A Proofs

Proof of Theorem 1: As the set $\mathcal{BN}_{\Theta\succeq\gamma}(G)$ is uncountably infinite, we cannot simply apply the standard techniques for PAC-learning a finite hypothesis set. We can, however, partition this uncountable space into a finite number $L = L(K, \gamma, \epsilon)$ of sets, such that any two BNs within a partition have similar conditional log-likelihood scores. We can then, in essense, simultaneously estimate the scores of all members of $\mathcal{BN}_{\Theta\succeq\gamma}(G)$ if we collect enough query instances to estimate the score for one representative of each partition.

Now for the details: We prove below that, if the CPtables for two BNs $\Theta^{(1)}, \Theta^{(2)} \in \mathcal{BN}_{\Theta\succeq\gamma}(G)$ have similar CPtables $\Theta^{(1)} = \{\theta^{(1)}_{d_i|\mathbf{f}_i}\}_i$ and $\Theta^{(2)} = \{\theta^{(2)}_{d_i|\mathbf{f}_i}\}_i$, then they will have similar LCL-scores wrt any query; *i.e.*,

$$\text{if} \quad \left|\theta^{(1)}_{d_i|\mathbf{f}_i} - \theta^{(2)}_{d_i|\mathbf{f}_i}\right| \leq \frac{\gamma\,\epsilon}{6\,K} \quad \text{then} \quad \forall c, \mathbf{e} \;\; |\ln(P_{\Theta^{(1)}}(\,c\,|\,\mathbf{e}\,)) \, - \, \ln(P_{\Theta^{(2)}}(\,c\,|\,\mathbf{e}\,))| \leq \frac{\epsilon}{6}\,. \tag{1}$$

This of course implies the same bound on the difference between their overall LCL-scores

$$|\mathrm{LCL}_k(\,\Theta^{(1)}\,) - \mathrm{LCL}_k(\,\Theta^{(2)}\,)| \quad \leq \quad \frac{\epsilon}{6}$$

for any distribution $\mathrm{LCL}_k(\,\cdot\,)$ — both for the "true" query distribution $\mathrm{LCL}(\,\cdot\,)$, and for the distribution associated with any empirical sample $\widehat{\mathrm{LCL}}(\,\cdot\,)$.

We therefore partition the $\mathcal{BN}_{\Theta\succeq\gamma}(G)$ space into $L = (\frac{6\,K}{\gamma\,\epsilon})^K$ disjoint sets (where any two BNs from any partition will have similar CPtable values), then define the set $R = \{\Theta_i\}_i$ to contain one representative from each partition. We prove below that a sample S of size

$$M\left(\frac{\epsilon}{6}, \frac{\delta}{L}\right) \quad = \quad 2\left(\frac{3N\log\gamma}{\epsilon}\right)^2 \ln\frac{2L}{\delta} \tag{2}$$

is sufficient to estimate each of these single representatives to within $\epsilon/6$ of correct, with probability of error at most δ/L; *i.e.*, such that, for each i,

$$P\left[\,\left|\widehat{\mathrm{LCL}}^{(S)}(\,\Theta_i\,) - \mathrm{LCL}(\,B_i\,)\right| > \frac{\epsilon}{6}\,\right] \quad < \quad \frac{\delta}{L}\,.$$

As there are L representatives, we have a total probability of at most $L\frac{\delta}{L} = \delta$ that *any* of the representative's scores are mis-estimated by more than $\epsilon/6$.

This means we have, in effect, estimated the scores on *any* $\Theta \in \mathcal{BN}_{\Theta\succeq\gamma}(G)$ to within $\epsilon/2$: For any $\Theta \in \mathcal{BN}_{\Theta\succeq\gamma}(G)$, let $\Theta' \in R$ be the representative in Θs partition. Observe

$$\begin{array}{rcccccc}
|\widehat{\mathrm{LCL}}(\,\Theta\,) - \mathrm{LCL}(\,\Theta\,)| \leq & |\widehat{\mathrm{LCL}}(\,\Theta\,) - \widehat{\mathrm{LCL}}(\,\Theta'\,)| & + & |\widehat{\mathrm{LCL}}(\,\Theta'\,) - \mathrm{LCL}(\,\Theta'\,)| & + & |\mathrm{LCL}(\,\Theta'\,) - \mathrm{LCL}(\,\Theta\,)| \\
\leq & \epsilon/6 & + & \epsilon/6 & + & \epsilon/6 \\
= & \epsilon/2\,. & & & &
\end{array}$$

This means, in particular, that our estimate of the scores of both $\widehat{\Theta}$ and Θ^* are within $\epsilon/2$, and so

$$\begin{array}{rcccccc}
\mathrm{LCL}(\,\widehat{\Theta}\,) - \mathrm{LCL}(\,\Theta^*\,) \leq & |\mathrm{LCL}(\,\widehat{\Theta}\,) - \widehat{\mathrm{LCL}}(\,\widehat{\Theta}\,)| & + & \widehat{\mathrm{LCL}}(\,\widehat{\Theta}\,) - \widehat{\mathrm{LCL}}(\,\Theta^*\,) & + & |\widehat{\mathrm{LCL}}(\,\Theta^*\,) - \mathrm{LCL}(\,\Theta^*\,)| \\
\leq & \epsilon/2 & + & 0 & + & \epsilon/2
\end{array}$$

To complete the proof, we need only prove Equations 1 and 2. For Equation 1: Consider the sequence of BNs $\Theta_0, \Theta_1, \ldots, \Theta_K$ where the first i of Θ_i's CPtables come from $\Theta^{(1)}$, and the remaining from $\Theta^{(2)}$ — *i.e.*,

$$\Theta_i \quad \sim \quad \{\theta^{(1)}_{d_1|\mathbf{f}_1}, \,\ldots,\, \theta^{(1)}_{d_i|\mathbf{f}_i}, \,\theta^{(2)}_{d_{i+1}|\mathbf{f}_{i+1}}, \,\ldots,\, \theta^{(2)}_{d_K|\mathbf{f}_K}\}\,.$$

Now observe

$$|\ln(P_{\Theta^{(1)}}(\,c\,|\,\mathbf{e}\,)) - \ln(P_{\Theta^{(2)}}(\,c\,|\,\mathbf{e}\,))| \quad \leq \quad \sum_{i=1}^{K} |\ln(P_{\Theta_i}(\,c\,|\,\mathbf{e}\,)) - \ln(P_{\Theta_{i-1}}(\,c\,|\,\mathbf{e}\,))|\,,$$

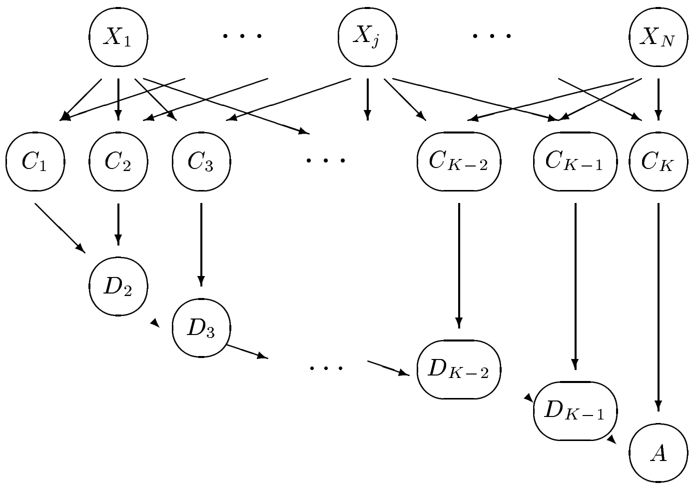

Figure 1: Belief Net structure corresponding to arbitrary SAT problem [Coo90]

and each $|\ln(P_{\Theta_i}(\,c\,|\,\mathbf{e}\,)) \;-\; \ln(P_{\Theta_{i-1}}(\,c\,|\,\mathbf{e}\,))|$ is based on changing a single CPtable entry. We therefore need only show $|\ln(P_{\Theta_i}(\,c\,|\,\mathbf{e}\,)) - \ln(P_{\Theta_{i-1}}(\,c\,|\,\mathbf{e}\,))| \;\le\; \frac{\epsilon}{6\,K}$. For any value of $z = \theta_{d_i|\mathbf{f}_i}$, let $f(z) = \ln(P_{\Theta[z]}(\,c\,|\,\mathbf{e}\,))$, where $\Theta[z]$ be the BN whose first $i-1$ CPtable entries come from $\Theta^{(1)}$, whose final $K-i-1$ entries come from $\Theta^{(2)}$, and whose i^{th} CPtable entries is z; hence $f(\theta^{(1)}_{d_i|\mathbf{f}_i}) = \ln(P_{\Theta_i}(\,c\,|\,\mathbf{e}\,))$, and $f(\theta^{(2)}_{d_i|\mathbf{f}_i}) = \ln(P_{\Theta_{i+1}}(\,c\,|\,\mathbf{e}\,))$. As this function is continuous, we know that

$$|f(a) - f(b)| \quad = \quad \frac{\partial f(z)}{\partial z}[b-a]$$

for some $z \in [a,b]$. As $f(z) = \ln(P_{\Theta[z]}(\,c,\mathbf{e}\,)) - \ln(P_{\Theta[z]}(\,\mathbf{e}\,))$, we see that

$$\begin{aligned}
\tfrac{\partial f(z)}{\partial z} &= \tfrac{1}{P_{\Theta[z]}(\,c,\mathbf{e}\,)} P_{\Theta[z]}(\,c,\mathbf{e}\,|\,d_i,\mathbf{f}_i\,) \times P_{\Theta[z]}(\,\mathbf{f}_i\,) \;-\; \tfrac{1}{P_{\Theta[z]}(\,\mathbf{e}\,)} P_{\Theta[z]}(\,\mathbf{e}\,|\,d_i,\mathbf{f}_i\,) \times P_{\Theta[z]}(\,\mathbf{f}_i\,) \\
&= \tfrac{1}{z}[P_{\Theta[z]}(\,d_i,\mathbf{f}_i\,|\,c,\mathbf{e}\,) - P_{\Theta[z]}(\,d_i,\mathbf{f}_i\,|\,\mathbf{e}\,)]
\end{aligned}$$

which means that $|\frac{\partial f(z)}{\partial z}| \;\le\; 1/z \;\le\; 1/\gamma$. (The second inequality follows from the assumption that we are only considering $\Theta \in \mathcal{BN}_{\Theta\succeq\gamma}(G)$.) Hence,

$$\begin{aligned}
|\ln(P_{\Theta_{i+1}}(\,c\,|\,\mathbf{e}\,)) - \ln(P_{\Theta_i}(\,c\,|\,\mathbf{e}\,))| &= |f(\theta^{(2)}_{d_i|\mathbf{f}_i}) - f(\theta^{(1)}_{d_i|\mathbf{f}_i})| \\
&\le \tfrac{1}{\gamma} \times |\theta^{(2)}_{d_i|\mathbf{f}_i} - \theta^{(1)}_{d_i|\mathbf{f}_i}| \quad \le \quad \tfrac{1}{\gamma} \times \tfrac{\gamma\,\epsilon}{6\,K} \quad = \quad \tfrac{\epsilon}{6\,K}\,.
\end{aligned}$$

To prove Equation 2: Observe first that the probability of any event must be at least the product of N CPtable entries, and hence $P_\Theta(\,c\,) \ge \gamma^N$ for any c and any $\Theta \in \mathcal{BN}_{\Theta\succeq\gamma}(G)$. This means the value of $-\ln(P_\Theta(\,c\,|\,\mathbf{e}\,))$, and hence $\mathrm{LCL}_{sq}(\,\Theta\,)$ for any distribution sq, is between 0 and $-N\ln\gamma$.

As the queries $q = P(\,c\,,\,\mathbf{e}\,)$ are drawn at random from a stationary distribution, we can view the quantity $\ln P_\Theta(\,q\,)$ as an iid random value, whose range is $[0,\,-N\ln\gamma]$ and whose expected value is $\mathrm{LCL}(\,\Theta\,)$. Hoeffding's Inequality bounds the chance that the empirical average score after M iid examples (here $\widehat{\mathrm{LCL}}^{(S)}(\,\Theta\,)$) will be far away from the true mean $\mathrm{LCL}(\,\Theta\,)$:

$$P(\,|\widehat{\mathrm{LCL}}^{(S)}(\,\Theta\,) - \mathrm{LCL}(\,\Theta\,)| > \frac{\epsilon}{6}\,) \quad < \quad 2\exp\left[-2M((\epsilon/6)/N\ln\gamma)^2\right]. \tag{3}$$

Here, we want the right-hand-side to be under δ/L, which requires $M = M(\epsilon,\delta) = 2\left(\frac{3\,N\,\ln\gamma}{\epsilon}\right)^2 \ln(\frac{2L}{\delta})$. ∎

Proof of Theorem 2: We reduce 3SAT to our task, using a construction similar to the one in [Coo90]: Given any 3-CNF formula $\varphi \equiv \bigwedge C_i$, where each $C_i \equiv \bigvee \pm X_{ij}$, we construct the network shown in Figure 1, with one node for each variable X_i and one for each clause C_j, with an arc from X_i to C_j whenever C_j involves X_i — *e.g.*, if

Table 1: Queries used in proof of Theorem 2

X_1	X_2	X_3	X_4	$\cdots$	X_n	A
0	1	0				0
0		0	1			0
	$\vdots$					$\vdots$
0		1		1		0
						1

$C_1 = x_1 \vee \neg x_2 \vee x_3$ and $C_2 = \neg x_1 \vee \neg x_3 \vee x_4$, then there are links to C_1 from each of X_1, X_2 and X_3, and to C_2 from X_1, X_3 and X_4. In addition, we include $K-1$ other boolean nodes, $\{D_2, \ldots, D_{K-1}, A\}$, where D_j is the child of D_{j-1} and C_j, where D_1 is identified with C_1, and A is used for D_K.

Here, we intend each C_i to be true if the assignment to the associated variables X_{i1}, X_{i2}, X_{i3} satisfies C_i; and A corresponds is the conjunction of those C_i variables. We do this using all-but-the-final instances in Table 1. (Note only 3 of the X_i variables are specified in each of these instances; the other $n-3$ X_is are not, nor are any C_js nor D_ks.) There is one such instance for each clause, with exactly the assignment (of the 3 relevant variables) that falsifies this clause. Hence, the first line corresponds to $C_1 \equiv x_1 \vee \neg x_2 \vee x_3$. The final instance is just stating that the prior value for A should $P(\,+a\,) = 1.0$. The "label" of each instance always corresponds to the single variable A.

We now prove, in particular, that

> There is a set of parameters for the structure in Figure 1, producing the $\widehat{\mathrm{LCL}}(\,\cdot\,)$-score, over the queries in Table 1, of 0
>
> *iff*
>
> there is a satisfying assignment for the associated φ formula.

⇐: Just set the CPtable for each C_i to be the disjunction of the associated X_{i1}, X_{i2}, X_{i3} variables (its parents), with the appropriate $\pm$ parity. *E.g.*, using $C_1 \equiv x_1 \vee \neg x_2 \vee x_3$, then C_1's CPtable would be

x_1	x_2	x_3	$P(\,+c_1 \mid x_1, x_2, x_3\,)$
0	0	0	1.0
0	0	1	1.0
0	1	0	0.0
0	1	1	1.0
1	0	0	1.0
1	0	1	1.0
1	1	0	1.0
1	1	1	1.0

Similarly set the CPtables for the D_j to correspond to the conjunction of its 2 parents $D_j = D_{j-1} \wedge C_j$; *e.g.*,

D_4	C_5	$P(\,+d_5 \mid D_4, C_5\,)$
0	0	0.0
0	1	0.0
1	0	0.0
1	1	1.0

.

Finally, set X_i to correspond to the satisfying assignment; *i.e.*, if $X_1 = 1$, then $\boxed{\dfrac{P(\,+x_1\,)}{1.0}}$; and if *i.e.*, if $X_4 = 0$, then $\boxed{\dfrac{P(\,+x_4\,)}{0.0}}$. Note that these CPtable values satify all $k+1$ of the labeled instances.

⇒: Here, we assume there is no satisfying assignment. Towards a contradiction, we can assume that there is a 0-LCL set of CPtable entries. This means, in particular, that $P(\,+a \mid x_{i1}, x_{i2}, x_{i3}\,) = 0$, where x_{i1}, x_{i2}, x_{i3} correspond to the assignment that violates the ith constraint. (*E.g.*, for $C_1 \equiv x_1 \vee \neg x_2 \vee x_3$, this would be $X_1 = 0$, $X_2 = 1$, $X_3 = 0$.)

Now consider the final labeled instance, $P(\, a\,)$. As there is no satisfying assignment, we know that each assignment $\mathbf{x}$ violates at least one constraint. For notation, let $\gamma^{\mathbf{x}}$ refer to one of these violations (say the one with the smallest index). So if $\mathbf{x} = \langle 0, 1, 0, \ldots \rangle$, then $\gamma^{\langle 0,1,0,\ldots\rangle} = \langle X_1 = 0,\ X_2 = 1,\ X_3 = 0 \rangle$ corresponds to the violation of the first constraint C_1. We also let $\beta^{\mathbf{x}}$ refer to the rest of the assignment.

Now observe

$$\begin{aligned} P(\,+a\,) &= \textstyle\sum_{\mathbf{x}} P(\,+a,\ \mathbf{x}\,) \\ &= \textstyle\sum_{\mathbf{x}} P(\,+a \,|\, \gamma^{\mathbf{x}}\,) \cdot P(\,\gamma^{\mathbf{x}}\,) \cdot P(\,\beta^{\mathbf{x}} \mid +a,\ \gamma^{\mathbf{x}}\,) \\ &= \textstyle\sum_{\mathbf{x}} \quad 0 \quad \cdot P(\,\gamma^{\mathbf{x}}\,) \cdot P(\,\beta^{\mathbf{x}} \mid +a,\ \gamma^{\mathbf{x}}\,) \quad = \quad 0\,, \end{aligned}$$

which shows that the final instance will be mislabeled. This proves that there can be no set of CPtable values that produce 0 LCL-score when there are no satisfying assignments. ∎

Proof of Proposition 3: Below, we will use $P(\,\chi\,)$ to refer to $P_\Theta(\,\chi\,)$, the value the belief net with parameters Θ will assign to the χ event. In general, for any assignment Z,

$$P(\,Z\,) \quad = \quad \sum_{\mathbf{f}'} \sum_{d'} P(\,Z \,|\, D = d', \mathbf{F} = \mathbf{f}'\,)\, P(\,D = d' \,|\, \mathbf{F} = \mathbf{f}'\,)\, P(\,\mathbf{F} = \mathbf{f}'\,)\,. \tag{4}$$

As we assume the different CPtable rows are estimated independently, and $\mathbf{F}$ is the set of parents of D, this means

$$\frac{\partial\, P(\,Z\,)}{\partial\, \beta_{d|\mathbf{f}}} \quad = \quad \sum_{d'} P(\,Z \,|\, d', \mathbf{f}\,)\, \frac{\partial\, P(\,d' \,|\, \mathbf{f}\,)}{\partial\, \beta_{d|\mathbf{f}}} P(\,\mathbf{f}\,)\,.$$

Recalling $\theta_{d|\mathbf{f}} = P(\,d \,|\, \mathbf{f}\,) = e^{\beta_{d|\mathbf{f}}} / \sum_{d'} e^{\beta_{d'|\mathbf{f}}}$, observe that $\frac{\partial\, P(\,d\,|\,\mathbf{f}\,)}{\partial\, \beta_{d|\mathbf{f}}} = \theta_{d|\mathbf{f}}(1 - \theta_{d|\mathbf{f}})$, and when $d \neq d'$, $\frac{\partial\, P(\,d'\,|\,\mathbf{f}\,)}{\partial\, \beta_{d|\mathbf{f}}} = -\theta_{d|\mathbf{f}}\, \theta_{d'|\mathbf{f}}$. This means $\frac{\partial\, P(\,Z\,)}{\partial\, \beta_{d|\mathbf{f}}} = P(\,Z, d, \mathbf{f}\,) - \theta_{d|\mathbf{f}} P(\,Z, \mathbf{f}\,)$.

Hence, as $\ln P(\,c \,|\, \mathbf{e}\,) = \ln P(\,c, \mathbf{e}\,) - \ln P(\,\mathbf{e}\,)$,

$$\begin{aligned} \frac{\partial\, \ln P(\,c \,|\, \mathbf{e}\,)}{\partial\, \beta_{d|\mathbf{f}}} &= \frac{\partial\, \ln P(\,c, \mathbf{e}\,)}{\partial\, \beta_{d|\mathbf{f}}} - \frac{\partial\, \ln P(\,\mathbf{e}\,)}{\partial\, \beta_{d|\mathbf{f}}} \\ &= \frac{1}{P(\,c, \mathbf{e}\,)} \frac{\partial\, P(\,c, \mathbf{e}\,)}{\partial\, \beta_{d|\mathbf{f}}} - \frac{1}{P(\,\mathbf{e}\,)} \frac{\partial\, P(\,\mathbf{e}\,)}{\partial\, \beta_{d|\mathbf{f}}} \\ &= \frac{1}{P(\,c, \mathbf{e}\,)} [P(\,c, \mathbf{e}, d, \mathbf{f}\,) - \theta_{d|\mathbf{f}} P(\,c, \mathbf{e}, \mathbf{f}\,)] - \frac{1}{P(\,\mathbf{e}\,)} [P(\,\mathbf{e}, d, \mathbf{f}\,) - \theta_{d|\mathbf{f}} P(\,\mathbf{e}, \mathbf{f}\,)] \\ &= [P(\,d, \mathbf{f} \,|\, c,\ \mathbf{e}\,) - P(\,d, \mathbf{f} \,|\, \mathbf{e}\,)] - \theta_{d|\mathbf{f}} [P(\,\mathbf{f} \,|\, c,\ \mathbf{e}\,) - P(\,\mathbf{f} \,|\, \mathbf{e}\,)]\,. \end{aligned}$$

∎

References

[Coo90] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, 1990.