Age-Layered EM for Parameter Learning in Bayesian networks
Avneesh Saluja, Priya Krishnan Sundararajan & Ole J. Mengshoel

Slow EM Convergence

The most common way to learn parameters in a Bayesian network with hidden nodes is through the Expectation Maximization (EM) algorithm. However, since the algorithm converges to a local optimum, its performance suffers when a large number of local, non-global optima are present. Thus, multiple initializations are often done with EM.

There is also a significant difference between the average number of iterations for all runs and the average number of iterations for successful, or optimal, runs.

These graphs show the average number of iterations for 200 EM runs, across all hidden variable and sample size configurations for two Bayesian networks (Carstarts and Alarm). Notice the peaks at particular hidden variable configurations. We found that high average number of iterations over EM runs was more problematic than the relative likelihood shortfall.

Multiple starting points are expensive and wasteful. Is there a way to speed up the use of multiple initialization, using EM, with minimal degradation in solution quality?

ALEM Algorithm

Algorithm : ALEM (N, L, M)
procedure CHECKRUNS(Γ1, ρ1)
if |Γ1| > M1
then
Insert k EM
γ1 = γ1 ∪ {ρ1, ρ1, ..., ρ1}
end
Start traditional EM algorithm for each ρi = γ1
else
Randomly initialized
end

procedure ALEMCHECK(N, L, M)
comment: x denotes the number of terminated runs
x = 0
while x < N
for i = 1 to L
if |Γi| < M1 and |Γi| < M2
then
Start traditional EM algorithm for each ρi = Γi
else
Randomly initialized
end
end

ALEM Intuition and Experiments

- M trades off computational efficiency and solution quality. Higher M gives higher solution quality.
- A main assumption is that log likelihood comparison between similarly aged runs is a reliable indicator of comparisons at convergence.
- Evaluation on 2 Bayesian networks: Carstarts and Alarm. Sample sizes ranged from 100 to 4,000. Multiple hidden variable configurations were tested.
- Focused on problematic hidden variable configurations when comparing Traditional EM and Age-Layered EM (ALEM).

Results

Speedup as a function of hidden variables and sample size for the Carstarts & Alarm networks with the ALEM approach:

\[ speedup = \frac{\text{traditional iterations}}{\text{ALEM iterations}} \]

Highest speedups are in bold. Configurations with low speedups typically have, for Traditional EM, low average number of iterations.

Variation in the number of iterations ALEM runs undergo, on average, as a function of the minimum runs parameter M for both networks. The lower M is, the fewer iterations undergone. Black: M = 1, Red: M = 2, and Blue: M = 3.

Wall clock time comparison between traditional EM and ALEM. ALEM is, for larger sample sizes, significantly faster and the variation amongst runs tends to be much smaller. Traditional EM is in black, ALEM is in red.

Figure 1: Pseudocode for ALEM algorithm. The values for β, M, and c can be set depending on the nature of the Bayesian network. In our experiments, we have set w = 5, u = 1000, N = 500, M1 = 5 for the bottom layer, M2 = N for the top layer, M1 = 2 for L < 2 to L - 1 and c = 0.00001.