Appendix: Scaling Bayesian Network Parameter Learning with MapReduce and Age-Layered Expectation Maximization

Erik B. Reed
Carnegie Mellon University
NASA Research Park
Moffett Field, CA 94035
erikreed@cmu.edu

Ole J. Mengshoel
Carnegie Mellon University
NASA Research Park
Moffett Field, CA 94035
ole.mengshoel@sv.cmu.edu

1 Age Layered Expectation Maximization (ALEM) for Bayesian Network Parameter Learning

Consider a Bayesian Network (BN) $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\theta})$, where $\boldsymbol{X}$ are the nodes, $\boldsymbol{W}$ are the edges, and $\boldsymbol{\theta}$ are the parameters/CPTs. Let $\boldsymbol{E} \subset \boldsymbol{X}$ be the evidence nodes, and $\boldsymbol{e}$ the evidence. A BN factorizes a joint distribution $\Pr(\boldsymbol{X})$, and allows for different probabilistic queries to be formulated and supported by efficient algorithms; they all assume that all nodes in $\boldsymbol{E}$ are clamped to values $\boldsymbol{e}$. Computation of most probable explanation (MPE) amounts to finding a most probable explanation over the remaining nodes $\boldsymbol{R} = \boldsymbol{X} - \boldsymbol{E}$, or MPE($\boldsymbol{e}$). Computation of marginals (or beliefs) amounts to inferring the posterior probabilities over one or more query nodes $\boldsymbol{Q} \subseteq \boldsymbol{R}$, specifically BEL($Q, \boldsymbol{e}$) where $Q \in \boldsymbol{Q}$. Marginals may be used directly or used to compute most likely values (MLVs) simply by picking, in BEL($Q, \boldsymbol{e}$), a most likely state.

The Expectation Maximization (EM) algorithm can be summarized as follows:

1. Initialize parameters $\boldsymbol{\theta}^{(0)}$
2. **E-step:** Using parameters $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{E}$, generate the likelihood $\ell^{(t)}$ for the hidden nodes $\boldsymbol{R}$.
3. **M-step:** Modify the parameters to $\boldsymbol{\theta}^{(t+1)}$ to maximize the data likelihood.
4. While $|\ell^{(t)} - \ell^{(t-1)}| > \epsilon$, where ϵ is the tolerance, go to 2.

To formalize ALEM: let $\boldsymbol{L}$ be a set of k layers $\boldsymbol{L} = \{L_1, L_2, ..., L_k\}$, where L_i is a set of EM runs, where each layer has $R_j = [0, k]$ EM runs, $\sum_j R_j = k$. L_{ij} denotes jth EM run in layer i. Each layer L_i has an age limit $\beta_i \in \mathbb{N}$, which determines the maximum number of iterations. When an EM run L_{ij} reaches the maximum number of iterations, it ascends to the next layer. That is, L_{ij} is removed from L_i and put in L_{i+1}. The number of iterations of an EM run L_{ij} is denoted $\eta(L_{ij})$ parameter (log) likelihood of is denoted by $\ell(L_{ij})$. Consequently, $\beta_i \geq \eta(L_{ij})$ for $\forall i \forall j$. β is assigned to be an exponential function $\beta_i = \alpha 2^{i-1}$ for $\forall i \in [1, k-1]$, where α is the *Age Gap*, a constant influencing the maximum number of iterations between layers, or age difference. The max number of iterations for the last layer is s: $\beta_k = \omega$, where ω is equivalent to the max number of iterations in standard EM.

Each layer L_i also has a maximum number of runs $M_i \in \mathbb{N}$ for $\forall i \in [1, k]$. The maximum runs of the lowest layer M_1 is the initial population when ALEM initializes. When EM runs reach the maximum number of iterations for their layer β_i, they move to layer L_{i+1}, which can result in competition if their are more than M_{i+1} EM runs in L_{i+1}. When this occurs, the best likelihoods remain: the EM run with the lowest likelihood is removed from L_{i+1}. This has been termed as ALEM culling. That is:

$$L_{i+1} = \{L_{i+1} - L_{i+1, \arg\min_j \ell(L_{i+1,j})}\}$$

With the introduction of ω, we note there are three ways in which an EM run L_{ij} can terminate in ALEM:

1. L_{ij} reaches the maximum number of iterations ω i.e. $\eta(L_{ij}) = \omega$
2. The likelihood of L_{ij} has changed by an amount less than ϵ from the previous iteration. i.e. $|\ell^{(t)}(L_{ij}) - \ell^{(t-1)}(L_{ij})| \leq \epsilon$
3. ALEM culling: if $\|L_i\| > M_i$ and $L_{ij} = L_{i,\arg\min_j \ell(L_{i+1,j})}$