To appear in *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, Florida, July 18-22, 1999

Exploiting the Architecture of Dynamic Systems

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

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

Abstract

Consider the problem of monitoring the state of a complex dynamic system, and predicting its future evolution. Exact algorithms for this task typically maintain a *belief state*, or distribution over the states at some point in time. Unfortunately, these algorithms fail when applied to complex processes such as those represented as dynamic Bayesian networks (DBNs), as the representation of the belief state grows exponentially with the size of the process. In (Boyen & Koller 1998), we recently proposed an efficient approximate tracking algorithm that maintains an *approximate* belief state that has a compact representation as a set of independent factors. Its performance depends on the error introduced by approximating a belief state of this process by a factored one. We informally argued that this error is low if the interaction between variables in the processes is "weak". In this paper, we give formal information-theoretic definitions for notions such as weak interaction and sparse interaction of processes. We use these notions to analyze the conditions under which the error induced by this type of approximation is small. We demonstrate several cases where our results formally support intuitions about strength of interaction.

Introduction

Consider an intelligent agent whose task is to monitor a complex dynamic system such as a freeway system with multiple vehicles (Forbes *et al.* 1995). Tracking the state of such systems is a difficult task: their dynamics are noisy and unpredictable, and their state is only partially observable. Stochastic processes provide a coherent framework for modeling such systems. In many cases, the state of the system is represented using a set of *state variables*, where individual state are assignments of values to these variables. *Dynamic Bayesian networks (DBNs)* (Dean & Kanazawa 1989) allow complex systems to be represented compactly by exploiting the fact that each variable typically interacts only with few others.

Unfortunately, although this type of limited interaction helps us achieve a compact representation, it does not support effective inference. Consider the task of maintaining a *belief state* — a distribution over the current process state (Aström 1965). A naive representation of such a distribution is exponential in the number of state variables. Unfortunately, it can be shown that, unless the system is completely decoupled (i.e., composed of non-interacting subprocesses), any two variables will have some common influence in the past and will thus be correlated. The belief state therefore has no structure, and can only be represented as an explicit joint distribution over the system variables. This limitation renders algorithms that try to track the system exactly (Kjærulff 1992) impractical for complex problems.

However, one has a strong intuition that keeping track of these correlations is often unnecessary. While the variables might be correlated, this correlation is often very weak. In Herbert Simon's words, these are "nearly decomposable systems, in which the interactions among the subsystems are weak but not negligible" (Simon 1962). Simon argues that these "nearly decomposable systems are far from rare. On the contrary, systems in which each variable is linked with almost equal strength with almost all other parts of the system are far rarer and less typical."

In (Boyen & Koller 1998) — hereafter, BK — we propose an algorithm that exploits this idea of weak interaction by momentarily ignoring the weak correlations between the states of different system components. More precisely, the BK algorithm represents the belief state over the entire system as a set of localized beliefs about its parts. For example, it might represent the beliefs about the freeway as a set of independent beliefs about the state of the individual vehicles; or, more appropriately, the states of the vehicles might be represented as conditionally independent given the overall traffic load. The algorithm chooses a restricted class of factored belief states. Given a time t belief state in this class, it propagates it to time $t+1$; this step typically has the effect of inducing correlations between the subsystems. The algorithm projects the resulting distribution back into the restricted space. Note that the correlations between subsystems are not eliminated; they are merely "summarized" at every point in time by the projection step.

The analysis in BK shows that the stochasticity of the process prevents the repeated errors resulting from the projection steps at every time t from accumulating unboundedly. However, the amount of error resulting from the approximation is not quantified. Rather, the justification is based on the intuition that, if the processes interact only weakly, the error cannot be too large. In order to make this intuition precise, we must formally define what it means for processes to interact weakly, and show that weak interaction does allow us to bound the error introduced by this approximation.

We provide a formal information-theoretic notion of interaction, that corresponds to the amount of correlation between subsystems that is generated in a single step of the process. We then use this idea to provide a quantitative measure for the strength of interaction between systems. We also analyze the case of two processes whose correlation is largely mediated by a third; we show that such processes can be approximated as conditionally independent given the third if the latter evolves more slowly than they do, and thereby "remembers" its state. These notions allow us to determine the error induced by a decoupled approximation to the belief state.

We also analyze a new notion of *sparse interaction,* where subprocesses mostly interact only weakly, but have an occasional strong interaction. Indeed, this type of interaction might be more accurate as a fine-grained traffic model, as individual cars do occasionally have a very strong interaction (e.g., when one makes an emergency stop directly in front of another). In this case, the weak interaction assumption is warranted only part of the time. We extend the BK algorithm to settings such as this. The algorithm tailors the approximation it uses to the circumstances; after a strong interaction between two subsystems takes place, it stops decoupling their states. Thus, it temporarily resorts to a different approximation structure. If the strong interaction is momentary, then the processes go back to their usual mode of weak interaction. Hence, after some amount of time, the correlation attenuates. At that point, we can go back to decoupling the subprocess states. Our analysis shows how long this coupling needs to last in order to guarantee that we incur only small error by decoupling the processes.

Our results show how the architecture of a dynamic system can be exploited to provide an effective algorithm for reasoning about it. The system structure can be used to select an approximation scheme appropriate to it, and to adapt it as the system evolves. Our analysis provides error bounds, allowing a tradeoff between accuracy and computation.

Approximate inference in DBNs

In this section, we review the basic definitions of dynamic systems represented compactly as *dynamic Bayesian networks*. We also review the approximate inference of the BK algorithm, which is the starting point for our analysis.

A stochastic dynamic system is defined via a set of states, and a *transition model* that represents the way in which one state leads to the next. In complex systems, a state is best described using a set of *random variables* $A_1, \ldots, A_n$. We use U, V, W, X, Y, Z to denote sets of random variables, and their lower case version to denote instantiations of values for the variables in the set. The transition model is described via a directed acyclic graph $\mathcal{B}$. The network contains nodes $A_1, \ldots, A_n$ reprsenting the current state, and $A'_1, \ldots, A'_n$ representing the next state. Each node A'_i has a set of parents $\mathrm{Pa}(A'_i)$; nodes A_i have no parents. The network represents the qualitative structure of the transition model — the variables that directly influence the new value of each variable A'_i. The transition model is made quantitative by associating with each variable A'_i a conditional probability table $\boldsymbol{P}[A'_i \mid \mathrm{Pa}(A'_i)]$.

Our goal in many dynamic systems is *monitoring*: keeping track of the state of the system as it evolves. In general, we do not have access to the full state of the system. Rather, we get to observe only some subset of the state variables. Thus, the best we can do is to maintain a *belief state* — a probability distribution $\mu^{(t)}$ over the possible states at the current time t. In principle, the process of maintaining a belief state is straightforward. Having computed $\mu^{(t)}$, we propagate it forward using the transition model to obtain the expected next belief state $\mu^{(\bullet t+1)}$; we then condition $\mu^{(\bullet t+1)}$ on our time $t+1$ evidence to get $\mu^{(t+1)}$.

In practice, however, this process can be very computationally intensive. The problem is that $\mu^{(t)}$ is a distribution over all possible assignments of values to $A_1, \ldots, A_n$, i.e., an exponentially sized space. One might hope that this belief state can be represented compactly. After all, the transition model is structured; perhaps that also induces structure on the belief state, allowing a compact representation. Unfortunately, despite the limited interaction that the transition model induces between the variables, they all become correlated. Intuitively, unless the system is completely decoupled into noninteracting subprocesses, any two variables $A_i^{(t)}$ and $A_j^{(t)}$ will eventually be influenced by a common cause, somewhere in the history of the process. Regardless of how long ago that was, and how weak the correlation currently is, the variables are qualitatively correlated. As any decomposition of a distribution rests on some form of conditional independence structure, no factored representation of the belief state is possible.

In BK, we propose an approach for circumventing this problem. Our algorithm maintains an *approximate* belief state that admits a factored representation. Specifically, we consider belief states that fall into some restricted family of distributions Σ, e.g., ones where certain sets of variables are marginally independent. Let $\tilde{\mu}^{(t)} \in \Sigma$ be our current approximate to the belief state. When we transition it to the next time slice, the result is a distribution $\varphi^{(t+1)}$ which is usually not in Σ. We must therefore project it back into Σ. We now make this algorithm more precise.

Definition 1 A *cluster forest* (Jensen, Lauritzen, & Olesen 1990) $\mathcal{F}$ is an undirected forest whose nodes are *clusters* $F_1, \ldots, F_m \subset \{A_1, \ldots, A_n\}$ and whose edges are $E = \{(i,j)\}$. The forest has the *running intersection property* — if F_i and F_j are clusters such that $A_k \in F_i$ and $A_k \in F_j$, then every cluster on the path between F_i and F_j also contains A_k.

Definition 2 We say that a distribution ψ is *representable* over $\mathcal{F}$ if it is represented as a set of marginals ψ_i over the clusters F_i, which are *calibrated,* i.e., $\psi_i[F_i \cap F_j] =$

$\psi_j[F_i \cap F_j]$ for any i, j. The distribution ψ is defined as:

$$\psi(A_1, \ldots, A_n) = \frac{\prod_{i=1}^{m} \psi[F_i]}{\prod_{(i,j)\in E} \psi_i[F_i \cap F_j]}.$$

We define $\Sigma[\mathcal{F}]$ to be the set of distributions ψ that are representable over $\mathcal{F}$.

The BK algorithm takes the approximate belief state $\tilde{\mu}^{(t)}$ in $\Sigma[\mathcal{F}]$, and generates the approximate belief state $\tilde{\mu}^{(t+1)}$ in $\Sigma[\mathcal{F}]$, as follows. In the first phase, the algorithm propagates $\tilde{\mu}^{(t)}$ to $\varphi^{(t+1)}$ using the transition model. It then projects $\varphi^{(t+1)}$ into $\Sigma[\mathcal{F}]$, generating $\psi^{(t+1)}$. Finally, it conditions on the time $t+1$ evidence, resulting in $\tilde{\mu}^{(t+1)}$.

In order for this process to be performed correctly, we require that, if $A'_k \in \mathrm{Pa}(A'_l)$, i.e., A_l has a parent A_k in its own time slice, then there must be some cluster F_i such that A_k and A_l are both in F_i. That is, all intra-time-slice edges must be contained in some cluster. This assumption allows us to focus attention on inter-time-slice influences. We therefore define $\mathrm{Pa}^-(A'_l)$ to be $\mathrm{Pa}(A'_l) \cap \{A_1, \ldots, A_n\}$, and $\mathrm{Pa}^-(Y')$ for a set of variables Y' analogously.

The potential problem with this approach is that the repeated approximation at every time slice t could accumulate unboundedly, resulting in a meaningless approximation. In BK, we analyze this algorithm, and provide conditions under which the error remains bounded. The first condition is that the process is somewhat stochastic, so that errors from the past are "forgotten." The second is that each approximation step does not introduce too much error.

Definition 3 Let Y' be a cluster at time $t+1$, and X be $\mathrm{Pa}^-(Y')$. We define the *mixing rate* of the generalized transition $X \to Y'$ as

$$\gamma[X \to Y'] \triangleq \min_{x_1, x_2} \sum_y \min[\boldsymbol{P}[y \mid x_1], \boldsymbol{P}[y \mid x_2]].$$

If $X = V \cup W$, we also define the *mixing rate* of the conditional transition $V \to Y'$ as the minimal mixing rate obtained over all possible values of W;

$$\gamma[V \to Y' \mid W] \triangleq \min_w \min_{v_1, v_2} \left[\sum_y \min[\boldsymbol{P}[y \mid v_1, w], \boldsymbol{P}[y \mid v_2, w]] \right].$$

Intuitively, the mixing rate is the minimal amount of mass that two distributions over Y' are guaranteed to have in common: one is the distribution we would get starting at x_1 and the other the one starting at x_2. The minimal mixing rate in the conditional transition is similar, except that we now restrict to starting points that agree about the variables in W. From here on, we will often drop the explicit reference to W in the notation for mixing rate, as W is defined implicitly to be $\mathrm{Pa}^-(Y') \setminus V$.

The mixing rate can be used to bound the rate at which errors arising from approximations in the past are "forgotten." Let $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$ be the finest disjoint partition of $A_1, \ldots, A_n$, such that each cluster F_i is contained in some Q_j; i.e., each Q_j is one of the connected components — or *trees* — in the forest defined by $\mathcal{F}$. Let r be the maximum *inward connectivity* of the process relative to $\mathcal{Q}$, i.e., an upper bound, over all partitions Q'_j, on the number of partitions Q_i such that there is an edge from (a variable in) Q_i to (a variable in) Q'_j. Similarly, let q be the maximum *outward connnectivity* of the process relative to $\mathcal{Q}$, i.e., an upper bound, over all Q_i, on the number of Q'_j such that there is an edge from Q_i to Q'_j. We define:

$$\gamma^* \triangleq \left(\frac{1}{r} \min_l \gamma[Q_l \to Q'_l] \right)^q.$$

Based on this definition, we prove that the stochastic transition decreases the error between the two distributions, measured as their Kullback-Leibler divergence. The *KL divergence* (*relative entropy*) (Cover & Thomas 1991) between a reference distribution μ and another $\tilde{\mu}$ is:

$$\boldsymbol{D}[\mu \,\|\, \tilde{\mu}] \triangleq \boldsymbol{E}_\mu \left[\ln \frac{\mu(s)}{\tilde{\mu}(s)} \right] = \sum_s \mu(s) \cdot \ln \frac{\mu(s)}{\tilde{\mu}(s)}.$$

Theorem 1 (Boyen & Koller 1998)

$$\boldsymbol{D}[\mu^{(t+1)} \,\|\, \varphi^{(t+1)}] \le (1-\gamma^*) \cdot \boldsymbol{D}[\mu^{(t)} \,\|\, \tilde{\mu}^{(t)}].$$

Of course, the time $t+1$ approximation step — going from $\varphi^{(t+1)}$ to $\psi^{(t+1)}$ — introduces a new error into our approximate belief state. We can show that if this error is bounded, then the overall error in our approximation also remains bounded.

Definition 4 The *implicit projection error* of approximating φ by ψ with respect to a "true" distribution μ, as

$$\varepsilon_\mu(\varphi \mapsto \psi) \triangleq \boldsymbol{E}_\mu \left[\ln \frac{\varphi(s)}{\psi(s)} \right].$$

Theorem 2 (Boyen & Koller 1998) *Let ε^* be a bound on $\varepsilon_{\mu^{(\bullet t)}}(\varphi^{(t)} \mapsto \psi^{(t)})$ for all t. Then, on expectation over the sequence of observations, for all t:*

$$\boldsymbol{E}\, \boldsymbol{D}[\mu^{(t)} \,\|\, \tilde{\mu}^{(t)}] \le \varepsilon^* / \gamma^*.$$

Note that, since $\varphi^{(t)}$ and $\psi^{(t)}$ are distributions prior to conditioning on the time t observation, the implicit projection error should be taken relative to $\mu^{(\bullet t)}$ — the true time t belief state prior to the conditioning step.

Measuring interaction

In BK, we provide an analysis for the contraction rate γ^*, allowing it to be bounded in terms of parameters of the dynamic system. We do not provide a similar analysis for ε^*. Rather, we argue that, if the processes do not interact very strongly, such an approximation does not incur too large an error. Indeed, our experimental results support this prediction. Our goal in this paper is to try to analyze this error and to relate it to the structure of the process.

The problem with analyzing the implicit error is that it is, as its name suggests, implicit. As we do not have access to the true distribution $\mu^{(\bullet t)}$, we cannot measure the error incurred by the projection. Instead, we will analyze a closely related quantity — the KL divergence from φ to ψ. Then, we show how this projection error can be analyzed in terms of the architecture of the system and its dynamics.

Definition 5 We define the *projection error* of approximating φ by ψ as

$$\varepsilon(\varphi \mapsto \psi) \triangleq \boldsymbol{D}[\varphi \,\|\, \psi] = \boldsymbol{E}_{\varphi}\left[\mathit{ln}\, \frac{\varphi(s)}{\psi(s)} \right].$$

Although the projection error is not precisely the quantity that appears in the analysis of the BK algorithm, there are good reasons for believing it to be close. In particular, if our current approximation $\tilde{\mu}^{(t)}$ is fairly close to the true distribution $\mu^{(t)}$, then our estimate of the projection error relative to $\varphi^{(t)}$ is close to the implicit approximation error relative to $\mu^{(t)}$. In this case, if we guarantee that the projection error is small, we can show that $\tilde{\mu}^{(t+1)}$ remains close to $\mu^{(t+1)}$. We are currently working on formalizing this intuition. For now, we will analyze the projection error.

The key aspect of this analysis is based on a close relation between the projection error and mutual information between clusters in the cluster forest $\mathcal{F}$.

The *mutual information* between two (sets of) variables X and Y given Z, is defined as (Cover & Thomas 1991):

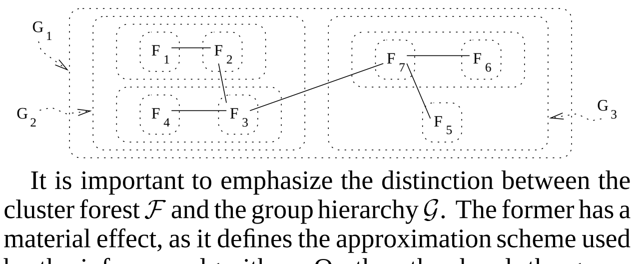
$$\boldsymbol{I}[X;Y \mid Z] \triangleq \boldsymbol{E}_Z \boldsymbol{D}[\boldsymbol{P}[X,Y \mid Z] \,\|\, \boldsymbol{P}[X \mid Z] \otimes \boldsymbol{P}[Y \mid Z]],$$

where $\otimes$ is the outer product.

We now show that the projection error ε can be decomposed according to our clustering. In fact, the decomposition is done on a related structure, which is a hierarchical grouping of the clusters of $\mathcal{F}$.

Definition 6 We define a *cluster hierarchy* $\mathcal{G}$ as a binary tree whose leaves are the clusters $F_1, \ldots, F_m$, such that, for every pair of sibling subtrees, there is at most one edge in the cluster forest between those clusters at the leaves of one sibling and those at the leaves of the other. Each interior node is associated with a *(cluster) group* G_i, which is the union of the clusters F_k at its leaves. We say that the groups G_i and G_j for sibling subtrees are *sibling groups*, and use R to denote the set of the $m-1$ pairs (i,j) such that G_i and G_j are siblings. For $(i,j) \in R$, we denote by M_{ij} the intersection $G_i \cap G_j$, and by $G_{i \setminus j}$ the difference $G_i \setminus M_{ij}$.

Intuitively, $\mathcal{G}$ is a recursive partition of $\mathcal{F}$, where each split divides up the clusters of a group G_k into a pair of "reciprocal" sub-groups G_i, G_j, so that no more than one edge of $\mathcal{F}$ is broken by the partition (in which case G_i and G_j will share the variables shared by the clusters on the broken edge). The following picture shows one possible cluster hierarchy (dotted lines) for a given cluster forest (solid lines).



It is important to emphasize the distinction between the cluster forest $\mathcal{F}$ and the group hierarchy $\mathcal{G}$. The former has a material effect, as it defines the approximation scheme used by the inference algorithm. On the other hand, the group hierarchy $\mathcal{G}$ is merely used for the purpose of analysis, and can be chosen freely once $\mathcal{F}$ is fixed. The nature of $\mathcal{F}$ and $\mathcal{G}$ is also different: the clusters in $\mathcal{F}$ may or may not be overlapping, and if they are, they must satisfy the running intersection property; in contrast, some groups in $\mathcal{G}$ are necessarily overlapping, since each group is a proper subset of its parent in the hierarchy. The key insight is that the approximation error for using the clusters in $\mathcal{F}$ decomposes nicely according to the structure of $\mathcal{G}$, as we now show.

Theorem 3 *Let φ be a distribution and ψ its projection on the cluster forest $\mathcal{F}$. Then the projection error*

$$\varepsilon(\varphi \mapsto \psi) = \sum_{(i,j)\in R} \boldsymbol{I}[G_i; G_j \mid M_{ij}],$$

where the mutual informations are computed with respect to the distribution φ.

Proof sketch We use an induction argument. Let G_k be any interior node, and G_i, G_j the children of G_k. Since M_{ij} is fully contained in some cluster F_l, we have $\psi[M_{ij}] = \varphi[M_{ij}]$. Therefore,

$$\begin{aligned}
&\boldsymbol{D}[\varphi[G_k] \,\|\, \psi[G_k]] \\
&\quad = \boldsymbol{E}_{\varphi(G_k)}[\ln(\varphi(G_k)/\varphi(M_{ij})\varphi(G_i \mid M_{ij})\varphi(G_j \mid M_{ij})) \\
&\qquad\quad + \ln(\varphi(M_{ij})\varphi(G_i \mid M_{ij})/\psi(M_{ij})\psi(G_i \mid M_{ij})) \\
&\qquad\quad + \ln(\varphi(M_{ij})\varphi(G_j \mid M_{ij})/\psi(M_{ij})\psi(G_j \mid M_{ij}))] \\
&\quad = \boldsymbol{I}[G_i; G_j \mid M_{ij}] \\
&\qquad + \boldsymbol{D}[\varphi(G_i) \,\|\, \psi(G_i)] + \boldsymbol{D}[\varphi(G_j) \,\|\, \psi(G_j)].
\end{aligned}$$

The claim follows by recursion on the last two terms, noticing that for any cluster F_l, $\boldsymbol{D}[\varphi(F_l) \,\|\, \psi(F_l)] = 0$. ∎

The key to exploiting this decomposition is the following. Recall that $\varphi^{(t+1)}$ is obtained from propagating a distribution $\psi^{(t)}$. The distribution $\psi^{(t)}$ is in the restricted space $\Sigma[\mathcal{F}]$, i.e., it satisfies the independence assumptions defined by $\mathcal{F}$. Intuitively, there is a limit to the amount of dependencies introduced by a factored stochastic transition on a factored distribution. This should result in bounds on each of the mutual information terms that appear in the theorem. Our analysis will make this idea formal.

It turns out that the notion of mixing rate, which captures the extent to which information is retained from a set of variables X to a set of variables Y', can also be viewed as representing the extent to which a set of variables influences another. In particular, we are interested in the extent to which one group G_i at time t influences another group G_j at time t. We therefore define

$$\gamma_{ij} \triangleq \gamma[(G_i \setminus M_{ij}) \to (G'_j \setminus M'_{ij})].$$

(As usual, the dependence on other parents of G'_j is implicit.)

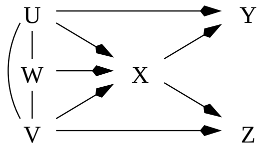
Theorem 3 shows that the projection error is decomposed as a sum of conditional mutual informations — one term for each pair of siblings G_i, G_j. In the rest of the paper, we shall derive bounds on those mutual information terms.

Weak interaction

We begin with analyzing the error for two groups that our approximation takes to be completely independent, i.e., groups contained in different connected components of the cluster forest. Intuitively, we would expect the error in this case to depend on the extent to which these two groups interact. In other words, if our system is such that the variables in these two groups interact only weakly, the error incurred by assuming them to be independent is small.

We first state a central lemma for this kind of analysis.

Lemma 4 *Let U, V, W, X, Y, Z be sets of random variables with the following dependency structure:*

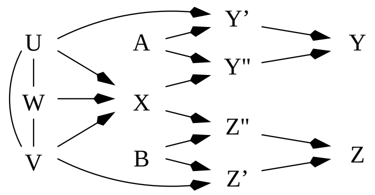


Then, writing γ_{UY} for $\gamma[U \to Y]$, etc.,

$$\begin{aligned}\boldsymbol{I}[Y;Z \mid W] &\le (1-\gamma_{UY})\cdot(1-\gamma_{VZ})\cdot \boldsymbol{I}[U;V \mid W] \\ &\quad +3\cdot \ln|\mathrm{dom}\, X|.\cdot(1-\min[\gamma_{XY},\gamma_{XZ}])\end{aligned}$$

Proof sketch We pose $\gamma_Y = \gamma[(U,X) \to Y]$ and $\gamma_Z = \gamma[(V,X) \to Z]$.

To start with, we observe that we can decompose the given system as



Each of Y' and Y'' either copies the value of its parent, or enters a special "contraction state" c, depending on the value of A. The domain of A has four values, respectively triggering the contraction for Y', Y'', both, or none. These values are distributed with probabilities $(\gamma_{UY}-\gamma_Y), (\gamma_{XY}-\gamma_Y), \gamma_Y, (1-\gamma_{UY}-\gamma_{XY}+\gamma_Y)$ respectively. We can show that all of these quantities are non-negative, so that they form a well-defined distribution.

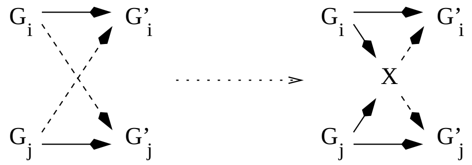
It suffices to show that: (i) One can construct the conditional probabilities of Y and Z in the new network so as to emulate the joint distribution specified by the original network. (ii) The mutual information $\boldsymbol{I}[Y;Z \mid W]$ computed in the new network is bounded as claimed. We defer details to a longer version of this paper. ∎

We are now ready to prove our theorem.

Theorem 5 *Let $\mathcal{F}$ and $\mathcal{G}$ be a cluster forest and group hierarchy, and G_i and G_j two siblings contained in different connected components of $\mathcal{F}$. Let $\tilde{\mu}$ be a distribution factored according to $\mathcal{F}$. Let φ be obtained from $\tilde{\mu}$ by propagation through the given transition model. Then, with respect to φ,*

$$\boldsymbol{I}[G'_i;G'_j] \le 3\cdot\ln|\mathrm{dom}\,(G_i\cup G_j)|\cdot(1-\min[\gamma_{ij},\gamma_{ji}]).$$

Proof sketch We consider the transition model involving G_i and G_j at t and $t+1$. The idea is to transform it into an equivalent model by introducing a "mediator" variable X through which the cross-interaction between G_i and G_j is funneled.



Specifically, X simply copies the values of G_i and G_j, and the new edges from X to G'_i and G'_j reproduce the previous cross-dependences $G_i \to G'_j$ and $G_j \to G'_i$. Notice that γ_{Xi} and γ_{Xj} in the new model are respectively equal to γ_{ji} and γ_{ij} in the original one. Then, an application of Lemma 4 to the new structure gives

$$\begin{aligned}&\boldsymbol{I}[G'_i;G'_j] \\ &\quad\le 3\cdot(1-\min[\gamma_{Xi},\gamma_{Xj}])\cdot\ln|\mathrm{dom}\,X| + c\cdot\boldsymbol{I}[G_i;G_j] \\ &\quad= 3\cdot(1-\min[\gamma_{Xi},\gamma_{Xj}])\cdot\ln|\mathrm{dom}\,(G_i\cup G_j)|,\end{aligned}$$

where we have used the fact that $\boldsymbol{I}[G_i;G_j]=0$ since G_i and G_j are independent in the belief state representation at time t. ∎

Note that, as G_i and G_j are disjoint, we have $M_{ij}=\emptyset$. Thus, the term $\boldsymbol{I}[G'_i;G'_j]$ bounded in this theorem is precisely the term $\boldsymbol{I}[G_i;G_j \mid M_{ij}]$ that appears in Theorem 3. In other words, Theorem 5 gives us a bound on the error introduced by two specific groups G'_i, G'_j. To get the overall bound on the error, we simply add the contributions of all pairs of siblings.

The bound for G'_i and G'_j closely matches our intuition regarding their "strength of interaction." To understand this, consider the term γ_{YX} for two groups X and Y that are "weakly interacting". In this case, we believe that Y is not a strong influence on X, i.e., $\boldsymbol{P}[X' \mid x, y_1]$ is close to $\boldsymbol{P}[X' \mid x, y_2]$ for any x, y_1, and y_2. But in this case, $\sum_{x'}\min[\boldsymbol{P}[x' \mid x,y_1], \boldsymbol{P}[x' \mid x,y_2]]$ is close to one for all x, y_1, y_2, and hence so is γ_{YX}. If both γ_{ij} and γ_{ji} are close to one, the error bound in our analysis will be close to zero.

To illustrate, consider the process composed of a number of cars on a highway. In normal circumstances, the cars interact weakly with each other, so we want to place each car in a separate cluster F_i in our belief state representation. We can use the above theorem to justify this, as the weak interaction between the cars will ensure that each $\gamma_{ij} \simeq 1$ in any group hierarchy $\mathcal{G}$ that we choose. In fact, since the choice of $\mathcal{G}$ is arbitrary given a clustering $\mathcal{F}$, we can choose $\mathcal{G}$ to maximize the various γ_{ij}. In our highway example, it is reasonable to assume that only neighboring cars may experience any kind of (weak) interaction. We can maximize γ_{ij} by minimizing the number of neighboring cars belonging to any two siblings G_i and G_j. This is very intuitive: we simply group cars according to their proximity.

Conditional weak interaction

The previous section analyzed the error of approximating clusters of variables as completely independent. However, as we show experimentally in BK, we can sometimes obtain much lower errors by approximating distributions (or parts of them) as *conditionally independent*. For example, it may be much more reasonable to have an approximate belief states where the states of individual cars are conditionally independent given the overall traffic on the road. In this case, our cluster forest would contain a cluster for each vehicle, which also contains the *Traffic* random variable. In this case, the clusters are overlapping, which will cause some siblings to overlap in $\mathcal{G}$. We therefore analyze the error bound for two groups that need not be disjoint.

The *conditional entropy* of X given Y, denoted $\boldsymbol{H}[X \mid Y]$, is defined as

$$\boldsymbol{H}[X \mid Y] \triangleq \boldsymbol{E}_Y \boldsymbol{E}_{X|Y}\left[\ln \frac{1}{\boldsymbol{P}[X \mid Y]}\right].$$

Lemma 6 *Let W, X, Y, Z four sets of random variables with an arbitrary dependency structure. Then,*

$$|\boldsymbol{I}[Y;Z \mid X] - \boldsymbol{I}[Y;Z \mid W]| \leq \boldsymbol{H}[X \mid W] + \boldsymbol{H}[W \mid X].$$

Theorem 7 *Let $\mathcal{F}$ be a cluster forest, $\mathcal{G}$ a cluster hierarchy, and G_i and G_j two siblings in $\mathcal{G}$. Let $\tilde{\mu}$ and φ be defined as in Theorem 5. Then, with respect to φ, we have*

$$\begin{aligned} &\boldsymbol{I}[G'_i; G'_j \mid M'_{ij}] \\ &\quad \leq \quad 3 \cdot \ln|\mathrm{dom}\,(G_{i\setminus j} \cup G_{j\setminus i})| \cdot (1 - \min[\gamma_{ij}, \gamma_{ji}]) \\ &\qquad\quad + \boldsymbol{H}[M_{ij} \mid M'_{ij}] + \boldsymbol{H}[M'_{ij} \mid M_{ij}]. \end{aligned}$$

Proof sketch The proof is based on a similar construction as in Theorem 5, introducing a mediator variable X to capture the cross-interactions between $G_{i\setminus j}$ and $G_{j\setminus i}$. Using Lemma 4, we obtain

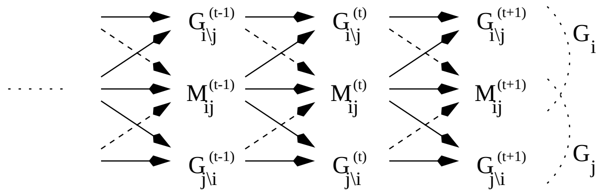
$$\begin{aligned} &\boldsymbol{I}[G'_i; G'_j \mid M_{ij}] \\ &\quad \leq \quad 3 \cdot \ln|\mathrm{dom}\,(G_{i\setminus j} \cup G_{j\setminus i})| \cdot (1 - \min[\gamma_{ij}, \gamma_{ji}]). \end{aligned}$$

Applying Lemma 6, we get

$$\begin{aligned} \boldsymbol{I}[G'_i; G'_j \mid M'_{ij}] \quad \leq \quad & \boldsymbol{I}[G'_i; G'_j \mid M_{ij}] \\ & + \boldsymbol{H}[M_{ij} \mid M'_{ij}] + \boldsymbol{H}[M'_{ij} \mid M_{ij}], \end{aligned}$$

so the claim follows. ∎

Let us examine the result of the theorem from an intuitive standpoint. The first term was already present in Theorem 5 and represents the amount of correlation introduced by the weak interaction. The second term is new: it represents the amount by which conditioning on M'_{ij} instead of M_{ij} might change the mutual information. Intuitively, if M'_{ij} is a faithful (deterministic) copy of M_{ij}, then conditioning on one or the other should not make any difference. In this case, indeed, we would have both conditional entropies equal to zero. This behavior generalizes to more realistic situations, where M_{ij} does evolve over time, but more slowly than the two clusters it separates. More precisely, let us assume that G_i and G_j interact only through M_{ij}, and that M_{ij} tends to preserve its value from one step to the next. (In particular, this implies that all external influences on M_{ij} are weak.)



The processes G_i and G_j are conditionally independent given the entire sequence $M^{(\cdot)}_{ij}$. Assuming that $G_{i\setminus j}$ and $G_{j\setminus i}$ are mixing fast enough, $G^{(t)}_i$ and $G^{(t)}_j$ will be approximately independent given the values of $M^{(\cdot)}_{ij}$ in a vincinity of t. If M_{ij} evolves slowly (implying $\boldsymbol{H}[M_{ij} \mid M'_{ij}] \simeq 0$, $\boldsymbol{H}[M'_{ij} \mid M_{ij}] \simeq 0$), these values of $M^{(\cdot)}_{ij}$ will be approximately determined by the knowledge of $M^{(t)}_{ij}$, so that $G^{(t)}_i$ and $G^{(t)}_j$ are approximately independent given the single point $M^{(t)}_{ij}$. The same analysis holds if $G_{i\setminus j}$ and $G_{j\setminus i}$ do interact directly, but only weakly.

In the real world, we find many examples of processes whose primary interaction is via some more slowly evolving process. Our freeway example is a typical one: we can refine the model of the previous section if we consider external influences that affect some or all cars in mostly the same way, such as road work or the weather. Different stocks on the stock market is another: the price trend of different stocks is clearly correlated, but they are reasonably modeled as conditionally independent given the current market trends. In both cases, the conditioning variables fluctuate more slowly than the dependent subprocesses. In these examples, our model will contain a number of clusters $F_1, \ldots, F_k$ that all contain some variable W, which will therefore appear in the M_{ij} at one or more levels in the group hierarchy.

Sparse interaction

As we have argued, many systems are composed of interacting processes. However, the assumption of weak interaction throughout the entire lifetime of the system is an idealization. In many domains, we may have processes that interact weakly (if at all) most of the time, but may have an occasional strong interaction. In our example of cars on a freeway, the interaction of one car with another is very weak most of the time. However, there are momentary situations when the interaction becomes quite strong, e.g., if one car gets too close to another, or if one wants to change lanes into a position occupied by the other.

The interaction structure of the system is very different in these two situations. So long as the processes interact weakly, an approximation of the belief state as independent components is a very reasonable one. However, when a strong interaction occurs, this approximation incurs a large error. A naive solution would have us correlate the belief states of any two processes that can interact strongly. Unfortunately, in many systems, this solution greatly reduces our ability to select clusters of small sizes and achieve computational efficiency.

An alternative solution is to target our approximation to the context. When two processes have a momentary strong interaction, we should avoid decoupling them in the belief state. In fact, we must take care. The strong correlation between the two processes usually lasts for more than one time slice. But as the system evolves, and the two processes return to their standard weak interaction, the correlation decays. After some amount of time, we will be able to return to an approximation that decouples the states of these two processes.

We now proceed to extend both the inference algorithm and the analysis to account for this type of *sparse interaction*. First, we define the notion of sparse interaction. The idea is to consider two separate transition models, which will be applicable respectively in the standard and exceptional mode of interaction. Concretely, if X and Y interact sparsely, we define a binary random variable B_{XY} which will be a parent of both X' and Y', and will select their mode of interaction. We will speak of the *weak interaction model* and the *strong interaction model* to designate the portions of the conditional probability distributions of X' and Y' that are relevant to either value of B_{XY}.

The extended algorithm uses a different cluster forest $\mathcal{F}^{(t)}$ at each time slice t. If at time t the algorithm detects a strong interaction between the variables in two clusters, it couples them for some length of time. In fact, it couples all the variables in the two siblings which contained either of the two clusters. More precisely, at each time t, we maintain a set $C^{(t)} \subseteq R$ of *couplings* — pairs (i, j) of siblings. At time t, we define a new cluster forest $\mathcal{F}^{(t)}$, derived from the basic cluster forest $\mathcal{F}$. Each cluster $F_k^{(t)}$ in $\mathcal{F}^{(t)}$ is either a cluster of $\mathcal{F}$, or obtained by merging two siblings G_i, G_j. In other terms, each cluster F_l in $\mathcal{F}$ is assigned to some cluster $F_k^{(t)}$, in which case $F_l \subseteq F_k^{(t)}$. We require that if F_l is assigned to $F_k^{(t)}$, $F_l \subseteq G_i$, and $(i, j) \in C^{(t)}$, then $G_i \cup G_j \subseteq F_k^{(t)}$.

Note that, in general, the algorithm might not be able to observe directly the existence of a strong correlation. In some circumstances, there may be certain tests that are reliable indicators of such an event, e.g., a sensor detecting a car signalling a lane change with its blinkers. In other domains, the strong interaction may be due to an action taken by the agent tracking the system; in this case, correlations can be predicted. One general (but expensive) approach for discovering strong interactions is by evaluating the error that would be incurred by taking two clusters to be independent. Here, to simplify the analysis, we assume that strong interactions can be detected. When one occurs, the algorithm couples the two sibling groups involved. If no strong interaction occurs for some number of time slices d, the algorithm decouples them back.

We begin by analyzing the error of this algorithm in the case the groups are disjoint. Let G_i and G_j be two siblings decoupled at time $t + 1$. What do these groups contribute to the error at time $t + 1$? There are two cases. If G_i and G_j were decoupled at time t, then we assume that their most recent interaction was weak. In this case, the analysis reduces to that of Theorem 5. Otherwise, the groups were coupled at t, and we have chosen this time slice to decouple them. In this case, we have assumed that, for some number of time slices d, these groups have been coupled. For that period of time, no error has been incurred by this pair. We therefore need to estimate the extent to which the strong correlation that occurred d time slices ago has attenuated. Let $\gamma_{ij}^w = \gamma[G_i^{(t)} \to G_j^{(t+1)}]$ be the mixing rate using the conditional probabilities of the weak interaction model.

Theorem 8 *Let G_i and G_j be two disjoint reciprocal groups of clusters in $\mathcal{G}$, and assume that no strong interaction has occured between them since time slice $t - d$.*

1. *If G_i and G_j were decoupled at time t, then*

$$\boldsymbol{I}[G_i^{(t+1)}; G_j^{(t+1)}] \leq 3 \cdot \ln |\text{dom } (G_i \cup G_j)| \cdot (1 - \min[\gamma_{ij}^w, \gamma_{ji}^w]).$$

2. *If G_i and G_j were coupled at time t and have just been decoupled, then*

$$\boldsymbol{I}[G_i^{(t+1)}; G_j^{(t+1)}] \leq \ln |\text{dom } (G_i \cup G_j)| \cdot \left((1 - \gamma_{ii}^w)^d \cdot (1 - \gamma_{jj}^w)^d + \frac{3 \cdot (1 - \min[\gamma_{ij}^w, \gamma_{ji}^w])}{\gamma_{ii}^w + \gamma_{jj}^w - \gamma_{ii}^w \cdot \gamma_{jj}^w} \right).$$

Proof sketch The first case follows from Theorem 5. The general case is obtained by applying Lemma 4 d times, giving

$$\begin{aligned} \boldsymbol{I}[G_i^{(t+1)}; G_j^{(t+1)}] \leq& \\ & (1 - \gamma_{ii}^w)^d \cdot (1 - \gamma_{jj}^w)^d \cdot \boldsymbol{I}[G_i^{(t-d)}; G_j^{(t-d)}] \\ & + 3 \cdot \ln |\text{dom } (G_i \cup G_j)| \cdot (1 - \min[\gamma_{ij}^w, \gamma_{ji}^w]) \cdot \\ & \textstyle\sum_{k=0}^{d-1} (1 - \gamma_{ii}^w)^k \cdot (1 - \gamma_{jj}^w)^k \end{aligned}$$

Note that all the γ's are for the weak interaction model, since no strong interaction has occured since epoch $t - d$. Then, the claim follows from the fact that $\boldsymbol{I}[G_i^{(t-d)}; G_j^{(t-d)}] \leq \ln |\text{dom } (G_i \cup G_j)|$ and $\sum_{k=0}^{d-1} x^d \leq 1/(1 - x)$. ∎

Thus, the error induced by decoupling two groups d time slices after a strong correlation decreases exponentially with d. The analysis also tells us how long we need to couple two groups in order to guarantee a bound on the error.

To see how this theorem can be used, let us go back to our highway example, and assume that we observe a strong interaction between two vehicles (such as one cutting in front of the other). Our algorithm would then "couple" the reciprocal groups G_i and G_j to which these vehicles belong, which results in the merging of G_i and G_j (and all their subgroups) for a certain number of time slices, until the correlation induced by the strong correlation has sufficiently decayed that the groups can be decoupled without incurring too much error. Our theorem guarantees that this eventually happens, and gives us an upper bound on the time it takes.

We finally put all of our results together, and state the theorem for models that involve both sparse interactions and overlapping clusters.

Theorem 9 *Let G_i and G_j be two reciprocal groups of clusters in $\mathcal{G}$, and M_{ij} their intersection. Assume that no strong interaction has occured between them since time slice $t-d$. Then*

1. *If G_i and G_j were decoupled at time t, then*

$$\boldsymbol{I}[G_i^{(t+1)}; G_j^{(t+1)} \mid M_{ij}^{(t+1)}] \leq 3 \cdot \ln|\mathrm{dom}\,(G_i \cup G_j)| \cdot (1 - \min[\gamma_{ij}^w, \gamma_{ji}^w]) + \boldsymbol{H}[M_{ij}^{(t)} \mid M_{ij}^{(t+1)}] + \boldsymbol{H}[M_{ij}^{(t+1)} \mid M_{ij}^{(t)}].$$

2. *If G_i and G_j were coupled at time t and have just been decoupled, then* $\boldsymbol{I}[G_i^{(t+1)}; G_j^{(t+1)} \mid M_{ij}^{(t+1)}] \leq$

$$\ln|\mathrm{dom}\,(G_{i\setminus j} \cup G_{j\setminus i})| \cdot (1-\gamma_{ii}^w)^d \cdot (1-\gamma_{jj}^w)^d + \sum_{k=0}^{d-1} (1-\gamma_{ii}^w)^k \cdot (1-\gamma_{jj}^w)^k \cdot \begin{pmatrix} 3 \cdot (1 - \min[\gamma_{ij}^w, \gamma_{ji}^w]) \cdot \ln|\mathrm{dom}\,(G_{i\setminus j} \cup G_{j\setminus i})| \\ + \boldsymbol{H}[M_{ij}^{(t-k)} \mid M_{ij}^{(t-k+1)}] + \boldsymbol{H}[M_{ij}^{(t-k+1)} \mid M_{ij}^{(t-k)}] \end{pmatrix}$$

Discussion and conclusions

Our results show how various system properties, such as weak interaction, conditional weak interaction, and sparse interaction, can be exploited by our inference algorithm. We argue that these properties appear in many real-world systems. Complex systems are almost always hierarchically structured out of subsystems. For example, a freeway system is made up of individual roads, which are composed of many road segments; each segment has several vehicles on it. A computer network has multiple subnets, each with multiple devices, each in turn has several users, running many processes. From a different perspective, one could argue that, regardless of whether complex systems are actually hierarchical, people can deal with them only by decomposing their description into more manageable chunks.

Hierarchical dynamical systems, such as those investigated in (Friedman, Koller, & Pfeffer 1998), are ideally suited for the kind of decomposition provided by the algorithm. Let us consider the interaction structure of such a system. Most of the interaction in the system occurs within subsystems. The lower-level the system, the more tightly it is coupled. Besides this internal interaction, a subsystem usually interacts primarily with its enclosing system. Our results apply exactly to situations such as this. As our results demonstrate, if the interaction between subsystems in a level is weak (or sparse), the correlation it induces can (mostly) be ignored. The correlation induced by the enclosing system is often stronger. However, as Simon states, "the higher-frequency dynamics are associated with the subsystems, the lower-frequency dynamics with the larger systems." This is precisely the case to which our results for conditionally independent clusters apply. Thus, we can model the variables in the subsystems as conditionally independent given the state of the enclosing system. This decomposition can be extended to lower levels of a hierarchy, resulting in a hierarchical decomposition of the belief state analogous to that of the system.

Finally, in many settings, there may be an occasional strong interaction that crosses tradional boundaries. Our extended algorithm and analysis for sparse interactions are precisely designed to handle such situations.

Our results can also help guide the construction of models that will support effective approximation. For example, the decomposition described above relies on the existence of a slowly-evolving enclosing system that renders its subsystems almost independent. We may therefore wish to introduce such a component deliberately into our model, enabling such a decomposition. For example, we can introduce a variable that corresponds to aggregate properties of the system as a whole, e.g., the amount of overall traffic on the road, or stock market indicators. Such aggregate variables typically evolve very slowly, making them suitable to the type of analysis described above.

In summary, the results we have presented allow us to exploit the architecture of a dynamic system for efficient and accurate approximate inference. They also allow us to design the architecture of the system so as to support such an approximation. We therefore hope that they will help us to reason effectively about the very large complex systems that we encounter in the real world.

Acknowledgements This research was supported by ARO under the MURI program "Integrated Approach to Intelligent Systems", grant number DAAH04-96-1-0341, by DARPA contract DACA76-93-C-0025 under subcontract to Information Extraction and Transport, Inc., and through the generosity of the Powell Foundation and the Sloan Foundation.

References

Aström, K. 1965. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Applic.* 10.

Boyen, X., and Koller, D. 1998. Tractable inference for complex stochastic processes. In *Proc. UAI*.

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

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

Forbes, J.; Huang, T.; Kanazawa, K.; and Russell, S. 1995. The BATmobile: Towards a Bayesian automated taxi. In *Proc. IJCAI*.

Friedman, N.; Koller, D.; and Pfeffer, A. 1998. Structured representation of complex stochastic systems. In *Proc. AAAI*.

Jensen, F.; Lauritzen, S.; and Olesen, K. 1990. Bayesian updating in recursive graphical models by local computations. *Computational Statistical Quarterly* 4.

Kjærulff, U. 1992. A computational scheme for reasoning in dynamic probabilistic networks. In *Proc. UAI*.

Simon, H. 1962. The architecture of complexity. *Proc. Am. Phil. Soc.* 106:467–482.