Algorithm 19.3 The structural EM algorithm for structure learning

Procedure Structural-EM (
  $\mathcal{G}^0$, // Initial bayesian network structure over $X_1, \ldots, X_n$
  $\boldsymbol{\theta}^0$, // Initial set of parameters for $\mathcal{G}^0$
  $\mathcal{D}$ // Partially observed data set
)
1 **for** each $t = 0, 1 \ldots,$ until convergence
2 // Optional parameter learning step
3 $\boldsymbol{\theta}^{t'} \leftarrow$ Expectation-Maximization$(\mathcal{G}^t, \boldsymbol{\theta}^t, \mathcal{D})$
4 // Run EM to generate expected sufficient statistics for $\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}}$
5 $\mathcal{G}^{t+1} \leftarrow$ Structure-Learn$(\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}})$
6 $\boldsymbol{\theta}^{t+1} \leftarrow$ Estimate-Parameters$(\mathcal{D}^*_{\mathcal{G}^t, \boldsymbol{\theta}^{t'}}, \mathcal{G}^{t+1})$
7 **return** $\mathcal{G}^t, \boldsymbol{\theta}^t$

However, we must take care in interpreting this guarantee. Assume that we have already modified $\mathcal{G}_0$ in several ways, to obtain a new graph $\mathcal{G}$. Now, we are considering a new operator o, and are interested in determining whether that operator is an improvement; that is, we wish to estimate the delta-score: $\text{score}_{BIC}(o(\mathcal{G}) : \mathcal{D}) - \text{score}_{BIC}(\mathcal{G} : \mathcal{D})$. The theorem tells us that if $o(\mathcal{G})$ satisfies $\text{score}_{BIC}(o(\mathcal{G}) : \mathcal{D}^*_{\mathcal{G}_0, \tilde{\boldsymbol{\theta}}_0}) > \text{score}_{BIC}(\mathcal{G}_0 : \mathcal{D}^*_{\mathcal{G}_0, \tilde{\boldsymbol{\theta}}_0})$, then it is necessarily better than our original graph $\mathcal{G}_0$. However, it does not follow that if $\hat{\delta}_{\mathcal{D}^*_{\mathcal{G}_0, \tilde{\boldsymbol{\theta}}_0}}(\mathcal{G} : o) > 0$, then $o(\mathcal{G})$ is necessarily better than $\mathcal{G}$. In other words, we can verify that each of the graphs we construct improves over the graph used to construct the completed data set, but not that each operator improves over the previous graph in the sequence. Note that we are guaranteed that our estimate is a true lower bound for any operator applied directly to $\mathcal{G}_0$. Intuitively, we believe that our estimates are likely to be reasonable for graphs that are "similar" to $\mathcal{G}_0$. (This intuition was also the basis for some of the heuristics described in section 19.4.2.2.) However, as we move farther away, our estimates are likely to degrade. Thus, at some point during our search, we probably want to select a new graph and construct a more relevant complete data set.

structural EM

This observation suggests an EM-like algorithm, called *structural EM*, shown in algorithm 19.3. In structural EM, we iterate over a pair of steps. In the E-step, we use our current model to generate (perhaps implicitly) a completed data set, based on which we compute expected sufficient statistics. In the M-step, we use these expected sufficient statistics to improve our model. The biggest difference is that now our M-step can improve not only the parameters, but also the structure. (Note that the structure-learning step also reestimates the parameters.) The structure learning procedure in the M-step can be any of the procedures we discussed in section 18.4, whether a general-purpose heuristic search or an exact search procedure for a specialized subset of networks for which we have an exact solution (for example, a maximum weighted spanning tree procedure for learning trees). If we use the BIC score, theorem 19.10 guarantees that, if this search procedure finds a structure that is better than the one we used in the previous iteration, then the structural EM procedure will monotonically improve the score.