Tractable Inference for Complex Stochastic Processes

Xavier Boyen
Stanford University
Computer Science Dept.
xb@cs.stanford.edu

Daphne Koller
Stanford University
Computer Science Dept.
koller@cs.stanford.edu

Technical report
1998/03/17

Abstract — The monitoring and control of any dynamic system depends crucially on the ability to reason about its current status and its future trajectory. In the case of a stochastic system, these tasks typically involve the use of a *belief state*—a probability distribution over the state of the process at a given point in time. Unfortunately, the state spaces of complex processes are very large, making an explicit representation of a belief state intractable. Even in dynamic Bayesian networks (DBNs), where the process itself can be represented compactly, the representation of the belief state is intractable. We investigate the idea of utilizing a compact approximation to the true belief state, and analyze the conditions under which the errors due to the approximations taken over the lifetime of the process do not accumulate to make our answers completely irrelevant. We show that the error in a belief state *contracts* exponentially as the process evolves. Thus, even with multiple approximations, the error in our process remains bounded indefinitely. We show how the additional structure of a DBN can be used to design our approximation scheme, improving its performance significantly. We demonstrate the applicability of our ideas in the context of a monitoring task, showing that orders of magnitude faster inference can be achieved with only a small degradation in accuracy.

1 Introduction

The ability to model and reason about stochastic processes is fundamental to many applications [Forbes *et al.*, 1995; Jensen *et al.*, 1989; Dagum *et al.*, 1992; Provan, 1992]. For example, we may be observing a freeway traffic scene via a video camera mounted on a bridge, with the goal of understanding the current traffic situation and predicting its future evolution [Forbes *et al.*, 1995], or monitoring a patient's symptoms to design his treatment [Provan, 1992].

A number of formal models have been developed for describing situations of this type, including Hidden Markov Models [Rabiner and Juang, 1986], Kalman Filters [Kalman, 1960], and Dynamic Bayesian Networks (DBNs) [Dean and Kanazawa, 1989]. These very different models all share the same underlying *Markov assumption*, the fact that the future is conditionally independent of the past given the current state. Since the domain is stochastic and partially observable, the true state of the process is rarely known with certainty. However, most reasoning tasks can be performed by using a *belief state*, which is a probability distribution over the state of a system at a given time [Aström, 1965]. It follows from the Markov assumption that the belief state at time t completely captures all of our information about the past. In particular, it suffices both for predicting the probabilities of future trajectories of the system, and for making optimal decisions about our actions.

Consider, for example, the task of monitoring an evolving system. Given a belief state at time t which summarizes all of our evidence so far, we can generate a belief state for time $t+1$ using a straightforward procedure: We propagate our current belief state through the *state evolution model*, resulting in a distribution over states at time $t+1$, and then condition that distribution on the observations obtained at time $t+1$, getting our new belief state.

The effectiveness of this procedure (as well as of many others) depends crucially on the representation of the belief state. Certain types of systems, e.g., Gaussian processes, admit a compact representation of the belief state and an effective update process (via *Kalman filtering* [Kalman, 1960]). However, in other cases matters are not so simple. Consider, for example, a stochastic system represented as a dynamic Bayesian network (DBN). A DBN (see Figure 1), like a Bayesian network, allows a decomposed representation of the state via state variables, and a compact representation of the probabilistic model by utilizing conditional independence assumptions. Here, a belief state is a distribution over some subset of the state variables at time t. In general, not all of the variables at time t must participate in the belief state [Dagum *et al.*, 1992]; however, (at least) every variable whose value at time t directly affects its value at time $t+1$ must be included. In large DBNs, the obvious representation of a belief state (as a flat distribution over its state space) is therefore typically infeasible, particularly in time-critical applications.

However, our experience with Bayesian networks has led us to the tacit belief that structure is usually synonymous with easy inference. Thus, we may expect that, here also, the structure of the model would support a decomposed representation of the distribution, and thereby much more effective inference. Unfortunately, this hope turns out to be unfounded. Consider the DBN of Figure 1(a), and assume to begin with that the evidence variable did

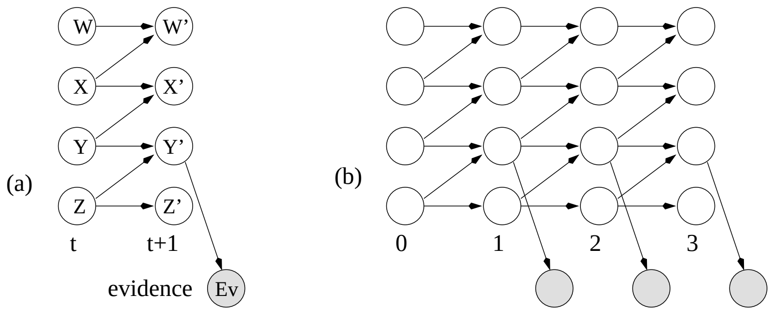

Figure 1: Example of a structured stochastic process: (a) two-time-slice DBN model; (b) resulting unrolled BN over the first three time slices.

not exist. At the initial time slice, all of the variables start out being independent. Furthermore, there are only a few connections between one variable and another. Nevertheless, at time 3, all the variables have become fully correlated: one can see from the unrolled DBN of Figure 1(b) that no conditional independence relation holds among any of them. Observing the evidence variable makes things even worse: in this case it takes only 2 time slices for all the state variables to become fully correlated. In general, unless a process decomposes into completely independent subprocesses,[1] the belief state will become fully correlated very early in time. As any factored decomposition of a distribution rests on some form of conditional independence, no decomposition of this belief state is possible. This phenomenon is perhaps the most serious impediment to applying probabilistic reasoning methods to dynamic systems.

The scope of this phenomenon seems unfair: had the W, X and Y, Z subprocesses in Figure 1 been completely independent, then the variables would have remained independent forever and a decomposed representation of the belief state would have been possible. Yet even a very weak dependency is enough to render this approach impossible. This observation leads naturally to the idea behind our work. Clearly, a weakly correlated process is not far from one with complete independence. In particular, if we start out with a decomposed belief state at time t (say one where all the variables are independent), and compute the resulting belief state at time $t+1$, that new belief state cannot be that far from being decomposed. We can therefore approximate it using one whose representation is decomposed, which allows us to continue this process to the next time slice.

This idea immediately suggests a new scheme for monitoring a stochastic process: We decide on some computationally tractable representation for an approximate belief state, e.g., one which decomposes into independent factors. We propagate the approximate belief state at time t through the transition model and condition it on our evidence at time $t+1$. In general, the resulting belief state for time $t+1$ will not fall into the class which we have

[1]Where independence implies not only that the state of one process does not influence the state of another, but also that two processes do not influence the same evidence variable.

chosen to maintain. We therefore approximate it using one that does, and continue. We note that this basic idea has many useful variants. For example, in processes with continuous state that do not fall under the umbrella of traditional Kalman filters, we may wish to approximate the belief state using a mixture of Gaussians with a fixed number of components.

This type of process must face the risk that our constant approximations cause errors to build up out of control over extended periods of time, either by accumulation due to the repeated approximations, or (worse) by spontaneous amplification due to some sort of instability. In this paper, we show that this problem does not occur: the mere stochasticity of the process serves to attenuate the effects of errors over time, preventing the accumulated error from growing unboundedly.

More precisely, we analyze the effect on the relative entropy distance of propagating two different distributions (e.g., the exact and approximate belief states) through the transition model. It is well known [Cover and Thomas, 1991] that the resulting distributions will be closer than the original ones. However, we also show that, under reasonable assumptions about the stochasticity of the process, this transition results in a *contraction* effect; i.e., each propagation through the transition model results in a constant factor reduction of the distance between two distributions. Our result is, (to our knowledge) the first quantitative and non-asymptotic analysis of the contraction behavior of Markov processes for relative entropy distance. Based on this contraction result, we show that the effect of errors due to previous approximations decreases exponentially, allowing us to prove that the overall error in our approximation remains bounded *indefinitely*.

Our approximation scheme applies to any discrete stochastic process and any approximate representation of the belief state. However, we show that even stronger results can be obtained in the case of structured processes and an approximation scheme that is tailored to the structure of the process. Specifically, we consider processes that are composed of several weakly interacting subprocesses, and an approximation scheme that decomposes each belief state as a product of independent belief states over the individual processes. Under these assumptions, our contraction rate improves dramatically; furthermore, there is a direct connection between the amount of interaction between the subprocesses and the quality of our bounds.

Finally, we return to the problem of approximate monitoring in complex processes. We show how our analysis can be used to help us select an appropriate representation for an approximate belief state. We also show how a DBN inference algorithm can be adapted to the task of maintaining an approximate belief state. We provide empirical evidence for the success of our approach on two practical DBNs, showing that we can achieve orders of magnitude faster inference with only a small degradation in accuracy.

2 Basic framework

Our focus in this paper is on discrete-time finite-state Markov processes. Such processes can be modeled explicitly, as a *Hidden Markov Model* or, if additional structure is present, more compactly as a *Dynamic Bayesian Network*. A discrete time Markov process evolves by moving from one state to the other at consecutive times points. We use $S^{(t)}$ with $S^{(t)} \in \boldsymbol{S} = \{s_1, ..., s_n\}$ to denote the state at time t. In the case of a DBN, $S^{(t)}$ may be described as an assignment of values to some set of state variables. The *Markov assumption*, inherent to all of the models we consider, asserts that the present state of the system contains enough information to make its future independent from its past, i.e., that $\boldsymbol{P}[S^{(t+1)} \mid S^{(0)}, ..., S^{(t)}] = \boldsymbol{P}[S^{(t+1)} \mid S^{(t)}]$. For simplicity, we also assume that the process is *time-invariant*, i.e., that the probability with which we have a transition from some state s_i at time t to another state s_j at time $t+1$ does not depend on t. Thus, we obtain that the process can be described via a *transition model* $\mathcal{T}$:

$$\mathcal{T}[s_i \rightsquigarrow s_j] \triangleq \boldsymbol{P}[s_j^{(t+1)} \mid s_i^{(t)}],$$

where we use $s_i^{(t)}$ to denote the event $S^{(t)} = s_i$. In the case of an HMM, $\mathcal{T}$ is often described explicitly as an $n \times n$ matrix; for a DBN, $\mathcal{T}$ is described more compactly as a fragment of a Bayesian network (see Section 5.2).

The Markov process is typically hidden, or *partially observable*, meaning that its state is not directly observable. Rather, we observe a *response* $R^{(t)} \in \boldsymbol{R} = \{r_1, ..., r_m\}$; in the case of a DBN, $R^{(t)}$ can be an assignment to some set of observable random variables. The response depends stochastically and exclusively on the state $S^{(t)}$; i.e., $R^{(t)}$ is conditionally independent of any $S^{(t')}$ and $R^{(t')}$ given $S^{(t)}$. Using $r_h^{(t)}$ to denote $R^{(t)} = r_h$, we obtain that the observability aspect of the process can be described via an *observation model* $\mathcal{O}$:

$$\mathcal{O}[s_i \hookrightarrow r_h] \triangleq \boldsymbol{P}[r_h^{(t)} \mid s_i^{(t)}].$$

The Markov assumption implies that all the historical information needed to monitor or predict the system's evolution is contained in (the available knowledge about) its present state. This knowledge can be summarized in a *belief state*—a probability distribution over the possible states. At each time point t, we distinguish between the *prior* and the *posterior* belief state, defined as follows:

Definition 1 The *prior belief state* at time t, denoted $\sigma^{(\bullet t)}$, is the distribution over the state at t given the response history up to but not including time t. Letting $r_{h_k}^{(k)}$ denote the response observed at time k,

$$\boldsymbol{\sigma}^{(\bullet t)}[s_i] \triangleq \boldsymbol{P}[s_i^{(t)} \mid r_{h_0}^{(0)}, ..., r_{h_{t-1}}^{(t-1)}].$$

The *posterior belief state* at time t, denoted $\sigma^{(t\bullet)}$, is the distribution over the state at time t given the response history up to and including time t:

$$\boldsymbol{\sigma}^{(t\bullet)}[s_i] \triangleq \boldsymbol{P}[s_i^{(t)} \mid r_{h_0}^{(0)}, ..., r_{h_{t-1}}^{(t-1)}, r_{h_t}^{(t)}].$$

The *monitoring* task is defined as the task of maintaining a belief state as time advances and new responses are observed. In principle, the procedure is quite straightforward. Assume we have a posterior belief state $\boldsymbol{\sigma}^{(t\bullet)}$ at time t. Upon observing the response r_h at time $t+1$, the new state distribution $\boldsymbol{\sigma}^{(t+1\bullet)}$ can be obtained via a two-stage computation, based on the two models $\mathcal{T}$ and $\mathcal{O}$. The prior belief state $\boldsymbol{\sigma}^{(\bullet t+1)}$ for the next time slice is obtained by propagating $\boldsymbol{\sigma}^{(t\bullet)}$ through the stochastic transition model, while the posterior $\boldsymbol{\sigma}^{(t+1\bullet)}$ is obtained by conditioning $\boldsymbol{\sigma}^{(\bullet t+1)}$ on the response r_h observed at time $t+1$:

$$\begin{aligned}
\boldsymbol{\sigma}^{(\bullet t+1)}[s_j] &= \sum_{i=1}^{n} \boldsymbol{\sigma}^{(t\bullet)}[s_i]\, \mathcal{T}[s_i \rightsquigarrow s_j]\ , \\
\boldsymbol{\sigma}^{(t+1\bullet)}[s_i] &= \frac{\boldsymbol{\sigma}^{(\bullet t+1)}[s_i]\, \mathcal{O}[s_i \hookrightarrow r_h]}{\sum_{l=1}^{n} \boldsymbol{\sigma}^{(\bullet t+1)}[s_l]\, \mathcal{O}[s_l \hookrightarrow r_h]}.
\end{aligned}$$

Abstractly, we can view $\mathcal{T}$ as a function mapping $\boldsymbol{\sigma}^{(t\bullet)}$ to $\boldsymbol{\sigma}^{(\bullet t+1)}$, and define $\mathcal{O}_{r_h}$ as the function mapping $\boldsymbol{\sigma}^{(\bullet t+1)}$ to $\boldsymbol{\sigma}^{(t+1\bullet)}$ upon observing the response r_h at time $t+1$. The one-step update rule for belief states is then given by: $\boldsymbol{\sigma}^{(t\bullet)} \overset{\mathcal{T}}{\mapsto} \boldsymbol{\sigma}^{(\bullet t+1)} \overset{\mathcal{O}_{r_h}}{\mapsto} \boldsymbol{\sigma}^{(t+1\bullet)}$, assuming r_h is observed at time $t+1$.

While exact monitoring is simple in principle, it can be quite costly. As we mentioned in the introduction, the belief state for a process represented compactly as a DBN is typically exponential in the number of state variables; it is therefore impractical in general even to feasibly store the belief state, far less to propagate it through the various update procedures described above.

Thus, we are interested in utilizing compactly represented approximate belief states in our inference algorithm. The risks associated with this idea are clear: the errors induced by our approximations may accumulate to make the results of our inference completely irrelevant. As we show in the next two sections, the stochasticity of the process prevents this problem from occurring.

3 Simple contraction

Consider the exact and estimated belief states $\boldsymbol{\sigma}^{(t\bullet)}$ and $\hat{\boldsymbol{\sigma}}^{(t\bullet)}$. Intuitively, as we propagate each of them through the transition model it "forgets" some of its information; and as the two distributions forget about their differences, they become closer to each other. As we will see in Section 5, in order for our errors to remain bounded, their effect needs to dampen exponentially quickly. That is, we need to show that $\mathcal{T}$ reduces the distance between two belief states $\boldsymbol{\sigma}^{(t\bullet)}$ and $\hat{\boldsymbol{\sigma}}^{(t\bullet)}$ by a constant factor.

In fact, this result is known for $\mathcal{T}$ if we use $\mathcal{L}_2$ norm as the distance between our distributions, namely $\parallel \mathcal{T}[\boldsymbol{\sigma}^{(t\bullet)}] - \mathcal{T}[\hat{\boldsymbol{\sigma}}^{(t\bullet)}] \parallel_2 \leq |\lambda_2| \cdot \parallel \boldsymbol{\sigma}^{(t\bullet)} - \hat{\boldsymbol{\sigma}}^{(t\bullet)} \parallel_2$, where λ_2 is the second largest eigenvalue of $\mathcal{T}$.[2] Unfortunately, $\mathcal{L}_2$ norm is inappropriate for our purposes. Recall

[2]The proof of this result is a simple extension to the proof of the result for the asymptotic convergence of a Markov process to a stationary distribution.

that there are two main operations involved in updating a belief state: propagation through the transition model, and conditioning on an observation. The behavior of $\mathcal{L}_2$ norm with respect to conditioning is not helpful: $\| \mathcal{O}_{r_h}[\boldsymbol{\sigma}^{(\bullet t)}] - \mathcal{O}_{r_h}[\hat{\boldsymbol{\sigma}}^{(\bullet t)}] \|_2$ can be arbitrarily larger than $\| \boldsymbol{\sigma}^{(\bullet t)} - \hat{\boldsymbol{\sigma}}^{(\bullet t)} \|_2$. In fact, one can construct examples where the observation of any response r_h will cause the $\mathcal{L}_2$ distance to grow.

Thus, we must search for an alternative distance measure for which to try and prove our contraction result. The obvious candidate is *relative entropy*, or *KL divergence*, which quantifies the loss or inefficiency incurred by using distribution $\boldsymbol{\psi}$ when the true distribution is φ [Cover and Thomas, 1991, p.18]:

Definition 2 If $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are two distributions over the same space Ω, the *relative entropy* of $\boldsymbol{\varphi}$ to $\boldsymbol{\psi}$ is

$$\mathbb{D}[\varphi \,\|\, \boldsymbol{\psi}] \triangleq \boldsymbol{E}_{\boldsymbol{\varphi}}[\mathit{ln}\, \frac{\boldsymbol{\varphi}}{\boldsymbol{\psi}}] = \sum_{\omega_i \in \Omega} \varphi[\omega_i]\, \mathit{ln}\, \frac{\varphi[\omega_i]}{\boldsymbol{\psi}[\omega_i]}.$$

Relative entropy is, for a variety of reasons detailed in [Cover and Thomas, 1991, ch.2], a very natural measure of discrepancy to use between a distribution and an approximation to it. Furthermore, and in contrast to $\mathcal{L}_2$, it behaves very reasonably with respect to conditioning:

Fact 1 *For any t,*

$$\boldsymbol{E}_{\boldsymbol{\rho}^{(t)}}[\mathbb{D}[\mathcal{O}_{r_h}[\boldsymbol{\sigma}^{(\bullet t)}] \,\|\, \mathcal{O}_{r_h}[\hat{\boldsymbol{\sigma}}^{(t\bullet)}]]] \leq \mathbb{D}[\boldsymbol{\sigma}^{(\bullet t)} \,\|\, \hat{\boldsymbol{\sigma}}^{(\bullet t)}],$$

where $\boldsymbol{\rho}^{(t)} = (\boldsymbol{\sigma}^{(\bullet t)}\, \mathcal{O})$ is the prior on the response at time t.

Unfortunately, we seem to have simply shifted the problem from one place to another. While relative entropy is better behaved with respect to $\mathcal{O}$, there is no known contraction result for $\mathcal{T}$. Indeed, until now, the only related properties that seems to have been known are that relative entropy never increases by transition through a stochastic process (i.e., that $\mathbb{D}[\mathcal{T}[\boldsymbol{\sigma}^{(t\bullet)}] \,\|\, \mathcal{T}[\hat{\boldsymbol{\sigma}}^{(t\bullet)}]] \leq \mathbb{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \hat{\boldsymbol{\sigma}}^{(t\bullet)}]$), and that it ultimately tends to zero for a very broad class of processes (i.e., $\mathbb{D}[\mathcal{T}^k[\boldsymbol{\sigma}^{(t\bullet)}] \,\|\, \mathcal{T}^k[\hat{\boldsymbol{\sigma}}^{(t\bullet)}]] \to 0$ as $k \to \infty$ when $\mathcal{T}$ is ergodic), see [Cover and Thomas, 1991, p.34]. Unfortunately, we need much stronger results if we wish to bound the accumulation of the error over time. One of the main contributions of this paper is the first proof (to our knowledge) that a stochastic transition contracts relative entropy at a geometric rate.

We now prove that a stochastic process $\mathcal{T}$ does, indeed, lead to a contraction in the relative entropy distance. It will be useful later on (for the case of DBNs) to consider a somewhat more general setting, where the sets of states 'before' and 'after' the stochastic transition are not necessarily the same. Thus, let $\Omega = \{\omega_1, ..., \omega_n\}$ be our *anterior* state space, and $\Omega' = \{\omega'_1, ..., \omega'_{n'}\}$ be our *ulterior* state space. Let $\mathcal{Q}$ be an $n \times n'$ stochastic matrix, representing a random process from Ω to Ω'. Let φ and $\boldsymbol{\psi}$ be two given anterior distributions over Ω, and let $\boldsymbol{\varphi}'$ and $\boldsymbol{\psi}'$ be the corresponding ulterior distributions induced over Ω' by $\mathcal{Q}$.

Our goal is to measure the minimal extent to which the process $\mathcal{Q}$ causes the two distributions to become the same. In the worst case, there is no common starting point at all: all of the mass in one distribution $\boldsymbol{\varphi}$ is on some state ω_{i_1} while all of the mass in the other distribution $\boldsymbol{\psi}$ is on some other state ω_{i_2}. However, the stochastic nature of the process causes each of them to place some weight on any posterior state ω'_j: the probability $\boldsymbol{\varphi}'[\omega'_j]$ is $\mathcal{Q}[\omega'_j \mid \omega_{i_1}]$ while $\boldsymbol{\psi}'[\omega'_j]$ is $\mathcal{Q}[\omega'_j \mid \omega_{i_2}]$. Thus, while none of the probability mass of φ and $\boldsymbol{\psi}$ was in agreement, φ' and $\boldsymbol{\psi}'$ must agree on ω'_j for a mass of $min[\mathcal{Q}[\omega'_j \mid \omega_{i_1}], \mathcal{Q}[\omega'_j \mid \omega_{i_2}]]$. Based on this insight, we have the following natural characterization of the mixing properties of our process:

Definition 3 For a Markov process with stochastic transition model $\mathcal{Q}$, the *minimal mixing rate* of $\mathcal{Q}$ is

$$\gamma_{\mathcal{Q}} \triangleq min_{i_1,i_2} \sum_{j=1}^{n'} min[\mathcal{Q}[\omega'_j \mid \omega_{i_1}], \mathcal{Q}[\omega'_j \mid \omega_{i_2}]].$$

Shortly, our first theorem will show the fundamental relationship between $\gamma_{\mathcal{Q}}$ and the reduction of relative entropy through stochastic propagation by $\mathcal{Q}$. We start by proving the following lemma, which allows us to isolate the probability mass in the two distributions which is guaranteed to mix:

Lemma 2 *For any $\gamma \leq \gamma_{\mathcal{Q}}$ and any $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$, the matrix $\mathcal{Q}$ admits an additive* contraction decomposition $\mathcal{Q} = \mathcal{Q}^{\Gamma} + \mathcal{Q}^{\Delta}$ *satisfying these three properties:*

1. *for all $i \in \{1, ..., n\}$ and $j \in \{1, ..., n'\}$, $0 \leq \mathcal{Q}^{\Gamma}_{i,j} \leq \mathcal{Q}_{i,j}$;*
2. *for all $i \in \{1, ..., n\}$, $\sum_{j=1}^{n'} \mathcal{Q}^{\Gamma}_{i,j} = \gamma$;*
3. *for all $j \in \{1, ..., n'\}$, $\sum_{i=1}^{n} \varphi[\omega_i]\, \mathcal{Q}^{\Gamma}_{i,j} = \sum_{i=1}^{n} \boldsymbol{\psi}[\omega_i]\, \mathcal{Q}^{\Gamma}_{i,j}$.*

The proof is based on the fact that, for at least a certain portion of their probability mass, φ' and $\boldsymbol{\psi}'$ must agree. Our goal is to capture in $\mathcal{Q}^{\Gamma}$ the fraction of $\mathcal{Q}$ that forces this agreement.

Proof Our first task is to establish a correspondence from the masses of one distribution to the other. If we can "map" some amount of mass in φ sent to ω'_j by $\mathcal{Q}$ to a comparable amount of mass in $\boldsymbol{\psi}$ also sent to ω'_j, we can be sure that this fraction of the mass will end up distributed identically. That part of the mass which is already the same in both distributions simply maps to itself. The following construction shows how to take care of the rest.

Let $p_i = \boldsymbol{\varphi}[\omega_i] - min[\varphi[\omega_i], \boldsymbol{\psi}[\omega_i]]$ and $q_i = \boldsymbol{\psi}[\omega_i] - min[\boldsymbol{\varphi}[\omega_i], \boldsymbol{\psi}[\omega_i]]$. Note that for every i, either $p_i = 0$ or $q_i = 0$ (or both). Intuitively, p_i is the probability mass of $\varphi[\omega_i]$ that needs to be mapped. Let $I_p = \{i : p_i > 0\}$ and $I_q = \{i : q_i > 0\}$. We now choose a set of numbers $\pi_{i_1 \to i_2} \geq 0$ such that $\sum_{i_2} \pi_{i_1 \to i_2} = p_{i_1}$ and $\sum_{i_1} \pi_{i_1 \to i_2} = q_{i_2}$. As $\sum_{i_1} p_{i_1} = \sum_{i_2} q_{i_2}$, the construction of such numbers is straightforward, and we omit details. Let also $\rho_{i_1,i_2}[j] \geq 0$ such that $\sum_j \rho_{i_1,i_2}[j] = \gamma$ and $\rho_{i_1,i_2}[j] \leq min[\mathcal{Q}[\omega'_j \mid \omega_{i_1}], \mathcal{Q}[\omega'_j \mid \omega_{i_2}]]$. Such a choice is always possible, by the definition of $\gamma_{\mathcal{Q}}$.

We now define the matrix $\mathcal{Q}^\Gamma$ as follows. For $i_1 \in I_p$, let $\mathcal{Q}^\Gamma_{i_1,j} = \sum_{i'_2} (\pi_{i_1 \to i'_2}/p_{i_1})\, \rho_{i_1,i'_2}[j]$ while for $i_2 \in I_q$, $\mathcal{Q}^\Gamma_{i_2,j} = \sum_{i'_1} (\pi_{i'_1 \to i_2}/q_{i_2})\, \rho_{i'_1,i_2}[j]$; as I_p and I_q are disjoint, there is no conflict. For $i_0 \notin I_p \cup I_q$, we take $\mathcal{Q}^\Gamma_{i_0,j} = \gamma\, \mathcal{Q}_{i_0,j}$. We now show that the construction satisfies our three properties. From the definition of $\pi_{i_1 \to i_2}$ it follows that $\mathcal{Q}^\Gamma_{i_1,j}$ is a weighted average of $\rho_{i_1,i_2}[j]$. Properties 1 and 2 now follow from the definition of $\rho_{i_1,i_2}[j]$. An analogous proof holds for $\mathcal{Q}^\Gamma_{i_2,j}$; and the result holds trivially for $\mathcal{Q}^\Gamma_{i_0,j}$. As for property 3, we have:

$$\begin{aligned}
&\sum_{i=1}^n \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Gamma_{i,j} - \sum_{i=1}^n \boldsymbol{\psi}[\omega_i]\, \mathcal{Q}^\Gamma_{i,j} \\
&\quad= \sum_{i=1}^n (\boldsymbol{\varphi}[\omega_i] - \boldsymbol{\psi}[\omega_i])\, \mathcal{Q}^\Gamma_{i,j} \\
&\quad= \sum_{i=1}^n (p_i - q_i)\, \mathcal{Q}^\Gamma_{i,j} \\
&\quad= \sum_{i_1 \in I_p} p_{i_1}\, \mathcal{Q}^\Gamma_{i_1,j} - \sum_{i_2 \in I_q} q_{i_2}\, \mathcal{Q}^\Gamma_{i_2,j} \\
&\quad= \sum_{i_1 \in I_p} p_{i_1} \sum_{i_2} \frac{\pi_{i_1 \to i_2}}{p_{i_1}}\, \rho_{i_1,i_2}[j] - \sum_{i_2 \in I_q} q_{i_2} \sum_{i_1} \frac{\pi_{i_1 \to i_2}}{q_{i_2}}\, \rho_{i_1,i_2}[j] \\
&\quad= \sum_{i_1 \in I_p} \sum_{i_2} \pi_{i_1 \to i_2}\, \rho_{i_1,i_2}[j] - \sum_{i_2 \in I_q} \sum_{i_1} \pi_{i_1 \to i_2}\, \rho_{i_1,i_2}[j] \\
&\quad= \sum_{i_1} \sum_{i_2} (\pi_{i_1 \to i_2} - \pi_{i_1 \to i_2})\, \rho_{i_1,i_2}[j],
\end{aligned}$$

where the last step is based on the fact (implied by the definition of $\pi_{i_1 \to i_2}$) that $\pi_{i_1 \to i_2} = 0$ when $i_2 \in I_q$ or when $i_1 \in I_p$. Since this last expression equals 0, the proof is complete. ∎

Based on this lemma, our contraction result now follows easily. Essentially, the argument is based on a construction that makes explicit the different behavior of the process corresponding to the two parts of the contraction decomposition $\mathcal{Q}^\Gamma, \mathcal{Q}^\Delta$. We split the process into two separate phases: in the first, the process "decides" whether to contract, and in the second the appropriate transition occurs. That is, we define a new intermediate state space $\Omega^\dagger$, which contains a new distinguished *contraction* state c and otherwise is identical to Ω. We separate out the cases in which the process is guaranteed to contract $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ by having them correspond to an explicit transition to c. From c, the two processes behave identically, according to $\mathcal{Q}^\Gamma$, while from the remaining states in $\Omega^\dagger$, they behave according to $\mathcal{Q}^\Delta$.

Theorem 3 *For $\mathcal{Q}$, $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$, $\boldsymbol{\varphi}'$, $\boldsymbol{\psi}'$ as above:*

$$\boldsymbol{D}[\boldsymbol{\varphi}' \,\|\, \boldsymbol{\psi}'] \leq (1 - \gamma_{\mathcal{Q}})\, \boldsymbol{D}[\boldsymbol{\varphi} \,\|\, \boldsymbol{\psi}].$$

Proof Fix $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$, and define a new two-phase process with one Markovian transition $\mathcal{W}^\Gamma$ from Ω to $\Omega^\dagger$ and another $\mathcal{W}^\Delta$ from $\Omega^\dagger$ to Ω', where $\Omega^\dagger = \{u_1, ..., u_n, c\}$ is a new state set (see Figure 2). Intuitively, the state u_i corresponds to ω_i while the state c corresponds

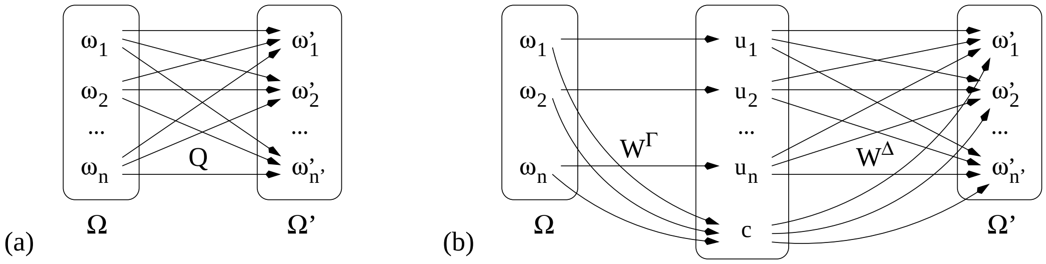

Figure 2: Decomposition used in Theorem 3: (a) generic Markov transition process; (b) two-stage process equivalent to it for $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$. In this diagram, arrows denote stochastic transitions between states.

to the state obtained if the process contracts. The process $\mathcal{W}^\Gamma$ transitions to the contract state with probability γ and preserves its state with probability $1-\gamma$; i.e., $\mathcal{W}^\Gamma[\omega_i \rightsquigarrow c] = \gamma$ while $\mathcal{W}^\Gamma[\omega_i \rightsquigarrow u_i] = 1-\gamma$. The process $\mathcal{W}^\Delta$ behaves like $\mathcal{Q}^\Delta$ from the states u_i; from c, it duplicates the aggregate behavior of $\boldsymbol{\varphi}$ on $\mathcal{Q}^\Gamma$; i.e., $\mathcal{W}^\Delta[u_i \rightsquigarrow \omega'_j] = \mathcal{Q}^\Delta_{i,j}/(1-\gamma)$ while $\mathcal{W}^\Delta[c, \omega'_j] = \sum_i \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Gamma_{i,j}/\gamma$.

We first need to show that the decomposed process is equivalent to the original process $\mathcal{Q}$ *for* $\boldsymbol{\varphi}$ *and* $\boldsymbol{\psi}$. (This will not be the case for other distributions.) Let $\boldsymbol{\varphi}^\dagger$ (resp. $\boldsymbol{\psi}^\dagger$) be the result of applying $\mathcal{W}^\Gamma$ to $\boldsymbol{\varphi}$ (resp. $\boldsymbol{\psi}$). Consider the distribution obtained by applying the two-phase process $\mathcal{W}^\Delta \circ \mathcal{W}^\Gamma$ to $\boldsymbol{\varphi}$. The probability of ω'_j is

$$
\begin{aligned}
&\frac{1}{1-\gamma} \sum_i \boldsymbol{\varphi}^\dagger[u_i]\, \mathcal{Q}^\Delta_{i,j} + \frac{\boldsymbol{\varphi}^\dagger[c]}{\gamma} \sum_{i'} \boldsymbol{\varphi}[\omega_{i'}]\, \mathcal{Q}^\Gamma_{i',j} \\
&\quad = \frac{1}{1-\gamma} \sum_i (1-\gamma)\, \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Delta_{i,j} + \frac{\gamma}{\gamma} \sum_{i'} \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Gamma_{i',j} \\
&\quad = \sum_i \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Delta_{i,j} + \sum_{i'} \boldsymbol{\varphi}[\omega_{i'}]\, \mathcal{Q}^\Gamma_{i',j} \\
&\quad = \sum_i \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}_{i,j}.
\end{aligned}
$$

Similarly, applying $\mathcal{W}^\Delta \circ \mathcal{W}^\Gamma$ to $\boldsymbol{\psi}$, we get $\sum_i \boldsymbol{\psi}[\omega_i]\, \mathcal{Q}^\Delta_{i,j} + \sum_i \boldsymbol{\varphi}[\omega_i]\, \mathcal{Q}^\Gamma_{i,j}$; and by property 3, this expression is also $\sum_i \boldsymbol{\psi}[\omega_i]\, \mathcal{Q}_{i,j}$.

To show the contraction property $\boldsymbol{D}[\boldsymbol{\varphi}' \,\|\, \boldsymbol{\psi}'] \leq (1-\gamma)\, \boldsymbol{D}[\boldsymbol{\varphi} \,\|\, \boldsymbol{\psi}]$, we first note that since $\mathcal{W}^\Delta$ is Markovian, we have $\boldsymbol{D}[\boldsymbol{\varphi}' \,\|\, \boldsymbol{\psi}'] \leq \boldsymbol{D}[\boldsymbol{\varphi}^\dagger \,\|\, \boldsymbol{\psi}^\dagger]$. On the other hand, we have

$$
\begin{aligned}
\boldsymbol{D}[\boldsymbol{\varphi}^\dagger \,\|\, \boldsymbol{\psi}^\dagger] &= \sum_i \boldsymbol{\varphi}^\dagger[u_i]\, \mathit{ln} \frac{\boldsymbol{\varphi}^\dagger[u_i]}{\boldsymbol{\psi}^\dagger[u_i]} + \boldsymbol{\varphi}^\dagger[c]\, \mathit{ln} \frac{\boldsymbol{\varphi}^\dagger[c]}{\boldsymbol{\psi}^\dagger[c]} \\
&= \sum_i (1-\gamma)\, \boldsymbol{\varphi}[\omega_i]\, \mathit{ln} \frac{\boldsymbol{\varphi}[\omega_i]}{\boldsymbol{\psi}[\omega_i]} + \gamma\, \mathit{ln} \frac{\gamma}{\gamma} \;=\; (1-\gamma)\, \boldsymbol{D}[\boldsymbol{\varphi} \,\|\, \boldsymbol{\psi}],
\end{aligned}
$$

which concludes the proof. ∎

4 Compound processes

So far we have shown that errors decrease geometrically in Markov processes. Later we shall see that the rate of decrease is directly linked to the quality of the approximation we obtain. Clearly, one cannot guarantee that this rate will always be large enough to be useful: for example, if the process has a component which is almost deterministic, γ can be quite small. Unfortunately, this situation also arises for complex processes with large number of variables, even if each of them is governed by nicely stochastic transition dynamics. As an extreme example, imagine a process composed of N binary variables evolving independently, flipping their value from one time slice to the next with probability δ. Each variable, viewed as a separate Markov process, has a mixing rate of 2δ. Thus, one may expect the mixing rate of the process as a whole to be as good; indeed, since all of the processes are independent, we could hardly expect otherwise.

However Theorem 3 tells a different story: computing γ for the transition matrix of the compound process as a whole, one gets a discouragingly small value which is $\leq (4\delta)^{N/2}$. Is our definition of the mixing rate simply too pessimistic? Unfortunately not. The fallacy is in our assumption that local mixing properties would automatically carry over to the compound process. Each subprocess is rapidly mixing for belief states over its own variable only. If the belief state of the compound process involves dependencies between variables belonging to different subprocesses, then our contraction rate can, indeed, be as bad as our prediction. Assume, for example, that the true distribution $\boldsymbol{\varphi}$ gives probability 1 to the state (i.e., assignment of values to variables) $(0, ..., 0)$ while the approximate distribution $\boldsymbol{\psi}$ gives probability p to that state and $1 - p$ to its opposite $(1, ..., 1)$. We can view the state space as a hypercube, and each of these distributions as a mass assignment to vertices of the hypercube. A single step through the transition matrix "diffuses" the mass of the two distributions randomly around their starting points. However, the probability that the diffusion process around one point reaches the other is exponentially low, since all bits have to flip in one or the other of the two distributions.

Given this example, we must face the question of whether our contraction theorem is even useful for large processes. As we have seen, even the assumption that the process is highly decomposable is not enough. One idea is to make some additional assumptions about the structure of the belief state distributions, e.g., that they decompose. In general, an analysis based on decomposing a relative entropy expression requires that we assume decomposability for the "true distribution" (the first argument of $\mathbb{D}[\cdot \,\|\, \cdot]$). In our context, an assumption such as this would be patently false; indeed, the lack of decomposability of the true belief state was the basis for our entire paper. As it happens, however, our example above shows that independence of the true distribution $\boldsymbol{\varphi}$ would not help in this context.

Surprisingly (and atypically), we gain significant advantage by making a decomposability assumption only for the *approximate* belief state. That is, we can show that if the process decomposes well, and the estimated belief state decomposes in a way that matches the structure of the process, then we get significantly better bounds on the error contraction coefficient, regardless of the true belief state. Thus, as far as error contraction goes, the properties of the true belief state are not crucial: only the approximate belief state and the process itself. This is very fortunate, as it is feasible to enforce decomposability properties on the approximate belief state, whereas we have no control over the true belief state.

Formally, it is most convenient to describe our results in the framework of factored HMMs [Smyth *et al.*, 1996]; in the next section, we discuss how they can be applied to dynamic Bayesian networks. We assume that our system is composed of several subprocesses $\mathcal{T}_l$. Each subprocess has a state with a Markovian evolution model. The state of subprocess l at time t is written $S_l^{(t)}$. The evolution model $\mathcal{T}_l$ is a stochastic mapping from the states of some set of processes at time t to the state of process l at time $t+1$. We say that subprocess l *depends on* subprocess l' if $\mathcal{T}_l$ depends on the value of $S_{l'}^{(t)}$. Our model also allows a set of response variables at each time t, which can depend arbitrarily on the states of the processes at time t; however, as we are interested primarily in the contraction properties of our system, the properties of the response variables are irrelevant to our current analysis.

4.1 Independent subprocesses

We begin by considering the simple case where our subprocesses are completely independent, i.e., where subprocess l depends only on subprocess l. We also assume that our approximate belief state decomposes along the same lines. As we shall see later on, this case forms the basis for our more general result. We now give a proof for this case, starting with the following lemma.

Lemma 4 *Let $\boldsymbol{\varphi}^\star$ and $\boldsymbol{\psi}^\star$ be two distributions over the same space, $\boldsymbol{W}$ and $\boldsymbol{Z}$ be two (sets of) random variables over this space, and E be some event in the space. Assume that, in $\boldsymbol{\psi}^\star$, $\boldsymbol{W}$ and $\boldsymbol{Z}$ are conditionally independent given E. Then:*

$$\boldsymbol{D}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid E] \,\|\, \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]] \leq \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{D}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E] \,\|\, \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]]].$$

Proof Letting $\boldsymbol{H}[\cdot]$ represent the standard entropy function, we have:

$$\begin{aligned}
&\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{D}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E] \,\|\, \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]]] \\
&\quad = \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid \boldsymbol{W},E]}[\ln \frac{\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]}{\boldsymbol{\psi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]}]] \\
&\quad = \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid \boldsymbol{W},E]}[\ln \boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]] - \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid \boldsymbol{W},E]}[\ln \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]]] \\
&\quad = \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[-\boldsymbol{H}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]]] - \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid \boldsymbol{W},E]}[\ln \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]]] \\
&\quad \geq -\boldsymbol{H}[\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid \boldsymbol{W}, E]]] - \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{W}\mid E]}[\boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid \boldsymbol{W},E]}[\ln \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]]] \\
&\quad = -\boldsymbol{H}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid E]] - \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid E]}[\ln \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]] \\
&\quad = \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid E]}[\ln \boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid E]] - \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid E]}[\ln \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]] \\
&\quad = \boldsymbol{E}_{\boldsymbol{\varphi}^\star[\boldsymbol{Z}\mid E]}[\ln \frac{\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid E]}{\boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]}] \\
&\quad = \boldsymbol{D}[\boldsymbol{\varphi}^\star[\boldsymbol{Z} \mid E] \,\|\, \boldsymbol{\psi}^\star[\boldsymbol{Z} \mid E]].
\end{aligned}$$

The third equality is by linearity of expectation and the conditional independence assumption for $\boldsymbol{\psi}^{\star}$. The inequality is a consequence of Jensen's inequality, based on the convexity of $-\boldsymbol{H}[\cdot]$. The next equality is based on a simple marginalization of the probability distribution. The remaining steps are largely definitional. ∎

Theorem 5 *Consider L independent subprocesses $\mathcal{T}_1, \ldots, \mathcal{T}_L$, and assume that each subprocess l depends only on l. Let γ_l be the mixing rate of the subprocess $\mathcal{T}_l$ and let $\gamma = \min_l \gamma_l$. Let $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ be distributions over $S_1^{(t)}, \ldots, S_L^{(t)}$, and assume that $\boldsymbol{\psi}$ renders the $S_l^{(t)}$ marginally independent. Then (using our notations from the previous section):*

$$\boldsymbol{D}[\boldsymbol{\varphi}' \,\|\, \boldsymbol{\psi}'] \leq (1-\gamma)\boldsymbol{D}[\boldsymbol{\varphi} \,\|\, \boldsymbol{\psi}].$$

Proof It suffices to show the result for two independent subprocesses. The general case follows by induction. Let $\mathcal{T}_X$ be a subprocess whose state is described by the variable X; let $\boldsymbol{X}$ be the state at time t and $\boldsymbol{X}'$ the state at time $t+1$. Let $\mathcal{T}_Y$ be a subprocess whose state is $\boldsymbol{Y}$, with analogous notations. Let $\boldsymbol{\varphi}[\boldsymbol{X}, \boldsymbol{Y}]$ and $\boldsymbol{\psi}[\boldsymbol{X}, \boldsymbol{Y}]$ be our true and approximate distributions over the anterior variables, and let $\boldsymbol{\varphi}^{\star}[\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{X}', \boldsymbol{Y}']$ and $\boldsymbol{\psi}^{\star}[\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{X}', \boldsymbol{Y}']$ be the distributions over both time slices induced by our anterior distributions and $\mathcal{T}_X$ and $\mathcal{T}_Y$. Note that $\boldsymbol{\varphi}' = \boldsymbol{\varphi}^{\star}[\boldsymbol{X}', \boldsymbol{Y}']$, and similarly for $\boldsymbol{\psi}'$.

By the standard decomposition properties of relative entropy, we have that

$$\begin{aligned}
&\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{X}', \boldsymbol{Y}'] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{X}', \boldsymbol{Y}']] \\
&= \boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{X}'] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{X}']] + \boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}']}[\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}'] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}']]] \;. \qquad (1)
\end{aligned}$$

Using the contraction $\mathcal{T}_X$, we have that the first term is at most

$$(1-\gamma)\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{X}] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{X}]]. \qquad (2)$$

To simplify the second term, we apply Lemma 4 to the internal component of the expectation, substituting $\boldsymbol{Y}'$ for $\boldsymbol{Z}$, the specific value of $\boldsymbol{X}'$ for E, and $\boldsymbol{X}$ for $\boldsymbol{W}$. The conditions of the lemma hold due to our assumptions: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent in $\boldsymbol{\psi}^{\star}$, and the subprocesses evolve independently; therefore, $\boldsymbol{X}, \boldsymbol{X}'$ and $\boldsymbol{Y}, \boldsymbol{Y}'$ are independent in $\boldsymbol{\psi}^{\star}$; and certainly $\boldsymbol{X}$ and $\boldsymbol{Y}'$ are conditionally independent given $\boldsymbol{X}'$. We get that

$$\begin{aligned}
&\boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}']}[\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}'] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}']]] \\
&\leq \boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}']}[\boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X} \mid \boldsymbol{X}']}[\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}', \boldsymbol{X}] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}', \boldsymbol{X}]]]] \\
&= \boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}, \boldsymbol{X}']}[\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}]]] \\
&= \boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}]}[\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y}' \mid \boldsymbol{X}]]] \\
&\leq \boldsymbol{E}_{\boldsymbol{\varphi}^{\star}[\boldsymbol{X}]}[(1-\gamma)\,\boldsymbol{D}[\boldsymbol{\varphi}^{\star}[\boldsymbol{Y} \mid \boldsymbol{X}] \,\|\, \boldsymbol{\psi}^{\star}[\boldsymbol{Y} \mid \boldsymbol{X}]]], \qquad (3)
\end{aligned}$$

where the second equality follows from our conditional independence assumptions, and the last inequality follows from the contraction property for $\mathcal{T}_Y$, applied to each of the distribution pairs $\varphi^\star[Y \mid X]$ and $\psi^\star[Y \mid X]$ (for the different possible values of X). Putting together (1), (2) and (3), we get that

$$\begin{aligned}
&\mathbb{D}[\varphi^\star[X', Y'] \,\|\, \psi^\star[X', Y']] \\
&\quad\leq\ (1-\gamma)\, \mathbb{D}[\varphi^\star[X] \,\|\, \psi^\star[X]] + (1-\gamma)\, \mathbb{E}_{\varphi^\star[X]}[\mathbb{D}[\varphi^\star[Y \mid X] \,\|\, \psi^\star[Y \mid X]]] \\
&\quad=\ (1-\gamma)\, \mathbb{D}[\varphi^\star[X, Y] \,\|\, \psi^\star[X, Y]],
\end{aligned}$$

as required. ∎

Thus, if we have a set of independent subprocesses, each of which contracts, *and* our approximate belief state decomposes along the same lines as the process, then the contraction of the process as a whole is no worse than the contraction of the individual subprocesses. Since each subprocess involves a much smaller number of states, its transition probabilities are likely to be reasonably large (assuming it is stochastic enough). This analysis usually results in a much better mixing rate.

The above result is not really useful in itself, because if our subprocesses are really independent, our belief state would never become correlated in the first place, and we would not need to approximate it. The main purpose of this result is to lay the foundation to the general case to be described next.

4.2 Conditionally independent subprocesses

Our goal is now to relax the decomposition in independent subprocesses to conditional independence relations only; i.e., we generalize the above result to the more realistic case where the states of different subprocesses can depend on each other.

Assume that subprocess l depends on subprocesses $l_1, ..., l_k$. Then, $\mathcal{T}_l$ defines a probability $P(S_l \mid S_{l_1}, ..., S_{l_k})$. This transition probability can be defined as a transition matrix, but one whose anterior and ulterior state spaces can be different. Luckily, in Section 3, we allowed for exactly this possibility, so that the mixing rate of $\mathcal{T}_l$ is well-defined. Let γ_l be the mixing rate of $\mathcal{T}_l$, and let $\gamma = min_l \gamma_l$. If our approximate belief state, as before, respects the process structure, then we can place a bound on the mixing rate of the entire process. This bound depends both on γ and on the process structure. We illustrate the basic construction for a simple example; generalization to arbitrary structures is straightforward.

Consider two processes $\mathcal{T}_X$ and $\mathcal{T}_Y$ as above, and assume that $\mathcal{T}_Y$ depends on $\mathcal{T}_X$. Our basic construction follows the lines of the proof of Theorem 3: we split the transition of each process into two phases where the first chooses whether or not to contract and the second concludes the transition in a way that depends on whether the process has contracted. Note, however, that the variable X plays a role both in $\mathcal{T}_X$ and in $\mathcal{T}_Y$, and that the transitions of these two subprocesses are conditionally independent given X. Thus, X cannot make a single decision to contract and apply it in the context of both processes. We therefore introduce two separate intermediate variables for $\mathcal{T}_X$, $X_X^\dagger$ and $X_Y^\dagger$, where the first decides

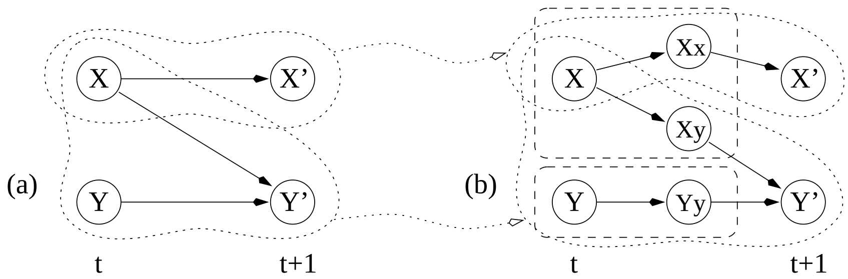

Figure 3: Principle of the construction used in Theorem 6: (a) some factored Markov transition process; (b) decomposition in a structured two-stage process. The dotted contours delineate subprocesses that, for $\boldsymbol{\varphi}[\boldsymbol{X},\boldsymbol{Y}]$ and $\boldsymbol{\psi}[\boldsymbol{X},\boldsymbol{Y}]$, behave equivalently in (a) and (b). The dashed boxes subdivide the first stage of (b) in fully independent subprocesses.

whether $\boldsymbol{X}$ contracts in the context of $\mathcal{T}_X$ and the second whether it contracts in the context of $\mathcal{T}_Y$, as shown in Figure 3.

We define our transitions for the partitioned model so as to make sure that the two-phase process

$$\boldsymbol{X} \to \boldsymbol{X}^\dagger_X \to \boldsymbol{X}'$$

induces the same behavior as $\mathcal{T}_X$ for the distributions $\boldsymbol{\varphi}[\boldsymbol{X}]$ and $\boldsymbol{\psi}[\boldsymbol{X}]$. This construction is essentially identical to that of Theorem 3, except that that the transition from $\boldsymbol{X}$ to the distinguished contraction state c^X_X of $\boldsymbol{X}^\dagger_X$ is taken with some probability λ_1 to be specified below. Similarly, we ensure that the two phase process

$$\begin{array}{lllll} \boldsymbol{X} & \to & \boldsymbol{X}^\dagger_{\boldsymbol{Y}} & \searrow & \\ \boldsymbol{Y} & \to & \boldsymbol{Y}^\dagger & \to & \boldsymbol{Y}' \end{array}$$

induces the same behavior as $\mathcal{T}_Y$ for the distributions $\boldsymbol{\varphi}[\boldsymbol{X},\boldsymbol{Y}]$ and $\boldsymbol{\psi}[\boldsymbol{X},\boldsymbol{Y}]$. Here, we have to deal with a slight subtlety: The transition from $\boldsymbol{X}^\dagger_Y, \boldsymbol{Y}^\dagger$ to $\boldsymbol{Y}'$ must behave as if the first phase contracted if *either* of the two antecendents $\boldsymbol{X}^\dagger_Y, \boldsymbol{Y}^\dagger$ is in the contraction state. (There is no choice, because if even one is in the contraction state, the process no longer has enough information to transition according to $\mathcal{Q}^\Delta_Y$.) Thus, if $\boldsymbol{X} \to \boldsymbol{X}^\dagger_Y$ contracts with probability λ_2 and $\boldsymbol{Y} \to \boldsymbol{Y}^\dagger$ with probability λ_3, the process as a whole contracts with probability $1-(1-\lambda_2)(1-\lambda_3)$.

Therefore, in order for us to be able to construct a contraction decomposition $\mathcal{Q}^\Gamma_Y, \mathcal{Q}^\Delta_Y$, we must select λ_1, λ_2, λ_3 so as to satisfy $\lambda_1 \le \gamma_{\mathcal{Q}_X}$ and $1-(1-\lambda_2)(1-\lambda_3) \le \gamma_{\mathcal{Q}_Y}$. Assuming that $\gamma_{\mathcal{Q}_X} = \gamma_{\mathcal{Q}_Y} = \gamma$, we have that $\lambda_2 = \lambda_3 = 1-\sqrt{1-\gamma}$ is one legitimate selection; observe that $1-\sqrt{1-\gamma} \ge \gamma/2$.

To analyze the actual contraction rate of the process as a whole, we first analyze the contraction from the initial variables $\boldsymbol{X}, \boldsymbol{Y}$ to the intermediate variables $\boldsymbol{X}^\dagger_X, \boldsymbol{X}^\dagger_Y, \boldsymbol{Y}^\dagger$; the contraction for the two phases is no smaller. Our analysis uses a somewhat different process structure than the one used to show the correctness of the partition. We analyze the contraction rate for the transition $\mathcal{W}_X$ from $\boldsymbol{X}$ to $\{\boldsymbol{X}_X^\dagger, \boldsymbol{X}_Y^\dagger\}$ and for the transition $\mathcal{W}_Y$ from $\boldsymbol{Y}$ to $\boldsymbol{Y}^\dagger$. These two processes are independent (by design); further, we have assumed that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent in $\boldsymbol{\psi}$. Thus, the conditions of Theorem 5 apply[3] and the contraction of the process from $\boldsymbol{X}, \boldsymbol{Y}$ to $\boldsymbol{X}_X^\dagger, \boldsymbol{X}_Y^\dagger, \boldsymbol{Y}^\dagger$ is the minimum of the contraction of $\mathcal{W}_X$ and of $\mathcal{W}_Y$. Straightforwardly, the contraction of $\mathcal{W}_Y$ is λ_3. However, $\mathcal{W}_X$ contracts—loses all information about its original state—only when both $\boldsymbol{X}_X$ and $\boldsymbol{X}_Y$ enter their contraction state. These events are independent, hence the probability that they both occur is the product $\lambda_1\lambda_2$.

Thus, interconnectivity between the processes costs us in terms of our contraction ratio. Consider a subprocess $\mathcal{Q}_l$ whose contraction ratio is γ_l, and assume that it depends on r subprocesses. We must choose the contraction factor λ for each influencing process (assuming for simplicity that all are chosen to be equal) to be $1 - \sqrt[r]{1-\gamma}$, which is $\geq \gamma/r$. Thus, the cost of "inward" connectivity is a linear reduction in the contraction rate. The cost of "outward" connectivity is much higher. Each influence of a subprocess $\mathcal{Q}_l$ on another subprocess involves the construction of another intermediate variable, each of which contracts independently. The total contraction of $\mathcal{W}_{X_l}$ is the product of the individual contractions. Thus, the cost of outward connectivity is an exponential reduction in the contraction factor. Intuitively, this phenomenon makes sense: if a process influences many others, it is much less likely that its value will be lost. This analysis is the basis for the following theorem.

Theorem 6 *Consider a system consisting of L subprocesses $\mathcal{T}_1, \ldots, \mathcal{T}_L$, and assume that each subprocess depends on at most r others; each subprocess influences at most q others; and each $\mathcal{T}_l$ has minimum mixing rate $\gamma_l \geq \gamma$. Let $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ be distributions over $S_1^{(t)}, \ldots, S_L^{(t)}$, where the $S_l^{(t)}$ are independent in $\boldsymbol{\psi}$. Then:*

$$\boldsymbol{D}[\boldsymbol{\varphi}' \,\|\, \boldsymbol{\psi}'] \leq (1-\gamma^*)\, \boldsymbol{D}[\boldsymbol{\varphi} \,\|\, \boldsymbol{\psi}],$$

where $\gamma^* = (\frac{\gamma}{r})^q$.

Thus, if we have a system which is composed of several sparsely interacting subprocesses each of which is fairly stochastic and we use an approximate belief state which makes the states of the subprocesses independent, then our process as a whole contracts at a reasonable rate. We will soon exploit this property to devise an approximate monitoring strategy that achieves both accuracy and efficiency.

Before moving on, one should note that the independence assumption on $\boldsymbol{\psi}$ is not a peculiarity of our proof, but a much needed condition: in particular, relaxing it to conditional independence will not work without additional assumptions. To see this, recall our example of the N independent processes at the beginning of Section 4, and observe that although the $\boldsymbol{\psi}$ there already satisfied all possible conditional independence relations (for any (sets of)

[3]The theorem was stated for processes where the anterior and ulterior state space is identical, but the same proof applies to the more general case.

variables A, B, C, we had that A and B were ψ-independent given C whenever $C \neq \emptyset$), the lack of unconditional independence in ψ was enough to make the whole process contract at an exponentially small rate as mentioned.

5 Efficient monitoring

As we suggested throughout the paper, one of the main applications of our results is to the task of monitoring a complex process. As we mentioned, the exact monitoring task is often intractable, requiring that we maintain a very large belief state. Our results suggest an alternative approach, where we maintain instead an *approximate* belief state. If, for example, our process is composed of some number of weakly interacting subprocesses—e.g., several cars on a freeway—it may be reasonable to ignore correlations between the different components in our belief state. In essence, we could represent our beliefs about the entire process via our beliefs about its parts, e.g., our beliefs about the individual vehicles. For other types of processes, other approximate belief state representations may be more suitable. In continuous processes, for example, we may want to use a fixed-size mixture of Gaussians.

5.1 Approximate monitoring

Recall the notation of Section 2 about belief states, and denote by $\tilde{\boldsymbol{\sigma}}^{(t)}$ our compactly represented approximate belief state at time t. It is updated using the same process as $\boldsymbol{\sigma}^{(t\bullet)}$: we propagate it through the transition model, obtaining $\hat{\boldsymbol{\sigma}}^{(\bullet t+1)}$, and condition on the current response, obtaining $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$. However, $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$ does not usually admit a compact representation. In order to maintain the feasibility of our update process, we must approximate $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$, typically by finding a "nearby" distribution that admits a compact representation; the result is our new approximate belief state $\tilde{\boldsymbol{\sigma}}^{(t+1)}$. In our freeway domain, for example, we may compute our new beliefs about the state of each vehicle by projecting $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$, and use the cross product of these individual belief states as our approximation. In our continuous process, we could project back into our space of allowable belief states by approximating the distribution using a fixed number of Gaussians.

We begin by analyzing the error resulting from this type of strategy, i.e., the distance between the true posterior belief state $\boldsymbol{\sigma}^{(t+1\bullet)}$ and our approximation to it $\tilde{\boldsymbol{\sigma}}^{(t+1)}$. Intuitively, this error results from two sources: the "old" error which we "inherited" from the previous approximation $\tilde{\boldsymbol{\sigma}}^{(t)}$, and the "new" error derived from approximating $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$ using $\tilde{\boldsymbol{\sigma}}^{(t+1)}$. Suppose that each approximation introduces an error of ε, increasing the distance between the exact belief state and our approximation to it. However, the contraction resulting from the state transitions serves to drive them closer to each other, reducing the effect of old errors by a factor of γ. The various observations move the two even closer to each other on expectation (averaged over the different possible responses). Therefore, the expected error accumulated up to time t would behave as $\varepsilon + (1-\gamma)\,\varepsilon + (1-\gamma)^2\,\varepsilon + ... + (1-\gamma)^{t-1}\,\varepsilon \leq$

$\varepsilon \sum_i (1-\gamma)^i = \frac{\varepsilon}{\gamma}$.

To formalize this result, we first need to quantify the error resulting from our approximation. Our new approximate belief state $\tilde{\boldsymbol{\sigma}}^{(t)}$ is an approximation to $\hat{\boldsymbol{\sigma}}^{(t\bullet)}$. Most obviously, we would define the error of the approximation as the relative entropy distance between them—$\boldsymbol{D}[\hat{\boldsymbol{\sigma}}^{(t\bullet)} \,\|\, \tilde{\boldsymbol{\sigma}}^{(t)}]$. However, our error is measured relative to $\boldsymbol{\sigma}^{(t\bullet)}$ and not to $\hat{\boldsymbol{\sigma}}^{(t\bullet)}$. Therefore, we use the following definition:

Definition 4 We say that an approximation $\tilde{\boldsymbol{\sigma}}^{(t)}$ of $\hat{\boldsymbol{\sigma}}^{(t\bullet)}$ *incurs error* ε *relative to* $\boldsymbol{\sigma}^{(t\bullet)}$ if

$$\boldsymbol{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \tilde{\boldsymbol{\sigma}}^{(t)}] - \boldsymbol{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \hat{\boldsymbol{\sigma}}^{(t\bullet)}] \leq \varepsilon.$$

Our final theorem now follows easily by induction on t:

Theorem 7 *Let $\mathcal{T}$ be a stochastic process whose mixing rate is γ, assume that we have an approximation scheme that, at each phase t, incurs error ε relative to $\boldsymbol{\sigma}^{(t\bullet)}$. Then for any t, we have:*

$$\boldsymbol{E}_{\boldsymbol{\rho}^{(1,\ldots,t)}}[\boldsymbol{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \tilde{\boldsymbol{\sigma}}^{(t)}]] \leq \varepsilon/\gamma,$$

where the expectation is taken over the possible response sequences $r_{h_1}, ..., r_{h_t}$, with the probability ascribed to them by the process $\mathcal{T}$.

Of course, it is not trivial to show that a particular approximation scheme will satisfy the accuracy requirement of ε. The main difficulty stems from the fact that the notion of error in Definition 4 depends on the true belief state $\boldsymbol{\sigma}^{(t\bullet)}$, which is usually not known. Nevertheless, it is easy to show that if we have

$$max_i \frac{\hat{\boldsymbol{\sigma}}^{(t\bullet)}[s_i]}{\tilde{\boldsymbol{\sigma}}^{(t)}[s_i]} = \eta \tag{4}$$

then necessarily $\boldsymbol{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \tilde{\boldsymbol{\sigma}}^{(t)}] - \boldsymbol{D}[\boldsymbol{\sigma}^{(t\bullet)} \,\|\, \hat{\boldsymbol{\sigma}}^{(t\bullet)}] \leq \varepsilon = ln[\eta]$. Expression (4) is simply the maximum relative error caused by the approximation scheme at time t, a value which is often easy to assess for a given approximation step.

Note that the above theorem only provides a bound on the expected error. The bounds it provides for specific sequences of evidence are much weaker; in particular, the error after a very unlikely sequence of evidence might be quite large. Fortunately, our contraction result holds for arbitrary distributions, no matter how far apart. Thus, even if momentarily $\boldsymbol{\sigma}^{(t\bullet)}$ and $\tilde{\boldsymbol{\sigma}}^{(t)}$ are very different, the contraction property will reduce this error at an exponential rate.

5.2 Monitoring in DBNs

Let us apply this idea to the problem of monitoring a process described as a DBN. In this case, the process is specified in terms of an ordered set of state variables $X_1, ..., X_n$. The probability model of a DBN is typically described using a *2-TBN* (a two time-slice temporal Bayesian network), as shown in Figure 1. The 2-TBN associates each variable with a conditional probability distribution $\mathbb{P}[X_k^{(t+1)} \mid \mathit{Parents}(X_k^{(t+1)})]$, where $\mathit{Parents}(X_k^{(t+1)})$ can contain any variable at time t and such variables at time $t+1$ that preceed X_k in the total ordering. This model represents the conditional distribution over the state at time $t+1$ given the variables at time t. A *persistent state variable* is a variable whose value at time t directly or indirectly affects some variable at time $t+1$. The *canonical set of state variables* is the set $\{X : X^{(t)} \in \cup_k \mathit{Parents}(X_k^{(t+1)})\}$. Clearly, the canonical variables are all persistent. A 2-TBN is in *canonical form* if only the canonical variables are represented at time t.

To capture the idea of a subprocess, we partition the set of canonical state variables into disjoint subsets $\boldsymbol{X}_1, ..., \boldsymbol{X}_L$. Our partition must satisfy the requirement that no $\boldsymbol{X}_l$ may be affected by another $\boldsymbol{X}_{l'}$ *within the same time slice*; i.e., if $X \in \boldsymbol{X}_l$, then $X^{(t+1)}$ cannot have as an ancestor a variable $Y^{(t+1)}$ for $Y \in \boldsymbol{X}_{l'} \neq \boldsymbol{X}_l$. Note that a time slice may also contain non-persistent variables, e.g., sensor readings; but since none of them may ever be an ancestor of a canonical variable, we allow them to depend on any persistent variable in their time slice. Since the various clusters $\boldsymbol{X}_l$ correspond to our subprocesses from Theorem 6, we shall maintain an approximate belief state in which the $\boldsymbol{X}_l$ are independent, as prescribed.

The approximate monitoring procedure for DBNs follows the same lines as the general procedure described in Section 2: At each point in time, we have an approximate belief state $\tilde{\boldsymbol{\sigma}}^{(t)}$, in which the $\boldsymbol{X}_l$ are all independent. We propagate $\tilde{\boldsymbol{\sigma}}^{(t)}$ through the transition model, and then condition the result on our observations at time $t+1$. We then compute $\tilde{\boldsymbol{\sigma}}^{(t+1)}$ by projecting $\hat{\boldsymbol{\sigma}}^{(t+1\bullet)}$ onto each $\boldsymbol{X}_l$; i.e., we define $\tilde{\boldsymbol{\sigma}}^{(t+1)}[\boldsymbol{X}_l] = \hat{\boldsymbol{\sigma}}^{(t+1\bullet)}[\boldsymbol{X}_l]$, and the entire distribution as a product of these factors.

In the case of DBNs, we can actually accomplish this update procedure quite efficiently. We first generate a clique tree [Lauritzen and Spiegelhalter, 1988] in which, for every l, some clique contains $\boldsymbol{X}_l^{(t)}$ and some clique contains $\boldsymbol{X}_l^{(t+1)}$. A standard clique tree propagation algorithm can then be used to compute the posterior distribution over every clique. Once that is done, the distribution over $\boldsymbol{X}_l^{(t+1)}$ is easily extracted from the appropriate clique. Further savings can be obtained if we assume a stationary DBN and a static approximation scheme. Then it is possible to compute the topology of the clique tree in advance, and initialize it once and for all with the numerical information contained in the DBN. This results in a "proto-clique-tree" that can serve as a starting point for each propagation. The method is summarized in the following algorithm.

Algorithm 1 Approximate DBN monitoring.

INPUTS: (i) a 2-TBN in canonical form; (ii) a partition of the canonical variables $\{\boldsymbol{X}_l\}$; (iii) an initial belief state $\tilde{\boldsymbol{\sigma}}^{(t_0)}$; (iv) a stream of observations $\boldsymbol{r}_{h_1}^{(1)}, \boldsymbol{r}_{h_2}^{(2)}, \boldsymbol{r}_{h_3}^{(3)}, ...$

OUTPUT: a stream of approximate belief states $\tilde{\boldsymbol{\sigma}}^{(1)}, \tilde{\boldsymbol{\sigma}}^{(2)}, \tilde{\boldsymbol{\sigma}}^{(3)}, ...$

METHOD: the following procedure is based on standard clique tree inference; we refer the reader to [Huang and Darwiche, 1994] for background information.

1. Construct a clique tree from the 2-TBN, requiring that every $\boldsymbol{X}_l^{(t)}$ and $\boldsymbol{X}_l^{(t+1)}$ be contained in full in at least one clique.
2. Initialize each clique factor to the constant function $= 1$.
3. Incorporate the conditional probability tables of the DBN into the appropriate factors. Let Υ_0 be the resulting (uncalibrated) "proto-clique-tree".
4. For $t = 0, 1, 2, \ldots$:
   a. Let Υ be a working copy of Υ_0. Create $\tilde{\boldsymbol{\sigma}}^{(t+1)}$ as a new empty belief state.
   b. For each l, incorporate the marginal $\tilde{\boldsymbol{\sigma}}^{(t)}[\boldsymbol{X}_l^{(t)}]$ in the appropriate factor of Υ.
   c. Incorporate the evidence $\boldsymbol{r}_{h_t}^{(t+1)}$ in Υ.
   d. Calibrate the potentials in Υ.
   e. For each l, query Υ for the marginal over $\boldsymbol{X}_l^{(t+1)}$ and store it in $\tilde{\boldsymbol{\sigma}}^{(t+1)}$.
   f. Discard Υ and output $\tilde{\boldsymbol{\sigma}}^{(t+1)}$.

In order to apply this generic procedure to a particular problem, we must define a partition of the canonical variables, i.e., choose a partition of the process into subprocesses. Our analysis in the previous sections can be used to evaluate the alternatives. The tradeoffs, however, are subtle: Subprocesses with a small number of state variables allow more efficient inference. They also have a smaller transition matrix and therefore their mixing rate is likely to be better. On the other hand, our subprocesses need to be large enough so that there are no edges between subprocesses within a single time slice. Furthermore, making our subprocesses too small will increase the error incurred by the approximation of assuming them to be independent. Specifically, if we have two (sets of) variables that are highly correlated, splitting them into two separate subprocesses is probably not a good idea.

5.3 Experimental results

We have validated this algorithm in the context of two real-life DBNs: the WATER network [Jensen *et al.*, 1989], used for monitoring the biological processes of a water purification plant; and the BAT network [Forbes *et al.*, 1995], used for monitoring freeway traffic (see Figure 4). We have added a few evidence nodes to WATER, which did not have any; these duplicate a few of the state variables with added noise. The experimentation methodology is as follows: starting from the same initial prior (the uniform prior), we monitor the process evolution using our approximate method with some fixed decomposition, and compare it at every t to the exact inference, which is emulated by using our algorithm with the trivial partition (all canonical state variables in a single cluster). Observations are simulated by sampling the evidence variables according to the exact distribution.

Figure 5(a) shows the evolution of relative entropy for the BAT network on a typical run, using all the shadowed nodes on the right end of Figure 4(a) as evidence nodes. In this network, the canonical belief state consists of 10 variables roughly partitioned in two weakly interacting groups of 5: we naturally choose this partition for our approximation scheme. On an UltraSparc 2, the approximate monitoring took about 0.11 seconds per time

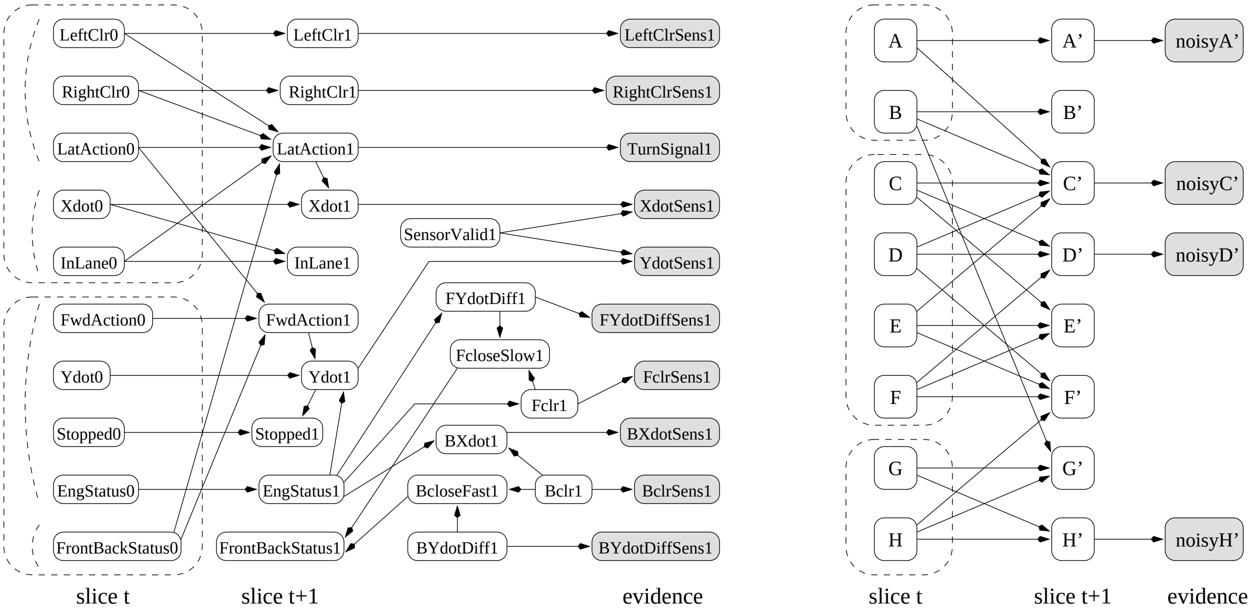

Figure 4: The canonical form 2-TBNs for the DBNs used in our experiments: (a) the BAT network; (b) the WATER network. The dotted lines indicate some of the clusterings used in our approximations.

slice, as compared to 1.72 for the exact inference, yielding a 15-fold speedup. In terms of accuracy, the error averages at 0.0007, remaining very low most of the time with, as we would expect, a few sparsely distributed spikes to somewhat larger values, peaking at 0.065. We also note that it does not appear to grow over the length of the run, as predicted by our analysis. Since in practical applications the emphasis is often on a few selected variables, we also computed the $\mathcal{L}_1$ errors for the beliefs over the two variables 'LateralAction' and 'ForwardAction' (i.e., the belief states marginalized over each of them). Their qualitative pattern was roughly similar to Figure 5(a), they respectively averaged 0.00013 and 0.0019, and remained bounded by 0.02 and 0.07 over the 1000-step run of our experiment.

Similar tests were conducted on the augmented WATER network shown on Figure 4(b), and using a decomposed belief over the 3 clusters A-B, C-D-E-F, and G-H. Over a run of length 3000, the error remained bounded by 0.06 with the exception of one outlier to 0.14, and averaged 0.006 over that run. Running times were 0.19 sec/slice for the approximation, vs. 6.02 sec/slice for the reference (a 31-fold speedup).

To investigate the effect of our approximate belief state representation, we tried three different approximation clusterings on the BAT network. The results are shown in Figure 5(b) (averaged over 8 different runs and plotted on logarithmic scale). The lower curve corresponds to the "5+5" clustering used above: at an average error of 0.0006, its error is always lower than a "3+2+4+1" clustering obtained by further breaking down each of the 5-clusters (medium curve) for which the error averages 0.015 (and whose speedup is 20 compared to

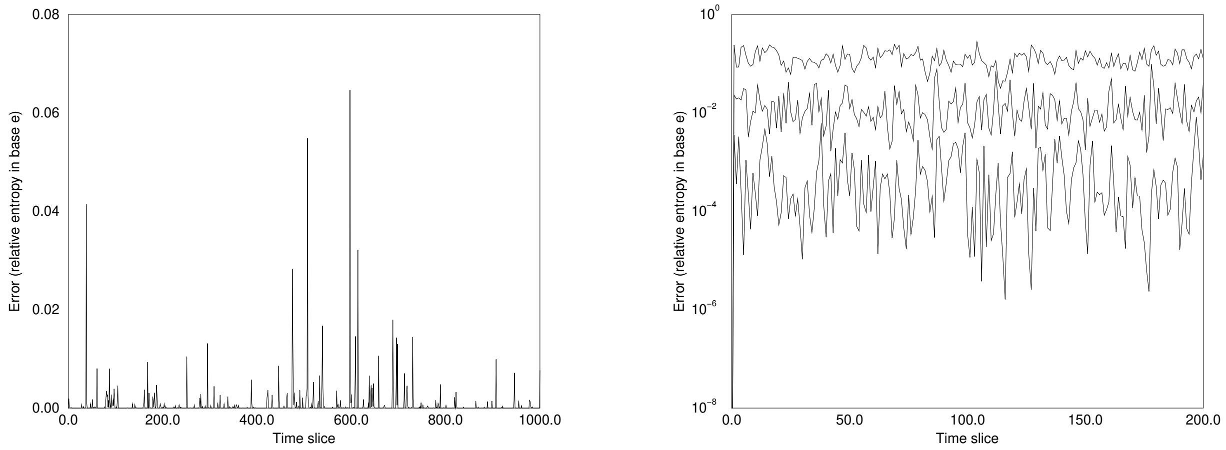

Figure 5: Experimental results: (a) relative entropy error for a typical run using the BAT network; (b) comparison of relative entropy error for three different approximate belief state representations (BAT network, curves averaged over 8 runs).

15). Both are clearly superior to the third case, a "3+3+4" clustering that bears no relationship to the connectivity of the variables in the network, and for which the average error reaches 0.13 (with a comparable speedup of 20). We observed qualitatively similar behavior in the WATER network.

Interestingly, the accuracy can be further improved by using *conditionally* independent approximations.[4]In the WATER network, for example, using a belief state decomposed into the overlapping clusters A-B-C-D-E, C-D-E-F-G, G-H yields, for the same sequence of observations as above, an average error of just 0.0015. Also, the error now remains bounded by 0.018 throughout, reducing the maximum error by a factor of 8. Approximate inference here took 0.47 sec/slice, a 13-fold speedup over exact inference.

We also tested the effect of evidence on the quality of our approximation. Without evidence, the error curve is very smooth, and converges relatively soon to a constant error, corresponding to the error between the correct stationary distribution of the process and our approximation to it. With evidence, the error curve has a much higher variance, but its average is typically much lower, indicating that evidence further boosts contraction (as opposed to merely being harmless). The effect of evidence was more beneficial with better clusterings.

[4]The conditional independence assumption requires a few modifications to Algorithm 1, since $\{\boldsymbol{X}_l\}$ no longer forms a partition of the canonical variables, but a covering with overlapping clusters. Assuming the graph of these clusters forms a tree satisfying the running intersection property, the full belief state is now given by $\varphi[X_1, ..., X_n] = \prod_l \varphi[\boldsymbol{X}_l] / \prod_{l_1 < l_2} \varphi[\boldsymbol{X}_{l_1} \cap \boldsymbol{X}_{l_2}]$, which differs from our original expression by the introduction of the denominator. Hence, in order to maintain a belief state, we also need to maintain the marginals over all non-empty intersections, and incorporate the inverse factors $(\boldsymbol{\varphi}[\boldsymbol{X}_{l_1} \cap \boldsymbol{X}_{l_2}])^{-1}$ in addition to the $\boldsymbol{\varphi}[\boldsymbol{X}_l]$ in the clique tree Υ before calibrating it.

6 Conclusion and extensions

In this paper, we investigated the effect of approximate inference in a stochastic process. We showed that the stochastic nature of the process tends to make errors resulting from an approximation disappear rapidly as the process continues to evolve. We applied this idea to the task of monitoring a stochastic process, i.e., continuously maintaining a belief state over the state at the current time. This task is known to be infeasible in complex stochastic processes involving a large number of subprocesses, since exact inference (e.g., [Kjærulff, 1992]) is forced into intractability by the full correlation of the belief state that occurs even in highly structured processes. Our approach allows us to maintain an approximate belief state in a way that guarantees that the errors from our approximations do not accumulate. Indeed, our error is significantly reduced in such processes, if we match the structure of our approximation to the structure of the process. We have shown that our approach works extremely well for real-life processes. Indeed, we get order of magnitude savings even for small processes, at the cost of a a very low error in our approximation. For larger processes, we expect the savings to be much greater.

There has been fairly little work on approximate inference in complex temporal models. [Kanazawa *et al.*, 1995] utilizes random sampling, ignoring the structure of the process entirely. [Provan, 1992] considers the idea of using domain knowledge to simply eliminate some of the variables from each time slice. [Ghahramani and Jordan, 1996] and [Saul and Jordan, 1995] utilize mean field approximation in the context of various types of HMMs. Of these approaches, [Ghahramani and Jordan, 1996] is the closest to our work. There, the compound process is also approximated as being composed of independent subprocesses, whose parameters are chosen in a way that depends on the evidence. However, this approach applies only to situations where the processes are, in fact, independent, and only become correlated due to observations. Our work applies to much richer models of partially decomposed systems. Furthermore, none of these other approaches provide an analysis of the error resulting from the approximation. Thus, our result is the first to show in a general setting how the decomposability of a stochastic process can be utilized in an inference algorithm.

One important direction in which our results can be extended relates to different representations of the belief state and of the process. Our current analysis requires that the belief state representation make the states of the subprocesses completely independent. Clearly, in many situations a more refined belief state representation is more appropriate. For example, as our experiments show, it can be very beneficial to make the states of two processes in our approximate belief state *conditionally independent* given a third; we would like to derive formal conditions for this to happen. We may also want our approximation to adjust the belief state representation to the current situation, placing more focus on representing "important" (e.g., more likely) parts of the distribution. In general, our basic contraction result applies to any approximation scheme for belief state representation. We are currently examining alternative belief state representations that might be more suitable for stochastic processes. Finally, we would like to apply our analysis to other types of processes requiring a compact representation of the belief state, e.g., processes with a continuous state space.

Another priority is to improve our analysis of the effect of evidence. Our current results only show that the evidence does not hurt too much. Our experiments, however, show (as we would expect) that the evidence can significantly reduce the overall error of our approximation. We would like to analyze this effect formally.

There are many other tasks to which our techniques can be applied. For example, our analysis applies as is to the task of predicting the future evolution of the system. It also applies to tasks where the transition model depends on some action taken by an agent. Our contraction analysis is done on a phase-by-phase basis, and therefore does not require that the transition model be the same at each time slice. Thus, so long as the transition associated with each action is sufficiently stochastic, we can use our technique to effectively predict the effects of a plan or a policy. A more speculative idea is in the domain of reasoning in POMDPs (Partially Observable Markov Decision Processes), where a policy is a mapping from belief states to actions. Perhaps our more compact belief state representation will turn out to be a better basis for representing a policy.

We hope to generalize our contraction result to the case of reasoning backwards in time, showing that the influence of approximations in the future cannot significantly affect our beliefs about the distant past. By using an approximation that simply ignores future observations, we could obtain accurate beliefs about a time slice from a fairly small window on both sides. The applications of this idea to the task of learning the process are particularly intriguing. In this case, we are often faced with long sequences of observations corresponding to long trajectories of the system. We would prefer to be able to learn the process online, without having to first reason about the entire trajectory. We believe that our ideas will allow the process to be learned effectively from small windows, with only minimal degradation of quality.

Acknowledgements

We gratefully acknowledge Eric Bauer, Lise Getoor, and Uri Lerner for work on the software used in the experiments, Raya Fratkina for help with the network files, and Tim Huang for providing us with the BAT network. Many thanks to Tom Cover, Nir Friedman, Alex Kozlov, Uri Lerner, and Stuart Russell for useful discussions and comments.

References

[Aström, 1965] K.J. Aström. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Applic.*, 10:174–205, 1965.

[Cover and Thomas, 1991] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[Dagum *et al.*, 1992] P. Dagum, A. Galper, and E. Horwitz. Dynamic network models for forecasting. In *Proc. UAI*, 1992.

[Dean and Kanazawa, 1989] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Comp. Int.*, 5(3), 1989.

[Forbes *et al.*, 1995] J. Forbes, T. Huang, K. Kanazawa, and S.J. Russell. The BATmobile: Towards a Bayesian automated taxi. In *Proc. IJCAI*, pages 1878–1885, 1995.

[Ghahramani and Jordan, 1996] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. In *NIPS 8*, 1996.

[Huang and Darwiche, 1994] C. Huang and A. Darwiche. Inference in belief networks: a procedural guide. *Int. J. Approx. Reas.*, 11:1–158, 1994.

[Jensen *et al.*, 1989] F.V. Jensen, U. Kjærulff, K.G. Olesen, and J. Pedersen. An expert system for control of waste water treatment— a pilot project. Technical report, Judex Datasystemer A/S, Aalborg, Denmark, 1989. In Danish.

[Kalman, 1960] R.E. Kalman. A new approach to linear filtering and prediction problems. *J. of Basic Engineering*, 1960.

[Kanazawa *et al.*, 1995] K. Kanazawa, D. Koller, and S.J. Russell. Stochastic simulation algorithms for dynamic probabilistic networks. In *Proc. UAI*, pages 346–351, 1995.

[Kjærulff, 1992] U. Kjærulff. A computational scheme for reasoning in dynamic probabilistic networks. In *Proc. UAI*, pages 121–129, 1992.

[Lauritzen and Spiegelhalter, 1988] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Stat. Soc.*, B 50:157–224, 1988.

[Provan, 1992] G. Provan. Tradeoffs in constructing and evaluating temporal influence diagrams. In *Proc. UAI*, pages 40–47, 1992.

[Rabiner and Juang, 1986] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE Acoustics, Speech & Signal Processing*, 1986.

[Saul and Jordan, 1995] L.K. Saul and M.I. Jordan. Exploiting tractable substructure in intractable networks. In *NIPS 7*, 1995.

[Smyth *et al.*, 1996] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1996.