Cyclic Causal Models: Markov Chain Equilibrium Semantics and Factored Inference

David Poole and Mark Crowley

Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada

Abstract

We analyze the foundations of cyclic causal models, and compare structural equation models (SEMs) to an alternative semantics as the equilibrium (stationary) distribution of a Markov chain. We show that in general, cyclic SEMs cannot have independent noise; even in the simplest case, cyclic structured equation models imply constraints on the noise. We give a formalization of the alternative Markov chain equilibrium semantics which requires not only the causal graph, but also a sample order. We present a new algorithm for inference in the equilibrium distribution that exploits the structure of the causal network. We show when it is exact and empirically evaluate a case where it is an approximation.

Keywords: causality, inference

1. Introduction

In this paper we describe a novel approach to modelling probabilistic causal distributions, specifically in the case where cycles are present. We motivate and describe a method for modelling cyclic causal distributions as the equilibrium of a Markov chain.

In (Crowley and Poole, 2011) we investigated spatial planning problems where correlated actions need to be decided for many different locations. This type of problem is very common in environmental and natural resource planning where there may be a higher value assigned to policies which don't take similar actions in neighbouring locations for example.

We wanted to define a stochastic spatial policy which was locally interprettable yet represented a globally consistent distribution over the exponentially large joint actions space.

It seemed that a natural approach to modelling this kind of policy was as a cyclic causal model where each variable is a decision for a location which depends on the decisions at other neighbouring locations. The distribution represented by the cyclic causal model is in fact the equilbirium distribution of a related Markov chain. In (Crowley and Poole, 2011) we use an MCMC simulation approach to estimate the policy over joint actions of around 2000 decision locations.

In this paper we provide/review a theoretical grounding for the equilibrium as the correct interpretation of cyclic causal models, investigate the properties of these models

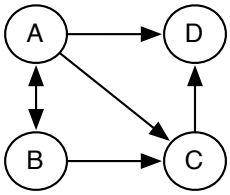

Figure 1: An example of a cyclic causal network for a simple gold market.

and present an iterative algorithm for computing the marginal distributions from the equilibrium of arbitrary subsets of variables. This algorithm can sometimes compute the exact marginals but at the expense of a more complex formulation that increases treewidth of the graph.

For the discussion here we use as a running example a simple economic market composed of four agents: Alfred, Betty, Cindy and Doug. Every day each agent they decided whether they will buy or sell a fixed amount of gold. We ignore complexities such as price and external information or news about gold; we simply want to model the behaviour of a number of agents given fixed policies for their behaviour and how they ineract with each other which itself can be quite complex. The only information agents will use to make their decisions is their own internal desire to own gold conditioned on their reaction to how other agents behave. To avoid runaway purchasing we can assume that each agent has a fixed number of units of gold they can hold at any time.

When deciding whether to buy a unit of gold today or not, each agent can take into account the most recent action of any other agents. Alfred and Betty trust each other and pay close attention to each other's activities and have a high probability of following each other's lead. Meanwhile, Doug and Cindy's activities are ignored by everyone but they both pay close attention to the activities of Betty and Alfred.

The entire set of causal relationships are shown in Figure 1 as a causal network containing cycles. Causal relationships are the most natural representation here and cycles indicate that an agent is reacting to the latest information they have about others. The goal is to compute a distribution over the amount of gold being bought or sold over time by each individual agent or in total.

An important question now is can such a local definition be used to define a consistent distribution over joint actions for all agents? The afirmative answer was explained by Strotz and Wold (1960), the correct distribution is the fixed point of the causal interactions over time.

A more difficult question to answer is how to compute this distribution over joint actions in an efficient manner and in what cases it can be computed exactly. When we began investigating this we found that existing representations of causal distributions do not adequately take advantage of the structure in cyclic causal models. Addressing these questions is the focus of this paper.

It is not surprising that a model which applies well to an economic market could be best represent an equilibrium system, equilibria are a core topic in economics **?**. How-

ever, popular causal modelling methods cannot easily deal with some of hte difficulties which can arise in cyclic causal models like this. Also, there are strong connections between how these equilibria can be computed and advances in filtering of dynamic Bayesian networks that to the best of our knowledge have not been addressed in the literature. In this paper we attempt to rectify these two situations by providing a new representation of causal models that can deal naturally with cycles and demonstrating a deterministic, iterative algorithm for computing exact or approximate equilibrium distributions from cyclic causal models.

2. Modelling Cyclic Causality with SEMs

We begin by characterizing the distribution we want as a *cyclic causal model(CCM)*. CCDs can be represented using standard causal modelling languages but they are not commonly studied in the literature. Cyclic causal models define distributions that are equilibria of rolled out Markov chains. We show that some common assumptions about causal models do not hold in the cyclic case. We demonstrate a method for computing the marginal distributions for queries on cyclic causal models.

Pearl (2009) proposed structural equation models (SEMs) as a representation for causality. A structural equation model consists of a deterministic function for each variable in terms of other variables and (independent) noise inputs. A modal logic for SEMs was presented by Halpern (2000).

An alternative to SEMs is an equilibrium model (Strotz and Wold, 1960), where the causes of each variable form a transition model of a Markov chain, and we are interested in the equilibrium distribution of this Markov chain. Strotz (1960) describes the contribution of that paper as:

> *If a causal interpretation of an interdependent system is possible it is to be provided in terms of a recursive system. The interdependent system is then either an approximation to the recursive system or a description of its equilibrium state.*

The SEM and the equilibrium structure of Iwasaki and Simon (1994) can be seen as an equilibrium where the *values* of the variables are invariant. In the Markov chain semantics the equilibrium is on the distribution of the variables.

One of the properties of the causal theories of Pearl is that local causal models are sufficient to predict all combinations of interventions (including the case of no interventions). For each variable X, with parents π_X, and for each combination of values, $\bar{v}$, to π_X, the probabilities $P(X|do(\pi_X = \bar{v}))$ fully specify the model.

3. A Simple Cyclic Example

We write variables in upper case and assignments in lower case. For a Boolean variable, we write the $true$ value as the lower-case of the variable, for example, $A = true$ is written as a and $A = false$ is thus $\neg a$.

Example 1. Consider the simple cyclic causal model with two Boolean variables A and B, each dependent on the other. The causal model can be defined in terms of 4 parameters:

$$p_1 = P(a|do(b))$$
$$p_2 = P(a|do(\neg b))$$
$$p_3 = P(b|do(a))$$
$$p_4 = P(b|do(\neg a))$$

This can be represented as a structural equation model:

$$a \leftrightarrow (b \wedge u_1) \vee (\neg b \wedge u_2) \tag{1}$$
$$b \leftrightarrow (a \wedge u_3) \vee (\neg a \wedge u_4) \tag{2}$$

where the U_i are independent exogenous Boolean variables and $P(u_i) = p_i$. An exogenous variable is **extreme** if its probability distribution contains zeros, and is **non-extreme** if its probabilities are all strictly between 0 and 1.

This model gives the anticipated result for all interventions, except for the case of no interventions:

Proposition 1. *The noise variables $u_1, \ldots, u_4$ in the SEM cannot be non-extreme and independent.*

Proof. The assignment $u_1 = true$, $u_2 = false$, $u_3 = false$, $u_4 = true$ is logically inconsistent, as it implies $(a \leftrightarrow b) \wedge (b \leftrightarrow \neg a)$, and so must have probability 0. For the u_i to be independent $P(u_1 \wedge \neg u_2 \wedge \neg u_3 \wedge u_4) = P(u_1) \times (1 - P(u_2)) \times (1 - P(u_3)) \times P(u_4) = 0$. The probability of one of them must be zero. Similarly, $u_1 = false$, $u_2 = true$, $u_3 = true$, $u_4 = false$ must have probability 0, the and so probability of another proposition must be zero. □

This does not rely on there being two binary variables, but has to do with the cyclic causality. The following result shows that it happens quite generally.

Proposition 2. *If there is a variable X with two different values v_1 and v_2, and there is a set of assignments $u_1 \ldots u_k$ to exogenous variables $U_1 \ldots U_k$ such that $P(u_i) > 0$ for each i and $u_1 \ldots u_k \models (X{=}v_1) \rightarrow (X{=}v_2)$ and $u_1 \ldots u_k \models (X{=}v_2) \rightarrow (X{=}v_1)$, then the variables $U_1 \ldots U_k$ cannot be probabilistically independent.*

Proof. $P(u_1 \ldots u_k) = 0$ as $u_1 \ldots u_k$ are logically inconsistent. If they were independent, $P(u_1 \ldots u_k) = \prod_i P(u_i) > 0$. □

Note that this just requires a weak but sound reasoning procedure (that implements $\models$). If there is a stronger logic behind it, such as $\mathscr{L}^+(S)$ of Halpern (2000), the results still hold, but the logic may be able to prove the inconsistency directly. Note that Halpern's logic does not include exogenous variables and so is not directly applicable.

Example 2. One way to avoid inconsistency is to make the noise variables dependent, for example to make $u_2 \rightarrow u_1$, which makes $u_2 \wedge \neg u_1$ inconsistent. This can be modelled by making $u_2 = u_1 \wedge u_5$ for some noise u_5. Equation (1) becomes:

$$a \leftrightarrow (b \wedge u_1) \vee (\neg b \wedge u_1 \wedge u_5).$$

This can be reduced to:

$$a \leftrightarrow (b \vee u_5) \wedge u_1.$$

This is the style of many of the SEMs of Pearl (2009), for example on page 29. This (with the corresponding equation for B) incorporates prior knowledge that A and B are positively correlated, as making one true can only increase the probability of the other being true. This is not appropriate if it is possible that A and B are negatively correlated. It also does not result in a unique probability for A or for B.

3.1. Equilibrium Models

An alternative semantics for the distribution is in terms of the equilibrium distribution (also called the stationary distribution or the fixed-point distribution) of a Markov chain (Bremaud, 1999). For Example 1, this semantics is defined in terms of a Markov chain with variables $A^0, A^1, \ldots$ and $B^0, B^1, \ldots$, where the superscript represents a time point, with transition probabilities such as:

$$\begin{aligned} p_1 &= P(a^t | b^t) \\ p_2 &= P(a^t | \neg b^t) \\ p_3 &= P(b^t | a^{t-1}) \\ p_4 &= P(b^t | \neg a^{t-1}) \end{aligned}$$

where a^t is the proposition that A is true at time t. This can be specified like an SEM, but variables on the right hand sides can refer to a previous time (in such a way that there are no cycles in the temporally extended graph). E.g.:

$$a^t \leftrightarrow (b^t \wedge u_1^t) \vee (\neg b^t \wedge u_2^t) \tag{3}$$

$$b^t \leftrightarrow (a^{t-1} \wedge u_3^t) \vee (\neg a^{t-1} \wedge u_4^t) \tag{4}$$

where for all t, U_i^t are independently identically distributed variables with probability p_i. The use of the previous time is used to avoid cycles in the temporally extended models. The aim is to determine the equilibrium distribution — the distribution over the variables that does not change in time.

This Markov chain has an equilibrium that satisfies:

$$P(a) = p_1 P(b) + p_2(1 - P(b)) \tag{5}$$

$$P(b) = p_3 P(a) + p_4(1 - P(a)) \tag{6}$$

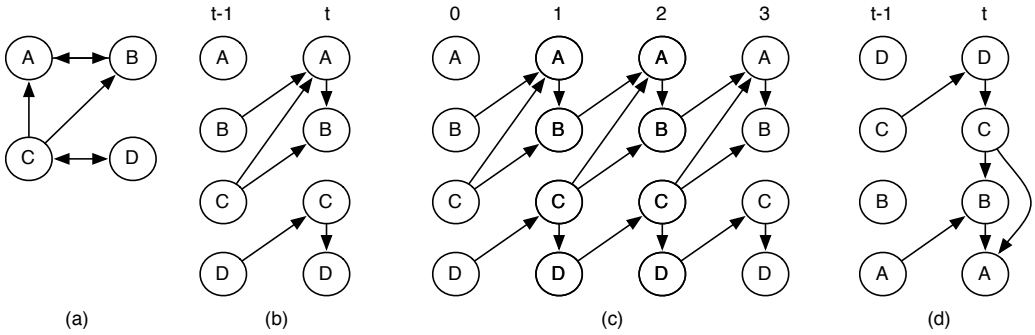

Figure 2: A causal network, its 2-stage DBN for sample ordering $A < B < C < D$, its unrolled DBN, and the 2-stage DBN for sample ordering $D < B < C < D$

Solving the simultaneous equations gives:

$$P(a) = \frac{p_1 p_4 + p_2(1-p_4)}{1-(p_1-p_2)(p_3-p_4)} \tag{7}$$

$$P(b) = \frac{p_3 p_2 + p_4(1-p_2)}{1-(p_1-p_2)(p_3-p_4)} \tag{8}$$

which are well defined for all $p_i \in [0,1]$, except for the two cases: $p_1 = 1,\ p_2 = 0,\ p_3 = 1,\ p_4 = 0$ (which corresponds to $a \leftrightarrow b$) and $p_1 = 0,\ p_2 = 1,\ p_3 = 0,\ p_4 = 1$ (which corresponds to $a \leftrightarrow \neg b$). In these cases, there is an equilibrium for every value in $[0,1]$. For the rest of this discussion, we ignore extreme probabilities that give these two extreme cases.

In the Markov chain, the A's at different times are different variables. There is no logical inconsistency that leads to the problem in the proof of Proposition 1.

To specify Equations (3) and (4), we need not only specify that A and B are dependent, but also that B depends on the previous value of A, and A depends on the current value of B. Intuitively, for each time, we sample B then A.

4. Markov Chain Equilibrium Models

In this section we define Markov chain equilibrium models as an alternative to SEMs for representing causal knowledge. These models are slightly more complex as the equilibrium distribution depends on the order in which the variables are sampled as well as when the distribution is sampled.

For this paper, we assume finitely many discrete-valued variables, and that all conditional probabilities are non-extreme. The non-extreme assumption is reasonable for learned models, where we may not want to a priori assume that any transition is impossible, but may not be appropriate for all domains. It simplifies the discussion as all of the Markov chains are then ergodic, with a unique equilibrium distribution, independent of the starting state (Bremaud, 1999).

If X is a variable, the **parents** of X are defined to be a minimal set of variables $\mathbf{Y}$ such that for all sets of variables $\mathbf{Z}$ such that $\{X\}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint sets

$$P(X|do(\mathbf{Y})) = P(X|do(\mathbf{Y},\mathbf{Z})).$$

That is, for all interventions where the variables in $\mathbf{Y}$ are set to particular values, changing the value of any other variables $\mathbf{Z}$ does not affect X. This is like the standard definition of conditional independence, but involves interventions, not observations. It is easy to show that the set of parents of X is unique.

This parent relation induces a directed graph that can contain cycles, but is irreflexive (there is no arc from a variable to itself). This is not a restriction since a reflexive graph can always be turned into an irreflexive one by absorbing the dependence of the variable on itself into its conditional probabilities.

Define a **causal network** to be an irreflexive directed graph where the nodes are random variables, together with a causal mechanism for each variable X that consists of a conditional probability $P(X|do(\pi_X))$ where π_X is the set of parents of X in the causal network.

To represent an intervention on a variable X, the causal mechanism for X is replaced by $P(X{=}v) = 1$ when we $do(X) = v$ (Pearl, 2009).

To define the post-intervention semantics, we construct a two stage dynamic Bayesian network (DBN) (Dean and Kanazawa, 1989). A 2-stage DBN specifies for each variable how a variable at the current stage depends on variables at the current stage and variables at the previous stage.

This DBN depends on both the causal network and a **sample ordering** which is a total ordering of the variables, say $X_1 < X_2 < \cdots < X_n$. For each variable X, define π_X^- to be the set of those parents of X that are less than X in the sample ordering, and π_X^+ to be the set of those parents of X that are greater than X in the sample ordering. Thus $\pi_X = \pi_X^- \cup \pi_X^+$.

Intuitively, each X_i, depends on its parents in π_X^- at the current stage, and on its parents in π_X^+ at the previous stage.

A causal network with variables $\{X_1,\ldots,X_n\}$ and sample ordering $X_1 < X_2 < \cdots < X_n$ defines a decomposition of a discrete-time Markov chain where the state S^t at time t can be described by the variables $X_1^t,\ldots,X_n^t$ for each time t, and for each causal variable X for each time t, the Markov chain variable X^t has parents $\{Y^t : Y \in \pi_X^-\} \cup \{Y^{t-1} : Y \in \pi_X^+\}$. X^t is independent of all variables $Z^{t'}$ for $t' < t$ given these parents and is independent of all variables Z^t where $Z < X$ given these parents in the Markov chain. Thus the causal network with the sample ordering defines the decomposition of the state transition function:

$$\begin{aligned} P(S^t|S^{t-1}) &= P(X_1^t,\ldots,X_n^t|S^{t-1}) \\ &= \prod_{i=1}^{n} P(X_i|X_1\ldots X_{i-1}S^{t-1}) \\ &= \prod_{i=1}^{n} P(X_i|(\pi_{X_i}^-)^t(\pi_{X_i}^+)^{t-1}) \end{aligned} \tag{9}$$

The conditional probabilities for the Markov chain, the $P(X_i|(\pi_{X_i}^-)^t(\pi_{X_i}^+)^{t-1})$, are the $P(X|\pi_X)$ in the causal network.

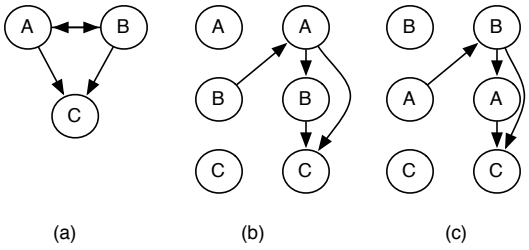

Figure 3: A causal network and two induced 2-stage Bayesian networks

Example 3. Consider the causal network in Figure 2 (a), where the double-ended line segment represents two arcs. In this example, the parents of A are B and C, the parents of B are A and C, the parent of C is D, and the parent of D is C.

Figure 2 (b) shows the 2-stage DBN with sample ordering $A < B < C < D$. The left nodes represent the variables at time $t-1$ and the right nodes represent the variables at time t. This represents the Markov chain given in Figure 2 (c), where the structure is repeated indefinitely to the right. Each of the conditional probabilities is defined as part of the causal network.

Figure 2 (d) shows the 2-stage DBN for the same causal network with sample ordering $D < C < B < A$.

We define the distribution of the causal model (after interventions) to be the equilibrium (stationary) distribution of the induced Markov chain.

4.1. Dependence on Sample Ordering 1

The following example shows that the equilibrium distribution can depend on the sample ordering:

Example 4. Consider the causal network of Figure 3 (a), with the causal probabilities:

$$P(a|do(b)) = 0.1 \;\; P(a|do(\neg b)) = 0.9$$
$$P(b|do(a)) = 0.9 \;\; P(b|do(\neg a)) = 0.1$$
$$P(c|do(a \wedge b)) = P(c|do(\neg a \wedge \neg b)) = 0.9$$
$$P(c|do(\neg a \wedge b)) = P(c|do(a \wedge \neg b)) = 0.1$$

One way to see this is that doing B tends to change A to be different to B, and doing A tends to change B to be the same as A. C has high probability if A and B have the same value.

Figure 3 (b) shows the 2-stage DBN with the sample ordering $A < B < C$. Figure 3 (c) shows the 2-stage DBN with the sample ordering $B < A < C$.

In the equilibrium distribution of Figure 3 (b), $P(c) = 0.82$, whereas in the equilibrium distribution of (c), $P(c) = 0.18$. Intuitively, in (b), A is sampled, then B is sampled, based on that value of A, and so they tend to have the same value and so C tends to be true. Whereas in (c), B is sampled, then A is sampled, based on that value of B, and so they tend to have different values and so C tends to be false.

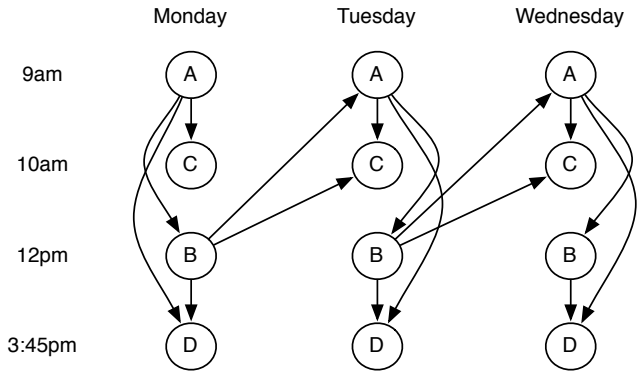

Figure 4: Example schedule for gold trades of four people with a cyclic causal interaction based on figure 1.

By changing the probabilities of the causal network, for each sample ordering, the probability distribution of C in the equilibrium can be made to have an arbitrary non-extreme distribution. There are no bounds to the effect that sample ordering can have.

4.2. Dependence on Sample Ordering 2

In a cyclic causal model there are situations where the ordering of variables in the Markov chain makes a difference on the estimated marginal distributions. Consider the example in figure 1. If each of the four people always buy or sell gold at the same time each day this will induce a consistent causal ordering on the information each other person has over time as they make their trading decisions. One possible schedule is shown in figure 4 with the appropriate causal arcs from the model in figure 1 added. Recall that Betty tends to follow Alfred's actions from that morning while Alfred often acts opposite to what Betty did yesterday. Now Cindy, who always trades in the morning after Alfred always sees that Betty and Alfred's trades are differently, so Cindy's purchasing decisions will always be made on that understanding. Meanwhile Doug, who trades at the end of the day always sees Betty's behaviour largely in sync with Alfred's and acts on the understanding that they are in general agreement. Note that Cindy and Doug could have identical dependencies on Alfred and Betty yet their behaviour will be very different.

This effect of different orderings comes from the fact that there are instabilities in the equilibrium distribution. We can understand instabilities as arising from assymetric correlations between two variables. When the local causal models states that A's value is positively correlated with B's value but B's model states that B's value is *negatively correlated* with A's value.

In the logical case this instability produces a contradiction. In the stochastic case it produces a dependency on the causal ordering. The method presented here defines a spectrum of distributions between fully independent SEM models and asymmetric correlated causal models. In between these extremes we can still compute the distribution given a particular causal ordering but there is no one ordering that is more 'right' in

these cases. The maximum deviation that can occur due to different causal orderings is analyzed and bounded in relation to the structure of the local causal distribution.

5. Inference for MC-Equilibrium Causal Models

The inference problem we consider here is, given a causal network and a sample ordering, determine $P(X|do(Y),Z)$ for some sets of variables X, Y and Z, which means the posterior distribution of X after doing Y and then observing Z in the equilibrium distribution[1]. This can be computed by replacing the causal mechanisms of the variables in Y with the intervention values, computing the equilibrium distribution, conditioning on Z and marginalizing over the remaining variables.

One way to compute the equilibrium distribution is to sample from it, sampling each variable in turn according to the sample ordering. This is an instance of Markov Chain Monte Carlo (MCMC) sampling (Bremaud, 1999). In MCMC sampling we sample S^t from S^{t-1}, where S^t is the state at time t. The samples generated (after some burn-in period) can be considered as random samples from the equilibrium distribution, as long as there are sufficiently many.

For computing the equilibrium of a causal model, a state is an assignment of a value to each variable. If variables are selected according to the the sample ordering, the probabilities from the causal model can be directly used. To see this, suppose the sample ordering is $X_1 < X_2 < \cdots < X_n$, then we can use the decomposition of equation (9), and note that, when X_i is selected, $(\pi^-_{X_i})^t(\pi^+_{X_i})^{t-1}$ are the current values of these variables. Thus Gibbs sampling of each variable using the probabilities of the causal network, sampled according to the sample ordering, has exactly the same form as the equilibrium distribution itself.

It is possible to sample using a different ordering than the sample ordering, but this requires Bayes' rule when sampling a variable after one of its children at any time.

The equilibrium distribution can also be computed directly using Gaussian elimination, which take time polynomial in the number of states.

An alternative iterative method is to start with a probability distribution over states and repeatedly use the model to compute a distribution over the next state. This converges to the stationary distribution (unlike MCMC which gives samples that are distributed according to the stationary distribution) with geometric convergence (Bremaud, 1999). This algorithm is polynomial in state space as it entails computing the probability of each state.

AI research over the last few decades has demonstrated that we can typically do much better than polynomial in the state space by exploiting the sparseness of the representation. Exploiting conditional independence is the basis for efficient inference in Bayesian networks (Pearl, 1988a), using techniques such as variable elimination (Zhang and Poole, 1994).

[1]Note that this is *not* counterfactual reasoning (Pearl, 2009), which would be observing then doing. In general, there could be arbitrary sequences of observing and doing. It is what Dash (2005) calls the manipulated-equilibriated model, but our equilibrium is over distributions.

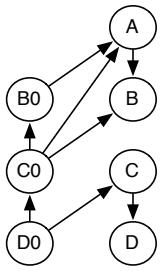

Figure 5: An equilibrium belief network

In the rest of this section we show a class of algorithms that construct a Bayesian network that can answer (approximately) conditional queries about the stationary distribution. We show when it gives exact answers, and empirically evaluate when it is not exact.

To understand the algorithm, notice that a two stage DBN is like a Bayesian network but the distribution over the "previous" variables is not specified. Suppose we were to create a belief network over these variables (i.e., we define a probability for each previous variable given only other previous variables as parents, such that this previous network is acyclic). If this previous network represented the equilibrium distribution, the current variables will also represent the equilibrium distribution. If the distribution over the previous variables was to represent an approximation to the equilibrium distribution, the current variables will represent a (better) approximation for the equilibrium distribution.

Example 5. Consider Example 3 with sample ordering $A < B < C < D$, and with the corresponding 2-stage DBN is shown in Figure 2 (b). Figure 5 shows a Bayesian network representation of the equilibrium distribution. Note that, unlike the 2-stage DBN, this is complete in itself, and is not a template for a temporally extended network.

Define a **previous ordering** to be a total ordering of the variables. This ordering will be used to order the variables at the previous stage and can be unrelated to the sample ordering.

An **equilibrium belief network** for causal network G, sample ordering O and previous ordering, is a Bayesian network, where for each variable X in G, there is a *current* node, X^1, and there is a *previous* node X^0 if X has a child earlier in the sample ordering. The parents of current nodes are derived from the causal network and the sample ordering: X^1 has parents $\{Y^1 : Y \in \pi_X^-\} \cup \{Y^0 : Y \in \pi_X^+\}$, and the conditional probability is specified as part of G. The parents of previous nodes can only be previous nodes that are earlier in the previous ordering. Thus an equilibrium belief network is like a 2-stage DBN but with a structure on the previous variables.

To answer queries about the equilibrium distribution, we:

- determine the structure of the previous stage
- iteratively compute the conditional probabilities for the previous stage

- condition on and query the current variables.

Given an equilibrium belief network, the **iterative improvement algorithm** repeatedly updates the conditional probability tables for the previous variables:

- For each previous variable X^0 in the EBN with parents $\mathbf{Y}^0$, compute $P(X^1|\mathbf{Y}^1)$ where X^1 and $\mathbf{Y}^1$ are the current counterparts of X^0 and $\mathbf{Y}^0$, and replace the value of $P(X^0|\mathbf{Y}^0)$ by the computed value.

If the subgraph on the previous variables is complete, the iterative improvement algorithm will converge geometrically to the equilibrium distribution, as it is equivalent to iteratively applying the transition matrix to a representation of the state, which converges geometrically (Bremaud, 1999). If the subgraph is not complete, this converges to an approximation to the equilibrium distribution.

When we have constructed the equilibrium belief network, and iteratively computed the conditional probability for each previous variable, we can condition on the variables at the current time, and query any current variables, as in standard probabilistic inference.

This algorithm is sometimes exact (converges to the actual equilibrium distribution). It is exact if the belief network of the previous variables can represent the equilibrium distribution. In particular, it is exact if the previous variables are fully connected. It is also exact for other cases described in the next section.

One way to view this algorithm is in terms of the monitoring results of Boyen and Koller (1998), but without observations. Essentially we are projecting the belief state onto a carefully constructed belief network representation of the belief state. However, we are representing the belief state using hidden variables (the previous variables are latent variables for the representation of the equilibrium).

6. Analysis

In this section, we characterize when the algorithm is exact.

Example 6. Figure 6 (a) shows a 5 node causal network. (b) is the induced 2-stage DBN under the sample ordering $A < B < C < D$, which expands to the temporally extended network shown in (c). (d) shows an equilibrium belief network where the previous nodes are disconnected. The iterative improvement algorithm converges to the correct marginals for all of the variables. However, it is not exact for marginals on non-singleton sets of variables (except for DE). (e) shows an equilibrium belief network under the previous ordering E, D, C, B, A where neighboring nodes are joined. For this EBN, the algorithm converges to the exact marginal for each singleton variable, and for AB, BC, CD and DE. However, the resulting network does not fully represent the equilibrium distribution as it gets the wrong answer, for example, for $P(A|D)$. To understand this, consider the unrolled DBN in (c). A and D are dependent, but this dependence cannot be fully represented by the equilibrium belief network (e).

A set $\mathbf{S}$ of nodes forms a **previous clique** in an equilibrium belief network if, for each node $N \in \mathbf{S}$, $\{M^0 : M \in \mathbf{S} \text{ and } N^0 < M^0 \text{ in the previous ordering}\} \subseteq \pi_{N^0}$. A set

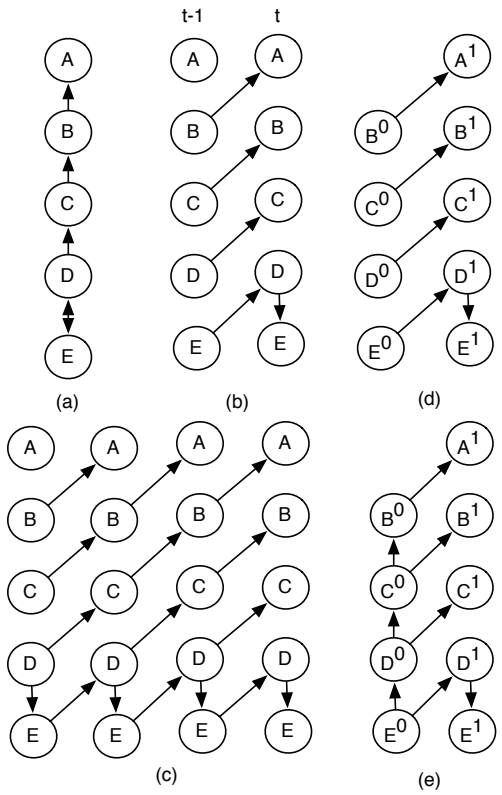

Figure 6: (a) A causal network, (b) its induced 2-stage DBN (c) the unfolded DBN and (d) and (e) two equilibrium belief networks

of nodes **S** is **represented** in an EBN if the set of previous nodes that are ancestors of elements of **S**, i.e., $\{M : \exists A \in \mathbf{S}$ such that M^0 is an ancestor of $A^1\}$, forms a previous clique.

In other words, a set of nodes S is represented if for every possible pairing of nodes N and M in the combined ancestors of all nodes in S, exactly one of the following is true:

1. M is a parent of N in the latent model: $M^0 \subseteq \pi_N^0$
2. M comes earlier than N in the causal ordering : $M < N$

This means that if we want to make queries on marginals of a set of nodes then all ancestors of that set of nodes needs to be available as previous nodes in the causal ordering or need to be interconnected in the latent model. Only then can all the cyclic correlations between nodes in **S** be modelled and exact marginals computed.

Consider the calculation of the marginal probability of a query set $\mathbf{S} \subseteq \mathbf{X}$ using an EBN $G^{\theta\gamma}$ where **S** is the set of all variable assignments, or *queries*, that will be modelled exactly by the EBN. Each $\mathbf{s} \in \mathbf{S}$ assigns a value to a subset of the modelled

variables in the EBN (eg. $\mathbf{s} = \{a^1, \neg b^1\}$).

The marginal distribution for the query set is

$$G^{\theta,\gamma}(\mathbf{S}) = \sum_{\mathbf{Y}}^{1} \sum_{\mathbf{Y}^0} G^{\gamma}(\mathbf{S}|\mathbf{Y}^1, \mathbf{Y}^0) G^{\gamma}(\mathbf{Y}^1|\mathbf{Y}^0) G^{\theta}(\mathbf{Y}^0) \tag{10}$$

for parameters θ and γ; γ are the fixed parameters defining the transition model of the Markov chain for all conditional probabilities between nodes *within* the current timestep as well as *between* the current and the previous steps; θ are the parameters for conditional probabilities *within the previous step*. $\mathbf{Y}^1$ and $\mathbf{Y}^0$ are the sets of ancestors within the current or previous step repspectively of any node in $\mathbf{S}$.

6.1. Expectation Propogation

Belief Propagation (BP) Pearl (1988b) was designed for inference in trees and computes the exact marginals of all variables in that case. It was discovered later that a fast process for computing error correcting codes, called Turbo codes, was equivalent to BP in cyclic graphs. This created interest in exploring the properties of loopy or iterated BP. A large literature has sprung up exploring the properties of this algorithm (**?**) and generalizing it (**????**).

? introduced the *Expectation Propogation (EP)* algorithm which generalizes the Boyen-Koller algorithm for filtering in DBNs which itself can be seen as a special case of BP. EP always has one or more fixed points when the approximation is in the exponential family. However, EP does not necessarily converge to that fixed point if started far away from the fixed point. It has been found in practice (**?**) that for BP a lack of convergence indicates that the approximation does not match the form of the true posterior well.

Conditional probability tables represent discrete multinomials which are in the exponential family.

In our case the approximate belief model is very tightly related to the true distribution through the causal model.

The algorithm iteratively updates the estimates on a product of estimator terms.

The iterative improvement algorithm on an EBN is an instance of the EP algorithm.

In BP the approximate belief state is modelled by a set of independent approximate marginals:

$$q(X_k) = \prod_k q_k(X_k) \tag{11}$$

In our model the approximate belief state is a Bayesian network over a subset of the nodes:

At each step t we recompute the approximate distribution by setting it to the best estimate of the current model (this automatically minimizes the KL-divergence $KL(p(x)||q(x))$

by setting $q_t(x) := \hat{p}_{t-1}(x)$)):

$$q_t(X_t^0) = p(X_t^1) \tag{12}$$

$$= \sum_{Y \in V_{t-1} - X_{t-1}^1} p(X_t^1|Y)p(Y) \tag{13}$$

$$= \sum_{Y_{t-1} \in V_{t-1} - X_{t-1}^1} p(X_t^1|Y) \prod_{Y_{t-1}^1} p(Y_{t-1}^1|pa(Y_{t-1}))p(Y_{t-1}^0) \tag{14}$$

$$= \sum_{Y \in V_{t-1} - X_{t-1}^1} p(X_t^1|Y) \prod_{Y^1} p(Y^1|pa(Y^1))q_{t-1}(Y_{t-1}^0) \tag{15}$$

$$\tag{16}$$

We begin $q_0(X_0^0)$ with random parameters θ and update them using the above formulation iteratively.

? describe a tree based approach to approximating the belief state which is related to (**?**). They suggest that EP will converge faster if more correlations between the ancestor variables are modelled in the approximation. This is exactly the approach that is usedin our discussion of the structure of the previous nodes. This is because the tree minimizes the inclusive KL-divergence at each step by trying to ensure that the approximation matches as closely as possible the true distribution. This is done automatically by computing new values for the parameters of the tree from the marginalized form of the current model. This is precisely what we are doing in our iterative approximation algortihm.

BP Message Ordering vs Causal Ordering. The ordering in BP that the messages are computed does not matter for determining the fixed points, they will be the same regardless. But if you choose to compute them in one particular order you have a higher risk of not converging, averaging them all together in a junction tree minimizes this risk. BP might still not converge for other reasons though.

The causal ordering we refer to is an entirely different concept from this message ordering. The causal ordering we use determines the order that nodes are considered from the cyclic causal model to build the DBN. This creates a properly normalized graphical model (unlike the causal network). But the ordering we choose determines the model itself, different orderings lead to different distributions because a probabilistic model requires a causal model plus an ordering. When the causal model is a tree this ordering is determined, but when it is cyclic it is meaningful and a fundamental requirement for specifying the model.

6.2. Message Passing in the Join Graph

A cyclic causal model implies a dbn which can be turned into a join graph over the cliques. If the join graph has cycles in it then BP will not necessarily converge to a good approximation. The previous clique rule explicitely models potentials over pairs of nodes. Each of these pairs implies a new variable should be added to cliques in the join graph. These nodes create new links in the join graph which which disconnect the loops in the join graph. If all the arcs required by the previous clique rule are added

then the join graph will become singly connected and BP will converge to the exact solution. This is why the iterative updating algorithm can converge to an exact answer for some marginals even if the latent model is not fully connected.

A given marginal will be computed exactly by BP as long as something involving summing out loops.

6.3. Effect of Causal Orderings

Two different causal orderings γ and α will only lead to different equilibrium distributions if all of the following conditions are met:

- there is a causal cycle containing nodes A and B with all other intervening nodes in the cycle contained in $\mathbf{W}$ (i.e. $\mathbf{W} \subseteq anc(A \cup B)$ and $\mathbf{W} \subseteq desc(A \cup B)$)
- there is at least one node $C \in V - (\mathbf{W} \cup A \cup B)$ which has the probability distribution $p(C|A,B,\mathbf{Y})$ where $\mathbf{Y}$ can contain any other nodes including $\mathbf{W}$
- the orderings γ and α have different relative orderings for A and B

6.4. Fixed Point Proof

We now prove that if the EBN is defined so that the query set $\mathbf{S}$ is represented in the previous step then the parameters θ will be sufficient to define the equilibrium distribution; there is one, unique θ^* which defines that equilibrium distribution; and our iterative improvement algorithm is guaranteed to converge geometrically to θ^*.

The following well known theorem from fixed point theory and other lemmas will be useful in showing the convergence of our algorithm.

Proposition 3 (Banach - (Agarwal et al., 2001))**.** *Let (X,d) be a complete metric space of points X and distance function $\mathscr{D}(x,y)$ for all $x,y \in X$. Define a transformation mapping $T : X \to X$ which is a* contraction*; meaning that there exists $k \in [0,1)$ such that*

$$\mathscr{D}(T(x),T(y)) \leq k\ \mathscr{D}(x,y) \tag{17}$$

for all x,y in X. Applying T iteratively n times is denoted as $T^n(x)$.

Then starting from any point $x \in X$ there is a unique fixed point, $x^* \in X$, which can be found by

$$\lim_{n\to\infty} T^n(x) = x^*. \tag{18}$$

The system converges by $k^n/1-k$ over n steps.

Proof. See Agarwal et al. (2001) for a standard proof of this theorem. □

Lemma 1. *For any binary distributions $p(X)$, $q(X)$, $f(X)$:*

$$\sum_{x\in X} f(x)(p(x)-q(x)) = |p(x)-q(x)|(f(x)-f(\neg x)) \tag{19}$$

We define a complete metric space $(\Theta, \mathscr{D})$ where Θ is the set of all latent model parametrizations of the EBN network and $\mathscr{D}$ is a distance measure. We use the L1-norm to define the distance between distributions represented by two EBNs G^{θ} and H^{ψ} with two latent model paramterizations $\theta, \psi \in \Theta$. The expanded form of the distance measure is:

$$\begin{aligned}
\mathscr{D}(\theta,\psi) &= \sum_{\mathbf{s}\in\mathbf{S}} \mathscr{D}(\theta,\psi)(\mathbf{s}) = \sum_{\mathbf{s}\in\mathbf{S}} \left| G^{\theta,\gamma}(\mathbf{s}) - H^{\psi,\gamma}(\mathbf{s}) \right| \\
&= \sum_{\mathbf{s}\in\mathbf{S}} \left| \sum_{\mathbf{Y}^1}\sum_{\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y}^1,\mathbf{Y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{Y^0}) G^{\theta}(\mathbf{Y}^0) - \sum_{\mathbf{Y}^1}\sum_{\mathbf{Y}^0} H^{\gamma}(\mathbf{s}|\mathbf{Y^1},\mathbf{Y}^0) H^{\gamma}(\mathbf{Y^1}|\mathbf{Y^0}) H^{\psi}(\mathbf{Y}^0) \right| \\
&= \sum_{\mathbf{s}\in\mathbf{S}} \left| \sum_{\mathbf{Y}^1}\sum_{\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y}^1,\mathbf{Y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{Y^0}) \left[G^{\theta}(\mathbf{Y}^0) - H^{\psi}(\mathbf{Y}^0) \right] \right|
\end{aligned}$$

For the following we assume there is no variable $Z \in \mathbf{y}^0$ which has all of the following qualities: Z is an ancestor of the query node s in the previous step, Z comes after s in the current step *and* Z and s share a common ancestor in the current step. This creates an effect like conditioning in the computation to update the parameters which breaks the current proof, but there may be some other way to prove it.

`M:Question: what happens with Z variables when computing the next step? A problem arises when some node in the previous clique is both an ancestor and descendent of query node s in the current state. Can this occur? Is itervention or observation used when computing next parameters? that might avoid the problem.`

One step of the iterative algorithm can be defined as a transition mapping, $T : \Theta \times \Gamma \to \Theta$, from all the EBN parameters to an updated set of latent model parameters. After applying the transition mapping one time the marginal distribution is updated as follows: `M:question: should` $\mathbf{S}$ `be` S `instead? proof is just for querying a single node at a time.`

$$\begin{aligned}
G^{T(\theta,\gamma),\gamma}(\mathbf{s}) &= \sum_{\mathbf{Y}^1}\sum_{\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y}^1,\mathbf{Y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{Y^0}) G^{T(\theta,\gamma)}(\mathbf{Y}^0) \\
&= \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0\in\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y^1},\mathbf{y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{y^0}) \\
&\quad \sum_{\hat{\mathbf{S}}^1}\sum_{\hat{\mathbf{X}}^1} G^{\gamma}(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1,\hat{\mathbf{X}}^1) G^{\gamma}(\hat{\mathbf{X}}^1) \sum_{\hat{\mathbf{Y}}^0}\sum_{\hat{\mathbf{Y}}^1} G^{\gamma}(\hat{\mathbf{S}}^1|\hat{\mathbf{Y}}^1,\hat{\mathbf{Y}}^0) G^{\gamma}(\hat{\mathbf{Y}}^1|\hat{\mathbf{Y}}^0) G^{\theta}(\hat{\mathbf{Y}}^0) \\
&= \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0\in\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y^1},\mathbf{y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{y^0}) \sum_{\hat{\mathbf{S}}^1}\sum_{\hat{\mathbf{X}}^1} G^{\gamma}(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1,\hat{\mathbf{X}}^1) G^{\gamma}(\hat{\mathbf{X}}^1) G^{\theta,\gamma}(\hat{\mathbf{S}}^1) \qquad (20)
\end{aligned}$$

For $\mathbf{s} \notin \mathbf{Y^0}$

$$\begin{aligned}
&G^{T(\theta,\gamma),\gamma}(\mathbf{s}) \\
&= \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0\in\mathbf{Y}^0} G^{\gamma}(\mathbf{s}|\mathbf{Y^1},\mathbf{y}^0) G^{\gamma}(\mathbf{Y^1}|\mathbf{y^0}) \sum_{\hat{\mathbf{X}}^1} G^{\gamma}(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1 = \mathbf{y}^0,\hat{\mathbf{X}}^1) G^{\gamma}(\hat{\mathbf{X}}^1) G^{\theta,\gamma}(\hat{\mathbf{S}}^1 = \mathbf{y}^0)
\end{aligned} \qquad (21)$$

where $\hat{\mathbf{Z}}^1 = \mathbf{Y}^0 - \mathbf{S}$ are all the current step counterparts of the query ancestor nodes from the previous step $\mathbf{Y}^0$; $\hat{\mathbf{X}}^1$ are any ancestor nodes of $\hat{\mathbf{Z}}^1$ other than the query node. $\hat{\mathbf{S}}^1$ is a duplicate of the query node, its value may be provided if $\mathbf{S} \subset \mathbf{Y}^0$, otherwise it will be marginalized out leading to the two forms of the update.

To show convergence of the algorithm it suffices to show that T is a contraction with respect to $\mathscr{D}$.

For the update function in (20) the distance between two distributions after applying one step of the algorithm on a particular value of the query node is:

$$\begin{aligned}
\mathscr{D}(T(\theta),T(\psi))(s) &= \left| G^{T(\theta),\gamma}(s) - G^{T(\psi),\gamma}(s) \right| \\
&= |\sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \sum_{\hat{\mathbf{S}}^1} \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1,\hat{\mathbf{X}}^1) G^\gamma(\hat{\mathbf{X}}^1) G^\theta(\hat{\mathbf{S}}^1) - \\
&\quad \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \sum_{\hat{\mathbf{S}}^1} \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1,\hat{\mathbf{X}}^1) G^\gamma(\hat{\mathbf{X}}^1) G^\psi(\hat{\mathbf{S}}^1)| \\
&= \left| \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \sum_{\hat{\mathbf{S}}^1} \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{\mathbf{S}}^1,\hat{\mathbf{X}}^1) G^\gamma(\hat{\mathbf{X}}^1) \left[G^\theta(\hat{\mathbf{S}}^1) - G^\psi(\hat{\mathbf{S}}^1) \right] \right|
\end{aligned} \tag{22}$$

Using Lemma 1 we can express $\mathscr{D}(T(\theta),T(\psi))(s)$ in terms of $\mathscr{D}(\theta,\psi)(s)$:

$$\begin{aligned}
&\mathscr{D}(T(\theta),T(\psi))(s) \\
&= \left| \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \right| G^\theta(\hat{s}^1) - G^\psi(\hat{s}^1) \Big| \\
&\qquad \left[\sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{s}^1,\hat{\mathbf{X}}^1) G^\gamma(\hat{\mathbf{X}}^1) - \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\neg\hat{s}^1,\hat{\mathbf{X}}^1) G^\gamma(\hat{\mathbf{X}}^1) \right] \Big| \\
&= \Big| \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \mathscr{D}(\theta,\psi)(s) \\
&\qquad \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{X}}^1) \left[G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{s}^1,\hat{\mathbf{X}}^1) - G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\neg\hat{s}^1,\hat{\mathbf{X}}^1) \right] \Big| \\
&= \mathscr{D}(\theta,\psi)(s) \Big| \sum_{\mathbf{Y^1}} \sum_{\mathbf{y}^0 \in \mathbf{Y}^0} G^\gamma(s|\mathbf{Y^1},\mathbf{y}^0) G^\gamma(\mathbf{Y^1}|\mathbf{y^0}) \\
&\qquad \sum_{\hat{\mathbf{X}}^1} G^\gamma(\hat{\mathbf{X}}^1) \left[G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\hat{s}^1,\hat{\mathbf{X}}^1) - G^\gamma(\hat{\mathbf{Z}}^1 = \mathbf{y}^0|\neg\hat{s}^1,\hat{\mathbf{X}}^1) \right] \Big|
\end{aligned} \tag{23}$$

$$\mathscr{D}(T(\theta),T(\psi))(s) = k\, \mathscr{D}(\theta,\psi)(s) \tag{24}$$

Lemma 2. *The estimate of the marginal distribution of a represented node in an EBN will converge to a unique fixed point using the iterative improvement algorithm.*

Proof. A constant k in terms of the fixed parameters γ is given by (24). If $0 \leq k < 1$ then the transistion mapping defined by the iterative improvement algorithm is a contraction

mapping. By Banach's theorem the transition will have a unique fixed point θ^* that it will converge to iteratively at a rate of $k/1-k$ per iteration. Note that if some of the conditional probabilities are deterministic it is possible that $k = 1$ which would remove the guarantee of convergence. □

Proposition 4. *If all previous cliques are represented in an EBN, the iterative improvement algorithm will converge to the marginal of the exact distribution for each represented set of nodes.*

Proof. (sketch) Part I: If the exact equilibrium distribution is projected onto the previous nodes, the represented nodes will have exact marginals. In computing the distribution on a represented clique, the nodes that are not ancestors can be pruned, we can then sum out the other current variables, and we end up with the probability of the corresponding previous clique, for which we have the exact distribution. Part II: (see Lemma 2) The iterative improvement algorithm converges to a unique equilibrium and the exact answer is an equilibrium, it must converge to the exact answer on the represented marginals. □

7. Evaluation

In this section we empirically evaluate the accuracy of the algorithm when it is not exact. If the structure on the previous variables does not match the condition of the proposition, we do not expect the equilibrium to be exact, but expect it to be an approximation. The following example gives an empirical evaluation of the approximation.

`M:should we be more specific about how we compute the exact answer?`

Example 7. One case where the previous variables are fully connected to satisfy Proposition 4 is when the causal network is a double linked chain:

$$A \leftrightarrow B \leftrightarrow C \leftrightarrow D \leftrightarrow \ldots$$

The iterative improvement algorithm converges to the exact answer when the previous variables are fully connected. If the previous variables are not fully connected, it converges to an approximation. To test how good various approximations are, we created such a chain with 16 binary variables. We then constructed 1000 random parametrizations (with various levels of skewed distributions, where the probabilities were of the form r^k or $1-r^k$ for r a random number and k an integer in $[1,4]$). We then chose the 10 parametrizations where the approximations were worst. Here we report on the worst parametrization (as this was typical of the others).

Instead of considering all of the ancestors in the definition of *represented*, we only considered the ancestors to a depth bound. For this example, when the depth bound is 1, the previous variables are disconnected. When the depth bound in 2, each variable, except the last one, has one parent. When the depth bound is b, each variable (other than the boundary cases) has $b-1$ parents. Figure 7 gives the number of iteration until convergence (the probabilities change by less than 10^{-15}), the sum of the errors on the marginals of the individual variables, and the maximum error for individual variables (to two significant digits) as a function of the bound:

bound	#iters	total error	max error
1	82	1.7	0.36
2	381	0.072	0.023
3	374	0.0031	0.00066
4	369	0.00016	3.8e-05
5	385	5.0e-05	1.6e-05
6	387	4.7e-07	1.7e-07
7	384	5.0e-08	1.8e-08
8	382	5.5e-09	1.2e-09
9	383	4.4e-10	2.1e-10
10	379	8.2e-11	3.6e-11

Figure 7: Empirical results from Example 7

8. Relation to DBN Belief State Monitoring

This work is related to the problem of inference in Dynamic Bayesian Networks (DBN)(Dean and Kanazawa, 1989), here we present an overview of research in that area and how our work differs from it.

Exact inference in DBNs is as complex as in normal Bayesian networks but since new variables are essentially added at each time step this can become a problem very quickly. In fact, inference quickly becomes intractable even when there are completely independent nodes at each time step if they influence each other over time. The network can quickly reach a point where all of the nodes are fully interrelated. Evaluating the probability distribution of the state of the DBN after each time step requires inference on the entire network. So the cost of inference continues increasing as time moves on and the size of the DBN increases. This *entanglement* of variables over time is a central research problem in DBNs.

8.1. Boyen-Koller algorithm

Boyen and Koller (1999) demonstrated a method, known as the *BK algorithm*, to reduce the entanglement of variables in a DBN by breaking the conditional link to the past at each step and approximating the current state. They showed that this can be done without having the approximation error increase at each step if the dynamics are stochastic.

In the BK algorithm the current state is projected to an *approximate belief state* using a factored representation. This factored, approximate belief state removes the links between clusters of interrelated nodes *within a given time slice*. This is a reasonable approximation as long as interaction between variables in different clusters is sufficiently weak Pfeffer (2006). The approximate belief state is structured so that the variables that make up the state space are divided up into small, weakly interacting components conditioned on some other aggregate variable.

The hidden model in our EBN can bee seen as a kind of projection at each step of a more complex DBN where each time slice contains the entire, two-stage EBN. Boyen and Koller (1999) provide a convergence proof for their algorithm which is related to

our but needs to also deal with observations on variables which does not happen in our model.

An example use is the distribution of cars on a highway being roughly independent given a value for overall traffic level. As long as the correlations between variables in different clusters are weak, this produces a reasonable approximation of the distribution of cars. The system dynamics may be simulated forward, projecting back to the approximate belief state at each step without increasing error unboundedly, as long as the transition dynamics have enough noise. If there are occasional strong interactions between some variables then this can also be handled by altering the projection to allow those variables to interact for some time. In the car example, cars are roughly independent most of the time until two cars come very close to each other or crash. These cars then have a very strong influence on each other's state which will continue until the cars move apart.

`M:More general background` The Boyen-Koller algorithm for approximate filtering of DBNs is a special case of EP using single cariable marginals. Pfeffer's extensions allowed sets of variables but only if only one at a time were influencing the child. It is in fact a special case of BP where a single loop is carried out at each step. EP allows for arbitrary aproximation schemes over hte variables.

8.1.1. Separability of DBNs

Pfeffer (2006) has analysed the type of factorizations that can be used effectively in the BK algorithm in terms of approximate separability by defining groups of sufficient variables needed to compute marginals. He points out that what makes inference hard in DBNs are observations which tie everything together. Our EBN factorization has a similar feel to Pfeffer (2006)'s approach however the sufficient variables we construct are not built with a notion of seperability in mind rather than a notion of covering all the correlations in the equilibrium distribution which are relevant for the represented queries.

Pfeffer showed how to compute the marginals for variables in any future state by using separable models without observations. This corresponds to each variable getting input from just one of its number of parents form the previous timestep but we do not know which parent. The conditional distributions need to be defined such that only one of the parents of a variable ever influences its distribution. He points out that *causal independence* such as noisy-or allows a similar breakup of the influence of parents. However, causally independent models allow all parents to influence through some aggregation function whereas separability models the case where a single parent from the previous timestep is the influence, but we do not know which parent it is.

> Unfortunately, observations tend to break sufficiency. The problem is that even if we have the correct marginals over a family of subsets of variables, we do not have the joint distribution over all the variables.
> –Pfeffer (2006)

Our approach takes this observation at face value and attempts to compute exactly only certain queries of marginal distributions. Causally independent and separable models are quite different from the problem we address here. We can model multiple

interrelated influences from the previous step but with a constrained structure over them given by the causal model.

8.2. How Our Approach Differs [better title needed]

Existing DBN inference research with which we are aware assumes that the goal is always to produce a full joint distribution and that projections which are used to reduce interrelation of variables can only involve existing variables. We have demonstrated that if a new set of variables with a different structure are allowed then exact monitoring of marginal distributions is possible. We also found that in some cases the full joint can be computed with an compact representation of these latent variables.

Unlike the standard DBN projection methods, the iterative improvement algorithm and EBN model presented here has the ability to compute exact distributions over a range of sets of query nodes from the marginals on individual variables up to the full joint distribution over all variables.

9. Conclusion

The idea that a causal model means the equilibrium of a Markov chain is not new; Strotz (1960) argued that when there are variables that are interdependent in a cyclic ordering, the fixed point in values was a specification error. Others (e.g. Fisher, 1970) followed up by giving conditions for the equilibrium to be well defined. There is also a vast literature on learning cyclic causal models that focuses on the interpretation that causal cycles are caused by unmeasured latent variables (Glymour and Spirtes, 1988; Schmidt and Murphy, 2009, e.g.,) which is largely unrelated to the discussion here.

This discussion also provides a theoretical grounding for the use of cyclic causal equilibrium models as the basis for multi-action stochastic policies. These policies can model multi-agent behaviour such as the gold market example or spatial policies such as land use in environmental planning. The ability to model consistent global distributions using only local, conditional distributions is an attractive property. The approximate belief model we propose makes full use of the structure of the causal model and provides a natural interpretation for approximate forms in terms of proximty of each 'agent' to its relevant neighbours.

If SEMs are the right model for causality, they should work for simple cases. In this paper, we have argued that they impose undesirable dependencies, and proposed MC equilibrium models as an alternative. We gave an algorithm for constructing a network that can answer queries about the equilibrium.

It should also be noted that the counterexample of Neal (2000) to Pearl and Decter (1996) does not work for the equilibrium semantics. D-separation holds with the MC equilibrium semantics as that uses a directed network.

While this paper has assumed a fixed sample ordering, a distribution over sample orderings will allow for more flexible modelling.

The EBN iterative improvement algorithm presented here can provide exact solutions and bounded error approximations for cyclic causal models. This provides grounding for the feasibility of using cyclic causal models for representing spatial, stochastic policies. While we can represent spatial stochastic policies with cyclic causal

models the EBN method described here cannot yet be used directly for large planning problems due to the computational cost of exact inference on models with many variables. This method could provide improvements when used as a submodule within stochastic simulation method such as in the inner inference loop within a block-Gibbs sampler.

References

Agarwal, R., Meehan, M., O'Regan, D., 2001. Fixed Point Theory and Applications. Cambridge University Press.

Boyen, X., Koller, D., 1998. Tractable inference for complex stochastic processes. In: Fourteenth Annual Conference on Uncertainty in AI (UAI). No. 42. p. 33.

Boyen, X., Koller, D., 1999. Exploiting the architecture of dynamic systems. In: AAAI '99.
URL

Bremaud, P., 1999. Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues. Springer.

Crowley, M., Poole, D., 2011. Policy gradient planning for environmental decision making with existing simulator (to appear). In: AAAI2011.

Dash, D., 2005. Restructuring dynamic causal systems in equilibrium. In: 10th Int. Workshop on AI and Stats. pp. 81–88.

Dean, T., Kanazawa, K., 1989. A model for reasoning about persistence and causation. Computational Intelligence 5, 142–150.

Fisher, F. M., 1970. A Correspondence Principle for Simultaneous Equation Models. Econometrica 38 (1).
URL

Glymour, C., Spirtes, P., 1988. Latent variables, causal models and overidentifying constraints. journal of Econometrics 39 (1-2), 175–198.

Halpern, J. Y., 2000. Axiomatizing Causal Reasoning. J. Artif. Intell. Res 12, 202–210.
URL

Iwasaki, Y., Simon, H. A., May 1994. Causality and model abstraction. Artificial Intelligence 67 (1), 143–194.
URL

Neal, R., 2000. On deducing conditional independence from d-separation in causal graphs with feedback (research note). JAIR 12, 87–91.

Pearl, J., 1988a. Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference. Morgan Kaufmann, San Mateo, California.

Pearl, J., 1988b. Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference. Morgan Kaufmann, San Mateo, California.

Pearl, J., 2009. Causality: Models, Reasoning and Inference, 2nd Edition. Cambridge University Press.

Pearl, J., Decter, R., 1996. Identifying independencies in causal graphs with feedback. In: 12th Conference on Uncertainty in AI. pp. 420 – 426.

Pfeffer, A., 2006. Approximate separability for weak interaction in dynamic systems. In: UAI'06.

Schmidt, M., Murphy, K., 2009. Modeling discrete interventional data using directed cyclic graphical models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. UAI '09. AUAI Press, Arlington, Virginia, United States, pp. 487–495.
URL

Strotz, R. H., 1960. Interdependence As a Specification Error. Econometrica 28 (2).
URL

Strotz, R. H., Wold, H. O. A., 1960. Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems). Econometrica 28 (2).
URL

Zhang, N., Poole, D., 1994. A simple approach to Bayesian network computations. In: Proceedings of the Tenth Canadian Conference on AI. pp. 171–178.