Minimax Prediction in Tree Ising Models

Guy Bresler
Massachusetts Institute of Technology
Cambridge, Massachusetts
Email: guy@mit.edu

Mina Karzand
University of Wisconsin – Madison
Madison, Wisconsin
Email: karzand@wisc.edu

***Abstract*—Graphical models are often used to facilitate efficient computation of posteriors in order to make predictions. With this objective in mind, we consider the problem of estimating the parameters of a graphical model with known structure from samples such that posteriors computed using the model are accurate. Focusing on tree-structured binary Markov random fields, our main result is a sharp characterization of the dependence on number of samples needed for all pairwise marginals (and hence posteriors of one variable given another) to be accurate: $n = \Theta(\eta^{-2} \log p)$ samples are necessary and sufficient to estimate model parameters such that all marginals of arbitrary order k are accurate to within $\sqrt{k}\,\eta$. The result implies that prediction error is bounded uniformly, with no dependence on the strength of interactions. We will also show that these guarantees are achievable using moment matching techniques.**

I. Introduction

Graphical models are ubiquitous in a wide variety of application domains including signal processing, computational biology, finance, and natural language processing [13], [24]. As the name suggests, these models use an underlying graph to represent structure in high-dimensional data: random variables are associated with nodes, and edges indicate interactions between the variables. For a given application, domain knowledge is often used to select the model *structure*, for instance one might use a Hidden Markov Model for time-series; the model *parameters* are then estimated from data.

From the statistical point of view, the fundamental objective is to determine how much data is needed to accurately estimate the model. There are two natural measures of accuracy: one measures closeness of estimated parameters to those of the model generating the data, for instance in squared error; the second measures accuracy of *predictions* computed using the model. Since in the typical machine learning application graphical models are used to make predictions, and moreover the first objective is by now largely understood when the graph structure is known, we focus on prediction.*

A graphical model allows to simultaneously make many predictions. One might wish to predict whether (or compute the probability that) a user in a recommendation system will buy each of a collection of items given feedback previously obtained from the user, and also to do this for many different users each of whom has rated a different set of items. Accuracy of the model thus entails accuracy of posteriors $P(X_i|X_S)$ for many choices of nodes i and sets S. This paper is motivated by the following basic question.

Motivating Question: How many samples generated from a graphical model with known structure are needed in order to estimate the model parameters θ such that all posteriors $P_\theta(X_i|X_S)$ computed using the estimated model are accurate, for sets S of size at most k?

As observed in [3], and can be seen by writing out the conditional probability, accuracy of posteriors is captured by the following *local* total variation, which we take as the loss function in this paper. For any two distributions P and Q and integer $k \geq 2$, let

$$\mathcal{L}^{(k)}(P,Q) = \max_{S:|S|\leq k} d_{\mathsf{TV}}(P_S, Q_S)\,. \tag{I.1}$$

Here P_S denotes the marginal of P on the subset of variables S (and similarly for Q_S). This loss function was used by [3] in the same context and previously by Rebeschini and Handel [17]. To connect this notion of distance to the prediction based on partial observations, note that given two distributions $P(X)$ and $Q(X)$ on $X \in \{-,+\}^p$, if $\mathcal{L}^{(k)}(P,Q) \leq \eta$, then for any $\mathcal{S} \subseteq \mathcal{V}$, such that $|\mathcal{S}| \leq k-1$ and any $i \in \mathcal{V}$ and $x \in \{-,+\}$, we have

$$\sum_{x_\mathcal{S}} P(x_\mathcal{S}) \big| P_{i|\mathcal{S}}(x|x_\mathcal{S}) - Q_{i|\mathcal{S}}(x|x_\mathcal{S}) \big| \leq 2\eta\,,$$

as proved in [3].

The main challenge in controlling this loss is that that the total variation must be bounded uniformly over all marginals, and moreover, depending on k and the number of edges in the model, there are far too few parameters in order to separately minimize the total variation for each of the marginals. What this means is that inherently one must trade-off accuracy of some marginals against others.

In this paper we give an essentially complete answer to the question posed above for the special case of Ising models (binary pairwise graphical models) with no external field on trees, with no restriction on the node and edge parameters. A probability distribution $P_X(x)$, denoted by $P(x)$ on $x \in \{-,+\}^p$ represented by the tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ in this class of models is represented as

$$P(x) = \exp\Big(\sum_{(i,j)\in\mathcal{E}} \theta_{i,j} x_i x_j - \Phi(\theta) \Big) \tag{I.2}$$

*The story for linear regression is parallel, with a large body of work studying both parameter estimation and prediction error [18].

where $\Phi(\theta)$ is a normalizing constant. The tree T determines the factorization of the distribution and the edge parameters $\theta_{i,j}$ quantify the pairwise marginal distributions on each edge.

What our results show is that for this class of models there is essentially no trade-off and it is possible to simultaneously guarantee accuracy for all marginals by optimizing the errors only across the edges. This is surprising because errors can accumulate and combine along paths in the tree, but this turns out not to matter.

The intuition for why accumulation of errors along paths does not result in worse error is that accumulation of errors can occur over large distances only if the interactions across the edges are quite strong, but this in turn reduces the variance in estimating the edge parameter and this compensates for the accumulation.

We show that given $n > C\eta^{-2}\log p/\delta$ samples with a universal constant C, for all $k \geq 2$, we have $\mathcal{L}^{(k)}(P,Q) \leq \sqrt{k}\,\eta$ where the distribution Q is the derived using moment matching across the edges of the given tree. Moreover Theorem IV.3 shows that this bound is information theoretically tight: There exists a constant c such that for any value of $k \geq 2$, if $n < c\eta^{-2}\max\{k, \log p\}$, then $\mathcal{L}^{(k)}(P,Q) > \eta$.

This paper gives *the first sharp guarantees for estimation in graphical models with a prediction-centric loss*. We view tree models as an important first step in this broader direction. Tree graphical models are important in a range of applications, in part because marginals can be efficiently computed exactly using belief propagation. Thus when actually carrying out posterior computations in a learned tree model, the error is entirely due to parameter estimation from data, and none of it from the need to approximate the computational task [23].

A. Related work

There is a large literature on structure learning of graphical models, where the goal is to infer the graph structure underlying the model given samples. There has been a great deal of recent progress in learning non-tree bounded-degree graphical models, including [2], [9], [12], [16], [22]. The seminal paper by Chow and Liu [5] gives a computationally efficient algorithm to find the maximum likelihood tree and the parameters over the tree, given samples from a tree structured graphical model. They algorithm guarantees to minimize the probability of error in recovering the underlying structure. Tan et. al [20], [21] generalized this algorithm for forest approximation purposes. They also analyzed the asymptotic number of samples required to guarantee correct recovery of the underlying tree.

Instead of focusing on correct structure recovery, several papers have considered various notions of approximation. Heinemann and Globerson [10] studied high-girth graphs with correlation decay to show that the loopy belief propagation guarantees accurate marginals. Narasimhan and Bilmes [15] proposed an algorithm to learn bounded tree-width graphical models with respect to KL-divergence. In this work, the error in computing a posterior is due to estimation of model parameters from noisy data. Other papers have studied errors arising from noisy *computation* in message passing algorithms on loopy graphs, e.g. [25].

Bresler and Karzand [3] showed that accurate estimation of all pairwise marginals does not require correct structure recovery. They analyzed the performance of the Chow-Liu algorithm and showed that even in the regime in which correct recovery of the structure is not possible, using a possibly incorrect tree (the Chow-Liu tree) provides accurate pairwise marginals. The sufficient number of samples for this purpose is a function of the maximum strength of interaction between variables. In this work, we show that given the correct tree, the sample complexity of learning a distribution which accurately estimates the pairwise marginals is independent of the interaction strengths.

Theorem 9 by Babichenko et. al. [1] is equivalent to bounding $\mathcal{L}^{(2)}$ loss in our setup in the specific case $P_i(x_i) = 1/2$ for all i. Their analysis claims that number of samples required to guarantee $\mathcal{L}^{(2)}(P,Q) < \eta$ with probability $1/2$ is at least $C/\eta^2\log(p/\eta)$. Theorem IV.1 in this paper shows a strict improvement over this number.

There is a nascent literature on *testing* graphical models, such as [4], [7], [8]. In goodness-of-fit testing, the goal is to determine whether or not samples are from a given model P; in equality testing, one attempts to determine whether samples are both from the same distribution P or from two different distributions P and Q. These testing questions are distinct from (but related to) estimation questions, and moreover the underlying metric is typically total variation or Kullback-Leibler divergence over the entire joint distribution.

II. Model and notation

We use the notation $P_{i,j}(x_i, x_j)$ in place of $P_{X_i,X_j}(x_i, x_j)$ and similarly $P_{i|j}(x_i \,|\, x_j)$ for $P_{X_i|X_j}(x_i \,|\, x_j)$. We denote by $\mathcal{P}_{\mathsf{T}}$ the set of all distributions factorizing according to tree T.

Given a tree T and n i.i.d. samples $X^{(1:n)} = (X^{(1)}, \ldots, X^{(n)}) \in \{-1,+1\}^{|\mathcal{V}|}$ generated according to $P \in \mathcal{P}_{\mathsf{T}}$ (i.e., P factorizes according to T), an estimator $\psi : X^{(1:n)} \to \mathcal{P}_{\mathsf{T}}$ returns a distribution factorized according to T.

a) Estimation.: The maximum likelihood estimator is a natural estimator and the one we will analyze in this paper. It was observed by Chow and Liu in their 1968 paper [5] that the maximum likelihood estimator gives a distribution Q matching the marginals *on the edges* to those of the empirical distribution $\widehat{P}$, i.e., for all $(i,j) \in \mathcal{E}$, and all $x_i, x_j \in \mathcal{X}$, $Q(x_i, x_j) = \widehat{P}(x_i, x_j)$. The full distribution Q is then obtained via the tree factorization (I.2). We use the notation $Q = \Pi_{\mathsf{T}}(\widehat{P})$ to denote the maximum likelihood estimator just described, which is equivalent to the reverse information projection of $\widehat{P}$ onto $\mathcal{P}_{\mathsf{T}}$ [6], to emphasize the role of the tree T.

III. Model and Notation

We use notation similar to the ones used in [3] For a given tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ let $\mathcal{P}$ be the set of Ising models (I.2) We denote by $\mu_{i,j} = \mathbb{E}_P X_i X_j$ the correlation between the variables corresponding to any pair of vertices $i, j \in \mathcal{V}$. For

an edge $e = (i, j)$ we write $\mu_e = \mu_{i,j}$ and similarly for a set of edges $\mathcal{A} \subseteq \mathcal{E}$, $\mu_{\mathcal{A}} = \prod_{e \in \mathcal{A}} \mu_e$. The empirical distribution of samples is denoted by $\widehat{P}(x) = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{X^{(l)} = x\}}$.

IV. Main Result

For distribution P on $\{-,+\}^p$ and tree T, we define $\Pi_{\mathsf{T}}(P) = \arg\min_{R \in \mathcal{P}_{\mathsf{T}}} D(P\|R)$ to be the reverse information projection of P onto the class of Ising models on T with no external field (where $D(P\|R)$ is the KL divergence between distributions P and R). Lemma 11.1 in [3] shows that $Q = \Pi_{\mathsf{T}}(P)$ has edge weights satisfying $\mathbb{E}_Q X_i X_j = \mathbb{E}_P X_i X_j$ for all edges $(i, j) \in \mathcal{E}$.

From here on, we define

$$Q = \Pi_{\mathsf{T}}(\widehat{P}) \tag{IV.1}$$

to be the reverse information projection of the empirical distribution $\widehat{P}$ on tree T. Hence, distribution Q is factorized according to the tree T (according to Eq. (I.2)) and satisfies $\mathbb{E}_Q X_i X_j = \mathbb{E}_{\widehat{P}} X_i X_j$ for all edges $(i, j) \in \mathcal{E}_{\mathsf{T}}$.

Theorem IV.1. *Let distribution Q be defined as in Eq.* (IV.1)*. Then, given $n > 80\eta^{-2} \log \frac{p}{\delta}$ samples, we have*

$$\mathcal{L}^{(k)}(P, Q) \leq \sqrt{k}\, \eta$$

for all $1 \leq k \leq p$ with probability $1 - \delta$.

Corollary IV.2. *Let distribution Q be defined as in Eq.* (IV.1)*. Then, given n samples, we have a universal constant C such that*

$$\mathbb{E}\big[\mathcal{L}^{(k)}(P, Q)\big] \leq C \sqrt{\frac{k}{n} \log p} \tag{IV.2}$$

The statement of the corollary can be derived using standard techniques from Theorem IV.1.

This main result implies that using moment matching techniques, one can achieve optimal accuracy in terms of k-wise marginals of the learned distribution for any $k \geq 2$. Note that by using moment matching, the distribution Q is enforced to be factorized according to the given tree T, and its parameters on the edges are computed by looking at the pairwise empirical distributions along the edges. This implies that by exploiting the tree structure, the local accuracy along the edges is sufficient to guarantee global accuracy of the learned distribution in terms of higher order marginals.

Theorem IV.3 (Necessary sampling for parameter estimation)**.** *Fix $2 \leq k \leq p$. Given $n < c\eta^{-2} \max\{k, \log p\}$ samples and given tree structure T, the worst case probability of $\mathcal{L}^{(k)}$ loss greater than η taken over distributions $R \in \widetilde{P}_{\mathsf{T}}$ is at least half for any algorithm, i.e.,*

$$\inf_{\phi} \sup_{P \in \widetilde{P}_{\mathsf{T}}} \mathbb{P}[\mathcal{L}^{(k)}(P, \phi(X^{(1:n)})) \geq \eta] > 1/2\,.$$

where $\phi(X^{(1:n)}$ is an estimator which provides a distribution given i.i.d. samples $X^{(1:n)}$.

Proof. We assume the tree T is a Markov chain, such that $X_1 \to X_2 \to \cdots, \to X_p$. If $k < \log p$, we choose p different models over this Markov chain each denoted by edge parameters $\theta^{(m)}$. We assume that for $i = 1, \cdots, p-1$, we have $\theta^{(1)}_{i,i+1} = \mathsf{atanh}(\eta)$. We also assume that in the m-th model, we have $\theta^{(m)}_{m,m+1} = \mathsf{atanh}(\eta)$ and the remaining edge weights are $\mathsf{atanh}(2\eta)$. Then, applying the Fano's inequality Similar to the proof of Theorem 3.3. in [3] gives the statement of the theorem. If $k \geq \log p$, we choose p different models over this Markov chain denoted by edge parameters $\theta^{(0)}$ and $\theta^{(1)}$ such that $\mathcal{L}^k$ between these two models is greater than η. To do so, assume $\theta^{(0)}_{i,i+1} = \mathsf{atanh}(\eta/k)$ for all $i \leq k+1$ and $\theta^{(0)}_{i,i+1} = 0$ for all i. □

V. Proof of the Main Result

A. Event of interest

We introduce an event of interest in this section and bound $\mathcal{L}^{(k)}$ on this event. Later on, we show that this event happens with large probability given $n > C\eta^{-2} \log p$.

Definition V.1. For $\eta \geq 0$, define the $\mathsf{E}^{\mathsf{pair}}(\eta)$ as follows:

$$\mathsf{E}^{\mathsf{pair}}(\eta) := \Bigg\{ 1 - \frac{\eta}{\sqrt{P(x_i, x_j)}} < \frac{Q(x_i, x_j)}{P(x_i, x_j)} < 1 + \epsilon_{ij} \quad \text{with } \epsilon_{ij} = 2 \max\Big\{ \frac{\eta}{\sqrt{P(x_i, x_j)}}, \frac{\eta^2}{P(x_i, x_j)} \Big\} \quad \text{for all } x_i, x_j \in \{-,+\} \text{ and } i \neq j \Bigg\} \tag{V.1}$$

Lemma V.2. *On the event $\mathsf{E}^{\mathsf{pair}}(\eta)$ (Definition V.1), we have $\mathcal{L}^{(k)}(P, Q) \leq 4\sqrt{k}\eta$ for all $k \geq 2$.*

Simple algebraic manipulation can prove the statement of lemma for $k = 2$. For $k > 2$, we prove the statement of lemma in Section V-C by bounding the TV distance between distributions with the Hellinger distance and then using subadditivity of the Hellinger distance along with the properties of the event $\mathsf{E}^{\mathsf{pair}}(\eta)$ as in Definition V.1. In particular, the statement of the lemma is a direct consequence of Equations (V.3) and (V.5) and Lemma V.3 (summarized in Eq. (V.6)) and Lemma V.4.

Later, in Corollary V.5 in Section V-D, we show that

$$\mathbb{P}[\mathsf{E}^{\mathsf{pair}}(\eta)] > 1 - 2p^2 \exp\big(-n\eta^2\big)\,. \tag{V.2}$$

Hence, the Theorem IV.1 is a direct consequence of Lemmas V.2 and V.5.

B. Marginal distribution over the subtree induced by $\mathcal{S}$

To bound $\mathcal{L}^{(k)}$, we need an upper bound for $d_{\mathsf{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}})$ for any $\mathcal{S} \subseteq \mathcal{V}$ such that $|\mathcal{S}| = k$. Fix $\mathcal{S} \subseteq \mathcal{V}$ with $|\mathcal{S}| = k$. To compare the marginals of distributions P and Q on $\mathcal{S}$, we look at the subtree $\widetilde{\mathsf{T}} = (\mathcal{V}_{\widetilde{\mathsf{T}}}, \mathcal{E}_{\widetilde{\mathsf{T}}})$ of tree T induced by the nodes in $\mathcal{S}$. It can be shown that $\widetilde{k} \triangleq |\mathcal{V}_{\widetilde{\mathsf{T}}}| \leq 2k$ and

$$d_{\mathsf{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}) \leq d_{\mathsf{TV}}(P_{\mathcal{V}_{\widetilde{\mathsf{T}}}}, Q_{\mathcal{V}_{\widetilde{\mathsf{T}}}})\,, \tag{V.3}$$

since $\mathcal{S} \subseteq \mathcal{V}_{\widetilde{\mathsf{T}}}$. First, we define an ordering over the $\widetilde{k}$ nodes of the tree $\widetilde{\mathsf{T}}$. Pick an arbitrary node of the tree $\widetilde{\mathsf{T}}$ as root and

label it as node 1. Define an ordering in the nodes of tree $\widetilde{\mathsf{T}}$ such that for any $\ell > 1$, the nodes $1, \cdots, \ell$ form a connected component in tree $\widetilde{\mathsf{T}}$. For each node i, define the parent of the node i, $\pi(i)$ as the first node in the path between node i and the root. With this choice of labeling over nodes, $\pi(i) < i$. We use this ordering over the nodes to define a decomposition of the distributions described by $\widetilde{\mathsf{T}}$. If $P_{\mathcal{V}_{\widetilde{\mathsf{T}}}}$ is described by $\widetilde{\mathsf{T}}$, then

$$P_{\mathcal{V}_{\widetilde{\mathsf{T}}}}(x) = P(x_1) \prod_{i=2}^{\widetilde{k}} P(x_i|x_{\pi(i)}) \,.$$

Throughout this section, the subscript will be omitted when clear from the argument, for instance we use $P(x_i|x_{\pi(i)})$ instead of $P_{i|\pi(i)}(x_i|x_{\pi(i)})$. We can decompose $Q_{\mathcal{V}_{\widetilde{\mathsf{T}}}}$ described by $\widetilde{\mathsf{T}}$ similarly.

$$Q_{\mathcal{V}_{\widetilde{\mathsf{T}}}}(x) = Q(x_1) \prod_{i=2}^{\widetilde{k}} Q(x_i|x_{\pi(i)}) \,.$$

Note that $Q(x_i|x_{\pi(i)})$ is the marginal distribution of Q defined in Eq. (IV.1) on nodes i and $\pi(i)$. If $(i, \pi(i)) \in \mathcal{E}_{\mathsf{T}}$ is an edge in the original tree, according to the moment matching, we have $Q(x_i|x_{\pi(i)}) = \widehat{P}(x_i|x_{\pi(i)})$. But when $(i, \pi(i)) \notin \mathcal{E}_{\mathsf{T}}$, the $Q(x_i|x_{\pi(i)})$ is derived by cascading the edges along the path between i and $\pi(i)$ with edge parameters based on $\widehat{P}$.

C. Subadditivity of Hellinger Distance (Proof of Lemma V.2)

Given two distributions $P(x)$ and $Q(x)$, the Hellinger distance $H(P, Q)$ is defined such that

$$\begin{aligned} H^2(P,Q) &= \frac{1}{2} \sum_x \Big(\sqrt{P(x)} - \sqrt{Q(x)} \Big)^2 \\ &= 1 - \sum_x \sqrt{P(x)\, Q(x)} \,. \end{aligned} \tag{V.4}$$

Hellinger distance satisfies the following inequalities:

$$H^2(P,Q) \le d_{\mathsf{TV}}(P,Q) \le \sqrt{2} H(P,Q) \,. \tag{V.5}$$

We use Hellinger distance and the subadditivity property of Hellinger distance described in Lemma V.3) to bound the TV distance between k-wise marginals of distributions P and Q.

Lemma V.3. *Using the notation defined in Section V-B,*

$$H^2(P_{\mathcal{V}_{\widetilde{\mathsf{T}}}}, Q_{\mathcal{V}_{\widetilde{\mathsf{T}}}}) \le \sum_{i=2}^{\widetilde{k}} H^2\big(P_{i,\pi(i)}, Q_{i,\pi(i)}\big) \,.$$

where $P_{i,\pi(i)}$ and $Q_{i,\pi(i)}$ are marginal distribution of P and Q on the variables X_i and $X_{\pi(i)}$

The proof is similar to the proof of Theorem 2.1 in [8]. We provide the variation of the proof with our notation for the sake of completeness in the Appendix.

Note that using the description of the tree $\widetilde{\mathsf{T}}$ in Section V-B, we have $\tilde{k} \le 2k$. Hence, using Equations (V.3) and (V.5) and Lemma V.3, for two distributions P and Q factorized according to the same tree T, we have

$$\mathcal{L}^{(k)}(P,Q) \le 2\sqrt{k} \,. \max_{i,j} H\big(P_{i,j}, Q_{i,j}\big) \tag{V.6}$$

where $P_{i,j}$ and $Q_{i,j}$ are marginal distribution of P and Q on the variables X_i and X_j.

Lemma V.4. *On the event $\mathsf{E}^{\mathsf{pair}}(\eta)$ (Definition V.1), for any $i \neq j$, we have*

$$H^2\big(P_{i,j}, Q_{i,j}\big) \le 8\eta^2 \,,$$

where $P_{i,j}$ and $Q_{i,j}$ are marginal distribution of P and Q on the variables X_i and X_j.

Proof. We use the shorthand $P(a,b) = P_{i,j}(a,b)$ and $Q(a,b) = Q_{i,j}(a,b)$ in the proof of this lemma. Using definition of Hellinger distance in Eq. (V.4),

$$H^2\big(P_{i,j}, Q_{i,j}\big) = \sum_{a,b \in \{-1,+1\}} P(a,b) \left[\sqrt{\frac{Q(a,b)}{P(a,b)}} - 1 \right]^2 .$$

We bound each of the above four terms in terms of η on the event $\mathsf{E}^{\mathsf{pair}}(\eta)$. Note that for some values of $a, b \in \{-1, +1\}$ we could have $Q(a,b) > P(a,b)$ or $Q(a,b) \le P(a,b)$.

a) Case 1, $Q(a,b) > P(a,b)$: On the event $\mathsf{E}^{\mathsf{pair}}(\eta)$

$$\begin{aligned} &Q(a,b) \le (1+\epsilon)\, P(a,b) \\ &\qquad \text{with } \epsilon = 2 \max\Big\{ \frac{\eta}{\sqrt{P(a,b)}}, \frac{\eta^2}{P(a,b)} \Big\} \end{aligned}$$

(according to Definition V.1) and since $(\sqrt{1+\epsilon} - 1)^2 \le \epsilon^2/(2+\epsilon)$ for all $\epsilon \ge 0$, we have

$$P(a,b)[\sqrt{1+\epsilon} - 1]^2 \le P(a,b) \frac{\epsilon^2}{2+\epsilon} \le 2\eta^2 \,.$$

b) Case 2, $Q(a,b) \le P(a,b)$: On the event $\mathsf{E}^{\mathsf{pair}}(\eta)$,

$$Q(a,b) > \left[1 - \frac{\eta}{\sqrt{P(a,b)}} \right] P(a,b) \,,$$

(according to Definition V.1) and since $1 - \sqrt{1-x} \le x$ for all $0 \le x \le 1$, we have

$$P(a,b) \left[1 - \sqrt{1 - \frac{\eta}{\sqrt{P(a,b)}}} \right]^2 \le \eta^2 \,. \quad \square$$

D. Concentration bounds (Proof of Lemma V.5)

Lemma V.5. *Given n samples, we have*

$$\mathbb{P}[\mathsf{E}^{\mathsf{pair}}(\eta)] \le 8p^2 \exp\big(-n\eta^2/4 \big) \,.$$

Proof. To prove this lemma, we show that for any $i \neq j$ and any x_i, x_j we have

$$\mathbb{P}\Big[Q(x_i,x_j) \ge (1+t) P(x_i,x_j) \Big] \le \exp\Big(-n \frac{t^2}{2+t} P(x_i,x_j) \Big)$$

for any $t \ge 0$. Also,

$$\mathbb{P}\Big[Q(x_i,x_j) \le (1-t) P(x_i,x_j) \Big] \le \exp\Big(-nt^2 P(x_i,x_j)/2 \Big)$$

for any $0 \le t \le 1$. Next, using a union bound on $i \neq j$ and $x_i, x_j \in \{-, +\}$ gives the statement of the lemma

The proof of these statements uses the definition of the distribution $Q(x) = \Pi_{\mathsf{T}}(\widehat{P})$ which implies that Q is factorized

according to the tree T (Eq. (I.2)) and pairwise marginals according to Q on the edges of T is the same as the pairwise marginals according to empirical distribution $\widehat{P}$. Hence, for $(i,j) \in \mathcal{E}$ and any $x_i, x_j \in \{-,+\}$, we have $Q(x_i,x_j) = \frac{1}{2}(1+x_i x_j \widehat{\mu}_{i,j})$. Also, for $(i,j) \notin \mathcal{E}$ and any $x_i, x_j \in \{-,+\}$, we have $Q(x_i,x_j) = \frac{1}{2}(1 + x_i x_j \prod_{e \in \mathsf{path}_{\mathsf{T}}(i,j)} \widehat{\mu}_e)$.

This characterization of distribution Q implies that the statement of the lemma is a direct application of Lemma V.6 on the deviation bound of product of binomial random variables. □

Lemma V.6. *Let $Z_1, \dots, Z_t$ be independent random variables such that for each $i \in [t]$, $Z_i \sim 2\mathsf{Bin}(n,\mu_i)/n - 1$. Let $\mu = \prod_{i=1}^n \mu_i$. Then*

$$\mathcal{P}\Big[\Big|\prod_{i=1}^t Z_i - \prod_{i=1}^t \mathbb{E}[Z_i]\Big| \geq t\Big] \leq 2\exp\Big(\frac{-nt^2}{2(1-\mu^2)+4t/3}\Big).$$

Proof. For each $i \in [t]$, write $Z_i = \frac{1}{n}\sum_{j=1}^n X_{ij}$, where the $X_{ij} \sim \mathsf{Rad}\big((1+\mu_i)/2\big)$ are independent. Then

$$\begin{aligned}\prod_{i=1}^t Z_i &= \frac{1}{n^t}\prod_{i=1}^t\sum_{j=1}^n X_{ij} \\ &= \frac{1}{n^t}\sum_{j_1,\dots,j_t\in[n]}\prod_{i=1}^t X_{ij_i} \\ &= \frac{1}{n^{t-1}}\sum_{s_2,\dots,s_t\in[n]}\Big(\frac{1}{n}\sum_{s_1\in[n]}\prod_{i=1}^t X_{i,(\sum_{l=1}^i s_l \mod n)}\Big).\end{aligned}$$

Note that for each choice of $s_2,\dots,s_t \in [n]$, the variable $Y_{s_2,\dots,s_t} := \frac{1}{n}\sum_{s_1\in[n]}\prod_{i=1}^t X_{i,(\sum_{l=1}^i s_l \mod n)}$ is equal in distribution to $Y := Y_{0,\dots,0} = \frac{1}{n}\sum_{j\in[n]}\prod_{i=1}^t X_{i,j}$.

Therefore for any $\lambda \in \mathcal{R}$,

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\Big(\prod_{i=1}^t Z_i - \mathbb{E}\Big[\prod_{i=1}^t Z_i\Big]\Big)\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\lambda}{n^{t-1}}\sum_{s_2,\dots,s_t\in[n]}\big(Y_{s_2,\dots,s_t} - \mathbb{E}[Y_{s_2,\dots,s_t}]\big)\right)\right] \\ &\overset{(a)}{\leq} \frac{1}{n^{t-1}}\sum_{s_2,\dots,s_t\in[n]}\mathbb{E}\big[\exp\big(\lambda(Y_{s_2,\dots,s_t} - \mathbb{E}[Y_{s_2,\dots,s_t}])\big)\big] \\ &= \mathbb{E}\big[\exp\big(\lambda(Y-\mathbb{E}[Y])\big)\big],\end{aligned}$$

where Inequality (a) is by convexity of the exponential function. Now we can bound the deviations of $\prod_{i=1}^t Z_i$ by Bernstein's inequality. Since $Y - \mathbb{E}[Y] = \frac{1}{n}\sum_{i\in[n]}\big(W_i - \mathbb{E}[W_i]\big)$ for $W_1,\dots,W_n \overset{\text{i.i.d.}}{\sim} \mathsf{Rad}\big((1+\mu)/2\big)$, the Bernstein bound reads

$$\mathbb{E}\big[\exp\big(\lambda(Y-\mathbb{E}[Y])\big)\big] \leq \exp\Big(\lambda^2 n(1-\mu^2)\frac{(e^{2\lambda}-1-2\lambda)}{4\lambda^2}\Big).$$

Optimizing λ and applying a Markov bound as in the proof of the Bernstein bound yields (similar to [19] and [11])

$$\mathbb{P}\Big[\Big|\prod_{i=1}^t Z_i - \prod_{i=1}^t \mathbb{E}[Z_i]\Big| \geq t\Big] \leq 2\exp\Big(\frac{-nt^2}{2(1-\mu^2)+4t/3}\Big).$$

□

ACKNOWLEDGMENT

We appreciate the help from Fredric Koehler and Enric Boix in pointing out a shorter proof of Lemma V.6 which appears in the paper.

APPENDIX: PROOF OF LEMMA V.3

Since $\pi(\ell) < \ell$, we can prove the following statement which gives the result in the lemma.

$$\begin{aligned}&H^2(P(x_1,\cdots,x_\ell), Q(x_1,\cdots,x_\ell)) \\ &\quad\leq H^2(P(x_1,\cdots,x_{\ell-1}), Q(x_1,\cdots,x_{\ell-1})) \\ &\quad\quad + H^2(P(x_\ell,x_{\pi(\ell)}), Q(x_\ell,x_{\pi(\ell)}))\,.\end{aligned}$$

A recursion on the above statement gives the statement of the lemma. For given $\ell > 0$, define $\mathcal{S} = \{1,\cdots,\ell-1\}$ and $x_{\mathcal{S}} = x_1,\cdots,x_{\ell-1}$. Then,

$$\begin{aligned}&H^2(P(x_1,\cdots,x_\ell), Q(x_1,\cdots,x_\ell)) \\ &= 1 - \sum_{x_{\mathcal{S}},x_\ell}\sqrt{P(x_{\mathcal{S}},x_\ell)\,Q(x_{\mathcal{S}},x_\ell)} \\ &= 1 - \sum_{x_{\mathcal{S}}}\sqrt{P(x_{\mathcal{S}})\,Q(x_{\mathcal{S}})}\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &= 1 - \sum_{x_{\mathcal{S}}}\frac{P(x_{\mathcal{S}})+Q(x_{\mathcal{S}})}{2}\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &\quad + \sum_{x_{\mathcal{S}}}\left(\frac{P(x_{\mathcal{S}})+Q(x_{\mathcal{S}})}{2} - \sqrt{P(x_{\mathcal{S}})\,Q(x_{\mathcal{S}})}\right) \\ &\quad\quad \times \sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})}\end{aligned}$$

Next, we bound each term separately

$$\begin{aligned}&\sum_{x_{\mathcal{S}}}\frac{P(x_{\mathcal{S}})+Q(x_{\mathcal{S}})}{2}\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &= \sum_{x_{\pi(\ell)}}\frac{P(x_{\pi(\ell)})+Q(x_{\pi(\ell)})}{2}\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &\geq \sum_{x_{\pi(\ell)}}\sqrt{P(x_{\pi(\ell)})Q(x_{\pi(\ell)})}\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &= 1 - H^2(P(x_\ell,x_{\pi(\ell)}), Q(x_\ell,x_{\pi(\ell)}))\end{aligned}$$

where we used AM-GM inequality.

$$\begin{aligned}&\sum_{x_{\mathcal{S}}}\left(\frac{P(x_{\mathcal{S}})+Q(x_{\mathcal{S}})}{2} - \sqrt{P(x_{\mathcal{S}})\,Q(x_{\mathcal{S}})}\right) \\ &\quad\quad\times\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &\quad= \frac{1}{2}\sum_{x_{\mathcal{S}}}\Big(\sqrt{P(x_{\mathcal{S}})} - \sqrt{Q(x_{\mathcal{S}})}\Big)^2\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \\ &\quad\leq \frac{1}{2}\sum_{x_{\mathcal{S}}}\Big(\sqrt{P(x_{\mathcal{S}})} - \sqrt{Q(x_{\mathcal{S}})}\Big)^2 = H^2(P(x_{\mathcal{S}}), Q(x_{\mathcal{S}}))\end{aligned}$$

where we used Cauchy-Schwartz to get $\sum_{x_\ell}\sqrt{P(x_\ell|x_{\pi(\ell)})\,Q(x_\ell|x_{\pi(\ell)})} \leq 1$.

REFERENCES

[1] BABICHENKO, Y., BARMAN, S., AND PERETZ, R. Empirical distribution of equilibrium play and its testing application. *arXiv preprint arXiv:1310.7654* (2013).

[2] BRESLER, G. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* (2015), ACM, pp. 771–782.

[3] BRESLER, G., AND KARZAND, M. Learning a tree-structured ising model in order to make predictions. *arXiv preprint arXiv:1604.06749* (2016).

[4] CANONNE, C., DIAKONIKOLAS, I., KANE, D., AND STEWART, A. Testing bayesian networks. *arXiv preprint arXiv:1612.03156* (2016).

[5] CHOW, C., AND LIU, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information theory 14*, 3 (1968), 462–467.

[6] CSISZÁR, I., AND SHIELDS, P. C. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[7] DASKALAKIS, C., DIKKALA, N., AND KAMATH, G. Testing ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (2018), SIAM, pp. 1989–2007.

[8] DASKALAKIS, C., AND PAN, Q. Square hellinger subadditivity for bayesian networks and its applications to identity testing. *arXiv preprint arXiv:1612.03164* (2016).

[9] HAMILTON, L., KOEHLER, F., AND MOITRA, A. Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems* (2017), pp. 2460–2469.

[10] HEINEMANN, U., AND GLOBERSON, A. Inferning with high girth graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 1260–1268.

[11] HOEFFDING, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association 58*, 301 (1963), 13–30.

[12] JALALI, A., JOHNSON, C., AND RAVIKUMAR, P. On learning discrete graphical models using greedy methods. *arXiv preprint arXiv:1107.3258* (2011).

[13] KOLLER, D., AND FRIEDMAN, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[14] LAURITZEN, S. *Graphical models*. Oxford University Press, 1996.

[15] NARASIMHAN, M., AND BILMES, J. PAC-learning bounded tree-width graphical models. *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence* (2004), 410–417.

[16] RAVIKUMAR, P., WAINWRIGHT, M., AND LAFFERTY, J. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics 38*, 3 (2010), 1287–1319.

[17] REBESCHINI, P., VAN HANDEL, R., ET AL. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability 25*, 5 (2015), 2809–2866.

[18] RIGOLLET, P. High-dimensional statistics. *Lecture notes for course 18S997* (2015).

[19] SRIDHARAN, K. A gentle introduction to concentration inequalities. *Dept. Comput. Sci., Cornell Univ., Tech. Rep* (2002).

[20] TAN, V. Y., ANANDKUMAR, A., TONG, L., AND WILLSKY, A. S. A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Transactions on Information Theory, 57*, 3 (2011), 1714–1735.

[21] TAN, V. Y., ANANDKUMAR, A., AND WILLSKY, A. S. Learning high-dimensional Markov forest distributions: Analysis of error rates. *The Journal of Machine Learning Research 12* (2011), 1617–1653.

[22] VUFFRAY, M., MISRA, S., LOKHOV, A., AND CHERTKOV, M. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems* (2016), pp. 2595–2603.

[23] WAINWRIGHT, M. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *The Journal of Machine Learning Research 7* (2006), 1829–1857.

[24] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning 1*, 1-2 (2008), 1–305.

[25] WAINWRIGHT, M. J., JAAKKOLA, T. S., AND WILLSKY, A. S. Map estimation via agreement on trees: message-passing and linear programming. *IEEE transactions on information theory 51*, 11 (2005), 3697–3717.