to include only those terms $\langle \log \phi_e(x_e) \rangle_j$ that depend on x_j. Otherwise they are constants, and would be cancelled by the normalization anyway.
2. Each update of the component distribution must increase the overall lower bound. This is a very useful property in practice and also helps debugging.

3. After updating another component distribution, the optimal setting of $P_j(x_j)$ may have changed. The updates need to be performed repeatedly. Repeating the same update before updating another component is useless, however, since the update was the optimal solution in the case where the other components were assumed fixed.
4. When none of the component distributions change as a result of the updates (or in practice change very little), we have found a locally optimal solution. This solution is characterized by the fixed point equations:

$$P_j(x_j) = \frac{1}{Z_j}\, e^{\sum_e \langle \log \phi_e(x_e)\rangle_j} \tag{64}$$

for all $j \in h$. These fixed point equations are known as *mean field equations*.
5. All locally optimal solutions satisfy the mean field equations. These equations thus characterize a set of solutions rather than just one. Which particular solution we arrive at after successively updating the component distributions depends on the initial setting of the component distributions as well as on the order in which the component distributions were updated. Note that there are no guarantees that the locally optimal solutions are all equally good.

- Note finally that all the derivations go through equally well if instead of pairwise potentials we have any undirected model where the potential functions depend on arbitrary sets of variables c:

$$P_0(x_v, x_h) = \prod_{c \in C} \phi_c(x_c) \tag{65}$$

The complexity of applying mean field approximation in this case is governed by the number of hidden variables contained in each x_c. In other words, if the variables take m values then the cost is on the order of $\mathcal{O}(\sum_c m^{|h \cap c|})$

2.3 Directed graphs

The mean field approximation for directed graphical models is very similar. Consider the original distribution P_0 of the form

$$P_0(x_v, x_h) = \prod_i P_0(x_i | x_{pa_i}) \tag{66}$$

where we have again tacitly divided the variables into visible and hidden in the right hand side expression.

The lower bound corresponding to the completely factorized variational distribution P is

$$\log P_0(x_v) \geq \sum_i \langle \log P_0(x_i|x_{pa_i}) \rangle + \sum_{i \in h} H(P_i) \tag{67}$$

The component distributions can be updated in the same way as before giving

$$P_j(x_j) = \frac{1}{Z_j}\, e^{\sum_i \langle \log P_0(x_i|x_{pa_i}) \rangle_j} \tag{68}$$

where the sum over i in the exponent can be limited to j and the children of j. The variables contained in these conditional probabilities define the *markov blanket* for x_j. Note that x_j is independent of all the other variables in the model given those in its markov blanket.

References

Parisi (1988). Statistical field theory.

Cover and Thomas (1996). Elements of information theory.

Levin and Tribus (eds.) (1978). The maximum entropy formalism.

Saul, Jaakkola, Jordan (1996). Mean field theory for sigmoid belief networks. Availabe from http://www.cs.washington.edu/research/jair/abstracts/saul96a.html

Peterson and Anderson (1987). A mean field theory learning algorithm for neural networks. Complex Systems 1, 995-1019.