Beyond Maximum Likelihood: Boosting the Chow–Liu Algorithm for Large Alphabets

Jiantao Jiao
EE Department, Stanford University
jiantao@stanford.edu

Yanjun Han
EE Department, Stanford University
yjhan@stanford.edu

Tsachy Weissman
EE Department, Stanford University
tsachy@stanford.edu

***Abstract*—We show that in high dimensional distributions, i.e., the regime where the alphabet size of each node is comparable to the number of observations, the Chow–Liu algorithm on learning graphical models is highly sub-optimal. We propose a new approach, where the key ingredient is to replace the empirical mutual information in the Chow–Liu algorithm with a minimax rate-optimal estimator proposed recently by Jiao, Venkat, Han, and Weissman [1]. We demonstrate the improved performance of the new approach in two problems: learning tree graphical models and Bayesian network classification.**

***Index Terms*—Chow–Liu algorithm, mutual information estimation, approximation theory, high dimensional statistics, nonsmooth functional estimation**

I. Introduction

Graphical models provide us with efficient computational tools to conduct inference in high dimensional data with potential structure, cf. [2] and references therein. Learning the structure and parameters of graphical models from empirical data is therefore the starting point for all these applications. It has been known that exact learning of a general graphical model is NP-hard [3], and there exist tractable sub-classes among which tree graphical models are the most famous. The seminal work of Chow and Liu [4] contributed an efficient algorithm to compute the Maximum Likelihood Estimator (MLE) of tree structured graphical model based on empirical data, and constitutes one of the very few cases where the exact MLE can be solved efficiently. There are various approaches towards learning more complex structures, for which we refer the reader to [5] for a review.

Concretely, the Chow–Liu algorithm (CL) addresses the following question. Given n i.i.d. samples of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$, where $X_i \in \mathcal{X}, |\mathcal{X}| < \infty$, we want to estimate the joint distribution of $\mathbf{X}$. Chow and Liu [4] assumed that $P_{\mathbf{X}}$ can be factorized as:

$$P_{\mathbf{X}} = \prod_{i=1}^{d} P_{X_{m_i}|X_{m_{j(i)}}}, \quad 0 \le j(i) < i, \tag{1}$$

where $(m_1, m_2, \ldots, m_d)$ is an unknown permutation of integers $1, 2, \ldots, d$, and $P_{X_i|X_0}$ is by definition equal to P_{X_i}. Then, CL outputs the distribution $P_{\mathbf{X}}$ that maximizes the likelihood of the observed data.

Proposed in 1968, The Chow–Liu algorithm is widely used in machine learning and statistics as a tool for dimensionality reduction, classification, and as a foundation for algorithm design in more complex dependence structures [5] in the theory of learning graphical models [2], [6]. It has also been widely adopted in applied research, and is particularly popular in systems biology. For example, the Chow–Liu algorithm is extensively used in the reverse engineering of transcription regulatory networks from gene expression data [7].

We begin by asking the following natural question:

Question. *Is the Chow–Liu algorithm optimal for learning tree graphical models?*

Since the Chow–Liu algorithm exactly solves the MLE, and has been widely used in many applications, its optimality seems to be tacitly assumed in much of the literature. However, a closer inspection of the statistical theory [8], [9] reveals that it is only known that the Chow–Liu algorithm performs essentially optimally when the number of samples n grows to ∞, while the number of states of the tree has fixed size. Indeed, the modern theory of the maximum likelihood estimation paradigm [10] only justifies the *asymptotic efficiency* of MLE, without general non-asymptotic guarantees when we have finitely many samples. In contrast, various modern data-analytic applications deal with datasets that do not have the luxury of too many observations compared to the alphabet size.

The main contribution of the present paper is the introduction of a new algorithm that provably significantly improves upon the CL algorithm when the alphabet size is comparable to the number of observations. The key ingredient in our improved algorithm is to replace the empirical mutual information employed in the Chow–Liu algorithm with the minimax rate-optimal estimator for mutual information proposed in [1]. In a nutshell, the performance of the new algorithm with n samples is essentially that of the original algorithms with $n \ln n$ samples.

We remark that we do not think nor try to imply that the schemes we present here are necessarily competitive with the state of the art for the applications we experimented with. Our point rather is that machine learning schemes which have a mutual information estimation component stand to benefit from significant performance boosts via use of improved near optimal estimators, such as those recently discovered by [1]. This is particularly true in the large-alphabet regimes where use of the latter estimators in lieu of the standard ones such as empirical mutual information can spell the difference between

consistency and complete divergence.

II. Improving the Chow–Liu algorithm

Chow and Liu [4] considered solving for the MLE under the constraint that the joint distribution factors as a tree. Interestingly, this optimization problem can be efficiently solved after being transformed into a Maximum Weight Spanning Tree (MWST) problem. In particular, they showed that the MLE of the tree structure boils down to the following expression:

$$E_{\mathrm{ML}} = \arg \max_{E_Q : Q \text{ is a tree}} \sum_{e \in E_Q} I(\hat{P}_e), \tag{2}$$

where $I(\hat{P}_e)$ is the mutual information associated with the empirical distribution of the two nodes connected via edge e, and E_Q is the set of edges of a tree distribution Q (i.e., Q factors as a tree). In words, it suffices to first compute the empirical mutual information between any two nodes (in total $\binom{d}{2}$ pairs), and the maximum weight spanning tree is the tree structure that maximizes the likelihood. To obtain estimates of distributions on each edge, Chow and Liu [4] simply assigned the empirical distribution.

To explain the insights underlying our improved algorithm, we revisit equation (2) and note that if we were to replace the empirical mutual information with the true mutual information, the output of the MWST would be the true edges of the tree. In light of this, the CL algorithm can be viewed as a "plug-in" estimator that replaces the true mutual information with an estimate of it, namely the empirical mutual information. Naturally then, it is to be expected that a better estimate of the mutual information would lead to smaller probability of error in identifying the tree. However, how bad can the empirical mutual information be as an estimate for the true mutual information? The following theorem in [1] implies that it can be highly sub-optimal in high dimensional regimes.

Theorem 1. *Suppose we have two random variables $X_1, X_2 \in \mathcal{X}, |\mathcal{X}| < \infty$. The minimax sample complexity in estimating the mutual information $I(X_1; X_2)$ under mean squared error is $\Theta(|\mathcal{X}|^2 / \ln |\mathcal{X}|)$, while the sample complexity required by the empirical mutual information to be consistent is $\Theta(|\mathcal{X}|^2)$.*

In words, Theorem 1 implies that for the minimax rate-optimal estimator, it suffices to take $n \gg |\mathcal{X}|^2 / \ln |\mathcal{X}|$ samples to consistently estimate the mutual information $I(X_1; X_2)$ for any underlying distributions. At the same time, unless $n \gg |\mathcal{X}|^2$, there exist distributions for which the error of the *empirical* mutual information would be bounded away from zero.

Theorem 1 sheds light on a possible improvement over CL. How can we construct computationally efficient mutual information estimators that require only $|\mathcal{X}|^2 / \ln |\mathcal{X}|$ samples? Here we take a slight detour and review recent work in entropy and mutual information estimation, which assists us in the improving the classical CL approach.

A. Recent advances in functional estimation

To simplify the notation, we use S and $|\mathcal{X}|$ interchangeably to denote the alphabet size of a discrete distribution.

Recently, [1] proposed a general methodology of constructing minimax estimators for functionals, and showed that the MLE is generally far from minimax optimality[1] [11]. In particular, [1] showed that this methodology achieves the minimax rates for estimating the Shannon entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$, mutual information, as well as the functional $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha$, for any $\alpha > 0$. In particular, an interesting observation therein is that the performance under L_2 risk of the optimal estimators with n samples is essentially that of the MLE with $n \ln n$ samples. It has been shown that this *effective sample size enlargement* phenomena generalize to more functionals [12]–[15]. On the practical side, the estimators in [1] have complexity linear in sample size n.

Specifically, for the estimation of the Shannon entropy, Valiant and Valiant [16], [17] were the first to show that it is necessary and sufficient to take $n = \Theta(S/\ln S)$ samples. More recently, [1] and [18] developed schemes based on approximation theory that achieve the optimal rates of convergence for the entropy. In contrast to all these schemes which require $n = \Theta(S/\ln S)$ samples, the MLE requires $n = \Theta(S)$ samples [11], [19]. It was recently shown in [20] that the seemingly natural approach of first using Dirichlet prior to obtain a smoothed distribution, and then plugging-in the entropy functional also requires $n = \Theta(S)$ samples. A comprehensive review of this problem can be found in [1].

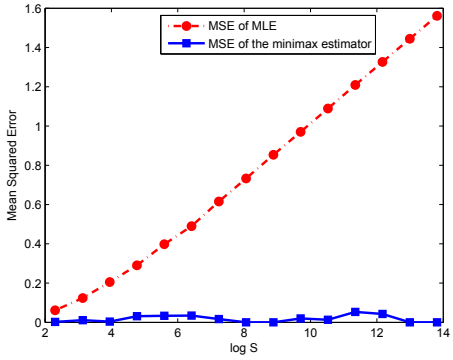

Fig. 1. The empirical MSE of the estimator in [1] and the MLE along sequence $n = 5S/\ln S$, where S is sampled equally spaced logarithmically from 10 to 10^6. The horizontal line is $\ln S$, and the vertical line is the MSE obtained using 20 Monte Carlo simulations from sampling a uniform distribution supported on S elements.

Figure 1 compares the performance of the essentially minimax optimal estimator in [1] and the MLE, and shows that this improvement can in fact be significant in practice. Intriguing as these findings are theoretically, they are valuable also

[1] An estimator is called minimax optimal if its maximum risk (e.g. expected L_2 error) is the minimum among all possible estimators. We refer the readers to [10] for more details.

to the practitioner encountering problems beyond functional estimation, as we illustrate next.

B. New algorithm for learning tree graphical models

The mutual information estimator in [1] can be shown to achieve the minimax sample complexity shown in Theorem 1. It is thus natural to suspect that using the latter in lieu of the empirical mutual information in the CL algorithm would lead to performance boosts. It is gratifying to find this intuition confirmed in all the experiments that we conducted. In the following experiment, we fix $d = 7, |\mathcal{X}| = 300$, construct a star tree (i.e. all random variables are conditionally independent given X_1), and generate a random joint distribution by assigning independent Beta$(1/2, 1/2)$-distributed random variables to each entry of the marginal distribution P_{X_1} and the transition probabilities $P_{X_k|X_1}, 2 \leq k \leq d$ (with normalization). Then, we increase the sample size n from 10^3 to 5.5×10^4, and for each n we conduct 20 Monte Carlo simulations.

Note that the true tree has $d - 1 = 6$ edges, and any estimated set of edges will have at least one overlap with these 6 edges because the true tree is a star graph. We define the wrong-edges-ratio in this case as the number of edges different from the true set of edges divided by $d - 2 = 5$. Thus, if the wrong-edges-ratio equals one, it means that the estimated tree is maximally different from the true tree and, in the other extreme, a ratio of zero corresponds to perfect reconstruction. We compute the expected wrong-edges-ratio over 20 Monte Carlo simulations for each n, and the results are exhibited in Figure 2.

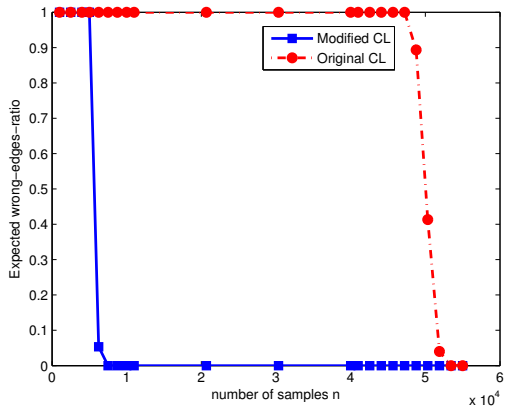

Fig. 2. The expected wrong-edges-ratio of our modifed algorithm and the original CL algorithm for sample sizes ranging from 10^3 to 5.5×10^4.

Figure 2 reveals intriguing phase transitions for both the modified and the original CL algorithm. When we have fewer than 3×10^3 samples, both algorithms yield a wrong-edges-ratio of 1, but soon after the sample size exceeds 6×10^3, the modified CL algorithm begins to reconstruct the network perfectly, while the original CL algorithm continues to fail maximally until the sample size exceeds 47×10^3, 8 times the sample size required by the new algorithm, which we temporarily call "Modified Chow–Liu" algorithm. The theoretical properties of these sharp phase transitions will be considered in future work.

III. Application: Bayesian Network Classifiers

Given n training samples, each of which has d attributes $\mathbf{X} = (X_1, X_2, \cdots, X_d), X_i \in \mathcal{X}_i$ and a class label $C \in \mathcal{C}$, we are interested in constructing a classifier to assign a class label to a test instance characterized by its attributes. One important class of classifiers is the Bayes classifier [21], which learns from training data the conditional joint distribution of $\mathbf{X}$ given the class label C. Then classification is done by applying the Bayes rule to compute the posterior probability of each class given the attribute vector. Many classifiers fall into this class under various assumptions. For example, the Naive Bayes classifier assumes that all the attributes are conditionally independent given the class label. This assumption was further relaxed by Friedman et al. [22] that $\mathbf{X}$ satisfies the order-one dependence conditioning on the class label, i.e., the joint probability of $\mathbf{X}$ given C can be factorized into the product of the probabilities of each attribute conditioning on another attribute and the class label. To be precise, $P_{\mathbf{X}|C}$ can be factorized as

$$P_{\mathbf{X}|C} = \prod_{i=1}^{d} P_{X_{m_i}|X_{m_{j(i)}},C}, \quad 0 \leq j(i) < i, \tag{3}$$

where, as in the preceding section, $(m_1, \ldots, m_d)$ is an unknown permutation of the integers $1, 2, \ldots, d$.

A. TAN classifier and CL classifier

In light of the CL algorithm, Friedman et al. [22] proposed the tree-augmented naive Bayes (TAN) classifier. To construct the TAN classifier, the tree graphical model is established first using the CL algorithm, with a slight difference that the empirical mutual information $I(\hat{P}_e)$ in (2) is replaced by the conditional empirical mutual information $I(\hat{P}_e|C)$. Once the tree graphical model has been obtained, the empirical distributions $\hat{P}_C$ and $\hat{P}_{X_i|X_{\pi(i)},C}$ are used to estimate P_C and $P_{\mathbf{X}|C}$, respectively, and both are substituted into

$$f(\mathbf{x}) \triangleq \arg\max_{c \in \mathcal{C}} P_C(c) P_{\mathbf{X}|C}(\mathbf{x}|c), \tag{4}$$

which is the maximum a posteriori (MAP) estimator of the class label given attribute vector $\mathbf{x}$ using the Bayes rule.

Since we have demonstrated in the preceding section that the CL algorithm based on MLE is far from optimal, it is reasonable that we can harvest a performance gain in classification problems by simply using our better estimate of the mutual information for learning the tree graphical model. Specifically, we estimate the conditional mutual information via

$$\begin{aligned} I(\hat{P}_{X_i X_j}|C) = \hat{H}(X_i, C) + \hat{H}(X_j, C) \\ - \hat{H}(C) - \hat{H}(X_i, X_j, C), \end{aligned} \tag{5}$$

where $\hat{H}$ is the entropy estimator in [1]. In this way, we construct a modified TAN classifier, and we remark that this construction does not impose an increased implementation burden since the computational complexity of the improved estimator $\hat{H}$ is linear in the number of observations [1]. If the mutual information is conditioned on a concrete realization of the class label, i.e.,

$$I(\hat{P}_{X_iX_j}|C=c) = \hat{H}(X_i|C=c) + \hat{H}(X_j|C=c) - \hat{H}(X_i, X_j|C=c), \tag{6}$$

the TAN classifier is turned into the Chow-Liu (CL) classifier [22]. In other words, in the construction of the CL classifier, different tree structures under different class labels are allowed.

In our real experiments, we also used the smoothing idea. Since we may be in an undersampling position for a robust estimation of the conditional probability $P_{\mathbf{X}|C}$, the performance of the original classifiers can be further improved by the introduction of an additional smoothing operation [22]. The identical smoothing method is also adopted in the modified classifiers which updates the conditional probability using some intuitive priors, i.e.,

$$p_s(A|B) = \frac{N_0 \cdot \hat{p}(A) + n \cdot \hat{p}(A \cap B)}{N_0 + n \cdot \hat{p}(B)}, \tag{7}$$

for any events A, B, where $\hat{p}(\cdot)$ denotes the empirical probability, and $N_0 \geq 0$ is the smoothing parameter.

B. Experiments and Results

Now we evaluate the performance gain of our modified classifiers in terms of the classification error via experimentation on a total of 26 datasets. All of the datasets are popular datasets from the UCI repository [23], and the first 25 datasets are identical to those used by Friedman et al. [22] for comparison. The last dataset *pendigits* is selected due to its property that its attribute alphabet size $\max_{1\leq i\leq d} |\mathcal{X}_i|$ is large. We refer to the full version [24] on the detailed description of the datasets and our preprocessing techniques.

First, we implement the 5-fold random cross validation repeatedly on all datasets for 100 times, and in each cross validation, the classification errors of naive Bayes, original and modified TAN, and original and modified CL classifiers are recorded separately. The full table consisting of all classification errors is referred to [24].

Figure 3 shows intriguing properties of the modified TAN classifier relative to the original one, where our modified TAN classifier uniformly outperforms the original one in terms of classification errors. Furthermore, a closer inspection reveals that the top eight datasets with largest classification error reduction all share a common feature that the squared maximum alphabet size is comparable to the number of observations, i.e., $S^2 \cong n$, where $S = \max_{1\leq i\leq d} |\mathcal{X}_i|$. It is consistent with Theorem 1, which indicates that the minimax rate-optimal estimators would perform significantly better than the empirical mutual information in the regime $S^2 \cong n$.

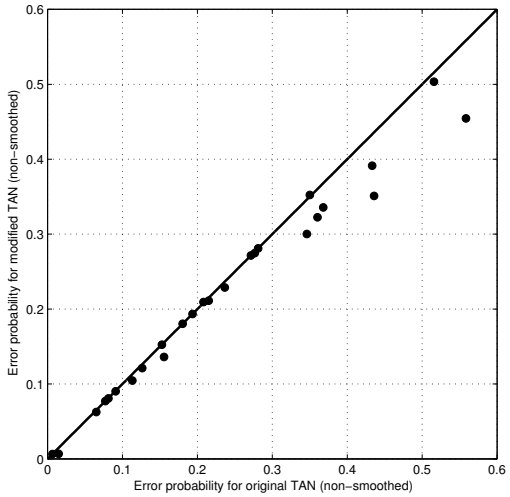

Fig. 3. The scatter plot comparing non-smoothed original TAN classifier (x-axis) with non-smoothed modified TAN classifier (y-axis). In this scatter plot, points above the diagonal line corresponds to datasets where original TAN classifier performs better and points below the diagonal line corresponds to datasets where modified TAN classifier performs better.

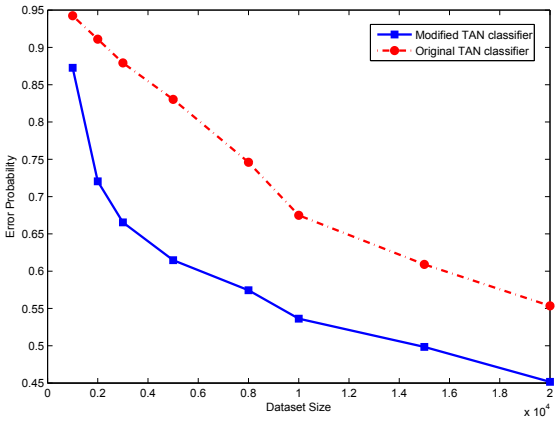

Fig. 4. The error probability decay curve comparing non-smoothed original TAN classifier (dashed line) with non-smoothed modified TAN classifier (solid line) in random subsets of the dataset *letter*. The x-coordinates of all squares (and circles) correspond to the subset size of 1000, 2000, 3000, 5000, 8000, 10000, 15000 and 20000, respectively.

Second, to further convey the point that the modified classifiers require fewer samples to achieve an acceptable classification error, we conducted another experiment to compare the error probability decay curves under different classifiers. Specifically, sample sizes from 1000 to 20000 are selected, and for each sample size n, the preceding classification experiment on a training sample of size $4n/5$ and a testing sample of size $n/5$ is implemented 20 times, where in each time the training sample is a subset randomly generated from the training data in dataset *letter*. Figure 4 displays the relationship between the average classification errors and training sample sizes.

Figure 4 exhibits a remarkable error reduction over the

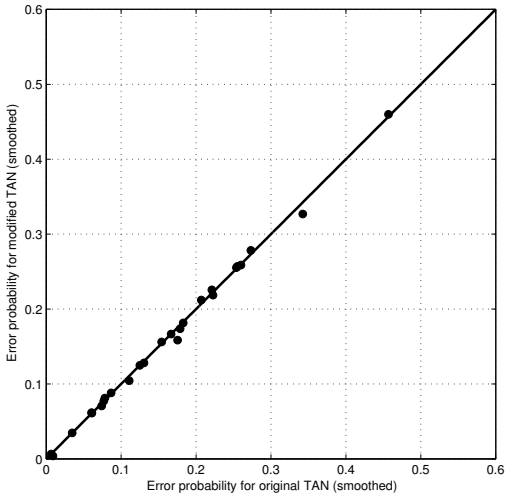

Fig. 5. The scatter plot comparing smoothed original TAN classifier (x-axis) with smoothed modified TAN classifier (y-axis). In this scatter plot, points above the diagonal line corresponds to datasets where original TAN classifier performs better and points below the diagonal line corresponds to datasets where modified TAN classifier performs better.

original scheme, uniformly over all sample sizes. For example, to achieve probability of error 0.7, the sample size required by the modified TAN classifier is about 2000, while that for the original one is about 10000. We remark that a random guess in this dataset would result in at least classification error 95%.

Now we come to the comparison in the smoothed scheme. Fig. 5 illustrates the comparison of the original and the modified TAN classifiers in the smoothed scheme, and unlike the non-smoothed scheme, the gap between two classifiers almost vanishes, i.e., most scatter points lie very close to the diagonal line. A closer inspection of the experimental details reveals that the output of the classifier does not differ much even though the tree structures generated by two classifiers differ considerably. Therefore, the involvement of the smoothing priors makes the classification problem insensitive to the tree structure. However, despite the similarity of these two classifiers, it can still be observed that for those datasets with $S^2 \cong n$, e.g., *letter* and *glass2*, the classification error reduction is still significant.

We have also compared the original CL classifier with our modified CL classifier in both the non-smoothed and smoothed schemes, and the results form a similar pattern as the TAN classifiers. In summary, we observe that

1) no matter how well the tree structure fits the data, replacing the empirical mutual information with the mutual information estimator of [1] results in uniformly better classification accuracy;
2) our modified TAN and CL classifiers require fewer training data to establish the correct dependence tree with the same classification error;
3) the uniformly better performance still holds when adding the smoothing operation, though this operation significantly reduces the sensitivity of the classification results to the selection of the tree.

REFERENCES

[1] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[2] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[3] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2001, pp. 392–401.

[4] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *Information Theory, IEEE Transactions on*, vol. 14, no. 3, pp. 462–467, 1968.

[5] Y. Zhou, "Structure learning of probabilistic graphical models: a comprehensive survey," *arXiv preprint arXiv:1111.6925*, 2011.

[6] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[7] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP journal on bioinformatics and systems biology*, vol. 2007, 2007.

[8] C. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions (corresp.)," *Information Theory, IEEE Transactions on*, vol. 19, no. 3, pp. 369–371, 1973.

[9] V. Y. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A large-deviation analysis of the maximum-likelihood learning of Markov tree structures," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1714–1735, 2011.

[10] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998, vol. 31.

[11] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of information measures," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 839–843.

[12] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating rényi entropy," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2015, pp. 1855–1869.

[13] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of KL divergence between discrete distributions," *arXiv preprint arXiv:1605.09124*, 2016.

[14] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence between large-alphabet distributions," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1118–1122.

[15] J. Jiao, Y. Han, and T. Weissman, "Minimax estimation of the L_1 distance," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 750–754.

[16] G. Valiant and P. Valiant, "Estimating the unseen: an $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the 43rd annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.

[17] ——, "The power of linear estimators," in *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*. IEEE, 2011, pp. 403–412.

[18] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[19] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[20] Y. Han, J. Jiao, and T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1367–1371.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[22] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[23] P. Murphy and D. W. Aha, "UCI repository of machine learning databases–a machine-readable repository," 1995.

[24] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.