Cluster variation method

Bert Kappen
SNN University of Nijmegen, Netherlands

June 2013
A calculus for reasoning

Probabilities as

- frequencies (in ensemble or repeated trials)

\[
p(ab|c) = p(a|bc)p(b|c) \quad p(a|b) + p(\bar{a}|b) = 1
\]

- reasonable expectation.

  - Many proposals
  - Probabilistic approach correct (Cox, 1948)

A consequence is Bayes’ rule:

\[
p(c, e) = p(e|c)p(c)
\]

\[
p(c|e) = \frac{p(e|c)p(c)}{p(e)}, \quad p(e) = \sum_c p(e|c)p(c)
\]
Graphical models

Probabilistic approach to reasoning:

\[ p(A, S, T, L, B, E, X, D) = p(A)p(S)p(T|A) \]
\[ p(L|S)p(B|S)p(E|T, L)p(X|E)p(D|E, B) \]

What are the node probabilities given evidence?
Junction tree algorithm

Convert the graph to a tree of cliques.

Inference in trees is linear in the number of nodes.

Junction tree is exponential in the size of the largest clique.
Maximum Entropy

Let $x = (x_1, \ldots, x_n)$ be a set of discrete variables. Consider a probability distribution of the form

$$p_H(x) = \frac{\exp -H(x)}{Z}.$$ 

Special cases are DAGs with evidence:

$$p(x|e) = \frac{p(x, e)}{p(e)}, \quad p(x, e) = \prod_i p(x_i|\pi_i)$$

Define the free energy

$$F_H(p) = \sum_x p(x) H(x) + \sum_x p(x) \log p(x)$$

$p_H$ can be obtained by minimizing $F_H(p)$ with respect to $p(x)$ under the constraint of normalisation $\sum_x p(x) = 1$:

$$p_H(x) = \arg\min_{p(x)} F(p(x))$$
Free energy for Gaussians

Suppose \( p(x) \) is a Gaussian with mean zero and unit variance:

\[
p_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)
\]

The corresponding free energy is

\[
F(p) = \int dx \frac{x^2}{2} p(x) + \int dx p(x) \log p(x)
\]
Cluster variation method

The cluster variation method replaces the probability distribution $p_H(x)$ by a large number of (possibly overlapping) probability distributions, each describing the interaction between a small number of variables.

$$p(x) \approx \{p_\alpha(x_\alpha), \alpha = 1, \ldots\}$$

Due to the one-to-one correspondence between a probability distribution and the minima of a free energy we can define approximate probability distributions by constructing an approximate free energy

$$F(p) = \sum_x H(x)p(x) + \sum_x p(x) \log p(x) \approx F(\{p_\alpha\})$$

and compute its minimum, subject to normalization and consistency constraints.
Interactions and clusters

A given graphical model has interactions

\[ p(x) \propto \exp(-H(x)), \quad H(x) = \sum_{\alpha \in P} H^\dagger_\alpha(x_\alpha) \]

We choose clusters and all their intersections. For each cluster, we define a cluster probability \( p_\alpha(x_\alpha), \alpha \in U \). The expected energy can be expressed exactly in terms of \( p_\alpha(x_\alpha) \):

\[
\sum_x H(x)p(x) = \sum_{\alpha} \sum_{x_\alpha} H^\dagger_\alpha(x)p_\alpha(x) = \sum_{\alpha} \sum_{\beta \supseteq \alpha} a_\beta \sum_{x_\alpha} H^\dagger_\alpha(x_\alpha)p_\alpha(x_\alpha)
\]

\[
= \sum_{\beta} a_\beta \sum_{\alpha \subseteq \beta} \sum_{x_\alpha} H^\dagger_\alpha(x_\alpha)p_\alpha(x_\alpha) = \sum_{\beta} a_\beta \sum_{x_\beta} H_\beta(x_\beta)p_\beta(x_\beta)
\]

\[ H_\alpha(x_\alpha) = \sum_{\beta \supset \alpha} H^\dagger_\beta(x_\beta) \quad 1 = \sum_{\beta \supset \alpha} a_\beta, \quad \forall \alpha \]

the Moebius numbers.
Cluster entropy

The entropy contribution to the free energy can only be approximately expressed in terms of cluster probabilities:

\[ S = - \sum_x p(x) \log p(x) \approx \sum_{\alpha \in U} a_\alpha S_\alpha \]

The approximation is exact for trees.
Example: 2d grid

Consider a 2d grid, with outer clusters the plaquetes.

The overlaps are pairs \((ij)\) and single nodes \(i\).

The Moebius formula \(1 = \sum_{\beta \supseteq \alpha} a_{\beta}\) gives:

\[
\begin{align*}
\alpha &= \square \quad 1 = a_{\square} \\
\alpha &= (ij) \quad 1 = a_{\square} + a_{\square} + a_{(ij)} \quad a_{(ij)} = -1 \\
\alpha &= i \quad 1 = 4a_{\square} + 4a_{(ij)} + a_{i} \quad a_{i} = 1
\end{align*}
\]

The entropy is approximated as

\[
\begin{align*}
S &= \sum_{\square} S_{\square} - \sum_{(ij)} S_{(ij)} + \sum_{i} S_{i} \\
S_{\alpha} &= -\sum_{x_{\alpha}} p_{\alpha}(x_{\alpha}) \log p_{\alpha}(x_{\alpha})
\end{align*}
\]
CVM Free Energy

The approximate free energy becomes a function of the marginals $p_\alpha, \alpha \in U$:

$$F_{cvm} = \sum_{\alpha \in U} a_\alpha \sum_{x_\alpha} p_\alpha(x_\alpha) \log \frac{p_\alpha(x_\alpha)}{\psi_\alpha(x_\alpha)}$$

with $\psi_\alpha(x_\alpha) = \exp(-H_\alpha(x_\alpha))$, subject to

$$\sum_{x_\alpha} p_\alpha(x_\alpha) = 1$$

$$p_\alpha(x_\beta) = p_\beta(x_\beta), \quad \beta \subset \alpha$$

$$p_\alpha(x_\alpha) \geq 0$$

$F_{cvm}$ is a difference of entropy terms, ie. a non-convex function. The constraints are linear.
CVM Free Energy

\[ F_{\text{cvm}} = \sum_{\alpha \in U} a_\alpha \sum_{x_\alpha} p_\alpha(x_\alpha) \log \frac{p_\alpha(x_\alpha)}{\psi_\alpha(x_\alpha)} \]

\[ + \sum \lambda_\alpha \left( \sum_{x_\alpha} p_\alpha(x_\alpha) - 1 \right) + \sum_{\alpha \rightarrow \beta} \sum_{x_\beta} \lambda_{\alpha\beta}(x_\beta) \left( p_\alpha(x_\beta) - p_\beta(x_\beta) \right) \]

\[ p_\alpha(x_\alpha) = \frac{1}{Z_\alpha} \psi_\alpha(x_\alpha) \exp \left( \sum_{\beta \leftarrow \alpha} \lambda_{\alpha\beta}(x_\beta) \right), \quad \alpha \in B \]

\[ p_\beta(x_\beta) = \frac{1}{Z_\beta} \psi_\beta(x_\beta) \exp \left( -\frac{1}{a_\beta} \sum_{\alpha \rightarrow \beta} \lambda_{\alpha\beta}(x_\beta) \right), \quad \beta \in M \]
Lagrange multipliers

\[
\min_{x, g(x)=0} f(x) \iff \min_x \max_\lambda f(x) + \lambda g(x)
\]

\[
f(x) = x^2 + y^2 \quad g(x) = x + y + 0.4
\]
When $f$ and $g$ are convex, we can interchange $\max_\lambda$ and $\min_x$:

$$\min_x \max_\lambda L(x, \lambda) = \max_\lambda \min_x L(x, \lambda)$$

with $L(x, \lambda) = f(x) + \lambda g(x)$.

We can then solve the constrained optimisation problem by

- minimize wrt $x$, yielding $x^*(\lambda)$
- maximizing $L(x^*(\lambda), \lambda)$ wrt $\lambda$, yielding $\lambda^*$. This is equivalent to finding $\lambda$ such that $x^*(\lambda)$ satisfies the constraints:
  $$g(x^*(\lambda)) = 0.$$
- The solution is $x^*(\lambda^*)$. 
In general for non-convex functions this does not work:

$$\max_{\lambda} \min_x L(x, \lambda) \leq \min_x \max_{\lambda} L(x, \lambda)$$

Game theory: Two player can either play 'yes' or 'no'. The profit for $\lambda$ is the loss for $x$

<table>
<thead>
<tr>
<th>$\lambda \backslash x$</th>
<th>'yes'</th>
<th>'no'</th>
<th>$\min_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>'yes'</td>
<td>5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>'no'</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\max_{\lambda}$</td>
<td>10</td>
<td>6</td>
<td>6(\frac{5}{6})</td>
</tr>
</tbody>
</table>

$$\max_{\lambda} \min_x L(\lambda, x) \leq \min_x \max_{\lambda} L(\lambda, x) = 6 \quad (x \text{ first})$$

The optimal solution for $x$ depends on the order of playing.
Two approaches

$F_{cvm}(p)$ is not convex in $p$. There are two popular approaches:

**Ignore non-convexity**
1. Interchange $\min_p$ and $\max_\lambda$.
2. Compute $p(\lambda)$ by setting $\frac{\partial F_{cvm}}{\partial p} = 0$.
3. Find $\lambda$ such that $p$ satisfies all constraints through fixed point iteration. This method is known as (loopy) belief propagation.

This approach is exact when the graph has no loops.
The method also works surprisingly well with loops, but convergence is not guaranteed.

**Treat non-convexity**
This leads to the so-called double loop approach, which is guaranteed to converge, but can be slow.
Asia network

CVM: Clusters are defined according to the conditional probability tables.

Maximal error on the marginals is 0.0033. Maxerror on potentials: 0.006. Fixed point iteration requires 2 iterations.
Clusters are defined according to the conditional probability tables.

Maximal error on the marginals is 0.029. Maxerror on potentials: 0.048.

Fixed point iteration requires 8 iterations.

MF/TAP gives no reproducible results. BP gives 0.25 maximal error on marginals.
Random directed graphs

Each node is randomly connected to \( k = 5 \) parents. The entries of the probability tables are randomly generated between zero and one.

Clusters are defined according to the conditional probability tables. Therefore, maximal cluster size in \( B \) is \( k + 1 \).

Fixed point iteration does not converge on these problems for \( n \geq 10 \).

| \( n \) | \(|C| \) | Potential error | Margin error | Constraint error |
|-------|-------|-----------------|--------------|-----------------|
| 10    | 8     | 0.018           | 0.004        | 1.2e-50         |
| 20    | 12    | 0.018           | 0.006        | 9.0e-12         |
| 30    | 16    | 0.016           | 0.006        | 6.5e-12         |
| 40    | 21    | -               | -            | 6.8e-12         |

Table 1: \(|C|\) is the tree width of the triangulated graph required for the exact computation. Potential error and margin error are the maximum absolute distance (MAD) in any of the cluster probabilities and single variable marginals computed with CVM, respectively.
Network of binary (±1) variables with pair interactions. CVM clusters are all pairs. Left: ferromagnetic, Center: Anti-ferromagnetic, Right: SG. Local bias is zero. From top to bottom: (I) spectrum of the local BP stability matrix $F'$ at the trivial fixed point $\nu = 0$ and $\beta = 1$. (II) minimal eigenvalue of stability matrix at the paramagnetic solution, as a function of inverse temperature; (III) number of undamped, parallel BP iterations needed for convergence as a function of inverse temperature (dotted line in antiferromagnetic case shows the number of iterations for a sequential update scheme). The underlying graph $G$ is a random graph with Poissonian degree distribution, $N = 50$ and average degree $d = 4$; the local fields are zero.
Convergence of message passing. Ferromagnetic case.

Parallel BP loses stability when largest eigenvalue (positive) crosses unit circle in complex plane. At this point, Hessian of BP free energy at paramagnetic solution loses stability.
Convergence of message passing. Anti-Ferromagnetic case.

Parallel BP loses stability when largest eigenvalue (negative) crosses unit circle in complex plane. Sequential BP converges up to larger $\beta$. At this point, Hessian of BP free energy at paramagnetic solution has minimal stability.
Convergence of message passing. Spin glass case.

Parallel BP loses stability when largest eigenvalue crosses unit circle in complex plane. At this point, Hessian of BP free energy at paramagnetic solution has minimal stability.

Mooij 2008
Accuracy and cpu time

$N = 100$, $d = 3$ regular graph, pairwise SG interactions. External fields size 2.
Double loop approach

Bound $f(x)$ by a convex function:

$$f(x) \leq \tilde{f}_{x_0}(x) \quad f(x_0) = \tilde{f}_{x_0}(x_0)$$

Then, optimizing $\tilde{f}_{x_0}(x)$ wrt $x$ under constraints is a convex problem that can be solved, and

$$f(x_0) = \tilde{f}_{x_0}(x_0) \geq \tilde{f}_{x_1}(x_0) \geq f(x_1)$$

'Like EM!!'.

Bert Kappen 24
Pseudo code

1. (Initialization) Choose $x_0$ random.

2. (Outer loop) For $t = 1, \ldots$
   (a) Compute $f_{x_t}(x)$
   (b) (Inner loop) solve the convex problem

\[
\min_{x, g(x) = 0} f_{x_0}(x) \rightarrow x_{t+1}
\]
Example: 2d grid

Consider a 2d grid, with outer clusters the plaquetes.

The entropy is approximated as

\[ S = \sum S_{\square} - \sum S_{(ij)} + \sum S_i \]

\[ S_\alpha = -\sum_{x_\alpha} p_\alpha(x_\alpha) \log p_\alpha(x_\alpha) \]

Free energy contains \(-S'\):

\[ -S = -\sum S_{\square} + \sum S_{(ij)} - \sum S_i \leq -\sum S_{\square} + \sum S'_{(ij)} - \sum S_i \]

\[ S'_\alpha = -\sum_{x_\alpha} p_\alpha(x_\alpha) \log p'_\alpha(x_\alpha) \]

with \( p'_\alpha(x_\alpha) \) the point \( x_0 \).
CPU times (seconds) for the 2d Edwards-Anderson model in the paramagnetic phase (left) and ferromagnetic phase (right).