Bayesian Analysis (0000) **00**, Number 0, pp. 1

Learning to Anticipate Indirect Causes in Dynamic Bayesian Networks Digital Supplement

Alexander Motzek[*,†], Ralf Möller[*]

Abstract. Modeling causal dependencies often demands cycles at a coarse-grained temporal scale. If Bayesian networks are used for representing uncertainty in temporal knowledge bases, cycles are frequently eliminated with dynamic Bayesian networks over time, which, however, spreads indirect dependencies over time as well, and enforces an infinitesimal resolution of time. Recently it has been shown that if indirect dependencies are spread over time, spurious results are returned and models can not represent causalities correctly. As a solution, it has been shown that some dynamic Bayesian networks, called ADBNs, are able to resolve cyclic dependencies intrinsically by a rapid adaptation to specific contexts at every timestep. To learn such networks from long, but incomplete datastreams, parameter and structure learning of DBNs are fused into one atomic phase in this paper. We show that classic (dynamic) Bayesian networks are unable to learn an anticipation of indirect causes in certain domains, and that classic approaches are not applicable to learning ADBNs. We proposes a learning approach for ADBNs considering rapidly adapting structures, while preserving (A)DBNs as a first-class representation of uncertainty in temporal knowledge bases.

1 Extended Smoothing Problem

Theorem 1 (Exact solution to the extended smoothing problem)**.** *Given a complete smoothing problem* $\texttt{ExtdSP}(B_0, B_\rightarrow, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$*, finding an exact solution is linear in* t*. Finding an exact solution is exponential in the maximal dimension of unobserved variables* $\vec{\zeta}^*$*,* $\vec{\beta}^*$ *in a timestep* $0 < * \leq t$*, and in the largest domain* $\mathrm{dom}(\zeta_+)$*,* $\mathrm{dom}(\beta_+)$ *of all random variables* $\zeta_+ \in \vec{\zeta}^{0:t}$*,* $\beta_+ \in \vec{\beta}^{1:t}$*. Finding an exact solution is exponential in the dimension of number of random variables* $|\vec{X}^{t-1:t}|$*,* $|\vec{\mathcal{A}}^{t-1:t}|$ *and a respective maximal domain size* $\mathrm{dom}(X_+)$*,* $\mathrm{dom}(A_{++})$ *of all random variables* $X_+ \in \vec{X}^{t-1:t}$*,* $A_{++} \in \vec{\mathcal{A}}^{t-1:t}$*.* ▲

Theorem 1 is proven by showing that an algorithm exists that finds an exact solution to $\texttt{ExtdSP}(B_0, B_\rightarrow, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$ in time-complexity $\mathcal{O}(t \cdot |\mathrm{dom}(X_+)|^{|\vec{X}^t|} \cdot |\mathrm{dom}(A_{++})|^{|\vec{\mathcal{A}}^t|} \cdot |\mathrm{dom}(\zeta_+)|^{|\vec{\zeta}^*|} \cdot |\mathrm{dom}(\beta_+)|^{|\vec{\beta}^*|})$ and with $\mathcal{O}(|\mathrm{dom}(X_+)|^{|\vec{X}^{t-1:t}|} \cdot |\mathrm{dom}(A_{++})|^{|\vec{\mathcal{A}}^{t-1:t}|})$ space-complexity for storing one extended smoothing distributions.

[*]Universität zu Lübeck, Institut für Informationssysteme, Germany. motzek@ifis.uni-luebeck.de moeller@uni-luebeck.de.
[†]Corresponding author

 DOI: 0000

Proof of Theorem 1 (extended smoothing problem). An algorithm obtaining solution for an extended smoothing problem $\texttt{ExtdSP}(B_0, B_{\rightarrow}, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$ is similar to an algorithm derived for obtaining a solution to classical smoothing problems in DBNs. For the case of a dense intra-timeslice ADBN an exact solution to an extended smoothing problem is given by straight marginalization from a JPD as

$$P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal}, \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal}) = \alpha \cdot \sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} \sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal})$$

using the JPD of a dense intra-timeslice ADBN, one obtains

$$\begin{aligned} &P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal}, \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal}) \\ &\qquad = \alpha \cdot \sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} \sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{0:t-1^\intercal}, \vec{\mathcal{A}}^{1:t-1^\intercal}) \\ &\qquad\qquad \cdot \prod_{X_i^t \in \vec{X}^t} P(X_i^t | \vec{X}^{t^\intercal} \backslash X_i^t, A_i^{t^\intercal}, X_i^{t-1}) \cdot \prod_{A_{ij}^t \in \vec{\mathcal{A}}^t} P(A_{ij}^t) \ . \end{aligned} \tag{1.1}$$

Using an intermediate joint probability distribution definition, one obtains

$$\begin{aligned} &P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal}, \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal}) \\ &= \alpha \cdot \sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} \sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{0:k^\intercal}, \vec{\mathcal{A}}^{1:k^\intercal}) \cdot P(\vec{X}^{k+1:t^\intercal}, \vec{\mathcal{A}}^{k+1:t^\intercal}) \\ &= \alpha \cdot \sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} P(\vec{X}^{0:k^\intercal}, \vec{\mathcal{A}}^{1:k^\intercal}) \cdot \sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{k+1:t^\intercal}, \vec{\mathcal{A}}^{k+1:t^\intercal}) \\ &= \alpha \cdot \sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} P(\vec{X}^{0:k-1^\intercal}, \vec{\mathcal{A}}^{1:k-1^\intercal}) \cdot P(\vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal}) \\ &\qquad\qquad \cdot \sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{k+1:t^\intercal}, \vec{\mathcal{A}}^{k+1:t^\intercal}) \\ &= \alpha \cdot \left(\sum_{\vec{\zeta}^{0:k-2}} \sum_{\vec{\beta}^{1:k-2}} P(\vec{X}^{0:k-1^\intercal}, \vec{\mathcal{A}}^{1:k-1^\intercal}) \right) \cdot \left(P(\vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal}) \right) \\ &\qquad\qquad \cdot \left(\sum_{\vec{\zeta}^{k+1:t}} \sum_{\vec{\beta}^{k+1:t}} P(\vec{X}^{k+1:t^\intercal}, \vec{\mathcal{A}}^{k+1:t^\intercal}) \right) \\ &= \alpha \cdot P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal} | \vec{z}^{0:k^\intercal}, \vec{b}^{1:k^\intercal}) \cdot P(\vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal}) \cdot P(\vec{z}^{k+1:t^\intercal}, \vec{b}^{k+1:t^\intercal} | \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal}) \ . \end{aligned} \tag{1.2}$$

Evaluating the extended smoothing Equation 1.2 for all instantiations of $\vec{X}^{k-1}, \vec{\mathcal{A}}^{k-1}, \vec{X}^k, \vec{\mathcal{A}}^k$ decrementally for descending $k = t \ldots 0$ is an algorithm that gives an exact solution to the extended smoothing problem $\texttt{ExtdSP}(B_0, B_{\rightarrow}, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$. The decremental evaluation allows for storing the backward message: an intermediate result $P(\vec{z}^{k+1:t^\intercal}, \vec{b}^{k+1:t^\intercal} | \vec{X}^{k+1^\intercal}, \vec{\mathcal{A}}^{k+1^\intercal})$ from an evaluation of $P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal}, \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal})$ is needed in an upcoming evaluation

of $P(\vec{X}^{k-2^\intercal}, \vec{\mathcal{A}}^{k-2^\intercal}, \vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal})$. Thus, obtaining the last term of Equation 1.2 is constant in t for every evaluation. Obtaining the middle term $P(\vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal})$ is linear in the number of random variables of timeslice k.

The first term $P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal} | \vec{z}^{0:k-1^\intercal}, \vec{b}^{1:k-1^\intercal})$ of the extended smoothing equation poses a filtering problem, for which a solution is found in $\mathcal{O}(1)$ in a storage from a solution to an offline filtering problem $\texttt{OffFP}(B_0, B_\rightarrow, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$, which is found in $\mathcal{O}(t)$ by the derived algorithm for solving filtering problems in Motzek and Möller (2015).

For a fixed $B_0, B_\rightarrow$ and a fixed number and domain size of unobserved variables per timeslice in $\vec{z}^{1:t}, \vec{b}^{1:t}$ the algorithm has linear time-complexity $\mathcal{O}(t)$ and linear space-complexity $\mathcal{O}(t)$. It requires storage for one distribution $P(\vec{X}^{k-1^\intercal}, \vec{\mathcal{A}}^{k-1^\intercal}, \vec{X}^{k^\intercal}, \vec{\mathcal{A}}^{k^\intercal} | \vec{z}^{0:t^\intercal}, \vec{b}^{1:t^\intercal})$ and storage for a solution of $\texttt{OffFP}(B_0, B_\rightarrow, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$. □

2 Derivation of EM Procedure

We prove Theorem 2 by showing that the proposed procedure in fact maximizes the likelihood of the dataset, i.e., the probability of observing the dataset under the optimized parameters. To do so, we analytically solve Eq. 3 (given for reference below). We derive an ADBN EM procedure for the most general, i.e., dense, intra-timeslice ADBN, which encapsulates all possible intra-timeslice DBN structures and onto which all other intra-timeslice (A)DBNs are reducable.

Proof of Theorem 2 (EM procedure). Optimized parameters $\vec{\vartheta}^*$ are obtained by

$$\vec{\vartheta}^* = \arg\max_{\vec{\Theta}} P_{\vec{\Theta}}(\vec{d}) = \arg\max_{\vec{\Theta}} \log\left(P_{\vec{\Theta}}(\vec{d})\right) \ , \tag{2.1}$$

i.e., by definition maximize the likelihood of observing a dataset under a given parameter set. In the following, we explicitly derive one parameter.

One obtains

$$\log\left(P_{\vec{\Theta}}(\vec{d})\right) = \log\left(\sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot P_{\vec{\Theta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal})\right) \ ,$$

where Jensen's inequality for the concav function log, i.e.,

$$\log\left(\sum_i p_i x_i\right) \geq \sum_i p_i \log(x_i)$$

is applicable. Under Jensen's inequality one obtains

$$\log\left(P_{\vec{\Theta}}(\vec{d})\right) = \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \log\left(P_{\vec{\Theta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal})\right) \ , \tag{2.2}$$

as $\sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) = 1$, if the dataset yields a well-defined ADBN, i.e., if the dataset only contains regular instantiations. Using the definition of the JPD of a dense intra-timeslice ADBN (Eq. 1) yields

$$\log\Big(P_{\vec{\Theta}}(\vec{d})\Big) = \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \Big(\sum_{X_k^0 \in \vec{X}^0} \mathfrak{P}_\Theta(X_k^0) + \sum_{i=1}^{t} \big(\sum_{X_k^i \in \vec{X}^{\,i}} \mathfrak{P}_\Theta(X_k^i | \vec{X}^{i^\intercal} \backslash X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) + \sum_{A_{cv}^i \in \vec{\mathcal{A}}^i} \mathfrak{P}_\Theta(A_{cv}^i) \big) \Big) \ ,$$

with $\mathfrak{P}(\cdot) = \log P(\cdot)$ for brevity. $P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t})$ is a numerical value and values of random variables are uniquely identified by $\vec{d}^{0:t}$ or by summation over the unobserved variables in the dataset.

Under a stationary process, i.e., $P_\Theta(\cdot^i) = P_\Theta(\cdot^j) = P_\Theta(\cdot)$, an optimized parameter set $\vec{\vartheta}^*$ is obtainable in a closed form. Let $X_\lambda^t \in \vec{X}^t$ be some random variable and let $A_{\mu\nu}^t \in \vec{\mathcal{A}}^t$ be some activator. In the following, we extract one parameter $P_\Theta(x_\lambda^t | \vec{x}^{t^\intercal} \backslash x_\lambda^t, \vec{a}_\lambda^{t^\intercal}, x_\lambda^{t-1})$ from all sum-products to allow for partial derivation. First, one is able to represent the likelihood of data as

$$\begin{aligned}\log\Big(P_{\vec{\Theta}}(\vec{d})\Big) = &\sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \sum_{X_k^0 \in \vec{X}^0} \mathfrak{P}_\Theta(X_k^0) \\ &+ \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \sum_{i=1}^{t} \sum_{X_k^i \in \vec{X}^{\,i}} \mathfrak{P}_\Theta(X_k^i | \vec{X}^{i^\intercal} \backslash X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) \\ &+ \sum_{\vec{\zeta}^{0:t^\intercal}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \sum_{i=1}^{t} \sum_{A_{cv}^i \in \vec{\mathcal{A}}^i} \mathfrak{P}_\Theta(A_{cv}^i) \ ,\end{aligned}$$

factoring out further yields

$$\begin{aligned}\log\Big(P_{\vec{\Theta}}(\vec{d})\Big) = &\sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \sum_{X_k^0 \in \vec{X}^0} \mathfrak{P}_\Theta(X_k^0) \\ &+ \sum_{i=1}^{t} \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \sum_{X_k^i \in \vec{X}^{\,i}} \mathfrak{P}_\Theta(X_k^i | \vec{X}^{i^\intercal} \backslash X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) \\ &+ \sum_{i=1}^{t} \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \sum_{A_{cv}^i \in \vec{\mathcal{A}}^i} \mathfrak{P}_\Theta(A_{cv}^i) \ ,\end{aligned}$$

where one is able to explicitly represent the summation over all possible instantiations of X_λ^i by

$$\log\Big(P_{\vec{\Theta}}(\vec{d})\Big) = \sum_{\vec{\zeta}^{0:t}} \sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal} | \vec{d}^{0:t}) \cdot \sum_{X_k^0 \in \vec{X}^0} \mathfrak{P}_\Theta(X_k^0)$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\setminus X_\lambda^i}\sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, {}^{+}x_\lambda^i, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})$$

$$\cdot\left(\sum_{X_k^i\in\vec{X}^{\,i}\setminus X_\lambda^i}\mathfrak{P}_\Theta(X_k^i|\vec{X}^{i^\intercal}\setminus X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) + \mathfrak{P}_\Theta({}^{+}x_\lambda^i|\vec{X}^{i^\intercal}\setminus X_\lambda^i, \vec{A}_\lambda^{i^\intercal}, X_\lambda^{i-1})\right)$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\setminus X_\lambda^i}\sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \neg x_\lambda^i, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})$$

$$\cdot\left(\sum_{X_k^i\in\vec{X}^{\,i}\setminus X_\lambda^i}\mathfrak{P}_\Theta(X_k^i|\vec{X}^{i^\intercal}\setminus X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) + \mathfrak{P}_\Theta(\neg x_\lambda^i|\vec{X}^{i^\intercal}\setminus X_\lambda^i, \vec{A}_\lambda^{i^\intercal}, X_\lambda^{i-1})\right)$$

$$+\sum_{\vec{\zeta}^{0:t^\intercal}}\sum_{\vec{\beta}^{1:t}}\sum_{i=1}^{t} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})\cdot\sum_{A_{cv}^i\in\vec{\mathcal{A}}^i}\mathfrak{P}_\Theta(A_{cv}^i)\ ,$$

Note that an instantiation of X_λ^i might be part of $\vec{d}^{0:t}$ and therefore was not present in the previous summation, i.e., was not included in $\vec{\zeta}^{0:t^\intercal}$. Nevertheless, the extraction remains sound as a respective $P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, x_\lambda^i, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})$ is then 0. With $\gamma(X_\lambda = x_\lambda)$ as a shorthand for

$$\gamma(X_\lambda = x_\lambda) = \sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\setminus X_\lambda^i}\sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, x_\lambda^i, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})$$
$$\cdot\Big(\sum_{X_k^i\in\vec{X}^{\,i}\setminus X_\lambda^i}\mathfrak{P}_\Theta(X_k^i|\vec{X}^{i^\intercal}\setminus X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1}) + \mathfrak{P}_\Theta(x_\lambda^i|\vec{X}^{i^\intercal}\setminus X_\lambda^i, \vec{A}_\lambda^{i^\intercal}, X_\lambda^{i-1})\Big)\ .$$

one obtains

$$\log\left(P_{\vec{\Theta}}(\vec{d})\right) = \sum_{\vec{\zeta}^{0:t}}\sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})\cdot\sum_{X_k^0\in\vec{X}^0}\mathfrak{P}_\Theta(X_k^0)$$
$$+\gamma(X_\lambda = {}^{+}x_\lambda) + \gamma(X_\lambda = \neg x_\lambda) + \sum_{\vec{\zeta}^{0:t}}\sum_{\vec{\beta}^{1:t}}\sum_{i=1}^{t} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})\cdot\sum_{A_{cv}^i\in\vec{\mathcal{A}}^i}\mathfrak{P}_\Theta(A_{cv}^i)\ .$$

Classically, to learn a specific parameter, i.e., to find an optimized parameter $P_\Theta(x_\lambda^i|\vec{x}^{i^\intercal}\setminus x_\lambda^i, \vec{a}_\lambda^{i^\intercal}, x_\lambda^{i-1}) \in \vec{\Theta}^*$ it must be explicitly extracted from all (nested) summations s.t. one is able to solve Eq. 2.1 by a partial derivation. By extracting this parameter from all summations one obtains,

$$\gamma(X_\lambda = x_\lambda)$$
$$= \sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\setminus X_\lambda^i}\sum_{\vec{\beta}^{1:t}} P_{\vec{\vartheta}}(\vec{X}^{0:t^\intercal}, x_\lambda^i, \vec{\mathcal{A}}^{1:t^\intercal}|\vec{d}^{0:t})\sum_{X_k^i\in\vec{X}^{\,i}\setminus X_\lambda^i}\mathfrak{P}_\Theta(X_k^i|\vec{X}^{i^\intercal}\setminus X_k^i, \vec{A}_k^{i^\intercal}, X_k^{i-1})$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\vec{\beta}^{1:t}\setminus\vec{x}^{i\intercal},\vec{a}^{i}_{\lambda},x^{i-1}_{\lambda}} P_{\vec{\vartheta}}(\vec{X}^{0:t\intercal},x^{i}_{\lambda},\vec{\mathcal{A}}^{1:t\intercal}|\vec{d}^{0:t})\mathfrak{P}_{\Theta}(x^{i}_{\lambda}|\vec{X}^{i\intercal}\setminus X^{i}_{\lambda},\vec{A}^{i\intercal}_{\lambda},X^{i-1}_{\lambda})$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:i-2}}\sum_{\vec{\beta}^{1:i-2}}\sum_{\vec{\zeta}^{i+1:t}}\sum_{\vec{\beta}^{i+1:t}}\sum_{\vec{\zeta}^{i-1}\setminus X^{i-1}_{\lambda}}\sum_{\vec{\beta}^{i-1}}\sum_{\vec{\beta}^{i}\setminus\vec{A}^{i}_{\lambda}}$$

$$P_{\vec{\vartheta}}(\vec{X}^{0:i-2\intercal},\vec{X}^{i-1\intercal}\setminus X^{i-1}_{\lambda},x^{i-1}_{\lambda},\vec{\mathcal{A}}^{1:i-1\intercal},\vec{x}^{i},\vec{a}^{i}_{\lambda},\vec{\mathcal{A}}^{i\intercal}\setminus\vec{A}^{i}_{\lambda},\vec{X}^{i+1:t\intercal},\vec{\mathcal{A}}^{i+1:t\intercal}|\vec{d}^{0:t})$$

$$\cdot\,\mathfrak{P}_{\Theta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})\;, \quad (2.3)$$

where $\sum_{\vec{\zeta}^{0:t}\vec{\beta}^{1:t}\setminus\vec{x}^{i},\vec{a}^{i}_{\lambda},x^{i-1}_{\lambda}}$ represents the summation over all possible instantiations of all unobserved variables excluding a specific instantiation $\vec{x}^{i},\vec{a}^{i}_{\lambda},x^{i-1}_{\lambda}$. However, Equation 2.3 extracts a too general parameter $P_{\Theta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i}_{\lambda},x^{i-1}_{\lambda})$ and does not consider activator criteria, which are needed in order to assure well-definedness in the transformation towards Equation 2.2.

To learn an activation criteria aware parameter $P_{\Theta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})$, we use Definition 5 (black squares) and extract the parameter from all summations as

$$\gamma(X_{\lambda}=x_{\lambda})$$

$$=\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\setminus X^{i}_{\lambda}}\sum_{\vec{\beta}^{1:t}}P_{\vec{\vartheta}}(\vec{X}^{0:t\intercal},x^{i}_{\lambda},\vec{\mathcal{A}}^{1:t\intercal}|\vec{d}^{0:t})\sum_{X^{i}_{k}\in\vec{X}^{\,i}\setminus X^{i}_{\lambda}}\mathfrak{P}_{\Theta}(X^{i}_{k}|\vec{X}^{i\intercal}\setminus X^{i}_{k},\vec{A}^{i\intercal}_{k},X^{i-1}_{k})$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:t}\vec{\beta}^{1:t}\setminus x^{i}_{\lambda},\vec{x}^{i}_{\blacksquare\lambda},\vec{a}^{i}_{\lambda},x^{i-1}_{\lambda}}P_{\vec{\vartheta}}(\vec{X}^{0:t\intercal},x^{i}_{\lambda},\vec{\mathcal{A}}^{1:t\intercal}|\vec{d}^{0:t})\mathfrak{P}_{\Theta}(x^{i}_{\lambda}|\vec{X}^{i\intercal}\setminus X^{i}_{\lambda},\vec{A}^{i\intercal}_{\lambda},X^{i-1}_{\lambda})$$

$$+\sum_{i=1}^{t}\sum_{\vec{\zeta}^{0:i-2}}\sum_{\vec{\beta}^{1:i-2}}\sum_{\vec{\zeta}^{i+1:t}}\sum_{\vec{\beta}^{i+1:t}}\sum_{\vec{\zeta}^{i-1}\setminus X^{i-1}_{\lambda}}\sum_{\vec{\beta}^{i-1}}\sum_{\vec{\beta}^{i}\setminus\vec{A}^{i}_{\lambda}}\sum_{\vec{X}^{i}_{\square\lambda}}$$

$$P_{\vec{\vartheta}}(\vec{X}^{0:i-2\intercal},\vec{X}^{i-1\intercal}\setminus X^{i-1}_{\lambda},x^{i-1}_{\lambda},\vec{\mathcal{A}}^{1:i-1\intercal},\vec{X}^{i\intercal}_{\square\lambda},\vec{x}^{i\intercal}_{\blacksquare\lambda},$$

$$\vec{a}^{i}_{\lambda},\vec{\mathcal{A}}^{i\intercal}\setminus\vec{A}^{i}_{\lambda},\vec{X}^{i+1:t\intercal},\vec{\mathcal{A}}^{i+1:t\intercal}|\vec{d}^{0:t})\mathfrak{P}_{\Theta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})\;.$$

Then, an optimized parameter $P^{*}_{\vartheta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})\in\vec{\vartheta}^{*}$ according to Eq. 2.1 is explicitly obtained by partial derivation as

$$\frac{\delta P_{\vec{\Theta}}(\vec{d})}{\delta P_{\Theta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})}\left(P^{*}_{\vartheta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})\right)=0\;,$$

which, under a stationary process reduces to

$$P^{*}_{\vartheta}(x^{i}_{\lambda}|\vec{x}^{i\intercal}\setminus x^{i}_{\lambda},\vec{a}^{i\intercal}_{\lambda},x^{i-1}_{\lambda})=\frac{\gamma'(X_{\lambda}=x_{\lambda})}{\gamma'(X_{\lambda}=+x_{\lambda})+\gamma'(X_{\lambda}=\neg x_{\lambda})}\;,$$

with

$$\gamma'(X_\lambda = x_\lambda) = \sum_{i=1}^{t} \sum_{\vec{\zeta}^{0:i-2}} \sum_{\vec{\beta}^{1:i-2}} \sum_{\vec{\zeta}^{i+1:t}} \sum_{\vec{\beta}^{i+1:t}} \sum_{\vec{\zeta}^{i-1} \setminus X_\lambda^{i-1}} \sum_{\vec{\beta}^{i-1}} \sum_{\vec{\beta}^{i} \setminus \vec{A}_\lambda^{i}} \sum_{\vec{X}_{\square\lambda}^{i}}$$
$$P_{\vec{\vartheta}}(\vec{X}^{0:i-2\intercal}, \vec{X}^{i-1\intercal} \setminus X_\lambda^{i-1}, x_\lambda^{i-1}, \vec{\mathcal{A}}^{1:i-1\intercal}, \vec{X}_{\square\lambda}^{i\intercal}, \vec{x}_{\blacksquare\lambda}^{i\intercal},$$
$$\vec{a}_\lambda^{i}, \vec{\mathcal{A}}^{i\intercal} \setminus \vec{A}_\lambda^{i}, \vec{X}^{i+1:t\intercal}, \vec{\mathcal{A}}^{i+1:t\intercal} | \vec{d}^{0:t}) \ .$$

in which an extended smoothing problem $\texttt{ExtdSP}(B_0, B_\rightarrow, \vec{z}^{0:t}, \vec{b}^{1:t}, t)$ is evident and can be written as in Theorem 2. Thus, the proposed procedure generates parameter instantiations $\vec{\vartheta}^*$ of $\vec{\Theta}$ that maximize the likelihood of the dataset. □

The proof for activator parameters is equivalent.

References

Motzek, A. and Möller, R. (2015). "Indirect Causes in Dynamic Bayesian Networks Revisited." In *IJCAI 2015: 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, July 25-31, 2015*, 703–709. 3