Supplementary Material: Variational Training for Large-Scale Noisy-OR Bayesian Networks

Geng Ji[1,2] **Dehua Cheng**[2] **Huazhong Ning**[2,3] **Changhe Yuan**[2,4]
Hanning Zhou[2] **Liang Xiong**[2] **Erik B. Sudderth**[1]

[1]UC Irvine [2]Facebook AI [3]WeRide.ai [4]CUNY Queens College

A VARIATIONAL BOUND IS CONCAVE IN r

For each node $i \in \{\mathcal{H} \cup \mathcal{O}^+\}$ of some document d, the subset of terms in the variational bound of Eq. (9) that depend on auxiliary variables r_i can be written as:

$$\mathcal{L}_{di}(r_i) = \sum_{k \in \mathcal{P}(i)} r_{k\to i} q_k \big[f\left(u_{k\to i}\right) - f(w_{0\to i}) \big].$$

The first partial derivative of this variational bound is

$$\frac{\partial \mathcal{L}_{di}}{\partial r_{k\to i}} = q_k \Big(f\left(u_{k\to i}\right) - f(w_{0\to i}) - \frac{w_{k\to i}}{r_{k\to i}} f'(u_{k\to i}) \Big),$$

and its second partial derivatives equal

$$\frac{\partial^2 \mathcal{L}_{di}}{\partial r_{k\to i} \partial r_{\ell\to i}} = 0, \quad \frac{\partial^2 \mathcal{L}_{di}}{\partial r_{k\to i}^2} = q_k \frac{w_{k\to i}^2}{r_{k\to i}^3} f''(u_{k\to i}).$$

Here, the function

$$f''(a) = \frac{-\exp(a)}{\big(\exp(a) - 1\big)^2} < 0$$

is the second derivative of $f(a)$. Thus on the convex set of auxiliary parameters defined by Eq. (7), the (diagonal) Hessian matrix of $\mathcal{L}_{di}$ is negative definite, and $\mathcal{L}_{di}(r_i)$ is a strictly concave function of r_i.

B INITIALIZATION OF r

We show that setting $r_{k\to i} \propto w_{k\to i}$ globally optimizes our variational objective whenever the activation probabilities q_k for all parent nodes $k \in \mathcal{P}(i)$ are equal. To prove this, note that optimizing Eq. (9) with respect to $r_{k\to i}$ is equivalent to maximizing

$$\sum_{k \in \mathcal{P}(i)} r_{k\to i} \Big[f\Big(w_{0\to i} + \frac{w_{k\to i}}{r_{k\to i}} \Big) - f(w_{0\to i}) \Big]. \tag{B.1}$$

Given the non-negativity and normalization constraints in Eq. (7), we can apply Jensen's inequality in the opposite direction of typical variational derivations:

$$\begin{aligned} &\sum_{k \in \mathcal{P}(i)} r_{k\to i} \Big[f\Big(w_{0\to i} + \frac{w_{k\to i}}{r_{k\to i}} \Big) - f(w_{0\to i}) \Big] \\ &\leq f\Big(\sum_{k \in \mathcal{P}(i)} r_{k\to i} \Big(w_{0\to i} + \frac{w_{k\to i}}{r_{k\to i}} \Big) \Big) - f(w_{0\to i}) \\ &= f\Big(w_{0\to i} + \sum_{k \in \mathcal{P}(i)} w_{k\to i} \Big) - f(w_{0\to i}). \end{aligned} \tag{B.2}$$

The bound in the second line of Eq. (B.2) is achieved with equality if and only if $w_{0\to i} + \frac{w_{k\to i}}{r_{k\to i}}$ is constant for all parent nodes, which occurs when $r_{k\to i} \propto w_{k\to i}$.