

Approximating Discrete Probability Distributions with Causal Dependence Trees

Christopher J. Quinn
Department of Electrical and
Computer Engineering
University of Illinois
Urbana, Illinois 61801
Email: quinn7@illinois.edu

Todd P. Coleman
Department of Electrical and
Computer Engineering
University of Illinois
Urbana, Illinois 61801
Email: colemant@illinois.edu

Negar Kiyavash
Department of Industrial and
Enterprise Systems Engineering
University of Illinois
Urbana, Illinois 61801
Email: kiyavash@illinois.edu

Abstract—Chow and Liu considered the problem of approximating discrete joint distributions with dependence tree distributions where the goodness of the approximations were measured in terms of KL distance. They (i) demonstrated that the minimum divergence approximation was the tree with maximum sum of mutual informations, and (ii) specified a low-complexity minimum-weight spanning tree algorithm to find the optimal tree. In this paper, we consider an analogous problem of approximating the joint distribution on discrete random processes with causal, directed, dependence trees, where the approximation is again measured in terms of KL distance. We (i) demonstrate that the minimum divergence approximation is the directed tree with maximum sum of directed informations, and (ii) specify a low-complexity minimum weight directed spanning tree, or arborescence, algorithm to find the optimal tree. We also present an example to demonstrate the algorithm.

I. INTRODUCTION

Numerous statistical learning, inference, prediction, and communication problems require storing the joint distribution for a large number of random variables. As the number of variables increases linearly, the number of elements in the joint distribution increases multiplicatively by the cardinality of the alphabets. Thus, for storage and analysis purposes, it is often desirable to approximate to the full joint distribution.

Bayesian networks is an area of research within which methods have been developed to approximate or simplify a full joint distribution with an approximating distribution [1]–[10]. In general, there are various choices of the structure of the approximating distribution. Chow and Liu developed a method of approximating a full joint distribution with a dependence tree distribution [11]. The joint distribution is represented as a product of marginals, where each random variable is conditioned on at most one other random variable. For Bayesian networks, graphical models are often used to represent distributions. Variables are represented as nodes and undirected edges between pairs of variables depict statistical dependence. A variable is statistically independent of all of the variables it does not share an edge with, given the variables it does share an edge with [10]. The dependence tree distributions have graphical representations as trees (a graph without any loops). Chow and Liu’s procedure efficiently computes the “best” approximating tree for a given joint distribution,

where “best” is defined in terms of the Kullback-Liebler (KL) divergence between the original joint distribution and the approximating tree distribution [11]. They also showed that finding the “best” fitting tree was equivalent to maximizing a sum of mutual informations [11].

This work will consider the specific setting where there are random processes with a common index set (interpreted as timing). In this setting, there are several potential problems with using the Chow and Liu procedure. First, applying the procedure would intermix the variables between the different processes. For some research problems, it is desired to keep the processes separate. Second, the procedure would ignore the index set, which would result in a loss of timing information and thus causal structure, which could be useful for learning, inference, and prediction problems. Third, the Chow and Liu procedure becomes computationally expensive. For example, if there are four processes, each of length 1000, the Chow and Liu procedure will search for the best fitting tree distribution over the 4000 variables (trees over 4000 nodes). Even if there are simple, causal relationships between the processes, the procedure will be computationally expensive.

We will present a procedure similar to Chow and Liu’s, but for this setting. To do so, we will use the framework of directed information, which was formally established 20 years ago by James Massey in his ISITA paper [12]. Analogous to the result of Chow and Liu, finding the “best” fitting causal dependence tree (between the processes) is equivalent to maximizing a sum of directed informations. Unlike the Chow and Liu procedure, however, our procedure will keep the processes intact and will not ignore the timing information. Thus, our procedure will be able to identify causal relationships between the processes. Also, we propose an efficient algorithm which will have a computational complexity that scales with the number of processes, but *not* the length of the processes. In the above example of four processes of length 1000, the search for the best fitting causal tree distribution will be over the four processes (directed trees over four nodes). Lastly, we present an example of the algorithm, where it identifies the optimal causal dependence tree distribution, in terms of KL distance, for a set of six random processes.

II. DEFINITIONS

This section presents probabilistic notations and information-theoretic definitions and identities that will be used throughout the remainder of the manuscript. Unless otherwise noted, the definitions and identities come from Thomas and Cover [13].

- For integers $i \leq j$, define $x_i^j \triangleq (x_i, \dots, x_j)$. For brevity, define $x^n \triangleq x_1^n = (x_1, \dots, x_n)$.
- Throughout this paper, $\mathcal{X}$ corresponds to a measurable space that a random variable, denoted with upper-case letters (X), takes values in, and lower-case values $x \in \mathcal{X}$ correspond to specific realizations.
- Define the probability mass function (PMF) of a discrete random variable by

$$P_X(x) \triangleq P(X = x).$$

- For a length n , discrete random vector, denoted as $X^n \triangleq (X_1, \dots, X_n)$, the joint PMF is defined as

$$P_{X^n}(x^n) \triangleq P(X^n = x^n).$$

Let $P_{X^n}(\cdot)$ denote $P_{X^n}(x^n)$ (when the argument is implied by context).

- For two random vectors X^n and Y^m , the conditional probability $P(X^n|Y^m)$ is defined as

$$P(X^n|Y^m) \triangleq \frac{P(X^n, Y^m)}{P(Y^m)}.$$

- The chain rule for joint probabilities is

$$P_{X^n|Y^n}(x^n|y^n) = \prod_{i=1}^n P_{X_i|X^{i-1}, Y^n}(x_i|x^{i-1}, y^n).$$

- Causal conditioning $P_{X^n||Y^n}(x^n||y^n)$, introduced by Kramer [14], is defined by the property

$$P_{X^n||Y^n}(x^n||y^n) \triangleq \prod_{i=1}^n P_{X_i|X^{i-1}, Y^i}(x_i|x^{i-1}, y^i). \quad (1)$$

- For two probability distributions P and Q on $\mathcal{X}$, the Kullback-Leibler divergence is given by

$$D(P||Q) \triangleq \mathbb{E}_P \left[\log \frac{P(X)}{Q(X)} \right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \geq 0.$$

- The mutual information between random variables X and Y is given by

$$I(X; Y) \triangleq D(P_{XY}(\cdot, \cdot) || P_X(\cdot) P_Y(\cdot)) \quad (2a)$$

$$= \mathbb{E}_{P_{XY}} \left[\log \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] \quad (2b)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_{Y|X}(y|x)}{P_Y(y)}. \quad (2c)$$

The mutual information is known to be symmetric: $I(X; Y) = I(Y; X)$.

- We denote a random process by $\mathbf{X} = \{X_i\}_{i=1}^n$, with associated $P_{\mathbf{X}}(\cdot)$ which induces the joint probability distribution of all finite collections of $\mathbf{X}$.

- The directed information from a process $\mathbf{X}$ to a process $\mathbf{Y}$, both of length n , is defined by

$$I(\mathbf{X} \rightarrow \mathbf{Y}) \quad (3a)$$

$$\triangleq \frac{1}{n} I(X^n \rightarrow Y^n) \quad (3b)$$

$$= \frac{1}{n} \sum_{i=1}^n I(Y_i; X^i | Y^{i-1}) \quad (3c)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_{X^i, Y^i}} \left[\log \frac{P_{Y_i|X^i, Y^{i-1}}(Y_i|X^i, Y^{i-1})}{P_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})} \right] \quad (3d)$$

$$= \frac{1}{n} \mathbb{E}_{P_{X^n, Y^n}} \left[\log \frac{P_{Y^n||X^n}(Y^n||X^n)}{P_{Y^n}(Y^n)} \right] \quad (3e)$$

$$= \frac{1}{n} D(P_{Y^n||X^n}(Y^n||X^n) || P_{Y^n}(Y^n)). \quad (3f)$$

Directed information was formally introduced by Massey [12]. It was motivated by Marko's work [15]. Related work was independently done by Rissanen [16]. It is philosophically grounded in Granger causality [17]. It has since been investigated in a number of research settings, and shown to play a fundamental role in communication with feedback [12], [14], [15], [18]–[20], prediction with causal side info [16], [21], gambling with causal side information [22], [23], control over noisy channels [18], [24]–[27], and source coding with feed forward [23], [28]. Conceptually, mutual information and directed information are related. However, while mutual information quantifies correlation (in the colloquial sense of statistical interdependence), directed information quantifies *causation*.

- Denote permutations on $\{1, \dots, n\}$ by $\pi(\cdot)$.
- Define functions, denoted by $j(\cdot)$, on $\{1, \dots, n\}$ to have the property $1 \leq j(i) < i \leq n$.

III. BACKGROUND: DEPENDENCE TREE APPROXIMATIONS

Given a set of n discrete random variables $X^n = \{X_1, X_2, \dots, X_n\}$, possibly over different alphabets, the chain rule is

$$\begin{aligned} P_{X^n}(\cdot) &= P_{X_n|X^{n-1}}(\cdot) P_{X_{n-1}|X^{n-2}}(\cdot) \cdots P_{X_1}(\cdot) \\ &= P_{X_1}(\cdot) \prod_{i=2}^n P_{X_i|X^{i-1}}(\cdot). \end{aligned}$$

For the chain rule, the order of the random variables does not matter, so for any permutation $\pi(\cdot)$ on $\{1, \dots, n\}$,

$$P_{X^n}(\cdot) = \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \dots, X_{\pi(1)}}(\cdot).$$

Chow and Liu developed an algorithm to approximate a known, full joint distribution by a product of second order distributions [11]. For their procedure, the chain rule is applied to the joint distribution, and all the terms of the form $P_{X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \dots, X_{\pi(1)}}(\cdot)$ are approximated (possibly exactly) by $P_{X_{\pi(i)}|X_{\pi(j(i))}}(\cdot)$ where $j(i) \in \{1, \dots, i-1\}$, such that the conditioning is on at most one variable. This product

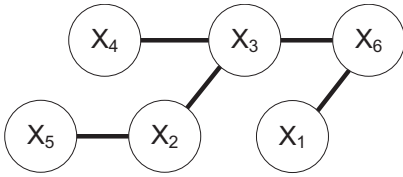


Fig. 1. Diagram of an approximating dependence tree structure. In this example, $\hat{P}_T \approx P_{X_6}(\cdot)P_{X_1|X_6}(\cdot)P_{X_3|X_6}(\cdot)P_{X_4|X_3}(\cdot)P_{X_2|X_3}(\cdot)P_{X_5|X_2}(\cdot)$.

of second order distributions serves as an approximation of the full joint.

$$P_{X^n}(\cdot) \approx \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(j(i))}}(\cdot).$$

This approximation has a tree dependence structure. Dependence tree structures have graphical representations as trees, which are graphs where all the nodes are connected and there are no loops. This follows because application of the chain rule induces a dependence structure which has no loops (e.g., no terms of the form $P_{A|B}(a|b)P_{B|C}(b|c)P_{C|A}(c|a)$), and this is a reduction of that structure, so it does not introduce any loops. An example of an approximating tree dependence structure is shown in Figure 1. In general, the approximation will not be exact. Denote each tree approximation of $P_{X^n}(x^n)$ by $\hat{P}_T(x^n)$. Each choice of $\pi(\cdot)$ and $j(\cdot)$ over $\{1, \dots, n\}$ completely specifies a tree structure T . Thus, the tree approximation of the joint using the particular tree T is

$$\hat{P}_T(x_1^n) \triangleq \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(j(i))}}(x_{\pi(i)}|x_{\pi(j(i))}). \quad (4)$$

Denote the set of all possible trees T by $\mathcal{T}$.

Chow and Liu's method obtains the "best" such model $T \in \mathcal{T}$, where the "goodness" is defined in terms of Kullback-Liebler (KL) distance between the original distribution and the approximating distribution [11].

Theorem 1:

$$\arg \min_{T \in \mathcal{T}} D(P_{X^n} || \hat{P}_T) = \arg \max_{T \in \mathcal{T}} \sum_{i=1}^n I(X_{\pi(i)}; X_{\pi(j(i))}) \quad (5)$$

See [11] for the proof. The optimization objective is equivalent to maximizing a sum of mutual informations.

They also propose an efficient algorithm to identify this approximating tree [11]. Calculate the mutual information between each pair of random variables. Now consider a complete, undirected graph, in which each of the random variables is represented as a node. The mutual information values can be thought of as weights for the corresponding edges. Finding the dependence tree distribution that maximizes the sum (5) is equivalent to the graph problem of finding a tree of maximal weight [11]. Kruskal's minimum spanning tree algorithm [29] can be used to reduce the complete graph to a tree with the largest sum of mutual informations [11]. If mutual information values are not unique, there could be multiple solutions. Kruskal's algorithm has runtime of $\mathcal{O}(n \log n)$ [30].

IV. MAIN RESULT: CAUSAL DEPENDENCE TREE APPROXIMATIONS

In situations where there are multiple random processes, the Chow and Liu method can be used. However, it will consider all possible arrangements of all the variables, "mixing" the processes and timings to find the best approximation. An alternative approach, which would maintain causality and keep the processes separate, is to find an approximation to the full joint probability by identifying causal dependencies between the processes themselves. In particular, consider finding a causal dependence tree structure, where instead of conditioning on a variable using one auxiliary variable as in Chow and Liu, the conditioning is on a *process* using one auxiliary process. A causal dependence tree has a corresponding graphical representation as a directed tree graph, or "arborescence." An arborescence is a graph with all of the nodes connected by directed arrows, such that there is one node with no incoming edges, the "root," and all other nodes have exactly one incoming edge [30].

Consider the joint distribution $P_{\mathbf{A}^M}$ of M random processes $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_M$, each of length n . Denote realizations of these processes as $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_M$ respectively. The joint distribution of the processes can be approximated in an analogous manner as before, except that *instead of permuting the index of the set of random variables, consider permutations on the index set of the processes themselves*. For a given joint probability distribution $P_{\mathbf{A}^M}(\vec{a}^M)$ and tree T , denote the corresponding approximating causal dependence tree induced probability to be

$$\hat{P}_T(\vec{a}^M) \triangleq \prod_{h=1}^M P_{\mathbf{A}_{\pi(h)} || \mathbf{A}_{\pi(j(h))}}(\vec{a}_{\pi(h)} || \vec{a}_{\pi(j(h))}). \quad (6)$$

Let $\mathcal{T}_C$ denote the set of all causal dependence trees.

As before, the goal is to obtain the "best" such model T , where the "goodness" is defined in terms of KL distance between the original distribution and the approximating distribution.

Theorem 2:

$$\arg \min_{T \in \mathcal{T}_C} D(P || \hat{P}_T) = \arg \max_{T \in \mathcal{T}_C} \sum_{h=1}^N I(\mathbf{A}_{\pi(j(h))} \rightarrow \mathbf{A}_{\pi(h)}) \quad (7)$$

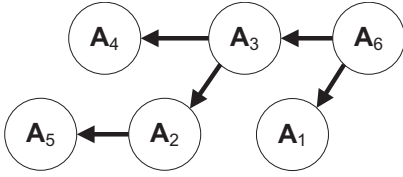


Fig. 2. Diagram of an approximating causal dependence tree structure. In this example,

$$\hat{P}_T \approx P_{A_6(\cdot)} P_{A_1||A_6(\cdot)} P_{A_3||A_6(\cdot)} P_{A_4||A_3(\cdot)} P_{A_2||A_3(\cdot)} P_{A_5||A_2(\cdot)}.$$

Proof:

$$\arg \min_{T \in \mathcal{T}_C} D(P||\hat{P}_T)$$

$$= \arg \min_T E_P \left[\log \frac{P_{\mathbf{A}^M}(\mathbf{A}^M)}{\hat{P}_T(\mathbf{A}^M)} \right] \quad (8)$$

$$= \arg \min_{T \in \mathcal{T}_C} E_P \left[\log \frac{1}{\hat{P}_T(\mathbf{A}^M)} \right] \quad (9)$$

$$= \arg \min_{T \in \mathcal{T}_C} \sum_{h=1}^M -E_P \left[\log P_{\mathbf{A}_{\pi(h)}||\mathbf{A}_{\pi(j(h))}}(\mathbf{A}_{\pi(h)}||\mathbf{A}_{\pi(j(h))}) \right] \quad (10)$$

$$= \arg \max_{T \in \mathcal{T}_C} \sum_{h=1}^M E_P \left[\log \frac{P_{\mathbf{A}_{\pi(h)}||\mathbf{A}_{\pi(j(h))}}(\mathbf{A}_{\pi(h)}||\mathbf{A}_{\pi(j(h))})}{P_{\mathbf{A}_{\pi(h)}}(\mathbf{A}_{\pi(h)})} \right] \quad (11)$$

$$+ \sum_{h=1}^M E_P [\log P_{\mathbf{A}_{\pi(h)}}(\mathbf{A}_{\pi(h)})]$$

$$= \arg \max_{T \in \mathcal{T}_C} \sum_{h=1}^M I(\mathbf{A}_{\pi(j(h))} \rightarrow \mathbf{A}_{\pi(h)}), \quad (12)$$

where (8) follows from definition of KL distance; (9) removes the numerator which does not depend on T; (10) uses (6); (11) adds and subtracts a sum of entropies; (12) uses (3) and that the sum of entropies is independent of T. ■

Thus, finding the optimal causal dependence tree in terms of KL distance is equivalent to maximizing a sum of directed informations.

V. ALGORITHM FOR FINDING THE OPTIMAL CAUSAL DEPENDENCE TREE

In Chow and Liu's work, Kruskal's minimum spanning tree algorithm performs the analogous optimization procedure efficiently, after having computed the mutual information between each pair [11]. A similar procedure can be done in this setting. First, compute the directed information between each ordered pair of processes. This can be represented as a graph, where each of the nodes represents a process. This graph will have a directed edge from each node to every other node (thus is a complete, directed graph), and the value of edge from node $\mathbf{X}$ to node $\mathbf{Y}$ will be $I(\mathbf{X} \rightarrow \mathbf{Y})$. An example of an approximating causal dependence tree, which is depicted as an arborescence, is in Figure 2. There are several efficient algorithms which can be used to find the maximum weight (sum of directed informations) arborescence of a directed graph [31], such as Chu and Liu [32] (which was independently discovered by

Edmonds [33] and Bock [34]) and a distributed algorithm by Humblet [35]. Note that in some implementations, a root is required a priori. For those, the implementation would need to be applied for each node in the graph as a root, and then the arborescence which has maximal weight among all of those would be selected. Chu and Liu's algorithm has runtime of $\mathcal{O}(M^2)$ [31]

VI. EXAMPLE

We will now illustrate the proposed algorithm with an example of jointly gaussian random processes. For continuously valued random variables, KL divergence and mutual information are well-defined and have the same properties as in the discrete case [13]. Since directed information is a sum over mutual informations (3), it too has the same properties, hence Theorem 2 analogously applies. Additionally, for jointly gaussian random processes, differential entropy, which is used in this example, is always defined [13].

Let $\mathbf{A}_1, \mathbf{A}_2, \dots$, and $\mathbf{A}_6$ denote six zero-mean, jointly gaussian random processes, each of length $n = 50$, which we constructed. Denote the full joint PDF as $f(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_6)$, and let $f(z)$ denote the corresponding marginal distribution of any subset $z \in \{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_6\}$. Let $\mathbf{Z}$ be an arbitrary vector of jointly gaussian random variables. Let K_Z denote $\mathbf{Z}$'s covariance matrix and $|K_Z|$ its determinant. Letting m denote the number of variables in $\mathbf{Z}$, the differential entropy $h(\mathbf{Z})$ is $\frac{1}{2} \log [(2\pi e)^m |K_Z|]$ [13]. For two jointly gaussian random processes X^n and Y^n , the causally conditioned differential entropy $h(Y^n||X^n)$ is

$$\begin{aligned} h(Y^n||X^n) &= \sum_{i=1}^n h(Y_i|Y^{i-1}, X^i) \\ &= \sum_{i=1}^n h(Y^i, X^i) - h(Y^{i-1}, X^i) \\ &= \sum_{i=1}^n \frac{1}{2} \log [(2\pi e)^{2i} |K_{Y^i, X^i}|] \\ &\quad - \frac{1}{2} \log [(2\pi e)^{2i-1} |K_{Y^{i-1}, X^i}|] \\ &= \sum_{i=1}^n \frac{1}{2} \log \left[(2\pi e) \frac{|K_{Y^i, X^i}|}{|K_{Y^{i-1}, X^i}|} \right]. \end{aligned}$$

The directed information values can be calculated as fol-

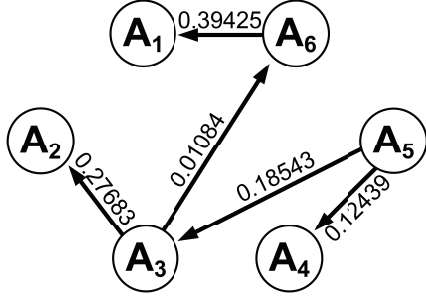


Fig. 3. Diagram of the optimal, approximating causal dependence tree structure.

lows. For jointly gaussian processes X^n to Y^n ,

$$I(X^n \rightarrow Y^n) = \sum_{i=1}^n I(Y_i; X^i | Y^{i-1}) \quad (13)$$

$$= \sum_{i=1}^n h(Y_i | Y^{i-1}) + h(X^i | Y^{i-1}) - h(Y_i, X^i | Y^{i-1}) \quad (14)$$

$$= \sum_{i=1}^n [h(Y^i) - h(Y^{i-1})] + [h(X^i, Y^{i-1}) - h(Y^{i-1})] - [h(Y^i, X^i) - h(Y^{i-1})] \quad (15)$$

$$= \sum_{i=1}^n h(Y^i) - h(Y^{i-1}) + h(X^i, Y^{i-1}) - h(Y^i, X^i) \quad (16)$$

$$= \sum_{i=1}^n \frac{1}{2} \log \left[\frac{|K_{Y^i}| |K_{X^i, Y^{i-1}}|}{|K_{Y^{i-1}}| |K_{X^i, Y^i}|} \right], \quad (17)$$

where (14) and (15) use identities [13].

The normalized, pairwise directed information values for all pairs of processes are listed in Table I. The identified maximum weight spanning tree, using Edmond's algorithm [33], [36], is shown in Figure 3. The resulting structure corresponding to the optimal causal dependence tree approximation is given by

$$f_T(\vec{a}_1, \vec{a}_2, \dots, \vec{a}_6) = f(\vec{a}_5) f(\vec{a}_4 | \vec{a}_5) f(\vec{a}_3 | \vec{a}_5) \cdot f(\vec{a}_2 | \vec{a}_3) f(\vec{a}_6 | \vec{a}_3) f(\vec{a}_1 | \vec{a}_6).$$

The KL distance $D(f || f_T)$ between the full PDF $f(\cdot)$ and the tree approximation $f_T(\cdot)$ can be computed as follows.

$$\begin{aligned} D(f || f_T) &= E_f \left[\log \frac{f(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_6)}{f_T(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_6)} \right] \\ &= E_f \left[\log \frac{1}{f_T(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_6)} \right] - h(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_6) \\ &= h(\mathbf{A}_5) + h(\mathbf{A}_4 | \mathbf{A}_5) + h(\mathbf{A}_3 | \mathbf{A}_5) + h(\mathbf{A}_2 | \mathbf{A}_3) \\ &\quad + h(\mathbf{A}_6 | \mathbf{A}_3) + h(\mathbf{A}_1 | \mathbf{A}_6) - h(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_6). \end{aligned}$$

In this example, the normalized KL distance is $\frac{1}{6n} D(f || f_T) = 0.1968$. (The normalization is $\frac{1}{6n}$ because there are $6n$ total variables.)

TABLE I
DIRECTED INFORMATION VALUES FOR EACH PAIR OF THE SIX JOINTLY GAUSSIAN RANDOM PROCESSES

$\nearrow$	A	B	C	D	E	F
A	0.00000	0.02339	0.00241	0.00028	0.00026	0.03487
B	0.15180	0.00000	0.17765	0.00305	0.00322	0.00438
C	0.03242	0.27683	0.00000	0.00385	0.00747	0.01084
D	0.00995	0.20656	0.15782	0.00000	0.10509	0.00287
E	0.00979	0.19596	0.18543	0.12439	0.00000	0.00233
F	0.39425	0.01782	0.00702	0.00278	0.00265	0.00000

VII. CONCLUSION

This work develops a procedure, similar to Chow and Liu's, for finding the "best" approximation (in terms of KL divergence) of a full, joint distribution over a set of random processes, using a causal dependence tree distribution. Chow and Liu's procedure had been shown to be equivalent to maximizing a sum of mutual informations, and the procedure presented here is shown to be equivalent to maximizing a sum of directed informations. An efficient algorithm is proposed to find the optimal causal dependence tree, analogous to an algorithm proposed by Chow and Liu. An example with six processes is presented to demonstrate the procedure.

REFERENCES

- [1] J. Pearl, *Causality: models, reasoning, and inference*, 2nd ed. Cambridge University Press, 2009.
- [2] W. Lam and F. Bacchus, "Learning Bayesian belief networks: An approach based on the MDL principle," *Computational intelligence*, vol. 10, no. 3, pp. 269–293, 1994.
- [3] N. Friedman and M. Goldszmidt, "Learning Bayesian networks with local structure," *Learning in graphical models*, pp. 421–460, 1998.
- [4] N. Friedman and D. Koller, "Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks," *Machine Learning*, vol. 50, no. 1, pp. 95–125, 2003.
- [5] D. Heckerman, "Bayesian networks for knowledge discovery," *Advances in knowledge discovery and data mining*, vol. 11, pp. 273–305, 1996.
- [6] M. Koivisto and K. Sood, "Exact Bayesian structure discovery in Bayesian networks," *The Journal of Machine Learning Research*, vol. 5, p. 573, 2004.
- [7] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu, "Learning Bayesian networks from data: an information-theory based approach," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 43–90, 2002.
- [8] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. dissertation, UNIVERSITY OF CALIFORNIA, 2002.
- [9] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.
- [10] D. Heckerman, "A tutorial on learning with Bayesian networks," *Innovations in Bayesian Networks*, pp. 33–82, 2008.
- [11] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.
- [12] J. Massey, "Causality, feedback and directed information," in *Proc. 1990 Intl. Symp. on Info. Th. and its Applications*. Citeseer, 1990, pp. 27–30.
- [13] T. Cover and J. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.
- [14] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, University of Manitoba, Canada, 1998.
- [15] H. Marko, "The bidirectional communication theory—a generalization of information theory," *Communications, IEEE Transactions on*, vol. 21, no. 12, pp. 1345–1351, Dec 1973.
- [16] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series (Corresp.)," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 598–601, 1987.

- [17] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
- [18] S. Tatikonda and S. Mitter, "The Capacity of Channels With Feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, 2009.
- [19] H. Permuter, T. Weissman, and A. Goldsmith, "Finite State Channels With Time-Invariant Deterministic Feedback," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, 2009.
- [20] J. Massey and P. Massey, "Conservation of mutual and directed information," in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, 2005, pp. 157–158.
- [21] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, 2010, accepted.
- [22] H. Permuter, Y. Kim, and T. Weissman, "On directed information and gambling," in *IEEE International Symposium on Information Theory, 2008. ISIT 2008*, 2008, pp. 1403–1407.
- [23] —, "Interpretations of Directed Information in Portfolio Theory, Data Compression, and Hypothesis Testing," *Arxiv preprint arXiv:0912.4872*, 2009.
- [24] N. Elia, "When bode meets shannon: control-oriented feedback communication schemes," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1477 – 1488, sept. 2004.
- [25] N. Martins and M. Dahleh, "Feedback control in the presence of noisy channels: "bode-like" fundamental limitations of performance," *Automatic Control, IEEE Transactions on*, vol. 53, no. 7, pp. 1604 – 1615, aug. 2008.
- [26] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.
- [27] S. Gorantla and T. Coleman, "On Reversible Markov Chains and Maximization of Directed Information," *submitted to IEEE International Symposium on Information Theory (ISIT)*, Jan 2010.
- [28] R. Venkataramanan and S. Pradhan, "Source coding with feed-forward: rate-distortion theorems and error exponents for a general source," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2154–2179, 2007.
- [29] J. Kruskal Jr, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.
- [30] J. Evans and E. Minieka, *Optimization algorithms for networks and graphs*, 2nd ed. Dekker, 1992.
- [31] H. Gabow, Z. Galil, T. Spencer, and R. Tarjan, "Efficient algorithms for finding minimum spanning trees in undirected and directed graphs," *Combinatorica*, vol. 6, no. 2, pp. 109–122, 1986.
- [32] Y. Chu and T. Liu, "On the shortest arborescence of a directed graph," *Science Sinica*, vol. 14, no. 1396-1400, p. 270, 1965.
- [33] J. Edmonds, "Optimum branchings," *J. Res. Natl. Bur. Stand., Sect. B*, vol. 71, pp. 233–240, 1967.
- [34] F. Bock, "An algorithm to construct a minimum directed spanning tree in a directed network," *Developments in operations research*, vol. 1, pp. 29–44, 1971.
- [35] P. Humblet, "A distributed algorithm for minimum weight directed spanning trees," *Communications, IEEE Transactions on*, vol. 31, no. 6, pp. 756–762, 1983.
- [36] A. Tofigh and E. Sjölund, "Edmond's Algorithm," <http://edmonds-alg.sourceforge.net/>, 2010, [Online; accessed 12-July-2010].