LEARNING TREE-STRUCTURED MODELS FROM NOISY DATA

by

KONSTANTINOS NIKOLAKAKIS

A dissertation submitted to the
School of Graduate Studies—New Brunswick
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Electrical and Computer Engineering

Written under the direction of
Anand D. Sarwate
and approved by

New Brunswick, New Jersey
May, 2021
ABSTRACT OF THE DISSERTATION

Learning Tree-Structured Models from Noisy Data

by Konstantinos Nikolakakis
Dissertation Director: Anand D. Sarwate

Graphical models are a widely-used and powerful tool for analyzing high-dimensional structured data [1, 2]. In those models, variables are represented as nodes of a graph, the edges of which indicate conditional dependencies among the corresponding nodes. We consider the problem of learning acyclic undirected graphs or tree-structured Markov Random Fields (MRFs), on the basis of noisy data. In particular, we study the well-known Chow-Liu (CL) Algorithm [3], which, given a dataset of samples drawn from a tree-structured distribution, returns an estimate of the original tree. The importance of the CL algorithm stems from its efficiency, its low computational complexity, and also its optimality in terms of sample complexity (matching information-theoretic limits) for a variety of statistical settings, e.g., such involving Gaussian data [4], binary data [5], as well as for non-parametric models with discrete alphabets [6, 7].

In our work, we present the first sample complexity analysis for structure and predictive learning of hidden tree-structured models. Specifically, we provide high-probability sample complexity guarantees for exact structure recovery and accurate predictive learning using noise-corrupted samples from an acyclic (tree-shaped) graphical model. To begin with, the parametric setting under consideration includes hidden variables that follow a tree-structured Ising model distribution, such that the observable variables are generated by a binary symmetric channel taking the hidden variables as its
input (flipping each bit independently with some constant probability $q \in [0, 1/2]$). This simple model arises naturally in a variety of applications, such as in physics, biology, computer science, and finance. In the absence of noise, the structure learning (CL algorithm) problem was recently studied by Bresler and Karzand (2020); the results in this dissertation quantify how noise in the hidden model impacts the sample complexity of structure learning and marginal distributions’ estimation by proving upper and lower bounds on the sample complexity. Our theorems generalize state-of-the-art bounds reported in prior work, and they exactly recover the noiseless ($q = 0$) case. Specifically, for any tree with $p$ vertices and probability of incorrect recovery $\delta > 0$, the sufficient number of samples remains logarithmic as in the noiseless case, i.e., $O(\log(p/\delta))$, while the dependence on $q$ is $O(1/(1 - 2q)^4)$ for both aforementioned tasks. We also present a new equivalent of Isserlis’s Theorem for sign-valued tree-structured distributions, yielding a new low-complexity algorithm for higher order moment estimation. In a similar way to the Ising model, we derive an expression of the sufficient number of noisy samples (AWGN) for exact hidden tree structure recovery from Gaussian data as well.

In addition to the parametric cases of Ising and Gaussian models, we deal with the problem of hidden structure learning under minimal assumptions on the distribution of the hidden variables and under general noisy channels (non-parametric case). We provide high probability finite sample complexity guarantees for non-parametric structure learning of tree-shaped graphical models whose nodes are discrete random variables with either finite or countable alphabets. Asymptotic sample complexity bounds for the noiseless case have been shown by Tan, Anandkumar, and Willsky [4] and Tan et al. [7]. However, based on the existing literature, it remains unclear how the sample complexity is affected under the presence of noise for generic tree-structured hidden models. As we show, the difficulty of the problem is summarized in a statistic that we call information threshold. The information threshold arises naturally from the error analysis of the Chow-Liu algorithm and, as we discuss, provides explicit necessary and sufficient conditions on sample complexity, by effectively summarizing the difficulty of the tree-structure learning problem. Specifically, we show that finite sample complexity of the Chow-Liu algorithm for ensuring exact (hidden) structure recovery is inversely proportional to the squared
information threshold, and scales almost logarithmically relative to the number of nodes over a given probability of failure, also matching relevant asymptotic results in the literature. Conversely, in the noisy case, we show that, the necessary number of samples is also inversely proportional to the squared information threshold. As consequence of the matching between the lower and upper sample complexity bounds, the information threshold is a fundamental quantity for the problem of hidden tree-structure learning. Lastly, as a byproduct of our analysis, we resolve the problem of tree structure learning in the presence of non-identically distributed observation noise, providing conditions for convergence of the Chow-Liu algorithm under this setting, as well.
Acknowledgements

First, I would like to express my gratitude to my adviser professor Anand Sarwate and to professor Dionysios Kalogerias for our collaboration, as well as for their support all along my Ph.D. research journey.

Secondly, I would like to express my appreciation to Emina Soljanin, Roy Yates, and Salim El Rouayheb for providing valuable comments and suggestions, which have substantially improved the presentation of the results and the overall quality of the dissertation.

Finally, I would like to thank Rutgers University and the professors with whom I collaborated with. Specifically, I would like to thank professor Andrzej Ruszczynski for his valuable lectures on Dynamic Programming, as well as professor Richard F. Gundy for constructive discussions and for his lectures on measure theory and real analysis.
Table of Contents

Abstract ................................................................. ii
Acknowledgements ...................................................... v
List of Tables ............................................................ x
List of Figures ............................................................ xi

1. Introduction ............................................................ 1
   1.1. Motivating Examples and Applications ......................... 2
   1.2. Dissertation Outline ........................................... 3
   1.3. Related Work .................................................. 4
   1.4. Notation ....................................................... 6
   1.5. Contributions ................................................ 8

2. Learning Binary and Gaussian Tree-Structured Models from Noisy Data ...................................................... 13
   2.1. Tree-Structured Ising Models, Preliminaries .................. 14
       2.1.1. Undirected Graphical Models .......................... 14
       2.1.2. Sign-Valued Markov Fields on Trees .............. 15
       2.1.3. Hidden Sign-Valued Tree-Structured Models .......... 17
       2.1.4. Hidden Structure Estimation .......................... 18
       2.1.5. Evaluating the Accuracy of the Estimated Distribution .... 19
   2.2. Models and Problem Statement .................................. 24
       2.2.1. Hidden Tree-structured Ising Models ............... 24
       2.2.2. Tree structured Gaussian models ................. 24
3.5.4. Analysis of the event $E_{\text{cascade}}^{\gamma}$ ................................. 90
3.5.5. Predictive Learning, Proof of Theorem 5 and Theorem 6 ........ 116
3.5.6. Theorem 9: KL-Divergence Loss ................................. 124
3.5.7. Proof of Theorem 7 ................................................. 125
3.5.8. Supplementary Discussion ........................................ 126
    3.5.8.1. The Gap between the Upper and Lower Bounds ....... 126
    3.5.8.2. A Structure-Preserving Case ................................. 128

4. Optimal Non-Parametric Structure Learning for Markov Trees ....... 129

4.1. Model and Problem Statement ........................................ 131
    4.1.1. Tree-Structured and Hidden Tree-Structured Models ...... 131
    4.1.2. The Classical Chow-Liu Algorithm ............................. 133
    4.1.3. Information Threshold and Properties ........................ 135

4.2. Recovering the Structure from Noisy Data ........................... 136

4.3. Converse: Information Threshold as a Fundamental Quantity .... 142

4.4. Recovering the Structure from Noisy Data ........................... 148
    4.4.1. IOP and the Importance of Pre-Processing ..................... 148
        4.4.1.1. Convergence of CL Algorithm under the IOP ............ 149
        4.4.1.2. Enforcing the IOP through pre-processing ............. 150
    4.4.2. Examples and Applications on Specific Models ............... 154

4.5. Simulations .......................................................... 158

4.6. Conclusion & Future Directions ...................................... 158

4.7. Appendix ............................................................. 159
    4.7.1. Proof of Proposition 5: Positivity of the Noiseless Information Threshold ..................................................... 160
    4.7.2. Proof of Proposition 6: Locality of the Noiseless Information Threshold ..................................................... 160
    4.7.3. Fano's Inequality .................................................. 160
    4.7.4. Proofs of the Conditions in Section 4.4.2 ..................... 161
4.7.4.1. $M$-ary Erasure Channel ........................................ 161
4.7.4.2. Binary Symmetric Channel with Non-Identically Distributed Noise ........................................ 162
4.7.4.3. $M$-ary Symmetric Channel ................................. 164

5. Conclusions and Future Directions ................................. 173

REFERENCES ......................................................... 182
List of Tables

1.1. Notation/Definitions. .................................................. 7
3.1. Sufficient number of samples for accurate structure and predictive learning. 64
3.2. Necessary number of samples for structure and predictive learning. . . 64
List of Figures

2.1. Estimate of the number of mismatched edges $\mathcal{D}(T, T^{\text{CL}})$ as a function of number of samples. The upper graph is over 1000 independent runs and up to $10^4$ independent samples, while the down over 100 independent runs and up to $10^5$ independent samples. . . . . . . . . . . . . . . . . . . 36

2.2. Ising Model data. Estimating the probability of the error event, $\mathbb{P}(T^{\text{CL}} \neq T)$, as a function of number of samples. Upper graph: 100 independent runs. Down: 1000 independent runs. . . . . . . . . . . . . . . . . . . 37

2.3. Gaussian Data. Upper: $\mathcal{D}(T, T^{\text{CL}})$ as a function of number of samples. Down: Estimating the probability of the error event $T^{\text{CL}} \neq T$. Both estimates are evaluated through 1000 independent iterations. . . . . . . . 38

2.4. Left: Tree structure estimate $T^{\text{CL}}$ for the closing prices of 10 different stocks. $10^5$ number of samples have been considered, data have been retrieved from https://finance.yahoo.com. Right: $\mathcal{D}(T^{\text{CL}}, T^{\text{CL}})$ as a function of $q$ for different number of samples. . . . . . . . . . . . . . . . 39

3.1. The simulation corresponds to structure learning. Comparison of the experimental results (heat-map) and the theoretical bound of Theorem 1, the bound that yields (3.1). The colored regions denote different values of the estimated probability of error $\delta$ (at least one edge has been missed). The value of $\delta$ varies between 0 and 1 while the parameters $\alpha = 0.2, \beta = 1.1, p = 100$ are fixed. The red line shows the bound from Theorem 1 (the explicit form of Theorem 4). The code of the experiment is available at structure learning code. . . . . . . . . . . . . . . . . . . . 60
3.2. The simulation corresponds to predictive learning. Comparison of the experimental results (heat-map) and the theoretical bound of Theorem 5. The colored regions denote different values of the estimated probability of error $\delta$ (ssTV to be greater than a fixed number $\eta$). The value of $\delta$ varies between 0 and 1 while the parameters $\eta = 0.03$, $\beta = 1.1$, $p = 31$ are fixed. The code is available at [predictive learning code].

3.3. Probability of incorrect structure recovery, The theoretical bound is given by Theorem 1. The top view of the figure is Figure 3.1 and provides a clear comparison between the experimental and theoretical results.

3.4. Estimate of the probability of the ssTV to be greater than $\eta = 0.03$. The theoretical bound is given by Theorem 6. The top view of the figure is Figure 3.2 and provides a clear comparison between the experimental and theoretical results.

3.5. Estimate of the distribution error metric ssTV as a function of $q$ and $n$.

3.6. The probability of error in the predictive learning task for different values of $n$ and estimation error of the parameter $q$. For both figures we consider $p = 31$, $\alpha = 0.2$, $\beta = 1$, $q = 0.1$ and averaging over 500 independent runs. Left: $\hat{q} \in [0.05, 0.15]$, $\eta = 0.1$, Right: $\hat{q} \in [0, 0.2]$, $\eta = 0.12$.

3.7. Stream mapping of the results.


4.1. $\delta \equiv \hat{P}_n(T_{CL}^T \neq T)$ for 3-node hidden structure whose $I_{i}^o < 0$ (non-identically distributed noise). "Weak" and "Strong" refer to correlations between hidden variables: $|\mathbb{E}[X_1X_2]| < |\mathbb{E}[X_2X_3]|$.

4.2. Hidden tree-structured models $M_0$, $M_1$, $M_2$.

4.3. Representation of the 3 nodes hidden model and non-identically distributed noise. Pre-processing is required to to guarantee that $I_{i}^o > 0$. 
4.4. Estimate of the probability of incorrect recovery for a hidden Markov model with 3 nodes, and $I_0^0 < 0$ (originally), for different values of $n \in [10^3 \times 10^4]$, before and after pre-processing. Right: $q_1 = 0.01$, $q_2 = q_3 = 0.3$, Left: $q_1 = 0.2$, $q_2 = q_3 = 0.25$.

4.5. Left, center: estimating the probability $P(T^{CL} \neq T)$ for different values of $I^0$ and $n$ through $5 \times 10^3$ independent runs with noiseless data. Right: estimating the probability $P(T^{CL} \neq T)$ through $10^4$ samples and $5 \times 10^3$ independent runs with noisy data.
Chapter 1
Introduction

Probabilistic graphical models are a useful tool for modeling statistical properties in high-dimensional settings. The graph illustrates the conditional dependence between random variables represented as the nodes of the graph. The edge set corresponds to (often physical) interactions between variables, and the absence of an edge corresponds to conditional independence. There is a long and deep literature on graphical models (see Koller and Friedman [1] for a comprehensive introduction), and they have found wide applications in areas such as image processing and vision [8, 9, 10, 11, 12, 13], artificial intelligence more broadly [14, 15], signal processing [16, 17], and gene regulatory networks [18, 19], to name a few.

An undirected graphical model, or Markov random field (MRF) in particular, is defined in terms of a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, that models the Markov properties of a joint distribution on $p \triangleq |\mathcal{V}|$ node variables $(X_1, X_2, \ldots, X_p) \triangleq \mathbf{X}$. A tree-structured graphical model is one in which $\mathcal{G}$ is a tree. Further, we consider the setting for which noisy observables $(Y_1, Y_2, \ldots, Y_p) \triangleq \mathbf{Y}$ are generated as random transformation of $\mathbf{X}$.

In this dissertation we focus on the following problems: learning the hidden tree structure of $\mathbf{X}$ given samples from $\mathbf{Y}$ and learning a set of marginal distributions (predictive learning) of the hidden variables $\mathbf{X}$, when only noisy data from $\mathbf{Y}$ are available. In fact, for the structure learning problem we study the case of a Ising and Gaussian models and we provide the sample complexity for exact structure estimation. For the predictive learning task we consider the case of tree-structured Ising model and we analyze the sample in this problem as well. Later we present sample complexity bounds for generic hidden trees and noisy channels. We continue by providing motivating applications of our work and then we proceed with the outline of the dissertation.
1.1 Motivating Examples and Applications

Models for joint distributions characterized by pairwise variable interactions have found many applications, with the Ising model being a popular model for binary variables. Our work is primarily motivated by examples of Ising models corrupted by noise. In many cases, the underlying graph-structured process cannot be observed directly; instead, only a noisy version of the process is available; examples abound in physics, computer science, biology, medicine, psychology, social sciences, and finance. Some applications motivating this work include the following:

1) Statistical mechanics of population, social and pedestrian dynamics [20, 21]: The Ising model can be used to represent the statistical properties of the spreading of a feeling, behavior or the change of an emotional state among individuals in a crowd, where each individual interacts with his neighbors.

2) Epidemic dynamics and epidemiological models [22, 23]: Disease spread can be modeled through the Ising model, where each individual is susceptible (spin down) or ineffective (spin up).

3) Neoplastic transitions and related applications in biology [24]: Each cell interacts with neighboring cells. Different cases are studied in the literature, for instance, healthy versus cancerous cells, malignant versus benign cells, where both can be modeled as spin up and spin down observations. The probability of diagnostic error is not zero which gives rise to the hidden model that we consider.

4) Differential Privacy [25]: In computer science, differential privacy is used to guarantee privacy for individuals. A hidden model describes data gathered using a privacy-preserving mechanism such as randomized response (Section 2.7, Appendix).

5) Trading and related applications in economics [26, 27]: The Ising model has been considered in the literature to model increasing (spin up) or decreasing (spin down) price trends in a market. In Section 2.6, we consider the closing prices of ten equities to demonstrate the performance of Chow-Liu algorithm.

6) Recommender System: A recommendation system is an example of an application that involves finite and noisy data, and requires accurate marginal distributions in
order to make predictions. Specifically, based on users’ choices we would like to use the estimated model to predict future preference. Note that users’ choices include noise because of imperfect user behaviour [28].

1.2 Dissertation Outline

The remaining sections of Chapter 1 include an illustration of the related work for graphical models on structure learning, predictive learning and the notational convention that we will use. Later, we provide a detailed discussion on the contributions of our work.

The next two chapters of this dissertation, Chapter 2 and Chapter 3, consider the parametric setting of binary and Gaussian data tree-structured data. Specifically, we are interested in answering the following general question: *How does noise affect the sample complexity of the structure learning (Chapter 2) and predictive learning (Chapter 3) procedure?* That is, given only noisy observations, our goal is to learn the tree structure of the hidden layer in a well-defined and meaningful sense. For the structure learning problem we study the cases of binary data under the Ising model assumption and continuous data under the assumption of Gaussian observations. The maximum likelihood estimate of the structure\(^1\) from tree-structured (noiseless) data is the output of the Chow-Liu algorithm [3]. However, the MLE-structure from noisy data is not consistent with the hidden structure in general because the graphical model of the observables is a complete graph. Further, the (latent) MLE of the actual interaction parameters \(\theta\) of the hidden layer is intractable. In Sections 2.1.4 and 2.1.6 we explain the importance of Chow-Liu algorithm in our setting, we show why the classical MLE approach fails, and we discuss the connection between the output of the Chow-Liu algorithm and an alternative, projection-based MLE approach. For the predictive learning problem we consider the setting of hidden tree structured Ising models.

The estimated structure is an essential statistic for estimating the underlying distribution of the hidden layer, allowing for predictive learning (Chapter 3). We consider

\(^1\)For the definition of MLE-structure and further discussion see Section 2.1.6 and prior work [3, 29].
this problem under the assumption of tree-structured Ising model for the hidden layer. Specifically, based on the structure estimate, we are interested in appropriately approximating the tree-structured distribution under study, which can then be used for accurate predictions. We also consider the problem of hidden layer higher-order moment estimation of tree-structured Ising models and, in particular, how such estimation can be efficiently performed, on the basis of noisy observations.

In Chapter 4 we consider the problem of learning tree-structures from noisy observations under minimal assumptions on hand. In fact, the data are general discrete tree-structured variables with countable and uncountable alphabets. To characterize the required number of samples for exact hidden structure recovery, we study a fundamental quantity that we call the \textit{information threshold} and provide upper and lower sample complexity bounds for the classical CL algorithm \cite{3}.

Finally, in Chapter 5 we conclude by summarizing our contributions, the impact of our results and we provide additional discussion for possible future work.

\section*{1.3 Related Work}

For a detailed review of methods for structure learning involving undirected and directed graphical models, see the relevant article by Drton and Maathuis \cite{30}. In general, learning the structure of a graphical model from samples can be intractable \cite{31,32}. For general graphs, neighborhood selection methods \cite{33,34,35} estimate the conditional distribution for each vertex in order to learn the neighborhood of each node and therefore the full structure. These approaches may use greedy search or $\ell_1$ regularization. For Gaussian or Ising models, $\ell_1$-regularization \cite{36}, the GLasso \cite{37,38}, or coordinate descent approaches \cite{39} have been proposed, focusing on estimating the non-zero entries of the precision (or interaction) matrix. Model selection can also be performed using score matching methods \cite{40,41,42,43}, or Bayesian information criterion methods \cite{44,45,46}. Other works address non-Gaussian models such as elliptical distributions, t-distribution models or latent Gaussian data \cite{47,48,49,50}, or even mixed data \cite{51}.

For tree- or forest-structured models, exact inference and the structure learning
problem are significantly simpler: the Chow-Liu algorithm provides an estimate of the tree or forest structure of the underlying graph \([3, 6, 52, 53, 54, 55, 29]\). Furthermore, marginal distributions and maximum values are simpler to compute using a variety of algorithms (sum-product, max-product, message passing, variational inference) \([2, 14, 52, 56]\).

The noiseless counterpart of the Ising model considered in this dissertation was studied recently by Bresler and Karzand \([29]\); in our work, we extend their results to the hidden case, where samples from a tree-structured Ising model are passed through a binary symmetric channel with crossover probability \(q \in [0, 1/2]\). Of course, in the special case of a linear graph, our model reduces to a hidden Markov model. Latent variable models are often considered in the literature when some variables of the graph are deterministically unobserved \([57, 58, 59, 60]\). Our model is most similar to that studied by Chaganty et al. \([61]\), in which a hidden model is considered with a discrete exponential distribution and Gaussian noise. They solve the parameter estimation problem by using moment matching and pseudo-likelihood methods; the structure can be recovered indirectly using the estimated parameters.

**Connection with Phylogenetic Estimation.** In phylogenetic estimation problems, the goal is to learn the structure of tree given only observations form the leaves \([62]\). The sample complexity of phylogenetic reconstruction algorithms grows exponentially with respect to the depth of the tree \([62]\), however if we are interested in reconstructing only parts of the tree which are “close” to the leaves, then the depth of tree does not affect the sample complexity \([63]\). The hidden structure learning problem that we consider is a special case of phylogeny estimation problem with constant depth; there is exactly one noisy observable for each hidden node of the tree. In contrast with phylogenetic estimation approaches, Chow-Liu algorithm is simple and computationally more efficient, while the sample complexity is of the same order\(^2\) with the well-known phylogenetic reconstruction methods, to name a few “Dyadic Closure” method \([62]\), the “Contractor-Extender” and “Cherry-picking” algorithms \([63, 64, 65]\). On the other hand, the

---

\(^2\)while considering the depth fixed
approach of distribution estimation by matching the structure and the correlations \[29\] has not been considered in the phylogenetic estimation literature. Based on the above discussion, the following interesting question naturally rises: How well can we estimate the distribution of a hidden tree structured model while having access only to the leaves of the tree? The latter remains open problem for future work.

1.4 Notation

In this section we provide the notational convention and symbols that are consistent with the rest of dissertation. We will introduce further notation at the beginning of each Chapter if it is required. Boldface indicates a vector or tuple and calligraphic face for sets and trees. The sets of even and odd natural numbers are \(2\mathbb{N}\) and \(2\mathbb{N} + 1\) respectively.

For an integer \(n\), define \([n] \triangleq \{1, 2, \ldots n\}\). The indicator function of a set \(A\) is \(\mathbf{1}_A\). For a graph \(G = (\mathcal{V}, \mathcal{E})\), \(\mathcal{V} = [p]\) indexes the set of variables \(\{X_1, X_2, \ldots, X_p\}\), for any pair of vertices \(i, j \in \mathcal{V}\) the correlation \(\mu_{ij} \triangleq \mathbb{E}[X_i X_j]\) and for any edge \(e = (i, j) \in \mathcal{E}\) it is \(\mu_e \triangleq \mathbb{E}[X_i X_j]\). The correlation coefficient is denoted as \(\rho_{i,j} \triangleq \mathbb{E}[X_i X_j]/\sqrt{\mathbb{E}[X_i^2] \mathbb{E}[X_j^2]}\).

For two nodes \(w, \tilde{w}\) of a tree, the term \(\text{path}(w, \tilde{w})\) denotes the set of edges in the unique path with endpoints \(w\) and \(\tilde{w}\). Further, \(\text{BSC}(q)^p\) denotes a binary symmetric channel with crossover probability \(q\) and block-length \(p\). The \(\text{BSC}(q)^p\) is a conditional distribution from \(\{-1, 1\}^p \to \{-1, 1\}^p\) that acts componentwise independently on \(X\) to generate \(Y\), such that \(X_i = N_i Y_i\) and \(N\) is a vector of i.i.d. Rademacher variables equal to +1 with probability \(1 - q\). We use the symbol \(\dagger\) to indicate the corresponding quantity for the observable (noisy) layer. For instance, \(p_\dagger(\cdot)\) is the probability mass function of \(Y\) and \(\mu_{i,j}^\dagger \triangleq \mathbb{E}[Y_i Y_j]\) corresponds to the correlation of variables \(Y_i, Y_j\), where \(Y_i\) generates noisy observations of \(X_i\), for any \(i \in \mathcal{V}\). For our readers’ convenience, we summarize the notation in Table 1.1.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>number of variables nodes in the tree</td>
</tr>
<tr>
<td>$p(x)$</td>
<td>$\exp \left( \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right) / Z(\theta)$, $x \in {-1, +1}^p$, $Z(\theta)$: partition function</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>minimum $</td>
</tr>
<tr>
<td>$\beta$</td>
<td>maximum $</td>
</tr>
<tr>
<td>$T$</td>
<td>Original tree of the model</td>
</tr>
<tr>
<td>$\mathcal{P}_T(\alpha, \beta)$</td>
<td>set of tree-structured Ising models with $\alpha \leq</td>
</tr>
<tr>
<td>$n$</td>
<td>number of samples</td>
</tr>
<tr>
<td>$q$</td>
<td>crossover probability of the BSC, $q \in [0, 1/2)$</td>
</tr>
<tr>
<td>$c_q$</td>
<td>$1 - 2q$</td>
</tr>
<tr>
<td>$p(\cdot)$</td>
<td>distribution of the hidden node variables, $X \sim p(\cdot) \in \mathcal{P}_T(\alpha, \beta)$</td>
</tr>
<tr>
<td>$p_1(\cdot)$</td>
<td>distribution of the observable node variables $Y \sim p_1(\cdot)$</td>
</tr>
<tr>
<td>$1_A$</td>
<td>indicator function of the set $A$</td>
</tr>
<tr>
<td>$D_{KL}$</td>
<td>KL divergence</td>
</tr>
<tr>
<td>$S_{KL}$</td>
<td>symmetric KL divergence</td>
</tr>
<tr>
<td>$I(X,Y)$</td>
<td>mutual information of $X,Y$</td>
</tr>
<tr>
<td>$d_{TV}$</td>
<td>total variation distance</td>
</tr>
<tr>
<td>$\mathcal{L}^{(2)}(P,Q)$</td>
<td>$\sup_{i,j \in V} d_{TV}(P_{ij}, Q_{ij})$, and $P_{ij}, Q_{ij}$ the pairwise marginals of $P,Q$</td>
</tr>
<tr>
<td>$X^{1:n}$</td>
<td>$n$ independent observations of $X$</td>
</tr>
<tr>
<td>$Y^{1:n}$</td>
<td>$n$ independent observations of $Y$</td>
</tr>
<tr>
<td>$T_{CL}$</td>
<td>Chow-Liu-estimated structure from noiseless data $X^{1:n}$</td>
</tr>
<tr>
<td>$\hat{T}_{\text{CL}}$</td>
<td>Chow-Liu-estimate of the hidden tree structure $T$ from noisy data $Y^{1:n}$</td>
</tr>
<tr>
<td>path$_T(w, \tilde{w})$</td>
<td>the set of edges which connects the nodes $w, \tilde{w} \in V_T$</td>
</tr>
<tr>
<td>$\hat{\mu}_{i,j}$</td>
<td>$\frac{1}{n} \sum_{k=1}^n X_i^{(k)} X_j^{(k)}$</td>
</tr>
<tr>
<td>$\hat{\mu}_1^{i,j}$</td>
<td>$\frac{1}{n} \sum_{k=1}^n Y_i^{(k)} Y_j^{(k)}$</td>
</tr>
<tr>
<td>$\Pi_{T_{\text{CL}}}^1(\hat{p}_1)$</td>
<td>estimator of the distribution $p(\cdot)$ from noisy data $Y^{1:n}$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>maximum error on the distribution estimation: $\mathcal{L}^{(2)}(P, \hat{P}) \leq \eta$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>maximum probability of error, the notation depends on the task in structure estimation: $P(T_{\text{CL}}^1 \neq T) \leq \delta$ in predictive learning: $P \left( \mathcal{L}^{(2)}(p(\cdot), \Pi_{T_{\text{CL}}}^1(\hat{p}_1)) \leq \eta \right) \geq 1 - \delta$.</td>
</tr>
</tbody>
</table>

Table 1.1: Notation/Definitions.
1.5 Contributions

Learning Binary and Gaussian Tree-Structured Models from Noisy Data:
Recall that the first question that we are interested to answer is as follows: *What is the impact of observation noise on the sample complexity of learning a tree structured graphical model?* For the binary case, we sample variables $Y$ generated by $X$, which follows a tree-structured Ising model distribution and randomly flipping each sign independently with probability $q$. A typical example is classification, where a subset of the data might be misclassified. Then, corrupted data are observed, however, we are still able to retrieve the underlying structure by considering the appropriate number of samples. Under the Gaussian model assumption, we sample $Y$, the output of an AWGN channel with Gaussian input $X$. Similarly, our goal is to learn the hidden tree structure. The contributions of the results in Chapter 2 follow.

- A lower bound on the sufficient number of samples needed to recover the exact hidden structure with high probability, by using the Chow-Liu algorithm. Although the graphical model of the observables is a complete graph, we show that the Chow-Liu algorithm (with input a finite number of noisy samples) returns the exact tree of the hidden layer with high probability and we characterize its sample complexity.

- We prove an upper bound (Theorem 2) on the number of binary samples necessary for any algorithm to recover a tree structure. The proof of the upper bound uses the same construction of the approach in Section 7.1 by Bresler and Karzand [29] but requires the combination of Fano’s inequality and a strong data processing inequality (SDPI) by Polyanskiy and Wu [66]. Specifically, we show that SDPI’s can be a useful tool to derive minimax bounds when closed form expressions or upper bounds of the KL-divergence are hard to be found. The later is of independent interest and it can be applied to other machine learning problems that involve noisy observations.

- For the Gaussian case, we provide a lower bound (Theorem 3) on the number of
samples that are sufficient for the Chow-Liu algorithm to recover the underlying structure. This is a general result which also reduces to the noiseless case as the noise variance goes to zero. In particular, we show that the order of the necessary number of samples is \( \text{polylogarithmic in } p/\delta \), i.e., \( \mathcal{O}(\log^4(p/\delta)) \).

Our results strictly generalize the noiseless case \([29, \text{Theorem 3.1, Theorem 3.2}]\) for that of a hidden model. Our proof strategy is similar, but the hidden model presents additional complexity, which presents extended technical challenges, requiring the development of new arguments. In particular, the corresponding graph of the observable layer, \( Y \), is no longer a tree, so the Markov property does not hold for the observable nodes. In addition, closed-form bounds of the KL divergence can not be computed for the set of output distributions \( p_\hat{Y}(\cdot) \). To overcome this, we combine Bresler’s and Karzand’s method \([29]\) and a strong data processing inequality by Polyanskiy and Wu \([66]\), to derive an upper bound on the sample complexity.

The above results have been published in \([67]\).

**Predictive Learning on Hidden Tree-Structured Ising Models:**

In Chapter 3 we extend the results of prior on predictive learning by Bresler and Karzand \([29]\) to the noisy setting. The estimated structure together with the pairwise correlations of edge-pair variables are sufficient statistics for accurate distribution estimation of the hidden model. Similarly to the binary hidden model in Chapter 2, we sample variables \( Y \) generated by the hidden layer variable \( X \), which follows a tree-structured Ising model distribution, while the observables are the output of binary symmetric channels (randomly flipping each sign independently with probability \( q \)). A summary of contributions for the Chapter 3 follows.

- Determination of the sufficient and necessary number of samples for accurate predictive learning. We analyze the sample complexity of learning distribution estimates, which can accurately provide predictions on the hidden tree. The estimates are computed using the noisy data. The sample complexity analysis of predictive learning under noisy samples is challenging because structural properties
such as the independence of random variables \( X_i, X_j \) and correlation estimates \( \hat{E}[X_i X_j] \) for \((i, j) \in E\) do not hold for the noisy observable \( Y \). To overcome this, we evaluate the required conditional distributions of the dependent variables, construct a martingale difference sequence, and prove a high probability bound of the event that involves these variables by applying a concentration bound for supermartingales (generalized Bennet’s inequality [68]). We refer the reader to Section 3.3.1 for a detailed discussion about the technical contributions and a sketch of proof of the main result.

- A closed-form expression and a computationally efficient method for higher-order moment estimation in tree-structured Ising models. This result corresponds to an equivalent statement of Isserlis’ theorem for sign-valued tree models. Given pair-wise correlations and the tree (or estimates of both, from noisy or noiseless data) we provide an algorithm that runs on the tree and returns the expression of high-order moments. The proof involves the existence and identification of (minimum length) disjoint paths among any set of pairs of nodes. The proposed algorithm (Algorithm 2) identifies these paths that yield the expression of the moments. The results may be of independent interest for a computational efficient exact or approximated higher-moment evaluation.

The main results of Chapter 3 provide the amount of finite samples needed for exact structure recovery and accurate predictive learning with high probability. Although we are interested in the finite sample complexity bounds, our results are also asymptotically optimal. That is, for any fixed (constant) \( q \in [0, 1/2) \) the order of the upper bound (necessary number of samples) matches the corresponding (lower) minimax bound. The sample complexity bounds that we provide are the extended form of state of the art (noiseless setting) bounds by Bresler and Karzand [29]. By setting \( q = 0 \), our bounds reduce to the noiseless setting bounds. Further, the explicit version of our results (see Section 3.2) are continuous functions of the cross-over probability \( q \).

These results are reported as a preprint [69] and are currently under review.
Optimal Structure Learning for Non-Parametric Markov Trees: In Chapter 4 we study the problem of learning non-parametric hidden tree structures. The node variables are discrete, and for finite alphabet variables, there are no assumptions on the distribution. However for countable alphabet node variables, we have to consider certain assumptions on tails of distributions to guarantee convergence of the mutual information estimates. In this non-parametric setting the estimates of the mutual information for pairs of nodes act as edge weights, and we consider the classical Chow-Liu algorithm [3]. Our results include upper and lower sample complexity bounds for hidden tree structure learning, and these bound involve a fundamental quantity that we call information threshold $I^o$. The importance of the results lies in the fact that we can simply run the CL algorithm on the noisy observations in the same way as we would do if noiseless observations $X^{1:n}$ were given. Indeed, in a variety of cases the model of the noise might be unknown, or we might be unaware of the existence of the noise altogether; still, we can learn the hidden structure efficiently. Our contributions are as follows.

- We provide the first results on non-parametric hidden tree-structure learning, where the distributions of the hidden and observable layers are unknown and the mappings between the hidden and observed variables are general. In fact, as with the noiseless case, we show that, as long as $I^o > 0$, the sample complexity of the CL algorithm with respect to $p/\delta$ scales as $O\left(\log^{1+\zeta}(p/\delta)\right)$, for any $\zeta > 0$; also, with respect to the associated noisy information threshold, the same sample complexity is of the order of $O\left(1/(I^o)^{2(1+\zeta)}\right)$, for any $\zeta > 0$ (Theorem 12).
- For an absolute constant $C > 0$, we show that if the number of samples is less than $C/(I^o)^2$, then no algorithm can recover the hidden tree structure with probability greater than one half (Theorem 16). The rate of of upper and lower bounds with respect to $I^o$ are essential identical, and CL algorithm in the noisy setting is optimal up to a logarithmic factor.
- Additionally, we show that, if $I^o \leq 0$, then structure recovery from raw data is not possible. Still, whenever $I^o < 0$, by introducing suitable pre-processing on the noisy observations and enforcing appropriate conditions on the hidden model (Definition 5), we show that the CL algorithm can be made convergent (with the
aforementioned sample complexity). Such conditions can be satisfied for a variety of interesting observation models. Essentially, our results confirm that the CL algorithm is an effective universal estimator for tree-structure learning, on the basis of either noiseless or noisy data. We lastly explicitly illustrate how our results capture certain interesting scenarios involving generalized $M$-ary erasure and symmetric channels (i.e., observation noise models). Our framework extends and unifies recent results proved earlier for the binary case [5, 67], by considering non-identically distributed noise as well. Note that this latter problem remains open for general graph structures; see e.g., [70, Section 6].

These results are reported as a preprint in [71].
Chapter 2
Learning Binary and Gaussian Tree-Structured Models from Noisy Data

In this Chapter we study concurrently two distinct problem settings. First, we consider binary models on $2p$ variables $(X, Y)$, where the joint distribution $p(\cdot)$ of $X$ is a tree structured model distribution and $Y = (Y_1, Y_2, \ldots, Y_p)$ constitutes a noisy version of $X$. Specifically, we assume that $X$ follows a tree structured Ising model, whereas $Y$ is the output of a binary symmetric channel with crossover probability $q$, and input $X$. Under this setting, our objective is to exactly recover the underlying tree structure of the hidden layer $X$ by only using noisy observables $Y$. This is non-trivial, as $Y$ does not itself follow any tree structure.

We also consider the case where $X$ follows a Gaussian tree-structured distribution, and $Y$ is the output $X$ measured through an Additive White Gaussian Noise (AWGN) channel. This is similar to more traditional nonlinear filtering, where a Markov process of known distribution (and thus, of known structure) is observed through noisy measurements [72, 73, 74, 75, 76].

Under both settings, we use Chow-Liu algorithm [3] to reconstruct the hidden tree. The main results in this chapter are upper and lower sample complexity bounds for exact structure recover. The lower bounds (sufficient number of samples) characterize the sample complexity of Chow-Liu algorithm for both cases of Ising and Gaussian Models. For reader’s convenience we proceed by providing several properties and definitions for tree-structured models and also for tree-structured Ising models. These properties will be useful for the rest of dissertation.
2.1 Tree-Structured Ising Models, Preliminaries

In this section, we introduce our model of hidden sign-valued Markov random fields on trees.

2.1.1 Undirected Graphical Models

We consider sign-valued graphical models where the joint distribution $p(\cdot)$ has support $\{-1, +1\}^p$. Let $X = (X_1, X_2, \ldots, X_p) \in \{-1, +1\}^p$ be a collection of sign-valued (binary) random variables. Then, $1_{X_i = x_i} \equiv (1 + x_i X_i)/2$, and the distribution of $X$ is

$$p(x) = \mathbb{E} \left[ \prod_{i=1}^p 1_{X_i = x_i} \right] = \frac{1}{2^p} \left[ 1 + \sum_{k \in [p]} \sum_{S \subseteq \mathcal{V} : |S| = k} \mathbb{E} \left[ \prod_{s \in S} X_s \prod_{s \in S} x_s \right] \right], \quad x \in \{-1, +1\}^p. \tag{2.1}$$

When we refer to an Ising model, we assume that the marginal distributions of the $X_i$ are uniform for the rest of the dissertation, that is,

$$\mathbb{P}(X_i = \pm 1) = \frac{1}{2}, \quad \forall i \in \mathcal{V}. \tag{2.2}$$

Thus, $\mathbb{E}[X_i] = 0$, for all $i \in \mathcal{V}$. A distribution is Markov with respect to a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if for every node $i$ in the set $\mathcal{V}$ it is true that $\mathbb{P} \left( X_i | x_{\mathcal{V} \setminus \{i\}} \right) = \mathbb{P} \left( X_i | x_{\mathcal{N}(i)} \right)$, where $\mathcal{N}(i)$ is the set of neighbors of $i$ in $\mathcal{G}$. One subclass of distributions for which the Markov property holds is the Ising model, in which the random variables $X_i$ are sign-valued and the hypergraph is a simple undirected graph, indicating that variables have only pairwise and unary interactions. The joint distribution for the Ising model with zero external field is given by

$$p(x) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in \mathcal{E}} \theta_{st} x_s x_t \right\}, \quad x \in \{-1, +1\}^p. \tag{2.3}$$

$\{\theta_{st} : (s, t) \in \mathcal{E}\}$ are parameters of the model representing the interaction strength of the variables and $Z(\cdot) \in (0, \infty)$ is the partition function. These interactions are expressed...
through potential functions \( \exp(\theta_s x_s x_t) \) that ensure that the Markov property holds with respect to the graph \( G = (V, E) \). Next, we discuss the properties of distributions of the form of (2.1), which are Markov with respect to a tree.

### 2.1.2 Sign-Valued Markov Fields on Trees

From prior work by Lauritzen [2], it is known that any distribution \( p(\cdot) \) that is Markov with respect to a tree (or forest) \( T = (V, E) \) factorizes as

\[
p(x) = \prod_{i \in V} p(x_i) \prod_{(i,j) \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \quad x \in \{-1, +1\}^p;
\]

and we call \( p(\cdot) \) as tree (forest) structured distribution, to indicate the factorization property. If the distribution \( p(\cdot) \) has the form of (2.1) with \( P(X_i = \pm 1) = 1/2, \) for all \( i \in V \), and is Markov with respect to a tree \( T \), then

\[
p(x) = \frac{1}{2} \prod_{(i,j) \in E} (1 + x_i x_j \mathbb{E}[X_i X_j])
\]

and

\[
\mathbb{E}[X_i X_j] = \prod_{e \in \text{path}(i,j)} \mu_e, \quad \text{for all } i, j \in V.
\]

(see Appendix A, Lemma 6). Additionally, let us state the definition of the so-called Correlation (coefficient) Decay Property (CDP), that will be of central importance in our analysis.

**Definition 1.** The CDP holds if and only if \( |\mathbb{E}[X_i X_k]| \geq |\mathbb{E}[X_\ell X_m]| \) for all tuples \( \{i, k, \ell, m\} \subset V \) such that \( \text{path}(i, k) \subset \text{path}(\ell, m) \).

The CDP is a well known attribute of acyclic Markov fields (see, e.g., [77], [29]). Further, it is true that the products \( X_i X_j \) for all \( (i, j) \in E \) are independent and the CDP holds for every \( p(\cdot) \) of the form of (2.1), that factorizes with respect to a tree (see Lemma 7, Appendix A). This is a consequence of property (2.6) and the inequality \( |\mu_e| \leq 1, \) for all \( e \in E \). We can interpret the CDP as a type of data processing inequality.
(see Cover and Thomas [78]). The connection is clear through the relationship between the mutual information \( I(X_i, X_j) \) and the correlations \( \mathbb{E}[X_i X_j] \), namely,

\[
I(X_i, X_j) = \frac{1}{2} \log_2 \left( \left( 1 - \mathbb{E}[X_i X_j] \right)^{1-\mathbb{E}[X_i X_j]} \left( 1 + \mathbb{E}[X_i X_j] \right)^{1+\mathbb{E}[X_i X_j]} \right),
\]

(2.7)

for any pair of nodes \( i, j \in V \). This expression shows that the mutual information is a symmetric function of \( \mathbb{E}[X_i X_j] \) and increasing with respect to \( \mathbb{E}[X_i X_j] \) (see also Lemma 10, Appendix A).

**Tree-structured Ising models:** Despite its simple form, the Ising model has numerous useful properties. In particular, (2.5), (2.6) hold for any tree-structured Ising model with uniform marginal distributions. Furthermore,

\[
\mathbb{E}[X_i X_j] = \tanh \theta_{ij}, \quad \forall (i, j) \in E_T,
\]

(2.8)

equation (2.8) implies that

\[
p(x) = \frac{1}{2} \prod_{(i,j) \in E_T} \frac{1 + x_i x_j \tanh \theta_{ij}}{2}, \quad x \in \{-1, 1\}^p, \quad \alpha \leq |\theta_{ij}| \leq \beta,
\]

(2.9)

\[
\mathbb{E}[X_i X_j] = \prod_{e \in \text{path}(i, j)} \mu_e = \prod_{e \in \text{path}(i, j)} \tanh(\theta_e), \quad \forall i, j \in V.
\]

(2.10)

A short argument showing (2.8) and (2.9) is included in Appendix A, Lemma 8. For the rest of the dissertation, we assume a tree-structured Ising model for the hidden variable \( X \), that is, the distribution of \( X \) has the form of (2.5). We also impose a reasonable compactness assumption on the respective interaction parameters, as follows.

**Assumption 1.** There exist \( \alpha \) and \( \beta \) such that for the distribution \( p(\cdot) \), \( 0 < \alpha \leq |\theta_{st}| \leq \beta < \infty \) for all \( (s, t) \in E \).

For a fixed tree structure \( T \), and for future reference, we hereafter let \( \mathcal{P}_T(\alpha, \beta) \) be the class of Ising models satisfying Assumption 1. The assumption \( \alpha > 0 \) ensures that the tree is not disconnected.
2.1.3 Hidden Sign-Valued Tree-Structured Models

The problem considered in this chapter is that of learning a tree-structured model from corrupted observations. Because we have no access to the original samples \( \mathbf{X}^{1:n} \), we obtain the noisy observations \( \mathbf{Y}^{1:n} \). To formalize this, consider a hidden Markov random field whose hidden layer \( \mathbf{X} \) is an Ising model with respect to a tree, i.e., \( \mathbf{X} \sim p(\cdot) \in \mathcal{P}_T(\alpha,\beta) \), as defined in (2.9). The observed variables \( \mathbf{Y} \) are formed by setting \( Y_r = N_r X_r \) for all \( r \in \mathcal{V} \), where \( \{N_r\} \) are i.i.d. Rademacher(\( q \)) random variables.

Let \( p_\dagger(\cdot) \) be the distribution of the observed variables \( \mathbf{Y} \). We can think of \( \mathbf{Y} \) as the result of passing \( \mathbf{X} \) through a binary symmetric channel BSC(\( q \)) \( p \). We have the following expressions

\[
E[N_r] = 1 - 2q \triangleq c_q, \quad \forall r \in \mathcal{V}, \text{ and } q \in [0,1/2), \tag{2.11}
\]

\[
\mu^\dagger_{r,s} \triangleq E[Y_r Y_s] = E[N_r X_r N_s X_s] = (1 - 2q)^2 E[X_r X_s], \quad \forall r, s \in \mathcal{V}. \tag{2.12}
\]

The distribution \( p_\dagger(\cdot) \) of \( \mathbf{Y} \) also has support \( \{-1, +1\}^p \), and so the joint distribution satisfies the general form (2.1). Since the marginal distribution of each \( Y_r \) is also uniform, \( E[Y_r] = 0 \) for all \( r \in \mathcal{V} \), (2.1) and (2.11) yield

\[
p_\dagger(\mathbf{y}) = E \left[ \prod_{i=1}^{p} 1_{Y_i = y_i} \right] = \frac{1}{2^p} \left[ 1 + \sum_{k \in [p]\cap 2\mathbb{N}} c_q^k \sum_{S \subseteq \mathcal{V}, |S| = k} E \left[ \prod_{s \in S} X_s \right] \prod_{s \in S} y_s \right], \quad \mathbf{y} \in \{-1, +1\}^p. \tag{2.13}
\]

The moments of the hidden variables \( E \left[ \prod_{s \in S} X_s \right] \) in (2.13) can be expressed as products of the pairwise correlations \( E[X_s X_t] \), for any \( (s,t) \in \mathcal{E}_T \) (Section 3.2.2, Theorem 8).

From (2.13) it is clear that the distribution \( p_\dagger(\cdot) \) of \( \mathbf{Y} \) does not factorize with respect to any tree, that is, \( p_\dagger(\cdot) \notin \mathcal{P}_T(\alpha,\beta) \) in general.\(^1\)
Algorithm 1 Chow – Liu [3]

**Require:** \( D = \{ y(1), y(2), \ldots, y(n) \} \in \{-1, 1\}^{p \times n} \), where \( y(k) \) is the \( k \)th observation of \( Y \)

1. Compute \( \hat{\mu}^\dagger_{i,j} \leftarrow \frac{1}{n} \sum_{k=1}^{n} y_i^{(k)} y_j^{(k)} \), for all \( i, j \in V \)
2. return \( T_{\text{CL}}^\dagger \leftarrow \text{MaximumSpanningTree} \left( \bigcup_{i \neq j} |\{i,j\}| \right) \)

### 2.1.4 Hidden Structure Estimation

We are interested in characterizing the *sample complexity* of structure recovery: given data generated from \( p(\cdot) \in \mathcal{P}_T(\alpha, \beta) \) for an unknown tree \( T \), what is the minimum number \( n_t^\dagger \) of samples \( \{y(i), i \in [n_t^\dagger]\} \) from \( p_t(\cdot) \) needed to recover the (unweighted) edge set of \( T \) with high probability? In particular, we would like to quantify how \( n_t^\dagger \) depends on the crossover probability \( q \). Intuitively, noise makes “weak” edges to appear “weaker”, and the sample complexity is expected to be an increasing function of \( q \). Because the distribution \( p_t(\cdot) \) of the observable variables does not factorize according to any tree, this problem does not follow directly from the noiseless case. Although the classical MLE is the standard approach for the noiseless case, for the noisy setting the MLE estimation of parameters \( \theta \) of the hidden model is intractable, due to the summation over the support of \( X \). Additionally, the MLE structure estimate from noisy data is not in general consistent with the hidden structure as we explain in Section 2.1.6.

For the above reasons we consider the Chow-Liu algorithm. The algorithm we study is the classical Chow-Liu algorithm (Algorithm 1), which requires as input the set of noisy observations \( \{y(1), y(2), \ldots, y(n_t)\} \), computes the estimates \( \hat{\mu}^\dagger_{i,j} \), and returns a tree structure \( T_{\text{CL}}^\dagger \), that is an estimate of \( T \) by running a maximum spanning tree routine\(^2\). Additionally, for the models that we consider in our work, the projected-MLE estimate of the observables onto the space of tree-structured models gives a consistent structure estimate. Additionally, that structure estimate is identical to the output of

\(^1\)Lemma 17 shows the structure preserving property for the observable layer holds for the special case of single-edge forests.

\(^2\)The maximum spanning tree routine in Algorithm 1 is a greedy algorithm that returns the maximum spanning tree of an undirected graph with edge weights \( |E[X_iX_j]| \) for \( (i, j) \in V \). As an example see Kruskal’s algorithm that has computational complexity \( \mathcal{O}(|E| \log |V|) \).
Chow-Liu algorithm (Algorithm 1) from noisy data. We refer the reader to Section 2.1.6 for the discussion about the MLE and the connection with the noisy Chow-Liu algorithm.

In this work, we use and analyze the sample complexity of the classical Chow-Liu algorithm (Algorithm 1) for the following reasons: We show that given finite number of noisy data as input, the Chow-Liu algorithm recovers the original tree \( T \) with high probability. Further the sample complexity is asymptotically optimal for fixed \( q < \frac{1}{2} \) (see Tables 3.1 and 3.2), and its order remains \( \mathcal{O}(\log p) \) in the high dimensional regime. The algorithm is computationally efficient in comparison to other optimization techniques and it does not require the value \( q \) to be known. The above reasons and our finite sample complexity bound Theorems 4 and 1 suggest that Algorithm 1 is an excellent approach for tree-structure learning from noisy data.

### 2.1.5 Evaluating the Accuracy of the Estimated Distribution

In addition to recovering the graph structure, we are interested in the “goodness of fit” of the estimated distribution. We study this problem in Chapter 3. Let \( P_S, Q_S \) be the marginal distributions of \( P, Q \) on the set \( S \subset \mathcal{V} \), let \( d_{TV} \) denote the total variation distance, and fix \( k = 2 \). We measure the error of distribution estimator through the “small set Total Variation” (or ssTV) distance as defined by Bresler and Karzand [29]:

\[
\mathcal{L}^{(k)} (P, Q) \triangleq \sup_{S: |S| = k} d_{TV} (P_S, Q_S).
\] (2.14)

If \( Q \) is an estimate of \( P \), the norm \( \mathcal{L}^{(k)} \) guarantees predictive accuracy because [29, Section 3, page 720]

\[
\mathbb{E}_{X_S} \left[ \left| P \left( X_i = +1 | X_S \right) - Q \left( X_i = +1 | X_S \right) \right| \right] \leq 2 \mathcal{L}^{(|S|+1)} (P, Q).
\] (2.15)

The estimated (from noisy data) distribution of the hidden variables in (3.3) is a simple extension of the noiseless estimate. In fact the estimated distribution factorizes according to the estimated from noisy data tree structure, that is the output of Algorithm 1.
Further, the pairwise correlations are normalized by the constant \((1 - 2q)\). As a result, the estimator is consistent because if \(n \to \infty\) then \(T_{CL}^\dagger \to T\), \(\hat{\mu}_{i,j}/(1 - 2q) \to \mu_{i,j}\), and as a consequence the estimated distribution that factorizes according to \(T_{CL}^\dagger\) converges to the original distribution \(p(\cdot)\) of \(X\) (see Chapter 3). Our main result gives a lower bound on the number of samples needed to guarantee accurate estimation (in the sense of small ssTV), with high probability.

2.1.6 Maximum Likelihood Estimate

A natural first place to start in estimation is the maximum-likelihood estimate (MLE). We explain why this is problematic and show a method (the projected-MLE) which turns out to be equivalent to the Chow-Liu algorithm. This motivates why we study the Chow-Liu algorithm in the first place. To begin, the distribution of the observables parametrized over the interaction parameters \(\theta\) of the hidden layer is

\[
p_{\dagger}(y) = \sum_{x \in \{-1, +1\}^p} \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in E_G} \theta_{st} x_s x_t \right\} p(y|x), \quad y \in \{-1, 1\}^p. \tag{2.16}
\]

It is known that above expression is intractable in closed form and it can be evaluated only through approximations [79]. Secondly, the log-likelihood of \(Y\) can be written as

\[
\log p_{\dagger}(y) = \sum_{x \in \{-1, +1\}^p} p(y|x) \prod_{i \in V} p(x_i) \prod_{(i,j) \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \quad y \in \{-1, 1\}^p, \tag{2.17}
\]

and the logarithm of the summation cannot be expressed as summation of logarithms. Therefore we see the classical MLE structure estimation approach is not applicable for hidden models. Specifically, the structure of the observable layer is a complete graph and not a tree (there is no conditional independence between \(Y\’s\)). The maximum likelihood structure estimate with respect to the parameters \(\theta'\) of the observables in general will return a complete graph. Specifically, let \(G = (V, E_G)\) be the graph (which is complete) of the observable layer, then the distribution \(p_{\dagger}(\cdot)\) is an Ising-Model distribution and it
can be written as
\[
p_\dagger(y) = \frac{1}{Z'\left(\theta'\right)} \exp \left\{ \sum_{(s,t) \in E_G} \theta'_{st} y_s y_t \right\}, \quad y \in \{-1,1\}^p.
\] (2.18)

Since all the edges exist in the edge set, none of the values \(\theta'_{st}\) is zero. As a consequence, even asymptotically \((n \to \infty)\) the maximum likelihood estimates of the parameters \(\theta'_{st}\) gives a complete graph. Recall that we want to recover the structure of the hidden layer which is a tree. Thus, the maximum likelihood structure estimate directly applied on (2.18) is not consistent, because of the different hidden and observables’ structure.

To overcome the inconsistency that is introduced by the noise, we can project the distribution \(p_\dagger(y)\) to a set of tree-structured distributions and then find the maximum likelihood structure estimate. We denote the projection of \(p_\dagger(y)\) onto the space of trees as \(p_\dagger^T(y)\) and we call the MLE with respect to \(p_\dagger^T(y)\) as projected-MLE (PMLE). Then the following questions are natural: Is the PMLE always consistent with respect to structure of the hidden layer? Is the PMLE asymptotically optimal \((n \to \infty)\)? Is the PMLE optimal for finite values of \(n\)? (by optimal we mean that the sample complexity bound matches the minimax bound). We continue by answering the questions above.

First we present the structural consistency and then we continue by discussing the asymptotic optimality and optimality for finite \(n\).

Although, the PMLE is not in general consistent with structure of the hidden layer (see also related work by [80]), for the setting of the BSC channel with i.i.d noise we do have \(\hat{T}_{PMLE} \to T\) when \(n \to \infty\). In fact, the projected distribution as \(p_\dagger^T(y)\) is given by
\[
p_\dagger^T(y) \triangleq \arg\min_{Q(.) \in \mathcal{P}_T(\alpha,\beta)} D_{KL}(p_\dagger(y)||Q(y)).
\] (2.19)

For the rest of the paper we denote by \(H(.)\) and \(H_B(.)\) the entropy and the binary entropy respectively. The proof of the claim follows by a standard argument (see also Lemma 1 and Lemma 2 by [29, Supplementary material, Appendix A]) and it gives
\[
D_{KL}(p_\dagger(y)||p_\dagger^T) = 1 - H(p_\dagger(y)) + \sum_{(i,j) \in E} H_B \left( \frac{1 + (1 - 2q)^2 \mu_{i,j}}{2} \right).
\] (2.20)
As a consequence the projected-MLE $\hat{T}_{\text{PMLE}}$ is

$$\hat{T}_{\text{PMLE}} = \arg\min_{T \in \mathcal{T}} \sum_{(i,j) \in E_T} H_B \left( \frac{1 + \mu^\dagger_{i,j}}{2} \right) \equiv T^\dagger_{\text{CL}},$$

(2.21)

and the following

$$\arg\min_{T \in \mathcal{T}} \sum_{(i,j) \in E_T} H_B \left( \frac{1 + (1 - 2q)^2 \mu_{i,j}}{2} \right) \equiv \arg\min_{T \in \mathcal{T}} \sum_{(i,j) \in E_T} H_B \left( \frac{1 + \mu_{i,j}}{2} \right)$$

(2.22)

gives that $\hat{T}_{\text{PMLE}} \equiv T^\dagger_{\text{CL}} \rightarrow T$ (almost surely) when $n \rightarrow \infty$. Although, the above discussion of the consistency for $n \rightarrow \infty$ shows the connection with MLE, our results for instance Theorem 3.1 shows that the Chow-Liu algorithm returns the original tree for finite $n$ with probability $1 - \delta$. The proof of (2.21) follows.

**Proof of** $\hat{T}_{\text{PMLE}} = T^\dagger_{\text{CL}}$. Recall that $Y \sim p^\dagger(y)$, and the fact that $p^\dagger(y)$ does not factorize with respect to any tree. To derive a projected maximum likelihood estimate of the structure, first we project the distribution of the observables onto the space of tree-structured distributions (by minimizing the KL Divergence we evaluate the Reverse-Information projection) and then we find the MLE-structure. Further for any tree-structured distribution $Q(\cdot) \in \mathcal{P}_T(\alpha, \beta)$ the following holds

$$D_{KL}(p^\dagger(y)||Q(y)) = -E_{p^\dagger} \left[ \log \frac{p^\dagger(y)}{Q(y)} \right]$$

$$= -H(p^\dagger) - E_{p^\dagger} \left[ \log \frac{1}{2} \prod_{(i,j) \in E_Q} \frac{1 + y_iy_j \tanh \theta_{ij}^Q}{2} \right]$$

$$= \log 2 - H(p^\dagger) - \sum_{(i,j) \in E} E_{p^\dagger} \left[ \log Q(y_iy_j) \right]$$

$$= 1 - H(p^\dagger) + \sum_{(i,j) \in E} H_B(p^\dagger(y_iy_j)) + \sum_{(i,j) \in E} D_{KL}(p^\dagger(y_iy_j)||Q(y_iy_j)),$$

(2.23)

and $H_B(\cdot)$ is the binary entropy. Now by replacing $Q(\cdot)$ with $p^T_\dagger(y)$ given in (2.19), we
find
\[
D_{\text{KL}}(p^\dagger(y)||p^T) = 1 - H(p^\dagger(y)) + \sum_{(i,j)\in\mathcal{E}} H_B(p^\dagger(y_iy_j))
\]
\[
= 1 - H(p^\dagger(y)) + \sum_{(i,j)\in\mathcal{E}} H_B\left(\frac{1 + \mu_{i,j}^\dagger}{2}\right)
\]
\[
= 1 - H(p^\dagger(y)) + \sum_{(i,j)\in\mathcal{E}} H_B\left(\frac{1 + (1 - 2q)^2\mu_{i,j}}{2}\right).
\]

As a consequence the projected-MLE is
\[
\hat{T}_{\text{PMLE}} = \arg\min_{T \in \mathcal{T}} \sum_{(i,j)\in\mathcal{E}_T} H_B\left(\frac{1 + \hat{\mu}_{i,j}^\dagger}{2}\right) \equiv T_{\text{CL}}^\dagger.
\] (2.24)

Finally
\[
\arg\min_{T \in \mathcal{T}} \sum_{(i,j)\in\mathcal{E}_T} H_B\left(\frac{1 + (1 - 2q)^2\mu_{i,j}}{2}\right) \equiv \arg\min_{T \in \mathcal{T}} \sum_{(i,j)\in\mathcal{E}_T} H_B\left(\frac{1 + \mu_{i,j}}{2}\right) \quad (2.25)
\]
gives that \(\hat{T}_{\text{PMLE}} \equiv T_{\text{CL}}^\dagger \to T\) (almost surely) when \(n \to \infty\), and completes the proof.

Additionally, the PMLE is asymptotically optimal, however for finite \(n\) it may be not optimal. For our structure/predictive learning problem our bounds are asymptotically optimal (up to constants). That is, for fixed \(q\) the upper and lower bounds match as \(n \to \infty\). Nevertheless for finite \(n\) the PMLE is not optimal in general. It is known that under the presence of noise the MLE approach may be non-robust and sub-optimal and extra steps should be considered including pre-processing, statistical learning of the noise by using pilot samples, and detecting and rejecting bad samples (for further information see also Zoubir et al. [81, page 62] and Nikolakakis, Kalogerias, and Sarwate [80]). The reason that we consider Chow-Liu algorithm in our work is that it computationally efficient, while its sample complexity remains logarithmic with respect to \(p\) even when noise exists. The latter makes the Chow-Liu algorithm useful in practice when only noisy observations are available. Finally, to give further insight about the gap between
the upper and lower bounds we present an example in Section 3.5.8.1 (Appendix), for which perfect denoising is possible for \( p \to \infty \) before running the Chow-Liu algorithm. As consequence, for \( p \to \infty \) the bounds in Propositions 1 and 2 reduce to the noiseless case as they should. This example is a marginal case (since perfect denoising is not possible in general) and it affects our converse results which are universal and owe to include corner cases.

2.2 Models and Problem Statement

We start by presenting the models under consideration and their properties.

2.2.1 Hidden Tree-structured Ising Models

For the first structure learning problem of the dissertation we assume binary tree-structured data. Recall that Section 2.1 provides a detailed discussion on the hidden tree-structured Ising models with zero external field. Specifically, (2.3) is the distribution of the Ising model under general graphs. Under the assumption of a Markov tree structure the distribution is pasteurized over the correlations \( \mathbb{E}[X_iX_j] \) for \((i, j) \in E\) according to (2.4). The hidden layer is a tree-structured model in the set of distributions \( \mathcal{P}_T(\alpha, \beta) \) (see Assumption 1), while the observable variables are generated by a binary symmetric channel BSC\((q)p\), see Section 2.1.4.

2.2.2 Tree structured Gaussian models

The next family of distribution under consideration includes Gaussian tree-structured models. Let \( \mathbf{X} = (X_1, X_2, \ldots, X_p) \) be a Gaussian random vector with distribution \( \mathcal{N}(0, \Sigma) \). The nonzero entries of the precision matrix \( \Sigma^{-1} \) indicate the existence of the corresponding edge in the underlying graph. The following assumption holds on the Gaussian data.

**Assumption 2.** The variances of variables \( X_i \) are equal to 1, for all \( i \in V \). Furthermore,
there exist numbers $\rho_m, \rho_M$, such that

$$0 < \rho_m \leq |\rho_{i,j}| \leq \rho_M < 1, \quad \forall (i,j) \in \mathcal{E}.$$  \hspace{1cm} (2.26)

Hereafter, we use the notation $\mathcal{N}_T^{m,M}$ to denote the set of tree structured Gaussian distributions satisfying Assumption 2. We also use the notation $\mu_{i,j}$ for the correlation coefficient, since $\rho_{i,j} = \mathbb{E}[X_iX_j]$ under Assumption 2.

### 2.2.3 Hidden Gaussian Model

For the Gaussian setting, and for $\mathbf{N} \sim \mathcal{N}(0, \sigma^2 \mathcal{I})$, the noisy output variables of the hidden model are taken as $\mathbf{X} + \mathbf{N} = \tilde{\mathbf{Y}} \sim \mathcal{N}(0, \Sigma + \sigma^2 \mathcal{I})$. Then the correlation coefficient of the observable data is

$$\rho^\dagger_{i,j} = \mathbb{E} \left[ \frac{\tilde{Y}_i}{\sqrt{1 + \sigma^2}} \frac{\tilde{Y}_j}{\sqrt{1 + \sigma^2}} \right], \quad \forall i,j \in \mathcal{V}. \hspace{1cm} (2.27)$$

The random variables $Y_i \triangleq \tilde{Y}_i / \sqrt{1 + \sigma^2}$ are normalized Gaussian with variance equal to 1. To simplify the analysis, we use normalized samples, instead of samples directly from $\tilde{\mathbf{Y}}$, which correspond to the variable $\mathbf{Y}$ with distribution

$$\mathbf{Y} \sim \mathcal{N} \left( 0, \frac{\Sigma + \sigma^2 \mathcal{I}}{1 + \sigma^2} \right). \hspace{1cm} (2.28)$$

Thus, $\mathbb{E}[Y_iY_j] = \rho^\dagger_{i,j}$. For the rest of the dissertation we use the notation $\mu^\dagger_{i,j}$, where $\mu^\dagger_{i,j} \equiv \mathbb{E}[Y_iY_j]$ and its corresponding estimated value is denoted as $\hat{\mu}^\dagger_{i,j}$.

### 2.3 Main Results

The first main result of our work is presented below, providing a finite sample complexity bound on the sufficient number of samples guaranteeing exact structure recovery for the hidden Ising model under consideration.

Recall that our goal is to learn the tree structure $T$ of an Ising model with parameters $|\theta_{st}| \in [\alpha, \beta]$, when the nodes $X_i$ are hidden variables and we observe $Y_i \triangleq N_iX_i$, $i \in \mathcal{V}$,
where \(N_i \sim \text{Rademacher}(q)\) are i.i.d, for all \(i \in \mathcal{V}\) and for all \(q \in [0, 1/2)\). We derive the estimated structure \(T^\text{CL}_i\) by applying the Chow-Liu algorithm (Algorithm 1) \[3\].

Instead of mutual information estimates, our Chow-Liu algorithm (Algorithm 1) requires correlation estimates; these are sufficient statistics because of (2.7). Further, it can consistently recover the hidden structure through noisy observations. The latter is true because of the order preserving property of the mutual information. That is, the stochastic mapping \(X \xrightarrow{\text{BSC}(q)} Y\) allows structure recovery of \(X\) by observing \(Y\), because for any tuple \(X_i, X_j, X_i', X_j'\) such that \(I(X_i; X_j) \leq I(X_i', X_j')\), it is true that \(I(Y_i; Y_j) \leq I(Y_i', Y_j')\). The proof directly comes from (2.7) and (2.12). In addition, the monotonicity of mutual information with respect to the absolute values of correlations allows us to apply the Chow-Liu algorithm directly on the estimated correlations \(\hat{\mu}_{i,j}^1 \triangleq 1/n_i \sum_{k=1}^{n_i} (Y_i)^{(k)} (Y_j)^{(k)}\). Notice that because of (2.12), \(\hat{\mu}_{i,j}^1\) can be used as an alternative of \(\hat{\mu}_{i,j}\). The algorithm returns the maximum spanning tree \(T^\text{CL}_i\). The following theorem provides the sufficient number of samples for exact structure recovery through noisy observations.

**Theorem 1** (Sufficient number of samples for structure learning). Let \(Y\) be the output of a BSC(q)p, with input variable \(X \sim p(\cdot) \in P_T(\alpha, \beta)\). Fix a number \(\delta \in (0, 1)\). If the number of samples \(n_i\) of \(Y\) satisfies the inequality

\[
n_i \geq 32 \frac{\left[1 - (1 - 2q)^4 \tanh \beta\right]}{(1 - 2q)^4 (1 - \tanh \beta)^2 \tanh^2 \alpha} \log \frac{2p^2}{\delta},
\]

(2.29)

then Algorithm 1 returns \(T^\text{CL}_i = T\) with probability at least \(1 - \delta\).

Theorem 1 characterizes the finite-sample performance of the Chow-Liu estimator and by taking \(n \to \infty\) we can see that Algorithm 1 is consistent in the noisy setting. As a consequence of (2.29) and the inequality \(1 - \tanh(\beta) \geq e^{-2\beta}\), if the number of samples satisfies the following bound

\[
n > C \frac{e^{2\beta}}{\tanh^2(\alpha)} \left[ I_{q=0} + e^{2\beta} \left( (1 - 2q)^{-4} - \tanh(\beta) \right) I_{q \neq 0} \right] \log(p/\delta),
\]

(2.30)

then the structure is exactly recovered with probability at least \(1 - \delta\). The latter gives
the statement of Theorem up to constant factors (see Chapter 3 Theorem 4).

Complementary to Theorem 1, our next result characterizes the necessary number of samples required for exact structure recovery. Specifically, we prove a lower bound on the sample complexity that characterizes the necessary number of samples for any estimator $\psi$.

**Theorem 2** (Necessary number of samples for structure learning). Let $Y$ be the output of a BSC$(q)^p$, with input variable $X \sim p(\cdot) \in \mathcal{P}_T(\alpha, \beta)$. If the given number of samples of $Y$ satisfies the inequality

$$n^\dagger < \frac{1 - (4q(1 - q))^p} {16\alpha \tanh(\alpha)} e^{2\beta \log(p)} ,$$

then for any estimator $\psi$, it is true that

$$\inf_{\psi} \sup_{T \in \mathcal{T}} \sup_{p(\cdot) \in \mathcal{P}_T(\alpha, \beta)} \mathbb{P}\left( \psi(Y_{1:n^\dagger}) \neq T \right) > \frac{1}{2} .$$

It can be shown that the right hand-side of (2.29) is greater than the right-hand side of (2.31) for any $q$ in $[0, 1/2)$ (and for all possible values of $p, \beta, \alpha$), by simply comparing the two terms. Theorems 1 and 2 reduce to the noiseless setting by setting $q = 0$ (Bresler and Karzand [29]). The sample complexity is increasing with respect to $q$, and structure learning is always feasible as long as $q \neq 1/2$. Let $n$ denote the required samples under a noiseless setting assumption, then for a fixed probability of exact recovery, we always need $n^\dagger \geq n$ because

$$\frac{\left[ 1 - (1 - 2q)^4 \tanh(\beta) \right]} {\left[ (1 - 2q)^4 (1 - \tanh(\beta)) \right]} \geq 1 , \quad \forall q \in \left[ 0, \frac{1}{2} \right) \text{ and } \beta \in \mathbb{R} .$$

Furthermore,

$$\frac{1}{1 - (4q(1 - q))^p} \geq 1 , \quad \forall q \in [0, 1/2) \text{ and } p \in \mathbb{N} ,$$

the latter shows that the sample complexity in a hidden model is greater than the noiseless case ($q = 0$), for any measurable estimator (Theorem 2). When $q$ approaches
1/2, the sample complexity approaches infinity, \( n_\dagger \to \infty \), and structure learning becomes impossible. Theorem 2 extends Theorem 3.1 by Bresler and Karzand [29] to our hidden model. Our results combines Bresler’s and Karzand’s method and a strong data processing inequality (SDPI) by Polyanskiy and Wu [66, Evaluation of the BSC].

Upper bounds on the symmetric KL divergence for the output distribution \( p_\dagger(\cdot) \) cannot be found in a closed form. However, by using the SDPI, we manage to capture the dependence of the bound on the parameters \( \alpha, \beta, q \) and derive a non-trivial result. When \( p \to \infty \), the bound becomes trivial since \( \lim_{p \to \infty} 1/[1 - (4q(1-q))^p] \to 1 \), giving the classical data processing inequality (contraction of KL divergence for finite alphabets, [82, 66]). While direct application of the SDPI is simple and provides an upper bound which is almost insensitive to \( q \) (for sufficiently large \( p \)), it introduces a gap between the lower and upper bounds. Nevertheless, it is important because it indicates a possible non-optimal performance of the classical Chow-Liu algorithm (under a hidden model).

We conjecture that the sample complexity bounds in (Theorem 2 and Theorem 7) are tight only under the low temperature regime \( |\theta_{i,j}| = \beta \to \infty \) for all \( i, j \in \mathcal{E} \), while in general \( (\theta_{i,j} \in [\alpha, \beta]) \) the inequalities hold but they are not tight. For further explanation related to the gap between the upper and lower bounds see Section 2.1.6. The latter is a consequence of the SDPI, which is tight for the repetition code [66, Evaluation for the BSC, page 12]. We performed extensive simulations (c.f. Figures 3.1, 3.2) that suggests that our bound does indeed accurately characterize the performance of Chow-Liu. These simulations choose \( p = 100 \), but our evidence shows that the dependence on \( q \) is not affected for larger \( (p = 200) \) or smaller \( (p = 50) \) values of \( q \). We believe that the term \( 1/[1 - (4q(1-q))^p] \) does not characterize the Chow-Liu algorithm, but possibly a more complicated algorithm.

Our third result characterizes the sample complexity for achieving exact structure recovery, when only noisy observations from a Gaussian hidden model (as defined above) are available.

**Theorem 3.** Let \( \mathbf{Y} \) be the output of a Gaussian channel, \( \mathbf{Y} = \mathbf{X} + \mathbf{N} \), where \( \mathbf{X} \sim p(\cdot) \in \mathcal{N}^m_M \) and \( \mathbf{N} \sim \mathcal{N}(0, \sigma^2) \). Fix a number \( \delta \in (0,1) \). The Chow-Liu algorithm recovers the structure, \( \mathbf{T} = \mathbf{T}_CL^\dagger \) with probability at least \( 1 - \delta \), if the number of samples
satisfies the inequality

\[
n \geq \frac{R^2 \left[ 7(1 + \sigma^2)^2 + \rho M \right] \log^4 \left( \frac{e^2 p^2}{\delta} \right)}{\rho_n^2 (1 - \rho M)^2}, \tag{2.35}
\]

and \( R \) is a positive constant.

Theorem 3 gives a lower bound on the sufficient number of samples needed for exact structure recovery, for the case of the Gaussian model. The required amount of observations increases as the power of noise increases. For \( \sigma = 0 \), the bound provides the sample complexity of the corresponding noiseless setting, while for \( \sigma \to \infty \) the structure learning task becomes impossible.

### 2.4 Comparison with the Noiseless Setting

Theorems 1 and 2 strictly generalize noiseless tree-structure recovery [5, Theorem 3.1, Theorem 3.2] for our hidden model; the noiseless results correspond to \( q = 0 \). In particular, it very interesting to observe that, in the presence of noise, the dependence of our complexity bounds on \( p \) is still logarithmic, that is, of the order of \( \mathcal{O}(\log(p/\delta)) \). To make the connection between sufficient conditions more explicit, it is true that, in the noiseless case, if the weakest edge satisfies the inequality

\[
\tanh \alpha \geq \frac{4\epsilon}{\sqrt{1 - \tanh \beta}}, \tag{2.36}
\]

and \( \epsilon \) is defined as \( \epsilon \triangleq \sqrt{2 \log \left( \frac{2p^2}{\delta} \right)} / n \), yielding

\[
n \geq \frac{32}{\tanh^2 \alpha (1 - \tanh \beta)} \log \left( \frac{2p^2}{\delta} \right), \tag{2.37}
\]

then the structure is recovered exactly with probability \( 1 - \delta \) by the Chow-Liu algorithm. For our hidden model, the respective condition for the weakest edge is

\[
\tanh \alpha \geq \frac{4\epsilon \sqrt{1 - (1 - 2q)^4 \tanh \beta}}{(1 - 2q)^2 (1 - \tanh \beta)}, \tag{2.38}
\]
and \( \epsilon_\dagger \) is similarly defined as 
\[
\epsilon_\dagger \triangleq \sqrt{\frac{2 \log (2p^2/\delta)}{n_\dagger}}.
\]

Note that, for \( q = 1/2 \), the mutual information of the hidden and observable variables is zero, that is, \( \textbf{X} \) and \( \textbf{Y} \) are independent, so structure recovery is impossible.

To make the relevant connection between necessary conditions, it holds that [29], in the noiseless case, if the number of samples satisfies
\[
n < \frac{1}{16} e^{2\beta \alpha^{-2} \log (p)},
\]
then for any (measurable) algorithmic mapping \( \psi \), it is true that
\[
\inf_{\psi} \sup_{T \in T} \sup_{P \in \mathcal{P}_{T}(\alpha,\beta)} \mathbb{P}(\psi((X_{1:n})) \neq T) > \frac{1}{2}.
\]

When \( q = 0 \), we retrieve the noiseless result, while for any \( q \in (0, 1/2) \) the sample complexity increases since \( [1 - (4q(1-q)p)^{-1}] > 1 \) in (2.31) and for \( q \to 1/2 \) the required number of samples \( n_\dagger \to \infty \) which makes structure learning impossible. The ratio between the noiseless and noisy necessary conditions indicates the gap between the hidden model and the noiseless one
\[
\frac{n_\dagger}{n} \leq [1 - (4q(1-q)p)^{-1}] \leq \frac{1}{\eta_{KL}},
\]
where the right hand-side inequality is the strong data processing inequality for the binary symmetric channel by Polyanskiy and Wu [66, Equation (39)].

As far as Theorem 3 is concerned, this reduces to the noiseless setting for \( \sigma = 0 \). Recently, the performance of the Chow-Liu algorithm for the noiseless Gaussian case was studied by Tavassolipour [83]. In this work, a lower complexity bound is derived, closely resembling the noiseless Ising model. The approach of [83] might potentially drive further improvement of our hidden Gaussian model bound (Theorem 3), and is the subject of our future work.
2.5 Analysis: Proof Sketches

Herein, we provide a sketch of the proof for each result of this Chapter. The definition of necessary events and the complete proofs can be found in Section 2.9 (Appendix).

**Theorem 1.** To analyze the Chow-Liu algorithm, we consider the error event \([5]\), at least one edge to be missed; if an edge \(f = (w, \bar{w}) \in T\) and \(f \notin T^{\text{CL}}_1\) (i.e. the edge is incorrectly not inferred), then there exists an edge \(g \in T^{\text{CL}}_1\) and \(g \notin T\) such that \(f \in \text{path}_T(u, \bar{u}), g \in \text{path}_{T^{\text{CL}}_1}(w, \bar{w})\), and

\[
\left( \sum_{i=1}^{n_+} Z_{f,u,\bar{u}}^{(i)} \right) \left( \sum_{i=1}^{n_+} M_{f,u,\bar{u}}^{(i)} \right) < 0, \tag{2.42}
\]

where \(Z_{f,u,\bar{u}} \triangleq Y_w Y_{\bar{w}} - Y_u Y_{\bar{u}}\) and \(M_{f,u,\bar{u}} \triangleq Y_w Y_{\bar{w}} + Y_u Y_{\bar{u}}\). Thus, to show that the reconstruction is successful, we need to show that this event does not happen, with high probability.

By using Bresler and Karzand’s method \([84, \text{Lemmas 9.6, 9.7}]\) under the error event (at least one incorrect edge in the estimated tree structure \(T^{\text{CL}}_1\)) we have \(\left| \hat{\mu}_f \right| \leq \left| \hat{\mu}_g \right|\), which gives

\[
0 \geq \left| \hat{\mu}_f \right|^2 - \left| \hat{\mu}_g \right|^2
= \left( \hat{\mu}_f - \hat{\mu}_g \right) \left( \hat{\mu}_f + \hat{\mu}_g \right)
= \frac{1}{n_+^2} \left( \sum_{i=1}^{n_+} \left( N_{w}^{(i)} X_{w}^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} - N_{u}^{(i)} X_{u}^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)} \right) \right)
\times \left( \sum_{i=1}^{n_+} \left( N_{w}^{(i)} X_{w}^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} + N_{w}^{(i)} X_{u}^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)} \right) \right)
= \frac{1}{n_+^2} \left( \sum_{i=1}^{n_+} Z_{f,u,\bar{u}}^{(i)} \right) \left( \sum_{i=1}^{n_+} M_{f,u,\bar{u}}^{(i)} \right), \tag{2.43}
\]

Notice that the random variables \(Z_{f,u,\bar{u}}^{(i)}, M_{f,u,\bar{u}}^{(i)}\) are functions of noisy observables. To understand how these quantities behave we use Bernstein’s inequality, which produces a factor of \((1 - 2q)^2\) to account for the variance in the noisy samples. The concentration
of measure results we need are for $Z_f^{(i)}$, $M_f^{(i)}$. Defining events $E_Z$ and $E_M$ as

$$E_Z \triangleq \bigcap_{(w, \bar{w}) \in \mathcal{E}, u, \bar{u} \in \mathcal{V}} E_Z^{(w, \bar{w}), u, \bar{u}}, \quad (2.44)$$

$$E_M \triangleq \bigcap_{(w, \bar{w}) \in \mathcal{E}, u, \bar{u} \in \mathcal{V}} E_M^{(w, \bar{w}), u, \bar{u}}, \quad (2.45)$$

and

$$E_Z^{(w, \bar{w}), u, \bar{u}} \triangleq \left\{ \left| \frac{1}{n_1^-} \sum_{i=1}^{n_1^-} Z_e^{(i)} - \mathbb{E} [ Z_e, u, \bar{u}] \right| \right. \leq \max \left\{ 8 \epsilon_1^2, 4 \epsilon_1 \sqrt{1 - \mu_A^1} \right\}, \quad (2.46)$$

$$E_M^{(w, \bar{w}), u, \bar{u}} \triangleq \left\{ \left| \frac{1}{n_1^-} \sum_{i=1}^{n_1^-} M_e^{(i)} - \mathbb{E} [ M_e, u, \bar{u}] \right| \right. \leq \max \left\{ 8 \epsilon_1^2, 4 \epsilon_1 \sqrt{1 + \mu_A^1} \right\}, \quad (2.47)$$

it is possible to show that each occurs with probability at least $1 - \delta'/2$ and $1 - \delta''/2$ respectively, where $\epsilon_1 = \sqrt{2/n_1 \log (2p^2/\delta)}$ and $A = \text{path}_1 (u, \bar{u}) \setminus \{e\}$. A union bound over all pairs $w, \bar{w}, u, \bar{u}$ and finally for the events $E_Z$, $E_M$ shows that the event $E_Z \cup E_M$ happens with probability at least $1 - \delta$, where $\delta'/2 + \delta''/2 \leq 2 \max \{ \delta'/2, \delta''/2 \} \triangleq \delta$.

**Theorem 2.** To show a (minimax) lower bound on tree structure estimation we follow the standard information-theoretic recipe using Fano’s inequality [85, Corollary 2.6]. As for noiseless models [84, Section 8.1], we consider difficult instances of the problem correspond to graphs which are nearly chains, see also Section A.2 (Appendix). First, we define $P_{\theta^0}$ to be an Ising model distribution with underlying structure a chain with $p$ nodes and parameters $\theta^0_{j, j+1} = \alpha$ when $j$ is odd and $\theta^0_{j, j+1} = \beta$, when $j$ is even. The rest of family is constructed as follows: the elements of each $\theta^i$ are equal to the elements of $\theta^0$ apart from two elements $\theta^i_{i, i+1} = 0$ and $\theta^i_{i, i+2} = \alpha$ for each odd value of $i$. There are $(p + 1)/2$ distributions in the constructed family. To find a non-trivial upper bound for the quantity $S_{KL}(P_{\theta^i}^i || P_{\theta^i}^i)$, (where $P_{\theta^i}^i$ is the distribution of the observable variables of the $i$th model) we require new techniques, namely the strong data processing inequality
Consider our hidden model in which \( X \) is drawn from \( p(\cdot) \in \mathcal{P}_T(\alpha, \beta) \), corrupted by multiplicative Rademacher(\( q \)) noise, \( N_i \). We have [66]:

\[
\eta_{\text{KL}} \triangleq \sup_Q \sup_{P:0 < D_{\text{KL}}(P||Q) < \infty} \frac{D_{\text{KL}} \left( P_{Y|X} \circ P \parallel P_{Y|X} \circ Q \right)}{D_{\text{KL}}(P||Q)} \leq 1 - (4q(1 - q))^p. \tag{2.48}
\]

We combine this with an upper bound on the symmetric KL divergence between any pair of noiseless models in a specially constructed set of \( M + 1 \) trees:

\[
S_{\text{KL}}(P_{\theta^i} || P_{\theta^0}) \leq 4\alpha^2 e^{-2\beta}, \quad \forall i \in [M]. \tag{2.49}
\]

This in turn yields the factor \( 1 - (4q(1 - q))^p \) in the final bound.

**Theorem 3.** The proof of Theorem 3 differs from that of Theorem 1 at two points. First, the correlation decay property holds as a correlation coefficient decay property, which makes the normalization of \( \tilde{Y} \) an essential step for the analysis (see (2.28)). The correlation coefficient decay property [4] can be stated as follows. If \( X \sim p(\cdot) \in \mathcal{N}_T^{m,M} \), then

\[
\rho_{i,j} = \prod_{e \in \text{path}_T(i,j)} \rho_e, \quad \forall (i,j) \in \mathcal{V}, \tag{2.50}
\]

where \( \rho_{i,j} = \mathbb{E}[X_i X_j]/\sqrt{\mathbb{E}[X_i^2] \mathbb{E}[X_j^2]} \). Furthermore, to bound the probability of the events analogous to (2.44) and (2.45) in the Gaussian case from above, we need concentration of measure inequalities for polynomials of dependent continuous random variables. For that purpose, we use a recent concentration result by Shudi and Sviridenko [86, Theorem 1.10]. This results in the polylogarithmic sample complexity in the Gaussian Case.
2.6 Experiments

For the experimental part we consider synthetic data. To demonstrate the performance of the Chow-Liu algorithm experimentally, we present the decay of the probability of incorrect recovery (Fig. 2.3), while the number of samples increases, for fixed values of the parameters $\alpha, \beta, p, \rho_M, \rho_m$. These results illustrate how noisy observations can degrade performance, unless we increase the sample size. Based on the experiments, more observations are required for the Gaussian model than the Ising model to provide an estimate $T^\dagger_{CL}$ with small $\mathbb{P}\left( T^\dagger_{CL} \neq T \right)$ (compared under equivalent noise levels). The probability (Figures 2.2 and 2.3) approaches zero with less observations in the case of the Ising model than the Gaussian model, for instance compare the lines for $q = 0$, which corresponds to $\text{SNR} = \infty$. We define the distance between two tree structures $T = (V, E), T' = (V, E')$ with identical node set and possibly different edge sets as

$$D_T(T, T') \triangleq \frac{|E \Delta E'|}{2},$$  \hspace{1cm} (2.51)

where the symbol $\Delta$ denotes the symmetric difference between two sets. Note that $0 \leq D_T(T, T') \leq \max\{|E|, |E'|\}$. In particular, $D_T(T, T^\dagger_{CL})$ counts the number of incorrect edges for the estimated structure. Similar metrics can be found in the literature. A closely related one is "false positive and false negative rates", which has been considered by Liu et al. [87].

**Synthetic Data.** To demonstrate the performance of the algorithm experimentally, we present the decay of the error based on the metric $D_T(T, T^\dagger_{CL})$ (Figure 2.1) and the probability of the error event $\{T \neq T^\dagger_{CL}\}$ (Figure 2.2), while the number of samples increases, for fixed values of the parameters $\alpha, \beta, p$, while the crossover probability $q$ varies between 0 and 1/2. Specifically, for the plots in Figures 2.1 and 2.2, we have chosen $\alpha = \text{arctanh}(0.25)$, $\beta = \text{arctanh}(0.75)$, $p = 100$. These results illustrate how noisy observations can significantly degrade performance unless we increase the sample size significantly. We consider synthetic Gaussian data for the plots of Figure 2.3. These show how the error $D_T(T, T^\dagger_{CL})$ and the probability of the not exact recovery $\{T \neq T^\dagger_{CL}\}$ varies as the number of observations increases and for different values of the signal to
noise ration (SNR).

**Real Data.** We consider as observations the increase (spin up) or decrease (spin down) of the closing prices for 10 stocks. The estimated tree structure $T_{\text{CL}}$ is found by applying Chow-Liu’s algorithm, Figure 2.4. Noisy data are generated by flipping each observation with probability $q$. Then the structure $T_{\text{CL}}^\dagger$ is estimated by taking into consideration (semi-synthetic) noisy data. The error $D_T(T_{\text{CL}}, T_{\text{CL}}^\dagger)$ is plotted as function of $q$ in figure 2.4. Notice that for hidden model structure estimates, where $q \in (0, 1/2)$, we see that small noise levels lead to a modest increase in sample complexity for a target error probability, but as the channel gets worse, the sample complexity explodes.

### 2.7 Connections with Differential Privacy

One way in which a hidden model can arise is in inference from data released under differential privacy. Suppose that data about individuals can be modeled as drawn from an Ising model: the $j$-th sample from the population has data $X(j)$ drawn according to $p(\cdot)$ representing $p$ correlated features characterizing the individual. Because of privacy concerns, the analyst is only given access to $Y(j)$, where each feature is randomly flipped with probability $q$. The noisy data guarantees differential privacy [88]: we can think of this process as a form of vectorized randomized response. More formally, the noisy samples guarantees $\epsilon$-differential privacy if for all $c, c', c'' \in \{0, 1\}^p$,

$$\frac{P(Y = c|X = c')}{P(Y = c|X = c'')} \leq e^\epsilon. \quad (2.52)$$

For our choice of $q$,

$$\frac{P(Y = c|X = c')}{P(Y = c|X = c'')} = \frac{(1-q)^{p-\ell}q^\ell}{(1-q)^{p-\ell'}q^{\ell'}} = \left[ \frac{1-q}{q} \right]^{\ell'-\ell},$$

where $\ell, \ell'$ is the number of different elements of the pairs $c, c'$ and $c, c''$ respectively, for any $c, c', c'' \in \{-1, +1\}^p$. Since $\ell, \ell \in \{1, 2, \ldots, p\}$ and $q \in [0, 1/2]$, we may write

$$\frac{P(Y = c|X = c')}{P(Y = c|X = c'')} \leq \max_{\ell, \ell'} \left[ \frac{1-q}{q} \right]^{\ell'-\ell} = \left[ \frac{1-q}{q} \right]^p. \quad (2.53)$$
Figure 2.1: Estimate of the number of mismatched edges $D_T(T, T^{CL}_T)$ as a function of number of samples. The upper graph is over 1000 independent runs and up to $10^4$ independent samples, while the down over 100 independent runs and up to $10^5$ independent samples.
Figure 2.2: Ising Model data. Estimating the probability of the error event, $P(T_{\text{CL}} \neq T)$, as a function of number of samples. Upper graph: 100 independent runs. Down: 1000 independent runs.
Figure 2.3: Gaussian Data. Upper: $D_T(T, T_{\text{CL}}^*)$ as a function of number of samples. Down: Estimating the probability of the error event $T_{\text{CL}}^* \neq T$. Both estimates are evaluated through 1000 independent iterations.
Thus for $\epsilon_o = p \log \left( (1 - q)/q \right)$ we guarantee $\epsilon_o$-local differential privacy.

We can interpret the results of our work, in terms of differential privacy, as characterizing the trade-off between privacy and sample complexity in inference from data protected by differential privacy. In this simplified mechanism, however, each individual data sample is perturbed, which is a form of local differential privacy or (alternatively) input perturbation. An interesting question would be the trade-off in standard differential privacy, where the algorithm releases only the estimated tree.

2.8 Conclusion

We have analyzed the problem of perfect reconstruction for hidden tree-structures from noisy observations, using the well-known Chow-Liu algorithm. In particular, we have focused on two distinct cases, namely, hidden Ising models observed in multiplicative $\pm 1$ binary noise, and hidden Gaussian graphical models observed in additive Gaussian noise. For the case of a hidden Ising model, our results (Theorem 1 and Theorem 2) give lower and upper bounds on the sample complexity of accurately inferring the latent tree structure, strictly generalizing the previously-studied noiseless case. In particular, the lower bound shows that the number of samples needed to estimate the tree grows only as $O(\log(p/\delta))$, where $\delta > 0$ is the probability of incorrect recovery. For hidden
Gaussian models, we provide an extension of our method, and derive a lower bound on the number of sufficient samples for exact structure recovery, which is polylogarithmic in $p/\delta$ (Theorem 3). Experiments illustrate the impact of the noise and how properly accounting for noisy samples can lead to more accurate structure inference.
2.9 Appendix

2.9.1 Proof of Theorem 1 (Sufficient number of samples)

Lemmas 1, 2, 3 together with Lemma 4 give the required number of samples for exact structure recovery when observations from a hidden model are given. To analyze the error event we use the "Two Trees Lemma" of Bresler and Karzand \[84, Lemmas 10.1, 10.2\]. For two different spanning trees on same set of nodes. Informally, if two maximum spanning trees $T, T'$ have a pair of nodes connected in a different way then there exist at least one edge in $\mathcal{E}_T$ which does not exist in $\mathcal{E}_{T'}$ and vice versa.

**Lemma 1.** Let $f = (w, \bar{w})$ be an edge such that $f \in T$ and $f \notin T_{CL}^\dagger$. Then there exists an edge $g \in T_{CL}^\dagger$ and $g \notin T$ such that $f \in \text{path}_T(u, \bar{u})$ and $g \in \text{path}_{T_{CL}^\dagger}(w, \bar{w})$ and the following holds under the error event $T \neq T_{CL}^\dagger$:

$$
\left(\sum_{i=1}^{n^\dagger} Z_{f,u,\bar{u}}^{(i)}\right) \left(\sum_{i=1}^{n^\dagger} M_{f,u,\bar{u}}^{(i)}\right) < 0, \quad (2.54)
$$

where $Z_{f,u,\bar{u}} = Y_w Y_{\bar{w}} - Y_u Y_{\bar{u}}$ and $M_{f,u,\bar{u}} = Y_w Y_{\bar{w}} + Y_u Y_{\bar{u}}$.

**Proof.** Using the same argument as the noiseless case \[84, Lemmas 9.6, 9.7\] we see that $|\hat{\mu}_f^\dagger| \leq |\hat{\mu}_g^\dagger|$ implies

$$
0 \geq |\hat{\mu}_f^\dagger|^2 - |\hat{\mu}_g^\dagger|^2
= \left(\hat{\mu}_f^\dagger - \hat{\mu}_g^\dagger\right)\left(\hat{\mu}_f^\dagger + \hat{\mu}_g^\dagger\right)
= \frac{1}{n^\dagger} \left(\sum_{i=1}^{n^\dagger} N_{w}^{(i)} X_{w}^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} - N_{u}^{(i)} X_{u}^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)}\right)
\times \left(\sum_{i=1}^{n^\dagger} N_{w}^{(i)} X_{w}^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)} + N_{u}^{(i)} X_{u}^{(i)} N_{\bar{u}}^{(i)} X_{\bar{u}}^{(i)}\right)
= \frac{1}{n^\dagger} \left(\sum_{i=1}^{n^\dagger} Z_{f,u,\bar{u}}^{(i)}\right) \left(\sum_{i=1}^{n^\dagger} M_{f,u,\bar{u}}^{(i)}\right). \quad (2.55)
$$

$\square$
Notice that the random variables $Z_{f,u,\bar{u}}^{(i)}$, $M_{f,u,\bar{u}}^{(i)}$ are functions of observations of the observable variables (noisy observations). These differ from the corresponding terms in the noiseless case and require a new analysis.

In Lemmas 2, 3, we derive two concentration of measure inequalities for the variables $Z_{f,u,\bar{u}}^{(i)}$ and $M_{f,u,\bar{u}}^{(i)}$. In fact, we have that the event $E_Z$ in (2.44) as

$$E_Z \triangleq \bigcap_{(w,\bar{w}) \in E, u, \bar{u} \in \mathcal{V}} E_{Z_{(w,\bar{w})},u,\bar{u}}^{(w,\bar{w})},$$

(2.56)

and

$$E_{Z_{(w,\bar{w})},u,\bar{u}}^{(w,\bar{w})},u,\bar{u} \triangleq \left\{ \left| \frac{1}{n^\dagger} \sum_{i=1}^{n^\dagger} Z_{e,u,\bar{u}}^{(i)} - \mathbb{E} \left[ Z_{e,u,\bar{u}} \right] \right| \leq \max \left\{ 8\epsilon^2, 4\epsilon^\dagger \sqrt{1 - \mu_A^\dagger} \right\} \right\},$$

(2.57)

happens with probability at least $1 - \delta^2$ and the event $E_M$, which is defined as

$$E_M \triangleq \bigcap_{(w,\bar{w}) \in E, u, \bar{u} \in \mathcal{V}} E_{M_{(w,\bar{w})},u,\bar{u}}^{(w,\bar{w})},$$

(2.58)

and

$$E_{M_{(w,\bar{w})},u,\bar{u}}^{(w,\bar{w})},u,\bar{u} \triangleq \left\{ \left| \frac{1}{n^\dagger} \sum_{i=1}^{n^\dagger} M_{e,u,\bar{u}}^{(i)} - \mathbb{E} \left[ M_{e,u,\bar{u}} \right] \right| \leq \max \left\{ 8\epsilon^2, 4\epsilon^\dagger \sqrt{1 + \mu_A^\dagger} \right\} \right\}$$

(2.59)

happens with probability at least $1 - \delta'$. The threshold variable $\epsilon^\dagger$ is a decreasing function of $n^\dagger$, both $\epsilon^\dagger$, $\mu_A$, which are defined below. Finally, we apply union bound to guarantee that the event $E_Z \cup E_M$ happens with probability at least $1 - \delta$, where

$$\frac{\delta'}{2} + \frac{\delta''}{2} \leq 2 \max \left\{ \frac{\delta'}{2}, \frac{\delta''}{2} \right\} \triangleq \delta.$$  

Then, we can apply the union bound over all pairs $w, \bar{w}, u, \bar{u}$ in Lemmas 2 and 3 and finally for the events $E_Z$ and $E_M$.

**Lemma 2.** For all pairs of vertices $u, \bar{u} \in \mathcal{V}$ and edges $e = (w, \bar{w})$ in the path $\mathsf{path}_T(u, \bar{u})$ from $u$ to $\bar{u}$, given $n^\dagger$ samples $Z_{e,u,\bar{u}}^{(1)}, Z_{e,u,\bar{u}}^{(2)}, ..., Z_{e,u,\bar{u}}^{(n^\dagger)}$ of $Z_{e,u,\bar{u}} = Y_wY_{\bar{w}} - Y_uY_{\bar{u}}$ we have

$$\mathbb{P} \left( \left| \sum_{i=1}^{n^\dagger} Z_{e,u,\bar{u}}^{(i)} - n^\dagger \mathbb{E} \left[ Z_{e,u,\bar{u}} \right] \right| \leq n^\dagger \max \left\{ 8\epsilon^2, 4\epsilon^\dagger \sqrt{1 - \mu_A^\dagger} \right\} \right) \geq 1 - \frac{\delta}{2},$$
and \( \epsilon^* = \sqrt{2/n^* \log (2p^2/\delta)} \) and \( A = \text{path}_T (u, \bar{u}) \setminus \{e\} \).

**Proof.** The proof is based on Bernstein’s inequality [89]. Expanding the definition of \( Z_{e,u,\bar{u}} \),

\[
Z_{e,u,\bar{u}} = X_{w}N_{w}X_{\bar{w}}N_{\bar{w}} - N_{w}X_{u}N_{\bar{u}}X_{\bar{u}} \tag{2.60}
\]

\[
= N_{w}X_{w}X_{\bar{w}} (1 - N_{w}X_{w}N_{\bar{w}}X_{u}N_{\bar{u}}X_{\bar{u}}). \tag{2.61}
\]

Then

\[
\mathbb{E} [Z_{e,u,\bar{u}}] = (1 - 2q)^2 \mathbb{E} [X_{w}X_{\bar{w}} - X_{u}X_{\bar{u}}] = (1 - 2q)^2 \mu_e (1 - \mu_A) \tag{2.62}
\]

and

\[
\text{Var} (Z_{e,u,\bar{u}})
\]

\[
= \mathbb{E} \left[ (Z_{e,u,\bar{u}})^2 \right] - \mathbb{E} \left[ (Z_{e,u,\bar{u}}) \right]^2
\]

\[
= \mathbb{E} \left[ (X_{w}N_{w}X_{\bar{w}}N_{\bar{w}} - N_{w}X_{u}N_{\bar{u}}X_{\bar{u}})^2 \right] - \left[ (1 - 2q)^2 \mathbb{E} [X_{w}X_{\bar{w}} - X_{u}X_{\bar{u}}] \right]^2
\]

\[
= \mathbb{E} [1 + 2X_{w}N_{w}X_{\bar{w}}N_{\bar{w}}X_{u}N_{\bar{u}}X_{\bar{u}} - (1 - 2q)^4 \mathbb{E} [X_{w}X_{\bar{w}} - X_{u}X_{\bar{u}}]^2
\]

\[
= 2 - 2 (1 - 2q)^4 \mathbb{E} [X_{w}X_{\bar{w}}X_{u}X_{\bar{u}}] - (1 - 2q)^4 \mathbb{E} [X_{w}X_{\bar{w}} - X_{u}X_{\bar{u}}]^2
\]

\[
= 2 - 2 (1 - 2q)^4 \mu_A - (1 - 2q)^4 \left( \mu_e (1 - \mu_A) \right)^2
\]

\[
= 2 - (1 - 2q)^4 \left[ 2\mu_A + \mu_e^2 (1 - \mu_A)^2 \right]. \tag{2.63}
\]

Using the expressions for the mean and the variance, we apply Bernstein’s inequality [89] for the noisy setting: for all \( i \in [n^*] \) we have

\[
\left| Z_{e,u,\bar{u}}^{(i)} - \mathbb{E} [Z_{e,u,\bar{u}}] \right| \leq M \text{ almost surely,}
\]
that is, for any $t > 0$

$$P \left[ \sum_{i=1}^{n^*} Z_{e,u,\bar{u}}^{(i)} - n^*E \left[ Z_{e,u,\bar{u}} \right] \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) + \frac{2}{3} Mt} \right), \tag{2.64}$$

$$P \left[ \sum_{i=1}^{n^*} Z_{e,u,\bar{u}}^{(i)} - n^*E \left[ Z_{e,u,\bar{u}} \right] \leq t \right] \geq 1 - 2 \exp \left( -\frac{t^2}{2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) + \frac{2}{3} Mt} \right). \tag{2.65}$$

Set

$$\frac{\delta}{2} = 2 \exp \left( -\frac{t^2}{2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) + \frac{2}{3} Mt} \right), \tag{2.66}$$

then

$$\log \frac{4}{\delta} = \frac{t^2}{2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) + \frac{2}{3} Mt}. \tag{2.67}$$

By solving with respect to $t$, we have

$$t_{1,2} = \frac{1}{3} M \log 4 \delta \pm \sqrt{\left( \frac{1}{3} M \log 4 \delta \right)^2 + 2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) \log 4 \delta}. \tag{2.68}$$

Since $t > 0$ and $M = 4$, we find

$$t = \frac{4}{3} \log 4 \delta + \sqrt{\left( \frac{4}{3} \log 4 \delta \right)^2 + 2n^* \text{Var} \left( Z_{e,u,\bar{u}} \right) \log 4 \delta}. \tag{2.69}$$

If the probability of the union of events

$$\bigcup_{u,\bar{u},w,\bar{w}:(w,\bar{w}) \in \text{path}_T (u,\bar{u})} \left\{ \left| \sum_{i=1}^{n^*} Z_{e,u,\bar{u}}^{(i)} - n^*E \left[ Z_{e,u,\bar{u}} \right] \right| \geq t \right\}$$
is at most $\frac{\delta}{2p^3}$, then the union bound gives probability at most $\frac{\delta}{2}$. Further,

$$\text{Var} \left( Z_{e,u,\bar{u}} \right) = 2 - (1 - 2q)^4 \left[ 2\mu_A + \mu^2_e (1 - \mu_A)^2 \right]$$
$$= 2 - (1 - 2q)^4 2\mu_A - (1 - 2q)^4 \mu^2_e (1 - \mu_A)^2$$
$$\leq 2 - (1 - 2q)^4 2\mu_A + 0$$
$$= 2 \left( 1 - (1 - 2q)^4 \mu_A \right)$$
$$= 2 \left( 1 - \mu^\dagger_A \right). \quad (2.70)$$

From (2.69) and (2.70) we get

$$t = \frac{4}{3} \log \frac{4p^3}{\delta} + \sqrt{\left( \frac{4}{3} \log \frac{4p^3}{\delta} \right)^2 + 4n^\dagger \left( 1 - \mu^\dagger_A \right) \log \frac{4p^3}{\delta}}$$
$$\leq \frac{8}{3} \log \frac{4p^3}{\delta} + \sqrt{4n^\dagger \left( 1 - \mu^\dagger_A \right) \log \frac{4p^3}{\delta}}, \text{ and}$$

$$t = n^\dagger \left( \frac{4}{3n^\dagger} \log \frac{4p^3}{\delta} + \sqrt{\left( \frac{4}{3n^\dagger} \log \frac{4p^3}{\delta} \right)^2 + \frac{4}{n^\dagger} \left( 1 - \mu^\dagger_A \right) \log \frac{4p^3}{\delta}} \right)$$
$$\leq n^\dagger \left( \frac{8}{3n^\dagger} \log \frac{4p^3}{\delta} + \sqrt{\frac{4}{n^\dagger} \left( 1 - \mu^\dagger_A \right) \log \frac{4p^3}{\delta}} \right). \quad (2.71)$$

Define $\epsilon^\dagger = \sqrt{\log \left( 2p^2/\delta \right)} 2/n^\dagger$, then we have

$$t \leq n^\dagger \left( 4\epsilon^2 + 2\epsilon^\dagger \sqrt{1 - \mu^\dagger_A} \right) \leq n^\dagger \max \left\{ 8\epsilon^2, 4\epsilon^\dagger \sqrt{1 - \mu^\dagger_A} \right\}. \quad (2.72)$$

This completes the proof.

Recall that $\epsilon^\dagger = \sqrt{2/n^\dagger \log \left( 2p^2/\delta \right)}$ and $A \triangleq \text{path}_T (u, \bar{u}) \setminus \{e\}$. The next Lemma gives the concentration of measure bound for the event $E_M$ defined in (2.58).

**Lemma 3.** For all pairs of vertices $u, \bar{u} \in V$ and edges $e = (w, \bar{w})$ in the path $\text{path}_T (u, \bar{u})$ from $u$ to $\bar{u}$, given $n^\dagger$ samples $M^{(1)}_{e,u,\bar{u}}, M^{(2)}_{e,u,\bar{u}}, \ldots, M^{(n)}_{e,u,\bar{u}}$ of $M_{e,u,\bar{u}} = Y_wY_{\bar{w}} + Y_uY_{\bar{u}}$, it is
true that

\[ P\left( \left| \sum_{i=1}^{n_t} M_{e,u,\bar{u}}^{(i)} - n_t \mathbb{E} [M_{e,u,\bar{u}}] \right| \leq n_t \max \left\{ 8\epsilon_t^2, 2 \epsilon_t \sqrt{1 + \mu_A^2} \right\} \right) \geq 1 - \frac{\delta}{2}. \]

**Proof.** We start by finding the expressions for \( \mathbb{E} [M_{e,u,\bar{u}}] \) and \( \text{Var} (M_{e,u,\bar{u}}) \) as follows,

\[ \mathbb{E} [M_{e,u,\bar{u}}] = (1 - 2q)^2 \mathbb{E} [X_w X_{\bar{w}} + X_u X_{\bar{u}}] = (1 - 2q)^2 \mu_e (1 + \mu_A), \quad (2.73) \]

and

\[
\text{Var} (M_{e,u,\bar{u}}) = \mathbb{E} \left[ (M_{e,u,\bar{u}})^2 \right] - \mathbb{E} \left[ (M_{e,u,\bar{u}}) \right]^2 \\
= \mathbb{E} \left[ (X_w N_w X_{\bar{w}} N_{\bar{w}} + N_u X_u N_{\bar{u}} X_{\bar{u}})^2 \right] - \left[ (1 - 2q)^2 \mathbb{E} [X_w X_{\bar{w}} + X_u X_{\bar{u}}] \right]^2 \\
= \mathbb{E} \left[ 1 + 2X_w N_w X_{\bar{w}} N_{\bar{w}} + N_u X_u N_{\bar{u}} X_{\bar{u}} \right] - (1 - 2q)^4 \mathbb{E} [X_w X_{\bar{w}} + X_u X_{\bar{u}}]^2 \\
= 2 + 2(1 - 2q)^4 \mathbb{E} [X_w X_{\bar{w}} X_u X_{\bar{u}}] - (1 - 2q)^4 \mathbb{E} [X_w X_{\bar{w}} + X_u X_{\bar{u}}]^2 \\
= 2 + 2(1 - 2q)^4 \mu_A - (1 - 2q)^4 (\mu_e (1 + \mu_A))^2 \\
= 2 + (1 - 2q)^4 \left[ 2\mu_A - \mu_e^2 (1 + \mu_A)^2 \right]. \quad (2.74)
\]

By applying Bernstein’s inequality, for any \( t > 0 \) we find the inequalities

\[
P\left( \left| \sum_{i=1}^{n_t} M_{e,u,\bar{u}}^{(i)} - n_t \mathbb{E} [M_{e,u,\bar{u}}] \right| \geq t \right) \leq 2 \exp \left( \frac{-t^2}{2n_t \text{Var} (M_{e,u,\bar{u}}) + \frac{2}{3} Mt} \right), \]

\[
P\left( \left| \sum_{i=1}^{n_t} M_{e,u,\bar{u}}^{(i)} - n_t \mathbb{E} [M_{e,u,\bar{u}}] \right| \leq t \right) \geq 1 - 2 \exp \left( \frac{-t^2}{2n_t \text{Var} (M_{e,u,\bar{u}}) + \frac{2}{3} Mt} \right). \quad (2.75)
\]

In the same way as done previously we get

\[
t \leq n_t \left( \frac{8}{3n_t} \log \frac{4p^3}{\delta} + \sqrt{\frac{2}{n_t} \text{Var} (M_{e,u,\bar{u}}) \log \frac{4p^3}{\delta}} \right). \quad (2.76)
\]
and

\[
\text{Var} \left( M_{e,u,\bar{u}} \right) = 2 + (1 - 2q)^4 \left[ 2\mu_A - \mu_e^2 (1 + \mu_A)^2 \right] \leq 2 + (1 - 2q)^4 2\mu_A \\
= 2 \left( 1 + \mu_A^\dagger \right). \quad (2.77)
\]

By setting \( \epsilon^\dagger = \sqrt{\log \left( \frac{2p^2}{\delta} \right)} \frac{2}{n^\dagger} \), we derive the following bound on \( t \)

\[
t \leq n^\dagger \left( 4\epsilon^2\dagger + 2\epsilon^\dagger \sqrt{1 + \mu_A^\dagger} \right) \leq n^\dagger \max \left\{ 8\epsilon^2\dagger, 4\epsilon^\dagger \sqrt{1 + \mu_A^\dagger} \right\}. \quad (2.78)
\]

the latter completes the proof. \( \square \)

In Lemma 4, we derive the set of strong edges for the hidden model. There is a threshold \( \frac{\tau^\dagger}{(1-2q)^2} \geq \tau \) as in the case where there was no noise [84] and the threshold was \( \tau \). Also we find a lower bound for the necessary number of samples for exact structure recovery. In fact we have \( n^\dagger \geq n \), as expected. By setting \( q = 0 \) (then the probability to flip a bit equals to zero) we derive the exact expressions for the threshold \( \tau \) and the sufficient number of samples defined in [84]. Under the event \( E^\dagger_{\text{strong}} (\epsilon^\dagger) \) only the strong edges are guaranteed to exist in the estimated structure \( T_{CL}^\dagger \).

**Lemma 4.** Define the set of strong edges: \( \left\{ (i,j) \in \mathcal{E}_T : \left| \tanh \theta_{ij} \right| \geq \frac{\tau^\dagger}{(1-2q)^2} \right\} \). Under the events defined in Lemmas 2 and 3 all the strong edges will be recovered from the Chow-Liu algorithm with probability at least \( 1 - \delta \). That is,

\[ \mathbb{P} \left[ E^\dagger_{\text{strong}} (\epsilon^\dagger) \right] \geq 1 - \delta = 1 - 2p^2 \exp \left( -\frac{n^\dagger \epsilon^2\dagger}{2} \right). \]

**Proof.** Lemma 1 gives

\[
\left( \sum_{i=1}^{n^\dagger} Z^{(i)}_{f,u,\bar{u}} \right) \left( \sum_{i=1}^{n^\dagger} M^{(i)}_{f,u,\bar{u}} \right) < 0 \implies \sum_{i=1}^{n^\dagger} Z^{(i)}_{f,u,\bar{u}} \leq 0 \text{ or } \sum_{i=1}^{n^\dagger} M^{(i)}_{f,u,\bar{u}} \leq 0 \implies
\]
\[
\left| \sum_{i=1}^{n_\dagger} Z_{f,u,\tilde{u}}^{(i)} - n_\dagger \mathbb{E} \left[ Z_{f,u,\tilde{u}}^{(i)} \right] \right| \geq n_\dagger \mathbb{E} \left[ Z_{f,u,\tilde{u}}^{(i)} \right] \quad \text{or} \\
\left| \sum_{i=1}^{n_\dagger} Y_{f,u,\tilde{u}}^{(i)} - n_\dagger \mathbb{E} \left[ Y_{f,u,\tilde{u}}^{(i)} \right] \right| \geq n_\dagger \mathbb{E} \left[ M_{f,u,\tilde{u}}^{(i)} \right],
\]

then (2.62),(2.73) and Lemmas 2, 3 give

\[
(1 - 2q)^2 \mu_f \frac{1}{1 - \mu_A} \leq \max \left\{ 8 \epsilon_\dagger^2, 4 \epsilon_\dagger \sqrt{1 - \mu_A} \right\} \quad \text{or} \\
(1 - 2q)^2 \mu_f \frac{1}{1 + \mu_A} \leq \max \left\{ 8 \epsilon_\dagger^2, 4 \epsilon_\dagger \sqrt{1 + \mu_A} \right\}
\]

which implies that

\[
\left| \frac{\mu_f^\dagger}{\mu_f^\dagger} \right| \leq (1 - \mu_A)^{-1} \max \left\{ 8 \epsilon_\dagger^2, 4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger} \right\} \quad \text{or} \\
\left| \frac{\mu_f^\dagger}{\mu_f^\dagger} \right| \leq (1 + \mu_A)^{-1} \max \left\{ 8 \epsilon_\dagger^2, 4 \epsilon_\dagger \sqrt{1 + \mu_A^\dagger} \right\}
\]

and the last yields to

\[
\left| \frac{\mu_f^\dagger}{\mu_f^\dagger} \right| \leq \max \left\{ \frac{8 \epsilon_\dagger^2}{(1 - \mu_A^\dagger)}, \frac{8 \epsilon_\dagger^2}{(1 + \mu_A^\dagger)}, \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger}}{(1 - \mu_A)}, \frac{4 \epsilon_\dagger \sqrt{1 + \mu_A^\dagger}}{(1 + \mu_A)} \right\} \Rightarrow \\left| \frac{\mu_f^\dagger}{\mu_f^\dagger} \right| \leq \max \left\{ \frac{8 \epsilon_\dagger^2}{(1 - \mu_A^\dagger)}, \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger}}{(1 - \mu_A)} \right\} \Rightarrow \left| \frac{\mu_f^\dagger}{\mu_f^\dagger} \right| \leq \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger}}{(1 - \mu_A)}. \quad (2.79)
\]

We get the last inequality for non-trivial values of the bound \(\frac{8 \epsilon_\dagger^2}{(1 - \mu_A^\dagger)} \leq 1\) and by using the following bound

\[
\frac{8 \epsilon_\dagger^2}{(1 - \mu_A)} \leq \frac{16 \epsilon_\dagger^2}{(1 - \mu_A)} \leq \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A}}{(1 - \mu_A)} = \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger}}{(1 - \mu_A)} \leq \frac{4 \epsilon_\dagger \sqrt{1 - \mu_A^\dagger}}{(1 - \mu_A)}.
\]
Finally, the function $f(\mu_A) = \frac{4\epsilon_t\sqrt{1-\mu_A^2}}{(1-\mu_A)} = \frac{4\epsilon_t\sqrt{1-(1-2q)^2\mu_A}}{(1-\mu_A)}$ is increasing with respect to $\mu_A$ (for all $\mu_A \leq 1$) and $\mu_A \leq \tanh \beta < 1$, thus we have

$$
\left| \mu_f^\dagger \right| \leq \frac{4\epsilon_t\sqrt{1-\mu_A^2}}{(1-\mu_A)} \leq \frac{4\epsilon_t\sqrt{1-(1-2q)^2\mu_A}}{(1-\mu_A)} \leq \frac{4\epsilon_t\sqrt{1-(1-2q)^4\tanh \beta}}{(1-\tanh \beta)} \triangleq \tau^\dagger. \quad (2.80)
$$

Notice that $\tau^\dagger > \tau = \frac{4\epsilon}{\sqrt{1-\tanh \beta}}$ when $n = n^\dagger$ (or $\epsilon = \epsilon^\dagger$). The weakest edge should satisfy the following property to guarantee the correct recovery of the tree under the event $E^\text{strong}_t(\epsilon_t)$

$$
\left| \mu_f^\dagger \right| \geq \tau^\dagger \implies (1-2q)^2 \tanh \alpha \geq \frac{4\epsilon_t\sqrt{1-(1-2q)^2\mu_A}}{(1-\mu_A)} \implies \tanh \alpha \geq \frac{4\epsilon_t\sqrt{1-(1-2q)^2\mu_A}}{(1-2q)^2(1-\tanh \beta)}. \quad (2.81)
$$

When there is no noise [84, Lemma 9.8], we can guarantee exact recovery with high probability under the event $E^\text{strong}_t(\epsilon)$ and the assumption that the weakest edge satisfies the inequality

$$
\tanh \alpha \geq \frac{4\epsilon}{\sqrt{1-\tanh \beta}}. \quad (2.82)
$$

Notice that (2.82) can be obtained by (2.81) when $q = 0$ and $n = n^\dagger$. When $q > 0$ and $n = n^\dagger$ it is clear that the set of trees which can be recovered from noisy observations is a subset of the set of trees that can be recovered from the original observations. Also, we have

$$
\epsilon \triangleq \sqrt{\frac{2\log(2p^2/\delta)}{n}} \implies n = \frac{2}{\epsilon^2} \log \left( \frac{2p^2}{\delta} \right) \quad \text{and} \\
\epsilon^\dagger \triangleq \sqrt{\frac{2\log(2p^2/\delta)}{n}} \implies n^\dagger = \frac{2}{\epsilon^\dagger^2} \log \left( \frac{2p^2}{\delta} \right). \quad (2.83)
$$

By combining (2.81) with (2.83) we found the number of samples that we need to recover the tree with probability at $1 - \delta$,.
\[ n_\dagger > \frac{32 \left[ 1 - (1 - 2q)^4 \tanh \beta \right]}{(1 - \tanh \beta)^2 (1 - 2q)^4 \tanh^2 \alpha} \log \frac{2p^2}{\delta}. \] (2.84)

On the other hand, when there is no noise [84] we need
\[ n > \frac{32}{\tanh^2 \alpha (1 - \tanh \beta)} \log \frac{2p^2}{\delta}. \] (2.85)

The last two inequalities give us how the number of samples scales as a function of the probability \( q \)
\[ n_\dagger \geq \frac{1 - (1 - 2q)^4 \tanh \beta}{(1 - \tanh \beta) (1 - 2q)^4} \left[ e^{2\beta} \left( (1 - 2q)^{-4} - 1 \right) + 1 + (1 - 2q)^{-4} \right]. \] (2.86)

From the above we can distinguish specific cases for values of \( q \). For instance when \( q \to \frac{1}{2} \) then we need \( n_\dagger \to \infty \) for exact structure recovery, when \( q \to 0 \) then we need at least \( n \) number of samples for exact structure recovery.

### 2.9.2 Proof of Theorem 2 (Necessary number of samples)

In this section, we use a strong data processing inequality together with a family of models (considered also by Bresler and Karzand [84]) to derive the proof of Theorem 2. Specifically, we combine the proofs of Theorem 3.2 by Bresler and Karzand [84, Lemma 8.1] and a strong data processing inequality result by Polyanskiy and Wu [66]. First, we consider the following variation of Fano’s inequality [85].

**Corollary 1.** [85, Corollary 2.6]: Assume that \( \Theta \) is a family of \( M + 1 \) distributions \( \theta_0, \theta_1, \ldots, \theta_M \) such that \( M \geq 2 \). Let \( P_{\theta_i} \) be the distribution of the variable \( X \) under the model \( \theta_i \), if
\[ \frac{1}{M + 1} \sum_{i=1}^{M} D_{KL} (P_{\theta_i} || P_{\theta_0}) \leq \gamma \log M, \] (2.87)

for any \( \gamma \in (0, \frac{1}{8}) \), then for the probability of the error \( p_e \) the following inequality holds:
\[ p_e \geq \frac{\log(M+1) - 1}{\log(M)} - \gamma. \] We restrict the values of \( \gamma \) to \((0, \frac{1}{8})\) because we are interested in
the case where \( p_e \geq \frac{1}{2} \), in general the above holds for all values of \( \gamma \in (0,1) \), see [85, Corollary 2.6].

At this point we consider Bresler and Karzand’s construction [84, section 8.1] of \( M + 1 \) different Ising model distributions \( \{ P_{\theta^i} : i \in \{0, \ldots, M\} \} \). This family of structured distributions is chosen such that the recovery task is sufficiently hard. First, we define \( P_{\theta^0} \) to be an Ising model distribution with underlying structure a chain with \( p \) nodes and parameters \( \theta_{j,j+1}^0 = \alpha \) when \( j \) is odd and \( \theta_{j,j+1}^0 = \beta \) when \( j \) is even. The rest of family is constructed as follows: the elements of each \( \theta^i \) are equal to the elements of \( \theta^0 \) apart from two elements \( \theta_{i,i+1}^i = 0 \) and \( \theta_{i,i+2}^i = \alpha \) for each odd value of \( i \). There are \( (p+1)/2 \) distributions in the constructed family. Bresler and Karzand evaluate the upper bound for the quantity \( S_{\text{KL}}(P_{\theta^0} \| P_{\theta^i}) \) for all \( i \in [M] \) under this family of distributions and we have [84, Section 8.1):

\[
S_{\text{KL}}(P_{\theta^0} \| P_{\theta^i}) = 2 \alpha \left( \tanh(\alpha) - \tanh(\alpha) \tanh(\beta) \right) \leq 4 \alpha \tanh(\alpha) e^{-2\beta}. \tag{2.88}
\]

For each distribution \( P_{\theta^i} \) and \( i \in \{0, \ldots, M\} \) we consider the distribution of the noisy variable in the hidden model \( P_{\theta^i}^\dagger \triangleq P_{Y|X} \circ P_{\theta^i} \) and we would like to find an upper bound for the quantities \( S_{\text{KL}}(P_{\theta^i}^\dagger \| P_{\theta^i}^\dagger) \). To do this we use a strong data processing inequality result [66] for any binary symmetric channel. The input random variable \( X \) is considered to have correlated elements while the noise variables \( N_i \) are i.i.d Rademacher(\( q \)) which is equivalent to the hidden model that we consider in this dissertation. In fact we have the following bound

\[
\eta_{\text{KL}} \leq 1 - (4q(1 - q))^p. \tag{2.89}
\]

The quantity \( \eta_{\text{KL}} \) is defined as:

\[
\eta_{\text{KL}} \triangleq \sup_Q \sup_{P: D_{\text{KL}}(P||Q) < \infty} \frac{D_{\text{KL}} \left( P_{Y|X} \circ P \| P_{Y|X} \circ Q \right)}{D_{\text{KL}} (P||Q)}, \tag{2.90}
\]

where \( P_{Y|X} \) is the distribution of the BSC and \( P, Q \) are any distributions of the input variable \( X \). Since (2.90) has the supremum over all possible distributions it covers any
pair of distributions in the desired family \( \{P_{\theta^j} : j \in \{0, \ldots, M\}\} \) and we have

\[
\frac{D_{\text{KL}}(P^{\dagger}_{\theta^0}||P^{\dagger}_{\theta^i})}{D_{\text{KL}}(P_{\theta^0}||P_{\theta^i})} \leq 1 - (4q(1-q))^p \implies (2.91)
\]

\[
S_{\text{KL}}(P^{\dagger}_{\theta^0}||P^{\dagger}_{\theta^i}) \leq [1 - (4q(1-q))^p]S_{\text{KL}}(P_{\theta^0}||P_{\theta^i}). \tag{2.92}
\]

(2.88) and (2.92) give

\[
S_{\text{KL}}(P^{\dagger}_{\theta^0}||P^{\dagger}_{\theta^i}) \leq [1 - (4q(1-q))^p]4\alpha^2 e^{-2\beta}. \tag{2.93}
\]

Finally, from (2.93) and Lemma 1 we derive the bound of Theorem 2.

2.9.3 Proof of Theorem 3, (Sufficient number of samples for a noisy Gaussian model)

Let \( X = (X_1, X_2, \ldots, X_p) \) be a Gaussian random vector with distribution \( N(0, \Sigma) \).

We assume that the Markov property holds such that the underlying graph is a tree \( T = (V, E) \). Also assumption 2 holds;

\[
\text{Var}(X_i) = \mathbb{E}[X_i^2] = 1, \quad \forall i \in V
\]

\[
0 < \rho_m \leq \left| \mathbb{E}[X_i X_j] \right| \leq \rho_M < 1, \quad \forall (i, j) \in E.
\]

We consider i.i.d. Gaussian noise \( N \sim N(0, \sigma^2 I) \). The noisy output variables of the hidden model are \( \tilde{Y} = X + N \sim N(0, \Sigma + \sigma^2 I) \). Then

\[
\rho_{i,j}^{\dagger} \triangleq \frac{\mathbb{E}[\tilde{Y}_i \tilde{Y}_j]}{\sqrt{\mathbb{E}[(\tilde{Y}_i)^2]\mathbb{E}[(\tilde{Y}_j)^2]}} = \frac{\mathbb{E}[\tilde{Y}_i \tilde{Y}_j]}{\sqrt{1 + \sigma^2}} = \mathbb{E} \left[ \frac{\tilde{Y}_i}{\sqrt{1 + \sigma^2}} \frac{\tilde{Y}_j}{\sqrt{1 + \sigma^2}} \right], \quad \forall i, j \in V. \tag{2.94}
\]

The random variables \( Y_i \triangleq \tilde{Y}_i/\sqrt{1 + \sigma^2} \) are normalized Gaussian with variance equal to 1. Instead of using \( \tilde{Y} \), we use the normalized variable \( Y \) with distribution

\[
Y \sim N \left( 0, \frac{\Sigma + \sigma^2 I}{1 + \sigma^2} \right). \tag{2.95}
\]
Then
\[ \text{Var}(Y_i) = \mathbb{E}[Y_i^2] = 1, \quad \forall i \in \mathcal{V}, \quad (2.96) \]

and
\[ \frac{\rho_m}{1 + \sigma^2} \leq |\mathbb{E}[Y_i Y_j]| = \left| \frac{\mathbb{E}[X_i X_j]}{1 + \sigma^2} \right| \leq \frac{\rho_M}{1 + \sigma^2}, \quad \forall (i, j) \in \mathcal{E}. \quad (2.97) \]

(2.97) shows that noise makes the edges "weaker", since \(1 + \sigma^2 > 1\) and for \(\sigma \to \infty\) we have \(|\mathbb{E}[Y_i Y_j]| \to 0\) which makes the structure learning task impossible. The next Lemma provides upper bounds on the probabilities of the sufficient events.

**Lemma 5.** Define
\[ f^{(1)}_{u,\tilde{u},c}(Y^{1:n}) \triangleq \sum_{i=1}^{n} Z^{(i)}_{f,u,\tilde{u}} = \sum_{i=1}^{n} \left(Y_{w}^{(i)} Y_{\tilde{w}}^{(i)} - Y_{u}^{(i)} Y_{\tilde{u}}^{(i)} \right), \]

and
\[ f^{(2)}_{u,\tilde{u},c}(Y^{1:n}) \triangleq \sum_{i=1}^{n} \tilde{Z}^{(i)}_{f,u,\tilde{u}} = \sum_{i=1}^{n} \left(Y_{w}^{(i)} Y_{\tilde{w}}^{(i)} + Y_{u}^{(i)} Y_{\tilde{u}}^{(i)} \right). \quad (2.98) \]

Then for some fixed \(R \in \mathbb{R}^+\), both of the probabilities below
\[ \mathbb{P}
\left\{ \bigcap_{u,\tilde{u},c} \left\{ \left| f^{(1)}_{u,\tilde{u},c}(Y^{1:n}) - \mathbb{E}[f^{(1)}_{u,\tilde{u},c}(Y^{1:n})] \right| \leq R \sqrt{\text{Var}(f^{(1)}_{u,\tilde{u},c}(Y^{1:n}))} \log^2 \left( \frac{p^3 e^2}{\delta} \right) \right\} \right\] \]

and
\[ \mathbb{P}
\left\{ \bigcap_{u,\tilde{u},c} \left\{ \left| f^{(2)}_{u,\tilde{u},c}(Y^{1:n}) - \mathbb{E}[f^{(2)}_{u,\tilde{u},c}(Y^{1:n})] \right| \leq R \sqrt{\text{Var}(f^{(2)}_{u,\tilde{u},c}(Y^{1:n}))} \log^2 \left( \frac{p^3 e^2}{\delta} \right) \right\} \right\] \]

are greater or equal than \(1 - \delta/2\).

**Proof.** We apply a concentration of measure Theorem by Schudy and Sviridenko [86,
Theorem 1.10; 

\[ P \left[ \left| f \left( Y^{1:n} \right) - \mathbb{E}[f \left( Y^{1:n} \right)] \right| \geq \lambda \right] \leq e^{2e} \left( \frac{\lambda}{R \sqrt{\text{Var}(f(Y^{1:n}))}} \right)^{1/q}, \quad \forall \lambda > 0, \]

and \( f \left( Y^{1:n} \right) = f(Y_1, \ldots, Y_n) \) is a \( q \) degree polynomial and the random variables \( Y_1, \ldots, Y_n \) are distributed according to a log-concave measure in \( \mathbb{R}^n \) and they are not necessarily independent. In our case \( f(Y) = \sum_{i=1}^{n} Z^{(i)}_{f,u,\bar{u}} \) or \( f(Y) = \sum_{i=1}^{n} \tilde{Z}^{(i)}_{f,u,\bar{u}} \) and we have \( q = 2 \). Then we choose the probability to be at least \( \delta \frac{p}{2p(\delta)} \), since we apply union bound for all pairs of nodes \( u, \bar{u} \) and edges \( e = (w, \tilde{w}) \in \text{path}_T(u, \bar{u}) \)

\[ \frac{\delta}{2p(\delta)} = e^{2e} \left( \frac{\lambda}{R \sqrt{\text{Var}(f(Y^{1:n}))}} \right)^{1/q} \]

\[ \implies \lambda = R \sqrt{\text{Var}(f(Y))} \log^2 \left( \frac{2pe^2(p)}{\delta} \right) < R \sqrt{\text{Var}(f(Y))} \log^2 \left( \frac{e^2p^3}{\delta} \right). \quad (2.99) \]

Then we calculate the variance \( \text{Var}(f(Y)) \) for both cases, when \( f(Y) = \sum_{i=1}^{n} Z^{(i)}_{f,u,\bar{u}} \) and \( f(Y) = \sum_{i=1}^{n} \tilde{Z}^{(i)}_{f,u,\bar{u}} \). For the case \( f(Y) = \sum_{i=1}^{n} Z^{(i)}_{f,u,\bar{u}} \), we can express the higher order moments in terms of the covariates by using Isserlis’ theorem [90]:

\[ \mathbb{E}[Y^2_{i}Y^2_{j}] = \mathbb{E}[Y^2_i] \mathbb{E}[Y^2_j] + 2\mathbb{E}^2[Y_iY_j], \quad (2.100) \]

\[ \mathbb{E}[Y_iY_jY_kY_l] = \mathbb{E}[Y_iY_j] \mathbb{E}[Y_kY_l] + \mathbb{E}[Y_iY_k] \mathbb{E}[Y_jY_l] + \mathbb{E}[Y_iY_l] \mathbb{E}[Y_jY_k], \quad (2.101) \]
and we get

\[ \text{Var} \left( f_{u, \tilde{u}_e}^{(1)} \right) \]
\[ \overset{\text{i.i.d.}}{=} \sum_{i=1}^{n} \text{Var} \left( Z_{f,u,\tilde{u}}^{(i)} \right) \]
\[ = n \text{Var} \left( Z_{f,u,\tilde{u}} \right) \]
\[ = n \left( \mathbb{E}[Y_w Y_{\tilde{u}} - Y_u Y_{\tilde{u}}]^2 \right) \]
\[ = n \left( \mathbb{E}[Y_w^2 Y_{\tilde{u}}^2] + \mathbb{E}[Y_u^2 Y_{\tilde{u}}^2] - 2 \mathbb{E}[Y_w Y_u Y_{\tilde{u}}] - \mathbb{E}[Y_w Y_{\tilde{u}}]^2 - \mathbb{E}[Y_u Y_{\tilde{u}}]^2 + 2 \mathbb{E}[Y_w Y_{\tilde{u}}] \mathbb{E}[Y_u Y_{\tilde{u}}] \right) \]
\[ = n \left( 2 + \mathbb{E}^2[Y_w Y_{\tilde{u}}] + \mathbb{E}^2[Y_u Y_{\tilde{u}}] - 2 \left( \mathbb{E}[Y_w Y_u] \mathbb{E}[Y_{\tilde{u}} Y_{\tilde{u}}] + \mathbb{E}[Y_w Y_{\tilde{u}}] \mathbb{E}[Y_u Y_{\tilde{u}}] \right) \right) \]
\[ \leq n \left( 6 + \mathbb{E}^2[Y_w Y_{\tilde{u}}] + \mathbb{E}^2[Y_u Y_{\tilde{u}}] \right) \]
\[ = n \left( 6 + \frac{1}{(1 + \sigma^2)^2} \mathbb{E}^2[X_w X_{\tilde{u}}] + \frac{1}{(1 + \sigma^2)^2} \mathbb{E}^2[X_u X_{\tilde{u}}] \right) \]
\[ = 6n + \frac{n}{(1 + \sigma^2)^2} \mathbb{E}^2[X_u X_{\tilde{u}}] \left( \prod_{e \in \text{path}(w, \tilde{u}) \setminus (u, \tilde{u})} \mu_e^2 + 1 \right), \]

and (2.102) comes from (2.96), (2.100), (2.101) and the last comes from the correlation coefficient decay property. In a similar way,
and we find

\[
\lambda < R \sqrt{\text{Var} \left( f \left( Y^{1:n} \right) \right)} \log^2 \left( \frac{e^2 p^3}{\delta} \right)
\]

\[
\leq R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2} \prod_{e \in \text{path}(w, \tilde{w}) \setminus (u, \tilde{u})} \mu^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right)}.
\]

Similarly, as in the case of Ising model, we start by stating the condition for the error event; Let \( f = (w, \tilde{w}) \) be an edge: \( f \in T \) and \( f \notin T_{\text{CL}}^+ \) then \( \exists g \in T_{\text{CL}}^+ \) and \( g \notin T \): \( f \in \text{path}_T (u, \tilde{u}) \) and \( g \in \text{path}_{T_{\text{CL}}^+} (w, \tilde{w}) \), then for the error event we have

\[
0 \geq \left| \hat{\rho}_f^+ \right|^2 - \left| \hat{\rho}_g^+ \right|^2 = \left( \hat{\rho}_f^+ - \hat{\rho}_g^+ \right) \left( \hat{\rho}_f^+ + \hat{\rho}_g^+ \right)
\]

\[
= \frac{1}{n^2} \left( \sum_{i=1}^{n} Y_w^{(i)} Y_{\tilde{w}}^{(i)} - Y_u^{(i)} Y_{\tilde{u}}^{(i)} \right) \times \left( \sum_{i=1}^{n} Y_w^{(i)} Y_{\tilde{w}}^{(i)} + Y_u^{(i)} Y_{\tilde{u}}^{(i)} \right)
\]

\[
= \frac{1}{n^2} \left( \sum_{i=1}^{n} Z_{f,u,\tilde{u}}^{(i)} \right) \left( \sum_{i=1}^{n} \tilde{Z}_{f,u,\tilde{u}}^{(i)} \right).
\]

the latter gives

\[
\sum_{i=1}^{n} Z_{f,u,\tilde{u}}^{(i)} \leq 0 \text{ or } \sum_{i=1}^{n} \tilde{Z}_{f,u,\tilde{u}}^{(i)} \leq 0
\]

and

\[
\left| \sum_{i=1}^{n} Z_{f,u,\tilde{u}}^{(i)} - \mathbb{E} \left[ \sum_{i=1}^{n} Z_{f,u,\tilde{u}}^{(i)} \right] \right| \geq \mathbb{E} \left| \sum_{i=1}^{n} Z_{f,u,\tilde{u}}^{(i)} \right| \quad \text{or}
\]

\[
\left| \sum_{i=1}^{n} \tilde{Z}_{f,u,\tilde{u}}^{(i)} - \mathbb{E} \left[ \sum_{i=1}^{n} \tilde{Z}_{f,u,\tilde{u}}^{(i)} \right] \right| \geq \mathbb{E} \left| \sum_{i=1}^{n} \tilde{Z}_{f,u,\tilde{u}}^{(i)} \right|. \quad (2.103)
\]
From Lemma 5 with probability at least $1 - \delta$ the following inequalities hold

$$
R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2}} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right)
$$

$$
\geq R \sqrt{\text{Var}(f_{u, \bar{u}, e}^{(1)}(Y^{1:n}))} \log^2 \left( \frac{e^2 p^3}{\delta} \right) \geq \left| \sum_{i=1}^{n} X_{f, u, \bar{u}}^{(i)} - \mathbb{E} \left[ \sum_{i=1}^{n} Z_{f, u, \bar{u}}^{(i)} \right] \right|
$$

and

$$
R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2}} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right)
$$

$$
\geq R \sqrt{\text{Var}(f_{u, \bar{u}, e}^{(2)}(Y^{1:n}))} \log^2 \left( \frac{e^2 p^3}{\delta} \right) \geq \left| \sum_{i=1}^{n} \tilde{Z}_{f, u, \bar{u}}^{(i)} - \mathbb{E} \left[ \sum_{i=1}^{n} \tilde{Z}_{f, u, \bar{u}}^{(i)} \right] \right|. \tag{2.104}
$$

We combine (2.103) and (2.104) and we have

$$
R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2}} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right) \geq \left| \mathbb{E} \left[ \sum_{i=1}^{n} Z_{f, u, \bar{u}}^{(i)} \right] \right|
$$

or

$$
R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2}} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right) \geq \left| \mathbb{E} \left[ \sum_{i=1}^{n} \tilde{Z}_{f, u, \bar{u}}^{(i)} \right] \right| \implies
$$

$$
R \sqrt{7n + \frac{n}{(1 + \sigma^2)^2}} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e^2 \log^2 \left( \frac{e^2 p^3}{\delta} \right)
$$

$$
\geq n \left| \mathbb{E}[X_{u, \bar{u}}] \right| \frac{1}{1 + \sigma^2} \prod_{e \in \text{path}(w, \bar{w}) \setminus (u, \bar{u})} \mu_e - 1 \right|
$$
or

\[
R \sqrt{\frac{\mathcal{E}[X_uX_{\tilde{u}}]}{1 + \sigma^2} \prod_{e \in \text{path}(w, \tilde{w}) \setminus (u, \tilde{u})} \mu_e} \log^2 \left( \frac{e^2p^3}{\delta} \right) \geq n \left| \mathbb{E}[X_uX_{\tilde{u}}] \right|
\]

From (2.105) we find the sufficient condition for the weakest edge: for exact structure recovery we need \( \rho_m \) to be greater than the following term

\[
R(1 + \sigma^2) \sqrt{\left( 7 + \frac{1}{(1 + \sigma^2)^2} \prod_{e \in \text{path}(w, \tilde{w}) \setminus (u, \tilde{u})} \mu_e^2 \right)} \log^2 \left( \frac{e^2p^3}{\delta} \right).
\]

The function \( f(x) = \frac{R(1 + \sigma^2) \sqrt{7 + \frac{1}{(1 + \sigma^2)^2} \prod_{e \in \text{path}(w, \tilde{w}) \setminus (u, \tilde{u})} \mu_e^2}}{\sqrt{1 - x}} \log^2 \left( \frac{e^2p^3}{\delta} \right) \) is increasing for all \( x \in [0, 1) \) and \( \prod_{e \in \text{path}(w, \tilde{w}) \setminus (u, \tilde{u})} \mu_e^2 \leq \rho_M \). Thus, it is sufficient to have

\[
\rho_m \geq \frac{R(7(1 + \sigma^2)^2 + \rho_M) \log^2 \left( \frac{e^2p^3}{\delta} \right)}{\sqrt{n}(1 - \rho_M)}
\]

and the sufficient number of samples is given by

\[
n \geq \frac{R^2 \left[ 7(1 + \sigma^2)^2 + \rho_M \right] \log^4 \left( \frac{e^2p^3}{\delta} \right)}{\rho_m^2 (1 - \rho_M)^2},
\]

for some positive constant \( R \).
Chapter 3

Predictive Learning on Hidden Tree-Structured Ising Models

In this chapter we consider the problem of learning accurately a set of marginal distributions (and conditional distributions) for the hidden tree-structured variables, while only noisy samples from are available. As we explain later, the estimation of conditional distributions will allow us to make predictions for the value of a node variable given the values of other nodes. A typical application of this setting is a recommendation system; based on users’ choices we would like to use the estimated model to predict future preference. The imperfect user behaviour \[28\] introduces noise, thus only noisy observables are available.

Herein, we consider binary models on \(2p\) variables \((X, Y)\) similarly to Chapter 2. The joint distribution \(p(\cdot)\) of \(X\) is a tree-structured Ising model distribution on \([-1, +1]^p\) and \(Y = (Y_1, Y_2, \ldots, Y_p)\) is a noisy version of \(X\), such that \(Y_i = N_iX_i\) and \\{\(N_i\)\} are independent and identically distributed (i.i.d.) Rademacher noise with \(\Pr(N_i = -1) = 1 - \Pr(N_i = +1) = q\), for all \(i \in V\). We refer to \(X\) as the hidden layer and \(Y\) as the observed layer.\(^1\) Under this setting, our objective is to accurately estimate the distribution of the hidden layer \(X\) (with high probability) using only the noisy observations \(Y\). This is non-trivial because \(Y\) does not itself follow any tree structure; this is similar to more traditional problems in nonlinear filtering, where a Markov process of known distribution (and thus, of known structure) is observed through noisy measurements \([72, 73, 74, 75, 76]\). The sample complexity of the noiseless version of our model was recently studied by Bresler and Karzand \[29\], where the well-known

\(^1\) For further details about the setting and premiliminaries on hidden tree-structured Ising models see Section 2.1, Chapter 2.
Figure 3.1: The simulation corresponds to structure learning. Comparison of the experimental results (heat-map) and the theoretical bound of Theorem 1, the bound that yields (3.1). The colored regions denote different values of the estimated probability of error $\delta$ (at least one edge has been missed). The value of $\delta$ varies between 0 and 1 while the parameters $\alpha = 0.2, \beta = 1.1, p = 100$ are fixed. The red line shows the bound from Theorem 1 (the explicit form of Theorem 4). The code of the experiment is available at structure learning code.

Chow-Liu algorithm [3] is employed for tree reconstruction. We continue by providing a summary of our results for hidden tree-structured Ising models: sample complexity bounds up to constant factors for structure and predictive learning. As a recap, in the next section we present Theorems 1 and 2 (Chapter 2) in a simplified form and then the predictive learning sample complexity bounds. This approach allows us to highlight the dependence of the sample complexity with respect to the parameters of interest, to compare the structure learning bounds with the predictive learning bounds, as well as, to illustrate the impact of noise, by comparing the noisy setting with the noiseless counterpart by Bresler and Karzand [29].

3.1 Summary of the Results

In this section, we present a summary of the main results of our work up to constant factors $C, C' > 0$. We refer the reader to Table 1.1 for the definition of the model parameters. We provide the explicit statements of the results, and we specify the constants in Section 3.2. Recall that, the random vector $Y \in \{-1, +1\}^p$ is the output.
of the binary symmetric channel $\text{BSC}(q)^p$ with input the random vector $X \sim p(\cdot) \in \mathcal{P}_T(\alpha, \beta)$.

### 3.1.1 Structure Learning

The first results provides the sufficient number of samples for exact structure recovery.

**Theorem 4** (Sample Complexity for Structure Learning). *The Chow-Liu algorithm with input $n$ noisy samples $Y_1^n$ exactly estimates the hidden tree structure $T_{\text{CL}}^\dagger = T$ with probability at least $1 - \delta \in (0, 1)$, as long as

$$n > C \frac{e^{2\beta (1+\mathbb{1}_{q \neq 0})}}{(1-2q)^4 \tanh^2(\alpha)} \log(p/\delta).$$  \hspace{1cm} (3.1)

The order with respect to $\beta$ is $O(e^{4\beta})$ for all $q > 0$. The bound in (3.1) exactly reduces to the noiseless case [29, Theorem 3.2]. Additionally, the explicit form of the result, Theorem 1, shows that the bound is also a continuous function of $q \in [0, 1/2)$. The next proposition gives the necessary number of samples for exact structure recovery.

**Proposition 1.** *No algorithm can recover the structure with probability great than $1/2$ if

$$n < C' e^{2\beta [1 - (4q(1-q))^p]^{-1}} \alpha \tanh(\alpha) \log(p).$$  \hspace{1cm} (3.2)

Note that the terms $(1-2q)^{-4}$ and $[1 - (4q(1-q))^p]^{-1}$ introduce a gap between the sample complexity of (3.1) and (3.2). However, the sample complexity of Theorem 4 is indeed accurate. To illustrate this experimentally, we show that the theoretical and experimental bounds exactly match, see Figure 3.1. The latter indicates that the Chow-Liu algorithm requires exactly the number of samples that our theoretical result suggest (see Figure 3.1). On the other hand, Proposition 1 provides the necessary number of samples, for any algorithm. Finally, we conjecture that the bound of Proposition 1.2 is tight only under the low temperature regime $|\theta_{i,j}| \to \infty$ for all $i, j \in \mathcal{E}$. The derivation of generalized tighter forms of the bound in (3.2) is challenging and left for future work.
3.1.2 Predictive Learning

To learn the tree-shaped distribution $p(\cdot) \in \mathcal{P}_T(\alpha, \beta)$ of $\mathbf{X}$ from $n$ noisy samples $\mathbf{Y}^{1:n}$, we first estimate the correlations $\hat{\mu}^i_{i,j}$ for all $i, j \in \mathcal{V}$. We then estimate the tree structure $T^\text{CL}_* \dagger$ by running the Chow-Liu algorithm with input the candidate edge weights $\hat{\mu}^i_{i,j}$ and finally evaluate the estimator of $p(\cdot)$ (by matching correlations) as follows:\footnote{The distribution in (3.3) is a function of $\mathbf{x}$, however we suppress the notation for consistency with prior work and for sake of space.}

$$\Pi_{T^\text{CL}_*}(\hat{p}_T) \triangleq \frac{1}{2} \prod_{(i,j) \in E_{T^\text{CL}_*}} \frac{1 + x_i x_j}{2} \frac{\hat{\mu}^i_{i,j}}{(1-2q)^2}, \quad \mathbf{x} \in \{-1, +1\}^p. \quad (3.3)$$

Note that one restriction of our approach is that the distribution estimator requires the value $q$ to be known. The same restriction appears in other structure learning from noisy data approaches [70]. However, in our setting $q$ is required only for the predictive learning, while the Chow-Liu algorithm and the structure estimation does not require $q$ to be known. Under the assumption that $q$ is unknown, one can first learn its value through an independent procedure [70, Section 5]. The accuracy of the estimated distribution in (3.3) is measured by the small-set Total Variation (ssTV), that captures the estimation error on the $k$th-order marginals [91, 92, 29]. Let $P_S, Q_S$ denote the marginals of $P, Q$ on a set $S \subset \mathcal{V}$, and $|S| = k$. Then the $k$th order ssTV of $P$ and $Q$ is defined as

$$\mathcal{L}^{(k)}(P, Q) \triangleq \sup_{S: |S| = k} d_{\text{TV}}(P_S, Q_S). \quad (3.4)$$

The next results provides the necessary number of samples for accurate distribution estimation by guaranteeing that the $\mathcal{L}^{(2)}$ is less than a small positive number $\eta$ with high probability. We provide guarantees on higher-order marginals ($k > 2$) in Section 3.2.2.

**Theorem 5** (Sample Complexity for Predictive Learning). Fix $\delta \in (0, 1)$. Choose $\eta > 0$
Figure 3.2: The simulation corresponds to predictive learning. Comparison of the experimental results (heat-map) and the theoretical bound of Theorem 5. The colored regions denote different values of the estimated probability of error $\delta$ (ssTV to be greater than a fixed number $\eta$). The value of $\delta$ varies between 0 and 1 while the parameters $\eta = 0.03, \beta = 1.1, p = 31$ are fixed. The code is available at predictive learning code.

\[ n = C \frac{\eta^2}{(1-2q)^4} \log \left( \frac{p}{\delta} \right) \]

(3.5)

where

\[ n \geq C \max \left\{ \frac{1}{\eta^2(1-2q)^4}, \frac{e^{2\beta(1+\mathbb{1}_{q\neq0})}}{(1-2q)^4}, \frac{e^{4\beta}}{\eta^2 \mathbb{1}_{q\neq0}} \right\} \log \left( \frac{p}{\delta} \right) \]

then

\[ \mathbb{P} \left( \mathcal{L}(2) \left( p(\cdot), \Pi_{\mathcal{CL}^+}(\hat{p}_t) \right) \leq \eta \right) \geq 1 - \delta. \]

(3.6)

Note that the dependence on $\beta$ is $O(e^{4\beta})$ for accurate distribution learning from noisy data (similarly to the structure learning task, Theorem 4). The bound in (3.5) exactly reduces to the noiseless setting bound by [29, Theorem 3.3]. Theorem 5 is a short version of the result. The explicit statement, Theorem 6, shows that the bound is also continuous at $q \to 0$.

Conversely, the following proposition gives an upper bound on the necessary number of samples for accurate marginal distributions’ estimation under the assumption $\beta > \alpha$.

**Proposition 2.** Fix $\eta > 0$ such that $\eta \leq (\tanh(\beta) - \tanh(\alpha))/2$. Then no algorithm can accurately estimate the distribution of the hidden variables (ssTV less than $\eta > 0$)
Sufficient Number of Samples

<table>
<thead>
<tr>
<th>Task/Setting</th>
<th>Noiseless (prior work)</th>
<th>Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure Learning</td>
<td>$\frac{e^{2\beta}}{\tanh^2(\alpha)} \log(p/\delta)$</td>
<td>$\frac{e^{2\beta(1+q\neq 0)}}{(1-2q)^2 \tanh^2(\alpha)} \log(p/\delta)$</td>
</tr>
<tr>
<td>Predictive Learning</td>
<td>$C \max{\eta^{-2}, e^{2\beta}} \log(p/\delta)$</td>
<td>$C \max\left{\frac{\eta^{-2}}{(1-2q)^2}, \frac{e^{2\beta(1+q\neq 0)}}{(1-2q)^2}, \frac{e^{4\beta} \eta^2}{\eta^2}\right} \log(p/\delta)$</td>
</tr>
</tbody>
</table>

Table 3.1: Sufficient number of samples for accurate structure and predictive learning.

<table>
<thead>
<tr>
<th>Task/Setting</th>
<th>Noiseless (prior work)</th>
<th>Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure learning</td>
<td>$C' \frac{e^{2\beta}}{\alpha \tanh(\alpha)} \log(p)$</td>
<td>$C' \frac{e^{2\beta(1-4q \neq 0)}}{\alpha \tanh(\alpha)} [1-(4q(1-q))^p]^{-1} \log(p)$</td>
</tr>
<tr>
<td>Predictive learning</td>
<td>$C' \eta^{-2} \log(p)$</td>
<td>$C' \eta^{-2} [1-(4q(1-q))^p]^{-1} \log(p)$</td>
</tr>
</tbody>
</table>

Table 3.2: Necessary number of samples for structure and predictive learning.

with probability greater than 1/2 if

$$n < C' \eta^{-2} [1-(4q(1-q))^p]^{-1} \log(p).$$

(3.7)

A quick comparison of (3.5) and (3.7) shows that there is a gap between the sufficient and necessary number of samples. Our experiments (Figure 3.2) confirm the accuracy of our theoretical results. For instance the bound of Theorem 5 exactly matches the experimental curve. For further discussion related to the gap between the upper and lower bounds see Section 2.1.6. Further, we conjecture that bound in (3.7) is tight only under the low temperature regime, similarly to the Proposition 1. The derivation of tighter characterization of the necessary number of samples Propositions 1 and 2 remains an problem for future work. Additional plots of the experiments are provided in Section 3.2.3. Finally, Table 3.1 and 3.2 summarize the state-of-the-art bounds of the noiseless setting by Bresler and Karzand [29] and the extended version under the noisy setting that we consider in this dissertation.

To summarize, the following holds for both structure and predictive learning: the dependence on the parameter $\beta$ is of the order $\mathcal{O}(e^{2\beta})$ for $q = 0$ and becomes $\mathcal{O}(e^{4\beta})$
for positive values of \( q \). Further, the bounds are continuous functions of \( q \), as our results suggest (for the continuity see the explicit form of the results Theorem 1 and Theorem 6.) Similarly to the noiseless case, the following statement holds when noise exists as well: Under the high temperature regime (\( \alpha \) close to zero), structure learning requires much more data than the predictive learning task, because of the \( \tanh^2(\alpha) \) in the denominator of the bound in (3.1). On the contrary, the required number of samples for predictive learning (3.5) does not depend on \( \alpha \). Specifically, exact structure recovery is not necessary for learning the distribution efficiently, that is, weak edges’ identification failure does not affect the predictive learning task. We refer the reader to Section 3.3.1 for the definition of weak/strong edges and additional explanation. Finally, for \( q > 0 \) an extra term that involves both \( \beta \) and \( \eta \) appears in the bound of Theorem 5, while for values of \( q \) close to zero and \( q = 0 \) vanishes.

The pairwise correlations of end-point vertices \( \langle E[X_i X_j]; (i,j) \in \mathcal{E} \rangle \) are sufficient statistics, and as expected, the accuracy of pairwise marginals corresponds to accuracy of higher order marginals and accurate estimation of higher order moments. In Sections 3.2.2 and 3.3.2 we provide a method for evaluating higher order moments (and marginals) from noisy observations. Our approach is based on an equivalent of Isserlis theorem for tree-structured Ising models that is also of independent interest.

### 3.2 Main Results

The main question asked in this work is as follows: what is the impact of noise on the sample complexity of learning a tree-structured graphical model in order to make predictions? This corresponds to sampling variables \( \mathbf{Y} \) generated by sampling \( \mathbf{X} \) from the model (2.3) and randomly flipping each sign independently with probability \( q \). We use the Chow-Liu algorithm to estimate the hidden structure using the noise-corrupted samples. We first find upper (Theorem 1) and lower bounds (Theorem 2) on the sample complexity for exact hidden structure recovery using the Chow-Liu algorithm on noisy observations.

Secondly, we use the structure statistic to derive an accurate estimate of the hidden
layer’s probability distribution. The distribution estimate is computed to be accurate under the ssTV utility measure, that was introduced by Bresler and Karzand [29]. Furthermore, the estimator of the distribution factorizes with respect to the structure estimate, while the ssTV metric ensures that the estimated distribution is a trustworthy predictor. Theorem 6 and Theorem 7 give the sufficient and necessary sample complexity for accurate distribution estimation from noisy samples. These theorems generalize the results for the noiseless case \( q = 0 \) by Bresler and Karzand [29] and lead to interesting connections between structure learning on hidden models and data processing inequalities [82, 66].

The third part of the results includes Theorem 8, which gives an equivalent of Isserlis’s theorem by providing closed form expressions for higher order moments of sign-valued Markov fields on trees. Based on Theorem 8 we propose a low complexity algorithm to estimate any higher order moment of the hidden variables given the estimated tree structure and estimates of the pairwise correlations (both evaluated from observations corrupted by noise).

Finally, Theorem 9 gives the sufficient number of samples for distribution estimation, when the symmetric KL divergence is considered as utility measure. These give rise to extensions of testing algorithms [55] under a hidden model setting.

### 3.2.1 Predictive Learning from Noisy Observations

In addition to recovering the structure of the hidden Ising model, we are interested in estimating the distribution \( p(\cdot) \in \mathcal{P}_T(\alpha, \beta) \) itself. If the \( \mathcal{L}^{(2)} \) distance between the estimator and the true distribution is sufficiently small, then the estimated distribution is appropriate for predictive learning because of (2.15). For consistency, this distribution should factorize according to the structure estimate \( T^{CL}_\dag \) and for the predictive learning part, the estimate \( T^{CL}_\dag \) is considered the output of the Chow-Liu algorithm (see Algorithm
1. We continue by defining the distribution estimator of \( p(\cdot) \) as

\[
\Pi_{T^*_{\text{CL}}} (\hat{p}^*) \triangleq \frac{1}{2} \prod_{(i,j) \in E_{T^*_{\text{CL}}}} \frac{1 + x_i x_j \hat{\mu}_{i,j}^*}{2 (1 - 2q)^2}.
\] (3.8)

The estimator (3.8) can be defined for any \( q \in [0, 1/2) \). For \( q = 0 \) it reduces to that in the noiseless case, since \( T^*_{\text{CL}} = T_{\text{CL}}, \hat{\mu}_{i,j}^* \equiv \hat{\mu}_{i,j} \), and thus \( \Pi_{T^*_{\text{CL}}} (\hat{p}^*) \equiv \Pi_{T_{\text{CL}}} (\hat{p}) \). It is also closely related to the reverse information projection onto the tree-structured Ising models [29, supplementary material, Appendix A], in the sense that

\[
\Pi_T (P) = \arg\min_{Q \in \mathcal{P}_T (\alpha, \beta)} D_{\text{KL}} (P || Q), \quad P \in \mathcal{P}_T (\alpha, \beta).
\] (3.9)

To compute \( \Pi_{T^*_{\text{CL}}} (\hat{p}^*) \), two sufficient statistics are required: the structure \( T^*_{\text{CL}} \) and the set of second order moments [3, 29], under the assumption that \( q \) is known. The next result provides a sufficient condition on the number of samples to guarantee that the \( L(2) \) distance between the true distribution and the estimated distribution is small with probability at least \( 1 - \delta \).

Note that the dependence on \( \beta \) changes from \( e^{2\beta} \) to \( e^{4\beta} \) when the data are noisy \( q > 0 \), while for \( q = 0 \) our bound exactly recovers the noiseless case [29]. A key component of the bound is the following function

\[
\Gamma (\beta, q) \triangleq \left( \frac{1 - (1 - 2q)^2}{1 - (1 - 2q)^4 \tanh^2 (\beta)} \right)^2, \quad \beta > 0 \text{ and } q \in [0, 1/2).
\] (3.10)

Further, notice that \( \Gamma (\beta, q) \in [0, 1] \) for all \( \beta > 0 \) and \( q \in [0, 1/2) \), and \( \Gamma (\beta, 0) = 0 \) for all \( \beta > 0 \). Additionally, we define the functions

\[
K (\beta, q) \triangleq \frac{10 (1 - \tanh^2 (\beta))}{9 + (1 - 2q)^2 - \tanh^2 (\beta) (1 - 2q)^2 (9 (1 - 2q)^2 + 1)},
\] (3.11)

and

\[
B (\beta, q) \triangleq \max \left\{ \frac{1}{K (\beta, q)}, \left( 1 + 2e^\beta \sqrt{2 (1 - q) q \tanh (\beta)} \right)^2 \right\}.
\] (3.12)
The latter constitute additional components of the main result that follows.

**Theorem 6.** Fix $\delta \in (0, 1)$ and choose $\eta > 0$. If

$$n \geq \max \left\{ \frac{512}{\eta^2 (1 - 2q)^4}, \frac{1152 e^{2\beta} B(\beta, q)}{(1 - 2q)^4}, \frac{48 e^{4\beta}}{\eta^2} \Gamma(\beta, q) \right\} \log \left( \frac{6p^3}{\delta} \right)$$

(3.13)

then

$$P \left( \mathcal{L}^{(2)} \left( p(\cdot), \Pi_{T_{\hat{\Theta}}} \left( \hat{\Theta} \right) \right) \leq \eta \right) \geq 1 - \delta.$$  

(3.14)

(3.13) and the inequalities $\Gamma(\beta, q) \leq 1_{q \neq 0}$, $B(\beta, q) \leq (1 + 3\sqrt{q})^2 e^{2\beta \eta \neq 0}$ give Theorem 5. We provide the proof of Theorem 5 and Theorem 6 in Section 3.5.5 (Appendix).

As we mentioned in Section 3.1.2, the sample complexity for accurate predictive learning does not depend on $\alpha$, that is, even in the high temperature regime $\alpha \to 0$ (and in contrast with the structure learning), the number of required samples does not increase.

Conversely, the following result provides the necessary number of samples for small $\mathcal{L}^{(2)}$ distance by a minimax bound, that characterizes any possible estimator $\psi$. In other words, it provides the necessary number of samples required for accurate distribution estimation, appropriate for predictive learning (small $\mathcal{L}^{(2)}(\cdot)$).

**Theorem 7** (Necessary number of samples for inference). Fix a number $\delta \in (0, 1)$. Choose $\eta > 0$ such that $\tanh(\alpha) + 2\eta < \tanh(\beta)$. If the given number of samples satisfies the inequality

$$n_{\hat{\Theta}} < \frac{1 - \left[ \tanh(\alpha) + 2\eta \right]^2}{16\eta^2 [1 - (4q(1 - q))^{\eta}]} \log p,$$

(3.15)

then for any algorithm $\psi$, it is true that

$$\inf_{\psi} \sup_{p(\cdot) \in \mathcal{F}_T(\alpha, \beta)} P \left( \mathcal{L}^{2} \left( p(\cdot), \psi(Y_{1:n}) \right) > \eta \right) > \frac{1}{2}.$$  

Theorems 6 and 7 reduce to the noiseless setting for $q = 0$, that has been studied earlier by Bresler and Karzand [29]. Similarly to our structure learning results, presented
previously (Theorem 1, Theorem 2), when \( q \to 1/2 \) we have \( n^\dagger \to \infty \), the latter indicates that the learning task becomes impossible for \( q = 1/2 \).

**Remark 1.** Theorem 7 requires the assumption \( \alpha < \beta \). The special case \( \alpha = \beta \) can be derived by applying the same proof technique of Theorem 7 combined with Theorem 3 by [29, supplementary material] and the SDPI by [66].

Further details and proof sketches of Theorems 6 and 7 are provided in Section 3.3.1.

### 3.2.2 Estimating Higher Order Moments of Signed-Valued Trees

A collection of moments is sufficient to represent completely any probability mass function. For many distributions, the first and second order moments are sufficient statistics; this is true, for instance, for the *Gaussian distribution* or the *Ising model* with unitary and pairwise interactions. Even further, in the Gaussian case, the well-known Isserlis Theorem (Isserlis [90]) gives a closed form expression for all moments of every order. As part of this work, we derive the corresponding moment expressions, for any tree-structured Ising model. To derive the expression of higher order moments, we first prove a key property of tree structures: for any tree structure \( T = (\mathcal{V}, \mathcal{E}) \) and a even-sized set of nodes \( \mathcal{V}' \subset \mathcal{V} \), we can partition \( \mathcal{V}' \) into \( |\mathcal{V}'|/2 \) pairs of nodes, such that the path along any pair is disjoint with the path of any other pair (see Appendix A, Lemma 9). We denote as \( C_T(\mathcal{V}') \) the set of distinct \( |\mathcal{V}'|/2 \) pairs of nodes in \( \mathcal{V}' \), such that \( \text{path}(u, u') \cap \text{path}(w, w') = \emptyset \), for all \( \{u, u'\}, \{w, w'\} \in C_T(\mathcal{V}') \). Let \( \mathcal{CP}_T(\mathcal{V}') \) be the set of all edges in all mutually edge-disjoint paths with endpoints the pairs of nodes in \( \mathcal{V}' \), that is,

\[
\mathcal{CP}_T(\mathcal{V}') \triangleq \bigcup_{\{w, w'\} \in C_T(\mathcal{V}')} \text{path}_T(w, w'). \tag{3.16}
\]

For any tree \( T \), the set \( \mathcal{CP}_T(\mathcal{V}') \) can be computed via the Matching Pairs algorithm, Algorithm 2. By using the notation above, we can now present the equivalent of Isserlis’s Theorem. The closed form expression of moments is given by the next theorem.

**Theorem 8.** For any distribution of the form of (2.4), which factorizes according to a
Algorithm 2 Matching Pairs

**Require:** Tree structure $T = (\mathcal{V}, \mathcal{E})$, any set $\mathcal{V}' \subset \mathcal{V}$ : $|\mathcal{V}'| \in 2\mathbb{N}$

1: $\mathcal{CP}_T \leftarrow \emptyset$
2: for $i \in \mathcal{V}$ do
3: if $i \in \mathcal{V}'$ then
4: $p(i) \leftarrow 1$
5: else
6: $p(i) \leftarrow 0$
7: end if
8: end for
9: for $k \in [d]$ do
10: Store all nodes at level $k$ to $L(k)$
11: end for
12: for $k \in [d]$ do
13: for $i \in L(d+1-k)$ do
14: if $p(i) = 1$ then
15: $\mathcal{V}' \leftarrow \mathcal{V}' \setminus \{i\}$
16: $\mathcal{CP}_T \leftarrow \mathcal{CP}_T \cup (i, \text{ancestor}(i))$
17: if $p(\text{ancestor}(i)) = 1$ then
18: $\mathcal{V}' \leftarrow \mathcal{V}' \setminus \{\text{ancestor}(i)\}$
19: $p(\text{ancestor}(i)) \leftarrow 0$
20: else
21: $p(\text{ancestor}(i)) \leftarrow 1$
22: end if
23: end if
24: if $\mathcal{V}' \equiv \emptyset$ then
25: return $\mathcal{CP}_T$
26: end if
27: end for
28: end for

Theorem 8 is an equivalent of Isserlis’s theorem for tree-structured sign-valued distributions. Equation (3.17) is used later to define an estimator of higher order moments that requires two sufficient statistics: the estimated structure $T_{\text{CL}}^{\dagger}$ and the correlation estimates $\hat{\mu}_e^{\dagger}$, for any $e \in T_{\text{CL}}^{\dagger}$. Together with the parameter $q$, the higher order moments completely characterize the distribution of the noisy variables of the
hidden model (2.13). We provide the proof of Theorem 8 in Appendix 3.5.

A similar expression to (3.17) has been introduced in prior work. Specifically, Algorithm 2 solves the problem of finding the optimal matching, see Definition 1 by [29, supplementary material]. The evaluation of higher order moments requires an explicit expression or a way to compute the set $\mathcal{CP}_T$. For a given tree $T = (\mathcal{E}, \mathcal{V})$ and a set $((i_1, i_2, \ldots, i_k) \subset \mathcal{V}$, there is a unique set $\mathcal{CP}_T(i_1, i_2, \ldots, i_k)$ (see Appendix 3.5, proof of Theorem 8). Given a set of edges $\mathcal{E}$, we show that the set $\mathcal{CP}_T$ can be evaluated by running a matching pair algorithm. For that purpose, we provide Algorithm 2 (with complexity $O(|\mathcal{E}|)$)) and we prove its consistency (See Appendix, Lemma 9). The latter yields to an explicit expression of higher order moments; the Theorem 8. Furthermore, it provides a concrete higher order moments estimator, that is based on the estimated structure $T^{CL}$ (or $T^{CL}_\dag$) and the set of estimated correlations $\{\hat{\mu}_e : e \in \mathcal{CP}_{T^{CL}}\}$.

**High Order Moments Estimator:** A higher order moment is the expected value of the product of the hidden tree-structured Ising model variables $\{X_i : i \in \mathcal{V}'\}$ where $\mathcal{V}' \subset \mathcal{V}$. Theorem 8 gives the closed form solution for such moments. We have the following estimator for higher order moments using only noisy observations and known $q$. In particular, we have

\[
\hat{E}[X_{i_1}X_{i_2} \ldots X_{i_k}] \equiv 0, \quad k \in 2\mathbb{N} + 1,
\]

\[
\hat{E}[X_{i_1}X_{i_2} \ldots X_{i_k}] \triangleq \prod_{e \in \mathcal{CP}_{T^{CL}}(i_1, i_2, \ldots, i_k)} \hat{\mu}_e^\dagger \frac{\hat{\mu}_e^\dagger}{(1 - 2q)^2}, \quad k \in 2\mathbb{N}.
\] (3.18) (3.19)

The estimated structure and pairwise correlations are sufficient statistics: given those, (3.19) suggests a computationally efficient estimator for higher order moments. First we run the classical Chow-Liu algorithm to estimate the tree structure $T^{CL}_\dag$, and then we run Algorithm 2 with input the estimate $T^{CL}_\dag$ to evaluate the set $\mathcal{CP}_{T^{CL}_\dag}$. Thus, by estimating $T^{CL}_\dag$, $\mathcal{CP}_{T^{CL}_\dag}$ and $\hat{\mu}_e^\dagger$ for any $e \in T^{CL}_\dag$, we can in turn estimate any higher
order moment through (3.19). Considering the absolute estimation error, we have

\[ \left| \mathbb{E} \left[ \prod_{s \in V'} X_s \right] - \mathbb{E} \left[ \prod_{s \in V'} X_s \right] \right| \leq 2|V'\mathcal{L}(2)\left( p(\cdot), \Pi_{\hat{T}^*} \right) \right). \tag{3.20} \]

Theorem 6 guarantees small ssTV and in combination with (3.20) gives an upper bound on the higher order moment estimate (3.19). In Section 3.3.2, we provide further details and discussion about Theorem 8, Algorithm 2, that computes the sets \( \mathcal{C}P_T(V'), \mathcal{C}P_{\hat{T}^*}(V') \), and the bound on the error of estimation (3.20).

So far we have studied the consistency of the estimator with respect to the \( \mathcal{L}(2) \) metric. We are also interested in sample complexity bounds for \( \phi \)-divergences. While general divergences may be challenging, the most widely-used is the KL-divergence, particularly in testing Ising models [55]. The next result gives a bound for the sufficient number of samples to guarantee a small symmetric KL divergence \( S_{KL}(P||Q) \triangleq D_{KL}(P||Q) + D_{KL}(Q||P) \) with high probability. For any Ising model distributions \( P, Q \) of the form (2.3) with respective interaction parameters \( \theta, \theta' \), we have

\[ S_{KL}(\theta||\theta') \triangleq S_{KL}(P||Q) = \sum_{s,t \in \mathcal{E}} (\theta_{st} - \theta'_{st}) (\mu_{st} - \mu'_{st}). \tag{3.21} \]

**Theorem 9** (Upper Bounds for the Symmetric KL Divergence). If the number of samples \( n_{\hat{T}} \) of \( Y \) satisfies

\[ n_{\hat{T}} \geq 4 \frac{\beta^2(p - 1)^2}{(1 - 2q)^4q^2} \log \left( \frac{p^2}{\delta} \right), \tag{3.22} \]

then for \( p(\cdot) \in \mathcal{P}_T(\alpha, \beta) \) we have

\[ \mathbb{P} \left( S_{KL} \left( p(\cdot)||\Pi_{\hat{T}^*(\hat{p}_T)} \right) \leq \eta_s \right) \geq 1 - \delta, \tag{3.23} \]

\( \hat{T}^* \) is the Chow-Liu tree estimate from noisy data, and the estimate \( \Pi_{\hat{T}^*}(\hat{p}_T) \) is given by (3.8).

The bound in (3.22) for the noiseless setting was recently studied by Daskalakis, Dikkala,
3.2.3 Simulations

We provide empirical results based on synthetic data to illustrate the probability of error $\delta$ as function of the cross-over probability $q$ and the number of samples $n$. For the simulations of this chapter the original tree structure $T$ is generated randomly by choosing the parent of each edge sequentially and uniformly distributed. First, we estimate the probability of error $\mathbb{P}(T^{CL}_{i} \neq T)$ (named as $\delta$) of the structure learning problem, Figure 3.3. For the structure learning experiments, the number of nodes is 100, $\beta = \text{arctanh}(0.8)$, and $\alpha = \text{arctanh}(0.2)$. Further, we considering 100 Monte Carlo runs for averaging, and we plot the estimated probability of incorrect structure recovery while $q$ and $n$ vary. As a next step, we would like to see how well the theoretical bound
Figure 3.4: Estimate of the probability of the ssTV to be greater than $\eta = 0.03$. The theoretical bound is given by Theorem 6. The top view of the figure is Figure 3.2 and provides a clear comparison between the experimental and theoretical results.

Figure 3.5: Estimate of the distribution error metric ssTV as a function of $q$ and $n$. 
Figure 3.6: The probability of error in the predictive learning task for different values of \( n \) and estimation error of the parameter \( q \). For both figures we consider \( p = 31, \alpha = 0.2, \beta = 1, q = 0.1 \) and averaging over 500 independent runs. Left: \( \hat{q} \in [0.05, 0.15], \eta = 0.1 \), Right: \( \hat{q} \in [0, 0.2], \eta = 0.12 \).

of Theorem 1 matches with the experimental results. To do this we plot the top view of Figure 3.3 to get Figure 3.1. Quite remarkably, the theoretical and experimental bounds exactly match. The latter suggests that our theoretical bound that we derive, sample complexity of the Chow-Liu algorithm (Theorem 1), is indeed accurate. Second, we plot the probability of error for the predictive learning task, that is the probability of the ssTV to be greater than a positive number \( \eta \) (Figure 3.4). For the simulation part, we restrict our attention to the case that the first of three terms in the maximization of (3.13) is the dominant. In fact, \( \eta = 0.03, p = 31 \), while \( \alpha \) and \( \beta \) are the same as the structure learning. Finally, Figure 3.5 presents the ssTV itself for different values of \( q \) and \( n \). Finally, the top view of Figure 3.4 is Figure 3.2, the latter suggest that the bound of our main result, Theorem 6 is accurate.

Finally, we provide experimental results for the case of unknown \( q \). Specifically, Figure 3.6 illustrates the relationship between the average probability of error and the relative error \( |\hat{q} - q|/q \) for the predictive learning task. We notice that the distribution can be approximated by using an estimate \( \hat{q} \) of \( q \) even for relative error 30% or 60%.

3.3 Discussion

In this section, we present sketches of proofs, and we further elaborate on Algorithm 1, Algorithm 2 and the error of higher order moment estimates. In the next Section 3.3.1,
we present the analysis and a sketch of proof for Theorem 6. Then, in Section 3.3.2, we provide further details about Theorem 8, discussion about the Matching Pairs algorithm (Algorithm 2) and the accuracy of the proposed higher order moments estimator (3.19).

3.3.1 Theorem 6: A Sketch of the Proof

Recall that the indices \( i, j \in V \) of the quantities \( \mu_{i,j}^\dagger \) and \( \hat{\mu}_{i,j}^\dagger \) are pair of nodes, and in fact they can be considered as one (pair) index. For sake of space we introduce the notation \( \mu_e^\dagger \) and \( \hat{\mu}_e^\dagger \) for some \( e \in \mathcal{E}_T \), that is consistent with our previous definition and \( e \) represents a pair of nodes. Theorem 6 guarantees that the estimated pairwise marginal distributions are close to the original distributions by ensuring that the \( \mathcal{L}^{(2)} \) is small.

In this section we provide a sketch of the proof of the Theorem and we mention the main differences between the hidden model and the noiseless case [29]. The intersection of three events is sufficient to guarantee that \( \mathcal{L}^{(2)} \) is upper bound by \( \eta > 0 \):

\[
E_{\text{corr}}^\dagger (\epsilon_\dagger) \triangleq \left\{ \sup_{i,j \in V} \left| \mu_{i,j}^\dagger - \hat{\mu}_{i,j}^\dagger \right| \leq \epsilon_\dagger \right\}, \quad (3.24)
\]

\[
E_{\text{strong}}^\dagger (\epsilon_\dagger) \triangleq \left\{ e \in \mathcal{E}_T : \left| \tanh \theta_e \right| \geq \frac{\tau^\dagger (\epsilon_\dagger)}{(1 - 2q)^2} \right\} \subseteq \mathcal{E}_{T_C}^\dagger, \quad (3.25)
\]

\[
E_{\text{cascade}}^\dagger (\gamma_\dagger) \triangleq \left\{ \prod_{e \in \text{path}_T(i,j)} \frac{\hat{\mu}_e^\dagger}{(1 - 2q)^2} - \prod_{e \in \text{path}_T(i,j)} \frac{\mu_e^\dagger}{(1 - 2q)^2} \leq \gamma_\dagger : i, j \in V \right\}, \quad (3.26)
\]

(2.80) gives the definition of \( \tau^\dagger (\epsilon_\dagger) \). The three events are equivalent events of the noiseless case, but they are modified accordingly to guarantee accurate estimation based on noisy data. The event \( E_{\text{corr}}^\dagger (\epsilon_\dagger) \) guarantees that the error of the correlation estimates is not greater than \( \epsilon_\dagger \). Under the event \( E_{\text{strong}}^\dagger (\epsilon_\dagger) \) all the strong edges are recovered by the Chow-Liu algorithm. Similarly to the noiseless setting, the event \( E_{\text{strong}}^\dagger (\epsilon_\dagger) \) requires the Chow-Liu algorithm to recover all the strong edges, while the weak edges (those that do not satisfy the inequality in (3.25)) do not affect the accuracy of the predictive learning, even if the Chow-Liu algorithm fails to recover them. In contrast with structure learning, exact structure recovery is not necessary for the predictive learning task. In other words, even if \( \alpha \) is extremely small, assume \( \tanh(\alpha) \leq \tau^\dagger (\epsilon_\dagger)/(1 - 2q)^2 \) the required...
number of samples for accurate predictive learning will remain unaffected.

Under the event $E_{\gamma}^\text{cascade}$ the end-to-end error along paths is no greater than $\gamma$. In fact, each path between two nodes of the tree can be considered a sequence of segments with strong and weak edges. The end-to-end path error is determined by the strong edge segments of the path through the parameter $\gamma$ for the $E_{\gamma}^\text{cascade}$ event, while the effect of weak edges parameters is controlled by the quantity $\tau(\epsilon)$ (for the segmentation of the tree and the detailed proof see 15). Our goal is to find sufficient conditions on the parameters $\epsilon$ and $\gamma$ that guarantee that the events $E_{\epsilon}^\text{corr}$, $E_{\gamma}^\text{cascade}$ and $E^\text{strong}$ occur with high probability.

Recall that our goal is to guarantee that the quantity $\mathcal{L}^2(p(\cdot), \Pi_{T_{CL}^\dagger}(\hat{p}))$ is smaller than a fixed number $\eta > 0$ with probability at least $1 - \delta$. To do this, we follow the technique of prior work by [29], the triangle inequality gives

\[
\mathcal{L}^2(p(\cdot), \Pi_{T_{CL}^\dagger}(\hat{p})) \leq \mathcal{L}^2(p(\cdot), \Pi_{T_{CL}}(p(\cdot))) + \mathcal{L}^2(\Pi_{T_{CL}}(p(\cdot)), \Pi_{T_{CL}^\dagger}(\hat{p})),
\]

and we find the required number of samples such that each of the two quantities $\mathcal{L}^2(p(\cdot), \Pi_{T_{CL}}(p(\cdot)))$ and $\mathcal{L}^2(\Pi_{T_{CL}}(p(\cdot)), \Pi_{T_{CL}^\dagger}(\hat{p}))$ is no greater than $\eta/2$ with probability at least $1 - \delta$. As we show the probability of the event $E_{\gamma}^\text{cascade}$ (Lemma 14, Appendix) and the $\mathcal{L}^2$ (between the true and estimated distribution) can be bounded by a constant uniformly over the set of all trees and is not affected by long paths. To prove these properties of the hidden model is non-trivial and ensures that the estimation error from noisy observations does not increase exponentially along paths as someone might expect. Specifically, the first quantity at the right hand-side of inequality (3.27) represents the loss due to graph estimation error, while the second term represents the loss due to parameter estimation error. Lemma 15 (Appendix) shows that under the event

\[
E_{\epsilon, \gamma} \triangleq E_{\epsilon}^\text{corr} \cap E_{\gamma}^\text{cascade} \cap E^\text{strong}.
\]
if 
\[ \gamma_t \leq \frac{\eta}{3} \text{ and } \epsilon_t \leq (1 - 2q)^2e^{-\beta} \left[ 20 \left(1 + 2e^{\beta}\sqrt{2(1 - q)q\tanh(\beta)} \right) \right]^{-1} \] (3.29)

then \( \mathcal{L}^{(2)}(\Pi_{T_{\textrm{CL}}}(\hat{p}(\cdot)), \Pi_{T_{\textrm{CL}}}((\hat{p}_t))) \leq \eta/2 \). Further Lemma 16 (Appendix) shows that if 
\[ \epsilon_t \leq \min \left\{ \frac{\eta}{16}(1 - 2q)^2, \frac{(1 - 2q)^2e^{-\beta}}{24 \left(1 + 2e^{\beta}\sqrt{2(1 - q)q\tanh(\beta)} \right)} \right\} \] (3.30)

then \( \mathcal{L}^{(2)}\left(p(\cdot), \Pi_{T_{\textrm{CL}}} (p(\cdot)) \right) \leq \frac{\eta}{2} \) under the event \( E_{\text{corr}}^t (\epsilon_t) \cap E_{\text{strong}}^t (\epsilon_t) \). Both conditions (3.29) and (3.30) should be satisfied, so it is necessary to have
\[ \gamma_t \leq \frac{\eta}{3} \text{ and } \epsilon_t \leq \min \left\{ \frac{\eta}{16}(1 - 2q)^2, \frac{(1 - 2q)^2e^{-\beta}}{24 \left(1 + 2e^{\beta}\sqrt{2(1 - q)q\tanh(\beta)} \right)} \right\} \] (3.31)

To guarantee that the errors \( \gamma_t \) and \( \epsilon_t \) are sufficiently small such that (3.31) is satisfied, we need to make sure that the number of samples \( n \) is sufficiently large. In fact, the upper bounds on the errors translate into lower bounds on the number of samples through the concentration bounds for the events. Specifically, Lemma 11 gives a sufficient sample size to ensure that the event \( E_{\text{corr}}^t (\epsilon_t) \) occurs with probability at least \( 1 - \delta \), Lemma 4 gives the concentration bound for the event \( E_{\text{strong}}^t (\epsilon_t) \) and Lemma 14 gives the concentration bound of the event \( E_{\text{cascade}}^t (\gamma_t) \). Lemma 11, Lemma 4 and Lemma 14 together with (3.31) give the final bound of the sample complexity (see the proof 11)
\[ n \geq \max \left\{ \frac{512}{\eta^2(1 - 2q)^4}, \frac{1152e^{2\beta}B(\beta, q)}{(1 - 2q)^4}, \frac{48e^{4\beta}}{\eta^2} \Gamma(\beta, q) \right\} \log \left( \frac{6p^3}{\delta} \right) \] (3.32)
and its simplified but looser bound
\[ n \geq \max \left\{ \frac{512}{\eta^2(1 - 2q)^4}, \frac{1152(1 + 3\sqrt{q})^2e^{2\beta(1 + t_{\delta,q}^\beta)}}{(1 - 2q)^4}, \frac{48e^{4\beta}}{\eta^2} \Gamma_{\delta,q} \right\} \log \left( \frac{6p^3}{\delta} \right) \] (3.33)
that provides the condition of Theorem 5. Although the general structure of our argument follows that of the noiseless case, the presence of noise introduces several
technical challenges whose solution may be of independent interest. In the sequel, we highlight the most important aspects of our approach that do not appear in the noiseless case.

The proof of Theorem 3.3 is significantly different and includes additional steps and techniques compared with the approach by [29]. Specifically, Lemma 12 is new and it is necessary for the hidden model and we use it later to prove (Lemma 14, Appendix). Lemma 13 is an non-trivial extension of the accurate estimation of edges’ correlation. Although the resulting expression seems complicated is important for the proof of Lemma 14. In fact Lemma 14, the proof of the concentration bound for the event $E_{\text{cascade}}^{\dagger}(\gamma_\dagger)$, is significantly more complicated and longer than the noiseless model (see Appendix E by [29] for comparison). To show this result we have to consider a martingale difference sequence and evaluate upper bounds for the conditional variance and bias of that sequence. The bias is crucial for the final result because it introduces an extra term in the final bound that does not exist in the noiseless case. It is interesting that this term does not involve any parameter related to the noise and shows how the result is affected by the structural inconsistency between the hidden and the observable layer. As a consequence, the expression of the bound (3.95) in Lemma 14 involves two inequalities to guarantee the high-probability bound. The first inequality which introduces restrictions on the parameter $\Delta$ (see (3.95)) is an attribute of the noisy case.

We continue by briefly explaining one of the main technical aspects of the proof.

To begin with, consider a path of length $d \geq 2$ in the original tree $T$, $X_1 - X_2 - \cdots - X_{d+1}$ and we denote the edge $(k, k + 1)$ as $e_k$, for some $k \in [d]$. Recall that $Y_k^{(i)}$ denotes the $i^{th}$ sample of $Y_k$ and $k \in [d + 1]$. We would like derive a concentration bound of the probability of the event $E_{\text{cascade}}^{\dagger}(\gamma_\dagger)$ (Lemma 14, Appendix). To do this, first we have to consider for all $\ell \in [n]$ and $k \in \{2, \ldots, d\}$ the random variables

$$Z_k^{(\ell)} \triangleq \left( \frac{(X_k N_k X_{k+1} N_{k+1})^{(\ell)}}{(1 - 2q)^2} - \frac{\hat{\mu}_{e_k}^\dagger}{(1 - 2q)^2} \right) \prod_{j=1}^{k-1} \frac{\hat{\mu}_{e_j}^\dagger}{(1 - 2q)^2} \prod_{j=k+1}^{d} \frac{\hat{\mu}_{e_j}^\dagger}{(1 - 2q)^2}. \quad (3.34)$$

Define the martingale difference sequence (MDS) \{\xi_k^{(i)}\} by setting $\xi_k^{(0)} \triangleq 0$, $\xi_k^{(1)} \triangleq Z_k^{(1)} - \mathbb{E}\left[Z_k^{(1)} | \hat{\mu}_{e_{k-1}}, \ldots, \hat{\mu}_{e_1}\right]$, $\xi_k^{(i)} \triangleq Z_k^{(i)} - \mathbb{E}\left[Z_k^{(i)} | Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{e_{k-1}}, \ldots, \hat{\mu}_{e_1}\right]$. Let
\( \mathcal{F}_{i-1}^k \) be the \( \sigma \)-algebra generated by \( Z_{-1}^{k(i-1)}, \ldots, Z_{-1}^{k(l)} \), \( \hat{\mu}_{e_{k-1}}, \ldots, \hat{\mu}_{e_1} \). Then the pair

\( (\xi_{k(i)^i}, \mathcal{F}_{i-1}^k)_{i=1,\ldots,n} \)

is a MDS. In contrast with the noiseless case, the conditional means are not zero, which makes the problem significantly harder. To proceed, we apply a concentration bound for supermartingales (generalized Bennett’s inequality) by [68].

Secondly we have to evaluate the following expression

\[
P \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = \pm 1 \mid \hat{\mu}_{e_{k-1}}, \ldots, \hat{\mu}_{e_1} \right) = \frac{1 \pm \mu_{e_k}^{\dagger}}{2} \frac{1 - \mu_{e_{k-1}}^{\dagger} \hat{\mu}_{e_{k-1}}^{\dagger} + \mu_{e_{k-1}}^{\dagger}}{1 - (\mu_{e_{k-1}}^{\dagger})^2} \frac{1 \pm \mu_{e_k}^{\dagger} \hat{\mu}_{e_{k-1}}^{\dagger} - \mu_{e_{k-1}}^{\dagger}}{1 - (\mu_{e_{k-1}}^{\dagger})^2}. \tag{3.35}
\]

In the noiseless case, the product variables \( X_k^{(\ell)} X_{k+1}^{(\ell)} \) are independent, leading to a simple expression for this probability (see Lemma 7, Appendix). The closed form expression of (3.35) is given by Lemma 12. Finally, the conditional expectations

\[
\mathbb{E} \left[ Z_{k}^{(i)} \mid Z_{k}^{(i-1)}, \ldots, Z_{k}^{(l)}, \hat{\mu}_{e_{k-1}}, \ldots, \hat{\mu}_{e_1} \right]
\]

are not zero, however when \( n \to \infty \), they approach zero. As a consequence, a bias exists that affects the sample complexity by introducing an additional term in the bound that does not appear in the noiseless case, the quantity \( e^{4\beta/\eta^2} \) (see (3.32) and (3.33)).

Finally, we continue by bounding the norm \( \mathcal{L}^{(2)} \) between the true and estimated distribution in Appendix E. The proof of Lemma 15 shows that in the noisy setting as well, the \( \mathcal{L}^{(2)} \) can be bounded by a constant uniformly over the set of all trees and it is not affected by long paths. This property of the hidden model is highly non-trivial and ensures that the estimation error from noisy observations does not increases along paths as someone might expect. Lemma 16 follows the corresponding approach of Lemma 6.1 by [29] and we provide only the required for the noisy setting differences. In Theorem 11, we combine the Lemmata of Appendices 3.5.4 and 3.5.5, we find the appropriate choice of the parameter \( \Delta \) that satisfies the necessary conditions of Lemma 14 and we derive the final sample complexity bound. For further details about the proof of Theorem 6 see Appendix, Section 3.5.4 and Section 3.5.5.
3.3.2 Estimating Higher Order Moments

Our results also provide an analogue of Isserlis’s Theorem (Theorem 8) and the Matching Pairs algorithm, which returns the set $CP_T(V')$ in (3.17). We provide a short proof sketch for the bound on the error of estimation (3.20).

Proof sketch of Theorem 8: We prove that $C_T(V')$ always exists (when $k$ is even) by induction (see Appendix A, Lemma 9). We define the set of edges $CP_T(V')$ as the union of the edge-disjoint paths $3 CP_T(V') = \bigcup_{w,w' \in C_T(V')} \text{path}(w,w')$. Combining the set $CP_T(V')$ together with the independent products property (see Lemma 7), we derive the final expression (see Appendix A, proof of Theorem 8). Given the tree structure $T$ and the correlations $\mu_e$ for all $e \in E$, we can calculate the higher order expectations. Notice that the collection of edge-disjoint paths $CP_T$ depends on the tree structure and as a consequence an algorithm is required to discover those paths. Different matching algorithms can be considered to find the set $CP_T$. We propose Algorithm 2 which is simple and has low complexity of $O(|E|)$.

Matching Pairs Algorithm: Algorithm 2 requires as input the tree and the set of nodes $V' \equiv \{i_1, \ldots, i_k\} \subset V$, and returns the set of edges $CP_T(V')$. For each node in the tree, a flag variable is assigned to each node and indicates if the corresponding node is a candidate for the final set $C_T(V')$ at the current step of the algorithm. The candidate nodes have to be matched with other nodes of the tree, such that the pairs generate edge-disjoint paths. Initially, the candidate nodes are the nodes of the set $V'$. Starting from the nodes which appear in the deepest level of the tree, we “move” them to their ancestor. At each step, if two candidate nodes appear at the same point, we match them as pair, we store the pair in the set $CP_T(V')$ and we remove both of them from the set $V'$. We continue until $V' \equiv \emptyset$. The complexity of Algorithm 2 is $O(|E|)$. Finally, Theorem 8 can be extended to any forest $F$ structure by considering the set $CP_F(V')$ instead of $CP_T(V')$, where $CP_F(V') \equiv \bigcup_i CP_{T_i}(V')$ and $T_i$ is the $i^{th}$ connected tree of the forest.

Estimation error of higher order moments: Inequality (3.20) bounds the error

---

3By edge-disjoint paths we refer to paths with no common edges.
of estimation by the small set Total Variation (ssTV), that is guaranteed to be less than \( \eta > 0 \) by Theorem 6. Additionally, the bound on the error of the estimation in (3.20) can be found as follows

\[
\left| \hat{E} \left[ \prod_{s \in V'} X_s \right] - E \left[ \prod_{s \in V'} X_s \right] \right|
\]

\[= \prod_{e \in CP_{T^{\dagger}_{CL}}(i_1,i_2,...,i_k)} \frac{\hat{\mu}_e^{\dagger}}{(1 - 2q)^2} - \prod_{e \in CP_{T}(i_1,i_2,...,i_k)} \mu_e \]  

\[= \prod_{e \in CP_{T^{\dagger}_{CL}}(i_1,i_2,...,i_k)} \frac{\hat{\mu}_e^{\dagger}}{(1 - 2q)^2} - \prod_{e \in CP_{T}(i_1,i_2,...,i_k)} \frac{\mu_e^{\dagger}}{(1 - 2q)^2} \]  

\[= \prod_{e \in \bigcup \{w,w'\} \in C_{T^{\dagger}_{CL}}(V')} \frac{\hat{\mu}_e^{\dagger}}{(1 - 2q)^2} - \prod_{e \in \bigcup \{w,w'\} \in C_T(V')} \frac{\mu_e^{\dagger}}{(1 - 2q)^2} \]  

\[\leq 2|V'||L^{(2)} \left( \hat{p}(\cdot), \Pi_{T^{\dagger}_{CL}}(\hat{p}_T) \right), \]  

(3.38)

where (3.36) holds due to (3.17) and (3.19), (3.37) comes from (3.16) and the last inequality (3.38) is being proved by Bresler and Karzand [29, Lemma 1, supplementary material]. Thus, if we can accurately estimate the distribution under the sense \( L^{(2)} \left( P, \Pi_{T^{\dagger}_{CL}}(\hat{p}_T) \right) \leq \eta' \), for a sufficiently small positive number \( \eta' \), then by using (3.19) and choosing \( \eta' \leq \eta/(2|V'|) \), Theorem 6 guarantees accurate estimates for higher order moments with probability at least \( 1 - \delta \).

### 3.4 Conclusion

We considered the problem of predictive learning on hidden tree-structures from noisy observations, using the well-known Chow-Liu algorithm. In particular, we derived sample
complexity guarantees for exact structure learning and marginal distributions estimation. Our bounds extend prior work (see [29]) to the hidden model, by introducing the cross-over probability $q$ of the BSC($q$)$^p$. Our results exactly reduce to the noiseless setting when $q = 0$, and the explicit expressions of the bounds are also continuous functions of $q$. Additionally, by applying a graph property for tree structures and a probabilistic property for Ising models, we derived an equivalent of the well-known Isserlis’s theorem for Gaussian distributions, which yields a consistent high-order moments estimator for Ising models. Further, we considered simulations based on synthetic data to validate our theoretical results. Our theoretical bounds exactly match with the experiment, indicating that our results correctly characterize the dependence on the model parameters.

Our results show that the estimated structure statistic $T^{\text{CL}}_1$ is essential for successful statistical inference on the hidden (or observable) layer, while the sample complexity with respect to number of nodes and probability of error remains strictly logarithmic, as in the noiseless case. Our hidden setting constitutes a first step towards more technically challenging and potentially more realistic statistical models, such as, for instance, structure and distribution learning when the noise is generated by an erasure channel, or when the underlying hidden tree structured distribution has a larger, or even uncountable, support.

3.5 Appendix

We start by providing an outline of the Appendix, a stream mapping of the proofs and we prove several properties of the model under consideration.

3.5.1 Outline of the Proof, Properties of the Models

The chart in Figure 3.7 shows the various dependencies of the Lemmata and intermediate results either considered or developed in this dissertation, and the resulting Theorems. The proofs can be found in the corresponding section of the Appendix.

For completeness, we start with some properties that hold for any distribution with support $\{-1, +1\}^p$ and tree-structured graphical model [2]. Later we derive explicit
formulas for the Ising model (2.3).

**Lemma 6.** Any distribution \( p(x) \) with respect to a forest \( F = (V, E) \), where \( x \in \{-1, 1\}^p \) and uniform marginals \( \mathbb{P}(X_i = \pm 1) = 1/2 \), for all \( i \in V \) can be expressed as

\[
p(x) = \frac{1}{2} \prod_{(i,j) \in E} \left( 1 + x_i x_j \mathbb{E}[X_i X_j] \right). \tag{3.39}
\]

**Proof.** We prove the result for an arbitrary tree \( T = (V, E) \) and then we extend it to any forest structure by applying cuts to \( E \). The distribution factorizes according to the tree structure \( T \) and under the assumption of no external field (uniform marginal distributions), we have

\[
\mathbb{P}(X = x) = \prod_{i \in V} p(x_i) \prod_{(i,j) \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} = 2^{p-2} \prod_{(i,j) \in E} \left( \frac{1 + x_i x_j \mathbb{E}[X_i X_j]}{2} \right) \tag{3.40}
\]

\[
= \frac{1}{2} \prod_{(i,j) \in E} \left( 1 + x_i x_j \mathbb{E}[X_i X_j] \right). \tag{3.41}
\]
(3.40) holds since the joint distribution of any pair \((X_i, X_j)\) of distinct nodes \(i, j \in V\) is
\[
p(x_i, x_j) = \mathbb{E} \left[ 1_{X_i=x_i} 1_{X_j=x_j} \right] = 1 + x_i x_j \mathbb{E}[X_iX_j].
\]
(3.42)

By setting \(\mathbb{E}[X_iX_j] = 0\) for some \((i, j) \in E\) we derive the distribution with respect to a forest generated by cutting the edge \((i, j)\) of \(T\).

In Lemma 7 we prove two fundamental properties of the model, the independence of the random variables \(\{X_iX_j : (i, j) \in E\}\) and the correlation decay property (CDP). To the best of our knowledge, these properties are known but there is no reference for the corresponding proofs in the literature.

**Lemma 7.** Let \(X\) be a random binary vector in \((-1, +1)^p\) drawn according to a forest-structured distribution \(p(\cdot)\) with uniform marginal distributions on each entry \(X_i\) for \(i \in [p]\). Then the elements of the collection of \(|E|\) random variables \(\{X_iX_j : (i, j) \in E\}\), are independent. Furthermore, we have
\[
\mathbb{E}[X_iX_j] = \prod_{e \in \text{path}(i, j)} \mu_e,
\]
so the Correlation Decay Property (CDP) holds since \(\mu_e \leq 1\) for all \(e \in E\).

**Proof.** Let \((i_r)_{r=1}^p\) be an arbitrary permutation of \(\ell = \{1, 2, \ldots, p\}\). Notice that the singletons \(\{i_r\}, r = 1, \ldots, p\) form a partition of \(\ell\). Then, the set of edges \(E\) is defined as
\[
E = (i_r, j_r)_{r=2}^p, \quad \text{and} \quad j_1 = \emptyset, \quad j_r \in \{i_1, \ldots, i_{r-1}\} \subset \ell.
\]
(3.44)

(3.44) defines a tree \(T = (\mathcal{V}, \mathcal{E})\) with root the node \(i_1\) (since \(j_1 = \emptyset\)). For the first part, it is sufficient to show that for any \(\{c_r : r = 2, 3, \ldots, p\} \in \{-1, +1\}^{p-1}\), the following holds
\[
\mathbb{P} \left( \bigcap_{r=2}^p \{x_i x_{j_r} = c_r\} \right) = \prod_{r=2}^p \mathbb{P}(x_i x_{j_r} = c_r).
\]
(3.45)
We have
\[
\mathbb{P} \left( \cap_{r=2}^p \{ X_{i_r} X_{j_r} = c_r \} \right) = \sum_{x: x_{i_r} x_{j_r} = c_r | r=2} p(x)
\]
\[
= \sum_{x: x_{i_r} x_{j_r} = c_r | r=2} \frac{1}{2} \prod_{r=2}^p \frac{1 + x_{i_r} x_{j_r} \mathbb{E} \left[ X_{i_r} X_{j_r} \right]}{2}
\]
\[
= \sum_{x: x_{i_r} = c_r x_{j_r} | r=2} \frac{1}{2} \prod_{r=2}^p \frac{1 + x_{i_r} x_{j_r} \mathbb{E} \left[ X_{i_r} X_{j_r} \right]}{2}
\]
\[
= \sum_{x_{i_1} \in \{-1, +1\}} \frac{1}{2} \prod_{r=2}^p \frac{1 + c_r \mathbb{E} \left[ X_{i_r} X_{j_r} \right]}{2} = \prod_{r=2}^p \mathbb{P} \left( X_{i_r} X_{j_r} = c_r \right),
\]
(3.47)

(3.46) comes from (3.44) and Lemma 6 and the last from (3.42). For the second part of the statement note that for all \( i, j \in \mathcal{V} \) there exists a unique path \( \{i, k_1, k_2, \ldots, k_\ell, j\} \) from \( i \) to \( j \). Define the variable \( 1_{(i,j)} \triangleq (X_{k_1} X_{k_1})(X_{k_2} X_{k_2}) \ldots (X_{k_\ell} X_{k_\ell}) \), which is equal to 1 almost surely, since \( X \in \{-1, +1\}^p \). Then, we have

\[
\mathbb{E}[X_i X_j] = \mathbb{E}[X_i 1_{(i,j)} X_j]
\]
\[
= \mathbb{E}[X_i (X_{k_1} X_{k_1})(X_{k_2} X_{k_2}) \ldots (X_{k_\ell} X_{k_\ell}) X_j]
\]
\[
= \mathbb{E}[X_i X_{k_1}] \left( \prod_{m=1}^{\ell-1} \mathbb{E}[X_{k_m} X_{k_m+1}] \right) \mathbb{E}[X_{k_\ell} X_j] = \prod_{e \in \text{path}(i,j)} \mu_e, \hspace{1cm} (3.48)
\]
and (3.48) comes from (3.45) and completes the proof.

The next lemma relates the pairwise correlations to the parameters of the Ising model.

**Lemma 8.** An equivalent expression of (2.3) is the following

\[
p(x) = \frac{\prod_{(i,j) \in \mathcal{E}} \left[ 1 + x_i x_j \tanh(\theta_{i,j}) \right]}{\sum_x \prod_{(i,j) \in \mathcal{E}} \left[ 1 + x_i x_j \tanh(\theta_{i,j}) \right]} \quad x \in \{-1, 1\}^p.
\]

Further, for a tree-structure Ising model \( \mathbb{E} \left[ X_i X_j \right] = \tanh(\theta_{i,j}), \) for all \( (i, j) \in \mathcal{E} \).}

\( ^4 \)1_{\cdot} should not be confused with \( 1_{A} \), where the last denotes the indicator function of a set \( A \).
Proof. We can write \( \exp(\theta_{ij}x_i x_j) \) as

\[
\exp(\theta_{ij}x_i x_j) = \frac{\exp(\theta_{ij}x_i x_j) + \exp(-\theta_{ij}x_i x_j)}{2} + \frac{\exp(\theta_{ij}x_i x_j) - \exp(-\theta_{ij}x_i x_j)}{2}
\]

(3.50)

\[
= \frac{\exp(\theta_{ij})}{2} + \frac{\exp(-\theta_{ij})}{2} + \frac{\exp(\theta_{ij})}{2} - \frac{\exp(-\theta_{ij})}{2} + x_i x_j \exp(\theta_{ij}) - \exp(-\theta_{ij})
\]

(3.51)

(3.50) holds because \( x_i x_j \in \{-1, +1\} \). The partition function can be written as

\[
Z(\theta) = \sum_x \prod_{(i,j) \in E} \exp(\theta_{ij}x_i x_j)
\]

\[
= \sum_x \prod_{(i,j) \in E} \cosh(\theta_{ij}) \left[ 1 + x_i x_j \tanh(\theta_{ij}) \right]
\]

\[
= \prod_{(i,j) \in E} \cosh(\theta_{ij}) \sum_x \prod_{(i,j) \in E} \left[ 1 + x_i x_j \tanh(\theta_{ij}) \right] = 2^p \prod_{(i,j) \in E} \cosh(\theta_{ij}). \quad (3.52)
\]

Notice that \( \sum_x \prod_{(i,j) \in E} \left[ 1 + x_i x_j \tanh(\theta_{ij}) \right] = 2^p \) under the tree-structure assumption. Then

\[
P(X = x) = \frac{\prod_{(i,j) \in E} \exp(\theta_{ij}x_i x_j)}{Z(\theta)} = \frac{\prod_{(i,j) \in E} \cosh(\theta_{ij}) \left[ 1 + x_i x_j \tanh(\theta_{ij}) \right]}{2^p \prod_{(i,j) \in E} \cosh(\theta_{ij})}
\]

(3.53)

\[
= \frac{1}{2} \prod_{(i,j) \in E} \frac{1 + x_i x_j \tanh(\theta_{ij})}{2},
\]

(3.54)

(3.51) and (3.52) give (3.53) and \( |E| = p - 1 \) gives (3.54). Finally

\[
\mathbb{E}\left[ X_i X_j \right] = \frac{\partial \ln Z(\theta)}{\partial \theta_{ij}} = \frac{\partial \ln \left[ 2^p \prod_{(i,j) \in E} \cosh(\theta_{ij}) \right]}{\partial \theta_{ij}} = \tanh(\theta_{ij}), \quad \forall (i,j) \in E,
\]

(3.55)

and the latter gives the second part of the Lemma.

\[ \square \]

Lemma 9. Let \( \mathcal{V}' \) be a set of nodes such that \( \mathcal{V}' \subset \mathcal{V} \) and \( |\mathcal{V}'| = 2N \). Then it exists a set \( \mathcal{C}_T(\mathcal{V}') \) of \( |\mathcal{V}'|/2 \) pairs of nodes of \( \mathcal{V}' \), such that any two distinct pairs \( (w, w'), (v, v') \)
Figure 3.8: Proof of the existence of $C_T(V')$, Lemma 9

in $C_T(V')$ are pairwise disjoint (their paths have no commons edge), that is,

$$path_T(w, w') \cap path_T(v, v') = \emptyset, \quad \forall (w, w'), (v, v') \in C_T(V') : (w, w') \neq (v, v'). \quad (3.56)$$

**Proof.** We prove the existence of $C_T(V')$ by contradiction. Assume that the two distinct paths $path_T(w, u'), path_T(u, w')$ share at least one edge. Let their common sub-path be $path_T(z, z')$, Figure 3.8 and note that $z$ and $z'$ do not necessarily differ from $w, w', u, u'$. Notice that the common sub-path is unique (acyclic graph). Then we can always consider the permutation of the endpoints which gives the edge-disjoint paths $path_T(w, u)$ and $path_T(w', u')$. Now the paths $path_T(w, u)$ and $path_T(w', u')$ are disjoint, however it is possible that one of them or both, contain sub-paths with common edges. Then, we similarly proceed by removing the common sub-paths as previously. The set of common edges strictly decreases through the process, which terminates when there are only paths with no common edge.

**Theorem 10** (Theorem 8). Assume $X \sim p(x) \in \mathcal{P}_T(\alpha, \beta), \{i_1, i_2, \ldots, i_k\} \subset V$, then

$$E \left[ X_{i_1}X_{i_2} \ldots X_{i_k} \right] = \begin{cases} \prod_{e \in CP_T(i_1, i_2, \ldots, i_k)} \mu_e, & \forall k \in 2\mathbb{N} \\ 0, & \forall k \in 2\mathbb{N} + 1 \end{cases} \quad (3.57)$$

Recall that the set of edges $CP_T(i_1, \ldots, i_k)$ is a collection of $k/2$ edge-disjoint paths with endpoints pairs of the nodes $i_1, \ldots, i_k$ for each path. Given a tree structure $T$, $CP_T(i_1, \ldots, i_k)$ is found by running Algorithm 2 on $T$.

**Proof.** Even $k$. We proceed by showing that the Algorithm 2 returns the unique set
When \( k = 2 \) the expression is proved in Lemma 7. For \( k > 2 \) we proceed by using Lemmas 7 and 9. For all \( i, j \in \mathcal{V} \) there exists a unique path \( \{i, k_1, k_2, \ldots, k_\ell, j\} \) from \( i \) to \( j \). Define as previously the variable \( 1_{(i,j)} \triangleq (X_{k_1}X_{k_2}) \ldots (X_{k_\ell}X_{k_\ell}) \), which is equal to 1 almost surely, and define the set of nodes \( \mathcal{V}' \triangleq \{i_1, i_2, \ldots, i_k\} \). Without loss of generality we assume that the variables in the product \( X_{i_1}X_{i_2} \ldots X_{i_k} \) are ordered such such that the pairs \( X_{i_j}, X_{i_{j+1}} \) for all \( j \in \{1, 3, 5, \ldots, k-1\} \triangleq [k-1]^{\text{odd}} \) form edge-disjoint paths (Lemma 9), in other words

\[
\text{path}(i_j, i_{j+1}) \cap \text{path}(i_{j'}, i_{j'+1}) = \emptyset, \forall j \neq j' \in [k-1]^{\text{odd}}. \tag{3.58}
\]

Then, we have

\[
\mathbb{E} [X_{i_1}X_{i_2} \ldots X_{i_k}] = \mathbb{E} \left[ X_{i_1} 1_{(i_1,i_2)} X_{i_2} 1_{(i_3,i_4)} X_{i_4} \ldots X_{i_{k-1}} 1_{(i_{k-1},i_k)} X_{i_k} \right] \tag{3.59}
\]

\[
= \prod_{j \in [k-1]^{\text{odd}}} \mathbb{E} [X_j 1_{(i_j,i_{j+1})} X_{i_{j+1}}] \tag{3.60}
\]

\[
= \prod_{j \in [k-1]^{\text{odd}}} \prod_{e \in \text{path}(i_j, i_{j+1})} \mu_e \tag{3.61}
\]

\[
= \prod_{e \in \mathcal{CP}_T(i_1, i_2, \ldots, i_k)} \mu_e, \tag{3.62}
\]

where (3.59) and (3.60) come from (3.43), and (3.61) holds because of (3.58).

**Odd \( k \):** Lemma 6 gives \( p(x) = 2^{-p} \prod_{(i,j) \in \mathcal{E}} (1 + x_i x_j \mathbb{E} [X_i X_j]) \). Then

\[
\mathbb{E} [X_{i_1}X_{i_2} \ldots X_{i_k}] = \frac{1}{2} \sum_{x \in \{-1, +1\}^k} x_{i_1} x_{i_2} \ldots x_{i_k} \prod_{(i,j) \in \mathcal{E}} \frac{1 + x_i x_j \mathbb{E} [X_i X_j]}{2} = 0, \tag{3.63}
\]

gives the second part of (3.57).

**Lemma 10.** The mutual information of \( X_i, X_j \in \{-1, +1\} \) is symmetric function of the correlation \( \mathbb{E} [X_i X_j] \) and increasing with respect to \( \mathbb{E} [X_i X_j] \),

\[
I(X_i, X_j) = \frac{1}{2} \log_2 \left( (1 - \mathbb{E} [X_i X_j])^{1-\mathbb{E} [X_i X_j]} (1 + \mathbb{E} [X_i X_j])^{1+\mathbb{E} [X_i X_j]} \right). \tag{3.64}
\]
The proof can be derived through the definition of $I(X_i, X_j)$ and the expression (3.42), under the assumption of uniform marginal distributions.

### 3.5.2 Bounding the probability of mis-estimating correlations

The following lemma bounds the probability that the estimated pairwise correlations in the graph deviate from their true values. This follows from standard concentration of measure arguments.

**Lemma 11.** Fix $\delta > 0$. Then for any $\epsilon > 0$, if

$$n_\epsilon \geq 2 \log \left( \frac{p^2}{\delta} \right) / \epsilon^2,$$

(3.65)

then the event $E_{\epsilon_1}^{\text{corr}}$ defined in (3.24) holds with high probability:

$$\mathbb{P} \left( E_{\epsilon_1}^{\text{corr}} \right) \geq 1 - \delta = 1 - p^2 \exp \left( -\frac{n_\epsilon \epsilon^2}{2} \right).$$

(3.66)

**Proof.** Let $Z^{(i)}$ be the $i$th sample of $Z_{\epsilon_1} = Y_w Y_{\bar{w}} = N_w X_w N_{\bar{w}} X_{\bar{w}}$. Then $\hat{\mu}_{i,j}^{\epsilon_1} = \frac{1}{n_\epsilon} \sum_{i=1}^{n_\epsilon} Z^{(i)} = \frac{1}{n_\epsilon} \sum_{i=1}^{n_\epsilon} N_w^{(i)} X_w^{(i)} N_{\bar{w}}^{(i)} X_{\bar{w}}^{(i)}$ for all $i \neq j \in V$. Then Hoeffding’s inequality and union bound over all pairs of nodes $(p^2 / 2) < p^2 / 2$ give (3.66). \( \square \)

### 3.5.3 Recovering strong edges

In Lemma 4, we define the set of strong edges for the hidden model and show that the event $E_{\epsilon_1}^{\text{strong}}$ defined in (3.25) occurs with high probability. That is, only the strong edges are guaranteed to exist in the estimated structure $T_{\epsilon_1}^{\text{CL}}$. We also find a lower bound for the necessary number of samples for exact structure recovery. In fact we have $n_\epsilon \geq n$, as expected. Our bounds coincide with the noiseless case [29] by setting the noise level $q = 0$.

### 3.5.4 Analysis of the event $E_{\epsilon_1}^{\text{cascade}}$ ($\gamma_1$)

**Lemma 12.** Consider a path of length $d \geq 2$ in the original tree $T$, and without loss of generality assume that path is $X_1 - X_2 - \cdots - X_{d+1}$. Recall that $Y_m^{(i)}$ is the $i$th sample
of $Y_m$ and $m \in [d+1]$ and $\hat{\mu}_k^\dagger \triangleq \frac{1}{n} \sum_{i=1}^{n} Y_k^{(i)} Y_{k+1}^{(i)}$, $k \in [d]$. Then

$$
P \left( Y_k^{(\ell)} Y_{k+1}^{(\ell)} = \pm 1 \bigg| \hat{\mu}^\dagger_{k-1}, \ldots, \hat{\mu}^\dagger_1 \right) = \frac{1 \pm (1 - 2q)^2 \mu_k}{2} \frac{1 - \mu_k^\dagger \hat{\mu}^\dagger_{k-1}}{1 - (\mu_k^\dagger)^2} + \mu_k \frac{1 \pm \mu_k \hat{\mu}^\dagger_{k-1} - \mu_k^\dagger}{2} \frac{1 - (\mu_k^\dagger)^2}{(\mu_k^\dagger)^2}.
$$

(3.67)

**Proof.** Note that

$$
\hat{\mu}^\dagger_k = \frac{1}{n} \sum_{i=1}^{n} Y_k^{(i)} Y_{k+1}^{(i)} = \frac{1}{n} \sum_{i=1}^{n} (X_k N_k X_{k+1} N_{k+1})^{(i)},
$$

(3.68)

where each term

$$
(X_k N_k X_{k+1} N_{k+1})^{(i)} \perp \hat{\mu}^\dagger_r \quad \forall r \in [1, 2, \ldots, k-2], \forall \ell \in [1, 2, \ldots, n]
$$

(3.69)

thus

$$
P \left( (X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1 \bigg| \hat{\mu}^\dagger_{k-1}, \ldots, \hat{\mu}^\dagger_1 \right) = P \left( (X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1 \bigg| \hat{\mu}^\dagger_{k-1} \right) = P \left( (X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1 \bigg| \hat{\mu}^\dagger_{k-1} = \frac{1}{n} \sum_{i=1}^{n} (X_{k-1} N_{k-1} X_k N_k)^{(i)} \right)
$$

$$
= P \left( \hat{\mu}^\dagger_{k-1} = \frac{1}{n} \sum_{i=1}^{n} (X_{k-1} N_{k-1} X_k N_k)^{(i)} \bigg| (X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1 \right)
$$

$$
\times P \left( (X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1 \right).
$$

(3.70)

First we compute the probability $P \left( \hat{\mu}^\dagger_{k-1} = \frac{1}{n} \sum_{i=1}^{n} (X_{k-1} N_{k-1} X_k N_k)^{(i)} \right)$. Define the Bernoulli random variable $Z_{k-1}$ as

$$
Z_{k-1} \triangleq \frac{X_{k-1} N_{k-1} X_k N_k + 1}{2} = \begin{cases} 
0, & \text{w.p. } 1 - (1-2q)^2 \mu_{k-1}/2 \\
1, & \text{w.p. } 1 + (1-2q)^2 \mu_{k-1}/2
\end{cases}
$$

(3.71)
Then

\[
\mathbb{P}\left( \hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1}^{n} (X_{k-1}N_{k-1}X_kN_k)^{(i)} \right) \\
= \mathbb{P}\left( \hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1}^{n} (2Z_{k-1} - 1)^{(i)} \right) \\
= \mathbb{P}\left( \sum_{i=1}^{n} Z_{k-1}^{(i)} = \frac{n}{2} + 1 \right) \\
= \left( \frac{n}{2} \right) \left( 1 - (1 - 2q)^{2\mu_{k-1}} \right) \left( 1 + (1 - 2q)^{2\mu_{k-1}} \right)^{n - \frac{n\mu_{k-1}^{1+1}}{2}}. (3.72)
\]

As a second step we compute the probability

\[
\mathbb{P}\left( \hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1}^{n} (X_{k-1}N_{k-1}X_kN_k)^{(i)} \bigg| (X_kN_kX_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right) (3.73) \\
= \mathbb{P}\left( \hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1, i \neq \ell}^{n} (X_{k-1}N_{k-1}X_kN_k)^{(i)} \right) \\
+ \frac{(X_{k-1}N_{k-1}X_kN_k)^{(\ell)}}{n} \left| (X_kN_kX_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right).
\]

Note that

\[(X_{k-1}N_{k-1}X_kN_k)^{(i)} \perp (X_kN_kX_{k+1}N_{k+1})^{(\ell)} , \quad \forall i \neq \ell, (3.74)\]

and we would like to find the conditional distribution of \((X_{k-1}N_{k-1}X_kN_k)^{(\ell)}\) under the
event \( \{(X_{k-1}N_{k-1}X_{k-1})^{(\ell)} = \pm 1\} \). We have

\[
\mathbb{P}\left( (X_{k-1}N_{k-1}X_{k-1})^{(\ell)} = c \left| (X_{k}N_{k}X_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right. \right) = \frac{\mathbb{P}\left( (X_{k}N_{k}X_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right)}{\frac{1 \pm c\mathbb{E}[X_{k-1}N_{k-1}X_{k-1}N_{k-1}]+c(1-2q)^{2}\mu_{k-1} \pm (1-2q)^{2}\mu_{k}}{4}}
\]

\[
= \frac{1 \pm c(1-2q)^{2}\mu_{k-1} + c(1-2q)^{2}\mu_{k-1} \pm (1-2q)^{2}\mu_{k}}{2 \left( 1 \pm (1-2q)^{2}\mu_{k} \right)} \quad c \in \{-1, +1\}.
\]

(3.75)

Define

\[
P_1 \triangleq \mathbb{P}\left( (X_{k-1}N_{k-1}X_{k})^{(\ell)} = +1 \left| (X_{k}N_{k}X_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right. \right) \quad (3.76)
\]

\[
P_2 \triangleq \mathbb{P}\left( (X_{k-1}N_{k-1}X_{k})^{(\ell)} = -1 \left| (X_{k}N_{k}X_{k+1}N_{k+1})^{(\ell)} = \pm 1 \right. \right), \quad (3.77)
\]
then

\[
\begin{align*}
\mathbb{P}(\hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1, i \neq \ell}^n (X_{k-1} N_{k-1} X_k N_k)^{(i)}) &+ \frac{(X_{k-1} N_{k-1} X_k N_k)^{(\ell)}}{n} \bigg| (X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1 \\
= \mathbb{P}(\hat{\mu}_{k-1}^\dagger = \frac{1}{n} \sum_{i=1, i \neq \ell}^n (2Z_{k-1} - 1)^{(i)} + \frac{(2Z_{k-1} - 1)^{(\ell)}}{n} \bigg| (X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1) \\
= \mathbb{P}\left(\sum_{i=1, i \neq \ell}^n (Z_{k-1})^{(i)} + (Z_{k-1})^{(\ell)} = n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2} \bigg| (X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1\right) \\
= \mathbb{P}\left((Z_{k-1})^{(\ell)} = 0 \bigg| (X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1\right) \mathbb{P}\left(\sum_{i=1, i \neq \ell}^n (Z_{k-1})^{(i)} = n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2}\right) \\
+ \mathbb{P}\left((Z_{k-1})^{(\ell)} = 1 \bigg| (X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1\right) \mathbb{P}\left(\sum_{i=1, i \neq \ell}^n (Z_{k-1})^{(i)} = n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2} - 1\right) \\
= P_2\left(n \frac{1 - (1 - 2q)^2 \mu_{k-1}}{2} \right)^{n-1-n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2}} \left(1 + (1 - 2q)^2 \mu_{k-1}\right)^{n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2}} \\
+ P_1\left(n \frac{1 - (1 - 2q)^2 \mu_{k-1}}{2} - 1\right) \left(1 + (1 - 2q)^2 \mu_{k-1}\right)^{n-1-n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2} + 1} \\
\left(1 + (1 - 2q)^2 \mu_{k-1}\right)^{n \frac{\hat{\mu}_{k-1}^\dagger + 1}{2} - 1}. 
\end{align*}
\]

(3.78)

Finally

\[
\mathbb{P}\left((X_k N_k N_{k+1} N_{k+1})^{(\ell)} = \pm 1\right) = \frac{1 \pm (1 - 2q)^2 \mu_k}{2},
\]

(3.79)
\[
\mathbb{P}\left(Y_k^{(\ell)} Y_{k+1}^{(\ell)} = \pm 1 \bigg| \mu_{k-1}^+, \ldots, \mu_{k}^+\right) \\
\mathbb{P}\left((X_k N_k X_{k+1} N_{k+1})^{(\ell)} = \pm 1\right) \\
= P_2 \left(\frac{n - 1}{\frac{n}{2} \mu_{k-1}^+ + 1}\right) \left(\frac{n}{2 \mu_{k-1}^+ + 1}\right)^{-1} \left(\frac{1 - \mu_{k-1}^+}{2}\right) \left(\frac{n-1 \mu_{k-1}^+ + 1}{2}\right) \left(\frac{1 + \mu_{k-1}^+}{2}\right) \\
+ P_1 \left(\frac{n - 1}{\frac{n}{2} \mu_{k-1}^+ + 1} - 1\right) \left(\frac{n}{2 \mu_{k-1}^+ + 1}\right)^{-1} \left(\frac{1 - \mu_{k-1}^+}{2}\right) \left(\frac{n-1 \mu_{k-1}^+ + 1}{2}\right) \left(\frac{1 + \mu_{k-1}^+}{2}\right) \\
= P_2 \frac{n - 1}{\frac{n}{2} \mu_{k-1}^+ + 1} \left(\frac{n}{2 \mu_{k-1}^+ + 1}\right)^{-1} \left(\frac{1 - \mu_{k-1}^+}{2}\right) \\
+ P_1 \left(\frac{n - 1}{\frac{n}{2} \mu_{k-1}^+ + 1} - 1\right) \left(\frac{n}{2 \mu_{k-1}^+ + 1}\right)^{-1} \left(\frac{1 + \mu_{k-1}^+}{2}\right) \\
= P_2 \frac{1}{1 - \mu_{k-1}^+} + P_1 \frac{1 + \mu_{k-1}^+}{1 + \mu_{k-1}^+}. \tag{3.80}
\]
The latter and the definition of $P_1, P_2$ (see (3.76), (3.77)) give

$$
P \left( Y_k^{(t)} Y_{k+1}^{(t)} \pm 1 \bigg| \hat{\mu}_{k-1}^+, \ldots, \hat{\mu}_1^+ \right)
= \left[ P_2 \frac{1 - \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} + P_1 \frac{1 + \hat{\mu}_{k-1}^+}{1 + \hat{\mu}_{k-1}^+} \right] P \left( (X_k N_k X_{k+1} N_{k+1})^{(t)} = \pm 1 \right)
= \left[ P_2 \frac{1 - \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} + P_1 \frac{1 + \hat{\mu}_{k-1}^+}{1 + \hat{\mu}_{k-1}^+} \right] \frac{1 \pm (1 - 2q)^2 \mu_k}{2} \frac{1 \pm (1 - 2q)^2 \mu_k}{2}
= \frac{1 \mp (1 - 2q)^2 \mu_{k-1} \mu_k - (1 - 2q)^2 \mu_{k-1} \pm (1 - 2q)^2 \mu_k}{4} \frac{1 \pm \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} \frac{1 \pm \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+}
+ \frac{1 \pm (1 - 2q)^2 \mu_{k-1} \mu_k + (1 - 2q)^2 \mu_{k-1} \pm (1 - 2q)^2 \mu_k}{4} \frac{1 \pm \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} \frac{1 \pm \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+}
= \frac{1 \pm (1 - 2q)^2 \mu_k}{4} \left( \frac{1 - \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} + \frac{1 + \hat{\mu}_{k-1}^+}{1 + \hat{\mu}_{k-1}^+} \right)
+ \frac{1 \pm (1 - 2q)^2 \mu_{k-1} \mu_k + (1 - 2q)^2 \mu_{k-1}}{4} \left( \frac{1 + \hat{\mu}_{k-1}^+}{1 + \hat{\mu}_{k-1}^+} - \frac{1 - \hat{\mu}_{k-1}^+}{1 - \hat{\mu}_{k-1}^+} \right)
= \frac{1 \pm (1 - 2q)^2 \mu_k}{2} \frac{1 - \hat{\mu}_{k-1}^+ \hat{\mu}_{k-1}^+}{1 - (\hat{\mu}_{k-1}^+)^2} + \frac{1 \pm (1 - 2q)^2 \mu_{k-1} \mu_k - \hat{\mu}_{k-1}^+}{2} \frac{1 \pm \mu_k \hat{\mu}_{k-1}^+ - \hat{\mu}_{k-1}^+}{1 - (\hat{\mu}_{k-1}^+)^2}.
\tag{3.81}
$$

Note that $(1 - 2q)^2 \mu_{k-1} = \hat{\mu}_{k-1}^+$, and the proof is completed.

Lemma 13. Define the function $K(\beta, q)$ as

$$
K(\beta, q) \triangleq \frac{10(1 - \tanh^2(\beta))}{9 + (1 - 2q)^2 - \tanh^2(\beta)(1 - 2q)^2(9(1 - 2q)^2 + 1)}
\tag{3.82}
$$

and the event $E_{e, t}^{\text{edge}}$ as

$$
E_{e, t}^{\text{edge}} \triangleq \left\{ \left| \hat{\mu}_e^+ - \mu_e^+ \right| \leq \gamma_e \right\}, \quad e \in E, \quad \gamma_e > 0,
\tag{3.83}
$$
and $E_{i}^{\text{edge}}(\mathcal{E}_T) \triangleq \cap_{e \in \mathcal{E}_T} E_{e,i}^{\text{edge}}$. If

$$n \geq \frac{108e^{2\beta} \log(2p/\delta)}{(1 - 2q)^4 K(\beta, q)}$$

and

$$\gamma_e = \sqrt{\frac{3}{nK(\beta, q)}} \frac{1 - \mu_e^2}{\log(2p/\delta)}$$

(3.84)

then $P\left[\left(\bigcap_{e \in \mathcal{E}_T} E_{e,i}^{\text{edge}}(\mathcal{E}_T)\right)^c\right] \leq \delta$.

Proof. The variance of $\hat{\mu}_e^\dagger$ is $(1 - (\mu_e^\dagger)^2)/n$ and by applying Bernstein’s inequality

$$P\left[\left(\bigcap_{e \in \mathcal{E}_T} E_{e,i}^{\text{edge}}(\mathcal{E}_T)\right)^c\right] \leq 2 \exp \left(-\frac{n\gamma_e^2}{2(1 - (\mu_e^\dagger)^2) + \frac{4}{3} \gamma_e}\right), \quad \forall \gamma_e > 0. \quad (3.85)$$

We choose $\gamma_e = \sqrt{\frac{3}{nK(\beta, q)}} \frac{1 - \mu_e^2}{\log(2p/\delta)}$ (because the parameter $\gamma_e$ is free, that is, Bernstein’s inequality holds for all $\gamma_e > 0$). If $n$ satisfies (3.84) then

$$\gamma_e \leq \sqrt{\frac{3}{108e^{2\beta}}} \frac{1 - \mu_e^2}{(1 - 2q)^4(1 - 2q)^2} \leq \frac{1 - 2q)^2}{6(1 - \mu_e^2)}, \quad (3.86)$$

and the last is true because $e^{-2\beta} \leq 1 - \tanh(\beta) \leq 1 - |\mu_e| \leq 1 - \mu_e^2$. By applying (3.84) and (3.86) on (3.85) we get

$$P\left[\left(\bigcap_{e \in \mathcal{E}_T} E_{e,i}^{\text{edge}}(\mathcal{E}_T)\right)^c\right] \leq 2 \exp \left(-\frac{n\gamma_e^2}{2(1 - (\mu_e^\dagger)^2) + \frac{4}{3} \gamma_e}\right) \leq 2 \exp \left(-\frac{3}{K(\beta, q)} \frac{1 - \mu_e^2}{2(1 - (1 - 2q)^4(1 - 2q)^2)(1 - \mu_e^2)} \log(2p/\delta)\right) \leq 2 \exp \left(-\frac{10}{9K(\beta, q)} \frac{1 - \mu_e^2}{2(1 - (1 - 2q)^2 - \mu_e^2(1 - 2q)^2(1 - 2q)(1 - 2q)^2 + \frac{3}{5}) \log(2p/\delta)}\right). \quad (3.87)$$
The following function

\[
f(x) = \frac{10}{9} \frac{1 - x}{1 + \frac{1}{9}(1 - 2q)^2 - x(1 - 2q)^2((1 - 2q)^2 + \frac{1}{9})}, \quad x \in [\tanh^2(\alpha), \tanh^2(\beta)]
\]

(3.88)

is strictly decreasing, thus we have \(f(\tanh^2(\beta)) \leq f(x)\) for all \(x \in [\tanh^2(\alpha), \tanh^2(\beta)]\). Also \(K(\beta, q) \equiv f(\tanh^2(\beta))\), the latter together with (3.87) give

\[
\begin{align*}
\mathbb{P}\left[ (E^\text{edge}_{e, \dagger})^c \right] & \leq 2 \exp \left( -\frac{1}{K(\beta, q)} f(\mu^2_e) \log(2p/\delta) \right) \\
& \leq 2 \exp \left( -\frac{1}{K(\beta, q)} f(\tanh^2(\beta)) \log(2p/\delta) \right) \\
& = 2 \exp \left( -\frac{1}{K(\beta, q)} K(\beta, q) \log(2p/\delta) \right) \\
& = \frac{\delta}{p}. 
\end{align*}
\]

(3.89)

Finally, by applying union over the \(p - 1\) edges of the tree we get \(\mathbb{P}\left[ (E^\text{edge}_{\dagger}(\mathcal{E}_T))^c \right] \leq \delta\).

The next Lemma is the extension of Lemma 8.7 by [29]. The sample complexity bound exactly recovers the noiseless case and its expression is continuous at \(q = 0\). Further, the bound is independent of the length of the longest path \(d\), similarly to the noiseless setting. Finally, we provide upper bounds on the functions that appear in the bound. The latter give a more tractable version of the result and a clear representation of the required number of samples as a function of the parameters.

**Lemma 14** (Concentration bound for the event \(E^\text{cascade}_{\dagger}(\gamma_\dagger)\)). For \(\beta > 0\) and \(q \in [0, 1/2)\)
we define the functions $S(\cdot), G(\cdot), K(\cdot), A(\cdot), \Delta(\cdot)$

\[
S(\beta, q) \triangleq 2 + \frac{(1 - 2q)^2}{6}(1 - (1 - 2q)^2) \tan^2(\beta) \leq 3 - (1 - 2q)^2 \triangleq S \quad (3.90)
\]

\[
A(\beta, q) \equiv A \triangleq (1 - 2q)^2[1 - \tan(\beta)(1 - (1 - 2q)^2)]
\]

\[
G(\beta, q) \triangleq \frac{3}{4(1 - 2q)^2} \left[d(1 - A) \left(\frac{A + 2}{3}\right)^d + 1\right] \leq \frac{3(3e^{-1}1_{q \neq 0} + 1)}{4(1 - 2q)^2} \triangleq G, \quad (3.92)
\]

and the inequality in (3.92) holds because the function $G(\beta, q)$ is bounded for all $d \in \mathcal{N} \setminus \{1\}$.

\[
K(\beta, q) \triangleq \frac{10(1 - \tanh^2(\beta))}{9 + (1 - 2q)^2 - \tanh^2(\beta)(1 - 2q)^2(9(1 - 2q)^2 + 1)} \geq e^{-2\beta_{q=0}} \triangleq K \quad (3.93)
\]

\[
\Delta \triangleq \frac{1 - (1 - 2q)^2}{1 - (1 - 2q)^4\tanh^2(\beta)} \sqrt{n \frac{3\log(2p^3/\delta)}{\log(2p^3/\delta)}} \tanh^2(\beta)e^{2\beta}. \quad (3.94)
\]

If $\Delta < \gamma_{\hat{t}} \leq S(\beta, q)G(\beta, q)/3 + \Delta$ and

\[
n \geq \max \left\{ \frac{S^2(\beta, q)G^2(\beta, q)}{0.3^2(\gamma_{\hat{t}} - \Delta)^2}, \frac{108e^{2\beta}}{(1 - 2q)^4K(\beta, q)} \right\} \quad (3.95)
\]

then for any path $\mathcal{A}_d = \{e_1, e_2, \ldots, e_d\}$ of $T$ with $d$ edges, it is true that

\[
\mathbb{P} \left( \left| \prod_{e \in \mathcal{A}_d} \hat{\mu}_e^\dagger (1 - 2q)^2 - \prod_{e \in \mathcal{A}_d} \hat{\mu}_e (1 - 2q)^2 \right| \geq \gamma_{\hat{t}} \right) \leq \frac{2\delta}{p^2}, \quad d > 2. \quad (3.96)
\]

**Proof.** For sake of space we proceed by using the notation $\hat{\mu}_k^\dagger$ and $\hat{\mu}_k^\dagger$ instead of $\mu_k^\dagger$ and $\hat{\mu}_k^\dagger$ for $k \in [d]$. Define the random variable

\[
M_i^\dagger \triangleq \left( \frac{\hat{\mu}_i^\dagger}{(1 - 2q)^2} - \frac{\hat{\mu}_i}{(1 - 2q)^2} \right) \prod_{j=1}^{i-1} \frac{\hat{\mu}_j^\dagger}{(1 - 2q)^2} \prod_{j=i+1}^{d} \frac{\hat{\mu}_j}{(1 - 2q)^2}. \quad (3.97)
\]

Then $\sum_{i=1}^{d} M_i^\dagger = \prod_{i=1}^{d} \frac{\hat{\mu}_i^\dagger}{(1 - 2q)^2} - \prod_{i=1}^{d} \frac{\mu_i^\dagger}{(1 - 2q)^2}$, and define the sequence of paths with length $k$ as $\mathcal{A}_k \triangleq \{e_1, e_2, \ldots, e_k\} \subset \mathcal{A}_d$, for $2 \leq k \leq d$. Although we provided the definition of the event $E_\dagger^{edge}(\cdot)$ in Lemma 13, we restate it below for completeness. For
some \( \gamma_e > 0 \) the definition follows

\[
\mathsf{E}_\text{edge}^*(A_k) \triangleq \bigcap_{e \in A_k} \left\{ |\hat{\mu}_e^+ - \mu_e^+| \leq \gamma_e \right\}.
\] (3.98)

The law of total probability gives

\[
P \left[ \left| \sum_{i=1}^d M_i^+ \right| \geq \gamma \right] \leq P \left[ \left| \sum_{i=1}^d M_i^+ \right| > \gamma \big| \mathsf{E}_\text{edge}^*(A_{d-1}) \right] + P \left[ \left( \mathsf{E}_\text{edge}^*(A_{d-1}) \right)^c \right].
\] (3.99)

For second term, Lemma 13 gives that

\[
P \left[ \left( \mathsf{E}_\text{edge}^*(A_{d-1}) \right)^c \right] \leq \delta/p^2 \text{ if } n \geq 108 \epsilon^{2\beta} \log(2p^3/\delta) \left(1 - 2q\right)^4 K(\beta, q),
\] (3.100)

we define the function \( K(\beta, q) \) in Lemma 13 (3.82). Here we will find an upper bound for the first term of the right hand-side of (3.99). Note that \( M_k^+ \) is written as

\[
M_k^+ = \left( \frac{1}{n} \sum_{\ell=1}^n (Y_kY_{k+1})^{(\ell)} - \frac{\mu_k^+}{(1 - 2q)^2} \right) \prod_{j=1}^{k-1} \frac{\hat{\mu}_j^+}{(1 - 2q)^2} \prod_{j=k+1}^d \frac{\mu_j^+}{(1 - 2q)^2}
\]

\[
= \frac{1}{n} \sum_{\ell=1}^n \left( \frac{(X_kN_kX_{k+1}N_{k+1})^{(\ell)}}{(1 - 2q)^2} - \frac{\mu_k^+}{(1 - 2q)^2} \right) \prod_{j=1}^{k-1} \frac{\hat{\mu}_j^+}{(1 - 2q)^2} \prod_{j=k+1}^d \frac{\mu_j^+}{(1 - 2q)^2}.
\] (3.101)

and we define

\[
Z_k^{(\ell)} \triangleq \left( \frac{(X_kN_kX_{k+1}N_{k+1})^{(\ell)}}{(1 - 2q)^2} - \frac{\mu_k^+}{(1 - 2q)^2} \right) \prod_{j=1}^{k-1} \frac{\hat{\mu}_j^+}{(1 - 2q)^2} \prod_{j=k+1}^d \frac{\mu_j^+}{(1 - 2q)^2}.
\] (3.102)

The random variables \( Z_k^{(\ell)} \) for \( \ell \in [n] \) and fixed \( k \in [d] \) are independent conditioned on the event \( \mathsf{E}_\text{edge}^*(A_{k-1}) \). However the conditional expectation \( \mathbb{E}[Z_k^{(i)} | Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{k-1}^+, \ldots, \hat{\mu}_1^+] \) is not zero. To apply a concentration of measure result on \( Z_k^{(\ell)} \) we use the extended Bennet's inequality for supermartingales [68].

**Martingale Differences:** Define \( \xi_k^{(0)} \triangleq 0, \xi_k^{(1)} \triangleq Z_k^{(1)} - \mathbb{E}[Z_k^{(1)} | \hat{\mu}_{k-1}^+, \ldots, \hat{\mu}_1^+] \).
\(\xi_k^{(i)} \triangleq Z_k - \mathbb{E} \left[ Z_k^{(i)} | Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{k-1}^\dagger, \ldots, \hat{\mu}_1^\dagger \right].\) Also, define as \(\mathcal{F}_{i-1}^k\) the \(\sigma\)-algebra generated by \(Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{k-1}^\dagger, \ldots, \hat{\mu}_1^\dagger\), then \((\xi_k^{(i)}, \mathcal{F}_{i-1}^k)_{i=1}^n\) is a Martingale Difference Sequence (MDS).

Additionally, conditioned on \(Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{k-1}^\dagger, \ldots, \hat{\mu}_1^\dagger\) we have

\[
Z_k^{(i)} = \begin{cases} 
\frac{1}{(1-2q)^{2d}} \left( 1 - \mu_k^\dagger \right) \prod_{j=1}^{k-1} \hat{\mu}_j^\dagger \prod_{j=k+1}^d \mu_j^\dagger, & \text{w.p. } \mathbb{P} \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = +1 | \hat{\mu}_{k-1}^\dagger \right) \\
- \frac{1}{(1-2q)^{2d}} \left( 1 + \mu_k^\dagger \right) \prod_{j=1}^{k-1} \hat{\mu}_j^\dagger \prod_{j=k+1}^d \mu_j^\dagger, & \text{w.p. } \mathbb{P} \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = -1 | \hat{\mu}_{k-1}^\dagger \right),
\end{cases}
\]

and we have proved (Lemma 12) that

\[
\mathbb{P} \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = \pm 1 | \hat{\mu}_{k-1}^\dagger, \ldots, \hat{\mu}_1^\dagger \right) = \frac{1}{2} + \frac{\mu_k^\dagger}{1 - (\mu_{k-1}^\dagger)^2} + \mu_k^\dagger \frac{1 - \mu_k^\dagger}{2} - \frac{\mu_k^\dagger - \mu_{k-1}^\dagger}{2} \left( 1 - (\mu_{k-1}^\dagger)^2 \right).
\]

(3.104)

Thus we have

\[
\mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}_{i-1}^k \right] = \mathbb{E} \left[ Z_k^{(i)} | Z_k^{(i-1)}, \ldots, Z_k^{(1)}, \hat{\mu}_{k-1}^\dagger, \ldots, \hat{\mu}_1^\dagger \right] \\
= \left[ (1 - \mu_k^\dagger) \mathbb{P} \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = +1 | \hat{\mu}_{k-1}^\dagger \right) - (1 + \mu_k^\dagger) \mathbb{P} \left( Y_{k}^{(\ell)} Y_{k+1}^{(\ell)} = -1 | \hat{\mu}_{k-1}^\dagger \right) \right] \\
\times \prod_{j=1}^{k-1} \hat{\mu}_j^\dagger \prod_{j=k+1}^d \mu_j^\dagger \frac{(1-2q)^{2d}}{(1-2q)^{2d}}.
\]

(3.105)
Note that (3.104) gives

\[
\left[ (1 - \mu_k^\dagger) P \left( Y_k^{(f)} Y_k^{(l)} | Y_{k+1} \right) - (1 + \mu_k^\dagger) P \left( Y_k^{(f)} Y_k^{(l)} | Y_{k+1} = -1 \right) \right] \\
= \left( (1 - \mu_k^\dagger) \frac{1 + \mu_k^\dagger}{2} - (1 + \mu_k^\dagger) \frac{1 - \mu_k^\dagger}{2} \right) \mu_{k-1}^\dagger \frac{1}{1 - (\mu_{k-1})^2} \\
+ \left( (1 - \mu_k^\dagger) \frac{1 + \mu_k}{2} - (1 + \mu_k^\dagger) \frac{1 - \mu_k}{2} \right) \mu_{k-1}^\dagger \frac{\mu_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1})^2} \\
= \left( (1 - \mu_k^\dagger) \frac{1 + \mu_k}{2} - (1 + \mu_k^\dagger) \frac{1 - \mu_k}{2} \right) \mu_{k-1}^\dagger \frac{\mu_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1})^2} \\
= \frac{1}{2} \left( (1 - \mu_k^\dagger + \mu_k - \mu_k^\dagger \mu_k - 1 + \mu_k - \mu_k^\dagger + \mu_k \mu_k^\dagger) \mu_{k-1}^\dagger \frac{\mu_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1})^2} \right) \\
= (\mu_k - \mu_k^\dagger) \mu_{k-1}^\dagger \frac{\mu_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1})^2}.
\]

(3.106)

Combine the latter with (3.105) to get

\[
\mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}^k_{i-1} \right] = \mu_{k-1}(\mu_k - \mu_k^\dagger) \mu_{k-1}^\dagger \frac{\tilde{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger \prod_{j=1}^{k-1} \tilde{\mu}_j \prod_{j=k+1}^{d} \mu_j}{(1 - (\mu_{k-1})^2)} \}, \quad i \in [n].
\]

(3.107)

If \( q = 0 \) then \( \mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}^k_{i-1} \right] = 0 \). Also \( \lim_{n \to \infty} \mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}^k_{i-1} \right] \to 0 \) for all \( q \in [0, 1/2] \) because \( \lim_{n \to \infty} \tilde{\mu}_{k-1}^\dagger \to \mu_{k-1}^\dagger \). Note that

\[
\mathbb{E} \left[ \left( \xi_k^{(i)} \right)^2 | \mathcal{F}^k_{i-1} \right] = \mathbb{E} \left[ \left( Z_k^{(i)} - \mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}^k_{i-1} \right] \right)^2 | \mathcal{F}^k_{i-1} \right] \\
= \mathbb{E} \left[ \left( Z_k^{(i)} \right)^2 | \mathcal{F}^k_{i-1} \right] - \mathbb{E} \left[ Z_k^{(i)} | \mathcal{F}^k_{i-1} \right].
\]

(3.108)

We compute \( \mathbb{E} \left[ \left( Z_k^{(i)} \right)^2 | \mathcal{F}^k_{i-1} \right] : \)

\[
\mathbb{E} \left[ \left( Z_k^{(i)} \right)^2 | \mathcal{F}^k_{i-1} \right] = \mathbb{E} \left[ \left( Z_k^{(i)} \right)^2 | Z_k^{(i-1)}, \ldots, Z_k^{(l)}, \tilde{\mu}_{k-1}, \ldots, \tilde{\mu}_1 \right] \\
= \left( 1 - \mu_k^\dagger \right)^2 P \left( Y_k^{(f)} Y_k^{(l)} | Y_{k+1} = +1 | \tilde{\mu}_{k-1}^\dagger \right) + (1 + \mu_k^\dagger)^2 P \left( Y_k^{(f)} Y_k^{(l)} | Y_{k+1} = -1 | \tilde{\mu}_{k-1}^\dagger \right) \\
\times \left( \prod_{j=1}^{k-1} \tilde{\mu}_j \prod_{j=k+1}^{d} \mu_j^\dagger \right)^2 \frac{(1 - 2q)^2d}{(1 - 2q)^2d}.
\]

(3.109)
We use (3.104) to find

\[
\left[ (1 - \mu_k^\dagger)^2 \mathbb{P} \left( Y_k^{(t)} Y_{k+1}^{(t)} = +1 \right) \right. + \left. (1 + \mu_k^\dagger)^2 \mathbb{P} \left( Y_k^{(t)} Y_{k+1}^{(t)} = -1 \right) \right]
\]

\[
= \left( (1 - \mu_k^\dagger) \frac{2 + \mu_k^\dagger}{2} + (1 + \mu_k^\dagger) \frac{2 - \mu_k^\dagger}{2} \right) \frac{1 - \mu_k^\dagger \hat{\mu}_{k-1}^\dagger}{1 - (\mu_k^\dagger)^2}
\]

\[
+ \left( (1 - \mu_k^\dagger) \frac{2 + \mu_k}{2} + (1 + \mu_k^\dagger) \frac{2 - \mu_k}{2} \right) \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
= (1 - (\mu_k^\dagger)^2) \frac{1 - \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} + (1 + (\mu_k^\dagger)^2 - 2\mu_k \hat{\mu}_k^\dagger) \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
= (1 - (\mu_k^\dagger)^2) \frac{1 - \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} + (1 - (\mu_{k-1}^\dagger)^2) \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
+ 2\mu_k^\dagger \left( \mu_k^\dagger - \mu_k \right) \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
= (1 - (\mu_k^\dagger)^2) \left[ \frac{1 - \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} + \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger \right] + 2\mu_k^\dagger \left( \mu_k^\dagger - \mu_k \right) \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
= (1 - (\mu_k^\dagger)^2) + 2\mu_k^\dagger \left( \mu_k^\dagger - \mu_k \right) \mu_{k-1}^\dagger \hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger
\]

\[
\text{(3.110)}
\]

Now we combine (3.107), (3.108), (3.109) and (3.110) to get

\[
E \left[ \left( \zeta_k^{(i)} \right)^2 \mathcal{F}_{i-1}^k \right]
\]

\[
= E \left[ \left( Z_k^{(i)} \right)^2 \mathcal{F}_{i-1}^k \right] - E^2 \left[ Z_k^{(i)} \mathcal{F}_{i-1}^k \right]
\]

\[
= \left[ (1 - (\mu_k^\dagger)^2) + 2\mu_k^\dagger \left( \mu_k^\dagger - \mu_k \right) \mu_{k-1}^\dagger \frac{\hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} - \left( \mu_{k-1}^\dagger \left( \mu_k^\dagger - \mu_k \right) \frac{\hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} \right)^2 \right]
\]

\[
\times \left( \frac{\Pi_{j=1}^{k-1} \hat{\mu}_{j}^\dagger \Pi_{j=k+1}^{d} \mu_{j}^\dagger}{(1 - 2q)^{2d}} \right)^2
\]

\[
= \left[ 1 - \left( \mu_k^\dagger + \mu_{k-1}^\dagger \left( \mu_k^\dagger - \mu_k \right) \frac{\hat{\mu}_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1}^\dagger)^2} \right)^2 \right] \left( \frac{\Pi_{j=1}^{k-1} \hat{\mu}_{j}^\dagger \Pi_{j=k+1}^{d} \mu_{j}^\dagger}{(1 - 2q)^{2d}} \right)^2.
\text{(3.111)}
\]
For sake of space we define the function

\[ f_{\mu_{k-1}}(q) \equiv f(\mu_{k}, \mu_{k-1}, \hat{\mu}_{k-1}, q) \triangleq \frac{\mu_{k-1}(\mu_{k} - \mu_{k}^{\dagger}) \hat{\mu}_{k-1} - \mu_{k-1}^{\dagger}}{1 - (\mu_{k-1}^{\dagger})^2}, \]  

(3.112)

and then (3.107) and (3.111) can be written as

\[
\mathbb{E}\left[Z_{i}^{(i)} \bigg| \mathcal{F}_{i-1}^{k}\right] = f_{\mu_{k-1}}(q) \prod_{j=1}^{k-1} \hat{\mu}_{j}^{\dagger} \prod_{j=k+1}^{d} \mu_{j}^{\dagger}, \quad i \in [n], \tag{3.113}
\]

\[
\mathbb{E}\left[(\xi_{k}^{(i)})^{2} \bigg| \mathcal{F}_{i-1}^{k}\right] = \left[1 - \left(\mu_{k}^{\dagger} + f_{\mu_{k-1}}(q)\right)^{2}\right] \left(\frac{\prod_{j=1}^{k-1} \hat{\mu}_{j}^{\dagger} \prod_{j=k+1}^{d} \mu_{j}^{\dagger}}{(1 - 2q)^{2d}}\right)^{2}, \quad i \in [n]. \tag{3.114}
\]

We would like to find an upper bound on the summation \(\sum_{k=1}^{d} \mathbb{E}\left[(\xi_{k}^{(i)})^{2} \bigg| \mathcal{F}_{i-1}^{k}\right]\).

Define \(A \equiv A(\beta, q) \triangleq (1 - 2q)^{2}[1 - \tanh(\beta)(1 - (1 - 2q)^{2})]\), for all \(\beta > 0\) and \(q \in [0, 1]\).

Then

\[
\left[1 - \left(\mu_{k}^{\dagger} + f_{\mu_{k-1}}(q)\right)^{2}\right]
\]

\[
= \left[1 - \left(\mu_{k}^{\dagger} + \mu_{k-1}(\mu_{k} - \mu_{k}^{\dagger}) \frac{\hat{\mu}_{k-1} - \mu_{k-1}^{\dagger}}{1 - (\mu_{k-1}^{\dagger})^2}\right)^{2}\right]
\]

\[
= \left[1 - \mu_{k}^{2} \left((1 - 2q)^{2} + \mu_{k-1}(1 - (1 - 2q)^{2}) \frac{\hat{\mu}_{k-1} - \mu_{k-1}^{\dagger}}{1 - (\mu_{k-1}^{\dagger})^2}\right)^{2}\right]
\]

\[
\leq \left[1 - \mu_{k}^{2} \left((1 - 2q)^{2} - |\mu_{k-1}|(1 - (1 - 2q)^{2}) \frac{\hat{\mu}_{k-1} - \mu_{k-1}^{\dagger}}{1 - (\mu_{k-1}^{\dagger})^2}\right)^{2}\right]
\]

\[
\leq \left[1 - \mu_{k}^{2} \left((1 - 2q)^{2} - (1 - 2q)^{2} \tanh(\beta)(1 - (1 - 2q)^{2})\right)^{2}\right] \tag{3.115}
\]

\[
\leq [1 - \mu_{k}^{2} A(\beta, q)], \tag{3.116}
\]

and (3.115) holds because \(|\hat{\mu}_{k-1}^{\dagger} - \mu_{k-1}^{\dagger}| \leq \gamma_{j} \leq (1 - 2q)^{2}(1 - \mu_{j}^{2})/6 \leq (1 - 2q)^{2}(1 - (\mu_{j}^{\dagger})^{2})/6\).
Then (3.114) and (3.116) give

\[
\sum_{k=1}^{d} \mathbb{E} \left[ \left( \xi_k^{(i)} \right)^2 \right] \mathcal{F}_{i-1}^k
\]

\[
\leq \sum_{k=1}^{d} \left[ 1 - \mu_k^2 A(\beta, q) \right] \left( \frac{\prod_{j=1}^{k-1} \mu_j^{\dagger} \prod_{j=k+1}^{d} \mu_j^{\dagger}}{(1 - 2q)^{2d}} \right)^2
\]

\[
\leq \frac{1}{(1 - 2q)^2} \sum_{k=1}^{d} \left[ 1 - \mu_k^2 A(\beta, q) \right] \prod_{j=1, j \neq k}^{d} \left[ \mu_j^2 + 2 \frac{\gamma_j^{\dagger}}{(1 - 2q)^2} \right]. \tag{3.117}
\]

The inequality (3.117) holds under the event \( \mathcal{E}_T^{\text{edge}}(\mathcal{E}_T) \) defined in (3.83) (see Lemma 13) because

\[
\left( \frac{\hat{\mu}_j^{\dagger}}{(1 - 2q)^2} \right)^2 \leq \left( \frac{\mu_j^{\dagger}}{(1 - 2q)^2} \right)^2 + 2 \frac{\gamma_j^{\dagger}}{(1 - 2q)^2}, \tag{3.118}
\]

since \(|\mu_j^{\dagger}| \leq (1 - 2q)^2, |\hat{\mu}_j^{\dagger}| \leq (1 - 2q)^2\) under the assumption of known \(q\). Next we define \(x_j \equiv \mu_j^2 + 2 \frac{\gamma_j^{\dagger}}{(1 - 2q)^2}\), then

\[
3(1 - A(\beta, q)x_j)/2 + (A(\beta, q) - 1)/2 \geq 1 - \mu_k^2 A(\beta, q)
\]

and (3.117) gives

\[
\sum_{k=1}^{d} \mathbb{E} \left[ \left( \xi_k^{(i)} \right)^2 \right] \mathcal{F}_{i-1}^k
\]

\[
\leq \frac{1}{(1 - 2q)^2} \sum_{k=1}^{d} \left[ \frac{3}{2} (1 - Ax_j) + \frac{A - 1}{2} \right] \prod_{j=1, j \neq k}^{d} x_j
\]

\[
\leq \frac{d}{(1 - 2q)^2} \left[ \frac{3}{2} (1 - Ax) + \frac{A - 1}{2} \right] x^{d-1}. \tag{3.119}
\]
The latter is maximized at \( x^* = (A + 2) \left(1 - \frac{1}{q}\right)/3 \) and \( A \in (0, 1] \), thus we have

\[
\frac{d}{(1 - 2q)^2} \left[ \frac{3}{2} (1 - Ax) + \frac{A - 1}{2} \right] x^{d-1}
\leq \frac{d}{(1 - 2q)^2} \left[ \frac{3}{2} - \frac{A(A + 2)}{3} + \frac{A - 1}{2} \right] (x^*)^{d-1} + \frac{A(A + 2)}{2(1 - 2q)^2} (x^*)^{d-1}
\]

\[
= \frac{2 - A^2 - A}{2(1 - 2q)^2} \left( \frac{A + 2}{3} \right)^{d-1} \left( 1 - \frac{1}{d} \right)^{d-1} + \frac{A(A + 2)}{2(1 - 2q)^2} \left( \frac{A + 2}{3} \right) \left( 1 - \frac{1}{2} \right)
\]

\[
\leq \frac{d(A + 2)(1 - A)}{4(1 - 2q)^2} \left( \frac{A + 2}{3} \right)^{d-1} + \frac{A(A + 2)}{12(1 - 2q)^2}
\]

\[
\leq d \frac{3(1 - A)}{4(1 - 2q)^2} \left( \frac{A + 2}{3} \right)^{d} + \frac{3}{4(1 - 2q)^2} \triangleq G(\beta, q)
\]

(3.120)

In (3.120) we define the function \( G(\beta, q) \) and we proved that it has an upper bound independent of \( d \in [p - 1] \),

\[
G(\beta, q) \triangleq d \frac{3(1 - A(\beta, q))}{4(1 - 2q)^2} \left( \frac{A(\beta, q) + 2}{3} \right)^{d} + \frac{3}{4(1 - 2q)^2} \leq \frac{3(3e^{-1} + 1)}{4(1 - 2q)^2} \leq \frac{3}{4(1 - 2q)^2}
\]

(3.121)

For the rest of the proof and the final result \( G(\beta, q) \) can be replaced by its upper bound in (3.121), however the definition of \( G(\beta, q) \) shows the continuity of the result for \( q \to 0 \).

The following inequality holds with probability 1 for all \( i \in [n] \) and \( k \in [d] \) under
the event $E_{i}^\text{edge}(\mathcal{E}_T)$ (Lemma 13),

$$
|\xi_{k}^{(i)}| \leq 2 + \left| \mathbb{E} \left[ Z_{k}^{(i)} | \mathcal{F}_{i-1}^k \right] \right|
= 2 + \left| \hat{f}_{\mu_{k-1}}(q) \prod_{j=1}^{k-1} \mu_{j}^\dagger \prod_{j=k+1}^{d} \mu_{j} \right| \frac{1}{(1 - 2q)^2}
\leq 2 + \left| \hat{f}_{\mu_{k-1}}(q) \right| \frac{1}{(1 - 2q)^2}
= 2 + \frac{1}{(1 - 2q)^2} \left| \mu_{k-1} \right| \left| \left( \mu_k - \mu_k^\dagger \right) \frac{\mu_{k-1}^\dagger - \mu_{k-1}^\dagger}{1 - (\mu_{k-1})^2} \right|
= 2 + \tan^2(\beta)(1 - (1 - 2q)^2) \frac{\gamma_{k-1}^\dagger}{1 - (\mu_{k-1})^2}
\leq 2 + \frac{(1 - 2q)^2}{6} (1 - (1 - 2q)^2) \tan^2(\beta) \equiv S(\beta, q). \tag{3.122}
$$

the last step comes form the inequality $\gamma_{e}^\dagger \leq \frac{(1 - 2q)^2}{6} (1 - (\mu_{e})^2)$ (3.86), which holds if the inequality $n > \frac{108e^2}{K(\beta,q)(1-2q)} \log(4p)$ holds (see 3.84). Recall that Lemma 13 gives

$$
\gamma_e = \sqrt{3 \frac{1 - \mu_e^2}{nK(\beta,q) \log(2p^2/\delta)}} \leq \sqrt{3 \frac{1 - \tan^2(\beta)}{nK(\beta,q) \log(2p^3/\delta)}} \leq \sqrt{\frac{3 \log(2p^3/\delta)}{n}} \equiv \bar{\gamma}.
\tag{3.123}
$$
Also, for all $i \in [n]$ we have

\[
\begin{align*}
&\sum_{k=1}^{d} \mathbb{E} \left[ Z^{(i)}_{k} | \mathcal{F}^{k}_{i-1} \right] \\
= &\sum_{k=2}^{d} \mathbb{E} \left[ Z^{(i)}_{k} | \mathcal{F}^{k}_{i-1} \right] \\
\leq &\sum_{k=2}^{d} \mathbb{E} \left[ Z^{(i)}_{k} | \mathcal{F}^{k}_{i-1} \right] \\
= &\sum_{k=2}^{d} ⌊\hat{\mu}_{k-1}^{d}(q)⌋ \frac{\prod_{j=1}^{k-1} |\hat{\mu}_{j}^{\dag}| \prod_{j=k+1}^{d} |\mu_{j}^{\dag}|}{(1 - 2q)^{2d-2}} \\
\leq &\frac{1}{(1 - 2q)^{2}} \sum_{k=2}^{d} |\mu_{k-1}^{\dag}| |\mu_{k}^{\dag}| \frac{|\hat{\mu}_{k-1}^{d} - \mu_{k}^{d}|}{1 - (\mu_{k-1}^{d})^{2}} \frac{\prod_{j=1}^{k-1} |\hat{\mu}_{j}^{\dag}| \prod_{j=k+1}^{d} |\mu_{j}^{\dag}|}{(1 - 2q)^{2}} \\
\leq &\tanh^{2}(\beta)(1 - (1 - 2q)^{2}) \frac{\gamma_{e}}{1 - (1 - 2q)^{4}} \sum_{k=2}^{d} \prod_{j=1}^{k-1} |\hat{\mu}_{j}^{\dag}| \prod_{j=k+1}^{d} |\mu_{j}^{\dag}| \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \prod_{j=1}^{d} |\hat{\mu}_{j}^{\dag}| \prod_{j=k+1}^{d} |\mu_{j}^{\dag}| \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \prod_{j=1}^{d} \tanh(\beta) + \frac{\gamma_{j}}{(1 - 2q)^{2}} \prod_{j=k+1}^{d} \tanh(\beta) \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \sum_{k=2}^{d} \prod_{j=1,j\neq k}^{d} \tanh(\beta) + \frac{\gamma_{j}}{(1 - 2q)^{2}} \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \left( \tanh(\beta) + \frac{1}{6}(1 - \tanh^{2}(\beta)) \right) \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \frac{(d - 1)}{3} \left( \frac{5}{3} - \frac{1}{6}(\tanh(\beta) - 3)^{2} \right)^{d-1} \\
&\leq \tanh^{2}(\beta) \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \frac{1}{\tanh^{2}(\beta)} - e \log \left( \frac{5}{3} - \frac{1}{6}(\tanh(\beta) - 3)^{2} \right) \\
&\leq \frac{(1 - (1 - 2q)^{2})\gamma_{e}}{1 - (1 - 2q)^{4}} \sqrt{\frac{3 \log(2p^{2}/\delta)}{n}} \triangleq \Delta, \tag{3.127}
\end{align*}
\]
where (3.124), (3.125), (3.126) come from Lemma 13 and (3.123). Finally, $0 < \frac{5}{3} - \frac{1}{6} \left( \tanh(\beta) - 3 \right)^2 < 1$ for all $\beta > 0$ and $-1/\log \left( \frac{5}{3} - \frac{1}{6} \left( \tanh(\beta) - 3 \right)^2 \right) \leq e^{2\beta}$.

We use the symbol $E_{A_{k-1}}[\cdot]$ to denote the conditional expectation given the event $E_{\text{edge}}(A_{k-1})$, for instance

$$E \left[ \exp \left( \lambda \sum_{i=1}^{k} M_i^\dagger \right) \bigg| E_{\text{edge}}(A_{k-1}) \right] \equiv E_{A_{k-1}} \left[ \exp \left( \lambda \sum_{i=1}^{k} M_i^\dagger \right) \right]. \quad (3.128)$$

Further we define the function $F(\cdot, \cdot)$ as

$$F(t, \lambda) = \log \left( \frac{1}{1+t} e^{-\lambda t} + \frac{t}{1+t} e^{\lambda} \right). \quad (3.129)$$
For any $k \leq d$ we have

$$
E \left[ \exp \left( \sum_{i=1}^{k} M_i^\dagger \right) \bigg| E_{\text{edge}}^\dagger (A_{k-1}) \right]
$$

$$
= E_{A_{k-1}} \left[ \exp \left( \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E \left[ \exp \left( \sum_{i=1}^{n} Z_i^{(i)} \big| \hat{\mu}_1, \ldots, \hat{\mu}_{k-1} \right) \right] \right]
$$

$$
= E_{A_{k-1}} \left[ \exp \left( \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E \left[ \exp \left( \sum_{i=1}^{n} Z_i^{(i)} \big| \hat{\mu}_1, \ldots, \hat{\mu}_{k-1} \right) \right] \right]
$$

$$
\leq \exp \left( \lambda \mathbb{E} \left[ Z_i^{(i)} \big| F_{i-1}^k \right] \right)
$$

$$
\times E_{A_{k-1}} \left[ \exp \left( \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E \left[ \exp \left( \sum_{i=1}^{n} Z_i^{(i)} \big| \hat{\mu}_1, \ldots, \hat{\mu}_{k-1} \right) \right] \right]
$$

$$
\leq \exp \left( \lambda \mathbb{E} \left[ Z_i^{(i)} \big| F_{i-1}^k \right] \right)
$$

$$
\times E_{A_{k-1}} \left[ \exp \left( \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E \left[ \exp \left( \sum_{i=1}^{n} Z_i^{(i)} \big| \hat{\mu}_1, \ldots, \hat{\mu}_{k-1} \right) \right] \right]
$$

$$
\leq \exp \left( \lambda \mathbb{E} \left[ Z_i^{(i)} \big| F_{i-1}^k \right] \right)
$$

$$
\times \exp \left\{ n F \left( \frac{\mathbb{E} \left[ \left( Z_i^{(i)} \right)^2 \big| F_{i-1}^k \right]}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right) \right\}
$$

$$
\leq \exp \left( \lambda \mathbb{E} \left[ Z_i^{(i)} \big| F_{i-1}^k \right] \right)
$$

$$
\times \exp \left\{ n F \left( \frac{\mathbb{E} \left[ \left( Z_i^{(i)} \right)^2 \big| F_{i-1}^k \right]}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right) \right\}
$$

(3.135)
The equation (3.130) comes from change of measure and tower property, the definitions (3.101) and (3.102) of $M_k^\dagger$ and $\xi_k^{(i)}$ respectively give (3.131) and (3.132). The (3.133) is derived by upper bounding the quantity $\left| E \left[ Z_k^{(i)} | \mathcal{F}_{i-1}^k \right] \right|$ similarly to (3.122), (3.134) is the upper bound on the moment generating function of the supermartingale [68]. To get a recurrence we proceed as follows:

$$
\begin{align*}
E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E^\mathrm{edge}_\dagger(A_{k-1}) \right] \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-1}) \right)
= E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \cap E^\mathrm{edge}_{e_{k-1},\dagger} \right] \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-1}) \right)
= E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \cap E^\mathrm{edge}_{e_{k-1},\dagger} \right] \times \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-1}) \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \cap E^\mathrm{edge}_{e_{k-1},\dagger} \right) \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-2}) \right)
= E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \cap E^\mathrm{edge}_{e_{k-1},\dagger} \right] \times \mathbb{P} \left( E^\mathrm{edge}_{e_{k-1},\dagger} \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \cap E^\mathrm{edge}_{e_{k-1},\dagger} \right) \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-2}) \right)
= E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \mathbf{1}_{E^\mathrm{edge}_{e_{k-1},\dagger}} \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \right] \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-2}) \right)
\leq E \left[ \exp \left( \lambda \sum_{i=1}^{k-1} M_i^\dagger \right) \bigg| E^\mathrm{edge}_\dagger(A_{k-2}) \right] \mathbb{P} \left( E^\mathrm{edge}_\dagger(A_{k-2}) \right). \quad (3.136)
\end{align*}
$$
By applying the recurrence $d$ times, we derive the following bound

$$
\mathbb{E} \left[ \exp \left( \lambda \sum_{i=1}^{k} M_i^1 \right) \bigg| \mathcal{E}_{\text{edge}}^{\text{edge}}(A_{k-1}) \right] 
\leq \exp \left( \lambda \sum_{k=1}^{d} \mathbb{E} \left[ Z_k^{(i)} \bigg| \mathcal{F}_{i-1}^k \right] \right) 
\times \exp \left\{ nd \frac{d}{d} \sum_{k=1}^{d} \frac{\mathbb{E} \left[ (\xi_k^{(i)})^2 \bigg| \mathcal{F}_{i-1}^k \right]}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right\} 
\leq \exp \left( \lambda \sum_{k=1}^{d} \mathbb{E} \left[ Z_k^{(i)} \bigg| \mathcal{F}_{i-1}^k \right] \right) 
\times \exp \left\{ nd \frac{d}{d} \sum_{k=1}^{d} \frac{\mathbb{E} \left[ (\xi_k^{(i)})^2 \bigg| \mathcal{F}_{i-1}^k \right]}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right\}.
$$

Further (3.121), (3.122), (3.127) and (3.137) give

$$
\mathbb{E} \left[ \exp \left( \lambda \sum_{i=1}^{k} M_i^1 \right) \bigg| \mathcal{E}_{\text{edge}}^{\text{edge}}(A_{k-1}) \right] 
\leq \exp (\lambda \Delta(\beta, q)) \exp \left\{ nd \frac{1}{d} \frac{G(\beta, q)}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right\} 
\leq \exp \left\{ nd \frac{1}{d} \frac{G(\beta, q)}{S(\beta, q)^2}, |\lambda| \frac{S(\beta, q)}{n} \right\} + \lambda \Delta(\beta, q) \right\}.
$$

For sake of space, we denote the functions $G(\beta, q)$, $S(\beta, q)$, and $\Delta(\beta, q)$ as $G$, $S$, and $\Delta$. 
respectively. It is true that

$$
\mathbb{E} \left[ \exp \left( \lambda \sum_{k=1}^{d} M_k^\dagger \right) \bigg| E_{\text{edge}}(A_{d-1}) \right] P \left[ E_{\text{edge}}(A_{d-1}) \cap E_q \right] \leq \exp \left\{ nd F \left( \frac{G}{d}, \frac{|\lambda| S}{n} \right) + \lambda \Delta \right\} .
$$

(3.139)

Under the assumption \( n > \frac{108e^{2\beta}}{(1-2q)^2 K(\beta, q)} \log(4p) \), we have

$$
P \left[ \left( E_{\text{edge}}(A_{d-1}) \right)^c \right] \leq \frac{1}{2} .
$$

(3.140)

The latter gives

$$
\mathbb{E} \left[ \exp \left( \lambda \sum_{k=1}^{d} M_k^\dagger \right) \bigg| E_{\text{edge}}(A_{d-1}) \right] \leq 2 \exp \left\{ nd F \left( \frac{G}{d}, \frac{|\lambda| S}{n} \right) + \lambda \Delta \right\} ,
$$

(3.141)

which implies that

$$
P \left[ \sum_{k=1}^{d} M_k^\dagger \geq \gamma \bigg| E_{\text{edge}}(A_{d-1}) \right] \leq 2 \min_{\lambda > 0} \exp \left\{ nd F \left( \frac{G}{d}, \frac{\lambda S}{n} \right) + \lambda \Delta - \lambda \gamma \right\}

= 2 \min_{\lambda > 0} \exp \left\{ nd F \left( \frac{G}{d}, \frac{\lambda S}{n} \right) + \lambda (\Delta - \gamma) \right\} ,
$$

(3.142)

and we define \( \gamma' \triangleq \gamma - \Delta \). The minimum value is attained at

$$
\lambda^* = \frac{n/S}{1 + G/d} \log \frac{1 + \frac{\gamma'}{dS}}{1 - \frac{\gamma'}{dS}}
$$

(3.143)
and by substituting the optimal value we get

\[
\exp \left\{ nd \left( \frac{G}{d}, \frac{\lambda S}{n} \right) - \lambda^* \gamma' \right\} \\
= \left[ \frac{1}{1 + \frac{\gamma'}{d}} \left( 1 + \frac{\gamma'}{GS} \right) \right. \\
\left. + \frac{G}{d} \frac{1}{1 + \frac{\gamma'}{d}} \left( 1 - \frac{\gamma'}{dS} \right) \left( 1 - \frac{\gamma'}{dS} \right) \right]^{nd}
\]

Then (3.142) and (3.144) give

\[
\mathbb{P} \left[ \sum_{k=1}^{d} M_k^+ \geq \gamma \right] \leq 2 \left[ \left( 1 + \frac{\gamma'}{GS} \right) \left( 1 - \frac{\gamma'}{dS} \right) \right]^{nd}.
\]
As a final step we want to express the upper bound as an exponential function of $\gamma$, we define $\zeta \triangleq \gamma'/(SG)$ and we proceed as follows:

$$d \left[ \left( \frac{G/d}{1 + G/d} + \frac{\gamma'/(dS)}{1 + G/d} \right) \log \left( 1 + \frac{\gamma' G}{G} \right) + \left( \frac{1}{1 + G/4d} - \frac{\gamma'/dy}{1 + G/4d} \right) \log \left( 1 - \frac{\gamma'}{dy} \right) \right]$$

$$\geq d \left[ \left( \frac{G/d}{1 + G/d} + \frac{\gamma'/(dS)}{1 + G/d} \right) \left( \frac{\gamma' G}{G} - \frac{1}{2} \left[ \frac{\gamma'}{G} \right]^2 \right) \right.$$

$$- \left( \frac{1}{1 + G/d} - \frac{\gamma'/(dS)}{1 + G/d} \right) \log \left( 1 + \frac{\gamma'/dS}{1 - \gamma'/dS} \right)$$

$$\geq \left[ \left( \frac{dG}{d + G} + \frac{d\gamma'/S}{(d + G)} \right) \left( \frac{\gamma' G}{G} - \frac{1}{2} \left[ \frac{\gamma'}{G} \right]^2 \right) \right.$$

$$- \left( \frac{d^2}{d + G} - \frac{d\gamma'/S}{(d + G)} \right) \left( \frac{\gamma'/dS}{1 - \gamma'/dS} - \frac{1}{2} \left( \frac{\gamma'/dS}{1 - \gamma'/dS} \right)^2 \right)$$

$$= \frac{d}{d + G} \left[ (G + \gamma'/S) \left( \frac{\gamma' G}{G} - \frac{1}{2} \left[ \frac{\gamma'}{G} \right]^2 \right) \right.$$

$$- (d - \gamma'/S) \left( \frac{\gamma'/dS}{1 - \gamma'/dS} - \frac{1}{2} \left( \frac{\gamma'/dS}{1 - \gamma'/dS} \right)^2 \right)$$

$$\geq \frac{d}{d + G} \left[ (G + \gamma'/(SG)) \left( \frac{\gamma' G}{G} - \frac{1}{2} \left[ \frac{\gamma'}{G} \right]^2 \right) \right.$$

$$- \gamma'/S$$

$$= \frac{dG}{d + G} \left[ (1 + \gamma'/(SG)) \left( \frac{\gamma' G}{G} - \frac{1}{2} \left[ \frac{\gamma'}{G} \right]^2 \right) \right.$$  

$$- \gamma'/(SG)$$

$$= \frac{dG}{d + G} \left[ (1 + \zeta) \left( \zeta - \frac{1}{2} \zeta^2 \right) - \zeta \right]$$

$$\geq \frac{2G}{2 + G} \left[ (1 + \zeta) \left( \zeta - \frac{1}{2} \zeta^2 \right) - \zeta \right]$$

$$\geq \left( \frac{3\gamma'}{10GS} \right)^2, \quad \forall \gamma' \in \left( 0, \frac{SG}{3} \right). \quad (3.146)$$

Recall that $\zeta \triangleq \gamma'/(SG)$, $\gamma' = \gamma - \Delta$. If $\Delta < \gamma \leq S(\beta, q)G(\beta, q)/3 + \Delta$ then (3.145) and (3.146) give

$$\mathbb{P} \left[ \sum_{k=1}^{d} M_k^t \geq \gamma \right| E_{t}^{\text{edge}}(A_{d-1}) \right] \leq 2 \exp \left( -0.3^2 n \frac{(\gamma - \Delta)^2}{S^2(\beta, q)G^2(\beta, q)} \right). \quad (3.147)$$
In a similar way we derive the bound
\[
\mathbb{P} \left[ \sum_{k=1}^{d} M_k^i \leq -\gamma \left| E^\text{dgc}_i (A_{d-1}) \right| \right] \leq 2 \exp \left( -0.3^2 n \frac{(\gamma - \Delta)^2}{S^2(\beta, q) G^2(\beta, q)} \right). \tag{3.148}
\]

Finally, we combine (3.99), Lemma 13, (3.147) and (3.148) to derive the bound (3.95) which guarantees that
\[
\mathbb{P} \left[ \left| \prod_{i=1}^{d} \frac{\hat{\mu}_i}{(1 - 2q)^2} - \prod_{i=1}^{d} \frac{\mu_i}{(1 - 2q)^2} \right| > \gamma \right] \leq 2 \delta/p^2, \quad \forall d \geq 2. \tag{3.149}
\]

To summarize we proved that the event \( E^\text{cascade}_i (\gamma_i) \) happens with probability at least \( 1 - 2\delta/p^2 \) by combining Bresler’s and Karzand’s technique, the Corollary 2.3 by [68] and Lemma 12.

\[\square\]

### 3.5.5 Predictive Learning, Proof of Theorem 5 and Theorem 6

Recall that our goal is to guarantee that the quantity \( \mathcal{L}^{(2)}(p(\cdot), \Pi_{\hat{t}}^{\text{CL}}(\hat{p}_t)) \) is smaller than a number \( \eta > 0 \) with probability at least \( 1 - \delta \). To do this, we use the triangle inequality as
\[
\mathcal{L}^{(2)}(p(\cdot), \Pi_{\hat{t}}^{\text{CL}}(\hat{p}_t)) \leq \mathcal{L}^{(2)}(p(\cdot), \Pi_{\hat{t}}^{\text{CL}}(p(\cdot))) + \mathcal{L}^{(2)}(\Pi_{\hat{t}}^{\text{CL}}(p(\cdot)), \Pi_{\hat{t}}^{\text{CL}}(\hat{p}_t)) \tag{3.150}
\]
and we find the required number of samples such that each of the terms \( \mathcal{L}^{(2)}(p(\cdot), \Pi_{\hat{t}}^{\text{CL}}(p(\cdot))) \) and \( \mathcal{L}^{(2)}(\Pi_{\hat{t}}^{\text{CL}}(p(\cdot)), \Pi_{\hat{t}}^{\text{CL}}(\hat{p}_t)) \) in (3.150) is less than \( \eta/2 \) with probability at least \( 1 - \delta \). The next Lemma provides the necessary bounds on \( \gamma_\hat{t} \) and \( \epsilon_\hat{t} \) that guarantee
\[
\mathcal{L}^{(2)}(\Pi_{\hat{t}}^{\text{CL}}(p(\cdot)), \Pi_{\hat{t}}^{\text{CL}}(\hat{p}_t)) \leq \eta/2.
\]

**Lemma 15.** If \( \gamma_\hat{t} \leq \frac{\eta}{3} \) and
\[
\epsilon_\hat{t} \leq (1 - 2q)^2 e^{-\beta} \left[ 20 \left( 1 + 2e^{\beta} \sqrt{2(1 - q) q \tanh \beta} \right) \right]^{-1},
\]
then $\mathcal{L}^{(2)}(\Pi_{T^{\text{CL}}}(p(\cdot)), \Pi_{T^{\text{CL}}}(\tilde{p})) \leq \eta/2$ under the event $E_i^{\text{corr}}(\epsilon_i) \cap E_i^{\text{cascade}}(\gamma_i) \cap E_i^{\text{strong}}(\epsilon_i)$.

**Proof.** The derivation of the bound is similar to the approach by Bresler and Karzand [29, Section 6.1] but with different calculations. In the hidden model, we consider the path between two nodes $i, j$ in the estimated structure $T^{\text{CL}}$, namely path $T^{\text{CL}}(i, j)$, to be $(F_0, e_1, F_1, e_1, \ldots, F_{t-1}, e_t, F_t)$, and $F_i$ are segments with all strong edges and $e_i$ are all weak edges. We consider the case of at least one weak edge to exist in the path. If there is no weak edge the bound reduces to the case of Lemma 14. The length of each sub-path $F_i$ is denoted as $d_i$, for all $i \in \{0, 1, \ldots, t\}$. Each segment (sub-path) $F_i$ has exactly $d_i$ edges, and the total number of edges in the path are $d$; thus $d = \sum_{i=0}^{t} d_i + t$.

Note that $t \geq 1$ and $d_i \geq 0$ for all $i \in \{0, 1, \ldots, t\}$. Recall that

$$
\Pi_{T^{\text{CL}}}(p(\cdot)) = \frac{1}{2} \prod_{(i,j) \in E^{\text{CL}}_T} \frac{1 + x_i x_j E[X_i X_j]}{2} = \frac{1}{2} \prod_{(i,j) \in E^{\text{CL}}_T} \frac{1 + x_i x_j \hat{E}[Y_i Y_j]}{(1-2q)^2} \quad (3.151)
$$

(the latter comes from (2.12)), and

$$
\Pi_{T^{\text{CL}}}(\tilde{p}) \triangleq \frac{1}{2} \prod_{(i,j) \in E^{\text{CL}}_T} \frac{1 + x_i x_j \hat{\tilde{E}}[Y_i Y_j]}{(1-2q)^2}. \quad (3.152)
$$

Further, for any tree-structured Ising model distributions $P, \tilde{P}$ with structures $T = (V, E)$ and $\tilde{T} = (\tilde{V}, \tilde{E})$ respectively, we have

$$
\mathcal{L}^{(2)}(P, \tilde{P}) \triangleq \sup_{i,j \in V} \frac{1}{2} \sum_{x_i, x_j \in \{-1, +1\}^2} \left| P(x_i, x_j) - \tilde{P}(x_i, x_j) \right| \quad (3.153)
$$

$$
= \sup_{i,j \in V} \frac{1}{2} \left| \prod_{e \in \text{path}_T(i,j)} \mu_e - \prod_{e' \in \text{path}_{\tilde{T}}(i,j)} \tilde{\mu}_{e'} \right|. \quad (3.154)
$$
To upper bound the quantity $\mathcal{L}^{(2)}(\Pi_{\text{T}^{\text{CL}}}(p(\cdot)), \Pi_{\text{T}^{\text{CL}}}(\hat{p}_t))$ we have

\[
2\mathcal{L}^{(2)}\left(\Pi_{\text{T}^{\text{CL}}}(p(\cdot)), \Pi_{\text{T}^{\text{CL}}}(\hat{p}_t)\right)
\]

\[
= \left| \prod_{e \in \text{path}_{\text{T}^{\text{CL}}}(i,j)} \frac{\mu_{e}^\dagger}{(1-2q)^2} - \prod_{e \in \text{path}_{\text{T}^{\text{CL}}}(i,j)} \frac{\mu_{e}^\dagger}{(1-2q)^2} \right| \]

\[
= \frac{1}{(1-2q)^{2d}} \left| \mu_{F_0}^\dagger \prod_{i=1}^{t} \mu_{F_i}^\dagger \mu_{e_i}^\dagger \right| \]

\[
\leq \frac{1}{(1-2q)^{2d}} \left[ \left| \mu_{F_0}^\dagger - \prod_{i=1}^{t} \mu_{F_i}^\dagger \right| \right]
\]

\[
= \frac{1}{(1-2q)^{2d}} \left[ \left| \mu_{F_0}^\dagger \prod_{j=1}^{t} \mu_{F_j}^\dagger \right| \right]
\]

\[
\leq \frac{1}{(1-2q)^{2d}} \left[ \left| \mu_{F_0}^\dagger \prod_{j=1}^{t} \mu_{F_j}^\dagger \right| \right]
\]

\[
\leq \gamma^t \left( \frac{\tau^t}{(1-2q)^2} \right)^{t-1} + \left( \frac{\tau^t + \epsilon^t}{(1-2q)^2} \right)^{t-1} \sum_{i=1}^{t} \left| \mu_{F_i}^\dagger \mu_{e_i}^\dagger \right| \]

\[
\leq \gamma^t \left( \frac{\tau^t}{(1-2q)^2} \right)^{t-1} + \left( \frac{\tau^t + \epsilon^t}{(1-2q)^2} \right)^{t-1} \sum_{i=1}^{t} \left( \gamma^t + \frac{\epsilon^t}{(1-2q)^2} \right)
\]

\[
\leq \gamma^t \left( \frac{\tau^t}{(1-2q)^2} \right)^{t-1} + \left( \frac{\tau^t + \epsilon^t}{(1-2q)^2} \right)^{t-1} \sum_{i=1}^{t} \left( \gamma^t + \frac{\epsilon^t}{(1-2q)^2} \right)
\]

\[
\leq \gamma^t \left( \frac{\tau^t + \epsilon^t}{(1-2q)^2} \right)^{t-1} \left( 2t + 1 \right) \max \left\{ \gamma^t, \frac{\epsilon^t}{(1-2q)^2} \right\}
\]

\[
\leq \left( \frac{4 \epsilon^t e^\beta (1 + 2e^\beta \sqrt{2(1-q)q \tanh \beta}) + \epsilon^t}{(1-2q)^2} \right)^{t-1} \left( 2t + 1 \right) \max \left\{ \gamma^t, \frac{\epsilon^t}{(1-2q)^2} \right\}
\]

\[
\leq \left( \frac{5 \epsilon^t e^\beta}{(1-2q)^2} \right)^{t-1} \left( 1 + 2e^\beta \sqrt{2(1-q)q \tanh \beta} \right)^{t-1} \left( 2t + 1 \right) \max \left\{ \gamma^t, \frac{\epsilon^t}{(1-2q)^2} \right\}
\]

\[
\leq \frac{2t + 1}{4 \epsilon^t - 1} \eta \left( \frac{3}{3} \right)
\]

\[
\leq \eta.
\]
Telescoping summation and triangle inequality give (3.155) and (3.156). We use the definition of $d$, $d = \sum_{i=0}^{t} d_i + t$ to get (3.157). The inequalities $|\mu_{x_i}^\dagger| \leq (1 - 2q)^{2d_i}$, $|\bar{\mu}_{x_i}^\dagger| \leq (1 - 2q)^{2d_i}$, $|\mu_{e_i}^\dagger| \leq \tau^\dagger + \epsilon^\dagger$ hold under $E_{\text{corr}}^{\dagger}(\epsilon^\dagger)$, $E_{\text{strong}}^{\dagger}(\epsilon^\dagger)$. Further, under event $E_{\text{cascade}}^{\dagger}(\gamma^\dagger)$ (Lemma 14) it is true that $|\bar{\mu}_{x_i}^\dagger - \mu_{x_i}^\dagger| \leq \gamma^\dagger$, the latter give (3.158) and (3.159). The bound $\tau^\dagger \leq 4\epsilon^\dagger e^\beta (1 + 2e^\beta \sqrt{2 (1 - q) q \tanh \beta})$ gives (3.160). Inequality (3.161) requires

$$\max \left\{ \frac{\epsilon^\dagger}{(1 - 2q)^2}, \gamma^\dagger \right\} \leq \frac{\eta}{3}$$

and

$$\epsilon^\dagger \leq (1 - 2q)^2 e^{-\beta} \left[ 20 \left( 1 + 2e^\beta \sqrt{2 (1 - q) q \tanh \beta} \right) \right]^{-1}.$$ (3.164)

Finally (3.162) holds for all $t \in \mathbb{N}$. The latter completes the proof. 

The next Lemma provides the set of values of $\epsilon^\dagger$ that guarantee $\mathcal{L}^{(2)}(p(\cdot), \Pi_{T^{\text{CL}}}(p(\cdot))) \leq \frac{\eta}{2}$ with high probability.

**Lemma 16.** If

$$\epsilon^\dagger \leq \min \left\{ \frac{\eta}{16} (1 - 2q)^2, \frac{(1 - 2q)^2 e^{-\beta}}{24 \left( 1 + 2e^\beta \sqrt{2 (1 - q) q \tanh \beta} \right)} \right\}.$$ (3.165)

then $\mathcal{L}^{(2)}(p(\cdot), \Pi_{T^{\text{CL}}}(p(\cdot))) \leq \frac{\eta}{2}$ under the event $E_{\text{corr}}^{\dagger}(\epsilon^\dagger) \cap E_{\text{strong}}^{\dagger}(\epsilon^\dagger)$.

**Proof.** Recall that

$$\mathcal{L}^{(2)}(p(\cdot), \Pi_{T^{\text{CL}}}(p(\cdot))) = \frac{1}{2} \left| \prod_{e \in \text{path}_T(w, \tilde{w})} \frac{\mu_e^\dagger}{(1 - 2q)^2} - \prod_{e \in \text{path}_{T^{\text{CL}}}(w, \tilde{w})} \frac{\mu_e^\dagger}{(1 - 2q)^2} \right|$$

$$= \frac{1}{2} \left| \prod_{e \in \text{path}_T(w, \tilde{w})} \mu_e - \prod_{e \in \text{path}_{T^{\text{CL}}}(w, \tilde{w})} \mu_e \right|.$$ (3.166)

We follow Bresler’s and Karzand’s technique “Loss due to graph estimation” [29, Section
and highlight the difference that appears in our setting. For the noisy case/hidden model, the argument changes slightly in the following manner:

\[
2\mathcal{L}^{(2)}\left(p(\cdot), \Pi_{\mathcal{T}_1} p(\cdot)\right) \leq |\mu_f \mu_A \bar{\mu} \mu_B \mu_B^*| |\mu_C^2 \mu_C^2 - 1| + |\mu_f| \left(\Delta(k) + \Delta(\hat{k}) + \Delta(\tilde{k}) \Delta(k)\right)
\]

\[
= \frac{\mu_f^*}{(1 - 2q)^2} |\mu_A \bar{\mu} \mu_B | \left| |\mu_C^2 \mu_C^2 - 1| + \frac{\mu_f^*}{(1 - 2q)^2} \left(\Delta(k) + \Delta(\hat{k}) + \Delta(\tilde{k}) \Delta(k)\right)\right|
\]

\[
\leq 8 \frac{\epsilon_t}{(1 - 2q)^2} + \frac{\tau_t}{(1 - 2q)^2} (2\eta + \eta^2)
\]

(3.167)

\[
\leq \eta.
\]

(3.168)

(3.167) holds since $|\mu_f^*| - |\mu_g^*| \leq 4\epsilon_t$, $|\mu_f| \left(1 - \mu_C^2 \mu_C^2\right) \leq 2|\mu_f| - 2|\mu_g|$, $|\mu_f^*| \leq \tau_t$, and (3.168) holds for all the values of $\epsilon_t$ that satisfy

\[
\epsilon_t \leq \min \left\{ \frac{\eta}{16(1 - 2q)^2}, \frac{(1 - 2q)^2 e^{-\beta}}{24 (1 + 2e^\beta \sqrt{2(1 - q) q \tanh \beta})} \right\}.
\]

(3.169)

The latter provides the statement of the Lemma.

The next Theorem provides the sufficient number of samples for predictive learning that recovers exactly the noiseless setting for $q = 0$. Note that the dependence on $\beta$ changes from $e^{2\beta}$ to $e^{4\beta}$ when the data are noisy. A key component of the bound is the following function

\[
\Gamma(\beta, q) \triangleq \left(\frac{1 - (1 - 2q)^2}{1 - (1 - 2q)^2 \tanh^2(\beta)}\right)^2, \quad \beta > 0 \text{ and } q \in [0, 1/2).
\]

(3.170)

Note that $\Gamma(\beta, q) \in [0, 1]$ for all $\beta > 0$ and $q \in [0, 1/2)$, and $\Gamma(\beta, 0) = 0$ for all $\beta > 0$.

Further we define

\[
B(\beta, q) \triangleq \max \left\{ \frac{1}{K(\beta, q)}, \left(1 + 2e^\beta \sqrt{2(1 - q) q \tanh \beta}\right)^2 \right\},
\]

(3.171)
and the expression of $K(\beta, q)$ is given by (3.93).

**Theorem 11.** Fix $\delta \in (0, 1)$. Choose $\eta > 0$ (independent of $\delta$). If

$$n \geq \max \left\{ \frac{512}{\eta^2(1-2q)^4}, \frac{1152e^{2\beta B(\beta, q)}}{(1-2q)^4}, \frac{48e^{4\beta}}{\eta^2} \Gamma(\beta, q) \right\} \log \left( \frac{6p^3}{\delta} \right),$$

(3.172)

then

$$\mathbb{P} \left( \mathcal{L}(2) \left( p(\cdot), \Pi_{TCL}(\hat{p}_\dagger) \right) \leq \eta \right) \geq 1 - \delta.$$  \hspace{1cm} (3.173)

Additionally, as a consequence of (3.172), if

$$n \geq \max \left\{ \frac{512}{\eta^2(1-2q)^4}, \frac{1152(1+3\sqrt{q})^2e^{2\beta(1+1_q)}(1+\sqrt{q})}{(1-2q)^4}, \frac{48e^{4\beta}}{\eta^2} \right\} \log \left( \frac{6p^3}{\delta} \right),$$

(3.174)

then

$$\mathbb{P} \left( \mathcal{L}(2) \left( p(\cdot), \Pi_{TCL}(\hat{p}_\dagger) \right) \leq \eta \right) \geq 1 - \delta.$$  \hspace{1cm} (3.175)

**Proof.** Recall that

$$\mathcal{L}(2) \left( p(\cdot), \Pi_{TCL}(\hat{p}_\dagger) \right) = \frac{1}{2} \left| \prod_{e \in \text{path}_T(w, \tilde{w})} \mu_e - \prod_{e \in \text{path}_{T_{TCL}}(w, \tilde{w})} \mu_e \right|.$$  \hspace{1cm} (3.176)

We combine the triangle inequality

$$\mathcal{L}(2) \left( p(\cdot), \Pi_{TCL}(\hat{p}_\dagger) \right) \leq \mathcal{L}(2) \left( p(\cdot), \Pi_{TCL}(p(\cdot)) \right) + \mathcal{L}(2) \left( \Pi_{TCL}(p(\cdot)), \Pi_{TCL}(\hat{p}_\dagger) \right),$$

(3.177)

Lemma 15, and Lemma 16 to get that $\mathcal{L}(2)(p(\cdot), \Pi_{TCL}(\hat{p}_\dagger)) \leq \eta$ with probability at least $1 - \delta$ if

$$\gamma_\dagger \leq \frac{\eta}{3} \text{ and } \epsilon_\dagger \leq \min \left\{ \frac{\eta}{16} (1-2q)^2, \frac{(1-2q)^2 e^{-\beta}}{24 \left( 1 + 2e^\beta \sqrt{2 (1 - q) q \tanh \beta} \right)} \right\}.$$  \hspace{1cm} (3.178)
First, we find the necessary number of samples such that for $\gamma \leq \eta/3$ the probability of the complement of $E_{\gamma}^\text{cascade}$ ($\gamma$) is not greater than $\delta/3$. Recall that

\begin{align}
G &\equiv \frac{3 (3e^{-1}I_{q \neq 0} + 1)}{4(1 - 2q)^2}, \\
S &\equiv 3 - (1 - 2q)^2.
\end{align}

(3.179)  
(3.180)

Recall that

$$\Gamma(\beta, q) \equiv \left( \frac{1 - (1 - 2q)^2}{1 - (1 - 2q)^4 \tanh^2(\beta)} \right)^2, \quad \beta > 0 \text{ and } q \in [0, 1/2).$$

(3.181)

Lemma 14 gives that for any $\Delta > 0$ and $\eta > \Delta$ if

$$n \geq \max \left\{ \frac{0.3^{-2}S^2G^2}{(\eta - \Delta)^2}, \frac{108e^{2\beta}}{(1 - 2q)^4K(\beta, q)} \frac{3e^{4\beta}}{\Delta^2} \Gamma(\beta, q) \right\} \log \left( \frac{6p^3}{\delta} \right),$$

(3.182)

then the probability of the complement of $E_{\gamma}^\text{cascade}$ ($\gamma$) is upper bounded by $\delta/3$ and we write

$$\mathbb{P} \left( \left( E_{\gamma}^\text{cascade} \right)^c \right) \leq \frac{\delta}{3}.$$

(3.183)

Second, we find the necessary number of samples such that the complements of the events $E_{\epsilon}^\text{strong}$ ($\epsilon$) and $E_{\epsilon}^\text{corr}$ ($\epsilon$) occur with probability not greater than $\delta/3$ each. In fact the upper bound on $\epsilon$ (3.178) and Lemma 11 gives that if

$$n \geq \max \left\{ \frac{512}{\eta^2(1 - 2q)^4}, \frac{1152e^{2\beta}}{(1 - 2q)^4} \left(1 + 2e^{\beta} \sqrt{2(1 - q)q \tanh(\beta)} \right)^2 \right\} \log \left( \frac{6p^3}{\delta} \right),$$

(3.184)

then $\epsilon$ satisfies the inequality in (3.178) with probability at least $1 - \delta/3$. Note that (3.182) holds for any $\Delta \in (0, \eta)$ and we will choose $\Delta = \eta/4$. Under the choice $\Delta = \eta/4$

$$\frac{0.3^{-2}S^2G^2}{(\eta - \Delta)^2} = \frac{0.3^{-2}S^2G^2}{(\eta - \eta/4)^2} < \frac{512}{\eta^2}, \quad \forall \eta > 0, q \in [0, 1/2).$$

(3.185)
Recall that

\[ B(\beta, q) \triangleq \max \left\{ \frac{1}{K(\beta, q)}, \left(1 + 2e^\beta \sqrt{2(1-q)q \tanh \beta} \right)^2 \right\}. \quad (3.186) \]

Combining (3.182), (3.184), (3.185) and (3.186) yields

\[ n \geq \max \left\{ \frac{512}{\eta^2 (1-2q)^4}, \frac{1152e^{2\beta} B(\beta, q)}{(1-2q)^4}, \frac{48e^{4\beta}}{\eta^2} \Gamma(\beta, q) \right\} \log \left( \frac{6p^3}{\delta} \right). \quad (3.187) \]

The latter gives the sample complexity for accurate predictive learning, it reduces exactly to the noiseless setting of prior work by [29] and it is continuous because

\[ \lim_{q \to 0^+} \Gamma(\beta, q) = \Gamma(\beta, 0) = 0, \quad \forall \beta > 0 \quad (3.188) \]

and

\[ \lim_{q \to 0^+} K(\beta, q) = K(\beta, 0) = 1, \quad \forall \beta > 0 \quad (3.189) \]

\[ \lim_{q \to 0^+} \left(1 + 2e^\beta \sqrt{2(1-q)q \tanh \beta} \right)^2 = 1 \quad (3.190) \]

thus

\[ \lim_{q \to 0^+} B(\beta, q) = \Gamma(\beta, 0) = 1, \quad \forall \beta > 0. \quad (3.191) \]

To derive a simplified version of (3.187) note that

\[ \frac{1}{K(\beta, q)} \leq e^{2\beta I_{q \neq 0}} \quad (3.192) \]

by the definition (3.93) of \( K(\beta, q) \) and

\[ \left(1 + 2e^\beta \sqrt{2(1-q)q \tanh \beta} \right)^2 \leq \left(e^{\beta I_{q \neq 0}} + 2e^{\beta I_{q \neq 0}} \sqrt{2(1-q)q \tanh \beta} \right)^2 \]

\[ \leq (1 + 3\sqrt{q})^2 e^{2\beta I_{q \neq 0}}. \quad (3.193) \]
Then (3.186), (3.192) and (3.193) give

\[ B(\beta, q) \leq (1 + 3\sqrt{q})^2 e^{2\beta \mathbb{1}_{q \neq 0}} \]  

(3.194)

and by the definition (3.181) \( \Gamma(\beta, q) \in [0, 1) \) and \( \Gamma(\beta, 0) = 0 \) for all \( \beta > 0 \), thus

\[ \Gamma(\beta, q) \leq \mathbb{1}_{q \neq 0}. \]  

(3.195)

Finally, we combine (3.187), (3.194), (3.195) to get

\[ n \geq \max \left\{ \frac{512}{\eta^2 (1 - 2q)^4}, \frac{1152 (1 + 3\sqrt{q})^2 e^{2\beta (1+\mathbb{1}_{q \neq 0})}}{(1 - 2q)^4}, \frac{48e^{4\beta}}{\eta^2} \mathbb{1}_{q \neq 0} \right\} \log \left( \frac{6p^3}{\delta} \right). \]  

(3.196)

This completes the proof.

3.5.6 Theorem 9: KL-Divergence Loss

Assume the Ising model tree distributions \( P_\theta \) according to a tree \( T_\theta = (V, E_\theta) \) and the estimate \( P_{\theta'} \) according a tree \( T_{\theta'} = (V, E_{\theta'}) \) \( \Theta \) is to upper bound the symmetric KL divergence

\[ S_{\text{KL}}(\theta || \theta') = \sum_{s,t \in E} (\theta_{st} - \theta'_{st}) (\mu_{st} - \mu'_{st}) \]

with high probability. Under the event \( E^\text{corr}(\epsilon) \) we can upper bound the quantity \( |\mu_{st} - \mu'_{st}| \) for all \( (s, t) \in E \) with high probability.

By using bounds \( |\theta_{st} - \theta'_{st}| \leq 2\beta \) and \( |\mu_{st} - \mu'_{st}| \leq \epsilon \) for all \( (s, t) \in E \) under the event
$E_{\text{corr}}(\epsilon)$, we have

$$S_{KL}(\theta||\theta') = \left| \sum_{s,t \in E} (\theta_{st} - \theta'_{st}) (\mu_{st} - \mu'_{st}) \right|$$

$$\leq \sum_{s,t \in E} |\theta_{st} - \theta'_{st}| |\mu_{st} - \mu'_{st}|$$

$$\leq (p-1) |\beta - (-\beta)| \epsilon$$

$$\leq \eta S,$$  \hspace{1cm} (3.197)

by assuming $\epsilon \leq \frac{\eta_s}{2\beta(p-1)}$. The sufficient number of samples satisfies the inequality

$$n \geq 4 \log \left( \frac{p^2/\delta}{\frac{\beta^2(p-1)^2}{\eta_s^2}} \right).$$  \hspace{1cm} (3.198)

Now assume that $n_\dagger$ samples of $Y$ are given, by using the estimate $P_{\theta'} = \Pi_\dagger^{\text{CL}}(\hat{p}_\dagger)$ defined in (3.152) under the event $E_{\text{corr}}(\epsilon_\dagger)$ we have $|\mu_{st} - \hat{\mu}'_{st}| \leq \frac{\epsilon_\dagger}{(1-2q)^2}$ from Lemma 11. In the same way by assuming $\epsilon_\dagger \leq \frac{\eta_s(1-2q)^2}{2\beta(p-1)}$, we get

$$n_\dagger \geq 4 \log \left( \frac{p^2/\delta}{\frac{\beta^2(p-1)^2}{(1-2q)^4\eta_s^2}} \right).$$  \hspace{1cm} (3.199)

### 3.5.7 Proof of Theorem 7

We combine Fano’s inequality and a Strong Data Processing Inequality to prove the necessary number of samples in the hidden model setting. \textbf{Proof of Theorem 7:}

Theorem 7 is the extended version of Theorem 3.4 by Bresler and Karzand [29] to the hidden model. Following a similar technique, we consider chain structured Ising models with parameters $\theta^j$ for $j \in [M]$ such that $\theta^j_{j,j+1} = \alpha$ and $\theta^j_{i,i+1} = \text{arctanh}(\text{tanh}(\alpha) + 2\eta)$, for all $i \neq j$. Then

$$\mathcal{L}^{(2)}(P_{\theta'}, P_{\theta'}) = \max_{s,t} \left| E_{\theta'}[X_sX_t] - E_{\theta'}'[X_sX_t] \right| \geq 2\eta$$  \hspace{1cm} (3.200)
and

$$S_{\text{KL}} \left( P_{\theta}, P_{\theta'} \right) \leq 2\eta \left[ \arctanh(\tanh(\alpha) + 2\eta) - \alpha \right] \leq 2\eta \frac{2\eta}{1 - \left[ \tanh(\alpha) + 2\eta \right]^2}, \quad (3.201)$$

where the last inequality is a consequence of Mean Value Theorem (see [29, Section 6.3] for the original statement). We derive the bound of Theorem 7 by combining the strong data processing inequality (2.90) with (2.89), (3.201), and Corollary 1.

### 3.5.8 Supplementary Discussion

In this section we provide supplementary material that supports the discussion in Sections 2.1.4 and 2.1.6. First, we present one marginal case for which perfect denoising is possible before applying the Chow-Liu algorithm. Then we show a structure-preserving case.

#### 3.5.8.1 The Gap between the Upper and Lower Bounds

We continue by analyzing the gap that appears between the upper and lower bounds for an example where perfect denoising can be applied on a specific class of tree models in the high-dimensional regime. This shows why the effect of noise vanishes in Theorems 1.2 and 1.4 for $p \to \infty$. Further, while it seems counter-intuitive that when $p \to \infty$ the problem becomes easier, we show below one example that this is the case. Our lower bound is directly affected by marginal cases like this, for instance see Proposition 1.4.

The gap is introduced by the terms $(1 - 2q)^4$ and $1 - (4q(1 - 2q))^p$ in the denominator of the lower and upper bounds respectively. Specifically, for $p \to \infty$ there exists a special case for which perfect denoising before running the Chow-Liu algorithm is possible, while in other cases that is not possible. Thus the minimax bound ought to be identical to noiseless case when $p \to \infty$ and $1 - (4q(1 - 2q))^p \to 1$ in the large dimensional regime.

We continue by providing the marginal case of a trivial tree structure and showing that perfect denoising is possible in this case before running the Chow-Liu algorithm.

First notice that if $p \to \infty$, then the sample size $n \to \infty$, even in the noiseless regime. Consider the case of $\mathbb{E}[X_iX_j] \to 1$ for all $(i, j) \in \mathcal{E}$. Because an infinite number
of samples are available, we can estimate perfectly the correlations of the observables and we find \( \hat{E}[Y_i Y_j] = E[Y_i Y_j] = (1 - 2q)^2 \) for all \((i, j) \in \mathcal{V}\). The latter as information is sufficient to find that \( E[X_i X_j] \to 1 \) for all \((i, j) \in \mathcal{E}\). The hidden layer \( X \) take two values, \((+1, +1, \ldots) \triangleq +1^p \) \((p \text{ values } +1)\) or \((-1, -1, \ldots) \triangleq -1^p \) \((p \text{ values } -1)\), because \( E[X_i X_j] \to 1 \) for all \((i, j) \in \mathcal{V}\) and the later allows us to denoise each sample. Define as \( d_H(X, Y) \) the Hamming distance between \( X \) and \( Y \). At this point we can perform perfect denoising for each sample \( y_s \) of infinite length \( p \) and find the hidden sample \( x_s \) with probability 1 because

\[
\mathbb{P}(X = x_s | Y = y_s) = \frac{\mathbb{P}(Y = y_s | X = x_s) \mathbb{P}(X = x_s)}{\sum_X \mathbb{P}(Y = y_s | X = x) \mathbb{P}(X = x)} = \frac{q^{d_H(x_s, y_s)}(1 - q)^{p-d_H(x_s, y_s)} - q^{d_H(x_s, y_s)}(1 - q)^{p-d_H(-x_s, y_s)}}{q^{d_H(x_s, y_s)}(1 - q)^{p-d_H(x_s, y_s)} + q^{d_H(-x_s, y_s)}(1 - q)^{p-d_H(-x_s, y_s)}}
\]

(3.202)

and the last holds for both of the cases \( x_s = +1^p \) or \( x_s = -1^p \) because of symmetry. Further for any observation \( y_s \) for any \( q \in (0, 1/2) \) we have

\[
\lim_{p \to \infty} \frac{d_H(-x_s, y_s) - d_H(x_s, y_s)}{p} \text{ a.s.} = 1 - q - q = 1 - 2q.
\]

(3.203)

We combine (3.202) and (3.203) to find

\[
\lim_{p \to \infty} \mathbb{P}(X = x_s | Y = y_s) = \lim_{p \to \infty} \frac{1}{1 + \left(\frac{q}{1-q}\right)\frac{d_H(-x_s, y_s) - d_H(x_s, y_s)}{p}} = 1
\]

(3.204)

and

\[
\lim_{p \to \infty} \mathbb{P}(X = -x_s | Y = y_s) = \lim_{p \to \infty} \frac{1}{1 + \left(\frac{1-q}{q}\right)\frac{d_H(-x_s, y_s) - d_H(x_s, y_s)}{p}} = 0.
\]

(3.205)

As a consequence there exists one case for which perfect denoising is possible before running the Chow-Liu algorithm. Because we want Theorems 1.2 and 1.4 to reduce to the noiseless case for \( p \to \infty \), the above best case scenario must be covered. However, perfect denoising is not possible in general (for instance \( E[X_i X_j] < 1 \) and finite \( p \)).
3.5.8.2 A Structure-Preserving Case

Lemma 17 considers a special case of tree structures for the hidden variables, the set of edges is a set with disconnected edges, no edge is connected to any other. Then we show that the same structure is preserved for the observable variables.

**Lemma 17.** Let $F = (\mathcal{V}, \mathcal{E})$ be a forest with $|\mathcal{V}| = p$ and $|\mathcal{E}| = p/2 \in \mathbb{N}$ such no edge is connected to any other edge. Assume that $X_i \in \{-1, +1\}$ and $\mathbb{E}[X_i] = 0$ for all $i \in [1, \ldots, p]$. If $Y$ is the output of the BSC channel (in the hidden model) with distribution $p_\dagger(y)$, then $p_\dagger(y)$ also factorizes with respect to $F$.

**Proof.** The pair variables $(Y_i, Y_j)$ for $(i, j) \in \mathcal{E}$ are independent because of the disconnected edges of the hidden layer. The latter directly gives the factorization as

$$p_\dagger(y) = \prod_{(i,j) \in \mathcal{E}} p_\dagger(y_i, y_j) = \prod_{i \in \mathcal{V}} p_\dagger(y_i) \prod_{(i,j) \in \mathcal{E}} \frac{p_\dagger(y_i, y_j)}{p_\dagger(y_i)p_\dagger(y_j)},$$

(3.206)

because $|\mathcal{V}| = p$, $|\mathcal{E}| = p/2$ and the marginal distributions are uniform. \qed
Chapter 4

Optimal Non-Parametric Structure Learning for Markov Trees

In this Chapter we consider the problem of learning hidden tree structures for general hidden trees and general noise models over discrete node alphabets. Our contribution centers on a rigorous characterization of a fundamental model-dependent statistic which we call the information threshold (denoted as $I^o$). We show that $I^o$ quantifies not only the sample complexity of the CL algorithm in particular, but also the complexity of the tree-structure learning problem in general. For the problem of learning tree-structures from noiseless data, the information threshold has been already appeared in fundamental prior work in the area (e.g., [93] and [94]). Precisely, in Tan et al. [93, Theorem 6], it is shown that if the information threshold is strictly positive, then the probability of incorrect structure recovery by the CL algorithm decays exponentially with respect to the number of samples. Additionally, the information threshold appears in the expression of the approximate error exponent [93, Theorem 8]. However, no explicit connection exists in the literature between the sample complexity of the CL algorithm (either finite or asymptotic) and a fixed positive value of the information threshold. For instance in [7, 93] the approximation of the error exponent is consistent with the true error exponent when $I^o \to 0$. Most importantly, based on prior works no conclusions can be derived for the rate of the sample complexity with respect to $I^o$ when the latter is bounded away from zero or when the data are noisy. Although it is true that the number of samples required for successful recovery scales logarithmically with respect to the number of nodes $p$ (in the asymptotic sense), the sample complexity can vary significantly for different tree-structured distributions or noise models (for fixed $p$), and fixed probability of failure $\delta$. The discussion above naturally raises the following basic
question: *Is there any essential statistic (e.g., the information threshold) that determines the sample complexity of the hidden tree-structure learning problem?* Here, we show that, indeed, the information threshold constitutes such a representative statistic; in fact, it provides explicit *necessary and sufficient* conditions on sample complexity, by effectively and completely summarizing the difficulty of the hidden tree-structure learning problem.

In many applications, the underlying physical or artificial phenomenon may be well-modeled by a (tree-structured) MRF, but the data acquisition device or sensor may itself introduce noise. We wish to understand how sensitive the performance of an algorithm is to noisy inputs. There are two recent works that study the impact of corrupted observations on both binary and Gaussian models [70, 67]. Goel et al. [70] extend the Interaction Screening Objective [95] for the case of Ising models, while our prior work [67] analyzes the performance of the CL algorithm for both noisy Ising and Gaussian models. In particular, this work showed the consistency of the CL algorithm under the assumption of independent and identically distributed noise.

Based on this prior work, it is tempting to think that one can successfully perform structure recovery by running the standard CL algorithm *directly* on noisy data. This is common in practice, because the distribution of the noise may be unknown. However, even for simplest models with binary observations, the CL algorithm with noisy input data may not be consistent (and divergence is guaranteed) without considering a pre-processing procedure. Figure 4.1 shows a simple binary-valued example of a 3-variable tree with non-identically distributed noise; even for this simple example, structure learning from raw data can be infeasible. However, if we denoise the observations first (a form of data pre-processing), CL will return the true structure with high probability.

This simple example raises another natural question: *Under which circumstances (e.g., using some form of pre-processing) can the CL algorithm be successful when used directly on noisy data?* We resolve this question by defining and analyzing the properties of the *noisy information threshold* (as termed herein, and denoted as $I_{\hat{o}}$), which, as it turns out, characterizes the finite sample complexity of the CL algorithm on hidden tree-structured MRFs.
4.1 Model and Problem Statement

First, we provide a complete description of our model including definitions, properties, and assumptions on the underlying distributions.

4.1.1 Tree-Structured and Hidden Tree-Structured Models

Similarly to the previous Chapters, we consider graphical models over \( p \) nodes with variables \( \{X_1, X_2, \ldots, X_p\} \equiv \mathbf{X} \) and finite or countable alphabet \( \mathcal{X}^p \). We assume that the distribution \( p(\cdot) \) of \( \mathbf{X} \) is given by a tree-structured Markov Random Field, (MRF). Recall that any distribution \( p(\cdot) \) which is Markov with respect to a tree \( T = (\mathcal{V}, \mathcal{E}) \) factorizes as

\[
p(\mathbf{x}) = \prod_{i \in \mathcal{V}} p(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{p(x_i, x_j)}{p(x_i)p(x_j)}, \quad \mathbf{x} \in \mathcal{X} ;
\]

we call such distributions \( p(\cdot) \) tree-structured.

The noisy node variables \( \mathbf{Y} = \{Y_1, Y_2, \ldots, Y_p\} \) are generated by a randomized mapping (noisy channel) \( \mathcal{F} : \mathcal{X}^p \to \mathcal{Y}^p \). We restrict attention here to mappings that act on each component (not necessarily independently): \( \mathcal{F}(\cdot) = \{F_1(\cdot), F_2(\cdot), \ldots, F_p(\cdot)\} \) and each \( F_i(\cdot) : \mathcal{X} \to \mathcal{Y} \) so \( Y_i = F_i(X_i) \) for all \( i \in \mathcal{V} \). Let \( \mathbb{P}(Y_i = y_i | X_i = x_i) \) be the
transition kernel associated with the randomized mapping $F_i(\cdot)$, then the distribution of the output is given by $p(\cdot) = \sum_{x_i \in \mathcal{X}} p(Y_i = y_i|X_i = x_i)p(x_i)$ for $y_i \in \mathcal{Y}$. Note that while the distribution of $X$ is tree-structured, the distribution of $Y$ does not factorize according to any tree. In general the Markov random field of $Y$ is a complete graph, which makes learning the hidden structure non-trivial [67, 96].

Given $n$ i.i.d observations $X^{1:n} \sim p(\cdot)$, our goal is to learn the underlying structure $T$. To do this, we use a plug-in estimate of the mutual information $I(X_i; X_j)$ between pairs of variables. In similar fashion, when noise-corrupted observations $Y^{1:n}$ are available, we aim to learn the hidden tree structure $T$ of $X$, by estimating the mutual information $I(Y_i; Y_j)$ between pairs of observable variables. Unfortunately, the plug-in mutual information estimate $\hat{I}(X_i; X_j)$ may converge slowly to $I(X_i; X_j)$ in certain cases with countable alphabets [97, Corollary 5], [98]. To avoid such ill-conditioned cases, we make the following assumption.

**Assumption 3.** For some $c > 1$ there exist $c_1, c_2 > 0$ such that $c_1/k^c \leq p_i(k) \leq c_2/k^c$, for $k \in \mathcal{X}$, and $c_1/(k\ell)^c \leq p_{i,j}(k, \ell) \leq c_2/(k\ell)^c$, for $k, \ell \in \mathcal{X}$ and for all $i, j \in \mathcal{V}$. That is, the tuple $\{c, c_1, c_2\}$ satisfies the assumption for all marginal and pairwise joint distributions.

Assumption 3 holds trivially for finite (fixed) alphabets, where the constants $c_1$ and $c_2$ depend on the minimum probability and the size of alphabet. The next assumption guarantees there is a unique tree structure $T$ with exactly $p$ nodes.

**Assumption 4.** $T$ is connected; $I(X_i; X_j) > 0$ for all $i, j \in \mathcal{V}$ and the distribution $p(\cdot)$ of $X$ is not degenerate.

Assumption 4 guarantees convergence for the CL algorithm; $T^{CL} \to T$. For a fixed tree $T$, we use the notation $\mathcal{P}_T(c_1, c_2)$ to denote the set of tree-structured distributions which satisfy Assumption 3 and Assumption 4. In particular, we assume that $X \sim p(\cdot) \in \mathcal{P}_T(c_1, c_2)$. The set of all trees on $p$ nodes is denoted as $\mathcal{T}$, and we call $\mathcal{P}_T$ the set of all tree-structured distributions that factorize according to (4.1) for some $T \in \mathcal{T}$. 

4.1.2 The Classical Chow-Liu Algorithm

We consider the classical version of the algorithm [3], due to the non-parametric nature of our model. Specifically, the difference between Algorithm 1 and Algorithm 3 is that the first considers the estimated correlations as edge weights while the second the estimated mutual information. Given \( n \) i.i.d samples of the node variables, we first find the estimates of the pairwise joint distributions and then we evaluate the plug-in mutual information estimates. Finally, a Maximum Spanning Tree (MST) algorithm (for instance Kruskal’s or Prim’s algorithm) returns the estimated tree. For the rest of the Chapter, we refer to Algorithm 3 by explicitly mentioning the input data set; if \( \mathcal{D} = \mathbf{X}^{1:n} \), then the input consist of noiseless data and we consider \( \mathbf{Z} \equiv \mathbf{X} \) (see Algorithm 3), furthermore the estimated structure \( \mathbf{T}^{\text{estimate}} \) is denoted by \( \mathbf{T}^{\text{CL}} \). Equivalently, if \( \mathcal{D} = \mathbf{Y}^{1:n} \), then the input consists of noisy data, \( \mathbf{Z} \equiv \mathbf{Y} \), and the estimated structure \( \mathbf{T}^{\text{estimate}} \) is named as \( \mathbf{T}^{\text{CL}} \). We compute \( \mathbf{T}^{\text{CL}} \) and \( \mathbf{T}^{\text{CL}} \) by running the MST algorithm on the edge weights \( \{ \hat{I}(X_i; X_j) : i,j \in \mathcal{V} \} \) and \( \{ \hat{I}(Y_i; Y_j) : i,j \in \mathcal{V} \} \) respectively. Note that the estimates \( \mathbf{T}^{\text{CL}} \) and \( \mathbf{T}^{\text{CL}} \) depend on the value \( n \), but for brevity, we write \( \lim_{n \to \infty} \mathbf{T}^{\text{CL}} = \mathbf{T} \) and \( \lim_{n \to \infty} \mathbf{T}^{\text{CL}} = \mathbf{T} \) respectively. We continue by analyzing the event of incorrect reconstruction \( \mathbf{T}^{\text{CL}} \neq \mathbf{T} \) (or \( \mathbf{T}^{\text{CL}} \neq \mathbf{T} \)), which yields a sufficient condition for exact structure recovery.

**Proposition 3.** The estimated tree \( \mathbf{T}^{\text{CL}} \neq \mathbf{T} \) if and only if there exist two edges \( e \equiv (w, \bar{w}) \in \mathbf{T} \) and \( g \equiv (u, \bar{u}) \in \mathbf{T}^{\text{CL}} \) such that \( e \notin \mathbf{T}^{\text{CL}} \), \( g \notin \mathbf{T} \) and \( e \in \text{path}_T(u, \bar{u}) \). Then also \( g \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \bar{w}) \).

Intuitively, exact recovery fails when there is at least one edge in the original tree \( \mathbf{T} \) which does not appears in the estimated tree \( \mathbf{T}^{\text{CL}} \). We refer the reader to the proof of Proposition 3 by Bresler and Karzand [5, Section F, “Two trees lemma”, Lemma F1]. For sake of space, we define the set \( \mathcal{E} \mathcal{V}^2 \).

**Definition 2 (Feasibility set \( \mathcal{E} \mathcal{V}^2 \)).** Let \( e \equiv (w, \bar{w}) \in \mathcal{E}_T \) be an edge and \( u, \bar{u} \in \mathcal{V}_T \) be a pair of nodes such that \( e \in \text{path}_T(u, \bar{u}) \) and \( |\text{path}_T(u, \bar{u})| \geq 2 \). The set of all such
tuples \((e, u, \bar{u})\), is denoted as \(\mathcal{E}V^2\),

\[
\mathcal{E}V^2 \triangleq \{e, u, \bar{u} \in \mathcal{E}_T \times \mathcal{V}_T \times \mathcal{V}_T : e \in \text{path}_T(u, \bar{u}) \text{ and } |\text{path}_T(u, \bar{u})| \geq 2\}. \tag{4.2}
\]

For the rest of the Chapter the pair of nodes \(w, \bar{w}\) denotes the edge \(e \equiv (w, \bar{w}) \in \mathcal{E}_T\). The error characterization of CL algorithm is expressed as if \(T^{\text{CL}}_{\dagger} \neq T\) then there exists \(((w, \bar{w}), u, \bar{u}) \in \mathcal{E}V^2\) such that\(^1\)

\[
\hat{I}(Y_w; Y_{\bar{w}}) \leq \hat{I}(Y_u; Y_{\bar{u}}).
\]

By negating the above statement, we get that if \(\hat{I}(Y_w; Y_{\bar{w}}) > \hat{I}(Y_u; Y_{\bar{u}})\) for all \(((w, \bar{w}), u, \bar{u}) \in \mathcal{E}V^2\) then \(T^{\text{CL}}_{\dagger} = T\). The latter yields the sufficient condition for accurate structure estimation.

**Sufficient condition for exact structure recovery:** For exact structure recovery we need \(\hat{I}(Y_w; Y_{\bar{w}}) > \hat{I}(Y_u; Y_{\bar{u}})\) for all \(((w, \bar{w}), u, \bar{u}) \in \mathcal{E}V^2\), or equivalently

\[
I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) > \hat{I}(Y_u; Y_{\bar{u}}) - I(Y_u; Y_{\bar{u}}) - \hat{I}(Y_w; Y_{\bar{w}}) + I(Y_w; Y_{\bar{w}}). \tag{4.3}
\]

The latter allows us to derive a sufficient condition based on the error estimation of the mutual information.

**Proposition 4.** If

\[
\left| \hat{I}(Y_\ell; Y_{\ell'}) - I(Y_\ell; Y_{\ell'}) \right| < \frac{1}{2} \min_{(e, u, \bar{u}) \in \mathcal{E}V^2} \left\{ I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \right\}, \tag{4.4}
\]

for all \(\ell, \ell' \in \mathcal{V}\) then \(T^{\text{CL}}_{\dagger} = T\).

In fact (4.4) implies (4.3) and (4.3) implies \(T^{\text{CL}}_{\dagger} = T\). Inequality (4.4) shows that if the error of mutual information estimates is less than a threshold statistic then exact structured recovery is guaranteed.

\(^1\)The event \(\{\hat{I}(Y_w; Y_{\bar{w}}) = \hat{I}(Y_u; Y_{\bar{u}})\}\) has non zero probability for certain cases, in this situation the MST arbitrarily chooses one of the edges \((w, \bar{w}), (u, \bar{u})\). The choice of \((u, \bar{u})\) yields the error event \(T^{\text{CL}}_{\dagger} \neq T\).
Algorithm 3 Chow-Liu (CL)

Require: Data set \( D = Z \in \mathbb{Z}^{[V] \times n} \)

1: \( \hat{p}_{i,j}(\ell, m) = \frac{1}{n} \sum_{k=1}^{n} 1\{Z_{i,k} = \ell, Z_{j,k} = m\}, \forall i, j \in V \)

2: \( \hat{I}(Z_i; Z_j) = \sum_{\ell, m} \hat{p}_{i,j}(\ell, m) \log_2 \frac{\hat{p}_{i,j}(\ell, m)}{\hat{p}_i(\ell) \hat{p}_j(m)} \)

3: \( T_{\text{estimate}} \leftarrow \text{MST} \left( \{ \hat{I}(Z_i; Z_j) : i, j \in V \} \right) \)

4.1.3 Information Threshold and Properties

We now define a new quantity for tree structured distributions, which we call the information threshold. As well will see shortly, our sample complexity bounds for exact structure recovery via the CL algorithm depend on the distribution only through the information threshold, \( I_o^\dagger \). We first define \( I_o^\dagger \) and then show how it affects the difficulty of the structure estimation problem.

Definition 3 (Information Threshold (IT)). Let \( e \equiv (w, \bar{w}) \in E_T \) be an edge and \( u, \bar{u} \in V_T \) be a pair of nodes such that \( e \in \text{path}_T(u, \bar{u}) \). The information threshold associated with the model \( p_\dagger(\cdot) \) (see Section 4.1.1) is defined as

\[
I_o^\dagger \equiv \frac{1}{2} \min_{(e, u, \bar{u}) \in E \setminus \{u, \bar{u}\}} \left( I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \right).
\] (4.4)

The minimization in (4.4) and (4.5) is with respect to the feasible set \( E \setminus \{u, \bar{u}\} \) of the hidden tree structure \( T \) of \( X \). Note that the distribution of \( Y \) does not factorize according to any tree [96].

When the data are noiseless, the gap between mutual informations that defines the information threshold will change. If the errors of the mutual information estimates of the noiseless variables satisfy the condition

\[
\left| \hat{I}(X_\ell; X_{\bar{\ell}}) - I(X_\ell; X_{\bar{\ell}}) \right| < \frac{1}{2} \min_{(e, u, \bar{u}) \in E \setminus \{u, \bar{u}\}} \left( I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right),
\] (4.6)

for all \( \ell, \bar{\ell} \in V \) then \( T_{\text{CL}} = T \), and (4.6) is derived similarly to (4.4). The definition of the noiseless information threshold naturally results from the previous condition.
Definition 4 (Noisy IT). The noisy information threshold is defined as

\[
\mathcal{I}^o \triangleq \frac{1}{2} \min_{(e, u, \bar{u}) \in \mathcal{E}\mathcal{V}^2} (I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})).
\] (4.7)

Similarly to the noisy case, if

\[
\left| \hat{I}(X_{\ell}; X_{\ell}) - I(X_{\ell}; X_{\ell}) \right| < \mathcal{I}^o \quad \forall \ell, \ell' \in \mathcal{V},
\] (4.8)

then \( T = T_{CL} \). The data processing inequality [78] shows that \( \mathcal{I}^o \geq 0 \). Also, Assumption 4 guarantees that \( \mathcal{I}^o > 0 \).

Proposition 5 (Positivity). If Assumption 4 holds then \( \mathcal{I}^o > 0 \).

Note that under the reasonable assumption that the values of \( I(X_i, X_j) \) for \( (i, j) \in \mathcal{E} \) are constant relative to \( p \) [6], \( \mathcal{I}^o \) does not depend on \( p \). The latter holds because of the locality property of \( \mathcal{I}^o \).

Proposition 6 (Locality). Assume that Assumption 4 holds. Let \((e^*, u^*, \bar{u}^*) \in \mathcal{E}\mathcal{V}^2 \) be a tuple such that

\[
(e^*, u^*, \bar{u}^*) = \arg\min_{((w, \bar{w}), u, \bar{u}) \in \mathcal{E}\mathcal{V}^2} I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}),
\] (4.9)

then \( u^* \equiv w^* \) or \( u^* \equiv \bar{w}^* \) and \( \bar{u} \in \mathcal{N}_T(w) \) or \( \bar{u} \in \mathcal{N}_T(\bar{w}) \).

We prove Propositions 5 and 6 in Section 4.7, Appendix.

4.2 Recovering the Structure from Noisy Data

We start by developing a finite sample complexity bound for exact structure learning with high probability, when noisy data are available. The structure learning condition in (4.8), combined with results on concentration of mutual information estimators [97], yields the following result.

Theorem 12. Assume that \( X \sim p(\cdot) \in \mathcal{P}_T \). Assume that noisy data \( Y \sim p_{\dagger}(\cdot) \) are generated by a randomized set of mappings \( \mathcal{F} = \{F_i(X_i) = Y_i : i \in [p]\} \), and \( p_{\dagger}(\cdot) \)
satisfies the Assumption 3 for some \( c \geq 2, c_1 > c_2 > 0 \) Fix a number \( \delta \in (0,1) \). If the number of samples of \( Y \) satisfies the inequalities

\[
\frac{n}{\log^2 n} \geq \max\{288 \log \left( \frac{p}{\delta} \right), 4C^2\} \tag{4.10}
\]

for a constant \( C > 0 \), then Algorithm 3 with input \( D = Y^{1:n} \) returns \( T_{\text{CL}}^\dagger = T \) with probability at least \( 1 - \delta \).

Before proceeding with the proof, we would like to provide some remarks. First, the constant \( C \) depends on the values of constants \( c, c_1, c_2 \) which are defined in Assumption 3. Specifically, \( C = 3c_2 + c^{-1} \int_{c_1}^\infty u^{1/c} - 2 \log \left( e^{u/c_1} + 1/c_1 \right) \). The derivation of \( C \) has been given by Antos and Kontoyiannis [97, Theorem 7]. Additionally, note that \( n/\log^2 n = \Omega \left( n^\epsilon \right) \) for any fixed \( \epsilon \in (0,1) \). Therefore, the required number of samples \( n \) with respect to \( p \) and \( \delta \) and for fixed \( I_\dagger^p \) scale as \( O(\log^{1+\zeta(p/\delta)}) \), for any choice of \( \zeta > 0 \), whereas, for fixed \( p \) and \( \delta \), the complexity is of the order of \( I_\dagger^p \) is \( O\left( (I_\dagger^p)^{-2(1+\zeta)} \right) \), for any \( \zeta > 0 \). The proof of Theorem 12 now follows.

Proof of Theorem 12. To calculate the probability of the exact structure recovery we use a concentration inequality quantifying the rate of convergence of entropy estimators from Antos and Kontoyiannis [97]. In particular, they show ([97, Corollary 1]) how the plug-in entropy estimator \( \hat{H}_n \) (say) is distributed around its mean \( \mathbb{E}[\hat{H}_n] \): For every \( n \in \mathbb{N} \) and \( \epsilon > 0 \),

\[
\mathbb{P}\left[ \left| \hat{H}_n - \mathbb{E}[\hat{H}_n] \right| > \epsilon \right] \leq 2e^{-n\epsilon^2/2\log^2 n}. \tag{4.11}
\]

The plug-in entropy estimator is biased and, actually, \( H \geq \mathbb{E}[\hat{H}_n] \). Under their Assumption 3, in Theorem 7 [97] they characterize the bias as follows. For \( c \in [2, \infty) \) (which is the case of interest in this proof),

\[
H - \mathbb{E}[\hat{H}_n] = O\left( n^{-1/2} \log n \right). \tag{4.12}
\]
and for \( c \in (1, 2) \),

\[
H - \mathbb{E}[\hat{H}_n] = \Theta\left(n^{\frac{1-c}{c}}\right). 
\]

(4.13)

Then, for \( \epsilon > Cn^{-1/2} \log n \), it is true that

\[
\mathbb{P}\left[|\hat{H}_n - H| > \epsilon\right] \\
= \mathbb{P}\left[|\hat{H}_n - \mathbb{E}[\hat{H}_n] + \mathbb{E}[\hat{H}_n] - H| > \epsilon\right] \\
\leq \mathbb{P}\left[|\hat{H}_n - \mathbb{E}[\hat{H}_n]| + |\mathbb{E}[\hat{H}_n] - H| > \epsilon\right] \\
= \mathbb{P}\left[|\hat{H}_n - \mathbb{E}[\hat{H}_n]| > \epsilon - |\mathbb{E}[\hat{H}_n] - H|\right] \\
\leq \mathbb{P}\left[|\hat{H}_n - \mathbb{E}[\hat{H}_n]| > \epsilon - \frac{C\log n}{\sqrt{n}}\right] \\
\leq 2e^{-n\left(\epsilon - C\frac{\log n}{\sqrt{n}}\right)^2/2\log^2 n}, 
\]

(4.14)

where the last inequality comes from (4.11) and (4.12). Notice that for non-trivial bounds we need the condition \( \epsilon > Cn^{-1/2} \log n \). Further, \( \epsilon \) is free parameter and we choose \( \epsilon = \frac{I^o_{\perp}}{3} \), driven by property (4.6). This requires than \( n \) has to be sufficiently large, such that the following inequality holds

\[
I^o_{\perp} > 3Cn^{-1/2} \log n. 
\]

(4.15)

Our goal is to find an upper on the probability of the event \( \{|\hat{I}(Y_i; Y_{\bar{i}}) - I(Y_i; Y_{\bar{i}})| > I^o_{\perp}\} \). By combining the above and applying the property \( I(X; Y) = H(X) + H(Y) - H(X, Y) \)
we have

\[
P \left[ \left| \hat{I} \left( Y_\ell; Y_{\bar{\ell}} \right) - I \left( Y_\ell; Y_{\bar{\ell}} \right) \right| > I_0^* \right] \\
= P \left[ \left| \hat{H} \left( Y_\ell \right) + \hat{H} \left( Y_{\bar{\ell}} \right) - \hat{H} \left( Y_\ell, Y_{\bar{\ell}} \right) - H \left( Y_\ell \right) + H \left( Y_{\bar{\ell}} \right) \right| > I_0^* \right] \\
\leq P \left[ \left| \hat{H} \left( Y_\ell \right) - H \left( Y_\ell \right) \right| + \left| \hat{H} \left( Y_{\bar{\ell}} \right) - H \left( Y_{\bar{\ell}} \right) \right| + \left| H \left( Y_\ell, Y_{\bar{\ell}} \right) - \hat{H} \left( Y_\ell, Y_{\bar{\ell}} \right) \right| > I_0^* \right] \\
\leq P \left[ \left| \hat{H} \left( Y_\ell \right) - H \left( Y_\ell \right) \right| > \frac{I_0^*}{3} \right] + \left\{ \left| \hat{H} \left( Y_{\bar{\ell}} \right) - H \left( Y_{\bar{\ell}} \right) \right| > \frac{I_0^*}{3} \right\} + \left\{ \left| H \left( Y_\ell, Y_{\bar{\ell}} \right) - \hat{H} \left( Y_\ell, Y_{\bar{\ell}} \right) \right| > \frac{I_0^*}{3} \right\} \\
\leq 6e^{-n \left( \frac{I_0^*}{3} - C \frac{\log n}{\sqrt{n}} \right)^2 / 2 \log^2 n},
\]  

(4.16)

and the last inequality is a consequence of (4.14). To guarantee that the condition in (4.6) holds, we apply the union bound on the events \( \left\{ \left| \hat{I} \left( Y_\ell; Y_{\bar{\ell}} \right) - I \left( Y_\ell; Y_{\bar{\ell}} \right) \right| > I_0^* \right\} \), for all \( \ell, \tilde{\ell} \in V \). Since there are \((p^2)\) pairs we define

\[
\delta \triangleq \left( \frac{p}{2} \right) 6e^{-n \left( \frac{I_0^*}{3} - C \frac{\log n}{\sqrt{n}} \right)^2 / 2 \log^2 n}.
\]  

(4.17)

To conclude, for some fixed \( \delta \in (0, 1) \) if

\[
\frac{n}{\log^2 n} \geq \frac{2 \log \left( \frac{6 \left( \frac{p}{2} \right)}{\delta} \right)}{\left( \frac{I_0^*}{3} - C \frac{\log n}{\sqrt{n}} \right)^2} \quad \text{and} \quad I_0^* > 3C \frac{\log n}{\sqrt{n}},
\]  

(4.18)

then the probability of exact recovery is at least \( 1 - \delta \). The latter combined with the inequalities \( 8 \log \left( p/\delta \right) > 2 \log \left( 6 \left( \frac{p}{2} \right) /\delta \right), p \geq 3 \) gives

\[
\frac{n}{\log^2 n} \geq \frac{72 \log \left( \frac{p}{\delta} \right)}{\left( I_0^* - C \frac{\log n}{\sqrt{n}} \right)^2} \quad \text{and} \quad I_0^* > C \frac{\log n}{\sqrt{n}},
\]  

(4.19)

thus it sufficient to have

\[
\frac{n}{\log^2 n} \geq \frac{72 \log \left( \frac{p}{\delta} \right)}{\left( I_0^* - C \frac{\log n}{\sqrt{n}} \right)^2} \quad \text{and} \quad I_0^* \geq \frac{C \log n}{\sqrt{n}},
\]  

(4.20)

\[
\frac{n}{\log^2 n} \geq \frac{288 \log \left( \frac{p}{\delta} \right)}{\left( I_0^* \right)^2} \quad \text{and} \quad \frac{n}{\log^2 n} \geq \left( \frac{2C}{I_0^*} \right)^2.
\]  

(4.21)
The latter gives the statement of the Theorem.

Theorem 12 characterizes the sample complexity for models with either countable or finite alphabets. By restricting our setting to finite alphabets Assumption 3 is not required and we have the following result. The proof is virtually identical to that of Theorem 12, and is omitted.

**Theorem 13.** Assume that the random variable $Y$ (as defined in Theorem 12) has finite support. Fix a number $\delta \in (0, 1)$. There exists a constant $C > 0$, independent of $\delta$ such that, if the number of samples of $Y$ satisfies the inequalities

\[
\frac{n}{\log^2 n} \geq \frac{288 \log \left( \frac{\delta}{\delta} \right)}{(I^o)^2} \quad \text{and} \quad n \geq \left( \frac{2C}{I^o} \right)^2,
\]

then Algorithm 3 with input $D = Y^{1:n}$ returns $T^{CL}_1 = T$ with probability at least $1 - \delta$.

Lastly, the corresponding variation of Theorem 12 when Assumption 3 holds for $c \in (1, 2)$ (in the general case of countable alphabets) follows.

**Theorem 14.** Assume that $X \sim p(\cdot) \in \mathcal{P}_T$. Assume that noisy data $Y \sim p^*(\cdot)$ are generated by a randomized set of mappings $\mathcal{F} = \{F_i(X_i) = Y_i : i \in [p]\}$, and $p^*(\cdot)$ satisfies the Assumption 3 for some $c \in (1, 2), c_1 > c_2 > 0$. Fix $\delta \in (0, 1)$. There exists a constant $C > 0$ independent of $\delta$ such that, if $I^o > 0$ and the number of samples $n$ of $Y$ satisfies the inequalities

\[
\frac{n}{\log^2 n} \geq \frac{288 \log \left( \frac{\delta}{\delta} \right)}{(I^o)^2} \quad \text{and} \quad n \geq \left( \frac{2C}{I^o} \right)^2 \frac{\varepsilon}{c-1},
\]

then Algorithm 3 with input the noisy data $D = Y^{1:n}$ returns $T^{CL}_1 = T$ with probability at least $1 - \delta$.

As a consequence of our analysis we also derive the sample complexity bound for the noiseless case as well. Now the bound involves the (noiseless) information threshold $I^o$. 
Theorem 15. Assume that \( X \sim p(\cdot) \in \mathcal{P}_{T}(c_1, c_2) \) for some \( c \in [2, \infty) \). Fix a number \( \delta \in (0, 1) \). There exists a constant \( C > 0 \), independent of \( \delta \) such that, if the number of samples of \( X \) satisfies the inequalities

\[
\frac{n}{\log^2 n} \geq \frac{288 \log \left( \frac{\delta}{2} \right)}{(I^o)^2} \quad \text{and} \quad \frac{n}{\log^2 n} \geq \left( \frac{2C}{I^o} \right)^2
\]  

(4.24)

then Algorithm 3 with input \( D = X^{1:n} \) returns \( T_{CL} = T \) with probability at least \( 1 - \delta \). The relationship between \( C' \) and \( c', c'_1, c'_2 \) is identical to that between \( C \) and \( c, c_1, c_2 \) in Theorem 15.

The latter allows us to compare the sample complexity of the noiseless and noisy setting. Let \( n \) and \( n^\dagger \) denote the sufficient number of samples of \( X \) and \( Y \) respectively and consider \( p, \delta \) fixed, then the ratio \( n/n^\dagger \) is \( O\left((I^o/I^o)^{-2(1+\zeta)}\right) \) for all \( \zeta > 0 \). The latter shows how \( n^\dagger \) changes relative to \( n \) under the same probability of success for both settings (noiseless and noisy). The proof of Theorem 15 is similar to the proof of Theorem 12.

Proof of Theorem 12. The difference is introduced by the event in (4.8). That is, the error on the mutual information estimates should be less than the noiseless information threshold \( I^o \). Note that for the entropy estimates of \( X \), equations (4.11) up to (4.14) hold with possibly different constants \( c', c'_1, c'_2, C' \) (see Assumption 3). Here we consider the case where \( c' \geq 2 \) (the case \( c' \in (1, 2) \) is similar, see also the proof of Theorem 15), and the corresponding bound on the estimation error \( \epsilon \) has to be at most \( I^o/3 \). Thus, (4.15) becomes

\[
I^o > 3C' n^{-1/2} \log n,
\]  

(4.25)

and (4.16) now is written as

\[
P \left[ \left| \hat{I} (X_{\ell}; X_{\tilde{\ell}}) - I (X_{\ell}; X_{\tilde{\ell}}) \right| > I^o \right] \leq 6 e^{-n \left( \frac{I^o}{3} - C' \log n \right)^2 / 2 \log^2 n}. \]  

(4.26)

Finally, by applying union bound over the pairs \( \ell, \tilde{\ell} \in \mathcal{V} \) we derive the statement of
Theorem 12, by following the equivalent steps of (4.17) and (4.18). The latter completes
the proof. □

4.3 Converse: Information Threshold as a Fundamental Quantity

In this section we provide the statement of the converse of our main result Theorem
12. We define the class of hidden tree-structured models with bounded (from below)
information threshold of the hidden layer by an absolute positive constant and bounded
absolute information threshold for the observable layer by a fixed $\Delta > 0$ as\(^2\)

$$\mathcal{C}_\Delta^T \doteq \{ M : |\Theta^o_{i,M}| \geq \Delta > 0 \}. \quad (4.27)$$

The next result provides a lower bound for the necessary number of samples for tree-
structure learning from noisy observations.

**Theorem 16.** There exist absolute constants $C', \tilde{C}, \epsilon_0 > 0$ such that for any $\epsilon \in (0, \epsilon_0)$,
$\Delta_\epsilon \doteq \tilde{C}\epsilon$ and for any estimator $\Phi : Y^{1:n} \rightarrow \mathcal{T}$, if $n < C'/\Delta_\epsilon^2$, then the worst-case
probability of incorrect structure recovery over all hidden tree-structured models in $\mathcal{C}_\Delta^T$ is at least 1/2. In other words, it is true that if $n < C'/\Delta_\epsilon^2$, then

$$\inf_{\Phi : Y^{1:n} \rightarrow \mathcal{T}} \sup_{M \in \mathcal{C}_\Delta^T} P(\Phi(Y^{1:n}_{T_M}) \neq T_M) \geq \frac{1}{2}. \quad (4.28)$$

Additionally, the supremum is attained.

**Proof of Theorem 16.** We define the set of hidden tree-structured models namely $M_0$,
$M_1$ and $M_2$ (see Figure 4.2) as follows: The hidden layer of $M_0$ is $X_1 - X_2 - X_3$, and
$\mathcal{E}_{M_0} = \{(1,2), (2,3)\}$, the hidden layer of $M_1$ is $X_2 - X_1 - X_3$ and $\mathcal{E}_{M_1} = \{(1,2), (1,3)\}$,
the hidden layer of $M_2$ is $X_2 - X_3 - X_1$ and $\mathcal{E}_{M_2} = \{(2,3), (1,3)\}$. Further for all
models, $X_i \in \{-1, +1\}$ for $i = 1, 2, 3$ and $\mathbb{E}[X_i X_j] \doteq c \in (0, 1)$ for any $(i,j) \in \mathcal{E}_{M_k}$ and
$k \in \{0,1,2\}$. Note that the information threshold $\Theta^o_{i,M}$ of the hidden layer is strictly

---

\(^2\)The information threshold $\Theta^o_{i,M}$ of the observable layer can be either positive or negative for each
model $M$ in the class $\mathcal{C}^T$. 

positive for any $c \in (0, 1)$, since

$$\Gamma^o_{M_0} = I_{M_0}(X_1; X_2) - I_{M_0}(X_1; X_3) = \frac{1}{2} \log \left( \frac{(1 - E[X_1X_2])^{1-E[X_1X_2]} (1 + E[X_1X_2])^{1+E[X_1X_2]}}{(1 - E[X_1X_3])^{1-E[X_1X_3]} (1 + E[X_1X_3])^{1+E[X_1X_3]}} \right)$$

(4.29)

and by contraction $\Gamma^o_{M_1} = I_{M_1}(X_1; X_2) - I_{M_1}(X_2; X_3) \equiv \Gamma^o_{M_0}$, $\Gamma^o_{M_2} = I_{M_2}(X_2; X_3) - I_{M_2}(X_1; X_2) \equiv \Gamma^o_{M_0}$. The observable data are generated by binary symmetric channels with cross-over probabilities $1 - q$, $1 - q'$ and $q = 0.75 + \epsilon$ and $q' = 0.75 - \epsilon$. The noise variables $N_i \in \{-1, +1\}$ are multiplicative binary, independent from each other and independent from the hidden variables and generate the corresponding observable as $Y_i = N_i \times X_i$, see Figure 4.2. Specifically, for the model $M_0$ it is that $P(N_1 = +1) = P(N_2 = +1) = q$ and $P(N_3 = +1) = q'$, for the model $M_1$ $P(N'_1 = +1) = P(N'_3 = +1) = q$ and $P(N'_2 = +1) = q'$, and for the model $M_2$ $P(N_2 = +1) = P(N_3 = +1) = q$ and $P(N_1 = +1) = q'.3$

\footnote{The construction of the hidden models $M_0, M_1, M_2$ is similar to \cite[Lemma 7.1]{99}. The crucial difference is that our construction involves hidden layers with $P' > 0$, while the in Lemma 7.1 by \cite{99} the information threshold of the hidden is zero and the construction is inappropriate for the structure learning problem.}
The distribution of any pair \( Y_i, Y_j \in \{-1, +1\} \) is

\[
p(y_i, y_j) = \frac{1 + \mathbb{E}[Y_i Y_j | y_i y_j]}{4}, \quad y_i, y_j \in \{-1, +1\}.
\] (4.31)

We find the second order moments \( \mathbb{E}[Y_i Y_j] \) of the model \( M_0 \) as follows,

\[
\mathbb{E}[N_1] = \mathbb{E}[N_2] = +1 \times \left( \frac{3}{4} + \epsilon \right) - 1 \times \left( \frac{1}{4} - \epsilon \right) = \frac{1}{2} + 2\epsilon,
\] (4.32)

\[
\mathbb{E}[N_3] = +1 \times \left( \frac{3}{4} - \epsilon \right) - 1 \times \left( \frac{1}{4} + \epsilon \right) = \frac{1}{2} - 2\epsilon.
\] (4.33)

\[
\mathbb{E}[Y_1 Y_2] = \mathbb{E}[N_1 X_1 N_2 X_2] = \mathbb{E}[N_1] \mathbb{E}[N_2] \mathbb{E}[X_1 X_2] = c \left( \frac{1}{2} + 2\epsilon \right)^2,
\] (4.34)

\[
\mathbb{E}[Y_2 Y_3] = \mathbb{E}[N_2 X_2 N_3 X_3] = c \left( \frac{1}{4} - 4\epsilon^2 \right),
\] (4.35)

\[
\mathbb{E}[Y_1 Y_3] = \mathbb{E}[N_1 X_1 N_3 X_3] = c^2 \left( \frac{1}{4} - 4\epsilon^2 \right).
\] (4.36)

We combine the above and we choose \( c = 1 - \epsilon \) to evaluate the information threshold \( I_{t,M_0}^o \) as

\[
I_{t,M_0}^o = I(Y_1; Y_2) - I(Y_1; Y_3) = H(Y_1; Y_3) - H(Y_1; Y_2)
\]

\[
= - \left( \frac{1 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{2} \right) \log \left( \frac{1 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
\]

\[
- \left( \frac{1 - c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{2} \right) \log \left( \frac{1 - c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
\]

\[
+ \left( \frac{1 + c \left( \frac{1}{2} + 2\epsilon \right)^2}{2} \right) \log \left( \frac{1 + c \left( \frac{1}{2} + 2\epsilon \right)^2}{4} \right)
\]

\[
+ \left( \frac{1 - c \left( \frac{1}{2} + 2\epsilon \right)^2}{2} \right) \log \left( \frac{1 - c \left( \frac{1}{2} + 2\epsilon \right)^2}{4} \right)
\]

\[
= \frac{9}{8} \log \left( \frac{5}{3} \right) \epsilon + O(\epsilon^2).
\] (4.37)
Similarly for the model $M_1$,

$$
\mathbb{E}[N_1] = \mathbb{E}[N_3] = +1 \times \left( \frac{3}{4} + \epsilon \right) - 1 \times \left( \frac{1}{4} - \epsilon \right) = \frac{1}{2} + 2\epsilon,
$$

(4.39)

$$
\mathbb{E}[N_2] = +1 \times \left( \frac{3}{4} - \epsilon \right) - 1 \times \left( \frac{1}{4} + \epsilon \right) = \frac{1}{2} - 2\epsilon,
$$

(4.40)

$$
\mathbb{E}[Y_1 Y_2] = \mathbb{E}[N_1 X_1 N_2 X_2] = \mathbb{E}[N_1] \mathbb{E}[N_2] \mathbb{E}[X_1 X_2] = c \left( \frac{1}{4} - 4\epsilon^2 \right),
$$

(4.41)

$$
\mathbb{E}[Y_2 Y_3] = \mathbb{E}[N_2 X_2 N_3 X_3] = c^2 \left( \frac{1}{4} - 4\epsilon^2 \right),
$$

(4.42)

$$
\mathbb{E}[Y_1 Y_3] = \mathbb{E}[N_1 X_1 N_3 X_3] = c \left( \frac{1}{2} + 2\epsilon \right)^2.
$$

(4.43)

The information threshold $I_{o^{\uparrow},M_1}$ is

$$
I_{o^{\uparrow},M_1} = I(Y_1; Y_2) - I(Y_2; Y_3) = H(Y_2; Y_3) - H(Y_1; Y_2)
$$

$$
= -2 \left( \frac{1 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right) \log \left( \frac{1 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
$$

$$
- 2 \left( \frac{1 - c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right) \log \left( \frac{1 - c^2 \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
$$

$$
+ 2 \left( \frac{1 + c \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right) \log \left( \frac{1 + c \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
$$

$$
+ 2 \left( \frac{1 - c \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right) \log \left( \frac{1 - c \left( \frac{1}{4} - 4\epsilon^2 \right)}{4} \right)
$$

$$
= \frac{\log \frac{5}{3}}{8} \epsilon + O(\epsilon^2).
$$

(4.44)

As a consequence, (4.38) and (4.44) give $I_{o^{\uparrow},M_0} = \Theta(\epsilon)$ and $I_{o^{\uparrow},M_1} = \Theta(\epsilon)$. In contrast, (4.30) gives $I_o = (\epsilon - 1)^2/2 + O((\epsilon - 1)^4)$. Further, the joint distributions of the models $M_0, M_1$ are given by

$$
p_{M_0}(y_1, y_2, y_3) = \frac{1}{8} \left[ 1 + \mathbb{E}[Y_1 Y_2] y_1 y_2 + \mathbb{E}[Y_2 Y_3] y_2 y_3 + \mathbb{E}[Y_1 Y_3] y_1 y_3 \right]
$$

$$
= \frac{1}{8} \left[ 1 + c \left( \frac{1}{2} + 2\epsilon \right)^2 y_1 y_2 + c \left( \frac{1}{4} - 4\epsilon^2 \right) y_2 y_3 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right) y_1 y_3 \right],
$$

(4.45)
and
\[
p_{M_1}(y_1, y_2, y_3) = \frac{1}{8} \left[ 1 + \mathbb{E}[Y_1 Y_2|y_1 y_2] + \mathbb{E}[Y_2 Y_3|y_2 y_3] + \mathbb{E}[Y_1 Y_3|y_1 y_3] \right]
= \frac{1}{8} \left[ 1 + c \left( \frac{1}{4} - 4\epsilon^2 \right) y_1 y_2 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right) y_2 y_3 + c \left( \frac{1}{2} + 2\epsilon \right)^2 y_1 y_3 \right],
\]
(4.46)

for \(y_1, y_2, y_3 \in \{-1, +1\}\). We use (4.45) and (4.46) to evaluate the probabilities
\[
p_{M_0}(+1, +1, +1) = p_{M_0}(-1, -1, -1) = \frac{7}{32} + \frac{1}{8} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_0}(-1, +1, -1) = p_{M_0}(+1, -1, +1) = \frac{3}{32} - \frac{3}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_0}(-1, +1, +1) = p_{M_0}(+1, -1, -1) = \frac{7}{32} + \frac{5}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_0}(-1, -1, +1) = p_{M_0}(+1, +1, -1) = \frac{3}{32} - \frac{5}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_0}(+1, +1, +1) = p_{M_0}(-1, -1, -1) = \frac{7}{32} + \frac{1}{8} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_0}(-1, -1, +1) = p_{M_0}(+1, -1, -1) = \frac{3}{32} - \frac{3}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_1}(+1, +1, +1) = p_{M_1}(-1, -1, -1) = \frac{7}{32} + \frac{1}{8} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_1}(-1, +1, -1) = p_{M_1}(+1, -1, +1) = \frac{3}{32} - \frac{5}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
\[
p_{M_1}(-1, -1, +1) = p_{M_1}(+1, +1, -1) = \frac{3}{32} - \frac{3}{16} \epsilon + \mathcal{O}(\epsilon^2)
\]
By the definition of the KL divergence
\[
D_{KL}(p_{M_1}||p_{M_0}) = \frac{73}{24} \epsilon^2 + \frac{1129}{72} \epsilon^3 + \mathcal{O}(\epsilon^4).
\]
(4.47)

Similarly for the model \(M_2\) we have
\[
I_{o,M_2} = I_{M_2}(Y_2; Y_3) - I_{M_2}(Y_1; Y_2) = \frac{9}{8} \log \left( \frac{5}{3} \right) \epsilon + \mathcal{O}(\epsilon^2) = \Theta(\epsilon),
\]
(4.48)

the distribution is
\[
p_{M_2}(y_1, y_2, y_3) = \frac{1}{8} \left[ 1 + c \left( \frac{1}{2} + 2\epsilon \right)^2 y_3 y_2 + c^2 \left( \frac{1}{4} - 4\epsilon^2 \right) y_2 y_3 + c \left( \frac{1}{4} - 4\epsilon^2 \right) y_1 y_3 \right],
\]
(4.49)
\[ y_1, y_2, y_3 \in \{-1, +1\}, \text{ and we find} \]
\[ D_{\text{KL}}(p_{M_2} || p_{M_0}) = \frac{73}{24} \epsilon^2 + \frac{1057}{72} \epsilon^3 + \mathcal{O}(\epsilon^4). \] \hspace{1cm} (4.50)

Next, we use Fano’s inequality (see [85, Corollary 2.6]). Fix \( L = 2 \) and let \( P_{M_0}, P_{M_1}, P_{M_2} \) denote the probability laws of \( Y \) under models \( M_0, M_1 \) and \( M_2 \) respectively, and consider \( n \) i.i.d. observations \( Y^{1:n} \). If

\[ n < \frac{\alpha \log L}{\frac{1}{L+1} \sum_{j=1}^{L} D_{\text{KL}}(P_{M_j} || P_{M_0})}, \] \hspace{1cm} (4.51)

with \( \alpha \in (0, 1) \) then

\[ \inf_{\Phi} \max_{0 \leq j \leq L} P_{M_j} \left( \Phi(Y^{1:n}) \neq T_{M_j} \right) \geq \frac{\log(L+1) - \log(2)}{\log(L)} - \alpha, \] \hspace{1cm} (4.52)

where the infimum is relative to all estimators (statistical tests) \( \Phi : \mathcal{Y}^n \to \{0, 1, \ldots, L\} \).

Specifically, for \( L = 2 \) and

\[ \tilde{\alpha} = \frac{\log(L+1) - \log(2)}{\log(L)} - \frac{1}{2} \in (0, 1), \] \hspace{1cm} (4.53)

if

\[ n < \frac{\tilde{\alpha} \log 2}{\frac{2}{3} \max \{ D_{\text{KL}}(P_{M_1} || P_{M_0}), D_{\text{KL}}(P_{M_2} || P_{M_0}) \}}, \] \hspace{1cm} (4.54)

then

\[ \inf_{\Phi} \sup_{M \in \mathcal{C}_{\Delta}} P(\Phi(Y^{1:n}) \neq T_M) \geq \inf_{\Phi} \max_{0 \leq j \leq L} P_{M_j} \left( \Phi(Y^{1:n}) \neq T_{M_j} \right) \geq \frac{1}{2}. \] \hspace{1cm} (4.55)

Finally, \( D_{\text{KL}}(p_{M_1} || p_{M_0}) = \mathcal{O}(\epsilon^2), D_{\text{KL}}(p_{M_2} || p_{M_0}) = \mathcal{O}(\epsilon^2) \) from (4.47), (4.50) and \( I_{1,M_0} = \Theta(\epsilon), I_{1,M_1} = \Theta(\epsilon), I_{1,M_2} = \Theta(\epsilon) \) from (4.38), (4.44), (4.48). Thus there exists \( \epsilon_0 \in (0, 1) \) and \( \tilde{C} > 0 \) such that for any \( \epsilon \in (0, \epsilon_0) \), \( \min \{ I_{1,M_0}, I_{1,M_1}, I_{1,M_2} \} \geq \tilde{C} \epsilon \triangleq \Delta(\epsilon) \).

The statement of the Theorem follows.
4.4 Recovering the Structure from Noisy Data

Recall that we defined the noisy information threshold $I_o^\dagger$ in (4.7), that appears in the condition of exact structure recovery of CL algorithm (4.6). The latter shows that structure recovery is feasible if $I_o^\dagger > 0$. To be more precise, for every $I_o^\dagger > 0$ there exists $N \in \mathbb{N}$ such that if $n > N$ then (4.6) holds with high probability. In fact, in Theorem 12 the condition $I_o^\dagger > 0$ is necessary. Under the assumption of $I_o > 0$ (see Assumption 4), it is not guaranteed that $I_o^\dagger > 0$. In fact, if $I_o^\dagger < 0$ then the CL is not consistent in the sense that

$$\mathbb{P}\left( \lim_{n \to \infty} E_{T_{\text{CL}}}^{\dagger}(D=Y_{1:n}) \neq E_T \right) = 1,$$

(4.56)

(see also Section 4.4.1). On the other hand if $I_o^\dagger = 0$, then ties are broken arbitrarily and the probability of missing an edge does not decrease as $n$ increases.

In what follows, we provide sufficient conditions which ensure that $I_o^\dagger > 0$; in particular, whenever $I_o^\dagger < 0$, we show that pre-processing on the the noisy data can be used as an extra step to overcome the inconsistency of the CL algorithm on the original noisy data. As a result, the output of CL algorithm (with input the processed data) will converge to the original tree $T$ of the hidden layer.

4.4.1 IOP and the Importance of Pre-Processing

Prior to running our estimation algorithm, we would like to know if recovering the hidden tree structure is possible given the noise model. Unfortunately, the definition of $I_o^\dagger$ involves the structure $T$ that we want to estimate. We first find a condition under which we can guarantee $I_o^\dagger > 0$ without any knowledge of $T$ beforehand. We state this in terms of the randomized mapping $F$ (see Section 4.1.1). The next property guarantees that $I_o > 0$ if and only if $I_o^\dagger > 0$, for any $T \in T$.

**Definition 5 (Information Order Preservation (IOP)).** Let $X \in X^p$ and $Y \in Y^p$ be random vectors such that $Y = F(X)$, with $F$ defined as in Section 4.1.1. We say that the randomized mapping $F$ is information order-preserving (IOP) relative to $X$ if and
only if, for every tuple \([(k, l), (m, r)]\) ∈ \(\mathcal{V}^2 \times \mathcal{V}^2\), such that \(k \neq l\), \(m \neq r\), \(\{k, l\} \neq \{m, r\}\), it is true that

\[
I(X_k; X_\ell) < I(X_m; X_r) \iff I(F_k(X_k); F_\ell(X_\ell)) < I(F_m(X_m); F_r(X_r)).
\]  
(4.57)

Note that if the randomized mapping \(F\) is IOP and \(I_o > 0\) then we can guarantee that \(I_o^* > 0\), without any knowledge of the hidden structure \(T\). On the other hand, if the IOP does not hold it is still possible that \(I_o^* > 0\). Thus, in general the condition (4.6) characterizes the error event of the CL algorithm (see Section 4.4.2).

In certain cases, by knowing only the noise distribution we can find if IOP holds. To make this clear we provide an example of an erasure channel for \(M\)-ary alphabets. Specifically, for each hidden variable \(X_i \in [M]\) and \(i \in \mathcal{V}\), the corresponding observable \(Y_i \in [M + 1]\) is either \(Y_i = X_i\) with probability \(1 - q_i\) or \(Y_i = 0\) (an erasure occurs) with probability \(q_i\). Further, it is true that \(I(Y_i; Y_j) = (1 - q_i)(1 - q_j)I(X_i; X_j)\) for all \(i, j \in \mathcal{V}\), and \(q_i, q_j \in [0, 1]\) (see Appendix, Section 4.7.4.1). As a consequence, for certain values of the pairs \((q_i, q_j)\) (4.57) holds, for instance consider the case \(q_i = q_j\). Additionally, we show that the IOP is satisfied for several important cases of hidden graphical models, as we will discuss later in Section 4.4.2. Next we show that CL algorithm is consistent when the IOP holds, while if IOP does not hold then an appropriate processing on the noisy data can help to overcome this issue.

### 4.4.1.1 Convergence of CL Algorithm under the IOP

If \(F\) is information order-preserving (IOP), then Algorithm 3 with input \(\mathcal{D} = \mathcal{Y}^{1:n}\) converges to the true tree \((T_{\text{CL}}^f \rightarrow T)\) (also see Theorem 12). For IOP randomized mappings, the following ordering with respect to all pairs of nodes

\[
I(X_{i_1}; X_{j_1}) < I(X_{i_2}; X_{j_2}) < \cdots < I(X_{i_r}; X_{j_r}),
\]
(4.58)
for $r = \binom{p}{2}$, $i_s, j_s \in \mathcal{V}$, $s \in [r]$, remains unchanged for the observable node variables $\mathbf{Y}$. That is, (4.58) through (4.57), implies that

$$I(Y_{i_1}; Y_{j_1}) < I(Y_{i_2}; Y_{j_2}) < \cdots < I(Y_{i_r}; Y_{j_r}),$$

(4.59)

for $r = \binom{p}{2}$, $i_s, j_s \in \mathcal{V}$, $s \in [r]$. Therefore, the maximum spanning tree algorithm with input weights in (4.58) returns the same tree structure $T$ if the input weights are changed to the corresponding mutual information values of $\mathbf{Y}$ in (4.59). Therefore, for sufficient large number of samples the order is also preserved for the estimates $\hat{I}(Y_i; Y_j)$, which ensures that Algorithm 3 is consistent; $T_{\hat{I}} \rightarrow T$ for $\mathcal{D} = \mathbf{Y}^{1:n}$. Thus, the information order preservation property is sufficient to guarantee convergence of the CL Algorithm, that is, $I_{\hat{I}} > 0$.

4.4.1.2  **Enforcing the IOP through pre-processing**

Although we are primarily interested in conditions ensuring a positive $I_{\hat{I}}$, a negative $I_{\hat{I}}$ is also informative. As we will show later, if we can find an appropriate pre-processing $\mathbf{Y} \rightarrow \mathbf{Z}$ such that $I_{\hat{I};Z} > 0$, Theorem 12 applies with $I_{\hat{I};Z} > 0$ as the new threshold, and $I_{\hat{I};Z}$ is defined by replacing $\mathbf{Y}$ variable with $\mathbf{Z}$ in (4.7). To further explain this, we demonstrate the pre-processing procedure by enforcing the IOP in the example that follows. Although we consider a 3-node hidden Markov chain with binary random variables for brevity, the same technique can be applied on larger trees with $p$ nodes. Additionally, the technique can be extended in certain models with larger alphabets.

To illustrate the case of $I_{\hat{I}} < 0$ and the effect of the pre-processing (on the input
data-set $\mathcal{D}$) we present a simple example of hidden tree-structured models for which Algorithm 3 succeeds with high probability, only if an appropriate pre-processing is being applied. Consider the smallest tree structure; let a three node Markov chain $X_1 - X_2 - X_3$ be the hidden layer and $Y_1, Y_2, Y_3$ be the variables of the observable layer (Figure 4.3). Specifically, $X_i, N_i \in \{-1, +1\}$ and $Y_i = N_i X_i$, for all $i \in \{1, 2, 3\}$. The distribution of the noise is

$$P(N_1 = -1) = 1 - P(N_1 = +1) = q' \in [0, 1/2),$$  \hspace{1cm} (4.60)

and for $i \in \{2, 3\}$

$$P(N_i = -1) = 1 - P(N_i = +1) = q \in (0, 1/2).$$  \hspace{1cm} (4.61)

Thus the noisy variables are generated by a BSC($q_i$) and the noise is not identically
distributed, because \( q \neq q' \), see Figure 4.3. Recall that the tree structure is \( T = (\mathcal{V}, \mathcal{E}) \), \( \mathcal{E} = \{(1, 2), (2, 3)\} \) and \( \mathcal{V} = \{1, 2, 3\} \). Further \( |\mathbb{E}[X_1X_2]|, |\mathbb{E}[X_3X_2]| \in (0, 1) \), and without loss of generality assume that

\[
|\mathbb{E}[X_1X_2]| \leq |\mathbb{E}[X_3X_2]|. \tag{4.62}
\]

The Markov property \([5, 69]\) of \( X_1, X_2, X_3 \in \{-1, +1\} \) gives

\[
\mathbb{E}[X_1X_3] = \mathbb{E}[X_1X_2]\mathbb{E}[X_3X_2]. \tag{4.63}
\]

The definition of \( I^\circ \) in (4.5) together with (4.62),(4.63) give

\[
I^\circ = I(X_1; X_2) - I(X_1; X_3), \tag{4.64}
\]

because \( I(X_i; X_j) \) is increasing with respect to \( |\mathbb{E}[X_iX_j]| \) (see Appendix, (4.84)). Additionally, it is true that \( I(X_2; X_3) \geq I(X_1; X_2) > I(X_1; X_3) \) because \( |\mathbb{E}[X_1X_3]| < |\mathbb{E}[X_1X_2]| \leq |\mathbb{E}[X_3X_2]| < 1 \). The latter guarantees that the Chow-Liu algorithm, with input \( D = X^{1:n} \) returns \( T_{\text{CL}} = T \) for \( n \to \infty \) with probability 1. However, structure recovery is not guaranteed from noisy data. In fact \( \mathbb{E}[N_i] = 1 - 2q_i, N_i \perp \! \! \! \perp X_i \) for all \( i \in \{1, 2, 3\} \) and

\[
\mathbb{E}[Y_1Y_2] = \mathbb{E}[N_1X_1N_2X_2] = (1 - 2q')(1 - 2q)\mathbb{E}[X_1X_2], \tag{4.65}
\]

\[
\mathbb{E}[Y_2Y_3] = \mathbb{E}[N_2X_2N_3X_3] = (1 - 2q)^2\mathbb{E}[X_2X_3]
\]

\[
\mathbb{E}[Y_1Y_3] = \mathbb{E}[N_1X_1N_3X_3] = (1 - 2q)\mathbb{E}[X_1X_3] = (1 - 2q')(1 - 2q)\mathbb{E}[X_1X_2]\mathbb{E}[X_2X_3]. \tag{4.66}
\]

Under a possible error event, CL algorithm replaces \( \{2, 3\} \) by \( \{1, 3\} \) and returns \( \mathcal{E}_{\text{CL}} = \{(1, 2), \{1, 3\}\} \). The latter occurs even for zero estimation error; \( I(Y_1; Y_2) = \hat{I}(Y_1; Y_2) \),
\[ I(Y_1; Y_3) = \hat{I}(Y_1; Y_3), \quad I(Y_2; Y_3) = \hat{I}(Y_2; Y_3) \] if and only if

\[ I(Y_1; Y_2) > I(Y_1; Y_3) > I(Y_2; Y_3), \tag{4.67} \]

and the last implies that \( I^*_1 = I(Y_2; Y_3) - I(Y_1; Y_3) < 0 \) (by the definition (4.7) of \( I^*_1 \)). Specifically, for \( n \to \infty \) the edge \((2, 3)\) will be replaced by \((1, 3)\) (w.p. 1) if and only if \( I(Y_1; Y_3) > I(Y_2; Y_3) \). The latter gives the locus of \( q, q' \) that yields an error in the structure estimation process as follows,

\[ I(Y_1; Y_3) > I(Y_2; Y_3) \iff |E[X_1X_2]| > |E[X_2X_3]| \iff (1 - 2q')(1 - 2q)|E[X_1]||E[X_2X_3]| > (1 - 2q)^2|E[X_2X_3]| \iff |E[X_1X_2]| > \frac{1 - 2q}{1 - 2q'}. \tag{4.68} \]

Further, note that (4.65) and (4.66) and guarantee that the left hand-side (first inequality) of (4.67) holds, thus

\[ |E[X_1X_2]| > \frac{1 - 2q}{1 - 2q'} \]

\[ \iff \tag{4.65),(4.66} I(Y_1; Y_2) > I(Y_1; Y_3) > I(Y_2; Y_3) \]

\[ \iff \tag{4.65),(4.66} P \left( \lim_{n \to \infty} E_{T^*_{CL}(D=Y)} \neq E_T \right) = 1. \tag{4.69} \]

As a consequence the structure learning from raw data is not guaranteed (with high probability) for all the values of pairs \((q, q')\) that satisfy (4.69) (even for \( n \to \infty \)). To overcome this we have enforce the IOP (Definition 5) by considering the following pre-processing for each sample of the variables \( Y_1, Y_2, Y_3 \)

\[ Z_1 \triangleq Y_1/(1 - 2q') \]
\[ Z_2 \triangleq Y_2/(1 - 2q) \]
\[ Z_3 \triangleq Y_3/(1 - 2q), \tag{4.70} \]

then it is true that \( I(Z_2; Z_3) \geq I(Z_1; Z_2) > I(Z_1; Z_3) \). The latter guarantees that IOP
holds and

\[ \mathbb{P}\left( \lim_{n \to \infty} \mathcal{E}^{\text{CL}}_{T}(D \overset{\text{z}}{\rightarrow} \mathbb{Z}^{1:n}) = \mathcal{E}_{T} \right) = 1, \]  

(4.71)

for any \( q \in (0, 1/2) \), \( q' \in [0, 1/2) \). Simulations on synthetic data verify our analysis (see Figure 4.4). As a final observation, even for \( q' = 0 \), (4.68) may hold as \( \mathbb{E}[X_1X_2] > 1 - 2q \)
and structure learning is infeasible by running the Chow-Liu on raw data. However if \( q = q' \neq 0 \) then (4.68) does not hold and structure learning from raw data is feasible which yields to the counterintuitive fact that introducing more noise (turning \( q' \) from 0 to \( q > 0 \)) can make structure learning feasible in certain scenarios.

### 4.4.2 Examples and Applications on Specific Models

Theorem 12 can be applied on a wide class models that satisfies the general Assumptions 3 and 4. To illustrate the effect of noise on the structure learning complexity, we consider two classical noisy channels in the hidden model: the \( M \)-ary erasure and the (generalized) symmetric channel. We show that a simple comparison of \( I^o \) and \( I^o_{\dagger} \) determines the impact of noise on the sample complexity, and we present the relationship between the two. As we explained, the CL algorithm can fail when noise is not i.i.d.. In the next examples, we present conditions for accurate structure estimation for certain model scenarios, while the number of nodes \( p \) is arbitrary and the noise is non-identically distributed.

**Example 1: \( M \)-ary Erasure Channel.** Assume that the randomized mappings \( F_i(\cdot) \) “erase” each variable independently with probability \( q \), so for all \( i \in [p] \), we have \( Y_i = X_i \) with probability \( 1 - q \) and \( Y_i = M + 1 \) (an erasure) with probability \( q \). Then \( I(Y_i; Y_j) = (1 - q)^2 I(X_i; X_j) \) for all \( i, j \in V \) (see Appendix 4.7.4.1). Therefore, \( I^o_{\dagger} = (1 - q)^2 I^o \leq I^o \) and the information order preservation (4.57) holds for any \( q \in [0, 1) \). The latter guarantees that if \( I^o > 0 \) then \( I^o_{\dagger} > 0 \). Given the values of \( p, \delta, q \) and \( I^o \), Theorem 12 provides the sample complexity for exact structure recovery from noisy observations. For fixed values of \( p \) and \( \delta \), the ratio of sufficient number of samples in the noiseless and noisy settings is \( O((1 - q)^{4(1 + \zeta)}) \), for any \( \zeta > 0 \). In contrast, consider the
scenario where the erasure probability is not the same for every node (non-identically
distributed noise). Each \( F_i \) erases the \( i^{\text{th}} \) node value with probability \( q_i \in [0, 1) \), so \( I(Y_i; Y_j) = (1 - q_i)(1 - q_j)I(X_i; X_j) \) for all \( i, j \in V \), and the condition (4.6) shows that Algorithm 3 with input \( D = Y^{1:n} \) converges; \( T_{CL}^* \to T \), if for all tuples \((w, w), u, \bar{u} \in \mathcal{E}V^2\)

\[
\frac{(1 - q_w)(1 - q_w)}{(1 - q_u)(1 - q_{\bar{u}})} > \frac{I(X_u; X_{\bar{u}})}{I(X_w; X_{\bar{w}})}.
\]

(4.72)

Define \( RI \triangleq \max_{(w, \bar{w}), u, \bar{u} \in \mathcal{E}V^2} I(X_u; X_{\bar{u}})/I(X_w; X_{\bar{w}}) \). Inequality (4.72) provides the following simplified sufficient condition for convergence of CL Algorithm with input noisy data; \( T_{CL}^* \to T \) if for all \( i, j \in V \)

\[
\frac{1 - q_i}{1 - q_j} \in \left( RI^{1/2}, RI^{-1/2} \right).
\]

(4.73)

Given the values \( q_i, q_j \) and \( RI \) or estimates of them, (4.73) provides rule for testing if structure estimation is possible directly from raw noisy data. If direct structure estimation is not possible then we can consider a pre-processing that enforces the IOP similarly to the example of Section 4.4.1.2. We continue by providing a feasibility condition for the binary symmetric channel with non-identically distributed noise. The identically distributed noise case was studied in our prior work [67].

**Example 2: Binary Symmetric Channel.** Assume that the hidden variables \( X \in \{-1, +1\}^p \) follow a tree-structured Ising model, the noisy observable variables \( Y \in \{-1, +1\}^p \) are generated by setting \( Y_i = X_i \) with probability \( 1 - q_i \) and \( Y_i = -X_i \) with probability \( q_i \in [0, 1/2) \), \( i \in [p] \), and \( q_i \) is the probability of a value to change sign. Structure recovery directly from noisy observations \( Y^{1:n} \) is feasible, because the information order preservation property holds when the crossover probability values \( q_i \) are equal for all \( i \in V \) [67]. To guarantee convergence of Algorithm 1 (\( D = Y^{1:n} \)) for more general cases, the condition \( I_f^* > 0 \) should hold. As an example, consider the case of non identically distributed noise, that is, the probability \( q_i \) of a flip may differ for each node. In this case, the information order preservation property does not hold for all the possible sequences \( \{q_1, q_2, \ldots, q_p\} \in [0, 1/2]^p \) and we would like to know for which values of \( q_i \), \( i \in [p] \), Algorithm 3 with \( D = Y^{1:n} \) learns the hidden structure. The
Figure 4.5: Left, center: estimating the probability $P(T_{CL} \neq T)$ for different values of $I^o$ and $n$ through $5 \times 10^3$ independent runs with noiseless data. Right: estimating the probability $P(T_{CL}^{*} \neq T)$ through $10^4$ samples and $5 \times 10^3$ independent runs with noisy data.

Condition (4.6) implies that if for all $i, j \in V$

$$\frac{(1 - 2q_i)}{(1 - 2q_j)} \in \left( \max_{(i,j) \in E_T} |\mathbb{E}[X_i X_j]|, \frac{1}{\max_{(i,j) \in E_T} |\mathbb{E}[X_i X_j]|} \right)$$

then $T \rightarrow T_{CL}^*$ and the proof of (4.74) is given in Appendix 4.7.4.2. The condition (4.74) provides a testing rule for tree-structure estimation directly from raw noisy data, given the model parameters $q_i, q_j$ and $\max_{(i,j) \in E_T}$, or estimates of these parameters. Note that for identical noise $q_i = q_j = q$ it is true that $(1 - 2q_i)/(1 - 2q_j) = 1$ for all $i, j \in V$, thus (4.74) is always satisfied because $|\mathbb{E}[X_i X_j]| \in (0, 1)$, and structure learning is always feasible for this regime. If the condition (4.74) is not satisfied then structure recovery is still feasible by applying an appropriate pre-processing on the data $Y_{1:n}$.

The pre-processing procedure requires the values $q_i$ (or estimates of them) to be known, similarly to the example in Section 4.4.1.2. In contrast, Algorithm 3 does not require
any information related to the values \( q_i \), and its convergence is guaranteed under the condition \((4.74)\). In the next example we study an extension of the binary symmetric channel to alphabets of size \( M \).

**Example 3: Generalized Symmetric Channel.** We define the generalized symmetric channel as follows, assume \( X \sim \text{p}(\cdot) \in \mathcal{P}_T, X \in [M]^p \) and let \( Z_i \) for \( i \in [p] \) be i.i.d uniform random variables, such that \( \mathbb{P}(Z_i = k) = 1/M \), for all \( k \in [M] \). Also assume that \( Z \) and \( X \) are independent, then the \( i^{th} \) variable of the channel output \( Y \in [M]^p \) is defined as

\[
Y_i = F_i(X_i) = \begin{cases} 
X_i, & \text{with probability } 1 - q \\
Z_i, & \text{with probability } q,
\end{cases}
\]  

\[(4.75)\]

and \( q \in [0, 1) \). Note that the probability of a symbol to remain unchanged is given as \( \mathbb{P}(Y_i = X_i) = (1 - q) + q/M \), and for \( M = 2 \) it reduces to the binary symmetric channel. Theorem 12 can be directly applied for given values of \( p, \delta \) and \( I_o^\dagger \). However, closed form expressions of \( I_o^\dagger \) relative to \( I_o \) are unknown. As a consequence, an explicit comparison between the bounds of Theorems 15 and 12 is hard to evaluate in a closed-form expression for any \( q \in [0, 1) \). Nevertheless, we are able to derive approximations of the relationship of \( I_o \) and \( I_o^\dagger \) by considering sufficiently small \( q \). Lemma 22 (Section 4.7.4.3, Appendix) shows that in the small noise regime it is true that

\[
I_o^\dagger = (1 - q)^2 I_o - (1 - (1 - q)^2) \Delta_{KL} + \mathcal{O}(\epsilon^2),
\]

\[
\Delta_{KL} \triangleq KL(U||p(x^*_w, x^*_\bar{w})) - KL(U||p(x^*_u, x^*_\bar{u})),
\]

\[(4.76)\]

\( U \) is the uniform distribution on the alphabet \([M]^2\), the tuple \((w^*, \bar{w}^*), u^*, \bar{u}^*) \in \mathcal{E}^2 \) is the argument of the minimum in \((4.5)\), and \( \epsilon = (1 - (1 - q)^2) / M \). Whenever \( I_o^\dagger > 0 \) perfect reconstruction is possible and Theorem 12 provides the sufficient number of samples.
4.5 Simulations

To demonstrate the relationships between $\delta$, $n$, and $I^o$, $I^o_\dagger$, we estimate $T^{CL}$, $T^{CL}_\dagger$ and $\delta$ through $5 \times 10^3$ independent runs on tree-structured synthetic data, for different values of $n \in [10^3, 10^4]$ and $p = 10$. The variable nodes are binary and take values in the set $\{-1, +1\}$. Figure 4.5 (left) illustrates the relationship between the probability of incorrect reconstruction and $I^o$ and Figure 4.5 (center) the relationship between the log-probability of error as a function of the squared information threshold. We observe that probability of incorrect reconstruction decays exponentially with respect to $(I^o)^2$ as Theorem 15 suggests.

Lastly, Figure 4.5 (right) presents the effect of noise of a BSC for different values of the crossover $q$ and number of samples $n = 10^4$. Notice that the probability of incorrect reconstruction also decays exponentially with respect to $(I^o_\dagger)^2$, as Theorem 12 suggests, but with a significantly smaller rate than the noiseless case ($q = 0$).

4.6 Conclusion & Future Directions

In this chapter we showed how the information threshold characterizes the problem of recovering the structure of hidden tree-shaped graphical models. This quantity arises naturally in the error of Chow-Liu algorithm. As our main contribution, we introduced the first finite sample complexity bound on the performance of the CL algorithm for learning tree structured models from noisy data. More specifically, our sample complexity bounds show how the number of nodes $p$, the probability of failure $\delta$, and $I^o_\dagger$ are related for the problem of structure recovery. In fact, we provide matching upper and lower sample complexity bounds with respect to information threshold. The latter indicates that $I^o_\dagger$ is fundamental and CL algorithm is optimal up to a logarithmic factor $(\log p)$ for the hidden tree-structure learning. Our results demonstrate how noise affects the sample complexity of learning for a variety of standard models, including models for which the noise is not identically distributed.

Although we strictly consider the class of tree-structured models in our work, our approaches of Theorem 12 and Theorem 16 can be extended to the class of forests.
For that purpose, we should consider an generalization of \( I^p \) to forests and a modified version of CL the CLThres algorithm [6]. We leave this part for future as it is out of scope of this work.

Additionally, our approach is more generally applicable to the analysis of \( \delta \)-PAC Maximum Spanning Tree (MST) algorithms. At its root, our work shows how the error probability of MST algorithms (for example, Kruskal’s algorithm or Prim’s algorithm) behaves when edge weights are uncertain, i.e., when only (random) estimates of the true edge weights are known.

To conclude, the non-parametric graphical model setting presents interesting theoretical challenges that are connected with other statistical problems, out of the focus of this work. The relationship between \( I^p \) and \( I^p_\dagger \) is connected with open problems in information theory related to Strong Data Processing Inequalities [82, 66], for which tight characterizations are only known for a few channels. In our situation, a general analytical relationship may be similarly challenging. From a practical standpoint, we may wish to estimate the sample size needed to guarantee recovery with a pre-specified error probability. To do so would require knowing \( I^p_\dagger \) before collecting the full data; since \( I^p_\dagger \) depends on the noise model, we could find such a bound by considering a reasonable class of underlying models and taking the worst case. An interesting open question for future work is how to effectively estimate \( I^p_\dagger \) from (auxiliary) training data rather than relying on such a priori modeling assumptions. This may help design pre-processing methods that can make structure learning algorithms more robust against noise or adversarial attacks.

4.7 Appendix

We start by providing the proofs of Propositions 5 and 6.
4.7.1 Proof of Proposition 5: Positivity of the Noiseless Information Threshold

We consider the case $u^* \equiv w^*$ and $\bar{u} \in N_T(\bar{w})$, while the other three cases that are given by the locality property can be identically proved. The case $u^* \equiv w^*$ and $\bar{u} \in N_T(\bar{w})$ implies that $w^* - \bar{w}^* - \bar{u}^*$ is a subgraph of $T$. Assume for sake of contradiction that $I(X_w; X_{\bar{w}}) = I(X_w; X_{\bar{u}})$ then $I(X_w; X_{\bar{w}}|X_{\bar{u}}) = 0$. The latter implies that $w^* - \bar{u}^* - \bar{w}^*$ is also a subgraph of $T$ and it contradicts with the uniqueness of the structure (Assumption 4). □

4.7.2 Proof of Proposition 6: Locality of the Noiseless Information Threshold

Assume for sake of contradiction that $u^* \neq w^*$ and $u^* \neq \bar{w}^*$ or $\bar{u} \notin N_T(w)$ and $\bar{u} \notin N_T(\bar{w})$ and let $\nu$ be a node such that $\nu \in N_T(w) \cup N_T(\bar{w})$, then the data processing inequality [78] and Assumption 4 give

$$I(X_w; X_{\bar{w}}) - I(X_w; X_{\nu}) < I(X_w; X_{\bar{w}}) - I(X_{\bar{u}}; X_{\bar{u}})$$  \hspace{1cm} (4.77)

and

$$I(X_w; X_{\bar{w}}) - I(X_w; X_{\nu}) < I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}).$$  \hspace{1cm} (4.78)

The last two inequalities contradict the assumption (4.9). □

4.7.3 Fano’s Inequality

Lemma 18 (Fano’s Inequality, [85]). Fix $M \geq 2$ and let $\Theta$ be a family of models $\theta^0, \theta^1, \ldots, \theta^M$. Let $\mathbb{P}_\theta$ denote the probability law of $X$ under model $\theta^j$, and consider $n$ i.i.d. observations $X_1^n$. If

$$n < (1 - \delta) \frac{\log M}{\frac{1}{M+1} \sum_{j=1}^M D_{KL}(\mathbb{P}_\theta || \mathbb{P}_{\theta^j})},$$  \hspace{1cm} (4.79)
then it is true that

$$\inf_{\Phi} \max_{0 \leq j \leq M} \mathbb{P}_{\theta j} \left[ \Phi(X^{1:n}) \neq j \right] \geq \delta - \frac{1}{\log(M)}, \quad (4.80)$$

where the infimum is relative to all estimators (statistical tests) $\Phi : \mathcal{X}^{n \times n} \to \{0, 1, \ldots, M\}$.

### 4.7.4 Proofs of the Conditions in Section 4.4.2

In this subsection we provide the proof of the results and conditions that appear in Section 4.4.2.

#### 4.7.4.1 $M$-ary Erasure Channel

For the $M$-ary erasure channel, it is true that

$$I(Y_i; Y_j) = (1 - q_i)(1 - q_j)I(X_i; X_j). \quad (4.81)$$
for all $i,j \in V$ and $q_i, q_j \in [0,1)$. To prove this, we start by expanding the mutual information from the definition and pulling out the erasure event as follows

$$I(Y_i; Y_j)$$

$$= \sum_{y_i, y_j \in [M+1]^2} p^\dagger(y_i, y_j) \log \frac{p^\dagger(y_i, y_j)}{p^\dagger(y_i)p^\dagger(y_j)}$$

$$= \sum_{y_i, y_j \in [M]^2} p^\dagger(y_i, y_j) \log \frac{p^\dagger(y_i, M+1)}{p^\dagger(y_i)p^\dagger(y_j)}$$

$$+ \sum_{y_i \in [M]} p^\dagger(y_i, M+1) \log \frac{p^\dagger(y_i, M+1)}{p^\dagger(y_i)p^\dagger(y_j)}$$

$$+ \sum_{y_j \in [M]} p^\dagger(M+1, y_j) \log \frac{p^\dagger(M+1, y_j)}{p^\dagger(M+1)p^\dagger(y_j)}$$

$$+ p^\dagger(M+1, M+1) \log \frac{p^\dagger(M+1, M+1)}{p^\dagger(M+1)p^\dagger(M+1)}$$

$$= \sum_{y_i, y_j \in [M]^2} p^\dagger(y_i, y_j) \log \frac{p^\dagger(y_i ; y_j)}{p^\dagger(y_i)p^\dagger(y_j)}$$

$$= \sum_{x_i, x_j \in [M]^2} (1-q_i)(1-q_j)p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$$

$$= (1-q_i)(1-q_j)I(X_i; X_j) . \quad (4.82)$$

An erasure occurs independently on each node variable observable and independently with respect to the $X_i$, thus $p^\dagger(y_i, M+1) = p^\dagger(y_i)p^\dagger(M+1)$, for any $y_i \in [M+1]$ and $p^\dagger(M+1, y_j) = p^\dagger(M+1)p^\dagger(y_j)$ for any $y_j \in [M+1]$. The latter gives (4.82). □

### 4.7.4.2 Binary Symmetric Channel with Non-Identically Distributed Noise

Under the assumption of $I^\dagger > 0$, we wish to show that if

$$\frac{(1-2q_i)}{(1-2q_j)} \leq \left(\max_{(i,j)\in E_T} |E[X_iX_j]|, \frac{1}{\max_{(i,j)\in E_T} |E[X_iX_j]|} \right) , \quad (4.83)$$
for all \( i, j \in \mathcal{V} \) then \( I_{o}^{\ast} > 0 \). We start by finding the values of the sequence of crossover probabilities \( q_1, q_2, \ldots, q_k \in [0, 1/2) \) which guarantee that \( I_{o}^{\ast} > 0 \). The mutual information of two binary random variables \( Y_i, Y_j \in \{-1, +1\} \) (see [69]) is

\[
I(Y_i, Y_j) = \frac{1}{2} \log_2 \left( \left( 1 - \mathbb{E}[Y_i Y_j] \right)^{1 - \mathbb{E}[Y_i Y_j]} \left( 1 + \mathbb{E}[Y_i Y_j] \right)^{1 + \mathbb{E}[Y_i Y_j]} \right). \tag{4.84}
\]

The definition of \( I_{o}^{\ast} \) (Definition 4) and (4.84) give

\[
I_{o}^{\ast} = \frac{1}{2} \left\{ I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \right\} \tag{4.85}
\]

\[
= \frac{1}{2} \log_2 \left( \left( 1 - \mathbb{E}[Y_w Y_{\bar{w}}] \right)^{1 - \mathbb{E}[Y_w Y_{\bar{w}}]} \left( 1 + \mathbb{E}[Y_w Y_{\bar{w}}] \right)^{1 + \mathbb{E}[Y_w Y_{\bar{w}}]} \right) \left( \left( 1 - \mathbb{E}[Y_u Y_{\bar{u}}] \right)^{1 - \mathbb{E}[Y_u Y_{\bar{u}}]} \left( 1 + \mathbb{E}[Y_u Y_{\bar{u}}] \right)^{1 + \mathbb{E}[Y_u Y_{\bar{u}}]} \right).
\]

Define the function \( f(\cdot) \) as

\[
f(x) \triangleq (1 - x)^{1-x} (1 + x)^{1+x} \equiv f(|x|), \tag{4.86}
\]

then

\[
I_{o}^{\ast} = \frac{1}{2} \log_2 \frac{f(|Y_w Y_{\bar{w}}|)}{f(|Y_u Y_{\bar{u}}|)} \tag{4.87}
\]

and

\[
\mathbb{E}[Y_u Y_{\bar{u}}] = (1 - 2q_w)(1 - 2q_{\bar{w}}) \mathbb{E}[X_w X_{\bar{w}}], \tag{4.88}
\]

\[
\mathbb{E}[Y_u Y_{\bar{u}}] = (1 - 2q_w)(1 - 2q_{\bar{u}}) \mathbb{E}[X_w X_{\bar{w}}]
\times \prod_{(i,j) \in \text{path}_{T} \setminus \{w, \bar{w}\}} \mathbb{E}[X_i X_j] \tag{4.89}
\]

for the last equality we used the correlation decay property [5, 67]) and the fact that for \( \pm 1 \)-valued variables the binary symmetric channel can be consider as multiplicative binary noise [67]. Note that \( f(x) \) is increasing for \( x > 0 \). To guarantee that \( I_{o}^{\ast} > 0 \) we
need
\[
\frac{(1 - 2q_w)(1 - 2q_{\bar{w}})}{(1 - 2q_u)(1 - 2q_{\bar{u}})} > \prod_{(i,j) \in \text{path}_T(u,\bar{u}) \setminus (w,\bar{w})} |\mathbb{E}[X_iX_j]|. \tag{4.90}
\]
Recall that (4.90) should hold for all \((w,\bar{w}) \in \mathcal{E}\) and for all \(u,\bar{u} \in \mathcal{V}\) such that \((w,\bar{w}) \in \text{path}_T(u,\bar{u})\).

In addition,
\[
\prod_{(i,j) \in \text{path}_T(u,\bar{u}) \setminus (w,\bar{w})} |\mathbb{E}[X_iX_j]| \leq \left( \max_{(i,j) \in \mathcal{E}_T} |\mathbb{E}[X_iX_j]| \right)^{|\text{path}_T(u,\bar{u})|-1}.
\]

The last two inequalities give the sufficient condition
\[
\frac{(1 - 2q_w)(1 - 2q_{\bar{w}})}{(1 - 2q_u)(1 - 2q_{\bar{u}})} > \left( \max_{(i,j) \in \mathcal{E}_T} |\mathbb{E}[X_iX_j]| \right)^{|\text{path}_T(u,\bar{u})|-1}.
\]

As a consequence, if
\[
\frac{(1 - 2q_i)}{(1 - 2q_j)} \leq \left( \max_{(i,j) \in \mathcal{E}_T} |\mathbb{E}[X_iX_j]|, \frac{1}{\max_{(i,j) \in \mathcal{E}_T} |\mathbb{E}[X_iX_j]|} \right),
\]
for all \(i, j \in \mathcal{V}\) then \(I^*_q > 0\). Note that for the case of i.i.d. noise \(q_i = q_j\) for all \(i, j \in \mathcal{V}\) the inequality always holds because \(\max_{(i,j) \in \mathcal{E}_T} |\mathbb{E}[X_iX_j]| \in (0, 1)\). \(\square\)

### 4.7.4.3 \(M\)-ary Symmetric Channel

**Lemma 19.** Let \(A, B \in [L]\) be two discrete random variables, such that \(A \sim p_A(\cdot)\), \(B \sim p_B(\cdot)\) and \(H(A) < H(B)\). Assume \(A', B' \in [L]\) and \(A' \sim p_{A'}(\cdot)\) and \(B' \sim p_{B'}(\cdot)\) and \(q \in [0, 1/2)\) such that

\[
p_{A'}(\ell) = (1 - q)^2 p_A(\ell) + \frac{1 - (1 - q)^2}{L}, \quad \text{for all } \ell \in [L],
\]
\[
p_{B'}(\ell) = (1 - q)^2 p_B(\ell) + \frac{1 - (1 - q)^2}{L}, \quad \text{for all } \ell \in [L].
\]

Then \(H(A') < H(B')\) for sufficiently small values of \(q > 0\).
Proof. Define $\epsilon \equiv (1 - (1 - q)^2) / L$, so $0 < \epsilon < q$ for any $L \geq 2$.

\[ p_{A'}(\ell) \log_2 p_{A'}(\ell) \]
\[ = (1 - q)^2 p_A(\ell) + \epsilon \log_2 \left( (1 - q)^2 p_A(\ell) + \epsilon \right) \]
\[ = (1 - q)^2 p_A(\ell) \log_2 \left( (1 - q)^2 p_A(\ell) + \epsilon \right) + \epsilon \log_2 \left( (1 - q)^2 p_A(\ell) + \epsilon \right) \]
\[ = (1 - q)^2 p_A(\ell) \log_2 \left( (1 - q)^2 p_A(\ell) + \epsilon \right) + \epsilon \log_2 \left( (1 - q)^2 p_A(\ell) \right) + O(\epsilon^2). \tag{4.91} \]

To derive (4.91) recall that $\log(1 + x) = x + O(x^2)$ for $x < 1$, and set $x = \epsilon / (1 - q)^2 p_A(\ell)$ then the latter gives

\[ p_{A'}(\ell) \log_2 p_{A'}(\ell) \]
\[ = (1 - q)^2 p_A(\ell) \log_2 \left( (1 - q)^2 p_A(\ell) \right) \]
\[ + \epsilon \left( 1 + \log_2 \left( (1 - q)^2 p_A(\ell) \right) \right) + O(\epsilon^2). \tag{4.92} \]

Also,

\[ p_{B'}(\ell) \log_2 p_{B'}(\ell) \]
\[ = (1 - q)^2 p_B(\ell) \log_2 \left( (1 - q)^2 p_B(\ell) \right) \]
\[ + \epsilon \left( 1 + \log_2 \left( (1 - q)^2 p_B(\ell) \right) \right) + O(\epsilon^2). \tag{4.93} \]

Expanding both sides of the inequality $H(A) < H(B)$ we obtain the following:

\[ - \sum_{\ell=1}^{L} p_A(\ell) \log_2 p_A(\ell) < - \sum_{\ell} p_B(\ell) \log_2 p_B(\ell) \implies \]
\[ - \sum_{\ell=1}^{L} (1 - q)^2 p_A(\ell) \log_2 \left( (1 - q)^2 p_A(\ell) \right) \]
\[ < - \sum_{\ell} (1 - q)^2 p_B(\ell) \log_2 \left( (1 - q)^2 p_B(\ell) \right), \]
then for sufficiently small \( q \) and for any \( L > 2, p_A(\cdot), p_B(\cdot) \) there exist \( \epsilon > 0 \) such that

\[
-\sum_{\ell=1}^{L} \left[ (1-q)^2 p_A(\ell) \log_2 \left( (1-q)^2 p_A(\ell) \right) \right.
+ \epsilon \left( 1 + \log_2 \left( (1-q)^2 p_A(\ell) \right) \right) + \mathcal{O}(\epsilon^2) \bigg] \\
< -\sum_{\ell=1}^{L} \left[ (1-q)^2 p_B(\ell) \log_2 \left( (1-q)^2 p_B(\ell) \right) \right.
+ \epsilon \left( 1 + \log_2 \left( (1-q)^2 p_B(\ell) \right) \right) + \mathcal{O}(\epsilon^2) \bigg].
\]

This together with (4.92) and (4.93) give \( H(A') < H(B') \). \qed

We consider the extension an extension of the binary symmetric channel to alphabets of size \( M \) as follows. Assume \( X \sim P_T(c_1, c_2), X \in [M] \) and let \( Z_i \) for \( i \in [p] \) be i.i.d uniform random variables, \( P(Z_i = k) = 1/M, \) for all \( k \in [M] \). Also assume that \( Z \) and \( X \) are independent, then the noisy output variable \( Y \in [M]^p \) of the channel is defined for \( q \in [0, 1) \) as

\[
Y_i = F_i(X_i) = \begin{cases} 
X_i, & \text{with probability } 1-q \\
Z_i, & \text{with probability } q 
\end{cases} \] (4.94)

**Lemma 20.** The distribution of the two output variables \( Y_i, Y_j \) of the \( M \)-ary symmetric channel can be expressed as

\[
P(Y_i = y_i, Y_j = y_j) \\
= (1-q)^2 P(X_i = y_i, X_j = y_j) + \frac{1-(1-q)^2}{M^2}.
\]
Proof. This is a straightforward calculation

\[ p^\dagger(y_i, y_j) = \mathbb{P}(Y_i = y_i, Y_j = y_j) \]
\[ = (1 - q)^2 \mathbb{P}(X_i = y_i, X_j = y_j) + q(1 - q) \mathbb{P}(Z_i = y_i, X_j = y_j) + q(1 - q) \mathbb{P}(X_i = y_i, Z_j = y_j) + q^2 \mathbb{P}(Z_i = y_i, Z_j = y_j) \]
\[ = (1 - q)^2 \mathbb{P}(X_i = y_i, X_j = y_j) + q^2 \left( \frac{1 - (1 - q)^2}{M^2} \right), \]

and we are done. \( \square \)

**Lemma 21.** Consider \( X_k, X_\ell, X_m, X_r \) as four distinct inputs variables of the \( M \)-ary symmetric channel (defined in Section 4.4.2) with corresponding outputs \( Y_k, Y_\ell, Y_m, Y_r \). If the crossover probability \( q \) is sufficiently small and \( I(X_k; X_\ell) < I(X_m; X_r) \) then \( I(Y_k; Y_\ell) < I(Y_m; Y_r) \).

**Proof.** Note that the assumption of uniform marginal distributions for all four variables \( X_k, X_\ell, X_m, X_r \), implies that \( Y_k, Y_\ell, Y_m, Y_r \) also have uniform marginal distributions. Thus, it is sufficient to show that if \( H(X_k, X_\ell) > H(X_m, X_r) \) then

\[ H(Y_k, Y_\ell) > H(Y_m, Y_r). \quad (4.95) \]

Lemma 20 shows that

\[ p^\dagger(y_k, y_\ell) = (1 - q)^2 p(x_k, x_\ell) + \frac{1 - (1 - q)^2}{M^2}, \quad (4.96) \]
\[ p^\dagger(y_m, y_r) = (1 - q)^2 p(x_m, x_r) + \frac{1 - (1 - q)^2}{M^2}. \quad (4.97) \]

Then we consider \( M^2 = L \) and Lemma 19 gives (4.95). \( \square \)
Lemma 22. Consider $X_w, X_{\bar{w}}, X_u, X_{\bar{u}}$ as inputs variables of the $M$-ary symmetric channel (defined in Section 4.4.2) with corresponding outputs $Y_w, Y_{\bar{w}}, Y_u, Y_{\bar{u}}, X \sim p(\cdot) \in \mathcal{P}_T(c_1, c_2)$ and $((w, \bar{w}), u, \bar{u}) \in \mathcal{E}^2$. If the crossover probability $q$ is sufficiently small ($q \to 0$) then

$$I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) = (1 - q)^2 [I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})] - (1 - (1 - q)^2) [KL(U || p(x_w, x_{\bar{w}})) - KL(U || p(x_u, x_{\bar{u}}))] + O(\epsilon^2),$$

and $\epsilon = [1 - (1 - q)^2]/M^2$ and $U$ is the uniform distribution on the alphabet $[M]^2$.

Proof. Recall that the marginal distributions of each node variable $X_w, X_{\bar{w}}, X_u, X_{\bar{u}}$ is uniform and this implies that the marginal distributions of the corresponding outputs $Y_w, Y_{\bar{w}}, Y_u, Y_{\bar{u}}$ are uniform as well. Note that $Y \in [M]^p$ and $X \in [M]^p$ In addition, the pairwise joint distributions of $Y$ in terms of the corresponding joint pairwise distributions of $X$

$$\mathbb{P}(Y_w = x_w, Y_{\bar{w}} = x_{\bar{w}}) = (1 - q)^2 p(x_w, x_{\bar{w}}) + \frac{1 - (1 - q)^2}{M^2}, \quad x_w, x_{\bar{w}} \in [M]^2,$$

$$\mathbb{P}(Y_u = x_u, Y_{\bar{u}} = x_{\bar{u}}) = (1 - q)^2 p(x_u, x_{\bar{u}}) + \frac{1 - (1 - q)^2}{M^2}, \quad x_u, x_{\bar{u}} \in [M]^2.$$

We denote the probability mass function of the pairs $X_w, X_{\bar{w}}$ and $X_u, X_{\bar{u}}$ by $p(x_w, x_{\bar{w}}) \triangleq \mathbb{P}(X_w = x_w, X_{\bar{w}} = x_{\bar{w}})$ and $p(x_u, x_{\bar{u}}) \triangleq \mathbb{P}(X_u = x_u, X_{\bar{u}} = x_{\bar{u}})$ for $(x_w, x_{\bar{w}}), (x_u, x_{\bar{u}}) \in [M]^2$ and similarly, for the noisy versions $Y_w, Y_{\bar{w}}$ and $Y_u, Y_{\bar{u}}$, we use $p_\dagger(x_w, x_{\bar{w}}) \triangleq \mathbb{P}(Y_w = x_w, Y_{\bar{w}} = x_{\bar{w}})$ and $p_\dagger(x_u, x_{\bar{u}}) \triangleq \mathbb{P}(Y_u = x_u, Y_{\bar{u}} = x_{\bar{u}})$ for $(x_w, x_{\bar{w}}), (x_u, x_{\bar{u}}) \in [M]^2$.
Thus,

\[ I(X_w; X_{\tilde{u}}) - I(X_u; X_{\tilde{u}}) \]

\[ = -H(X_w, X_{\tilde{u}}) + H(X_u, X_{\tilde{u}}) \quad (4.98) \]

\[ = \sum_{x_w, x_{\tilde{w}} \in [M]^2} p(x_w, x_{\tilde{w}}) \log p(x_w, x_{\tilde{w}}) \]

\[ - \sum_{x_u, x_{\tilde{u}} \in [M]^2} p(x_u, x_{\tilde{u}}) \log p(x_u, x_{\tilde{u}}) \]

\[ = \sum_{x_w, x_{\tilde{w}} \in [M]^2} p(x_w, x_{\tilde{w}}) \log (1 - q)^2 p(x_w, x_{\tilde{w}}) \]

\[ - \sum_{x_u, x_{\tilde{u}} \in [M]^2} p(x_u, x_{\tilde{u}}) \log (1 - q)^2 p(x_u, x_{\tilde{u}}) \]

\[ = \left[ \sum_{x_w, x_{\tilde{w}} \in [M]^2} (1 - q)^2 p(x_w, x_{\tilde{w}}) \log (1 - q)^2 p(x_w, x_{\tilde{w}}) \right. \]

\[ - \left. \sum_{x_u, x_{\tilde{u}} \in [M]^2} (1 - q)^2 p(x_u, x_{\tilde{u}}) \log (1 - q)^2 p(x_u, x_{\tilde{u}}) \right] \]

\[ \times (1 - q)^{-2}, \quad (4.99) \]

and (4.98) holds because the marginal distributions are uniform. Define

\[ \epsilon \triangleq \frac{(1 - (1 - q)^2)}{M^2} \]

, so \( 0 < \epsilon < q \) for any \( M^2 \geq 2 \). Similarly to Lemma 19, for any \( x_w, x_{\tilde{w}} \in [M]^2 \)

\[ p(\hat{x}_w, x_{\tilde{w}}) \log_2 p(\hat{x}_w, x_{\tilde{w}}) \]

\[ = \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) \log_2 \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) \]

\[ = (1 - q)^2 p(x_w, x_{\tilde{w}}) \log_2 \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) \]

\[ + \epsilon \log_2 \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) \]

\[ = (1 - q)^2 p(x_w, x_{\tilde{w}}) \log_2 \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) + \epsilon \]

\[ + \epsilon \log_2 \left( (1 - q)^2 p(x_w, x_{\tilde{w}}) + \epsilon \right) + O(\epsilon^2). \quad (4.101) \]

Here the last equality holds because \( \log(1 + \epsilon / (1 - q)^2 p(x_w, x_{\tilde{w}})) = \epsilon / (1 - q)^2 p(x_w, x_{\tilde{w}}) + \)
\( \mathcal{O}(\epsilon^2) \) for \( \epsilon \) sufficiently small, while \( p(x_w, x_{\bar{w}}) \) and \( M \) are considered fixed. Also,

\[
\begin{align*}
p_t(x_u, x_{\bar{u}}) \log_2 p_t(x_u, x_{\bar{u}}) \\
= (1 - q)^2 p(x_u, x_{\bar{u}}) \log_2 (1 - q)^2 p(x_u, x_{\bar{u}}) + \epsilon \\
+ \epsilon \log_2 (1 - q)^2 p(x_u, x_{\bar{u}}) + \mathcal{O}(\epsilon^2).
\end{align*}
\] (4.102)

Now, we add and subtract the terms \((1 - q)^{-2} \sum_{x_w, x_{\bar{w}}} \epsilon \log_2 (2(1 - q)^2 p(x_w, x_{\bar{w}})) + \mathcal{O}(\epsilon^2)\) and \((1 - q)^{-2} \sum_{x_u, x_{\bar{u}}} \epsilon \log_2 (2(1 - q)^2 p(x_u, x_{\bar{u}})) + \mathcal{O}(\epsilon^2)\) in (4.99), and we get

\[
I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})
= \left[ \sum_{x_w, x_{\bar{w}} \in [M]^2} (1 - q)^2 p(x_w, x_{\bar{w}}) \log (1 - q)^2 p(x_w, x_{\bar{w}}) \\
+ \epsilon \log_2 (2(1 - q)^2 p(x_w, x_{\bar{w}})) + \mathcal{O}(\epsilon^2) \\
- \sum_{x_u, x_{\bar{u}} \in [M]^2} (1 - q)^2 p(x_u, x_{\bar{u}}) \log (1 - q)^2 p(x_u, x_{\bar{u}}) \\
+ \epsilon \log_2 (2(1 - q)^2 p(x_u, x_{\bar{u}})) + \mathcal{O}(\epsilon^2) \\
+ \sum_{x_u, x_{\bar{u}} \in [M]^2} \epsilon \log_2 (2(1 - q)^2 p(x_u, x_{\bar{u}})) \\
- \sum_{x_w, x_{\bar{w}} \in [M]^2} \epsilon \log_2 (2(1 - q)^2 p(x_w, x_{\bar{w}})) \right] (1 - 2q)^{-2} + \mathcal{O}(\epsilon^2).
\]
The latter and (4.101) and (4.102) give

\[
I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \\
= \frac{1}{(1-q)^2} \left[ -H(Y_w, Y_{\bar{w}}) + H(Y_u, Y_{\bar{u}}) \right] \\
+ \frac{\epsilon}{(1-q)^2} \left[ \sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2(p(x_u, x_{\bar{u}})) \right. \\
- \left. \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2(p(x_w, x_{\bar{w}})) \right] + O(\epsilon^2) \\
= \frac{1}{(1-q)^2} \left[ I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \right] \\
+ \frac{\epsilon}{(1-q)^2} \left[ \sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2(p(x_u, x_{\bar{u}})) \right. \\
- \left. \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2(p(x_w, x_{\bar{w}})) \right] + O(\epsilon^2),
\]

(4.103)

and the latter holds because the marginal distribution of each \(Y\) is uniform. The
definition of $\epsilon$, $\epsilon \triangleq \left(1 - (1 - q)^2\right)/M^2$ together with (4.103) give

\[
I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \\
= (1 - q)^2 \left[ I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right] \\
- \frac{1 - (1 - q)^2}{M^2} \left[ \sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left( p(x_u, x_{\bar{u}}) \right) \right. \\
- \left. \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left( p(x_w, x_{\bar{w}}) \right) \right] + O(\epsilon^2) \\
= (1 - q)^2 \left[ I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right] \\
- \frac{1 - (1 - q)^2}{M^2} \left[ \sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left( \frac{p(x_u, x_{\bar{u}})}{1/M^2} \right) \right. \\
- \left. \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left( \frac{p(x_w, x_{\bar{w}})}{1/M^2} \right) \right] + O(\epsilon^2) \\
= (1 - q)^2 \left[ I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right] \\
- (1 - (1 - q)^2) \left[ - \sum_{x_u, x_{\bar{u}} \in [M]^2} \frac{1}{M^2} \log_2 \left( \frac{1/M^2}{p(x_u, x_{\bar{u}})} \right) \right. \\
+ \left. \sum_{x_w, x_{\bar{w}} \in [M]^2} \frac{1}{M^2} \log_2 \left( \frac{1/M^2}{p(x_w, x_{\bar{w}})} \right) \right] + O(\epsilon^2) \\
= (1 - q)^2 \left[ I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right] \\
- (1 - (1 - q)^2) \left[ \text{KL}(U||p(x_w, x_{\bar{w}})) - \text{KL}(U||p(x_u, x_{\bar{u}})) \right] + O(\epsilon^2).
\]

This completes the proof. \qed
Chapter 5

Conclusions and Future Directions

In this dissertation we studied the problem of learning hidden tree structures. Specifically, we considered tree structured Ising models and Gaussian models and we provided sample complexity guarantees for exact structure recovery when noisy data are available. For the setting of Ising model we further provided sample complexity bounds for the problem of predictive learning and accurate distribution estimation in general. Additionally, we considered the case of non-parametric hidden tree structured models over discrete alphabets and we quantified the sufficient and necessary number of samples for exact structured recovery in this setting as well. For general hidden trees and noisy channels we showed that the fundamental quantity called information threshold suffices to characterize the complexity of tree structure learning problem.

As part of future work, a set of interesting problems remain open. For instance, a characterization of the required samples for predictive learning under general (non-parametric) hidden models, as well as, estimation and evaluation of the information threshold from auxiliary training data. The latter would provide a sense of the sample complexity for unknown tree models and channels before running the CL algorithm. Also, interesting theoretical challenges are connected to the problems that we considered in our work. Generalizations or extensions of the strong data processing inequality may possibly yield a refined version of the upper sample complexity bounds for Ising and Gaussian models. Finally, a closed form expression of the relationship between the noiseless and noisy information thresholds would be of independent interest.
REFERENCES


the curse of dimensionality?” In: *The Annals of Applied Probability* 25.5 (2015),
pp. 2809–2866.

Learning of Markov Tree Structures”. In: *IEEE Transactions on Information

[94] Vincent Yan Fu Tan. “Large-deviation analysis and applications of learning tree-
structured graphical models”. PhD thesis. Massachusetts Institute of Technology,

of Ising Models”. In: *Advances in Neural Information Processing Systems 29.*

Random Fields and Subfields”. In: *IEEE Transactions on Information Theory*

estimates for discrete distributions”. In: *Random Structures & Algorithms* 19.3-4

Results: Studying Plug-In Estimators”. In: *Entropy* 20.6 (2018), p. 397. url:

[99] Arnab Bhattacharyya et al. “Near-Optimal Learning of Tree-Structured Distribu-