On the Asymptotic Optimality of Maximum Margin Bayesian Networks

Sebastian Tschiatschek
Graz University of Technology, Austria

Franz Pernkopf
Graz University of Technology, Austria

1 MM PARAMETER LEARNING BY LINEAR PROGRAMMING

The convex optimization problem for learning MM parameters (Guo et al., 2005) is based on a relaxation of the normalization constraints inherent to learning probability distributions.

We exploit the renormalization property and combine ideas of Guo et al. (2005) and Pernkopf et al. (2012) to come up with a linear program for optimally learning MMBNs using the objective of Pernkopf et al. (2012). Assume a fixed graph $\mathcal{G}$ of the BNs. Then, we can express the joint probability $\mathrm{P}^{\mathcal{B}}(C, \mathbf{X})$ in (4) satisfying the independence properties implied by $\mathcal{G}$ as $\mathrm{P}^{\mathcal{B}}(C, \mathbf{X}) = \exp(\boldsymbol{\phi}(C, \mathbf{X})^T \mathbf{w})$. In this expression $\mathbf{w}$ is a vector containing the logarithms of all the conditional probabilities in the network, i.e. $w^i_{j|\mathbf{h}} = \log(\mathrm{P}(X_i = j | Pa(X_i) = \mathbf{h}))$, and $\boldsymbol{\phi}(c, \mathbf{x})$ is a binary vector indicating which entries of the log conditional probabilities $\log \mathrm{P}(X_i | Pa(X_i))$ are to be summed up when $C = c$ and $\mathbf{X} = \mathbf{x}$. Inserting the above expression into (14), the objective for learning MMBNs becomes

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \sum_{c,\mathbf{x}} \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) \min \Big(\gamma, \boldsymbol{\phi}(c, \mathbf{x})^T \mathbf{w} \qquad \text{(S1)} \\ & \qquad - \max_{c' \neq c} \boldsymbol{\phi}(c', \mathbf{x})^T \mathbf{w} \Big), \\ \text{s.t.} \quad & \sum_j \exp\left(w^i_{j|\mathbf{h}}\right) = 1 \quad \forall i, \mathbf{h} \in \mathrm{sp}(Pa(X_i)), \end{aligned}$$

where optimization is now solved over the log-parameters $\mathbf{w}$. The constraints ensure that $\mathbf{w}$ represents properly normalized conditional probabilities.

Problem (S1) can readily be expressed as the optimization problem

$$\begin{aligned} \underset{\mathbf{w}, \gamma_{(c,\mathbf{x})}}{\text{maximize}} \quad & \sum_{c,\mathbf{x}} \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) \gamma_{(c,\mathbf{x})} \qquad \text{(S2)} \\ \text{s.t.} \quad & \gamma_{(c,\mathbf{x})} \leq \gamma \quad \forall c, \mathbf{x} \\ & \gamma_{(c,\mathbf{x})} \leq [\boldsymbol{\phi}(c, \mathbf{x}) - \boldsymbol{\phi}(c', \mathbf{x})]^T \mathbf{w} \\ & \qquad \forall c, \mathbf{x}\, \forall c' \neq c \\ & \sum_j \exp\left(w^i_{j|\mathbf{h}}\right) = 1 \quad \forall i, \mathbf{h} \in \mathrm{sp}(Pa(X_i)). \end{aligned}$$

This problem is nonlinear and non-convex. To achieve convexity, Guo et al. (2005) relaxed the normalization constraints[1] to

$$\sum_j \exp\left(w^i_{j|\mathbf{h}}\right) \leq 1 \quad \forall i, \mathbf{h} \in \mathrm{sp}(Pa(X_i)). \qquad \text{(S3)}$$

These relaxed constraints have the disadvantage to cancel the effect of the margin-controlling parameter γ: To see this, consider the problem in (S2). If the normalization constraints are neglected, a linear program results. The dual of this linear program exhibits that its solutions in terms of $\mathbf{w}$ are independent of γ. However, every solution of the linear program can be transformed to a feasible solution of (S2) by subtracting a sufficiently large quantity from each component of $\mathbf{w}$. This subtraction does not change the objective and the induced classifier.

To achieve the desired effect of γ, we constrain the components of $\mathbf{w}$ to be smaller than 0 and use an ℓ_1-norm constraint on $\mathbf{w}$. The resulting linear program is

$$\begin{aligned} \underset{\mathbf{w}, \gamma_{c,\mathbf{x}}}{\text{maximize}} \quad & \sum_{c,\mathbf{x}} \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}) \gamma_{(c,\mathbf{x})} \qquad \text{(S4)} \\ \text{s.t.} \quad & \gamma_{(c,\mathbf{x})} \leq \gamma \quad \forall c, \mathbf{x} \\ & \gamma_{(c,\mathbf{x})} \leq [\boldsymbol{\phi}(c, \mathbf{x}) - \boldsymbol{\phi}(c', \mathbf{x})]^T \mathbf{w} \\ & \qquad \forall c, \mathbf{x}\, \forall c' \neq c \\ & - \sum_{i,j,\mathbf{h}} w^i_{j|\mathbf{h}} \leq 1, \quad \mathbf{w} \leq 0. \end{aligned}$$

[1] Guo et al. (2005) used a different objective function. However, the implications are the same.

A parameter vector $\mathbf{w}^*$ solving (S4) will in general not represent a properly normalized distribution. Whenever the renormalization Condition 1 is satisfied, normalization is possible without changing $h_{\mathbf{w}^*}$, where $h_{\mathbf{w}^*}$ is the classifier induced by $\mathbf{w}^*$. Roughly speaking, normalization can be achieved as follows (details are provided in (Wettig et al., 2003)): Due to the directed acyclic graph $\mathcal{G}$ assumed for the BN classifiers, the nodes of these classifiers can be topologically ordered. The conditional probabilities of the nodes can be sequentially normalized in a bottom up manner starting with the last node in the topological ordering. Multiplicative factors required for normalization are handed to the parent nodes. This does not affect the normalization of previous nodes.

If Condition 1 is not satisfied, the parameters can still be normalized. However, the resulting parameters are not guaranteed to maximize (14).

2 OPTIMAL MMBNS FOR THE THREE-CLASS EXAMPLE

2.1 Review of the Example

Consider a classifier with no features, i.e. $\mathbf{X} = \emptyset$, in a three-class scenario. Let the true distribution be defined by

$$\begin{aligned}\mathrm{P}^*(C=1) &= 0.4,\\ \mathrm{P}^*(C=2) &= 0.3, \text{and}\\ \mathrm{P}^*(C=3) &= 0.3.\end{aligned}$$

Hence, the Bayes optimal classifier would classify all instances as belonging to class 1. In this case however, any distribution inducing a Bayes optimal classifier has strictly smaller (larger) objective than the uniform distribution according to problem (14) (problem (17)). Consequently, any MM distribution induces an inconsistent classifier almost surely. In the remainder of this section, we assume that $\mathrm{P}^{\mathcal{T}}(C=1) > \mathrm{P}^{\mathcal{T}}(C=2)$, $\mathrm{P}^{\mathcal{T}}(C=1) > \mathrm{P}^{\mathcal{T}}(C=3)$ and $\mathrm{P}^{\mathcal{T}}(C=1) < \mathrm{P}^{\mathcal{T}}(C=2) + \mathrm{P}^{\mathcal{T}}(C=3)$ which holds asymptotically a.s.

2.2 Optimality of the Solution According to Guo et al.

For the considered example, MMBNs according to the formulation by Guo et al. can be found by minimizing

$$\begin{aligned}&\frac{1}{2\gamma^2} + BN\Bigg(\qquad\qquad (S5)\\ &\mathrm{P}^{\mathcal{T}}(C=1)\max\{0, \gamma - \min\{w_1 - w_2, w_1 - w_3\}\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=2)\max\{0, \gamma - \min\{w_2 - w_1, w_2 - w_3\}\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=3)\max\{0, \gamma - \min\{w_3 - w_1, w_3 - w_2\}\}\Bigg)\end{aligned}$$

with respect to w_1, w_2, w_3 and γ, under the constraints that $\gamma \geq 0$ and $\exp(w_1) + \exp(w_2) + \exp(w_3) \leq 1$.

Assuming a solution corresponding to the uniform distribution, i.e. $w_1 = w_2 = w_3 = \log(\frac{1}{3})$, the minimization problem becomes

$$\frac{1}{2\gamma^2} + BN\gamma, \tag{S6}$$

again, subject to $\gamma \geq 0$ and $\exp(w_1) + \exp(w_2) + \exp(w_3) \leq 1$. The latter constraint is clearly satisfied. The optimal value of γ can be determined easily, resulting in an objective value of

$$\frac{1}{2\,(BN)^{-\frac{2}{3}}} + \sqrt[3]{B^2N^2}. \tag{S7}$$

We now show by lower-bounding the objective that for any parameters w_1, w_2, w_3 corresponding to the Bayes optimal classifier, the objective is strictly larger than (S7). Assume that $w_1 > w_2$, $w_1 > w_3$ and without loss of generality, that $w_2 \geq w_3$ (this parameters correspond to a Bayes optimal classifier). Then the following chain of inequalities results:

$$\begin{aligned}&\frac{1}{2\gamma^2} + BN\Bigg(\\ &\mathrm{P}^{\mathcal{T}}(C=1)\max\{0, \gamma - \min\{w_1 - w_2, w_1 - w_3\}\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=2)\max\{0, \gamma - \min\{w_2 - w_1, w_2 - w_3\}\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=3)\max\{0, \gamma - \min\{w_3 - w_1, w_3 - w_2\}\}\Bigg)\\ =\ &\frac{1}{2\gamma^2} + BN\Bigg(\mathrm{P}^{\mathcal{T}}(C=1)\max\{0, \gamma - (w_1 - w_2)\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=2)\max\{0, \gamma - (w_2 - w_1)\}\\ &+ \mathrm{P}^{\mathcal{T}}(C=3)\max\{0, \gamma - (w_3 - w_1)\}\Bigg)\\ \overset{(a)}{\geq}\ &\frac{1}{2\gamma^2} + BN\Bigg(\mathrm{P}^{\mathcal{T}}(C=1)\max\{0, \gamma - (w_1 - w_2)\}\\ &+ \left(\mathrm{P}^{\mathcal{T}}(C=2) + \mathrm{P}^{\mathcal{T}}(C=3)\right)\max\{0, \gamma - (w_2 - w_1)\}\Bigg)\\ \overset{(b)}{\geq}\ &\frac{1}{2\gamma^2} + BN\Bigg(\mathrm{P}^{\mathcal{T}}(C=1)(\gamma - (w_1 - w_2))\\ &+ \left(\mathrm{P}^{\mathcal{T}}(C=2) + \mathrm{P}^{\mathcal{T}}(C=3)\right)(\gamma - (w_2 - w_1))\Bigg)\\ \overset{(c)}{>}\ &\frac{1}{2\gamma^2} + BN\gamma,\end{aligned} \tag{S8}$$

where (a) is because $w_2 - w_1 \leq w_3 - w_1$, (b) by selecting arbitrary elements instead of performing the maximum operations, and (c) because the empiric distribution satisfies $\mathrm{P}^{\mathcal{T}}(C=2) + \mathrm{P}^{\mathcal{T}}(C=3) > \mathrm{P}^{\mathcal{T}}(C=1)$ almost surely as $N \to \infty$. Consequently, an MMBN according to the formulation of Guo et al. must not be Bayes optimal for this example almost surely.

2.3 Optimality of the Solution According to Pernkopf et al.

For the considered example, MMBNs according to the formulation by Pernkopf et al. can be found by maximizing

$$\begin{aligned}&\mathrm{P}^{\mathcal{T}}(C=1)\min\left(\gamma, \log\theta_1 - \max\{\log\theta_2, \log\theta_3\}\right) \qquad \text{(S9)}\\ &+\mathrm{P}^{\mathcal{T}}(C=2)\min\left(\gamma, \log\theta_2 - \max\{\log\theta_1, \log\theta_3\}\right)\\ &+\mathrm{P}^{\mathcal{T}}(C=3)\min\left(\gamma, \log\theta_3 - \max\{\log\theta_1, \log\theta_2\}\right)\end{aligned}$$

with respect to $\theta_1, \theta_2, \theta_3$, where $\mathrm{P}^{\mathrm{MM}}(C=1) = \theta_1, \ldots, \mathrm{P}^{\mathrm{MM}}(C=3) = \theta_3$. In the case $\theta_1 = \theta_2 = \theta_3 = \frac{1}{3}$ the objective (S9) evaluates to zero.

We now show by calculation, that any $(\theta_1, \theta_2, \theta_3)$ that would correspond to a Bayes optimal classifier results in a strictly smaller objective. For this, assume that $\theta_1 > \theta_2$ and $\theta_1 > \theta_3$. Without loss of generality, additionally assume that $\theta_2 \geq \theta_3$. Consequently,

$$\begin{aligned}&\mathrm{P}^{\mathcal{T}}(C=1)\min\left(\gamma, \log\theta_1 - \max\{\log\theta_2, \log\theta_3\}\right)\\ &\quad+\mathrm{P}^{\mathcal{T}}(C=2)\min\left(\gamma, \log\theta_2 - \max\{\log\theta_1, \log\theta_3\}\right)\\ &\quad+\mathrm{P}^{\mathcal{T}}(C=3)\min\left(\gamma, \log\theta_3 - \max\{\log\theta_1, \log\theta_2\}\right)\\ &=\mathrm{P}^{\mathcal{T}}(C=1)\min\left(\gamma, \log\theta_1 - \log\theta_2\right)\\ &\quad+\mathrm{P}^{\mathcal{T}}(C=2)\min\left(\gamma, \log\theta_2 - \log\theta_1\right)\\ &\quad+\mathrm{P}^{\mathcal{T}}(C=3)\min\left(\gamma, \log\theta_3 - \log\theta_1\right)\\ &\overset{(a)}{\leq}\mathrm{P}^{\mathcal{T}}(C=1)\min\left(\gamma, \log\theta_1 - \log\theta_2\right)\\ &\quad+\left(\mathrm{P}^{\mathcal{T}}(C=2)+\mathrm{P}^{\mathcal{T}}(C=3)\right)\min\left(\gamma, \log\theta_2 - \log\theta_1\right)\\ &\overset{(b)}{<}\left(\mathrm{P}^{\mathcal{T}}(C=2)+\mathrm{P}^{\mathcal{T}}(C=3)\right)\min\left(\gamma, \log\theta_1 - \log\theta_2\right)\\ &\quad+\left(\mathrm{P}^{\mathcal{T}}(C=2)+\mathrm{P}^{\mathcal{T}}(C=3)\right)\min\left(\gamma, \log\theta_2 - \log\theta_1\right)\\ &=\left(\mathrm{P}^{\mathcal{T}}(C=2)+\mathrm{P}^{\mathcal{T}}(C=3)\right)\min\left(\gamma, \log\theta_1 - \log\theta_2\right)\\ &\quad-\left(\mathrm{P}^{\mathcal{T}}(C=2)+\mathrm{P}^{\mathcal{T}}(C=3)\right)\left(\log\theta_1 - \log\theta_2\right)\\ &\overset{(c)}{\leq} 0,\end{aligned}$$

where (a) is because $\theta_2 \geq \theta_3$ by assumption, (b) because $\mathrm{P}^{\mathcal{T}}(C=1) < \mathrm{P}^{\mathcal{T}}(C=2) + \mathrm{P}^{\mathcal{T}}(C=3)$, and (c) because $\log\theta_1 - \log\theta_2$ is bounded by γ. Hence, any MMBN must not be Bayes optimal for this example almost surely.

3 PROOF OF LEMMA 2

Proof. Similar to the proof of Lemma 1, we give a proof by contradiction and make the same assumptions. As the induced classifier $h_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ is assumed not to be optimal with respect to $\mathrm{P}^{\mathcal{T}}(C, \mathbf{X})$, there exists an instantiation of the features $\mathbf{x}^f$ that is not optimally classified by $h_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$, i.e. for which

$$[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})} \setminus [C|\mathbf{x}^f]_{\mathrm{P}^{\mathcal{T}}(C,\mathbf{X})} \neq \emptyset. \qquad \text{(S10)}$$

Because of the assumption of Lemma 2, the set $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathcal{T}}(C,\mathbf{X})}$ consists only of a single element. If $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ consists of a single element, a contradiction can be shown similar as in the binary-class case. If $[C|\mathbf{x}^f]_{\mathrm{P}^{\mathrm{MM}}(C,\mathbf{X})}$ consists of multiple elements, the sum

$$\sum_c \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}^f)\min\Bigg(\gamma, \log\mathrm{P}^{\mathrm{MM}}(c, \mathbf{x}^f) - \max_{c' \neq c}\log\mathrm{P}^{\mathrm{MM}}(c', \mathbf{x}^f)\Bigg) \qquad \text{(S11)}$$

in the MM-objective evaluates to at most zero.

Let $\{c^*\} = [C|\mathbf{x}^f]_{\mathrm{P}^{\mathcal{T}}(C,\mathbf{X})}$, i.e. c^* satisfies $\mathrm{P}^{\mathcal{T}}(c^*|\mathbf{x}^f) > \mathrm{P}^{\mathcal{T}}(c'|\mathbf{x}^f)$ for all $c' \neq c^*$. We generate a new distribution $\widetilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ that classifies $\mathbf{x}^f$ optimally and has higher objective. The distribution $\widetilde{\mathrm{P}}^{\mathrm{MM}}(C, \mathbf{X})$ is constructed from $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ by setting

$$\begin{aligned}\widetilde{\mathrm{P}}^{\mathrm{MM}}(c, \mathbf{x}) &= \mathrm{P}^{\mathrm{MM}}(c, \mathbf{x}) \qquad \forall \mathbf{x} \neq \mathbf{x}^f\ \forall c, &\text{(S12)}\\ \widetilde{\mathrm{P}}^{\mathrm{MM}}(c', \mathbf{x}^f) &= \frac{1}{|\mathrm{sp}(C)| - 1 + \exp(\gamma)} \cdot \mathrm{P}^{\mathrm{MM}}(\mathbf{x}^f)\\ &\qquad \forall c' \neq c^*, \text{ and} &\text{(S13)}\\ \widetilde{\mathrm{P}}^{\mathrm{MM}}(c^*, \mathbf{x}^f) &= \frac{\exp(\gamma)}{|\mathrm{sp}(C)| - 1 + \exp(\gamma)} \cdot \mathrm{P}^{\mathrm{MM}}(\mathbf{x}^f). &\text{(S14)}\end{aligned}$$

The terms in the objective that change their value, sum up to

$$\begin{aligned}&\sum_c \mathrm{P}^{\mathcal{T}}(c, \mathbf{x}^f)\min\Bigg(\gamma, \log\widetilde{\mathrm{P}}^{\mathrm{MM}}(c, \mathbf{x}^f) - \max_{c' \neq c}\log\widetilde{\mathrm{P}}^{\mathrm{MM}}(c', \mathbf{x}^f)\Bigg) \qquad \text{(S15)}\\ &= \gamma\left(\mathrm{P}^{\mathcal{T}}(c^*, \mathbf{x}^f) - \sum_{c' \neq c}\mathrm{P}^{\mathcal{T}}(c', \mathbf{x}^f)\right)\\ &> 0,\end{aligned}$$

where the inequality is due to the assumption of the Lemma. As the objective increases, $\mathrm{P}^{\mathrm{MM}}(C, \mathbf{X})$ is not an MMBN. □

References

Guo, Y., Wilkinson, D., and Schuurmans, D. (2005). Maximum margin Bayesian networks. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence*, pages 233–242. UAI Press.

Pernkopf, F., Wohlmayr, M., and Tschiatschek, S. (2012). Maximum margin Bayesian network classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):521–531.

Wettig, H., Grünwald, P., Roos, T., Myllymaki, P., and Tirri, H. (2003). When discriminative learning of Bayesian network parameters is easy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 491–496.