LEARNING MAX-WEIGHT DISCRIMINATIVE FORESTS

Vincent Y. F. Tan, John W. Fisher III and Alan S. Willsky

Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ABSTRACT

We present a method for sequential learning of increasingly complex graphical models for discriminating between two hypotheses. We generate forests for each hypothesis, each with no more edges than a spanning tree, which optimize an information-theoretic criteria. The method relies on a straightforward extension of the efficient Max-Weight Spanning Tree (MWST) algorithm by incorporating multi-valued edge-weights. Each iteration produces nested forests with increasing number of edges; each provably optimal as compared to alternative forests. Empirical results demonstrate superior probability of error as compared to generative approaches.

Keywords: Learning Graphical Models, Hypothesis Testing, Max-Weight Trees/Forests, Discriminative Learning.

1. INTRODUCTION

Reduced-order modeling of probability distributions is an important problem, particularly in high-dimensional spaces. The usual goal in learning graphical models is to construct probability models that are good *approximators* of an underlying *generative* distribution. For instance, the seminal paper by Chow and Liu [1] provides an elegant Max-Weight Spanning Tree (MWST) algorithm for learning a tree-structured distribution $\widehat{p}(x)$ that is closest, in the *Kullback-Leibler* (KL) divergence [2] sense, to the empirical (or any other given) distribution $p(x)$. This particular approach is of note in that the optimization relies solely on knowledge (or estimates) of marginal and pairwise distributions over elements of the probability model independent of the form of the source distribution. There has also been much work (e.g. [3, 4]) for learning thin graphical models that are, in some sense, optimal.

Here, we consider the problem of constructing tree approximations for the purposes of *discrimination*, that is, optimized for *hypothesis testing*. While the development focuses on the binary hypothesis testing case, the approach is easily extended to the M-ary case. We show, that with some modifications the approach of Chow and Liu can be extended to the discriminative case and that by redefining edge weights, Kruskal's MWST algorithm [5] can be utilized to efficiently generate a sequence of discriminative forest models of increasing order. In general, the discriminative tree approximations differ from the Chow-Liu derived generative models in their structure. In addition, whereas in the generative approach one can always improve the approximation by adding edges (subject to the tree constraint), in the discriminative case it is quite possible that early termination is optimal.

Vincent Tan (vtan@mit.edu) is supported by the Agency for Science, Technology and Research (A*STAR), Singapore.

John Fisher is supported by the AFOSR under Award No. FA9550-06-1-0324. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Air Force.

There are a variety of reasons for constructing reduced-order approximations, consequently we are also interested in building successively more complex probability models that are each optimal in an information theoretic sense. Given distributions p and q (in actuality one needs only the sets of marginal and pairwise distributions consistent with p and q), we learn thin *forests* $\widehat{p}$ and $\widehat{q}$ with fewer edges than a spanning tree. At each iteration, the resulting forests are nested. It can also be shown that the sequence of forest approximations provide increasingly tighter bounds on the symmetrized KL-divergence [2] as compared to the full source distributions. We validate the approach with some numerical experiments.

2. GRAPHICAL MODELS & HYPOTHESIS TESTING

A multivariate probability distribution $p(x)$ may be defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes of the graph $\mathcal{V}$ denote random variables and the edges $\mathcal{E} \subset \binom{\mathcal{V}}{2}$ indicate statistical dependencies between variables $\{x_s | s \in \mathcal{V}\}$. Graphical models [6] can be viewed as generalizations of Markov chains to arbitrary undirected graphs. The Markov property for general graphs (including chains) is that given its neighbors, any node is independent of the rest of the variables in the model i.e. $p(x_s|x_{\mathcal{N}(s)}) = p(x_s|x_{\mathcal{V}\setminus s})$.

Assume we are given two high-dimensional distributions $p(x)$ and $q(x)$ Markov with respect to graphs $\mathcal{G}_p$ and $\mathcal{G}_q$ (each possibly fully connected). We seek lower-order approximations $\widehat{p}(x)$ and $\widehat{q}(x)$ defined on graphs $\mathcal{G}_{\widehat{p}}$ and $\mathcal{G}_{\widehat{q}}$ such that these are good *classifiers*. Subsequently, the approximate models would be utilized within a likelihood ratio test (LRT). Given the binary hypothesis test

$$H_0 : x \sim p \quad \text{or} \quad H_1 : x \sim q \tag{1}$$

where $x = (x_1, \ldots, x_n)'$ is a length-n random vector, the LRT is approximated

$$\frac{P_0\, \widehat{p}(x)}{P_1\, \widehat{q}(x)} \underset{\text{declare } H_1}{\overset{\text{declare } H_0}{\gtrless}} 1. \tag{2}$$

To do this, we first consider a simpler but nonetheless related problem in the following section. This will provide us with key insights on how to construct increasingly complex probability models for hypothesis testing.

3. MODELING A DISTRIBUTION WITH A FOREST

In this section, we approximate general probability distributions defined on graphs with lower-order distributions. Let us define the set of acyclical graphs (with no cycles) with n nodes and containing k edges to be $\mathcal{T}^{(k)}$. Note that if $k = n-1$, then $\mathcal{T}^{(n-1)}$ would be the set of trees with n nodes. If $k < n-1$ then $\mathcal{T}^{(k)}$ would be a forest. Also, given edge weights w_{st} for all $s, t \in \mathcal{V}$, let us define the 'k-edge' MWST algorithm as the greedy Kruskal [5] algorithm with only $k \leq n-1$ edges selected.

Algorithm 1 The 'k-edge' MWST algorithm

Require: $1 \leq k \leq n-1$, w_{st};
$\mathcal{T}^{(k)} = \{\}$;
w_{st} = Sort(w_{st});
for $i = 1 : k$ **do**
  if (s,t) does not form a cycle in edges in tree **then**
    $\mathcal{T}^{(k)} \leftarrow \mathcal{T}^{(k)} \cup (s,t)$;
  end if
end for

Consider the following problem: We are provided with a distribution $p(x)$ defined on a graph $\mathcal{G}_p$ that is possibly fully connected. We would like to approximate $p(x)$ with a lower-order distribution $\widehat{p}_{\mathcal{T}}^{(k)}(x)$ that is Markov on $\mathcal{T}^{(k)} = (\mathcal{V}, \mathcal{E}^{(k)})$. For this, we choose to minimize the KL-divergence between $p(x)$ and $\widehat{p}_{\mathcal{T}}^{(k)}(x)$ i.e.

$$\widehat{p}_{\mathcal{T}}^{(k)}(x) = \underset{\widehat{p}(x) \in \mathcal{T}^{(k)}}{\operatorname{argmin}} \; D(p(x) \,\|\, \widehat{p}(x)). \tag{3}$$

Lemma 3.1 (Chow-Liu [1]) *$\widehat{p}_{\mathcal{T}}^{(k)}(x)$ can be optimally chosen via the 'k-edge' MWST algorithm with edge weights given by $w_{st} = I(x_s; x_t)$, the mutual information (MI) between x_s and x_t.*

Proof By the assumption that $\widehat{p}_{\mathcal{T}}^{(k)}(x) \in \mathcal{T}^{(k)}$,

$$D(p(x)||\widehat{p}(x)) = -\sum_{(s,t)\in\mathcal{E}^{(k)}} I(x_s;x_t) + \sum_{s\in\mathcal{V}} H(x_s) - H(x). \tag{4}$$

Since we are only concerned about optimizing over the choice of elements in the edge set $\mathcal{E}^{(k)}$, we can equivalently choose to maximize $\sum_{(s,t)\in\mathcal{E}^{(k)}} I(x_s;x_t)$. Thus, this reduces to a 'k-edge' MWST procedure. The 'k-edge' MWST returns the best forest (with k edges) at each iteration. □

While straightforward, Lemma 3.1 provides the main insight into the selection of $k \leq n-1$ edges for representing approximating an arbitrary graphical model. It implies that we can employ the canonical MWST algorithm and terminate once we have chosen k edges. The resulting graphical model would then be the optimal forest with k edges.

Corollary 3.2 *The edge sets $\mathcal{E}^{(k)}$ obtained from the minimization of the KL-divergence in Eqn (3) are nested i.e.*

$$\mathcal{E}^{(k)} \subset \mathcal{E}^{(k+1)}, \quad \forall\, k = 0, \ldots, n-2. \tag{5}$$

The 'k-edge' MWST algorithm gives us a series of nested trees with increasing number of edges which are optimal at each iteration, with respect to the criterion in Eqn (3). The observation that Kruskal's algorithm [5] produces an optimal sequence of k-edge forests is, in fact, well known. However, our application of it to discriminative forests is original.

4. LEARNING FORESTS SEQUENTIALLY FOR DISCRIMINATION

We now return to the discrimination problem. Given two known probability distributions $p(x)$ and $q(x)$ defined on arbitrary graphs, we would like to construct lower-order models $\widehat{p}(x)$ and $\widehat{q}(x)$ for the specific purpose of hypothesis testing. We proceed along the same line of argument as in section 3.

4.1. Formulation of objective function: The J-divergence

We formulate and maximize the J-divergence which, in turn, optimizes bounds on the probability of error for classifying new samples. We learn a sequence of models $\widehat{p}^{(k)}(x)$ and $\widehat{q}^{(k)}(x)$, each defined on forests $\mathcal{T}^{(k)}$, to discriminate between the two hypotheses.

Definition The J-divergence between two probability distributions $p(x)$ and $q(x)$ is defined as

$$J(p(x), q(x)) = D(p(x) \,\|\, q(x)) + D(q(x) \,\|\, p(x)). \tag{6}$$

Note that J is symmetric in its arguments, unlike the Kullback-Leibler divergence. We will be maximizing the J-divergence to select the elements in the edge sets $\mathcal{E}^{(k)}$ for hypothesis testing. On the use of J-divergence, it is worth noting the following upper and lower bounds on the probability of error $\Pr(\text{err})$ [7].

$$\frac{1}{2}\min(P_0, P_1) e^{-J} \leq \Pr(\text{err}) \leq \sqrt{P_0 P_1}\left(\frac{J}{4}\right)^{-1/4}, \tag{7}$$

where P_0 and P_1 are the prior probabilities of H_0 and H_1 respectively. Consequently, we conjecture that maximizing the J-divergence between the probability models we learn i.e. $\widehat{p}^{(k)}(x), \widehat{q}^{(k)}(x)$ minimizes an upper and a lower bound on $\Pr(\text{err})$. In section 5, we empirically show that this is indeed the case. In fact the actual probability of error $\Pr(\text{err})$ is reduced via discriminative learning.

As in section 3, maximizing the J-divergence is achieved by selecting a only subset of the edges $\mathcal{E}^{(k)}$ at the k^{th} iteration. More precisely, we seek lower-order models $\widehat{p}^{(k)}(x), \widehat{q}^{(k)}(x)$ such that

$$(\widehat{p}^{(k)}(x), \widehat{q}^{(k)}(x)) = \underset{\widehat{p},\widehat{q}\in\mathcal{T}^{(k)}}{\operatorname{argmax}} \; J(p,q). \tag{8}$$

In performing the optimization, we make extensive use of the easily shown identities.

$$\int p(x) \log\left(\frac{p(x_s)}{q(x_s)}\right) dx = D(p(x_s) \,||\, q(x_s)) \tag{9}$$

$$\int p(x) \log\left(\frac{p(x_s,x_t)}{q(x_s,x_t)}\right) dx = D(p(x_s,x_t) \,||\, q(x_s,x_t)) \tag{10}$$

where $s, t \in \mathcal{V}$.

4.2. Solution of the problem: Maximizing $J(\widehat{p}, \widehat{q})$

The source distributions p, q are Markov on general graphs $\mathcal{G}_p$ and $\mathcal{G}_q$ while the reduced-order distributions $\widehat{p}^{(k)}, \widehat{q}^{(k)}$ are Markov on *forests* $\mathcal{T}_p^{(k)}, \mathcal{T}_q^{(k)}$. Because of the tree assumption, the J-divergence admits a closed form solution in terms of marginal and pairwise information theoretic quantities.

Lemma 4.1 *The J-divergence of $\widehat{p}$ and $\widehat{q}$ can be expressed as*

$$J(\widehat{p}, \widehat{q}) = \int_x (p(x) - q(x)) \log\left(\frac{\widehat{p}(x)}{\widehat{q}(x)}\right) dx \tag{11}$$

$$= \sum_{s\in\mathcal{V}} J(p_s, q_s) + \sum_{(s,t)\in\mathcal{E}_p\cup\mathcal{E}_q} w_{st} \tag{12}$$

and the edge weights w_{st} are given by

$$w_{st} = \begin{cases} I_p(x_s;x_t) - I_q(x_s;x_t) \\ \quad + D(q_{s,t}||p_{s,t}) - D(q_s q_t||p_s p_t) & (s,t) \in \mathcal{E}_p \setminus \mathcal{E}_{pq} \\ I_q(x_s;x_t) - I_p(x_s;x_t) \\ \quad + D(p_{s,t}||q_{s,t}) - D(p_s p_t||q_s q_t) & (s,t) \in \mathcal{E}_q \setminus \mathcal{E}_{pq} \\ J(p_{st}, q_{st}) - J(p_s p_t, q_s q_t) & (s,t) \in \mathcal{E}_{pq} \end{cases} \tag{13}$$

where $p_s = p(x_s)$ and $p_{s,t} = p(x_s, x_t)$, $\mathcal{E}_{pq} = \mathcal{E}_p \cap \mathcal{E}_q$ denotes the intersection of the edge sets $\mathcal{E}_p$ and $\mathcal{E}_q$. $I_p(x_s; x_t)$, $I_q(x_s; x_t)$ denote the mutual information between nodes s and t under the p and q probability models, respectively.

Proof sketch: Since $\widehat{p}$ is defined as a tree distribution, it admits the factorization

$$\widehat{p}(x) = \prod_{s \in \mathcal{V}} p(x_s) \prod_{(s,t) \in \mathcal{E}_p} \frac{p(x_s, x_t)}{p(x_s)p(x_t)}. \tag{14}$$

$\widehat{q}$ has a similar factorization. These factorizations can be substituted into Eqn (6) and the KL-divergences can then be expanded. Finally, by using the identities in Eqns (9) and (10), we can group terms together to obtain the desired result. □

Note the close similarity between Eqns (4) and (12). In particular, the edge weights have been replaced by w_{st}. We can equivalently choose to maximize $\sum_{(s,t) \in \mathcal{E}_p \cup \mathcal{E}_q} w_{st}$. To do this, we use the same 'k-edge' MWST algorithm with edge weights given by w_{st}. In this case we must consider the maximum of the three possible values for w_{st}. Whichever is the maximum indicates one of three actions:

1. Place an edge between s, t for $\widehat{p}$ and **not** $\widehat{q}$ (corresponding to $(s,t) \in \mathcal{E}_p \setminus \mathcal{E}_{pq}$).
2. Place an edge between s, t for $\widehat{q}$ and **not** $\widehat{p}$ (corresponding to $(s,t) \in \mathcal{E}_q \setminus \mathcal{E}_{pq}$).
3. Place an edge between s, t for **both** $\widehat{p}$ and $\widehat{q}$ (corresponding to $(s,t) \in \mathcal{E}_{pq}$).

We now arrive at a direct analog of Lemma 3.1.

Lemma 4.2 *$\widehat{p}^{(k)}(x), \widehat{q}^{(k)}(x)$ with edge sets $\mathcal{E}_p^{(k)}$ and $\mathcal{E}_q^{(k)}$ can be optimally chosen via the 'k-edge' MWST algorithm with edge weights given by w_{st} in Eqn* (13)*.*

We would also expect the edge sets $\mathcal{E}^{(k)}$ to be nested just as in section 3, though in the discriminative case at each iteration the approximations for the two graphs $\mathcal{T}_p^{(k)}, \mathcal{T}_q^{(k)}$ may have a different number of edges. Additionally, the optimization may reach a point where it is advantageous to terminate at $k < n-1$ due to all of the remaining w_{st} being negative.

Corollary 4.3 *The edge sets $\mathcal{E}_p^{(k)}, \mathcal{E}_q^{(k)}$ obtained from the maximization of the J-divergence in Eqn (8) are nested i.e.*

$$\mathcal{E}_p^{(k)} \subset \mathcal{E}_p^{(k+1)}, \quad \forall\, k = 0, \ldots, n-2. \tag{15}$$

4.3. Bound on the J-divergence

The following result is evident from the definition of $\widehat{p}^{(k)}, \widehat{q}^{(k)}$.

Lemma 4.4 *Denoting $\widehat{p}, \widehat{q}$ the proper projections of p, q onto the graphical structure associated with $\widehat{p}, \widehat{q}$, the k-th forest approximation provides a lower bound for the full approximation $J(\widehat{p}, \widehat{q})$ i.e.*

$$J(\widehat{p}^{(k)}, \widehat{q}^{(k)}) \leq J(\widehat{p}, \widehat{q}). \tag{16}$$

In addition, if the source distributions p, q are tree distributions, then as $k \to n-1$, $J(\widehat{p}^{(k)}, \widehat{q}^{(k)}) \to J(p, q) = J(\widehat{p}, \widehat{q})$.

Hence, for tree-structured source distributions, we get increasingly accurate approximations to the actual J-divergence by using the optimal sequence of forests $\widehat{p}^{(k)}, \widehat{q}^{(k)}$ to approximate the source distributions p, q.

5. NUMERICAL EXPERIMENTS

In this section, we describe our numerical simulations that demonstrate the convergence of $J(\widehat{p}^{(k)}, \widehat{q}^{(k)})$ and the reduced $\Pr(\text{err})$. We only consider Gauss-Markov Random Fields (GMRFs) in this paper. GMRFs can be parameterized in the *information* form:

$$p(x) \propto \exp\left\{-\frac{1}{2}x'Jx + h'x\right\} \tag{17}$$

J is the information matrix (inverse covariance matrix) and its fill pattern [6] provides the Markov structure; x is Markov with respect to $\mathcal{G}$ if and only if $J_{s,t} = 0$ for all $(s,t) \notin \mathcal{E}$. Another useful quantity is the *conditional correlation coefficient* $\rho_{s,t}$ [6]. This is defined as the correlation coefficient of variables x_s and x_t conditioned on knowledge of all the other variables $x_{\mathcal{V} \setminus \{s,t\}}$ i.e.

$$\rho_{s,t} = \frac{\text{cov}(x_s, x_t | x_{\mathcal{V}\setminus\{s,t\}})}{\sqrt{\text{var}(x_s|x_{\mathcal{V}\setminus\{s,t\}})\text{var}(x_t|x_{\mathcal{V}\setminus\{s,t\}})}} = \frac{-J_{s,t}}{\sqrt{J_{s,s}J_{t,t}}}. \tag{18}$$

For our experiments, the source distributions p and q are described by two different sets of probability models.

1. *Grid*: p and q are both $n = 6 \times 6$ grid models with constant conditional correlation coefficients $\rho_p = 0.12$ and $\rho_q = -0.18$ respectively.
2. *Cycle*: p is a $n = 32$-node cycle graph with each edge connected to its 2 nearest neighbors and $\rho_p = 0.08$. q is another cycle graph of the same size with each edge connected to its 5 nearest neighbors and $\rho_q = -0.08$.

For both sets of models, the components of the mean vectors μ_p, μ_q are drawn independently from $\mathcal{N}(0,1)$. For clarity of exposition, we will denote the discriminatively and generatively learned models/forests as $(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})$ and $(\widehat{p}_G^{(k)}, \widehat{q}_G^{(k)})$ respectively.

5.1. Convergence of J-divergence

The structures of q under the grid and cycle models are shown in Fig 1 (top). The learned trees under the generative and discriminative approaches are also shown. For the cycle model, the discriminative model exploits the primary difference of longer range correlation mixed with shorter range correlations. The generative approach neglects this difference. It is harder make the same conclusion about the grid model other than the resulting tree structures are different. Finally, we note that some of the edge weights w_{st} may be identical so these structures may not be uniquely determined by w_{st}.

In Fig 2, we show the convergence of $J(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})/J(p,q)$ and $J(\widehat{p}_G^{(k)}, \widehat{q}_G^{(k)})/J(p,q)$ as a function of k for the grid and cycle models. Under both the grid and cycle models, the discriminative approach (in blue ○) provides a higher value of the J-divergence as compared to the generative approach (in red x).

Interestingly for the grid model, $J(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})$ is higher than $J(p,q)$ for some k as seen from Fig 2. This does not violate Lemma 4.4 because the lemma refers to the full approximate distributions $\widehat{p}, \widehat{q}$ and not the source distributions p, q. This observation shows that the *bounds* in Eqn (7) are lower. However, the $\Pr(\text{err})$ under the approximate models is obviously higher. Consequently, we conjecture that $J(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})$ is giving us an improved bound on the $\Pr(\text{err})$ when using the full model as well, since it must be contained in these bounds.

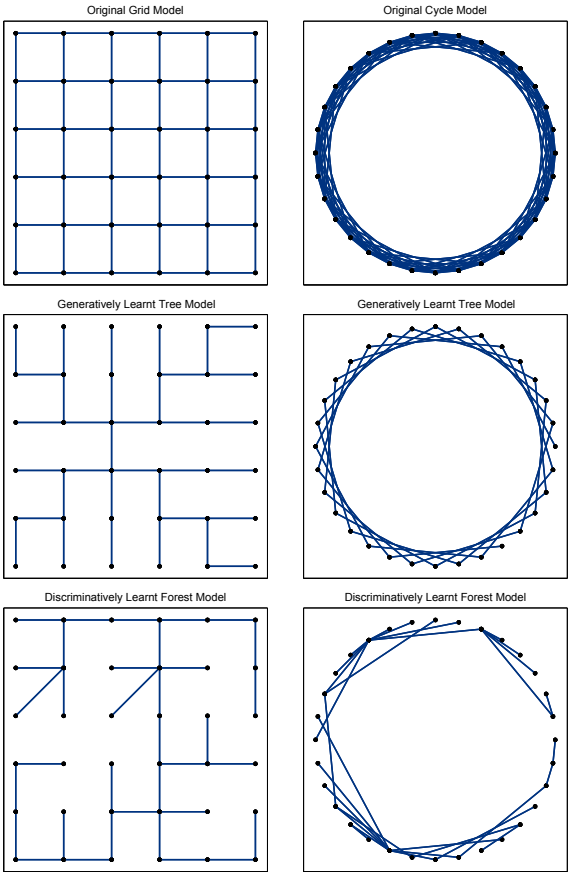

Fig. 1. Top plots: Original structures of $q(x)$. Middle plots: The tree approximation $\widehat{q}_G^{(n-1)}(x)$ using the *generative* learning model. Bottom plots: The $k = n-1$ forest approximation $\widehat{q}_D^{(n-1)}(x)$ using the *discriminative* model.

5.2. Reduced Probability of Error $\Pr(\text{err})$

Ultimately, $\Pr(\text{err})$ is to be minimized. To show that the $\Pr(\text{err})$ is indeed reduced, we generated 10000 new samples from the source distributions. In Fig. 3, we report the average $\Pr(\text{err})$ (over 2000 independent runs) on the source (p, q), the discriminative $(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})$ and the generative $(\widehat{p}_G^{(k)}, \widehat{q}_G^{(k)})$ distributions for all $k \leq n-1$. Under both the grid and cycle models, the discriminative learning approach (in blue ○) results in a lower $\Pr(\text{err})$ as compared to the generative learning approach (in red x).

6. CONCLUSION

We have a constructive method which optimizes an information measure (the J-divergence) which itself can be used to compute both upper and lower bounds on the $\Pr(\text{err})$. The 'k-edge' MWST algorithm provides a principled sequential algorithm to select edges of graphical models for the specific purpose of hypothesis testing. At each iteration, the nested forests are provably optimal with respect to the J-divergence. We have numerically verified the results on two sets of simple graphical models. Our experiments show that, as compared to the generative (Chow-Liu) approach, the *discriminative* learning approach results in higher J-divergences and hence better bounds on the $\Pr(\text{err})$. Most importantly, it is observed that the actual $\Pr(\text{err})$ is reduced for all the discriminative forest models learned.

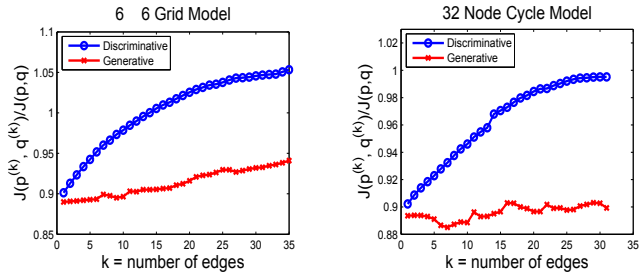

Fig. 2. $J(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})/J(p,q)$ and $J(\widehat{p}_G^{(k)}, \widehat{q}_G^{(k)})/J(p,q)$ for the grid and cycle models. Generative learning always results in a lower $J(\widehat{p}^{(k)}, \widehat{q}^{(k)})$ as compared to discriminative learning.

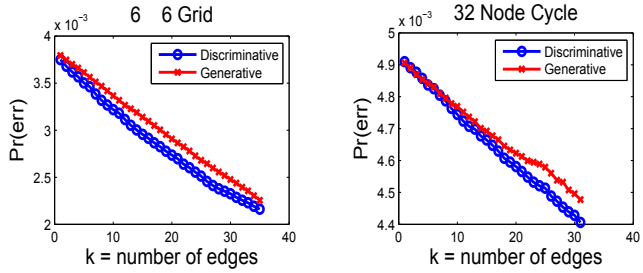

Fig. 3. $\Pr(\text{err})$'s for the grid and cycle models. At each k, learning the probability models $(\widehat{p}_D^{(k)}, \widehat{q}_D^{(k)})$ discriminatively reduces the $\Pr(\text{err})$. The $\Pr(\text{err})$ under the source distributions p, q for the grid and cycle models are 1.6×10^{-3} and 3.6×10^{-3} respectively.

7. REFERENCES

[1] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees.," *IEEE Trans. on Info. Th.*, vol. 14, no. 3, pp. 462–7, May 1968.

[2] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, New York, 1959.

[3] F. R. Bach and M. I. Jordan, "Thin Junction Trees," in *Neural Information Processing Systems*, 2002.

[4] S. Sanghavi, V. Y. F. Tan, and A. S. Willsky, "Learning graphical models for hypothesis testing," in *IEEE Statistical Signal Processing Workshop*, 2007.

[5] J. B. Kruskal, "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.

[6] S. L. Lauritzen, *Graphical Models*, Oxford Statistical Science Series, 1996.

[7] Basseville M., "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349 – 369, 1989.