Theoretical Study of Two Prediction-Centric Problems: Graphical Model Learning and Recommendations

by

Mina Karzand

B. Sc, University of Tehran (2007)

M.Sc, École Polytechnique Fédérale de Lausanne (2009)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Author . . . . . .
Department of Electrical Engineering and Computer Science
August 31, 2017

Certified by . . . . . .
Lizhong Zheng
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by . . . . . .
Guy Bresler
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Theoretical Study of Two Prediction-Centric Problems: Graphical Model Learning and Recommendations

by

Mina Karzand

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Motivated by prediction-centric learning problems, two problems are discussed in this thesis.

PART I. *Learning a tree-structured Ising model:*

We study the problem of learning a tree Ising model from samples such that subsequent predictions based on partial observations are accurate. Virtually all previous work on graphical model learning has focused on recovering the true underlying graph. We define a distance ("small set TV" or ssTV) between distributions P and Q by taking the maximum, over all subsets $\mathcal{S}$ of a given size, of the total variation between the marginals of P and Q on $\mathcal{S}$; this distance captures the accuracy of the prediction task of interest. We derive non-asymptotic bounds on the number of samples needed to get a distribution (from the same class) with small ssTV relative to the one generating the samples. An implication is that far fewer samples are needed for accurate predictions than for recovering the underlying tree.

PART II. *Optimal online algorithms for a latent variable model of recommendation systems:*

We consider an online model for recommendation systems, with each user being recommended an item at each time-step and providing 'like' or 'dislike' feedback. The user preferences are specified via a latent variable model: both users and items are clustered into types. The model captures structure in both the item and user spaces, and our focus is on simultaneous use of both structures. In the case when the type preference matrix is randomly generated, we provide a sharp analysis of the best possible regret obtainable by any algorithm.

Thesis Supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Guy Bresler
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I owe my deepest gratitude to my advisors Professor Lizhong Zheng and Professor Guy Bresler. Lizhong has been a true mentor through my graduate studies. His sincere sense of responsibility and concern for my education has helped me through the obstacles in this journey. He has been a true inspiration in building my taste in research and overcoming challenges in my academic life.

The completion of this thesis was not possible without the help and support of Professor Guy Bresler. Guy never hesitated to share his knowledge, experience and expertise with me. I could ask him for consult regarding the spectrum of challenges in my studies: whether I needed a consult on the broad picture in my research portfolio or I was just frustrated with a small detail in the proof of some lemma. I am grateful for the joyful experience of research I had through working with him.

I would like to gratefully acknowledge the members of my thesis committee Prof. Gregory Wornell and Prof. Muriel Médard for reviewing this manuscript in spite of their busy schedule and for their instructive comments.

I offer my sincere appreciation to all the members of MIT community whose help and support were essential for me to survive my graduate studies. In particular, I would like to thank the members of LIDS, IDSS and RLE for the amazing environment and community they have provided for me.

Most importantly, I would like to express my utmost appreciation for my parents who never hesitated to support me through this time. Their encouragement and faith in me has been my unique motivation to keep pushing regardless of the ups and downs of this path.

I would like to thank my brothers Masoud and Mohammad who have always been my rocks. They have been my cheer leaders even during the times that it was hard for me to stay confident.

The years I spent in Cambridge has been filled with love. I was blessed to share my life with amazing friends who touched my life and made me a different person; to share happiness, sad days, enthusiasm and frustrations with amazing friends.

Contents

List of Figures

Chapter 1

Introduction

In recent years, a new regime of interest and associated set of problems have become central to machine learning. Due to recent technological advances, the volume of data available has been growing steadily and this enables engineers to attack more complex problems. The data are often high-dimensional, meaning that there are many features associated with each datapoint. From the mathematical point of view, this high dimensionality has required an entirely new set of probabilistic and statistical tools.

The canonical problem in machine learning is the problem of model selection and parameter estimation within a class of statistical models. The proper choice of model allows one to draw conclusions, make predictions and take actions based on the data. Hence, the class of models for the data generation processes and the choice of loss function in the model selection process play key roles.

Our basic philosophy in this thesis is the following: "performance can be improved if the learning algorithm is tailored to the end-goal of the learning process". This thesis addresses two problems fitting into this theme.

In part I, we consider model selection with the goal of subsequently making predictions based on partial observations. More specifically, we study graphical model learning subject to a loss function which requires posterior distributions to be accurate. Accurate posterior distributions based on partial observations is a necessity to make good predictions, but guaranteeing correct recovery of underlying combinato-

rial structure is not required. Considering this observation in the context of learning graphical models, we show that far fewer samples are required to learn a model which is guaranteed to provide accurate posteriors compared to learning a model which guarantees correct structure recovery.

In part II, we were motivated by recommendation systems, a situation in which actions are taken based on information collected over time. The accuracy of the learned representation is relevant so far as it helps to inform good actions. In particular if the goal of learning is taking actions such as making recommendations, then we can adopt an exploration strategy to learn only the aspects of the model which help us to make meaningful recommendations.

In the following, we give more details about the problem setup and the results in each part:

Part I. In part I, our focus is on learning graphical models. In particular, we talk about learning tree-structured Ising models for the purpose of subsequent predictions. Structure learning refers to the problem of learning the underlying graph representing the conditional dependency relations between variable from samples generated from a distribution.

It is often the case that the learned graphical model is used for performing inference later on. This problem has been studied extensively in the literature, but the measure of performance used in the literature has been learning the true underlying graph with high probability. In this body of work, we propose a different measure of performance to guarantee accurate subsequent predictions using the learned graph irrespective of whether or not the underlying graph is recovered correctly. We introduce a loss function called *small set TV distance* (ssTV) between two distributions. We show that small ssTV distance between the original distribution generating the samples and the distribution learned from the samples guarantees that the outcome of any inference task on the graph based on a few observations given the learned distribution is accurate compared to the original distribution.

We provide theoretical guarantees on the outcome of the Chow-Liu algorithm for

tree-structured graphs. We show that structure learning with guarantees of small ssTV requires strictly smaller number of samples than structure learning with accurate recovery of underlying graph.

Part II. The second problem we study is online recommendation systems with a latent variable model. The underlying model of the preference of users for items is based on existence of some types of users and types of items where users of the same type like or dislike each item similar to each other and also items of the same type are liked or disliked similarly by all users. We provide optimal online algorithms and tight regret bounds for the regimes of interest in a model based on only structure in item space and a separate optimal algorithm based on structure only in user space. This is, to the best of our knowledge, one of the first works which provides optimality guarantees in the context of online recommendation systems.

Our analysis elucidates the system operating regimes in which existing algorithms are nearly optimal, as well as highlighting the sub-optimality of using only one of item or user structure (as is done in commonly used item-item and user-user collaborative filtering). This prompts a new hybrid algorithm that is nearly optimal in essentially all parameter regimes.

Part I

Learning a Tree-Structured Ising Model in order to Make Predictions

Chapter 2

Motivation and background

Markov random fields, or undirected graphical models, are a useful way to represent high-dimensional probability distributions [1, 2]. Given an undirected graph $G = (\mathcal{V}, \mathcal{E})$ a Markov random field defined on G is a set of random variables having a Markov property described by G: each random variable is associated with a vertex in G. A variable is conditionally independent of all other variables given its neighbors.

Their practical utility is in part due to 1) edges between variables capture direct interactions, which make the model *interpretable* and 2) the graph structure facilitates efficient approximate *inference* from partial observations, for example using loopy belief propagation or variational methods. The inference task of interest to us in this thesis is evaluation of conditional probabilities or marginals, e.g. $P(x_i|X_{\mathcal{S}} = x_{\mathcal{S}})$ or $P(x_{\mathcal{S}})$, having observed values $x_{\mathcal{S}}$ for a set of variables $\mathcal{S}$.

In applications it is often necessary to learn the model from data, and it makes sense to measure accuracy of the learned model in a manner corresponding to the intended use. While in some applications it is of interest to learn the graph itself, in many machine learning problems the focus is on making predictions. In the literature, learning the graph is called *structure learning*; this problem has been studied extensively in recent years, see e.g. [3, 4, 5, 6, 7, 8]. In this thesis we consider the problem of learning a good model *for the purpose of performing subsequent prediction from partial observations*. This objective has been called "inferning" [9], and has received significantly less attention. This thesis contains, to the best of our knowledge,

the first results on estimating graphical models with a prediction-centric loss that are applicable to the high-dimensional setting.

Structure learning becomes statistically more challenging, meaning more data is required when interactions between variables are very weak or very strong [10, 11, 12]. It is intuitively clear that very weak edges are difficult to detect, leading to non-identifiability of the model. The goal of this thesis is to show that accurate inference is possible even when structure learning is not.

With the goal of making predictions in mind, we introduce a loss function to evaluate learning algorithms based on the accuracy of low-order marginals. The loss between true distribution P and learned distribution Q is defined to be

$$\mathcal{L}^{(k)}(P, Q) \triangleq \max_{\mathcal{S}:|\mathcal{S}|=k} d_{\mathrm{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}) \,,$$

where $P_{\mathcal{S}}$ denotes the marginal on set $\mathcal{S}$. The total variation distance between two distributions $P(x)$ and $Q(x)$ for $x \in \mathcal{X}$ is defined as:

$$d_{\mathrm{TV}}(P, Q) \triangleq \frac{1}{2} \sum_{x \in \mathcal{X}} \big| P(x) - Q(x) \big| \,. \tag{2.1}$$

The small-set total variation is inherently far less stringent than the total variation over the entire joint distribution. This makes a crucial difference in high-dimensional models. This same local total variation metric was used by Rebeschini and van Handel in a somewhat different context [13] and has appeared earlier in Dobrushin's work on Gibbs measures [14]. As discussed in Section 3.2.2, small loss $\mathcal{L}^{(k)}$ guarantees accurate posterior distributions conditioned on sets of size $k-1$.

Tree-structured graphical models have been particularly well-studied. Aside from their theoretical appeal, there are two reasons for the practical utility of tree models: 1) Structure learning for trees can be accomplished with smaller sample and time complexity as compared to loopy graphs, and 2) Efficient exact inference (computation of marginals) is possible using belief propagation. Sum-product or max-product algorithm are two well-studied examples [15, 16, 1, 17] of inference algorithms on

trees. Hence, in this thesis, we focus on tree-structured models.

In this thesis, we restrict attention to tree-structured Ising models with no external field, defined as follows. For tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ on p nodes and edge parameters θ_{ij} for each edge $(i,j) \in \mathcal{E}$, each configuration $x \in \{-1,+1\}^p$ is assigned probability

$$P(x) = \exp\left(\sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_i x_j - \Phi(\theta) \right), \tag{2.2}$$

where $\Phi(\theta)$ is the normalizing constant. We assume throughout that $\alpha \leq |\theta_{ij}| \leq \beta$ for some $\alpha, \beta \geq 0$ for each edge $(i,j) \in \mathcal{E}$. An external field at node i can be captured by a term $\theta_i x_i$ in the exponent. No external field means that $\theta_i = 0$ for each i. This implies that $P(X_i = +1) = P(X_i = -1) = 1/2$ for all i.

Suppose we observe i.i.d. samples, generated from a tree-structured Ising model. The main question we address is how many samples are required in order to guarantee that subsequent inference computations are accurate. Since computation of marginals on trees is easy, the crux of the task is in learning a model with marginals that are close to those of the original model. One of the take-home messages is that learning for the purpose of making predictions requires dramatically fewer samples than is necessary for correctly recovering the underlying tree. The central technical challenge is that our analysis must also apply when it is impossible to learn the true tree, and this requires careful control of the sorts of errors that can occur.

Our main result gives lower and upper bounds on the number of samples needed to learn a tree Ising model to ensure small $\mathcal{L}^{(2)}$ loss, which in this setting is equivalent to accurate pairwise marginals. We emphasize that the task is to learn a model from the same class (tree-structured Ising) with these guarantees; this is sometimes called *proper* learning [18]. The main result concerns the maximum likelihood tree (also called Chow-Liu tree [19]), defined in Section 3.2. We give a preview here:

Theorem 2.1. *Fix $\eta > 0$. Given $n > C \max\{\eta^{-2} \log \frac{p}{\delta\eta}, e^{2\beta} \log \frac{p}{\delta}\}$ samples generated according to a tree Ising model P defined in (2.2) with $|\theta_{ij}| \leq \beta$. Denote the Chow-Liu tree by T^{CL}. The Ising model Q on T^{CL} obtained by matching correlations satisfies $\mathcal{L}^{(2)}(P,Q) \leq \eta$ with probability at least $1-\delta$. Conversely, if*

$n \leq C'\eta^{-1} atanh^{-1}\eta \log(p/\eta)$, then no algorithm can find a tree model Q such that $\mathcal{L}^{(2)}(P, Q) \leq \eta$ with probability greater than half.

The result shows that the Chow-Liu tree, which can be found in time $O(p^2 \log p)$, gives small $\mathcal{L}^{(2)}$ error. We next place the result in the context of related work.

2.1 Related work

Structure learning in general graphical models has been studied extensively. Information-theoretic bounds on the number of samples have been provided [10, 12, 11, 20, 21, 22]. Structure learning of trees has been studied by [19, 23]. Learning of generalizations of tree-structured models has been studied, including: forest approximations [24, 25], polytrees [26], bounded treewidth graphs [27, 28], loopy graphs with correlations decay [29, 30], and mixtures of trees [31, 8].

Loopy belief propagation yields accurate marginals in high girth graphs (locally tree-like graphs) with correlation decay. This fact is used by Heinemann and Globerson [9] to justify an algorithm which recovers all the edges from this family, given sufficiently many samples. The output of their algorithm can have extra edges, which are proved to be weak. Given number of samples scaling at least linearly with p, they guarantee that the learned distribution is within a constant of the true one in Kullback-Leibler divergence.

In [28], a polynomial time algorithm using a polynomial number of samples is proposed to learn bounded tree-width graphical models with respect to KL divergence. They use ideas from submodular optimization and proper factorization of the distribution over bounded treewidth graphs.

Learning of latent tree models has been well-studied in the phylogenetic reconstruction literature. Erdös et al. [32] studied sample and time complexity of tree metric based algorithms to reconstruct phylogenetic trees. Daskalakis et al. [33] and Mossel [34] use distorted tree metrics to get approximations of the phylogenetic trees when the exact reconstruction of the tree is not possible. In [34] a forest approximation of the tree is recovered. The maximum number of connected components in this

forest is given as a function of edge strengths, maximum distance between the leaves and the number of leaves. Daskalakis et al. [33] removed the prior assumptions on the phylogenetic tree and instead a forest structure is recovered which is guaranteed to recover edges that are sufficiently long and sufficiently close to the leaves.

The tree metric over p nodes is associated with a weighted spanning tree such that the distance between every pair of nodes is the sum of weights of the edges along the path between the nodes in the tree. In Agarwala et al. [35], given a pairwise distance matrix D over p nodes is approximated by a tree metric with induced distance matrix T. Let $\epsilon = \min_T\{\|T - D\|_\infty\}$ where T is a tree metric. An $O(p^2)$ algorithm is provided which produces $\widehat{T}$ with $\|\widehat{T} - D\|_\infty \leq 3\epsilon$. It is shown that finding a T with $\|T - D\|_\infty \leq 9/8\epsilon$ is NP-hard. Ambainis et al. in [36] studied the *leaf variational distance* between the original distribution on a latent tree under the Cavender-Faris (CF) model and the learned latent tree. Let tree T with leaves $\mathcal{V}$ be a CF-tree on p leaves with the property that all its edges are of length at least $1/\sqrt{n}$, (which is translated as $\log\tanh\alpha < 1/\sqrt{n}$ in our model). Then given n observations, their proposed algorithm produces a distribution Q on a tree T' with leaves $\mathcal{V}$ such that $\mathcal{L}^{(2)}(P_\mathcal{V}, Q_\mathcal{V}) = O(\sqrt{\frac{pe^{2\beta}}{n}})$ (Here $P_\mathcal{V}$ and $Q_\mathcal{V}$ are marginals of P and Q on leaves $\mathcal{V}$). We will discuss these results and compare them with our setup in detail in Section 6.

Wainwright [37] was motivated by the same general problem of learning a graphical model to be subsequently used for making predictions, but the focus was on computational rather than the statistical limits. For loopy graphs, both estimation of parameters and prediction based on finite observations are computationally challenging tasks. Hence, for both, approximate heuristic methods are often used. For given model parameters, approximate prediction can be achieved using heuristics such as reweighted sum-product (a convex relaxation). Intriguingly, when using such approximate prediction algorithms, an *inconsistent* procedure for estimation of parameters can give better predictions. The results elucidate asymptotic performance, but the analysis does not apply to high-dimensional setting with fewer samples n than dimension p.

2.2 Outline of part I

The next section provides background on the Ising model, tree models, and graphical model learning. Section 3.2.1 introduces the problem of learning tree-structured Ising models and specifies the sample complexity of exact recovery. Then, in Section 3.2.2 we define a loss function motivated by inference computation and state our main result in Section 3.2.3. Section 3.3.1 analyzes an illustrative example to provide intuition for the main result. Section 3.3.2 introduces the truncation algorithm and analyzes its performance in terms of ssTV and Section 3.4 contains a sketch of the full proof. Section 4 contains proof of the main result. Sections 4.5 through 4.7 contain further proofs.

Chapter 3

Learning for the purpose of making prediction

3.1 Preliminaries

3.1.1 Notation

For a given tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ and positive numbers α and β, let $\mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ be the set of Ising models (2.2) with the restriction $\alpha \leq |\theta_{ij}| \leq \beta$ on each edge $(i, j) \in \mathcal{E}$. Denote by $\mathcal{P}_{\mathsf{T}} = \mathcal{P}_{\mathsf{T}}(0, \infty)$ the set of Ising models on T with no restrictions on parameter strength.

Denote by $\mu_{ij} = \mathbb{E}_P X_i X_j$ the correlation between the variables corresponding to $i, j \in \mathcal{V}$. For an edge $e = (i, j)$, we write $\mu_e = \mu_{ij}$ and similarly for a set of edges $\mathcal{A} \subseteq \mathcal{E}$, $\mu_{\mathcal{A}} = \prod_{e \in \mathcal{A}} \mu_e$. Given n i.i.d. samples $X^{(1:n)} = X^{(1)}, \ldots, X^{(n)}$, the empirical distribution is denoted by $\widehat{P}(x) = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{X^{(l)}=x\}}$ and $\widehat{\mu}_{ij} = \mathbb{E}_{\widehat{P}} X_i X_j$ is the empirical correlation between nodes i and j.

3.1.2 Tree models

A probability measure P on $\mathcal{X}^{\mathcal{V}}$ is Markov with respect to a graph $G = (\mathcal{V}, \mathcal{E})$ if for all $i \in \mathcal{V}$, we have $P(x_i | x_{\mathcal{V} \setminus \{i\}}) = P(x_i | x_{\partial i})$ (here ∂i is the neighborhood of i in G). In this part, we are interested in distributions P that are Markov with respect to a

tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$. A consequence (see [15]) is that $P(x)$ factorizes as

$$P(x) = \prod_{i \in \mathcal{V}} P(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}. \tag{3.1}$$

Denote by $D(Q\|P)$ the Kullback-Leiber divergence between probability measures Q and P defined as $D(Q\|P) = \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{P(x)}$. For an arbitrary distribution P and tree T, the distribution

$$\widetilde{P}(x) = \underset{\substack{Q \text{ is factorized as (3.1)} \\ \text{according to} \mathsf{T}}}{\arg\min} D(P\|Q)$$

is the best approximation to P within the set of distributions Markov with respect to the tree T. It is proved by Chow and Liu in [19] that $\widetilde{P}$ is specified by matching the first and second order marginals to those of P, i.e., for all $(i,j) \in \mathcal{E}$, and all $x_i, x_j \in \mathcal{X}$, $\widetilde{P}(x_i, x_j) = P(x_i, x_j)$.

Let

$$\Pi_{\mathsf{T}}(P) = \underset{Q \in \mathcal{P}_{\mathsf{T}}}{\arg\min} D(P\|Q) \tag{3.2}$$

be the reverse information projection of P onto the class of Ising models on $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ with no external field. Hence, $\widetilde{P} = \Pi_{\mathsf{T}}(P)$ can be represented as Equation (2.2) for some $\widetilde{\theta}$ supported on T (*i.e.*, $\widetilde{\theta}_{ij} = 0$ if $(i,j) \notin \mathcal{E}$).

It is shown in the supplementary materials that $\widetilde{P} = \Pi_{\mathsf{T}}(P)$ has edge weights $\widetilde{\theta}_{ij}$ for each $(i,j) \in \mathcal{E}_{\mathsf{T}}$ satisfying $\tanh \widetilde{\theta}_{ij} = \mu_{ij} \triangleq \mathbb{E}_P X_i X_j$.

Denote the set of all trees on p nodes by $\mathcal{T}$. For some distribution $P \in \mathcal{P}_{\mathsf{T}}$, one observes n independent samples (configurations) $X^{(1)}, \ldots, X^{(n)} \in \{-,+\}^p$ from the Ising model (2.2). Restricting to trees, a *structure learning algorithm* is a (possibly randomized) map $\phi : \{-1,+1\}^{p \times n} \to \mathcal{T}$ taking n samples $X^{(1:n)} = X^{(1)}, \ldots, X^{(n)}$ to a tree $\phi(X^{(1:n)})$.

The maximum likelihood tree or Chow-Liu tree plays a central role in tree structure learning. Chow and Liu [19] observed that the maximum likelihood tree is the max-weight spanning tree in the complete graph, where each edge has weight equal to the empirical mutual information between the variables at its endpoints. The tree can

thus be found greedily via Kruskal's algorithm[19, 38], and the run-time is dominated by computing empirical mutual information between all pairs of nodes.

In the supplementary materials application of analysis similar to [19] to zero-field Ising models on trees is provided to support the following definition:

Definition 3.1 (Chow-Liu Tree)**.** Given n i.i.d samples $X^{(1:n)}$ from distribution $P \in \mathcal{P}_{\mathsf{T}}$, we define the Chow-Liu tree to be the maximum likelihood tree:

$$\mathsf{T}^{\mathsf{CL}} = \operatorname*{argmax}_{\mathsf{T}\in\mathcal{T}} \max_{P\in\mathcal{P}_{\mathsf{T}}} P(X^{(1:n)})\,.$$

This definition is slightly abusing the conventional terminology, as the Chow-Liu tree is classically the maximum likelihood tree assuming that the generative distribution is tree-structured [19], whereas in our definition we are assuming that the original distribution $P \in \mathcal{P}_{\mathsf{T}}$ can be described by Equation (2.2). Thus, it is not only tree-structured but also has uniform singleton marginals.

Note that maximizing the likelihood of i.i.d. samples corresponds to minimizing the KL divergence. Given the samples $X^{(1:n)}$ with empirical distribution

$$\widehat{P}(x) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}[X^{(m)} = x],$$

$\mathsf{T}^{\mathsf{CL}} = \arg\min_{\mathsf{T}\in\mathcal{T}} \min_{P\in\mathcal{P}_{\mathsf{T}}} D(\widehat{P}\|P)$. It is shown in the supplementary materials that

$$\mathsf{T}^{\mathsf{CL}} = \operatorname*{argmax}_{\{\text{spanning trees } \mathsf{T}'\}} \sum_{e\in\mathcal{E}_{\mathsf{T}'}} |\widehat{\mu}_e|\,, \tag{3.3}$$

where for $e = (i, j)$,

$$\widehat{\mu}_e = \mathbb{E}_{\widehat{P}} X_i X_j = \frac{1}{n}\sum_{m=1}^{n} X_i^{(m)} X_j^{(m)},$$

is the empirical correlation between variables X_i and X_j .

Chow and Wagner [39] showed that the maximum likelihood tree is consistent for structure learning of general discrete tree models, i.e., in the limit of large sample size n the correct graph structure is found. More recently, detailed analysis of error

exponents was carried out by Tan et al. [23, 24]. A variety of other results and generalizations have appeared, including for example Liu et al.'s work on forest estimation with non-parametric potentials [25] (we will not address general potentials in this thesis).

3.2 Learning trees to make predictions

In this section, we first state a specialization to our setting of the Chow-Liu algorithm. Next, in order to place the learning for predictions problem into context, we discuss the problem of exact structure learning and give tight (up to a multiplicative constant) sample complexity for that problem. Finally, in Section 3.2.2 we define the ssTV distance $\mathcal{L}^{(k)}$, explain how it relates to prediction, and then state our results.

3.2.1 Exact recovery of trees

The statistical performance of a structure learning algorithm is often measured using the zero-one loss,

$$\mathcal{L}^{0-1}(\mathsf{T}, \mathsf{T}') = \mathbf{1}_{\{\mathsf{T} \neq \mathsf{T}'\}}\,, \tag{3.4}$$

meaning that the exact underlying graph must be learned (see e.g., [4, 10, 24, 25]). The risk, or expected loss, under some distribution $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ is then given by the probability of reconstruction error, $\mathbb{E}_P \mathcal{L}^{0-1}(\mathsf{T}, \phi(X^{(1:n)})) = \mathsf{P}(\phi(X^{(1:n)}) \neq \mathsf{T})$, and for given α, β, p and n, the maximum risk is $\sup\{\mathsf{P}(\phi(X^{(1:n)}) \neq \mathsf{T}) : \mathsf{T} \in \mathcal{T}, P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)\}$.

The sample complexity of learning the correct tree underlying the distribution increases as edges become weaker, i.e., as $\alpha \to 0$, because weak edges are harder to detect. As the upper bound on maximum edge parameter β increases, there is a similar increase in sample complexity (as shown by [10, 12] for Ising models on general bounded degree graphs). In the context of tree-structured Ising models we give the following theorem:

Theorem 3.2 (Samples necessary for structure learning)**.** *Given* $n < \frac{1}{16} e^{2\beta} \alpha^{-2} \log p$ *samples, the worst-case probability of error over trees* $\mathsf{T} \in \mathcal{T}$ *and distributions* $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ *is at least half for any algorithm, i.e.,*

$$\inf_{\phi} \sup_{\substack{\mathsf{T} \in \mathcal{T} \\ P \in \mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} P(\phi(X^{(1:n)}) \neq \mathsf{T}) > 1/2 \,.$$

The proof, given in Section 4.5.1, applies Fano's inequality to a large set of trees that are difficult to distinguish. The next theorem gives an essentially matching sufficient condition.

Theorem 3.3 (Samples sufficient for structure learning)**.** *Fix an arbitrary tree* T *and Ising model* $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$*. If the number of samples is* $n > Ce^{2\beta} \tanh^{-2}(\alpha) \log(p/\delta)$*, then with probability at least* $1 - \delta$ *the Chow-Liu algorithm recovers the true tree, i.e.,* $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$.

The proof is presented in Section 4.5.2. Note that if $\alpha < 1/2$, then Theorems 3.2 and 3.3 give matching bounds (up to numerical constant) for the sample complexity of learning the tree structure of an Ising model with zero external field (defined immediately after Equation (2.2)). The necessary number of samples increases as the minimum edge weight α decreases. Hence, if edges can be arbitrarily weak, it is impossible to learn the tree given any bounded number of samples. If the goal is merely to make accurate predictions, it is natural to seek a less stringent, approximate notion of learning.

Several papers consider learning a model that is close in KL-divergence, e.g. [6, 25, 9, 40, 24]. The sample complexity of learning a model to within constant KL-divergence ϵ scales at least *linearly* with the number of variables p, an unrealistic requirement in the high-dimensional setting of interest. Using a number of samples scaling logarithmically in dimension requires relaxing the KL-divergence to scale linearly in p, but this does not imply a non-trivial guarantee on the quality of approximation for marginals of few variables (as done in this paper). The same observation is true for the total variation as the measure of distance. The TV distance between

the learned model the original joint distribution over p variables, scales linearly with p given fixed number of observations.

In the next section, we study estimation with respect to the small-set TV loss, which captures accuracy of prediction based on few observations. We will see that the associated sample complexity is independent of the minimum edge parameter α of the original model.

3.2.2 Small set total variation

For a subset of nodes $\mathcal{S} \subseteq [p]$, we denote by $P_{\mathcal{S}}$ the marginal distribution $P_{\mathcal{S}}(x_{\mathcal{S}}) = \sum_{x_{\mathcal{V}\setminus\mathcal{S}}} P(x)$.

Given two distributions P and Q on the same space, for each $k \geq 1$ the maximum total variation distance over all size k marginals is denoted by

$$\mathcal{L}^{(k)}(P, Q) \triangleq \max_{\mathcal{S}:|\mathcal{S}|=k} d_{\mathrm{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}})\,.$$

Note that $\mathcal{L}^{(k)}$ is non-decreasing in k. One can check that $\mathcal{L}^{(k)}$ satisfies the triangle inequality: for any three distributions P, Q, R,

$$\mathcal{L}^{(k)}(P, Q) + \mathcal{L}^{(k)}(Q, R) \geq \mathcal{L}^{(k)}(P, R)\,. \tag{3.5}$$

Closeness of P and Q in $\mathcal{L}^{(k)}$ implies that the respective posteriors conditioned on subsets of variables of size $k-1$ are close on average. To see this, suppose that we wish to compute $P(X_i = +|X_{\mathcal{S}})$. We measure performance of the approximation Q by the expected magnitude of error $|P(x_i = +|X_{\mathcal{S}}) - Q(x_i = +|X_{\mathcal{S}})|$ averaged over

$X_{\mathcal{S}}$:

$$
\begin{aligned}
&\mathbb{E}_{X_{\mathcal{S}}}|P(X_i = +|X_{\mathcal{S}}) - Q(X_i = +|X_{\mathcal{S}})| \\
&= \sum_{x_{\mathcal{S}}} |P(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) - Q(X_i = +|X_{\mathcal{S}} = x_{\mathcal{S}})P(X_{\mathcal{S}} = x_{\mathcal{S}})| \\
&\leq \sum_{x_{\mathcal{S}}} |P(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) - Q(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}})| + \sum_{x_{\mathcal{S}}} |Q(x_{\mathcal{S}}) - P(x_{\mathcal{S}})| \\
&\leq 2\mathcal{L}^{(|\mathcal{S}|+1)}(P, Q)\,.
\end{aligned}
$$

The last inequality is a consequence of monotonicity of $\mathcal{L}^{(k)}$ in k.

In this paper we focus on $\mathcal{L}^{(2)}$. Preliminary implications for $k \geq 3$ are discussed in the supplementary materials. For trees $\mathsf{T}, \widetilde{\mathsf{T}} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}$ and $\widetilde{P} \in \mathcal{P}_{\widetilde{\mathsf{T}}}$, we have

$$
\begin{aligned}
\mathcal{L}^{(2)}(P, \widetilde{P}) &= \max_{w, \widetilde{w} \in \mathcal{V}} \frac{1}{2} \sum_{x_w, x_{\widetilde{w}} \in \{-,+\}^2} |P(x_w, x_{\widetilde{w}}) - \widetilde{P}(x_w, x_{\widetilde{w}})| \\
&= \max_{w, \widetilde{w} \in \mathcal{V}} \frac{1}{2} \left| \prod_{e \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})} \mu_e - \prod_{e' \in \mathsf{path}_{\widetilde{\mathsf{T}}}(w, \widetilde{w})} \widetilde{\mu}_{e'} \right|, \qquad (3.6)
\end{aligned}
$$

where for $e = (i, j) \in \mathcal{E}_{\mathsf{T}}$, $\mu_e = \mathbb{E}_P X_i X_j$ and for $e' = (i, j) \in \mathcal{E}_{\widetilde{\mathsf{T}}}$, $\widetilde{\mu}_{e'} = \mathbb{E}_{\widetilde{P}} X_i X_j$. Since $P(x_i = +) = 1/2$, $P(x_w, x_{\widetilde{w}}) = [1 + x_w x_{\widetilde{w}} \mathbb{E}_P[X_w X_{\widetilde{w}}]]/4$ and analogously for $\widetilde{P}$. Also, from the Ising model definition (2.2) one can check that $\mathbb{E}_P X_w X_{\widetilde{w}} = \prod_{e \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})} \mu_e$. The same holds for $\widetilde{P} \in \mathcal{P}_{\widetilde{\mathsf{T}}}$, which gives (3.6).

Notice that the products of correlations appearing in Equation (3.6) are along possibly different paths when the corresponding trees T and $\widetilde{\mathsf{T}}$ are not equal.

3.2.3 Main result

Our main contribution is to provide upper and lower bounds on the number of samples required to obtain accurate pairwise marginals. To get the upper bound on the number of samples, we study the Chow-Liu algorithm and bound the expression in Equation (3.6). The Chow-Liu algorithm produces the maximum likelihood tree,

which minimizes the expected zero-one loss in (3.4). As shown in Theorem 3.4, the maximum likelihood tree also performs well in terms of accuracy of pairwise marginals.

Recall from (3.2) that $\Pi_{\mathsf{T}}(P) = \arg\min_{Q \in \mathcal{P}(\mathsf{T})} D(P\|Q)$ is the reverse information projection of the distribution P onto the set of zero-field Ising models on tree T.

Theorem 3.4 (Learning for inference using Chow-Liu algorithm)**.** *For $\mathsf{T} \in \mathcal{T}$, let the distribution $P \in \mathcal{P}_{\mathsf{T}}(0, \beta)$. Given $n > C \max\{e^{2\beta} \log \frac{p}{\delta}, \eta^{-2} \log \frac{p}{\eta\delta}\}$ samples, if the tree T^{CL} is the Chow-Liu tree as defined in* (3.3)*, then with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$.*

The main challenge is that the number of samples available in this regime is not sufficient for structure learning, as can be seen by comparing with Theorem 3.2. This means that accurate marginals must be computed *using possibly a wrong tree.* A sketch of the proof is provided in Section 3.4.

We also lower bound the number of samples necessary for small $\mathcal{L}^{(2)}$ loss. Let the learning algorithm be $\Psi : \{-1, +1\}^{p \times n} \to \mathcal{P}$ where $\mathcal{P} = \cup_{\text{trees}\mathsf{T}} \mathcal{P}_{\mathsf{T}}$ is the set of tree-structured Ising models with no external field defined in Equation (2.2).

Theorem 3.5 (Samples necessary for small ssTV)**.** *Suppose one observes*

$$n < \frac{1}{6\eta \operatorname{atanh}(\eta)} \log(p/\eta)$$

samples. Then the worst-case probability of $\mathcal{L}^{(2)}$ loss greater than η taken over trees $\mathsf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ is at least half for any algorithm, i.e.,

$$\inf_{\Psi} \sup_{\substack{\mathsf{T} \in \mathcal{T} \\ P \in \mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathsf{P}\left[\, \mathcal{L}^{(2)}(P, \Psi(X^{(1:n)})) > \eta \,\right] > 1/2\,.$$

In the statement of the above lower bound for sample complexity we assume $\tanh(\beta) - \tanh(\alpha) \geq \sqrt{\eta/p}$. The proof is provided in Section 4.4. The high-probability bounds from Theorems 3.4 and 3.5 imply the following bounds on risk.

Corollary 3.6 (Upper bound for risk)**.** *For $\mathsf{T} \in \mathcal{T}$, let the distribution $P \in \mathcal{P}_{\mathsf{T}}(0, \beta)$. Given n samples with empirical distribution $\widehat{P}$, if the tree T^{CL} is the Chow-Liu tree*

as defined in (3.3)*, then*

$$\mathbb{E}\Big[\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}))\Big] < C' p^2 \exp(-Cne^{-2\beta}) + C''\sqrt{\frac{\log n + \log p}{n}}\,.$$

Corollary 3.7 (Lower bound for risk)**.** *Suppose one observes n samples. Then the minimax risk over trees $\mathsf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ is lower bounded by*

$$\inf_{\Psi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathbb{E}\left[\mathcal{L}^{(2)}(P, \Psi(X^{(1:n)}))\right] > \min\left\{\frac{1}{24}\sqrt{\frac{\log n + \log p}{n}}, 1/4\right\}\,.$$

The upper bound in Corollary 3.6 has an extra term (that depends on β) relative to the lower bound of Corollary 3.7. We conjecture that the lower bound is tight.

3.3 Illustrative example and comparison with forest approximation

3.3.1 Three node Markov chain

A Markov chain with three nodes captures a few of the key ideas developed in this paper. Let $p = 3$ and $P(X_1, X_2, X_3)$ be represented by the tree T_1 in Figure 3.3.1 in which $X_1 \leftrightarrow X_2 \leftrightarrow X_3$ form a Markov chain with correlations μ_{12}, μ_{23} and $\mu_{13} = \mu_{12}\mu_{23}$. Without loss of generality, we assume $\mu_{12}, \mu_{23} > 0$. Suppose that for some small value ϵ, $\mu_{12} = 1 - \mu_{23} = \epsilon$.

Given n samples from P, the empirical correlations $\widehat{\mu}_{12}$, $\widehat{\mu}_{23}$ and $\widehat{\mu}_{13}$ are concentrated (to within $\approx \pm 1/\sqrt{n}$ with highprobability) around $\mu_{12} = \epsilon$, $\mu_{23} = 1 - \epsilon$ and $\mu_{13} = \mu_{12}\mu_{23} = \epsilon(1-\epsilon)$. Let $\widehat{\mu}_{12} = \mu_{12} + z_{12}$, $\widehat{\mu}_{23} = \mu_{23} + z_{23}$ and $\widehat{\mu}_{13} = \mu_{12}\mu_{23} + z_{13}$, where the fluctuations of z_{12}, z_{23} and z_{13} shrink as n grows. It is useful to imagine the typical fluctuations of z_{ij} to be on the order $\epsilon/10$.

Since $\max\{\mu_{12}, \mu_{13}\} = \max\{\epsilon\,,\, \epsilon(1-\epsilon)\} = \epsilon \ll \mu_{23}$, concentration bounds guarantee that with high probability $\widehat{\mu}_{23} > \max\{\widehat{\mu}_{12}\,,\, \widehat{\mu}_{13}\}$ and the (greedy implementation of) Chow-Liu algorithm described in (3.3) adds the edge $(2, 3)$ to T^{CL}. However,

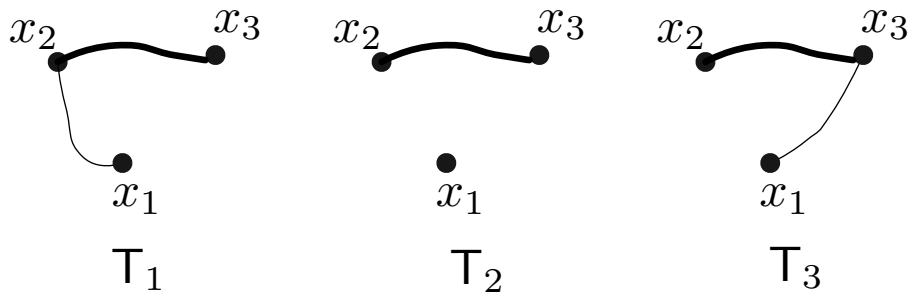

Figure 3.3.1: The original distribution factorizes according to T_1. The width of the edges indicate the coupling strength between the nodes. The forest approximation algorithm (defined later in Section 3.3.2) would recover $\widehat{\mathsf{T}} = \mathsf{T}_2$ as the correlation between node x_1 and all the other nodes is not strong enough to recover any edge confidently. Running Chow-Liu algorithm would recover T^{CL} to be T_1 or T_2 depending on the realization of the samples.

because $\mu_{12} - \mu_{13} = \epsilon^2$ is smaller than the fluctuations of z_{12} and z_{13} there is no guarantee that $\widehat{\mu}_{12} > \widehat{\mu}_{13}$. In particular, if $z_{13} - z_{12} > \epsilon^2$, then edge $(1, 3)$ is added and $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}_3$.

The preceding discussion provides the intuition underlying a statistical characterization of the possible errors made by the Chow-Liu algorithm. To make this quantitative, later on in the proof, we determine a value $\tau := \tau(n, \beta, \delta)$ so that any (strong) edge e with $|\mu_e| \geq \tau$ is recovered by the Chow-Liu algorithm with probability at least $1 - \delta$. Equivalently, if there is a mistake made by the Chow-Liu algorithm such that $e \in \mathcal{E}_{\mathsf{T}}$ but $e \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\mu_e| \leq \tau$ (i.e., missed edges are weak). This is going to play a key role in bounding the ssTV $\mathcal{L}^{(2)}$ for T^{CL}, whether or not it is equal to T.

In the regime where n is not large enough to guarantee the correct recovery of all the edges in the tree, there are two natural strategies:

I. *Forest approximation algorithm:* This algorithm attempts to recover a forest F which is a good approximation of the original tree T in the sense that $\mathcal{E}_{\mathsf{F}} \subseteq \mathcal{E}_{\mathsf{T}}$. This is accomplished by finding a forest consisting of sufficiently strong edges, i.e. having weight at least τ for an appropriate value of τ. As shown by Tan et al. in [24], such a forest can be obtained by running the Chow-Liu algorithm

and removing the edges with weight below τ. The details of this algorithm and its sample complexity will be discussed in Section 3.3.2.

II. *Chow-Liu Algorithm:* We can use the Chow-Liu tree as our estimated structure despite the fact that it may well be incorrect.

For our three-node example, the forest approximation algorithm would return $\widehat{\mathsf{T}} = \mathsf{T}_2$ in Figure 3.3.1, whereas Chow-Liu would give $\widehat{\mathsf{T}} = \mathsf{T}_1$ or $\widehat{\mathsf{T}} = \mathsf{T}_3$. To focus on the implication of graph structure estimation (as opposed to parameter estimation), we compare the *loss due to graph estimation error*, defined as $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(P))$, for the above cases:

$$\begin{aligned}
\widehat{\mathsf{T}} = \mathsf{T}_1 \quad &\rightarrow \quad \mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(P)) = 0\,, \\
\widehat{\mathsf{T}} = \mathsf{T}_2 \quad &\rightarrow \quad \mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(P)) = |\mu_{12}| = \epsilon\,, \\
\widehat{\mathsf{T}} = \mathsf{T}_3 \quad &\rightarrow \quad \mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(P)) = |\mu_{12}|(1 - \mu_{23}^2) = \epsilon^2(2 - \epsilon)\,.
\end{aligned}$$

Evidently, the loss due to graph estimation error in the forest approximation algorithm is bigger than the Chow-Liu algorithm, whether or not the latter recovers the true tree. This is because the Chow-Liu algorithm does not make arbitrary errors in estimating the tree: errors happen when both the original tree and the estimated tree describe the original distribution rather well. A formalization of this statement is given in Theorem 3.4.

3.3.2 Forest approximation

In the regime where exact recovery of the tree is impossible, a reasonable goal is to instead find a forest approximation to the tree. In this section we analyze a natural forest approximation algorithm, which thresholds to zero edges with correlation below a specified value τ.

There is extensive literature on estimating forests in the fully observed setting of this paper, including [24, 25]. Mossel in [34] studied the problem of learning phylogenetic forests, where samples are only observed at the leaves of the tree. They

quantified the idea that most edges of phylogenies are easy to reconstruct. In the regime that the sample complexity of structure learning is too high, they instead estimate a forest. An upper bound on the number of connected components in the forest is provided as a function of the number of leaves, the minimum edge weight, and the metric distortion bounds. Our results in this section are consistent with the asymptotic conditions on the thresholds given in [24] by Tan et al. for forest approximation of general distributions over trees.

The forest approximation algorithm considered in this section thresholds to zero the edges with correlations below $\tau = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ for $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. This is equivalent to finding the maximum-weight spanning forest over the complete graph with edge weights $|\widehat{\mu}_e| - \tau - \epsilon$. The output $\widehat{\mathsf{T}} = (\mathcal{V}, \mathcal{E}_{\widehat{\mathsf{T}}})$ is a truncated version of T^{CL} such that the empirical correlation between any pair of nodes $(i,j) \in \mathcal{E}_{\widehat{\mathsf{T}}}$ satisfies $|\widehat{\mu}_{ij}| \geq \tau + \epsilon$.

Proposition 3.8. *Given $n > C \max\{e^{2\beta}/\eta^2 \log \frac{p}{\delta}, 1/\eta^2 \log \frac{p}{\eta\delta}\}$ samples, the forest approximation algorithm guarantees that with probability $1-\delta$ we have $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$ and $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\widehat{P})) < \eta$.*

The proof is presented in the supplementary material. Comparing the above statement with Proposition 3.10 shows that in general the loss due to graph estimation using the truncation algorithm is (asymptotically) strictly greater than the loss incurred by the Chow-Liu algorithm, which continues adding edges that may be incorrect.

3.4 Outline of proof

We now sketch the argument for the main result, Theorem 3.4, guaranteeing accurate pairwise marginals in the Chow-Liu tree. The starting point is an application of the triangle inequality (3.5):

$$\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) \leq \mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) + \mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}))\,. \tag{3.7}$$

The first term on the right-hand side of Equation (3.7) represents the error due to the difference in the structure of T and T^{CL}, so we call this first term the *loss due to graph estimation error*. Equation (3.6) tells us that for each pair of nodes $u, v \in \mathcal{V}$, $\mathsf{path}_{\mathsf{T}}(u, v)$ and $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, v)$ must be compared.

The second term on the right-hand side of Equation (3.6) represents the propagation of error due to inaccuracy in estimated parameters. Recall that the estimated parameters on the Chow-Liu tree are obtained by matching correlations to the empirical values.

Theorem 3.4 follows by separately bounding each term on the right-hand side of Equation (3.7).

Proposition 3.9 (Loss due to parameter estimation error)**.** *Given*

$$n > C \max\{e^{2\beta} \log \frac{p}{\delta}, \eta^{-2} \log \frac{p}{\eta\delta}\}$$

samples, with probability at least $1 - \delta$ *we have* $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$.

Proposition 3.10 (Loss due to graph estimation error)**.** *Given*

$$n > C \max\{e^{2\beta},\, \eta^{-2}\} \log \frac{p}{\delta}$$

samples, with probability at least $1 - \delta$ *we have* $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) < \eta$.

These two propositions are proved in full detail in Sections 4.1 and 4.2. In the remainder of this section we define probabilistic events of interest and sketch the proofs. Proposition 3.9 has a relatively straightforward proof, while the major technical contribution is in Proposition 3.10. The latter uses a careful inductive argument to control the errors arising from computation of correlations using an incorrect tree.

We define three highly probable events $\mathtt{E}^{\text{corr}}(\epsilon), \mathtt{E}^{\text{strong}}(\epsilon)$ and $\mathtt{E}^{\text{cascade}}(\epsilon)$ as follows. Let $\mathtt{E}^{\text{corr}}(\epsilon)$ be the event that all empirical correlations are within ϵ of population values:

$$\mathtt{E}^{\text{corr}}(\epsilon) = \left\{ \max_{w, \widetilde{w} \in \mathcal{V}} |\mu_{w,\widetilde{w}} - \widehat{\mu}_{w,\widetilde{w}}| \leq \epsilon \right\} . \tag{3.8}$$

Let $\tau = 4\epsilon/\sqrt{1-\tanh\beta}$ and

$$\mathcal{E}_{\mathsf{T}}^{\text{strong}} = \{(i,j) \in \mathcal{E}_{\mathsf{T}} : |\tanh(\theta_{i,j})| \geq \tau\} \tag{3.9}$$

consist of the set of strong edges in tree T and let T^{CL} be the Chow-Liu tree defined in Equation (3.3). *Weak edges* are the ones that not strong, *i.e.*, $\mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}}^{\text{strong}}$. Let $\mathtt{E}^{\text{strong}}(\epsilon)$ be the event that all strong edges in T as defined in (3.9) are recovered using the Chow-Liu algorithm:

$$\mathtt{E}^{\text{strong}}(\epsilon) = \left\{\mathcal{E}_{\mathsf{T}}^{\text{strong}} \subset \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}\right\} . \tag{3.10}$$

Finally, define the event

$$\mathtt{E}^{\text{cascade}}(\epsilon) = \left\{\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}}(\widehat{P})) \leq \epsilon\right\} . \tag{3.11}$$

Recall that P factorizes according to T and $\widehat{P}$ is the empirical distribution which does not factorize according to any tree. This event controls *cascades* of errors in correlations computed along paths in T.

Lemma 3.11. *The following hold:*

$$\begin{aligned}
&\mathsf{P}[\mathtt{E}^{\text{corr}}(\epsilon)] \geq 1 - 2p^2 \exp(-n\epsilon^2/2) \\
&\mathsf{P}[\mathtt{E}^{\text{strong}}(\epsilon)] \geq 1 - 2p^2 \exp(-n\epsilon^2/2) \\
&\mathsf{P}[\mathtt{E}^{\text{cascade}}(\epsilon)] \geq 1 - 4p^2/\epsilon \exp(-\epsilon^2 n/32) .
\end{aligned}$$

This Lemma is the restatement of Lemmas 4.2, 4.8 and 4.11. Since we are interested in the situation that all three events hold, let

$$\mathtt{E}(\epsilon, \gamma) := \mathtt{E}^{\text{corr}}(\epsilon) \cap \mathtt{E}^{\text{strong}}(\epsilon) \cap \mathtt{E}^{\text{cascade}}(\gamma). \tag{3.12}$$

We use different values of γ and ϵ in $\mathtt{E}^{\text{cascade}}(\gamma)$ versus $\mathtt{E}^{\text{corr}}(\epsilon)$ and $\mathtt{E}^{\text{strong}}(\epsilon)$ because the bounds of Lemma 3.11 are different in nature for these events.

Sketch of proof of Proposition 3.9 The proof of Proposition 3.9 entails showing that on the event $\mathtt{E}(\epsilon, \gamma)$ (defined in (3.12)) with $\epsilon < \min\{\frac{\eta}{3}, e^{-\beta}/20\}$ and $\gamma \leq \eta/3$ we have the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$. Once we do this, we may apply Lemma 3.11 to see that the desired bound holds with probability at least $1-3\delta$ given $n > \max\{800e^{2\beta}\log\frac{4p^2}{\delta}, 300\eta^{-2}\log\frac{12p^2}{\eta\delta}\}$ samples, which gives the proposition.

To bound $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}))$ on event $\mathtt{E}(\epsilon, \gamma)$, we consider parameter estimation errors along paths in T^{CL}. First, observe that on event $\mathtt{E}^{\text{cascade}}(\gamma)$, according to (3.11), the end-to-end error for each path in $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \cap \mathcal{E}_{\mathsf{T}}$ is bounded by γ.

Next, we study parameter estimation error in paths containing (falsely added) edges, *i.e.*, those in $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \backslash \mathcal{E}_{\mathsf{T}}$. For any pair of nodes $w, \widetilde{w}$, denote by $t = |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) \backslash \mathcal{E}_{\mathsf{T}}|$ the number of falsely added edges in the path connecting them in T^{CL}. As discussed, later on in the proof, these edges correspond to missed edges in T and thus $\mathtt{E}^{\text{strong}}(\epsilon)$ guarantees that these edges are weak (as defined after (3.9)). These t weak edges break up the $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ into at most $t+1$ contiguous segments $\mathcal{F}_0, \mathcal{F}_1, \cdots, \mathcal{F}_t$, each entirely within $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \cap \mathcal{E}_{\mathsf{T}}$.

The error on each segment $\mathcal{F}_i$, is bounded by γ as already discussed, but now there are $t+1$ such segments and the errors add up. This effect is counterbalanced by the fact that the falsely added edges are weak and hence scale down the error in a multiplicative fashion. Therefore, the dependence on t of the error

$$\left| \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})} \widehat{\mu}_e - \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})} \mu_e \right|$$

has a linearly increasing factor γt and an exponentially decaying factor τ^t. If $\epsilon \leq e^{-\beta}/20$, then this gives a uniform upper bound of $3\max\{\epsilon, \gamma\}$ for this quantity for any t. Putting it all together shows that if $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$ and $\gamma \leq \eta/3$ then $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$ on the event $\mathtt{E}(\epsilon, \gamma)$.

Sketch of proof of Proposition 3.10 The goal is to show that the event $\mathtt{E}^{\text{corr}}(\epsilon) \cap \mathtt{E}^{\text{strong}}(\epsilon)$ with $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$ implies $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) < \eta$. Lemma 3.11 then implies that $n > 1200\max\{e^{2\beta}, \eta^{-2}\}\log\frac{2p^2}{\delta}$ suffices for $\mathtt{E}^{\text{corr}}(\epsilon) \cap \mathtt{E}^{\text{strong}}(\epsilon)$ to occur

with probability at least $1 - 2\delta$.

The proof of the proposition sets up a careful induction on the distance between nodes (computed in T^{CL}) for which we wish to bound the error in correlation. One of the ingredients is Lemma 4.12, a combinatorial statement relating trees T^{CL} and T. The lemma concerns any two spanning trees on p nodes and the paths between two given nodes w and $\widetilde{w}$. It proves the existence of at least one pair of corresponding edges $f \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ satisfying a collection of properties illustrated in Figure 4.2.1 (and specified in the lemma). A consequence is that the true correlation across g according to P can be expressed as a function of the correlation on f as $\mu_g = \mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{\mathcal{C}}}$, hence $|\mu_g| \leq |\mu_f|$. But $|\widehat{\mu}_g| \geq |\widehat{\mu}_f|$, since the Chow-Liu algorithm chose g in T^{CL} instead of f.

We obtain a recurrence for $\Delta(d)$, where $\Delta(d)$ provides an upper bound for the error due to graph estimation error for any pair of nodes $w, \widetilde{w}$ with $|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})| = d$. Using the relationship between the edges f and g, the paths $\mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$ and $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ are compared. It is proved that the recurrence $\Delta(d)$ increases linearly and decays exponentially with d. With $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$, it is shown that $\Delta(d)$ is uniformly upper bounded by η.

Chapter 4

Proofs of main results

As observed in Section 3.4, Theorem 3.3 is a direct consequence of Propositions 3.9 and 3.10, proved in Sections 4.1 and 4.2.

4.1 Loss due to parameter estimation (proof of Proposition 3.9)

By Lemma 3.11, the event $\mathtt{E}(\epsilon, \gamma)$ occurs with probability at least $1 - 3\delta$ provided

$$n > \max\{2/\epsilon^2 \log(4p^2/\delta), 32/\gamma^2 \log(4p^2/\gamma\delta)\}.$$

By assumption

$$n > \max\{800e^{2\beta} \log \frac{4p^2}{\delta}, 300\eta^{-2} \log \frac{12p^2}{\eta\delta}\},$$

hence we can take $\gamma \leq \eta/3$ and $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$. It remains to show that on the event $\mathtt{E}(\epsilon, \gamma)$ with $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$ and $\gamma \leq \eta/3$, $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$.

Let $\tau = 4\epsilon/\sqrt{1 - \tanh\beta}$ be the threshold used in (3.9) to define $\mathcal{E}_{\mathsf{T}}^{\text{strong}}$, the set of strong edges in T. For any pair of nodes $w, \widetilde{w}$, consider $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$. Let $0 \leq t < p$ be the number of weak edges $e_1, \cdots, e_t \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ such that $|\mu_{e_i}| < \tau$. There are at most $t + 1$ contiguous sub-paths in $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ consisting of strong edges. We call these segments $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_t$. If two weak edges e_i and e_{i+1} are adjacent in

$\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$, then $\mathcal{F}_i = \varnothing$, in which case we define $\mu_{\mathcal{F}_i} = \widehat{\mu}_{\mathcal{F}_i} = 1$. By definition of $\mathcal{F}_i$, all edges $f \in \mathcal{F}_i$ are strong.

Later, we prove Lemma 4.8 which shows that under the event $\mathtt{E}^{\text{strong}}(\epsilon)$ all strong edges in T are recovered in T^{CL}. Thus, $\mathcal{F}_i \subseteq \mathcal{E}_{\mathsf{T}}$ is a path not only in T^{CL} but also in T, which guarantees $|\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma$ under the event $\mathtt{E}^{\text{cascade}}(\gamma)$.

Note that if $t = 0$ then $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(x, \widetilde{w})$ consists of all strong edges for which Lemma 4.11 gives the desired bound. For $t \geq 1$ we have:

$$
\begin{aligned}
&\left| \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})} \widehat{\mu}_e - \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})} \mu_e \right| \\
&\overset{(a)}{=} \left| \widehat{\mu}_{\mathcal{F}_0} \prod_{i=1}^{t} \widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_0} \prod_{i=1}^{t} \mu_{\mathcal{F}_i} \mu_{e_i} \right| \\
&\overset{(b)}{\leq} |\widehat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \prod_{j=1}^{t} |\mu_{\mathcal{F}_j} \mu_{e_j}| \\
&\quad + \sum_{i=1}^{t} |\widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \cdot |\widehat{\mu}_{\mathcal{F}_0}| \prod_{j=1}^{i-1} |\widehat{\mu}_{\mathcal{F}_j} \widehat{\mu}_{e_j}| \prod_{k=i+1}^{t} |\mu_{\mathcal{F}_k} \mu_{e_k}| \\
&\overset{(c)}{\leq} \gamma \tau^t + (\tau + \epsilon)^{t-1} \sum_{i=1}^{t} |\widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \\
&\overset{(d)}{\leq} \gamma \tau^t + (\tau + \epsilon)^{t-1} \left[\sum_{i=1}^{t} |\mu_{\mathcal{F}_i} (\widehat{\mu}_{e_i} - \mu_{e_i})| + |\widehat{\mu}_{e_i} (\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i})| \right] \\
&\overset{(e)}{\leq} (\tau + \epsilon)^{t-1} (2t+1) \max\{\gamma, \epsilon\} \overset{(f)}{\leq} (\tau + \epsilon)^{t-1} (2t+1) \eta/3 \\
&\overset{(g)}{\leq} \left[\frac{4\epsilon}{\sqrt{1 - \tanh \beta}} + \epsilon \right]^{t-1} (2t+1)\eta/3 \overset{(h)}{\leq} [4\epsilon e^{\beta} + \epsilon]^{t-1} (2t+1)\eta/3 \\
&\overset{(i)}{\leq} \left(5\epsilon e^{\beta}\right)^{t-1} (2t+1)\eta/3 \overset{(j)}{\leq} \frac{2t+1}{4^{t-1}} \eta/3 \overset{(k)}{\leq} \eta \,.
\end{aligned}
$$

In (a) we use $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) = \{\mathcal{F}_0, e_1, \cdots, \mathcal{F}_t, e_t, \mathcal{F}_t\}$. (b) uses the bound

$$
|\prod_{i=1}^{t} a_i - \prod_{i=1}^{t} b_i| \leq \sum_{i=1}^{t} |a_i - b_i| \prod_{j=1}^{i-1} |a_j| \prod_{k=i+1}^{t} |b_k|
$$

obtained via telescoping sum and triangle inequality. In (c) we use $|\mu_{\mathcal{F}_i}| \leq 1$, $|\widehat{\mu}_{\mathcal{F}_i}| \leq 1$, $|\mu_{e_i}| \leq \tau$, $|\widehat{\mu}_{e_i}| \leq \tau + \epsilon$ under $\mathtt{E}^{\text{corr}}(\epsilon)$ and $|\widehat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \leq \gamma$ on $\mathtt{E}^{\text{cascade}}(\gamma)$. In (d)

triangle inequality is used. In (e), we use $|\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma$ on the event $\mathtt{E}^{\text{cascade}}(\gamma)$ and $|\widehat{\mu}_{e_i} - \mu_{e_i}| \leq \epsilon$ on the event $\mathtt{E}^{\text{corr}}(\epsilon)$. (f) is true under the assumption $\gamma, \epsilon \leq \eta/3$. (g) uses the definition of $\tau = 4\epsilon/\sqrt{1 - \tanh\beta}$. (h) and (i) use inequality $1 - \tanh\beta \geq e^{-2\beta}$ and $\beta \geq 0$. (j) uses the assumption $\epsilon \leq e^{-\beta}/20$. (k) holds for all $t \geq 1$. □

4.2 Loss due to graph estimation error (proof of Proposition 3.10)

Recall that P factorizes according to T. The error in correlation between any two variables $X_w, X_{\widetilde{w}}$ computed along the path in T^{CL} as compared to T is

$$\begin{aligned} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) &= \frac{1}{2}\left|\mathbb{E}_P X_w X_{\widetilde{w}} - \mathbb{E}_{\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)} X_w X_{\widetilde{w}}\right| \\ &= \frac{1}{2}\left|\prod_{e \in \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})} \mu_e - \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})} \mu_e\right| . \end{aligned} \tag{4.1}$$

Our goal is to bound $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) = \max_{w,\widetilde{w} \in \mathcal{V}} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$.

With

$$n > 1200 \max\{e^{2\beta}, \eta^{-2}\} \log(2p^2/\delta),$$

by Lemma 3.11, the event $\mathtt{E}^{\text{corr}}(\epsilon) \cap \mathtt{E}^{\text{strong}}(\epsilon)$ occurs with probability $1 - 2\delta$ for some

$$\epsilon < \min\{\eta/16, e^{-\beta}/24\}.$$

In the following we prove that with $\epsilon < \min\{\eta/16, e^{-\beta}/24\}$ on the event $\mathtt{E}^{\text{corr}}(\epsilon) \cap \mathtt{E}^{\text{strong}}(\epsilon)$, we have the inequality $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) < \eta$.

The core of the argument uses induction to derive a recurrence on the maximum of $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ in terms of the distance (as measured in T^{CL}) between the nodes $w, \widetilde{w}$. Define

$$\Delta(d) \triangleq \max_{\substack{w,\widetilde{w} \in \mathcal{V} \\ |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})| = d}} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) .$$

For nodes at distance one in T^{CL}, i.e. $|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})| = 1$, it follows that

$$\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) = 0$$

from the definition of the projected distribution $\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)$ (matching pairwise marginals across edges). Hence, $\Delta(1) = 0 \leq \eta$, and we define $\Delta(0) = 0$. For $d > 1$, we want to bound $\Delta(d)$ in terms of $\Delta(k)$ for $k < d$. Then we will show that on the event $\mathtt{E}^{\mathrm{corr}}(\epsilon) \cap \mathtt{E}^{\mathrm{strong}}(\epsilon)$ with $\epsilon < \min\{\eta/16, e^{-\beta}/24\}$, if $\Delta(k) \leq \eta$ for all $k < d$, then $\Delta(d) < \eta$ which gives the result.

Note that if $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) = \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$, then $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) = 0$ (again because correlations are matched across edges). Thus, we assume $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) \neq \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$. Lemma 4.12 in Section 4.7 shows the existence of a pair of edges $f = (u, \widetilde{u}) \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g = (v, \widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ with the following properties (See Figure 4.2.1):

- $f \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, \widetilde{u})$.
- $f \notin \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ and $g \notin \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$.
- $u, v \in \mathsf{SubTree}_{\mathsf{T},f}(w)$ and $\widetilde{u}, \widetilde{v} \in \mathsf{SubTree}_{\mathsf{T},f}(\widetilde{w})$.

Here $\mathsf{SubTree}_{\mathsf{T},f}(w) = \{i \in \mathcal{V}; f \notin \mathsf{path}_{\mathsf{T}}(w, i)\}$ is the set of nodes connected to w in T after removing edge f (see Figure 4.2.1).

We define several sub-paths:

$$\mathcal{A} = \mathsf{path}_{\mathsf{T}}(u, w) \cap \mathsf{path}_{\mathsf{T}}(u, v)\,, \qquad \mathcal{B} = \mathsf{path}_{\mathsf{T}}(u, w) \setminus \mathsf{path}_{\mathsf{T}}(u, v)\,,$$
$$\mathcal{C} = \mathsf{path}_{\mathsf{T}}(u, v) \setminus \mathsf{path}_{\mathsf{T}}(u, w)\,, \qquad \mathcal{D} = \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, v)\,.$$

Recall that for set of edges $\mathcal{S}$, we defined $\mu_{\mathcal{S}} = \prod_{e \in \mathcal{S}} \mu_e$. Since $\mathsf{path}_{\mathsf{T}}(w, v) = \mathcal{B} \cup \mathcal{C}$ and $\mathcal{B} \cap \mathcal{C} = \varnothing$, we have $\mu_{w,v} = \mu_{\mathcal{B}}\mu_{\mathcal{C}}$. Similarly, in $\mathsf{SubTree}_{\mathsf{T},f}(\widetilde{w})$ we define

$$\widetilde{\mathcal{A}} = \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{v})\,, \qquad \widetilde{\mathcal{B}} = \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{w}) \setminus \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{v})\,,$$
$$\widetilde{\mathcal{C}} = \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{v}) \setminus \mathsf{path}_{\mathsf{T}}(\widetilde{u}, \widetilde{w})\,, \qquad \widetilde{\mathcal{D}} = \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(\widetilde{w}, \widetilde{v})\,.$$

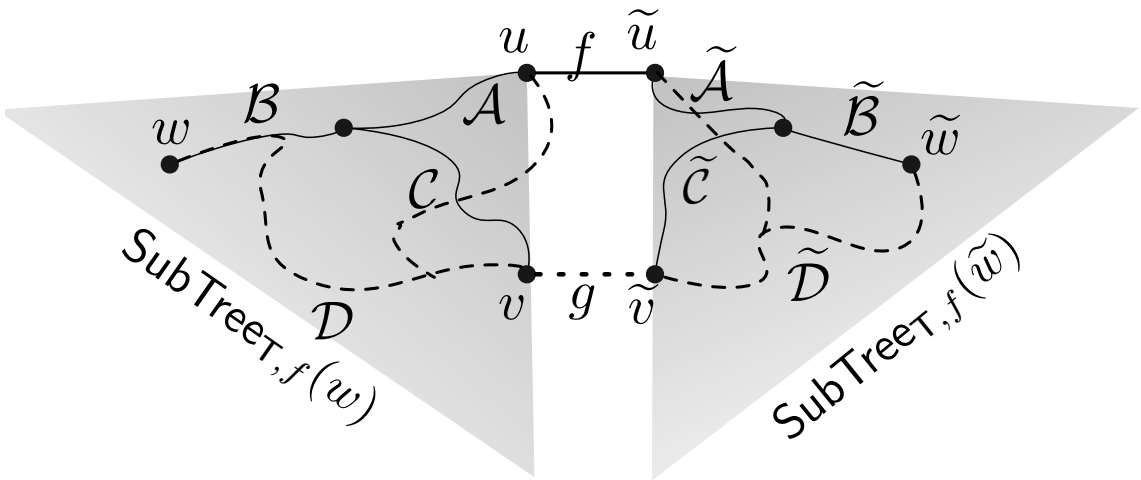

Figure 4.2.1: Schematic for the proof of Proposition 3.10. The solid lines represent paths in T and the dashed lines represent paths in T^{CL}. Edge $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, while $g \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$. The sets of edges $\mathcal{D}$ and $\widetilde{\mathcal{D}}$ may overlap with $\mathcal{A} \cup \mathcal{B}$ and $\widetilde{\mathcal{A}} \cup \widetilde{\mathcal{B}}$.

The sets are defined so that $\mathsf{path}_{\mathsf{T}}(v, \widetilde{v}) = \mathcal{C} \cup \mathcal{A} \cup \{f\} \cup \widetilde{\mathcal{C}} \cup \widetilde{\mathcal{A}}$ where $f = (u, \widetilde{u})$ and $g = (v, \widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$. Thus, $\mu_g = \mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{\mathcal{C}}}$. Since $\mathsf{path}_{\mathsf{T}}(w, \widetilde{w}) = \mathcal{A} \cup \mathcal{B} \cup \{f\} \cup \widetilde{\mathcal{A}} \cup \widetilde{\mathcal{B}}$ and $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) = \mathcal{D} \cup \{g\} \cup \widetilde{\mathcal{D}}$, our goal amounts to finding an upper bound for the quantity $|\mu_{\mathcal{D}} \mu_g \mu_{\widetilde{\mathcal{D}}} - \mu_{\mathcal{A}} \mu_{\mathcal{B}} \mu_f \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{\mathcal{B}}}|$.

Lemma 4.3 applied to $f = (u, \widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g = (v, \widetilde{v}) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, \widetilde{u})$ gives $|\widehat{\mu}_f| \leq |\widehat{\mu}_g|$. Also, $f \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$, hence $|\mu_g| \leq |\mu_f|$. On the event $\mathtt{E}^{\mathrm{corr}}(\epsilon)$,

$$|\mu_f| - 2\epsilon \leq |\widehat{\mu}_f| \leq |\widehat{\mu}_g| \leq |\mu_g| + 2\epsilon \leq |\mu_f| + 2\epsilon$$

which gives $|\mu_f| - 4\epsilon \leq |\mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{\mathcal{C}}}| \leq |\mu_f|$. Thus,

$$|\mu_f| \left(1 - |\mu_{\mathcal{C}}^2 \mu_{\widetilde{\mathcal{C}}}^2|\right) \leq 2|\mu_f| \left(1 - |\mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{\mathcal{C}}}|\right) \leq 8\epsilon\,. \tag{4.2}$$

Since $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, under the event $\mathtt{E}^{\mathrm{strong}}(\epsilon)$, f cannot be a strong edge as defined in Equation (3.9). It follows that $|\mu_f| \leq \tau$ for $\tau = 4\epsilon/\sqrt{1 - \tanh \beta}$.

Let $k = |\mathcal{D}| = |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, v)|$ and $\widetilde{k} = |\widetilde{\mathcal{D}}| = |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(\widetilde{w}, \widetilde{v})|$, so $d = k + \widetilde{k} + 1$. By definition of $\Delta(\cdot)$,

$$\begin{aligned} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, v) &= |\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}}| \leq \Delta(k)\,, \\ \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(\widetilde{w}, \widetilde{v}) &= |\mu_{\widetilde{\mathcal{D}}} - \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}| \leq \Delta(\widetilde{k})\,. \end{aligned} \tag{4.3}$$

We now prove that $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) \leq \eta$ assuming inductively that $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(i, j) \leq \eta$ for all pairs i, j such that $\mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(i, j) + 1 \leq d = \mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ (here, $\mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(i, j)$ denotes graph distance between i and j in T^{CL}). Using $\mu_g = \mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}$ and subsequently Equation (4.3),

$$\begin{aligned}
&\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w}) \\
&= |\mu_{\mathcal{D}}\mu_g\mu_{\widetilde{\mathcal{D}}} - \mu_{\mathcal{A}}\mu_{\mathcal{B}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{B}}}| \\
&= |\mu_{\mathcal{D}}\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}\mu_{\widetilde{\mathcal{D}}} - \mu_{\mathcal{A}}\mu_{\mathcal{B}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{B}}}| \\
&= |\mu_{\mathcal{A}}\mu_f\mu_{\widetilde{\mathcal{A}}}| \cdot |\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}} + \mu_{\mathcal{B}}\mu_{\mathcal{C}})(\mu_{\widetilde{\mathcal{D}}} - \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}} + \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}) - \mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}| \\
&\leq |\mu_{\mathcal{A}}\mu_f\mu_{\widetilde{\mathcal{A}}}| \cdot \big[|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\mathcal{B}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}} - \mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}| + |\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})| \\
&\qquad + |\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})| + |\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\mathcal{B}}\mu_{\mathcal{C}}(\mu_{\widetilde{\mathcal{D}}} - \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}})|\big] \\
&\overset{(a)}{\leq} |\mu_f\mu_{\mathcal{A}}\mu_{\widetilde{\mathcal{A}}}\mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}||\mu_{\mathcal{C}}^2\mu_{\widetilde{\mathcal{C}}}^2 - 1| + |\mu_f|\left(\Delta(k)\Delta(\widetilde{k}) + \Delta(k) + \Delta(\widetilde{k})\right) \\
&\overset{(b)}{\leq} 8\epsilon + \frac{4\epsilon}{\sqrt{1 - \tanh\beta}}\left(\Delta(k) + \Delta(\widetilde{k}) + \Delta(k)\Delta(\widetilde{k})\right) \\
&\overset{(c)}{\leq} 8\epsilon + 4\epsilon e^{\beta}(2\eta + \eta^2) \leq \eta\,.
\end{aligned}$$

Inequality (a) follows from (4.3). Inequality (b) uses Equation (4.2) and $|\mu_f| \leq \tau$. We showed that $\Delta(1) \leq \eta$. In (c) we use the inductive assumption $\Delta(k) \leq \eta$ for all $k < d$ and the assumption $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$. Since w and $\widetilde{w}$ were arbitrary, this proves $\Delta(d) \leq \eta$, and moreover this holds for all d. $\square$

4.3 Upper bound for risk (proof of Theorem 3.6)

We provide a proof for Theorem 3.6 on the risk, which is a direct application of Theorem 3.4. By Theorem 3.4, for any $0 \geq \eta \geq 1$, given n samples, $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) \geq \eta$ with probability at most $p\exp(-Cne^{-2\beta}) + p/\eta \exp(-Cn\eta^2)$. Hence,

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})\right] &= \int_0^1 \mathsf{P}\left[\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}) \geq \eta\right] d\eta \\
&\overset{(a)}{\leq} \int_0^1 \min\{1, p\exp(-Cne^{-2\beta}) + \frac{p}{\eta}\exp(-Cn\eta^2)\}\, d\eta \\
&\overset{(b)}{\leq} \eta^* + \int_{\eta^*}^1 p\exp(-Cne^{-2\beta}) + \frac{p}{\eta}\exp(-Cn\eta^2)\, d\eta \\
&\leq \eta^* + p\exp(-Cne^{-2\beta}) + \frac{p}{\eta^*}\int_{\eta^*}^{\infty} \exp(-Cn\eta^2)\, d\eta \\
&\overset{(c)}{\leq} \eta^* + p\exp(-Cne^{-2\beta}) + \frac{p}{\eta^*}\sqrt{\frac{\pi}{Cn}}\exp(-Cn{\eta^*}^2/2) \\
&\overset{(d)}{\leq} 3\eta^*,
\end{aligned}
$$

where (a) is a direct application of Theorem 3.4. Inequality (b) is true for any $0 < \eta^* \leq 1$. (c) uses the inequality $1/\sqrt{2\pi}\int_a^{\infty}\exp(-x^2/2)\,dx \leq \exp(-a^2/2)$. (d) holds for

$$\eta^* = p\exp(-Cne^{-2\beta}) + C'\sqrt{\frac{\log n + \log p}{n}}$$

with some numerical constant C'. To see that, note that $\eta^* \geq p\exp(-Cne^{-2\beta})$. Also, $\eta^* \geq C'\sqrt{\frac{\log n + \log p}{n}}$ gives

$$\frac{p}{\eta^*}\sqrt{\frac{\pi}{Cn}}\exp(-Cn{\eta^*}^2/2) \leq \frac{p}{C'(np)^{CC'^2/2}}\sqrt{\frac{\pi}{C\log(np)}} \leq C'\sqrt{\frac{\log(np)}{n}} \leq \eta^*$$

with proper choice of C'.

4.4 Necessary samples for accurate pairwise marginals (proof of Theorem 3.5)

We construct a family of trees that are difficult to distinguish from one another. Applying the version of Fano's inequality below in (4.5), (which can be found, for example, as Corollary 2.6 in [41]) gives a lower bound on the error probability. The bound on the sample complexity is in terms of the KL-divergence between pairs of points in the parameter space, where KL-divergence between two distributions P and Q on a space $\mathcal{X}$ is defined as $D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}$. The symmetrized KL-divergence between two zero-field Ising models with parameters θ and θ' has the convenient form

$$J(\theta\|\theta') \triangleq D(\theta\|\theta') + D(\theta'\|\theta) = \sum_{i<j}(\theta_{ij} - \theta'_{ij})(\mu_{ij} - \mu'_{ij})\,. \tag{4.4}$$

Here μ_{ij} and μ'_{ij} are the pairwise correlations between nodes i and j computed according to θ and θ', respectively.

Lemma 4.1 (Fano's inequality)**.** *Assume that $M \geq 2$ and that Θ is a family of models $\theta^0, \theta^1, \ldots, \theta^M$. Let Q_{θ_j} denote the probability law of the observation X under model θ_j. Given n i.i.d. samples from a model in the family Θ, if*

$$n < (1-\delta)\frac{\log M}{\frac{1}{M+1}\sum_{j=1}^{M} J(Q_{\theta^j}\|Q_{\theta^0})}\,, \tag{4.5}$$

then the probability of error for any algorithm is bounded as $p_e \geq \delta - \frac{1}{\log M}$.

Theorem 3.5. We consider a fixed tree structure given by a path, or in other words a Markov chain $x_1 \leftrightarrow x_2 \leftrightarrow \cdots \leftrightarrow x_p$. We choose $M+1$ different parameter vectors θ^m, $0 \leq m \leq M$, for $M = (p-1)\lfloor\frac{\tanh\beta - \tanh\alpha}{\eta}\rfloor + 1$. Given $\tanh\beta - \tanh\alpha \geq \sqrt{\eta/p}$, we have $\log M \geq \frac{1}{3}\log(p/\eta)$.

Let $\theta^0_{i,i+1} = \alpha$ for $i = 1, \cdots, p-1$. In the m-th model for $m = k\lfloor\frac{\tanh\beta-\tanh\alpha}{\eta}\rfloor + l - 1$ with $1 \leq k \leq p-1$ and $1 \leq l \leq \frac{\tanh\beta-\tanh\alpha}{\eta}$, we have $\theta^m_{k,k+1} = \operatorname{atanh}(\tanh\alpha + l\eta)$ and the remaining edge weights $\theta^m_{i,i+1} = \alpha$ for $i \neq k$. Thus, the number of models we try

to discriminate is $M \geq (p-1)\lfloor \frac{\tanh\beta - \tanh\alpha}{\eta} \rfloor$. For $m' \neq m$,

$$\max_{i,j} |\mathbb{E}_{\theta^{m'}}[X_i X_j] - \mathbb{E}_{\theta^m}[X_i X_j]| \geq \eta \quad \text{and} \quad J(\theta^m \| \theta^{m'}) \geq \eta \operatorname{atanh}(\eta)\,.$$

Fano's inequality (Lemma 4.1) gives the bound. □

4.5 Sample complexity for exact structure learning problem

4.5.1 Samples necessary for exact structure learning (proof of Theorem 3.2)

Suppose that p is odd (for simplicity) and let the graph T_0 be a path with associated parameters θ^0 given by $\theta^0_{i,i+1} = \alpha$ for odd values of i and $\theta^0_{i,i+1} = \beta$ for even values of i. For each odd value of $m \leq p-2$ we let θ^m be equal to θ^0 everywhere except $\theta^m_{m,m+1} = 0$ and $\theta^m_{m,m+2} = \alpha$. There are $(p+1)/2$ models in total (including θ^0). A small calculation using (4.4) leads to

$$J(\theta^0 \| \theta^m) = 2\alpha \tanh\alpha (1 - \tanh\beta) = 2\alpha \tanh\alpha \frac{2e^{-\beta}}{e^{\beta} + e^{-\beta}} \leq 4\alpha^2 e^{-2\beta}\,. \tag{4.6}$$

Here we used the inequality $\tanh\alpha \leq \alpha$ for $\alpha \geq 0$. Equation (4.6) can be plugged into Fano's inequality to complete the proof of Theorem 3.2.

4.5.2 Sufficient samples for structure learning (proof of Theorem 3.3)

Consider the original tree T with parameters $\alpha \leq |\theta_{ij}| \leq \beta$ for $(i,j) \in \mathcal{E}_{\mathsf{T}}$. Using Definition 3.10, the Chow-Liu algorithm recovers strong edges under the event $\mathsf{E}^{\text{strong}}(\epsilon)$, where edge (i,j) is strong if its parameter θ_{ij} satisfied $|\tanh\theta_{ij}| \geq 4\epsilon/\sqrt{1-\tanh\beta} = \tau$. Thus, if the edge parameter lower bound α in the original tree T satisfies $\tanh\alpha \geq \tau$, we can guarantee that with high probability $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$. Note that by Lemma 4.8,

the event $\mathtt{E}^{\text{strong}}(\epsilon)$ defined in Equation (3.10) occurs with probability at least $1-\delta$ for $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$.

The following bound on the number of samples guarantees this:

$$n > \frac{16}{\tanh^2(\alpha)(1-\tanh\beta)} \log \frac{2p^2}{\delta}.$$

As $\frac{1}{1-\tanh\beta} \leq e^{\beta}$, $n > 6\tanh^{-2}(\alpha)e^{2\beta}\log\frac{p}{\delta}$ guarantees $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$ with probability at least $1-\delta$. □

4.6 Control of events $\mathtt{E}^{\text{corr}}, \mathtt{E}^{\text{strong}}$ and $\mathtt{E}^{\text{cascade}}$ (proof of Lemma 3.11)

We state a standard form of Hoeffding's inequality [42] in Lemma A.1

Lemma 4.2. *The event* $\mathtt{E}^{\text{corr}}(\epsilon)$ *defined in* (3.8) *occurs with probability at least* $1 - 2p^2\exp(-n\epsilon^2/2)$.

Proof. For a given pair of nodes $w, \widetilde{w}$, we define $Z^{(i)} = X_w^{(i)} X_{\widetilde{w}}^{(i)}$ and apply Hoeffding's inequality (Lemma A.1) to get

$$\mathsf{P}\left[|\mu_{w,\widetilde{w}} - \widehat{\mu}_{w,\widetilde{w}}| > \epsilon\right] \leq 2\exp(-n\epsilon^2/2).$$

Applying the union bound over $\binom{p}{2}$ pairs $w, \widetilde{w} \in \mathcal{V}$ of nodes completes the proof. □

The following lemma is a well-known consequence of max-weight spanning tree algorithm [43] used to construct the Chow-Liu tree.

Lemma 4.3 (Error characterization in the Chow-Liu tree)**.** *Consider the complete graph on p nodes with weights $|\widehat{\mu}_{ij}|$ on each edge (i,j). Let T^{CL} be the maximum weight spanning tree of this graph. If edge $(u,\widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\widehat{\mu}_{u\widetilde{u}}| \leq |\widehat{\mu}_{ij}|$ for all $(i,j) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$.*

Proof. We include a proof for completeness. For edge $(u,\widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, if there is an edge $(i,j) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$ such that $|\widehat{\mu}_{u\widetilde{u}}| > |\widehat{\mu}_{ij}|$, then T^{CL} cannot be the maximum weight

spanning tree. To show that, consider the tree T' identical to T^{CL} except $(u,\widetilde{u}) \in \mathcal{E}_{\mathsf{T}'}$ and $(i,j) \notin \mathcal{E}_{\mathsf{T}'}$ (*i.e.*, $\mathcal{E}_{\mathsf{T}'} = (\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \{(i,j)\}) \cup \{(u,\widetilde{u})\}$.) Note that T' is a spanning tree and observe that:

$$\text{weight}(\mathsf{T}') \triangleq \sum_{e \in \mathcal{E}_{\mathsf{T}'}} |\widehat{\mu}_e| = \sum_{e \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}} |\widehat{\mu}_e| + |\widehat{\mu}_{u\widetilde{u}}| - |\widehat{\mu}_{ij}| > \text{weight}(\mathsf{T}^{\mathsf{CL}}).$$

□

We define a pair of random variables that will help to characterize the mistakes made by the Chow-Liu algorithm. For a given pair of nodes $v, \widetilde{v}$ and edge $f = (u,\widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$, let

$$Z_{f,v,\widetilde{v}} = X_u X_{\widetilde{u}} - X_v X_{\widetilde{v}} = X_u X_{\widetilde{u}} \left(1 - X_u X_v X_{\widetilde{v}} X_{\widetilde{u}}\right), \tag{4.7}$$

$$Y_{f,v,\widetilde{v}} = X_u X_{\widetilde{u}} + X_v X_{\widetilde{v}} = X_u X_{\widetilde{u}} \left(1 + X_u X_v X_{\widetilde{v}} X_{\widetilde{u}}\right). \tag{4.8}$$

These variables are used to study the event $\{f = (u,\widetilde{u}) \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}, (v,\widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}, f \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$ and $(v,\widetilde{v}) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})\}$. Applying Lemma 4.3, this happens only if $|\widehat{\mu}_{v\widetilde{v}}| \geq |\widehat{\mu}_f|$.

Lemma 4.4. *If there exists a pair of edges $f = (u,\widetilde{u})$ and $g = (v,\widetilde{v})$ such that $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, $g \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ and additionally $f \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$, then,*

$$\left(\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)}\right) \leq 0$$

Proof. Using Lemma 4.3, $f = (u,\widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g = (v,\widetilde{v}) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$ gives $|\widehat{\mu}_g| \geq |\widehat{\mu}_f|$. Hence $|\widehat{\mu}_g|^2 \geq |\widehat{\mu}_f|^2$ and

$$\begin{aligned}
0 &\geq |\widehat{\mu}_f|^2 - |\widehat{\mu}_g|^2 = (\widehat{\mu}_f - \widehat{\mu}_g)(\widehat{\mu}_f + \widehat{\mu}_g) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} X_u^{(i)} X_{\widetilde{u}}^{(i)} - X_v^{(i)} X_{\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^{n} X_u^{(i)} X_{\widetilde{u}}^{(i)} + X_v^{(i)} X_{\widetilde{v}}^{(i)}\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)}\right)
\end{aligned}$$

where in the last equality we used $f \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ and the definition of random variables $Z_{f,w,\widetilde{w}}$ and $Y_{f,w,\widetilde{w}}$ given in Equations (4.7) and (4.8). $\square$

Later on, to bound the probability of this event, deviation bounds on the random variables $Z_{f,v,\widetilde{v}}$ and $Y_{f,v,\widetilde{v}}$ are provided in Lemmas 4.5 and 4.6. To do so, we use the standard Bernstein's inequality quoted from [41] in Lemma A.2

Lemma 4.5. *For all pairs of nodes $v, \widetilde{v} \in \mathcal{V}$ and edges $f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$, given n i.i.d. samples $Z^{(1)}_{f,v,\widetilde{v}}, Z^{(2)}_{f,v,\widetilde{v}}, \cdots, Z^{(n)}_{f,v,\widetilde{v}}$ defined in (4.7), with probability at least $1 - \delta/2$*

$$\left| \sum_{i=1}^{n} Z^{(i)}_{f,v,\widetilde{v}} - n\mu_f(1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}) \right| \leq \max\left\{ 4n\epsilon^2, 4n\epsilon\sqrt{1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}} \right\}, \tag{4.9}$$

where $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$ and $\mathcal{A}_{f,v,\widetilde{v}} = \mathsf{path}_{\mathsf{T}}(v, \widetilde{v}) \setminus \{f\}$ so that $\mu_{v\widetilde{v}} = \mu_f \mu_{\mathcal{A}_{f,v,\widetilde{v}}}$.

Proof. We use the abbreviation $\mathcal{A}$ instead of $\mathcal{A}_{f,v,\widetilde{v}}$ in this proof. Applying Lemma 4.9, it follows from the fact that P is Markov with respect to T and $f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ that $X_u X_{\widetilde{u}}$, $X_v X_u$ and $X_{\widetilde{u}} X_{\widetilde{v}}$ are statistically independent random variables. Note that

$$\mu_f = \mathbb{E} X_u X_{\widetilde{u}} = P(X_u X_{\widetilde{u}} = +1) - P(X_u X_{\widetilde{u}} = -1) = 2P(X_u X_{\widetilde{u}} = +1) - 1.$$

Hence,

$$P(X_u X_{\widetilde{u}} = 1) = \frac{1 + \mu_f}{2} \quad \text{and} \quad P(X_u X_{\widetilde{u}} = -1) = \frac{1 - \mu_f}{2}.$$

Similarly, the distribution of $X_v X_u X_{\widetilde{u}} X_{\widetilde{v}}$ is function of $\mu_{\mathcal{A}}$. As a result, the random variable $Z_{f,v,\widetilde{v}} \in \{-2, 0, 2\}$ defined in (4.7) has the following distribution:

$$Z_{f,v,\widetilde{v}} = \begin{cases} -2 & \text{w.p.} \quad \frac{1-\mu_f}{2} \frac{1-\mu_{\mathcal{A}}}{2} \\ 0 & \text{w.p.} \quad \frac{1+\mu_{\mathcal{A}}}{2} \\ +2 & \text{w.p.} \quad \frac{1+\mu_f}{2} \frac{1-\mu_{\mathcal{A}}}{2}. \end{cases}$$

The first and second moments of $Z_{f,v,\widetilde{v}}$ are $\mathbb{E}[Z_{f,v,\widetilde{v}}] = \mu_f(1-\mu_{\mathcal{A}})$ and

$$\mathrm{Var}[Z_{f,v,\widetilde{v}}] = (1-\mu_{\mathcal{A}})[2-\mu_f^2(1-\mu_{\mathcal{A}})] \leq 2(1-\mu_{\mathcal{A}}).$$

By Bernstein's inequality (Lemma A.2), with probability at least $1-\delta/2$

$$\left|\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)} - n\mathbb{E}[Z_{f,v,\widetilde{v}}]\right| \leq n \max\left\{\frac{8}{3n}\log(\frac{4}{\delta}), \sqrt{\frac{4\mathrm{Var}[Z_{f,v,\widetilde{v}}]}{n}\log(\frac{4}{\delta})}\right\}.$$

Using a union bound, we show that for any pair of nodes $v, \widetilde{v}$ and any edge $f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$,

$$\left|\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)} - n\mu_f(1-\mu_{\mathcal{A}})\right| \leq n \max\left\{\frac{8}{3n}\log(\frac{4p^3}{\delta}), \sqrt{\frac{8(1-\mu_{\mathcal{A}})}{n}\log(\frac{4p^3}{\delta})}\right\}.$$

The definition of ϵ gives $\frac{8}{3n}\log(\frac{4p^3}{\delta}) \leq 4\epsilon^2$ and $\sqrt{\frac{8(1-\mu_{\mathcal{A}})}{n}\log(\frac{4p^3}{\delta})} \leq 4\epsilon\sqrt{1-\mu_{\mathcal{A}}}$ which gives the lemma. □

The proof of the following lemma is analogous to that of Lemma 4.5.

Lemma 4.6. *Given n samples, let $\epsilon = \sqrt{2/n\log(2p^2/\delta)}$. Then, with probability at least $1-\delta/2$, for all $v, \widetilde{v} \in \mathcal{V}$, $f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$, and $Y_{f,v,\widetilde{v}}^{(1)}, Y_{f,v,\widetilde{v}}^{(2)}, \cdots, Y_{f,v,\widetilde{v}}^{(n)}$ defined in* (4.8),

$$\left|\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)} - n\mu_f(1+\mu_{\mathcal{A}_{f,v,\widetilde{v}}})\right| \leq \max\left\{4n\epsilon^2, 4n\epsilon\sqrt{1+\mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\right\}. \tag{4.10}$$

Event $\mathtt{E}^{\mathrm{strong}}(\epsilon)$ in (3.10) occurs if all of the strong edges in T (defined in (3.9)) are recovered in T^{CL}. In Lemma 4.7 we show that the deviation bounds for the variables $Z_{f,v,\widetilde{v}}$ and $Y_{f,v,\widetilde{v}}$ stated in (4.9) and (4.10) imply that $\mathtt{E}^{\mathrm{strong}}(\epsilon)$ holds.

Lemma 4.7. *Under the events described in Lemmas 4.5 and 4.6, if there is an edge $f \in \mathcal{E}_{\mathsf{T}}$ missing from the Chow-Liu tree, $f \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\mu_f| \leq \tau = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ (i.e., $\mathtt{E}^{\mathrm{strong}}(\epsilon)$ defined in Equation* (3.10) *holds).*

Proof. Applying Lemma 4.12 to $f = (u, \widetilde{u})$ shows that for the edge $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, there

exists an edge $g = (v, \widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ such that, $f \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, \widetilde{u})$ (Figure 4.7.1). Applying Lemma 4.4, this implies that

$$\left(\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)}\right) \leq 0.$$

Note that $\mathbb{E}Z_{f,v,\widetilde{v}} = \mu_f(1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}})$ and $\mathbb{E}Y_{f,v,\widetilde{v}} = \mu_f(1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}})$ using the definition $\mathcal{A}_{f,v,\widetilde{v}} = \mathsf{path}_{\mathsf{T}}(v, \widetilde{v}) \setminus \{f\}$. Hence $\mathbb{E}Z_{f,v,\widetilde{v}}\,\mathbb{E}Y_{f,v,\widetilde{v}} = \mu_f^2(1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}^2) \geq 0$. Thus,

$$\left(\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)}\right) < 0$$

holds only if either one of the following inequalities holds:

$$\begin{aligned} &\Big|\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)} - n\mathbb{E}Z_{f,v,\widetilde{v}}\Big| \geq n|\mathbb{E}Z_{f,v,\widetilde{v}}| \qquad \text{or} \\ &\Big|\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)} - n\mathbb{E}Y_{f,v,\widetilde{v}}\Big| \geq n|\mathbb{E}Y_{f,v,\widetilde{v}}|\,. \end{aligned}$$

On the events described in Lemmas 4.5 and 4.6, there is an upper bound for $\big|\sum_{i=1}^{n} Z_{f,v,\widetilde{v}}^{(i)} - n\mathbb{E}Z_{f,v,\widetilde{v}}\big|$ and $\big|\sum_{i=1}^{n} Y_{f,v,\widetilde{v}}^{(i)} - n\mathbb{E}Y_{f,v,\widetilde{v}}\big|$. Hence, on these events, the property in above display holds only if either one of these inequalities holds:

$$\begin{aligned} &|\mu_f(1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}})| \leq \max\left\{4\epsilon^2, 4\epsilon\sqrt{1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\right\} \qquad \text{or} \\ &|\mu_f(1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}})| \leq \max\left\{4\epsilon^2, 4\epsilon\sqrt{1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\right\}\,, \end{aligned}$$

which is true if

$$\begin{aligned} |\mu_f| &\leq \max\left\{\frac{4\epsilon}{\sqrt{1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}}, \frac{4\epsilon^2}{1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}, \frac{4\epsilon}{\sqrt{1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}}, \frac{4\epsilon^2}{1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\right\} \\ &\leq \max\left\{\frac{4\epsilon}{\sqrt{1 - \tanh\beta}}, \frac{4\epsilon^2}{1 - \tanh\beta}\right\} = \max\{\tau, \tau^2\}\,, \end{aligned}$$

where τ is defined to be $\tau = 4\epsilon/\sqrt{1 - \tanh\beta}$. Note that if $\tau \geq 1$ then the bound on $|\mu_f| \leq 1 \leq \tau$ is trivial. If $\tau < 1$, then $\tau^2 < \tau$ which gives $|\mu_f| \leq \tau$. $\square$

Lemma 4.8. *With* $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$*, event* $\mathtt{E}^{\text{strong}}(\epsilon)$ *defined in* (3.10) *occurs with probability at least* $1 - \delta$*.*

Proof. Lemma 4.7 shows that, under the events described in Lemmas 4.5 and 4.6, the event $\mathtt{E}^{\text{strong}}(\epsilon)$ defined in Equation (3.10) holds. Lemmas 4.5 and 4.6 show that these events occur with probability at least $1 - \delta$. Thus, with probability at least $1 - \delta$, all edges $e \in \mathsf{T}$ such that $|\mu_e| \geq \tau$ are recovered by Chow-Liu algorithm $e \in \mathsf{T}^{\mathsf{CL}}$. □

Lemma 4.9. *Let the distribution* $P(x) \in \mathcal{P}(\mathsf{T})$ *be a zero-field Ising model on the tree* $\mathsf{T} = (\mathcal{V}, \mathcal{E})$*. For all* $e = (i, j) \in \mathcal{E}$ *let* $Y_e = X_i X_j$*. Then the set of random variables* Y_e *is jointly independent.*

This follows from the factorization of distribution $P(x) \in \mathcal{P}(\mathsf{T})$ in Equation (2.2).

Since edge-parities are independent random variables, the following lemma is applicable on paths in the tree T to give upper bound on the end-to-end error. Interestingly, this provides a dimension-free bound i.e., error guarantees independent of the length of path. The next lemma is a result of Babickenko et al. [44]. We include the proof for completeness.

Lemma 4.10. *[[44], Theorem 9] Let* $Y_1, \cdots, Y_d \in \{-1, +1\}^d$ *be* d *independent random variables with* $\mathbb{E}[Y_j] = \mu_j$*. Given* n *i.i.d. samples* $Y_j^{(i)}$ *for* $j = 1, \cdots, d$ *and* $i = 1, \cdots, n$*, let* $\widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y_j^{(i)}$ *be the empirical average of* Y_j*. Then for any* $\gamma > 0$ *,*

$$\mathbb{P}\Big[\big| \prod_{j=1}^{d} \widehat{\mu}_j - \prod_{j=1}^{d} \mu_j \big| \geq \gamma \Big] \leq \frac{8}{\gamma} \exp(-\gamma^2 n/32) \,.$$

Proof.

$$\prod_{j=1}^{d} \widehat{\mu}_j = \prod_{j=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} Y_j^{(i)} \right) = \frac{1}{n^d} \prod_{j=1}^{d} \sum_{i_j=1}^{n} Y_j^{(i_j)} = \frac{1}{n^d} \sum_{i_1, \cdots, i_d = 1}^{n} \prod_{j=1}^{d} Y_j^{(i_j)}$$

For every $l \in \{1, \cdots, n\}$, one can write

$$\prod_{j=1}^{d} \widehat{\mu}_j = \frac{1}{n^d} \sum_{i_1, \cdots, i_d = 1}^{n} \prod_{j=1}^{d} Y_j^{(i_j + l)}$$

where $i_j + l$ are taken modulo n. When we average over all l, we have

$$\prod_{j=1}^{d} \widehat{\mu}_j = \frac{1}{n^d} \sum_{i_1,\cdots,i_d=1}^{n} \frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_j+l)} \,. \tag{4.11}$$

For every multi-index $i_* = (i_1, \cdots, i_d)$, we define

$$d(i_*) = \begin{cases} 0 & \text{if } \left|\frac{1}{n}\sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_*+l)} - \prod_{j=1}^{d} \mu_j\right| \le \gamma/2 \,. \\ 2 & \text{otherwise} \,. \end{cases}$$

By definition of $d(i_*)$, we have

$$\left| \frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_*+l)} - \prod_{j=1}^{d} \mu_j \right| \le d(i_*) + \gamma/2 \,. \tag{4.12}$$

Note that for any fixed i_*, the random variables $\prod_{j=1}^{d} Y_j^{(i_*+1)}, \cdots, \prod_{j=1}^{d} Y_j^{(i_*+n)}$ are jointly independent with expectation $\prod_{j=1}^{d} \mu_j$. Therefore, using Hoeffding's inequality (Lemma A.1), we have

$$\mathbb{E}[d(i_*)] = 2\,\mathsf{P}\left[\left|\frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_*+l)} - \prod_{j=1}^{d} \mu_j\right| \ge \gamma/2 \right] \le 4\exp(-\gamma^2 n/32) \,. \tag{4.13}$$

We use Equations (4.11) and (4.12) to get the desired bound

$$\begin{aligned} \mathsf{P}\left[\left| \prod_{j=1}^{d} \widehat{\mu}_j - \prod_{j=1}^{d} \mu_j \right| \ge \gamma \right] &= \mathsf{P}\left[\left| \frac{1}{n^d} \sum_{i_1,\cdots,i_d=1}^{n} \frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_j+l)} - \prod_{j=1}^{d} \mu_j \right| \ge \gamma \right] \\ &\le \mathsf{P}\left[\frac{1}{n^d} \sum_{i_1,\cdots,i_d=1}^{n} \left| \frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{d} Y_j^{(i_j+l)} - \prod_{j=1}^{d} \mu_j \right| \ge \gamma \right] \\ &\le \mathsf{P}\left[\frac{1}{n^d} \sum_{i_1,\cdots,i_d=1}^{n} d(i_*) \ge \gamma/2 \right] \le \frac{8}{\gamma} \exp(-\frac{\gamma^2}{32}) \,, \end{aligned}$$

where in the last inequality we used Markov's inequality and (4.13). □

Lemma 4.11. *The event* $\mathtt{E}^{\text{cascade}}(\epsilon)$ *defined in* (3.11) *occurs with probability at least* $1 - \frac{4p^2}{\epsilon} \exp(-\epsilon^2 n/32)$.

Proof. For a given pair of nodes $w, \widetilde{w}$, let $\mathsf{path}_{\mathsf{T}}(w, \widetilde{w}) = \{e_1, e_2, \cdots, e_d\}$ and use the shorthand $\mu_i = \mu_{e_i}$ and $\widehat{\mu}_i = \widehat{\mu}_{e_i}$. Applying Lemma 4.10 to $\widehat{\mu}_i$'s (jointly independent random variables according to Lemma 4.9) gives

$$\mathsf{P}\left[\left|\prod_{j=1}^{d}\widehat{\mu}_j - \prod_{j=1}^{d}\mu_j\right| \leq \epsilon\right] \leq \frac{8}{\epsilon}\exp(-\epsilon^2 n/32)\,.$$

A union bound over $\binom{p}{2} \leq p^2/2$ pairs of nodes $w, \widetilde{w}$ provides the desired upper bound for probability of the event $\mathtt{E}^{\text{cascade}}(\epsilon)$. □

4.7 Two trees lemma

Lemma 4.12. *Let T_1 and T_2 be two spanning trees on set of nodes $\mathcal{V}$. Let $w, \widetilde{w}$ be a pair of nodes such that $\mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \neq \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$. Then there exists a pair of edges $f \triangleq (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w})$ and $g \triangleq (v, \widetilde{v}) \in \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$ such that*

(i) $f \notin \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$ and $g \notin \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w})$

(ii) $f \in \mathsf{path}_{\mathsf{T}_1}(v, \widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}_2}(u, \widetilde{u})$

Since $f \in \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}_1}(v, \widetilde{v})$, w and $\widetilde{w}$ (and respectively v and $\widetilde{v}$) are in different subtrees of T_1 after removing edge f, one can label the end points of the edges $f = (u, \widetilde{u})$ and $g = (v, \widetilde{v})$ such that $u, v \in \mathsf{SubTree}_{\mathsf{T}_1, f}(w)$ and $\widetilde{u}, \widetilde{v} \in \mathsf{SubTree}_{\mathsf{T}_1, f}(\widetilde{w})$ (See Figure 4.2.1).

Proof. We assume that T_1 and T_2 are two spanning trees with no common edges over a set of nodes $\mathcal{V}$. Otherwise, if T_1 and T_2 have any common edges, we can contract them and construct a new pair of spanning trees over set of nodes $\mathcal{V}'$ in a way that preserves paths. Note that by assumption $\mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \neq \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$, so $w\,, \widetilde{w} \in \mathcal{V}'$ are not merged in the contraction process explained above.

Let $u_1, \cdots, u_{K+1}$ be the vertices of $\mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w})$ in order with $u_1 = w$ and $u_{K+1} = \widetilde{w}$. Similarly, let $v_1, \cdots, v_{L+1}$ be the vertices of $\mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$ with $v_1 = w$ and $v_{L+1} =$

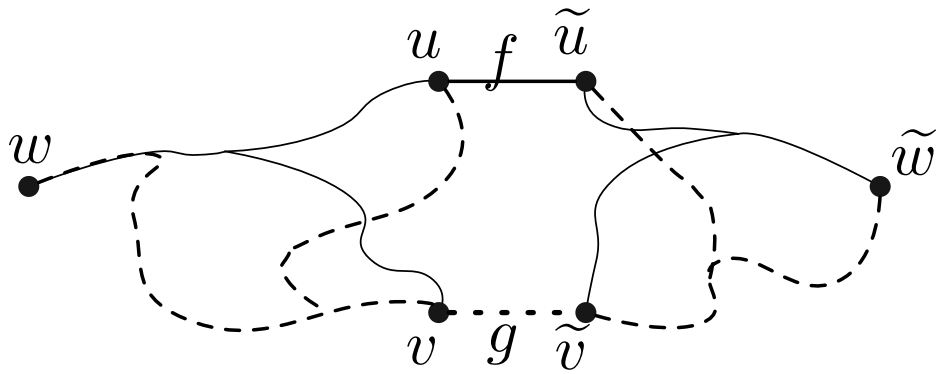

Figure 4.7.1: The schematics showing the pairs of edges f and g satisfying properties given in lemma 4.12. The solid lines represent the tree T_1 and the dashed lines represent the tree T_2. There always exists pairs of edges $f = (u, \widetilde{u})$ and $g = (v, \widetilde{v})$ such that $f \in \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}_1}(v, \widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}_2}(u, \widetilde{u})$.

$\widetilde{w}$. Now, for $1 \leq k \leq K$, let

$$h_{\mathsf{T}_1}(k) = \min\{t : (u_k, u_{k+1}) \in \mathsf{path}_{\mathsf{T}_1}(v_t, v_{t+1})\}\,.$$

We first show that $h_{\mathsf{T}_1}(k)$ is well-defined (i.e., for all $1 \leq k \leq K$, there exists a $1 \leq t \leq L$ such that $(u_k, u_{k+1}) \in \mathsf{path}_{\mathsf{T}}(v_t, v_{t+1})$). For each $1 \leq k \leq K$, let $\mathcal{S}_k = \mathsf{SubTree}_{\mathsf{T}_1,(u_k,u_{k+1})}(u_k)$ and $\widetilde{\mathcal{S}}_k = \mathsf{SubTree}_{\mathsf{T}_1,(u_k,u_{k+1})}(u_{k+1})$. Note that $w \in \mathcal{S}_k$ and $\widetilde{w} \in \widetilde{\mathcal{S}}_k$ for all k. Also, $(u_k, u_{k+1}) \in \mathsf{path}_{\mathsf{T}_1}(v_t, v_{t+1})$ for some $0 \leq t \leq L$ if and only if $v_t \in \mathcal{S}_k$ and $v_{t+1} \in \widetilde{\mathcal{S}}_{k+1}$. For any given k, if such t does not exist, then either $v_{t'} \in \mathcal{S}_k$ or $v_{t'} \in \widetilde{\mathcal{S}}_k$ for all $1 \leq t' \leq L+1$, which is a contradiction as $v_1 \in \mathcal{S}_k$ and $v_{L+1} \in \widetilde{\mathcal{S}}_k$. Hence $h_{\mathsf{T}_1}(k)$ is well-defined for all k. Equivalently, we could define $h_{\mathsf{T}_1}(k) = \min\{t : v_{t+1} \in \widetilde{\mathcal{S}}_k\}$. As $\widetilde{\mathcal{S}}_k \subseteq \widetilde{\mathcal{S}}_{k+1}$, the function $h_{\mathsf{T}_1}(k)$ is non-decreasing in k.

Similarly, the function $h_{\mathsf{T}_2}(l) = \min\{t : (v_l, v_{l+1}) \in \mathsf{path}_{\mathsf{T}_2}(u_t, u_{t+1})\}$ is well-defined and non-decreasing. To prove the result, we first show that there exists a pair of integers l^* and k^* such that $h_{\mathsf{T}_1}(k^*) = l^*$ and $h_{\mathsf{T}_2}(l^*) = k^*$. To do so, let function $h(l) : \{1, \cdots, L\} \to \{1, \cdots, L\}$ be $h(l) = h_{\mathsf{T}_1}(h_{\mathsf{T}_2}(l))$. This is a nondecreasing function over its domain. We now show that $h(l)$ has a fixed point. Assume that $h(1) > 1$ and $h(L) < L$, since otherwise the result is proven. Define $m = \min\{t : h(t) \leq t\}$

and note that $1 < m < L$. Thus, $h(m) \leq m$ while $h(m-1) > m-1$. Since $h(l)$ is non-decreasing, we have $m-1 < h(m-1) \leq h(m) \leq m$, which implies that $h(m) = m$.

Let l^* be a fixed point of $h(l)$. According to the definition of $h(l)$, the pair of integers l^* and $k^* = h_{\mathsf{T}_2}(l^*)$ satisfies $h_{\mathsf{T}_1}(k^*) = l^*$. Hence, the edge pair $(u_{k^*}, u_{k^*+1}) \in \mathsf{path}_{\mathsf{T}_1}(v_{l^*}, v_{l^*+1})$ and $(v_{l^*}, v_{l^*+1}) \in \mathsf{path}_{\mathsf{T}_2}(u_{k^*}, u_{k^*+1})$ satisfies the properties given in the Lemma. $\square$

4.8 Information projection lemmas

In the following, we provide two lemmas regarding the maximum likelihood tree and the parameter set given.

Lemma 4.13. *For a given tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$, the reverse information-projection of the distribution $P(x)$ onto the class of the Ising models on T with no external field (as in Equation (2.2)), $\widetilde{P}(x) = \Pi_{\mathsf{T}}(P) \triangleq \arg\min_{Q \in \mathcal{P}_{\mathsf{T}}} D(P\|Q)$ with the parameter $\widetilde{\theta}$ consistent with tree T (such that $\widetilde{\theta}_{ij} = 0$ for all $(i,j) \notin \mathcal{E}$), has the following property: $\tanh(\widetilde{\theta}_{ij}) = \mu_{ij}$ for all $(i,j) \in \mathcal{E}$, where $\mu_{ij} = \mathbb{E}_P X_i X_j$ is the pairwise correlation of the variables under the distribution $P(x)$. Also, we have $D(P\|\widetilde{P}) = -H(P) + \sum_{(i,j)\in\mathcal{E}_{\mathsf{T}}} H_B(\frac{1+\mu_{ij}}{2})$ where $H_B(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$ is the binary entropy function.*

Proof. This lemma is a direct corollary of Theorem 3.3 in [45]. Note that $\widetilde{P}(x)$ is the reverse I-projection of distribution $P(x)$ onto the exponential family with sufficient statistics $X_i X_j$ for $(i,j) \in \mathcal{E}$. It follows that $\mathbb{E}_{\widetilde{P}} X_i X_j = \mathbb{E}_P X_i X_j$ for $(i,j) \in \mathcal{E}$ which gives $\tanh(\widetilde{\theta}_{ij}) = \mu_{ij}$. Section 3.4.2 in [1] also addresses this problem in depth.

We provide the proof for the sake of completeness here: For the tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$

and any $Q \in \mathcal{P}_{\mathsf{T}}$ with parameter vector θ, using (2.2), Q factorizes as

$$\begin{aligned}
Q(x) &= \frac{\prod_{(i,j)\in\mathcal{E}} \exp(\theta_{ij} x_i x_j)}{\sum_{x\in\{-1,+1\}^p} \prod_{(i,j)\in\mathcal{E}} \exp(\theta_{ij} x_i x_j)} \\
&= \frac{\prod_{(i,j)\in\mathcal{E}} \exp(\theta_{ij} x_i x_j)}{\prod_{(i,j)\in\mathcal{E}} \sum_{x_i,x_j\in\{-1,+1\}} \exp(\theta_{ij} x_i x_j)} \\
&= \prod_{(i,j)\in\mathcal{E}} \frac{\exp(\theta_{ij} x_i x_j)}{\exp(\theta_{ij} x_i x_j) + \exp(-\theta_{ij} x_i x_j)} \\
&= \prod_{(i,j)\in\mathcal{E}} \frac{1}{2}\left(1 + \tanh \theta_{ij} x_i x_j\right)
\end{aligned}$$

where $\mathbb{E}_Q X_i X_j = \tanh \theta_{ij}$. Let $Y_e = X_i X_j$ for all $e = (i,j) \in \mathcal{E}$ so that $Q(Y_e = +) = (1 + \tanh \theta_{ij})/2$ and $P(Y_e = +) = (1 + \mu_{ij})/2$ where $\mu_{ij} = \mathbb{E}_P X_i X_j$.

Given probability mass function P on $\{-1,+1\}^p$ and any $Q \in \mathcal{P}_{\mathsf{T}}$ with parameters θ,

$$\begin{aligned}
D(P\|Q) = \mathbb{E}_P\left[\log \frac{P}{Q}\right] &= -H(P) - \mathbb{E}_P[\log Q(X)] \\
&= -H(P) - \mathbb{E}_P\left[\log \prod_{(i,j)\in\mathcal{E}} \frac{1 + \tanh \theta_{ij} X_i X_j}{2}\right] \\
&= -H(P) - \sum_{e=(i,j)\in\mathcal{E}} \mathbb{E}_P\left[\log \frac{1 + \tanh \theta_{ij} Y_e}{2}\right] \\
&= -H(P) - \sum_{e\in\mathcal{E}} \mathbb{E}_P[\log Q(Y_e)] \\
&= -H(P) + \sum_{e\in\mathcal{E}} H_B(P(Y_e)) + \sum_{e\in\mathcal{E}} D(P(Y_e)\|Q(Y_e))\,.
\end{aligned}$$

$\widetilde{P} = \arg\min_{Q\in\mathcal{P}_{\mathsf{T}}} D(P\|Q)$ with parameters $\widetilde{\theta}$ satisfies the property $\widetilde{P}(Y_e) = P(Y_e)$ for all $e \in \mathcal{E}$, hence $\tanh(\widetilde{\theta}_{ij}) = \mu_{ij}$ for all $(i,j) \in \mathcal{E}$. Consequently, $D(P\|\widetilde{P}) = -H(P) + \sum_{e\in\mathcal{E}} H_B\left(\frac{1+\mu_{ij}}{2}\right)$. □

Lemma 4.14. *Given empirical distribution $\widehat{P}$, the tree T^{CL} defined in Definition 3.1 can be found as the maximum weight spanning tree over a complete weighted graph where the weights of each edge (i,j) is $|\widehat{\mu}_{ij}|$.*

Proof. Using Lemma 4.13,

$$\mathsf{T}^{\mathsf{CL}} = \arg\min_{\mathsf{T}\in\mathcal{T}} \sum_{e\in\mathcal{E}_{\mathsf{T}}} H_B\big(\frac{1+\widehat{\mu}_e}{2}\big) - H(\widehat{P}) = \arg\min_{\mathsf{T}\in\mathcal{T}} \sum_{e\in\mathcal{E}_{\mathsf{T}}} H_B\big(\frac{1+\widehat{\mu}_e}{2}\big)\,.$$

The maximum weight spanning tree can be implemented greedily using Kruskal algorithm or Prim's algorithm [43]. So, finding the maximum weight spanning tree only depends on the sorted order of the edges of the graph.

$H_B\big(\frac{1+\widehat{\mu}_{ij}}{2}\big)$ is a monotonically increasing function of $|\widehat{\mu}_{ij}|$. So, sorting all the edges (i,j) in the complete graph based $H_B\big(\frac{1+\widehat{\mu}_{ij}}{2}\big)$ or $|\widehat{\mu}_{ij}|$ gives the same order. This gives

$$\mathsf{T}^{\mathsf{CL}} = \arg\min_{\mathsf{T}\in\mathcal{T}} \sum_{(i,j)\in\mathcal{E}_{\mathsf{T}}} H_B\big(\frac{1+\widehat{\mu}_{ij}}{2}\big) = \arg\min_{\mathsf{T}\in\mathcal{T}} \sum_{(i,j)\in\mathcal{E}_{\mathsf{T}}} |\widehat{\mu}_{ij}|\,. \qquad \square$$

4.9 Proof of forest approximation result

In the following, we show an upper bound for the sample complexity of

Proof. [Proposition 3.8] We will provide ϵ and γ such that on the event $\mathtt{E}(\epsilon,\gamma)$ the ssTV bound holds. All the weak edges $e \in \mathcal{E}_{\mathsf{T}}$ such that $|\mu_e| \le \tau$ (for which we cannot guarantee correct recovery by Chow-Liu algorithm) have $|\widehat{\mu}_e| \le \tau+\epsilon$ on event $\mathtt{E}^{\text{corr}}(\epsilon)$, hence these edges are truncated by the thresholding process (Remember that the forest approximation algorithm thresholds edges with $|\widehat{\mu}_{ij}| \le \tau+\epsilon$ to zero.) It follows that $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$.

On event $\mathtt{E}^{\text{strong}}(\epsilon)$, the strong edges defined in Equation (3.9) are recovered by the Chow-Liu algorithm, i.e., for all $e \in \mathcal{E}_{\mathsf{T}}$, with $|\mu_e| \ge \tau+2\epsilon$, $e \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$. On event $\mathtt{E}^{\text{corr}}(\epsilon)$, these edges also satisfy $|\widehat{\mu}_e| \ge \tau+\epsilon$ and these edges are retained. This implies that all of the strong edges in the original tree T (with $|\mu_e| \ge \tau+2\epsilon$) are recovered correctly by the truncation algorithm (i.e., $\mathcal{E}_{\mathsf{T}}^{\text{strong}} \subseteq \mathcal{E}_{\widehat{\mathsf{T}}}$).

To bound the ssTV, we decompose into two terms as per (3.7). We first study the loss due to graph estimation error $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\mathsf{P}))$. For any pair of nodes $w, \widetilde{w}$ in $\widehat{\mathsf{T}}$, since $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$, if there is a path between them, then this is the same as the path between them in T. Hence looking at (3.6), $\mathbb{E}_P[X_w X_{\widetilde{w}}] = \mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\mathsf{P})}[X_w X_{\widetilde{w}}]$ and the

error is zero. If there is no path between $w, \widetilde{w}$ in $\widehat{\mathsf{T}}$, then $\mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\mathsf{P})}[X_w X_{\widetilde{w}}] = 0$ and the error is $\mathbb{E}_P[X_w X_{\widetilde{w}}]$. This occurs whenever there is an edge $e \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$ such that $e \notin \mathcal{E}_{\widehat{\mathsf{T}}}$. All the missing edges in $\widehat{\mathsf{T}}$ are weak edges, hence $|\mu_e| \leq \tau + 2\epsilon$. This implies that $|\mathbb{E}_P[X_w X_{\widetilde{w}}]| \leq |\mu_e| \leq \tau + 2\epsilon$ and it follows that $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\mathsf{P})) \leq \tau + 2\epsilon$.

To bound the loss due to parameter estimation error $\mathcal{L}^{(2)}(\Pi_{\widehat{\mathsf{T}}}(\mathsf{P}), \Pi_{\widehat{\mathsf{T}}}(\widehat{P}))$, we again consider an arbitrary pair of nodes $w, \widetilde{w}$. If there is a path between them in $\widehat{\mathsf{T}}$, (since $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$), then occurrence of event $\mathtt{E}^{\text{cascade}}(\gamma)$ implies $|\mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\mathsf{P})}[X_w X_{\widetilde{w}}] - \mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\widehat{P})}[X_w X_{\widetilde{w}}]| \leq \gamma$. If there is no path between w and $\widetilde{w}$ in $\widehat{\mathsf{T}}$, then $\mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\mathsf{P})}[X_w X_{\widetilde{w}}] = \mathbb{E}_{\Pi_{\widehat{\mathsf{T}}}(\widehat{P})}[X_w X_{\widetilde{w}}] = 0$. Hence, $\mathcal{L}^{(2)}(\Pi_{\widehat{\mathsf{T}}}(\mathsf{P}), \Pi_{\widehat{\mathsf{T}}}(\widehat{P})) \leq \gamma$.

Combining the two error terms gives $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\widehat{P})) \leq \tau + 2\epsilon + \gamma$. If $\epsilon \leq \eta e^{-\beta}/8$ and $\gamma \leq \eta/4$ then $\tau = 4\epsilon/\sqrt{1 - \tanh\beta} \leq 4\epsilon e^{\beta} \leq \eta/8$ and $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\widehat{P})) \leq \tau + 2\epsilon + \gamma \leq \eta$. The event $\mathtt{E}(\epsilon, \gamma)$ with $\epsilon \leq \eta e^{-\beta}/8 \leq \eta/8$ and $\gamma < \eta/4$ occurs with probability at least $1 - 3\delta$ given

$$n > \max\{128 e^{2\beta}/\eta^2 \log \frac{2p^2}{\delta}, 512/\eta^2 \log \frac{32p^2}{\eta\delta}\}$$

samples (by Lemma 3.11), which gives the result. □

Chapter 5

Accurate k-wise marginals

We first describe our result, expressing marginals in terms of correlations, and then delineate implications to learning Ising models in order to make subsequent predictions from partial observations.

Let $\mathcal{S} \subset \mathcal{V}$ be a subset of variables. We are interested in the marginal $P(x_{\mathcal{S}})$ for an Ising model $P(x)$ of the form (2.2). Note that the variables have mean zero by symmetry of the model, hence $P(x_i) = \frac{1}{2}$ for each $i \in \mathcal{V}$. When $|\mathcal{S}| = 2$, the marginal over any pair of nodes $\mathcal{S} = \{i, j\}$ can be expressed as $P(X_i = x_i, X_j = x_j) = \frac{1}{4}[1 + x_i x_j \mu_{ij}]$ for $x_i, x_j \in \{-, +\}$.

We seek a simple closed form for the marginal distribution $P(x_{\mathcal{S}})$ as a function of pairwise correlations between pairs of nodes $i, j \in \mathcal{S}$. By the previous paragraph, we have already accomplished this when $|\mathcal{S}| = 2$. The general case $|\mathcal{S}| > 2$ is (as one might expect) rather more complicated and involves a certain maximization over matchings.

Let $\binom{\mathcal{V}}{2}$ denote the collection of pairs of nodes in $\mathcal{V}$. For $m \in \binom{\mathcal{V}}{2}$, if $m = \{i, j\}$, then we also write the correlation $\mu_m = \mathbb{E}_P X_i X_j$.

Definition 1. *For subset of nodes $\mathcal{C} \subseteq \mathcal{V}$ of even cardinality $|\mathcal{C}| = 2t$, we define the set $\mathcal{M}_{\mathcal{C}}$ to be the collection of all partitions of $\mathcal{C}$ into t disjoint pairs of nodes, i.e. the*

set of all perfect matchings in the complete graph with node set $\mathcal{C}$:

$$\mathcal{M}_{\mathcal{C}} = \left\{ M = \{m_1, \cdots, m_t\} : m_l \in \binom{\mathcal{C}}{2} \text{ for each } l \in [t],\, m_l \cap m_k = \varnothing \text{ for } l \neq k \right\}.$$

Definition 2. *Let $t \geq 1$ be a positive integer and $\mathcal{C} \subseteq \mathcal{V}$ be of size $|\mathcal{C}| = 2t$. The set-valued function $G(\mathcal{C})$ is defined as the set of matchings M in $\mathcal{M}_{\mathcal{C}}$ that maximize the product of correlations for elements in M:*

$$G(\mathcal{C}) = \arg\max_{M \in \mathcal{M}_{\mathcal{C}}} \prod_{m \in M} |\mu_m| \,. \tag{5.1}$$

We define $G^*(\mathcal{C})$ to be an arbitrary element in $G(\mathcal{C})$, and $G(\varnothing) = \varnothing$ and $G^*(\varnothing) = \varnothing$. We also use the notation $G^*(\mathcal{C}, P)$ to be optimal matching under the distribution P (maximizing (5.1)) to explicitly determine the underlying distribution.

We emphasize that often the maximum in (5.1) is not uniquely achieved and the set $G(\mathcal{C})$ can be large. In what follows we use the convention that the product $\prod_{m \in \mathcal{I}} \mu_m = 1$ whenever $\mathcal{I} = \varnothing$. Our main result is the following:

Theorem 5.1. *If $P(x)$ is the probability mass function for an Ising model (2.2) on any tree T, then the marginal $P(x_{\mathcal{S}})$ over the subset $\mathcal{S} \subseteq \mathcal{V}$ can be written as:*

$$P(x_{\mathcal{S}}) = 2^{-|\mathcal{S}|} \sum_{t=0}^{\lfloor |\mathcal{S}|/2 \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}| = 2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C})} \mu_m \,. \tag{5.2}$$

Proof. For subset of variables $\mathcal{S} \subseteq \mathcal{V}$, the marginal distribution $P_{\mathcal{S}}$ is

$$\begin{aligned}
P_{\mathcal{S}}(x_{\mathcal{S}}) &= \sum_{x_{\mathcal{S}^c}} P(x) \\
&\overset{(a)}{=} \sum_{x_{\mathcal{S}^c}} \prod_{(i,j) \in \mathcal{E}} \frac{1 + \mu_{ij} x_i x_j}{2} \\
&\overset{(b)}{=} \frac{1}{2^p} \sum_{x_{\mathcal{S}^c}} [1 + \sum_{\mathcal{F} \subseteq \mathcal{E}, \mathcal{F} \neq \varnothing} \prod_{(i,j) \in \mathcal{F}} \mu_{ij} x_i x_j] \\
&\overset{(c)}{=} \frac{1}{2^{|\mathcal{S}|}} [1 + \sum_{\substack{\mathcal{F} \subseteq \mathcal{E}, \\ \mathcal{F} \in \Omega(\mathcal{S}), \mathcal{F} \neq \varnothing}} \prod_{(i,j) \in \mathcal{F}} \mu_{ij} x_i x_j]
\end{aligned}$$

where (a) uses the Equation (3.1) for $P \in \mathcal{P}_{\mathsf{T}}$. In (b) we expand the product term over all the edges in the tree. In (c), $\Omega(\mathcal{S}) \subseteq 2^{\mathcal{E}}$ is defined such that for $\mathcal{F} \subseteq \mathcal{E}$, if $\mathcal{F} \in \Omega(\mathcal{S})$, then for all nodes $i \in \mathcal{S}^c$, there are even number of edges $e \in \mathcal{F}$ that have i as their end point. Equivalently, any $\mathcal{F} \in \Omega(\mathcal{S})$ is the collection of edges in the set of edge-disjoint paths between distinct pairs of nodes in $\mathcal{S}$. □

The utility of this expression is in part due to the fact that it depends on the tree topology T only through the optimization implicit in $G(\mathcal{C})$, which depends only on end-to-end correlations between variables in $\mathcal{C} \subseteq \mathcal{S}$. In particular, the expression is agnostic to the topology over variables not in $\mathcal{S}$: even the *number* of such variables is irrelevant. This is important because there can be several (non-isomorphic) subtrees which give the same marginal over $\mathcal{S}$. This means that it is in general impossible to detect the tree topology just by observing end-to-end correlations, but this expression is robust to this.

Lemma 5.2. *For any set $\mathcal{C} \subseteq \mathcal{V}$ with $|\mathcal{C}| = 2t$, and any pair of tree structured Ising distributions P and Q as in (2.2) (with possibly different underlying tree), let $\mu_{ij} = \mathbb{E}_P X_i X_j$ and $\widetilde{\mu}_{ij} = \mathbb{E}_Q X_i X_j$. If $|\mu_{ij} - \widetilde{\mu}_{ij}| \leq \epsilon$ for all $i, j \in \mathcal{C}$, then*

$$|\prod_{m \in G^*(\mathcal{C},P)} \mu_m - \prod_{m' \in G^*(\mathcal{C},Q)} \widetilde{\mu}_{m'}| \leq 2t\epsilon\,.$$

Proof. First, we will show that for all $M \in \mathcal{M}_{\mathcal{C}}$, $\prod_{m \in M} \mu_m$ takes the same sign (*i.e., for* $M, M' \in \mathcal{M}_{\mathcal{C}}$, $\prod_{m \in M} \mu_m \prod_{m' \in M'} \mu_{m'} \geq 0$). A similar statement holds for $\widetilde{\mu}$. To do so, note that for all $M, M' \in \mathcal{M}_{\mathcal{C}}$, one can make a sequence of matchings $M_0 = M, M_1, \cdots, M_r = M'$ such that M_r and M_{r+1} differ only at two pairs of nodes (*i.e.,* There exists $(i,j),(k,l) \in M_r$ such that $(i,k),(j,l) \in M_{r+1}$). The product of correlations on M_r and M_{r+1} have the same sign (*i.e.,* $\prod_{m \in M_r} \mu_m \prod_{m \in M_{r+1}} \mu_m \geq 0$). The same argument applies for all r and hence $\prod_{m \in M} \mu_m \prod_{m' \in M'} \mu_{m'} \geq 0$.

Next, we show that if $|\mu_{ij} - \widetilde{\mu}_{ij}| \leq \epsilon$ for all $i, j \in \mathcal{C}$, then for any

$$M = \{m_1, m_2, \cdots, m_t\} \in \mathcal{M}_{\mathcal{C}}$$

we have

$$|\prod_{m\in M}\mu_m - \prod_{m'\in M}\widetilde{\mu}_{m'}| \le t\epsilon\,.$$

Hence,

$$\begin{aligned}
|\prod_{i=1}^{t}\mu_{m_i} - \prod_{i=1}^{t}\widetilde{\mu}_{m_i}| &= |\sum_{i=1}^{t}\prod_{j=1}^{i-1}\mu_{m_j}\prod_{j'=i+1}^{t}\widetilde{\mu}_{m_{j'}}(\mu_i-\widetilde{\mu}_i)| \\
&\le \sum_{i=1}^{t}|\prod_{j=1}^{i-1}\mu_{m_j}\prod_{j'=i+1}^{t}\widetilde{\mu}_{m_{j'}}(\mu_i-\widetilde{\mu}_i)| \\
&\le \sum_{i=1}^{t}|\mu_i-\widetilde{\mu}_i| \le t\epsilon\,.
\end{aligned}$$

Let's define

$$a = \prod_{m\in G^*(\mathcal{C},P)}\mu_m,\quad a' = \prod_{m\in G^*(\mathcal{C},Q)}\mu_m,\quad b = \prod_{m\in G^*(\mathcal{C},P)}\widetilde{\mu}_m,\quad b' = \prod_{m\in G^*(\mathcal{C},Q)}\widetilde{\mu}_m\,.$$

Hence, we show that $aa' \ge 0\,, bb' \ge 0, |a-b| \le t\epsilon$ and $|a'-b'| \le t\epsilon$. We also know that $|a| \ge |a'|$ and $|b'| \ge |b|$ according to the definition of $G^*(\mathcal{C},P)$ and $G^*(\mathcal{C},Q)$. To bound $|a-b'|$, we study two different cases: Either $ab \ge 0$ in which case $|a-b'| = ||a|-|b'|| =\le \max\{|a-b|,|a'-b'|\} \le t\epsilon$ or $ab \le 0$ in which case $|a-b'| = |a|+|b'| \le |a-b|+|a'-b'| \le 2t\epsilon$. □

Proposition 1. *Let $\mathsf{T} = (\mathcal{V},\mathcal{E})$ be a tree and P an Ising model on T. Let Q be an Ising model on tree $\mathsf{T}' = (\mathcal{V},\mathcal{E}')$. If $|\mathbb{E}_P X_iX_j - \mathbb{E}_Q X_iX_j| \le \eta$ for all $i,j \in \mathcal{V}$ (i.e., $\mathcal{L}^{(2)}(P,Q) \le \eta/2$), then $\mathcal{L}^{(k)}(P,Q) \le k2^k\eta$.*

Proof. For $i,j \in \mathcal{V}$, let $\mu_{ij} \triangleq \mathbb{E}_P X_iX_j$ and $\widetilde{\mu}_{ij} \triangleq \mathbb{E}_P X_iX_j$. We also use the notation $G^*(\mathcal{C},P)$ to be optimal matching $G^*(\mathcal{C})$ under the distribution P.

To bound $\mathcal{S} \subseteq \mathcal{V}$ with $|\mathcal{S}| = k$, we use Lemma 5.2 to get:

$$
\begin{aligned}
& d_{\mathrm{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}) \\
&= \sum_{x_{\mathcal{S}} \in \{-,+\}^k} |P(x_{\mathcal{S}}) - Q(x_{\mathcal{S}})| \\
&= \frac{1}{2^k} \sum_{x_{\mathcal{S}} \in \{-,+\}^k} \left| \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}|=2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C},P)} \mu_m - \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}|=2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C},Q)} \widetilde{\mu}_m \right| \\
&\leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}|=2t}} \frac{1}{2^k} \sum_{x_{\mathcal{S}} \in \{-,+\}^k} \left| \prod_{c \in \mathcal{C}} x_c \right| \cdot \left| \prod_{m \in G^*(\mathcal{C},P)} \mu_m - \prod_{m \in G^*(\mathcal{C},Q)} \widetilde{\mu}_m \right| \\
&= \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}|=2t}} \left| \prod_{m \in G^*(\mathcal{C},P)} \mu_m - \prod_{m \in G^*(\mathcal{C},Q)} \widetilde{\mu}_m \right| \\
&\leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subset \mathcal{S} \\ |\mathcal{C}|=2t}} 2t\eta \leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \binom{k}{2t} 2t\eta \leq k 2^k \eta \,. \quad \square
\end{aligned}
$$

$\square$

Note that learning the tree T^{CL} from n samples using the variant of Chow-Liu algorithm has time complexity of $O((n+1)p^2)$. Finding low-order marginals and posteriors of any subset of variables on a tree-structured distribution takes $O(p)$ operations using belief-propagation algorithm.

Chapter 6

Discussion

Learning the structure of Ising models on trees with finite number of samples is impossible in the presence of weak edges. Yet, we proved that using the maximum likelihood tree enables accurate estimation of the the original distribution for the purpose of prediction. The sample complexity of this estimation is given in this paper.

There is a large literature on learning tree-structured Markov random fields, and it is useful to carefully compare the guarantees obtained by each when applied to our setting. In the following we review different approaches that could be taken toward learning a tree-structured distribution. We also review some known algorithms and their sample complexity. The input of these algorithm is n i.i.d. samples from a tree-structured distribution $P(x) \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ with $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ being a tree. The gurantees mentioned below are with high probability.

- *Guaranteed recovery of the structure*

  Algorithm: Chow-Liu algorithm [19]

  Output: T^{CL}

  Guarantee: $\mathsf{T} = \mathsf{T}^{\mathsf{CL}}$

  Requirement: $n > C\frac{e^{2\beta}}{\alpha^2} \log p$

  In Theorems 3.3 and 3.2 we provided the sample complexity of this problem

(tight up to a constant factor).

- *Guaranteed recovery of the structure and accurate pairwise marginal*

  Algorithm: Chow-Liu algorithm [19]

  Output: T^{CL} and $Q \in \mathcal{P}_{\mathsf{T}^{\mathsf{CL}}}$

  Guarantee: $\mathsf{T} = \mathsf{T}^{\mathsf{CL}}$ and $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

  Requirement: $n > C \max\{\frac{e^{2\beta}}{\alpha^2} \log p, \frac{1}{\eta^2} \log \frac{p}{\eta}\}$

- *Forest Approximation and accurate pairwise marginal*

  Algorithm: Chow-Liu algorithm with a proper threshold over the weight of edges

  Output: $\widehat{\mathsf{T}} = (\mathcal{V}, \mathcal{E}_{\widehat{\mathsf{T}}})$ and $Q \in \mathcal{P}_{\widehat{\mathsf{T}}}$

  Guarantee: $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$ and $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

  Requirement: $n > C \max\{e^{2\beta}/\eta^2 \log p, 1/\eta^2 \log \frac{p}{\eta}\}$

  Proved in Proposition 3.8

- *Accurate pairwise marginals*

  Algorithm: Chow-Liu Algorithm for the purpose of inference

  Output: $\mathsf{T}^{\mathsf{CL}} = (\mathcal{V}, \mathcal{E})$ and $Q \in \mathcal{P}_{\mathsf{T}^{\mathsf{CL}}}$

  Guarantee: $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

  Requirement: $n > C \max\{e^{2\beta} \log p, \eta^{-2} \log \frac{p}{\eta}\}$

  Proved in Theorem 3.4

- *Latent tree with accurate pairwise marginals*

  Algorithm: Agarwala, Bafna, Farach, Paterson, and Thorup [35].

Output: $S = (\mathcal{V} \cup \widetilde{\mathcal{V}}, \mathcal{E}_S)$ and $Q \in \mathcal{P}_S$ such that $\mathcal{V}$ is the set of leaves of S.

Guarantee: $\mathcal{L}^{(2)}(P, Q_{\mathcal{V}}) \leq \eta$.

Requirement: $n > C \frac{1}{\alpha^2 \eta^2} \log p$

Discussed in Section 2.1.

- *Latent tree with accurate joint distribution over the leaves*

  Algorithm: Ambainis, Desper, Farach, and Kannan [36].

  Output: $S = (\mathcal{V} \cup \widetilde{\mathcal{V}}, \mathcal{E}_S)$ and $Q \in \mathcal{P}_S$ such that $\mathcal{V}$ is the set of leaves of S.

  Guarantee: $\mathcal{L}^{(p)}(P, Q_{\mathcal{V}}) \leq \eta$.

  Requirement: $n > C \max\{1/\alpha^2, pe^{2\beta}/\eta^2\}$

  Discussed in Section 2.1.

The last two introduced problems are examples of improper learning in which the class of models from which we learn is bigger than the class of models assumed over the original distribution. In this case, we assume that the original distribution is tree-structured over the nodes in $\mathcal{V}$ whereas the learned model is represented with a latent tree such that its leaves are the nodes in $\mathcal{V}$. Specific applications determine the requirements over the learned models and the guarantees which are necessary for the subsequent purposes. In general proper learning is necessary when the identification of the qualitative structure behind the distributions is necessary. The extreme case of which is when one is only interested in the underlying structure which corresponds to the first problem in above category.

Note that every tree-structured distribution over p nodes in $\mathcal{V}$ can be equivalently represented using a latent tree with $\mathcal{V}$ as its leaves. To do so, one has to add some dummy variables and infinite strength edges to the original tree. The converse is not true: A distribution represented on the leaves of a latent tree cannot be factorized as a tree-structured distribution in general.

Part II

Online Recommendation System with a Latent Variable Model

Chapter 7

Background and motivation

7.1 Introduction

Options are good, but if there are too many options, we need help! It is increasingly the case that our interaction with content is mediated by recommendation systems. There are two main approaches taken in recommendation systems: *content filtering* makes use of features associated with items and users (e.g., age, location, gender of users and genre, director of movies). In contrast, *collaborative filtering* are made based on observed user preferences. Thus, for instance two users are thought of as similar if they have revealed similar preferences irrespective of their profile. Similarly, two items are thought of as similar if most users have similar preferences for them. More generally, collaborative filtering (CF) makes use of structure in the space of users and the space of items, for example low-rank matrix formulation [46, 47, 48, 49, 50, 51, 52, 53] .

An important aspect of most recommendation systems is that each recommendation influences what is learned which in turn determines the possible accuracy of the future recommendations. This introduces a tension between exploring to get information and exploiting existing knowledge to make good recommendations. This is exactly the phenomenon of interest in the substantial literature on the multi-armed bandit (MAB) problem and its variants [54, 55, 56]. In MAB problems, the optimal strategy is pulling the optimal arm repeatedly. But clearly this is a bad idea in rec-

ommendation systems as most people would not find it useful to be recommended the same movie over and over again.

It is common to think of recommendation systems as a matrix completion problem. Given a few observed entries, the matrix completion problem is to estimate the rest of matrix, where it is assumed that the matrix satisfies some properties. But this criteria is not matched to the motivating recommendation problem; more appropriate measure of performance is the number of good recommendations made by the algorithm.

The first contribution of this work is to provide a framework to understand various recommendation system algorithms. We focus on a latent variable model based on which each user is associated with a user type and each item is associated with an item type. Users who belong to the same user type share similar preferences for all items and items who belong to the same type are rated by everyone similarly. The measure of performance of recommendation algorithm in this framework is the expected number of bad recommendations per user made by the algorithm.

We focus on two well-studied categories of algorithm for recommendation systems: *user-user* algorithms [57, 58, 59] utilize the structure in user space to predict the preference of users for items they have not rated yet. To do so, the preference of user u for item i is estimated by the preference of another user u' (believed to be similar to u based on their previous ratings) for item i. Alternatively, *item-item* algorithms [60, 61, 62] utilize the structure in item space to predict the preference of users for items. To do so, the preference of user u for item i is estimated by the preference of user u for another item i' (believed to be similar to i based on their previous ratings). We provide two versions of user-user and item-item CF algorithms tailored to the latent variable model in the online recommendation system of interest. Performance guarantees for the proposed algorithms are provided based on the model introduced. Similar model of data to our approach in which there is an underlying clustering of rows and columns of data is studied in other problems including [63, 64].

Next, we focus on two extreme regimes of interest of the latent variable model: in user-structure only model, the parameters of the model are so that there is no structure in item space. Alternatively, in item-structure only model, the parameters

are so that there is not structure in user space. We provide information theoretic lower bounds for performance of any algorithm based on user-structure only model and item-structure only model. We show that our proposed user-user and item-item online CF algorithms are *almost* information theoretically optimal (within multiplicative logarithmic factors).

We use novel machinery in the proof of information theoretic lower bounds for performance of online CF algorithms. As an example, in the user-structure only model, we make two observations: recommending item i to user u at time t results in uncertain outcome if there is no user u' which has rated enough items similar to u by time $t-1$. In this case, we cannot be certain about similarity of u and any other user. The outcome of recommending i to u at time t is also uncertain if none of users similar to u has rated item i by time $t-1$. Such observations imply a lower bound on the necessary number of exploration recommendations before recommending something which is expected to be liked. Similar observations are made for the model based on item-structure only and later on the model considering structure in both user and item space.

Later, we propose a hybrid algorithm which utilizes the structure in item space and user space simultaneously. To do so, item space is explored until enough information is gathered to learn the similarity between users. After that, the effort of exploring item space is shared between similar users. Lower bounds show that the performance of our proposed hybrid algorithm is almost optimal in all regimes of interest. The performance of hybrid algorithm is also proved to be better than the performance of the user-user and item-item algorithms introduced before. Such hybrid algorithms have been studied before in [65, 66]. In [66] a latent variable model which features of users and items take values in real line and the the outcome of recommendation is a noisy observation of a function of features.

A few papers including [67, 62, 57] have provided theoretical guarantees for online or adaptive collaborative filtering problem.

In particular, the underlying model used in [67] is quite similar to our setup. But they focused on designing an algorithm which maximizes the probability of recom-

mending a liked item at finite horizon starting from any given time with some given history. In particular, they provide an algorithm which performs exploitation phase in a provably optimal fashion asymptotically, but their approach does not reveal any information on how to come up with the exploration phase which provides small regret. Similar results appear in [68] and [46].

[66] proposes a nonparametic regression over latent variable models. In this model features of users and items take values in real line. The the outcome of recommendations are a noisy observation of a function of features with the observation function being Lipschitz and bounded. They provide the sample complexity for the matrix completion problem. The Lipschitzness assumption of the observation function provides the guarantee results with sample complexity independent of the dimension of the feature space. They proposed two algorithm based on the structure in user space or item space. They also propose a hybrid algorithm which looks at both structure in user space and item space and exploits the one which has smaller variance for each entry. Their work is different from ours in the sense that the latent variable model is a bit more flexible (even though it does not capture our latent variable model as our model does not satisfy Lipschizness). But they study the matrix completion in the offline mode. They also did not provide the optimality guarantees.

7.1.1 Notation

For an integer a we write $[a] = \{1, \cdots, a\}$ and for real-valued x let $(x)_+ = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$.

We note here that variables or parameters, except those in Figure 7.2.1, may take different values in each section.

7.2 Model

7.2.1 Problem setup

There are N users in the system. At each time $t \geq 1$, the algorithm recommends an item $a_{u,t} \in \mathbb{N}$ to each user u and receives feedback $L_{u,a_{u,s}} \in \{+1, -1\}$ ('like' or 'dislike'). We impose the condition that each item may be recommended at most once to each user. The condition contrasts with the multi-armed bandit setup where the optimal algorithm converges to repeated play of the same arm: a recommendation system that repeatedly recommends the same movie, even if it is very good, is surely problematic! In order that the algorithm never run out of items to recommend, we suppose there are infinitely many items to draw from and we identify them with the natural numbers.

The history at time t, H_{t-1}, is the collection of actions and feedback up to time $t-1$, *i.e.*, $H_{t-1} = \{a_{u,s}, L_{u,a_{u,s}}, \text{ for } u \in [N], s \in [t-1]\}$. We are interested in online learning algorithms, in which the random variable $a_{u,t}$ is a (possibly random) function of H_{t-1}. This additional randomness is encoded in random variables $\zeta_{u,t}$, assumed to be independent of all other variables. In this way, $a_{u,t} = f_{u,t}(H_{t-1}, \zeta_{u,t})$, for some deterministic function $f_{u,t}$.

Algorithm performance will be evaluated after some arbitrary number of time-steps T. The performance metric we use is expected regret (simply called regret in what follows), defined as the expected number of disliked items recommended per user:

$$\texttt{regret}(T) = \mathbb{E} \sum_{t=1}^{T} \frac{1}{N} \sum_{u=1}^{N} \mathbb{1}[L_{u,a_{u,t}} = -1] \,. \tag{7.1}$$

The algorithms we describe depend on knowing the time-horizon T, but by standard techniques in the multi-armed bandit literature it is possible to convert these to algorithms achieving the same (up to constant factors) regret without this knowledge. This latter notion of regret, where the algorithm does not know the time-horizon of interest and must achieve good performance across all time-scales, is called *anytime regret* in the literature.

7.2.2 User preferences

There is a latent-variable model for the preferences ('like' or 'dislike') of the users for the items, based on the idea that there are relatively few *types* of users and/or few *types* of items. Each user $u \in [N]$ has a user type $\tau_U(u)$ i.i.d. uniformly distributed on $[\mathtt{q}_U]$, where $\mathtt{q}_U$ is the number of user types. Similarly, each item $i \in \mathbb{N}$ has a random item type $\tau_I(i)$ i.i.d. uniformly distributed on $[\mathtt{q}_I]$, where $\mathtt{q}_I$ is the number of item types. The random variables $\{\tau_U(u)\}_{1 \leq u \leq N}$ and $\{\tau_I(i)\}_{1 \leq i}$ are assumed to be jointly independent.

All users of a given type have identical preferences for all the items, and similarly all items of a given type are rated in the same way by any particular user. The entire collection of user preferences $(L_{u,i})_{u,i}$ is therefore encoded into a much smaller *type matrix* $\Xi = (\xi_{k,j}) \in \{-1, +1\}^{\mathtt{q}_U \times \mathtt{q}_I}$, which specifies the preference of each user type for each item type: The rating $L_{u,i}$ of a particular user $u \in [N]$ for item $i \in \mathbb{N}$ is the preference $\xi_{\tau_U(u),\tau_I(i)}$ of the associated user type $\tau_U(u)$ for the item type $\tau_I(i)$ given the type matrix Ξ, *i.e.*, $L_{u,i} = \xi_{\tau_U(u),\tau_I(i)}$.

We assume that the entries of Ξ are i.i.d., $\xi_{k,j} = +1$ w.p. $1/2$ and $\xi_{k,j} = -1$ w.p. $1/2$. Generalizing our results to i.i.d. entries with bias p is relatively straightforward, but the independence assumption is quite strong and an important future research direction is to obtain results for more realistic preference matrices.

7.2.3 Two regimes of interest

Throughout this part of the thesis, we are going to assume that the number of user types, $\mathtt{q}_U$ and the number of item types $\mathtt{q}_I$ satisfy the property $\mathtt{q}_U, \mathtt{q}_I > 40 \log N$. Note that the logarithmic dependency on N is rather a weak assumption. The statement of results as follows would holds by making these assumptions. But, we can additionally show the optimality of the proposed algorithms up to multiplicative logarithmic factors if there exists constants $0 < \alpha, \beta < 1$ such that $\mathtt{q}_U = N^\alpha$ and $\mathtt{q}_I = N^\beta$.

Two specific parameter regimes play a central role in this paper, capturing settings with structure only in user space or only in item space. As described in Section 7.3,

N	Number of users
T	Time horizon
q_U	Number of user types
q_I	Number of item types
$\tau_U(u)$	User type of user u
$\tau_I(i)$	Item type of item i
$a_{u,t}$	item recommended to user u at time t
$L_{u,i}$	Rating of user u for item i
$\xi_{k,j}$	Preference of user type k for item type j

Figure 7.2.1: Notation for the recommendation system model.

either user-user or item-item collaborative filtering is almost optimal in the corresponding regime.

Definition 7.1 (User structure only ($\mathsf{q}_I = 2^{\mathsf{q}_U}$))**.** The *user-structure model* refers to the case that there is no structure in the item space. To simplify matters, we assume that the type matrix $\Xi \in \{-1,+1\}^{\mathsf{q}_U \times 2^{\mathsf{q}_U}}$ is *nonrandom* and has all sequences in $\{-1,+1\}^{\mathsf{q}_U}$ in its columns. Effectively the same type matrix would arise (with high probability) with i.i.d. entries as specified above if q_I is much larger than 2^{q_U}.

Definition 7.2 (Item structure only ($\mathsf{q}_U = N$))**.** The *item-structure model* refers to the case that there is no structure in the user space. This happens roughly when q_U is much larger than N, since then most user types have no more than one user. For the purpose of proving near-optimality of item-item collaborative filtering, it suffices to take $\mathsf{q}_U = N$ (and we do so).

7.3 Main results

In this paper, we analyze a version of each of user-user and item-item collaborative filtering within the general setup described in Section 7.2. The resulting regret bounds appear as Theorems 7.3 and 7.5. These theorems are complemented by information theoretic lower bounds showing that no other algorithm can achieve better than this asymptotic regret bound (up to logarithmic factors) in the specific extreme parame-

ter regimes user-structure only and item-structure only. Next, we propose a hybrid algorithm which is almost optimal in all parameter regimes.

User-user collaborative filtering. User-user collaborative filtering exploits structure in the user space: the basic idea is to recommend items to a user that are liked by similar users. We analyze an instance of user-user CF described in Section 8.1.1, obtaining the regret bound given in the following theorem. Essentially, the algorithm clusters users according to type by recommending random items for an initial phase, and then uses this knowledge to efficiently explore each user type's preferences.

The random recommendations during the initial phase incur regret with slope 1/2. Subsequently, the users are clustered according to type. Recommending an item to $\mathtt{q}_U$ users, one from each type, gives us the preferences for all N users for this item, and each such recommendation is disliked with probability 1/2. This results in a slope of $\mathtt{q}_U/2N$ for regret in the second phase of the algortihm.

Theorem 7.3. *Consider the recommendation system model described in Section 7.2 with N users, $\mathtt{q}_U$ user types and $\mathtt{q}_I > 6\log(N\mathtt{q}_U^2)$ item types. There exists a numerical constant c so that Algorithm 8.1 achieves regret*

$$\mathtt{regret}(T) \leq \begin{cases} \frac{T}{2}\,, & \text{if } T \leq c\log N \\ \frac{c\log N}{2} + \frac{\mathtt{q}_U}{2N}T\,, & \text{if } T > c\log N\,. \end{cases}$$

When there is no structure in the item space, with $\mathtt{q}_U = N^\alpha$ user types for $0 < \alpha < 1$, the user-user CF algorithm is optimal up to a multiplicative constant.

Theorem 7.4. *There exist numerical constants c (same as in the previous theorem) and c_1 such that in the user structure model (Defn 7.1), any recommendation algorithm must incur regret*

$$\mathtt{regret}(T) \geq \begin{cases} 0.49T - 4\,, & \text{if } T \leq c\log \mathtt{q}_U \\ c_1(\log \mathtt{q}_U + \frac{\mathtt{q}_U}{N}T)\,, & \text{if } T > c\log \mathtt{q}_U\,. \end{cases}$$

The reasoning for the first part of the lower bound is as follows. If a user has been recommended fewer than $\log \mathsf{q}_U$ items, then its similarity with respect to other users cannot be determined. This implies that any recommendation made to this user has uncertain outcome. The second part of the lower bound is obtained by showing that when an item is recommended for the first time to a user from a given user type the outcome of that recommendation is uncertain, and lower bounding the number of such recommendations. This is where we use the condition that each item is recommended at most once to each user.

The lower bound shows that the poor initial performance of the user-user CF, as bad as simply recommending random items, is unavoidable in the setting with only user structure. In the recommendation systems literature the notion of *cold start* describes the difficulty of providing useful recommendations when insufficient information is available about user preferences. In [62] a formal definition of cold start time is provided, and it is shown that a version of item-item CF obtains much smaller cold start time than user-user CF in a model with item structure. Our results on item-item CF, described next, substantiate this.

Item-item collaborative filtering. Item-item collaborative filtering exploits structure in the item space: users are recommended items similar to those they have liked. We analyze an instance of item-item CF in Section 9.1.1, obtaining the regret bound given in the following theorem. The proposed algorithm recommends a few random items to all users. This is referred to as *self-exploration* of the item space by the users, because every user learns its preference for the types of explored items, which we will call *representatives*. The algorithm then compares a set of items to the representatives by recommending each to a subset of random users. We refer to this phase as *joint-exploration* of the item space, because the effort of classifying the items is shared among all the users. This yields a set of item clusters, one for each explored type. If the number of representatives is smaller than q_I, the number of item types, then the portion of items not matching any representative will be cast aside. Subsequently, users are recommended items from the clusters corresponding to liked types.

The number of items participating in the partitioning process is chosen so that there will be enough items to be recommended to each user by the horizon T. There is a tradeoff between the cost of paritiotning (the number of uncertain recommendations made to do the partitioning) and the cost of learning the item types. The choice of fraction of item space to be explored by all users is chosen by as the optimal point in the tradeoff to minimize regret as a function of the horizon T, the number of item types $\mathtt{q}_I$ and the total number of users N.

Theorem 7.5. *Consider the item-structure model with $\mathtt{q}_I = N^\beta$ item types. There are numerical constants c_1, c_2, c_3, c_4 and c_5 such that provided $N > N_0(\beta)$ Algorithm 9.1 obtains regret per user at time T upper bounded as*

$$\mathtt{regret}(T) \leq c_1 + c_2 \begin{cases} \sqrt{\frac{\mathtt{q}_I \log N}{N} T \log T}, & \text{if } c_3 \frac{N}{\mathtt{q}_I \log N} < T \log T < c_4 \frac{\mathtt{q}_I N}{\log N} \\ \frac{\log N}{N} T \log T, & \text{if } c_4 \frac{\mathtt{q}_I N}{\log N} \leq T \log T\,. \end{cases}$$

Theorem 7.6. *In the item-structure model with $\mathtt{q}_I = N^\beta$, there exist numerical constants c_1, c_2, c_3 and c_4 (possibly function of β) such that any recommendation algorithm must incur regret*

$$\mathtt{regret}(T) > \begin{cases} c_1 \sqrt{\frac{T \mathtt{q}_I \log \mathtt{q}_I}{N}}, & \text{if } T < c_4 \frac{\mathtt{q}_I N}{\log \mathtt{q}_I} \\ c_2 T \frac{\log \mathtt{q}_I}{N}, & \text{if } c_4 \frac{\mathtt{q}_I N}{\log \mathtt{q}_I} < T \end{cases}$$

The proof idea is based on two observations: The algorithms working based on structure in item space do not suffer from the cold-start time. Even in very short time horizon, they can guarantee nontrivial bounds on regret. In particular, the optimal algorithm proposed here, suffers from a constant value of regret for an initial period. Note that increasing the number of users, does not increase the upper bound on regret in this initial phase, but it increases the length of this phase. This implies that increasing the number of users makes recommending valuable items easier. This is the phenomena observed as time grows in the algorithm.

Hybrid collaborative filtering algorithm. A few previous works has been done to address the simultaneous utilization of item and user space in CF algorithm [69, 66]. But to the best of our knowledge, there has not been any proof of optimality of any algorithm based on the joint structure before.

Our proposed algorithm based on joint model is very similar to the item-item algorithm. It starts off by learning some carefully chosen number of item types, ℓ. For large T (long horizon) in which ℓ is large, after some time, the structure in user space becomes available for free. At that point the algorithm partition users in groups. This makes learning new item types much cheaper for the algorithm. At the exploitation phase, the algorithm recommends an item to a user if that user or someone with the same type has liked that item or some item with the same type before.

Theorem 7.7. *Given numerical constants c_1, c_2, c_3, C_1 and C_2, Algorithm 10.1 obtains regret per user at time T upper bounded as*

$$\texttt{regret}(T) \leq C_1 + C_2 \begin{cases} \log T, & \text{if } \; T \leq c_1 \frac{(\log T)^2 N}{q_I \log N} \\ \sqrt{\frac{q_I \log N}{N} T}, & \text{if } \; c_1 \frac{(\log T)^2 N}{q_I \log N} \leq T \leq c_2 \frac{N \log N}{q_I} \\ \log N (1 - \frac{q_U}{N}) + \frac{\sqrt{q_U q_I \log N}}{N} T, & \text{if } \; c_2 \frac{N \log N}{q_I} \leq T \leq c_3 \frac{q_I q_U}{\log N} \\ \log N + \frac{q_U}{N}(q_I - \log N) + \frac{q_I \log N}{N} T, & \text{if } \; c_3 \frac{q_I q_U}{\log N} \leq T \,. \end{cases}$$

The strategy for finding a lower bound for regret in the joint model is based on first, characterizing the recommendations whose outcome are uncertain. Then we carefully use a counting strategy in addition to some statistical tools to find a lower bound on the total number of uncertain recommendations.

Theorem 7.8. *In the joint structure model, given $N > 3 q_U \log(2 q_U)$, there are numerical constants c_1, c_2, c_3 and C such that any recommendation algorithm must incur*

regret

$$\texttt{regret}(T) \geq -1 + C \begin{cases} \sqrt{\frac{T\,q_I}{N}}, & \textit{if } T \leq c_1 \frac{N(\log q_U)^2}{q_I} \\ \frac{\sqrt{T\,q_I\,q_U}}{N}, & \textit{if } c_1 \frac{N(\log q_U)^2}{q_I} \leq T \leq c_2 \frac{q_I\,q_U}{(\log q_U)^2} \\ T\frac{\log q_I}{N}, & \textit{if } c_2 \frac{q_I\,q_U}{(\log q_U)^2} \leq T\,. \end{cases}$$

Chapter 8

User-user algorithm

This chapter focuses on exploiting the structure in user space in collaborative filtering. First, version of user-user algorithm is introduced and analyzed in Section 8.1. Later on, in Section 8.2, it is proved that the proposed algorithm is almost information-theoretically optimal in the setup based on user structure only, introduced in Definition 7.1.

8.1 User-user upper bound

In this section, we propose a version of user-user CF based on the latent variable model introduced in Section 7.2. To recover the structure in user space, the algorithm recommends random items to all users. Then the users are partitioned based on the similarity of their ratings. From then on, the effort of exploring new items is shared by the users who are in the same partition.

8.1.1 Algorithm

We propose the following algorithm (see Algorithm 8.1). In Step 1, random items are recommended to all of the users. The ratings of these items are used to construct a partition over the users. This operation recovers the user types correctly with high probability. In Step 2, users are recommended new random items until there is an

item that is liked. If the partition accurately describes the type, then this item is also liked by everyone in the partition. The liked item is then recommended to all other users in the same partition (*exploitation*). Step 2 (find and recommend items) is repeated indefinitely.

Algorithm 8.1 USER-USER(T, q_U, N)

Step 1: partition users

$\epsilon \leftarrow \frac{1}{2N}$, $\mathsf{r} \leftarrow \lceil 2\log(\frac{\mathsf{q}_U^2}{\epsilon})\rceil$
for $t = 1, \cdots, \mathsf{r}$ **do**
    Pick random item i
    $a_{u,t} \leftarrow i, \quad \forall u \in [N]$
Partition users into fewest possible groups such that each group agrees on all items. Let $\widehat{\tau}_U(u) \in [\mathsf{q}_U]$ be the label of user u's partition.
$\mathcal{P}_k = \{u \in [N] : \widehat{\tau}_U(u) = k\}, \quad \forall k \in [\mathsf{q}_U]$

Step 2: find and recommend items

$\mathcal{S}_k \leftarrow \varnothing, \quad \forall k \in [\mathsf{q}_U]$
for $\mathsf{r} < t \leq T$ **do**
    for $u \in [N]$ **do**
        if $\mathcal{S}_{\widehat{\tau}_U(u)} \setminus \{a_{u,1}, \ldots, a_{u,t-1}\} \neq \varnothing$ (i.e., u has not rated all of $\mathcal{S}_{\widehat{\tau}_U(u)}$) **then**
            $a_{u,t} \leftarrow$ an unrated item in $\mathcal{S}_{\widehat{\tau}_U(u)}$ (*exploit*)
        else
            $a_{u,t} \leftarrow$ random item not rated by any user in $\mathcal{P}_{\widehat{\tau}_U(u)}$ (*explore*)
            if $L_{u,a_{u,t}} = +1$ **then**
                $\mathcal{S}_{\widehat{\tau}_U(u)} \leftarrow \mathcal{S}_{\widehat{\tau}_U(u)} \cup \{a_{u,t}\}$

There are a few remarks regarding the algorithm:

- All the users with the same type rate all items similarly. Hence, users of the same type are always in the same group after the partitioning. In contrast, two or more user types might rate items in Step 1 of the algorithm similarly. This causes error in the partitioning as users of two distinct user types fall into a single group.

- The above remark implies that the total number of user groups in the partitioning is no more than q_U.

- The labeling of user groups in partitioning is unimportant in the implementation of the algorithm and it can be different from the type of users.

- In step 2 of the algorithm, the sets of items $\mathcal{S}_k$'s at each time contain the items exploitable by users in the k's group of users in the partition. The algorithm expects that all users in k-th group are going to like items in $\mathcal{S}_k$.

Theorem 8.1. *Consider the model introduced in Section 7.2 with N users, q_U user types and $\mathsf{q}_I > 6 \log N \mathsf{q}_U^2$ item types. Let $\mathsf{r} = \lceil 2 \log(\mathsf{q}_U^2/\epsilon) \rceil$ with $\epsilon = 1/(2N)$. Then Algorithm 8.1 achieves regret*

$$\texttt{regret}(T) \leq \begin{cases} \frac{1}{2}T\,, & \text{if } T \leq \mathsf{r} \\ \frac{1}{2}\mathsf{r} + \frac{\mathsf{q}_U+1}{N}T + 1\,, & \text{if } T > \mathsf{r}\,. \end{cases}$$

The above theorem states that up until time r, algorithm is making meaningless recommendations. After that, the algorithm achieves the asymptotic slope performance for which on average, q_U recommendations out of N are random.

We first bound the probability that the partition created by the algorithm is correct.

Lemma 8.2. *Let $B_{uv} = \{\mathbb{1}_{\{\widehat{\tau}_U(u)=\widehat{\tau}_U(v)\}} = \mathbb{1}_{\{\tau_U(u)=\tau_U(v)\}}\}$ be the event that users u and v are partitioned correctly in Step 1 of Algorithm* User-User*. If $\mathsf{q}_I > 6 \log \mathsf{q}_U^2/\epsilon$, then $\mathbb{P}[B_{uv}^c] \leq 2\epsilon/\mathsf{q}_U^2$. It follows that if $B = \bigcap B_{uv}$ is the event that all users are partitioned correctly, then $\mathbb{P}[B^c] \leq 2\epsilon$.*

Proof. If $\mathsf{q}_I > 6 \log \mathsf{q}_U^2/\epsilon$, then using Lemma A.4 with probability at least $1 - \exp(-\log \mathsf{q}_U^2/\epsilon) \geq 1 - \epsilon/\mathsf{q}_U^2$ there are $\log \mathsf{q}_U^2/\epsilon$ items with distinct types among the r items rated by each user. Two user types $k \neq k'$ rate s independently chosen items of distinct types in the same way with probability 2^{-s}. Plugging in $s \geq \log \frac{\mathsf{q}_U^2}{\epsilon}$ and the above statement give the first statement of lemma. The second statement in the lemma follows by the union bound. □

Proof of Theorem 8.1. For $t \leq \mathsf{r}$, the algorithm is recommending random independent items to all users, so for these times $\mathbb{P}[L_{u,a_{u,t}} = -1] = 1/2$. Hence for $T \leq \mathsf{r}$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{N} \sum_{u=1}^{N} \mathbb{1}[L_{u,a_{u,t}} = -1]\right] = \sum_{t=1}^{T} \frac{1}{N} \sum_{u=1}^{N} \mathbb{P}[L_{u,a_{u,t}} = -1] = \sum_{t=1}^{T} \frac{1}{2} = \frac{T}{2}\,. \tag{8.1}$$

At $t \geq \mathsf{r}$, by Lemma 8.2, the partitioning step recovers the user types correctly with probability $\mathbb{P}[B] \geq 1 - \epsilon$. On event B, since items in $\mathcal{S}_{\widehat{\tau}_U(u)}$ are liked by users of the same type as u and hence also by u,

$$\mathbb{E}\left[\sum_{t=r+1}^{T} \sum_{u \in [N]} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{S}_{\widehat{\tau}_U(u)}] \middle| B\right] = 0\,.$$

Because there are TN terms in the sum and $\mathbb{P}[B^c] \leq 2\epsilon$, it follows by law of total probability that

$$\mathbb{E}\left[\sum_{t=r+1}^{T} \sum_{u \in [N]} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{S}_{\widehat{\tau}_U(u)}]\right] \leq 2TN\epsilon\,. \tag{8.2}$$

Now, for $t > \mathsf{r}$ if $a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}$, then $a_{u,t}$ is an independent new random item (and in particular has uniformly random type $\tau_I(a_{u,t})$). Hence,

$$\mathbb{P}[L_{u,a_{u,t}} = -1 | a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] = \mathbb{P}[L_{u,a_{u,t}} = +1 | a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] = \frac{1}{2}.$$

This means that it suffices to bound the contribution from the case $L_{u,a_{u,t}} = +1$, as expressed in the following claim:

Claim 8.3. *The number of liked 'explore' recommendations at $t > \mathsf{r}$ can be bounded as*

$$\sum_{t=\mathsf{r}+1}^{T} \sum_{u \in [N]} \mathbb{1}[L_{u,a_{u,t}} = +1, a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] \leq T\,\mathsf{q}_U + N.$$

Proof. For user partition k and time t, define $\mathcal{S}_k^t$ to be the set of items denoted by $\mathcal{S}_k$ in the algorithm at time t, after making the set of time t recommendations.

An item is added to $\mathcal{S}_k$ precisely on the event $\{t > \mathsf{r}, \widehat{\tau}_U(u) = k, L_{u,a_{u,t}} = +1, a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}\}$. Thus

$$\sum_{t=\mathsf{r}+1}^{T} \sum_{u:\widehat{\tau}_U(u)=k} \mathbb{1}[L_{u,a_{u,t}} = +1, a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] = |\mathcal{S}_k^T|. \tag{8.3}$$

Now, the number of items added to $\mathcal{S}_k$ at time t is

$$|\mathcal{S}_{k,t}| - |\mathcal{S}_{k,t-1}| = \sum_{u:\widehat{\tau}_U(u)=k} \mathbb{1}[L_{u,a_{u,t}} = +1, a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] \leq \sum_{u:\widehat{\tau}_U(u)=k} \mathbb{1}[a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] \,.$$

If $|\mathcal{S}_k^{t-1}| \geq t$, then $\mathcal{S}_k^{t-1} \setminus \{a_{u,1}, \cdots, \cdots, a_{u,t-1}\} \neq \varnothing$. The exploration event $\{a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}\}$ happens only if there are no items left to exploit, i.e., $\mathcal{S}_k^{t-1} \setminus \{a_{u,1}, \cdots, \cdots, a_{u,t-1}\} = \varnothing$. Hence,

$$|\mathcal{S}_k^t| - |\mathcal{S}_k^{t-1}| \leq \sum_{u:\widehat{\tau}_U(u)=k} \mathbb{1}[a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}] \quad \begin{cases} = 0, & \text{if } |\mathcal{S}_k^{t-1}| \geq t \\ \leq |\mathcal{P}_k|, & \text{if } |\mathcal{S}_k^{t-1}| < t \,. \end{cases} \tag{8.4}$$

The bound $|\mathcal{P}_k|$ comes from the fact that each user in the partition is recommended one item in a given time-step. So, at most $|\mathcal{P}_k|$ users explore new items at a given time step.

Let $t^* = \max\{t : \mathsf{r} \leq t \leq T, |\mathcal{S}_k^t| < t\}$. Note that the set over which we take the maximum is nonempty since $|\mathcal{S}_k^{\mathsf{r}}| = 0$. By definition of t^*, $|\mathcal{S}_k^{t^*}| < t^* \leq T$. By (8.4), $|\mathcal{S}_k^{t^*+1}| - |\mathcal{S}_k^{t^*}|$ may be nonzero (and bounded by $|\mathcal{P}_k|$), but $|\mathcal{S}_k^{t+1}| - |\mathcal{S}_k^t| = 0$ for $t^* < t < T-1$, since using thedefinition of t^*, for $t > t^*$ we have $|\mathcal{S}_k^t| \geq t$. Thus

$$|\mathcal{S}_k^T| = |\mathcal{S}_k^{t^*}| + \sum_{t=t^*}^{T-1} (|\mathcal{S}_k^{t+1}| - |\mathcal{S}_k^t|) \leq T + |\mathcal{P}_k| \,.$$

Summing this inequality over the (at most) q_U partition indices proves the claim. □

We can now complete the proof of Theorem 8.1. By the preceding claim and (8.2) we get the bound

$$\begin{aligned} \mathbb{E}\left[\sum_{t=r+1}^{T} \sum_{u\in[N]} \mathbb{1}[L_{u,a_{u,t}} = -1]\right] &= \mathbb{E}\left[\sum_{t=r+1}^{T} \sum_{u\in[N]} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \notin \mathcal{S}_{\widehat{\tau}_U(u)}]\right] \\ &\quad + \mathbb{E}\left[\sum_{t=r+1}^{T} \sum_{u\in[N]} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{S}_{\widehat{\tau}_U(u)}]\right] \\ &\leq \mathsf{q}_U T + N + 2TN\epsilon \leq \mathsf{q}_U T + N + T \,, \end{aligned}$$

where in the last step we used the choice $\epsilon = 1/N$. For $T > \mathsf{r}$, we can now bound the regret by combining (8.1) with the previous display:

$$\begin{aligned}\texttt{regret}(T) &= \mathbb{E}\left[\sum_{t=1}^{r}\frac{1}{N}\sum_{u=1}^{N}\mathbb{1}[L_{u,t}=-1]\right] + \mathbb{E}\left[\sum_{t=r+1}^{T}\frac{1}{N}\sum_{u=1}^{N}\mathbb{1}[L_{u,a_{u,t}}=-1]\right] \\ &\leq \frac{1}{2}\mathsf{r} + \frac{\mathsf{q}_U+1}{N}T + 1.\square\end{aligned}$$

8.2 User-structure only: lower bound

8.2.1 Proof strategy

The high level idea in providing a lower bound for the regret of any online recommendation system based on the user-user model emerges from two main observations:

- *A good estimate of user types is necessary to make meaningful recommendations:*

  We show this necessity by using the concept of $(t,\epsilon)-$ column regularity as in Definition 8.5. With proper choice of t and ϵ, the preference matrix is column regular with high probability. This implies that for any subset of t items, user types are behaving almost uniformly and independently at random (within a factor of ϵ deviation). Using this property, we show that if one knows the preference of a given user for $t-1$ items, the preference of that given user for the $t-$th item is almost independent and uniformly distributed on $\{-1,+1\}$.

  Knowing the preference of a user for $t-1$ items corresponds to having a set of candidate user types as its type. If t is small, this set is large (Imagine, if a user has rated only one item, even given the preference matrix, there are almost $\mathsf{q}_U/2$ candidate user types for this user). If the cardinality of the candidate user types is large (meaning that a good estimation of user type is not available), the (t,ϵ) column regularity ensures that a good prediction of the preference of user for a new item is impossible (even if the preference matrix is given to the algorithm).

  It is shown that it takes no less than $\log \mathsf{q}_U$ recommendation to learn the user

types. This matches the information theoretic bound in source coding problem in which a random variable (user type of a user) with equiprobable distribution on q_U values requires at least $\log \mathsf{q}_U$ binary integer valued code.

- *Knowing user types, to know the preference of a user for an item, the preference of that user type for the item item should be learned:* In contrast, if the algorithm is given the user types of all users, still the preference of each user type for a specific new item should be learned to predict the preference of users for new items. This task is easily done by recommending a specific item to *one* user from a given user type. The rating of that (representative user) for that item would be the same as the preference of all other users of the same type for that item.

8.2.2 Lower bound theorem statement

Theorem 8.4. *Let $\delta > 0$ and define $\mathsf{r} = \lfloor \log \mathsf{q}_U - \log\left(6 \log \mathsf{q}_U \log \frac{2N \log \mathsf{q}_U}{\delta}\right) \rfloor$. In the user structure model with N users and q_U user types, any recommendation algorithm must incur regret*

$$\texttt{regret}(T) \geq \max\left\{\frac{\mathsf{q}_U}{2N} T, \frac{1-\delta}{2}\left[T - (T - \mathsf{r})_+\right] - 4\right\}.$$

We separately prove the two lower bounds in the maximum, starting with the second. The following lemma enforces a certain regularity property in submatrices of the preference matrix.

Definition 8.5 ((t, ϵ)-column regularity)**.** Let $A \in \{-1, +1\}^{m \times n}$. For ordered tuple of distinct (column) indices $w = (i_1, \ldots, i_t) \in [n]^t$, let $M = (A_{\cdot i})_{i \in w} \in \{-1, +1\}^{m \times t}$ be the matrix formed by concatenating columns of A indexed by w. For given row vector $b \in \{-1, +1\}^t$, let $K_{b,w}(A) \subseteq [m]$ be the set of rows in $M = (A_{\cdot i})_{i \in w}$ which are identical to the row b . The cardinality of $K_{b,w}(A)$ is denoted by $k_{b,w}(A)$. A is said

to be (t,ϵ)-*column regular* if

$$\max_{w,b}\left|k_{b,w}(A)-\frac{m}{2^t}\right|\le\epsilon\frac{m}{2^t}\,,$$

where the maximum is over tuples w of t columns and ±1 vectors b of size t.

We define $\Omega_{t,\epsilon}$ to be the set of (t,ϵ)- column regular matrices.

Claim 8.6. *If a matrix $A\in\{-1,+1\}^{m\times n}$ is* (t,ϵ)-column regular*, then it is also* (s,ϵ)-column regular *for all $s<t$.*

Proof. Suppose that A is (t,ϵ)-column regular. By induction it suffices to show that A is $(t-1,\epsilon)$-column regular. We will check that $(1-\epsilon)\frac{m}{2^{t-1}}\le k_{b,w}(A)\le(1+\epsilon)\frac{m}{2^{t-1}}$ for all size $t-1$ tuples w and vectors b. For any given $w\in[n]^{t-1}$ and $b\in\{-1,+1\}^{t-1}$, let $b^+=[b\ 1]\in\{-1,+1\}^t$ be obtained from b by appending $+1$. Similarly b^- is obtained from b by appending -1. If $w'=(w,i)\in[n]^t$ for any $i\notin w$, then $K_{b,w}=K_{b^+,w'}\cup K_{b^-,w'}$ and $K_{b^+,w'}\cap K_{b^-,w'}=\varnothing$, so $k_{b,w}=k_{b^+,w'}+k_{b^-,w}$. Since A is (t,ϵ)-column regular, $(1-\epsilon)\frac{m}{2^t}\le k_{b^+,w'},k_{b-,w'}\le(1+\epsilon)\frac{m}{2^t}$, hence $(1-\epsilon)\frac{m}{2^{t-1}}\le k_{b,w}\le(1+\epsilon)\frac{m}{2^{t-1}}$. □

Lemma 8.7. *Let matrix $A\in\{-1,+1\}^{m\times n}$ have i.i.d.* Bern(1/2) *entries. Then A is (t,ϵ)-column regular with probability at least*

$$1-2(2n)^t\max\left\{\exp\left(-\frac{\epsilon^2}{3}\frac{m}{2^t}\right),\exp\left(-\frac{\epsilon}{2}\frac{m}{2^t}\right)\right\}.$$

Proof. Note that for given column tuple w and row vector b, the expected number of times the row vector b appears is $\frac{m}{2^t}$. Using a Chernoff bound (Lemma A.3),

$$\mathbb{P}\Big[\big|k_{b,w}(A)-\frac{m}{2^t}\big|\ge\epsilon\frac{m}{2^t}\Big]\le2\max\left\{\exp\left(-\frac{\epsilon^2}{3}\frac{m}{2^t}\right),\exp\left(-\frac{\epsilon}{2}\frac{m}{2^t}\right)\right\}.$$

There are no more than n^t possible choices of column tuple w, and 2^t possible choices of row vector b; the union bound yields the proof. □

Lemma 8.8. *Let $\delta>0$ and define $\mathsf{r}=\lfloor\log\frac{q_U}{6\log q_U\log(\frac{2N\log q_U}{\delta})}\rfloor$. For any $T\le\mathsf{r}$, the*

regret is lower bounded by

$$\texttt{regret}(T) \geq (\frac{1}{2} - \delta)T - 4\,.$$

Proof. The approach of the proof is to show that for preference matrices satisfying column regularity, at any time $t \leq \mathsf{r}$, most users have probability roughly half of liking any given item given the feedback obtained thus far, even if the type matrix is known.

At time t, suppose that n items in total have been sampled by the algorithm ($n \leq Nt$ since each of the N users can rate one item per time-step). We label the set of items to be $[n] = \{1, \ldots, n\}$ since *a priori* the items are identical. Let A be the $\mathsf{q}_U \times n$ matrix indicating the preference of each user type for these n items. Each item i has type $\tau_I(i) \sim \text{Unif}([2^{\mathsf{q}_U}])$ and because the set of columns of the type matrix Ξ is precisely $\{-1, +1\}^{\mathsf{q}_U}$, the columns of A are independent and uniformly distributed in $\{-1, +1\}^{\mathsf{q}_U}$.

We now focus on a particular user u. Let $w = \{a_{u,s}\}_{s \in [t-1]}$ be the items recommended to user u up to time $t-1$, and let $b = (L_{u,a_{u,s}})_{s \in [t-1]}$ be the vector of feedback from user u for these items. We claim that conditional on the matrix A, vector b and w, the type $\tau_U(u)$ of user u at the end of time instant $t-1$ is uniformly distributed over the set of user types $K_{b,w}(A)$ consistent with this data.

Let $b^+ = [b\ 1] \in \{-1, +1\}^s$ be obtained from b by appending $+1$. $L_{u,a_{u,t}} = +1$ precisely when $\tau_U(u) \in K_{b^+,\{w,a_{u,t}\}}(A)$, which in words reads "user u is among those types that are consistent with first $t-1$ ratings of u and have preference vector with '+1' for item recommended to u at time t". It follows that for any matrix A corresponding to items $[n]$,

$$\begin{aligned}\mathbb{P}[L_{u,a_{u,t}} = +1 | H_{t-1}, A] &= \mathbb{P}[\tau_U(u) \in K_{b^+,\{w,a_{u,t}\}}(A) | H_{t-1}, A] \\ &= \frac{k_{b^+,\{w,a_{u,t}\}}(A)}{k_{b,w}(A)}\,.\end{aligned}$$

The second equality is due to: i) w, b, and $a_{u,t}$ are functions of H_{t-1}; ii) for fixed

w and b the set $K_{b,w}(A)$ is determined by A; iii) $\tau_U(u)$ is uniformly distributed on $K_{b,w}(A)$.

Recall that we define $\Omega_{t,\epsilon}$ to be the set of (t,ϵ)- column regular matrices. It now follows by tower property of conditional expectation that

$$\begin{aligned}\mathbb{P}[L_{u,a_{u,t}} = +1 | A \in \Omega_{t,\epsilon}] &= \mathbb{E}\big[\mathbb{P}[L_{u,a_{u,t}} = +1 | A, H_{t-1}] | A \in \Omega_{t,\epsilon}\big] \\ &= \mathbb{E}\left[\frac{k_{b^+,\{w,i\}}(A)}{k_{b,w}(A)} \middle| A \in \Omega_{t,\epsilon}\right] \\ &= \frac{1}{2}\frac{1+\epsilon}{1-\epsilon} \leq \frac{1}{2}(1+4\epsilon)\,.\end{aligned}$$

The last two inequalities are justified as follows: if $A \in \Omega_{t,\epsilon}$ then by Claim 8.6, $A \in \Omega_{t-1,\epsilon}$. By Definition 8.5, this means that $k_{b,w}(A) \geq (1-\epsilon)m/2^{t-1}$ and $k_{b^+,\{w,i\}}(A) \leq (1+\epsilon)m/2^t$. We pick $\epsilon < 1/2$ to get the last inequality.

Fix $\delta > 0$ and

$$\epsilon = \sqrt{3\frac{2^t}{\mathsf{q}_U}\log\frac{(2N\log\mathsf{q}_U)^t}{\delta}}\,.$$

Lemma 8.7 shows that at time $t \leq \mathsf{r}$, $A \in \Omega_{t,\epsilon}$ for this choice of ϵ, with probability at least $1-\delta$. We get the bound

$$\mathbb{P}[L_{u,a_{u,t}} = +1] \leq \mathbb{P}[L_{u,a_{u,t}} = +1, A \in \Omega_{t,\epsilon}] + \mathbb{P}[A \notin \Omega_{t,\epsilon}] \leq \frac{1}{2}(1+4\epsilon) + \delta\,.$$

From the last display it follows that for $T \leq \mathsf{r}$,

$$\begin{aligned}\texttt{regret}(T) &= \frac{1}{N}\sum_{t\in[T],u\in[N]} \mathbb{P}[L_{u,a_{u,t}} = -1] \\ &\geq (\frac{1}{2}-\delta)T - \sum_{t\in[T]} 2\sqrt{3\frac{2^t}{\mathsf{q}_U}\log\frac{(2N\log\mathsf{q}_U)^t}{\delta}} \\ &\geq (\frac{1}{2}-\delta)T - 4\sqrt{3\frac{2^{T+1}}{\mathsf{q}_U}\log\frac{(2N\log\mathsf{q}_U)^{\log\mathsf{q}_U}}{\delta}} \\ &\geq (\frac{1}{2}-\delta)T - 4\,.\square\end{aligned}$$

Lemma 8.9. *For any T,* $\texttt{regret}(T) \geq \frac{q_U}{2N}T$.

Proof. Let $\mathcal{B}^t_{\tau_U(u),i}$ be the event that some user of same type $\tau_U(u)$ as u has rated item i by time $t-1$.

$$\mathcal{B}^t_{\tau_U(u),i} = \{\exists v \in [N], s \in [t-1] : \tau_U(v) = \tau_U(u), a_{v,s} = i\}.$$

Claim 8.10. *If no user with the same type as u has rated item i by time $t-1$, the probability that user u likes item i conditional on history is 1/2:*

$$\mathbb{P}(L_{u,i} = -1 | (\mathcal{B}^t_{\tau_U(u),i})^c, H_{t-1}) = \frac{1}{2}\,.$$

Proof. Let $\tau_U(\cdot)$ be the sequence of user types for all users. Let A be the matrix of size $\mathsf{q}_U \times n$ for $n < Nt$ of the ratings of all user types for all the n items observed by time t. Let $A_{\cdot\setminus\{k,i\}}$ denote the set of elements if matrix A except the preference of user type k for item i.

We first show that conditional on $\tau_U(\cdot), A_{\cdot\setminus\{k,i\}}$ and $(\mathcal{B}^t_{\tau_U(u),i})^c$, the random variables $L_{u,i}$ is statistically independent of H_{t-1} (for the history in which $(\mathcal{B}^t_{\tau_U(u),i})^c$ holds as we are conditioning on this):

$$L_{u,i} \perp\!\!\!\perp H_{t-1} | \tau_U(\cdot), A_{\cdot\setminus\{\tau_U(u),i\}}, (\mathcal{B}^t_{\tau_U(u),i})^c\,.$$

To see this, note that given $\tau_U(\cdot)$, we have $L_{u,i} = A_{\tau_U(u),i}$. Also, $H_{t-1} = \{a_{u',s}, L_{u',a_{u',s}}, u' \in [N], s < t\}$ such that for all u' with $\tau_U(u') = \tau_U(u)$, we know that $a_{u,s} \neq i$ under the event $(\mathcal{B}^t_{\tau_U(u),i})^c$. Hence, $\{L_{u',a_{u',s}}, u' \in [N], s < t\}$ is a deterministic function of $\tau_U(\cdot), A_{\cdot\setminus\{\tau_U(u),i\}}$. According to the model, for any s, given H_{s-1} the recommendation $a_{u',s}$ is a deterministic function of an independent random variable $\zeta_{u',s}$. Hence, using induction, on the event $(\mathcal{B}^t_{\tau_U(u),i})^c$, given $\tau_U(\cdot), A_{\cdot\setminus\{\tau_U(u),i\}}$, the history H_{t-1} is independent of everything else. □

The lemma will be proved by lower bounding the number of items recommended with such a large error probability. Using this and the fact that $a_{u,t}$ is independent

of all other variables given H_{t-1},

$$\frac{\mathbb{P}[L_{u,i} = -1, a_{u,t} = i, (\mathcal{B}^t_{\tau_U(u),i})^c | H_{t-1}]}{\mathbb{P}[a_{u,t} = i, (\mathcal{B}^t_{\tau_U(u),i})^c | H_{t-1}]} = \mathbb{P}[L_{u,i} = -1 | (\mathcal{B}^t_{\tau_U(u),i})^c, H_{t-1}] = \frac{1}{2}\,, \tag{8.5}$$

multiplying both sides by $\mathbb{P}[a_{u,t} = i, (\mathcal{B}^t_{\tau_U(u),i})^c | H_{t-1}]$ and summing over i gives

$$\mathbb{P}[L_{u,a_{u,t}} = -1, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c] = \frac{1}{2}\mathbb{P}[(\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c]\,. \tag{8.6}$$

Claim 8.11. *The total number of times a new item is recommended to any user type by time T is at least $\mathtt{q}_U T$:* $\sum_{t=1}^{T}\sum_{u=1}^{N} \mathbb{1}[(\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c] \geq \mathtt{q}_U T$

Proof. At the end of time-step T each user has been recommended T items, hence each *user type* has been recommended at least T items. There are $\mathtt{q}_U$ number of user types, so the total number of times an item is recommended to a user type for the first time is at least $\mathtt{q}_U T$. □

Combining inequalities gives the bound

$$\begin{aligned}
&N[T - \mathtt{regret}(T)] \\
&= \mathbb{E} \sum_{t\in[T],u\in[N]} \mathbb{1}[L_{u,a_{u,t}} = +1] \\
&= \mathbb{E} \sum_{t\in[T],u\in[N]} \mathbb{P}[L_{u,a_{u,t}} = +1, \mathcal{B}^t_{\tau_U(u),a_{u,t}} | H_{t-1}] \\
&\quad + \mathbb{E} \sum_{t\in[T],u\in[N]} \mathbb{P}[L_{u,a_{u,t}} = +1, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c | H_{t-1}] \\
&\overset{Eq.\ (8.6)}{\leq} \mathbb{E} \sum_{t\in[T],u\in[N]} \mathbb{P}[\mathcal{B}^t_{\tau_U(u),a_{u,t}} | H_{t-1}] + \frac{1}{2}\mathbb{P}[(\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c | H_{t-1}] \\
&= NT - \frac{1}{2}\mathbb{E} \sum_{\substack{t\in[T],u\in[N]\\ i\in\mathbb{N}}} \mathbb{P}[(\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c | H_{t-1}] \overset{Claim 8.11}{\leq} NT - \frac{\mathtt{q}_U}{2} NT\,.
\end{aligned}$$

Hence, $\mathtt{regret}(T) \geq \frac{\mathtt{q}_U}{2N} T$ for all T. □

Chapter 9

Item-item algorithm

This chapter focuses on exploiting the structure in item space in collaborative filtering. First, a version of item-item algorithm tailored to the model described in Section 7.2 in proposed. Later, it is proved that the performance of the proposed algorithm is almost information theoretically optimal in the setup based on item-structure only (as introduced in Definition 7.2) when $\mathsf{q}_I = N^\alpha$ for some $0 < \alpha < 1$.

9.1 Item-item upper bound

9.1.1 Algorithm

The high-level description of the algorithm is as follows (see Algorithm 9.1). First, items are partitioned according to type; next, each user's preference for each item type is determined; finally, items from liked partitions are recommended. These steps are next described in a bit more detail.

Two sets $\mathcal{M}_1$ and $\mathcal{M}_2$ each consisting of M random items are selected. In the exploration step, each item is recommended to $\mathsf{r} = \lceil 2\log(\mathsf{q}_I/\epsilon)\rceil$ random users. The feedback from these recommendations later helps to partition the items according to type. The parameter r is chosen large enough to guarantee small probability of error in partitioning ($\leq \epsilon$ for each item).

An item from each of ℓ *explored types* is recommended to all users, where ℓ is a

parameter determined by the algorithm. It turns out that it is often beneficial to learn user preferences for only a subset of the types, in which case ℓ is strictly less than q_I. Each of the ℓ items chosen from $\mathcal{M}_1$ is thought of as a *representative* of its type. For each representative item i_j, all items in $\mathcal{M}_1$ that appear to be of the same type as i_j are stored in a set $\widetilde{\mathcal{S}}_j$ and removed from $\mathcal{M}_1$. The same is done for $\mathcal{M}_2$. This guarantees that at each time $\mathcal{M}_1$ does not contain items with the same type as any of previously studied representative items. Hence, new representative items are of distinct item types as previous ones.

For each user u, items in groups with liked representative are added to a set of exploitable items $\mathcal{R}_u$. Finally, in the exploitation phase, each user is recommended items from $\mathcal{R}_u$ (exploitable items for user u). Parameter M is chosen to make sure there are enough exploitable items in $\mathcal{R}_u$ for all users by time T. Parameter ℓ is chosen to minimize regret for the given horizon T.

The algorithm uses some notation, which we define next. For an item i and time $t > 0$, $\mathsf{rated}_t(i) = \{u \in [N] : a_{u,s} = i \text{ for some } s < t\}$ is the set of users that have rated the item i before time t. We use the notation $\mathsf{rated}(i)$ in the algorithm for $\mathsf{rated}_t(i)$ at the time it is used.

Theorem 9.1. *Consider the model introduced in Subsection 7.2 with q_I and $\mathsf{q}_U > 18\log(3N)$, Algorithm 9.1 obtains regret per user at time T upper bounded as*

$$\mathtt{regret}(T) < 4 + 57\max\left\{\log T\,,\ \sqrt{3\frac{\mathsf{q}_I\mathsf{r}}{N}T},\ 53\frac{\log N}{N}T\right\}.$$

9.1.2 Performance analysis of the algorithm

Proof. The basic error event is misclassification of an item. In Algorithm 9.2, $\mathcal{S}_j$ is the set of items which the algorithm posits are of the same type as the jth representative i_j. Let $\mathcal{E}_i$ be the event that item i was mis-classified,

$$\mathcal{E}_i = \{\exists j : i \in \mathcal{S}_j, \tau_I(i) \neq \tau_I(\mathrm{i}_j)\}\,. \tag{9.1}$$

Algorithm 9.1 ITEM-ITEM(T, q_I, N) (fixed time horizon)

$\ell = \max\left\{4, \min\left\{\left\lceil \max\left\{18\log T, \sqrt{330\frac{\mathsf{q}_I \mathsf{r}}{N}T}\ \right\}\right\rceil, \mathsf{q}_I\right\}\right\}$

$\epsilon = \frac{1}{2\mathsf{q}_I N}$; $\mathsf{r} = \left\lceil 2\log\frac{\mathsf{q}_I}{\epsilon}\right\rceil$; $M = \left\lceil \frac{64\mathsf{q}_I}{\ell}T\right\rceil$

$\mathcal{M}_1 = M$ random items; $\mathcal{M}_2 = M$ new random items;

$\mathcal{R}_u \leftarrow \varnothing, \quad u \in [N]$ (items exploitable by user u)

$\{\mathcal{R}_u\}_{u\in[N]} \leftarrow$ ITEMEXPLORE$(\mathcal{M}_1, \mathcal{M}_2, \ell)$

ITEMEXPLOIT$(\{\mathcal{R}_u\}_{u\in[N]})$

Algorithm 9.2 ITEMEXPLORE$(\mathcal{M}_1, \mathcal{M}_2, \ell)$

$\mathcal{S}_j \leftarrow \varnothing, \quad j \in [\mathsf{q}_I]$ (initialize sets of items of type j in $\mathcal{M}_2$)

$\widetilde{\mathcal{S}}_j \leftarrow \varnothing, \quad j \in [\mathsf{q}_I]$ (initialize sets of items of type j in $\mathcal{M}_1$)

Rate items by a few users (preliminary exploration)

Recommend each item in $\mathcal{M}_1$ and $\mathcal{M}_2$ to r users from $[N]$. (This is done over $\lceil(|\mathcal{M}_1| + |\mathcal{M}_2|)\mathsf{r}/N\rceil \le 2M\mathsf{r}/N+1$ time-steps, with extra recommendations being random new items.)

Partition items and learn item types (joint exploration)

for $j = 1, \cdots, \ell$ **if** $\mathcal{M}_1 \neq \varnothing$ **do**

  $\mathrm{i}_j \leftarrow$ a random item in $\mathcal{M}_1$ (representative item)

  $a_{u,t} \leftarrow \mathrm{i}_j$ if $u \notin \mathsf{rated}_t(\mathrm{i}_j)$, otherwise a random item not from $\mathcal{M}_1$.

  $\widetilde{\mathcal{S}}_j \leftarrow \{i \in \mathcal{M}_1 : U = \mathsf{rated}(i) \cap \mathsf{rated}(\mathrm{i}_j), L_{U,i} = L_{U,\mathrm{i}_j}\}$

  $\mathcal{M}_1 \leftarrow \mathcal{M}_1 \setminus \widetilde{\mathcal{S}}_j$

  $\mathcal{S}_j \leftarrow \{i \in \mathcal{M}_2 : U = \mathsf{rated}(i) \cap \mathsf{rated}(\mathrm{i}_j), L_{U,i} = L_{U,\mathrm{i}_j}\}$ (users agree on i vs. i_j)

  $\mathcal{M}_2 \leftarrow \mathcal{M}_2 \setminus \mathcal{S}_j$

$\mathcal{R}_u = \bigcup_{j\in[\ell]: L_{u,\mathrm{i}_j}=+1} \mathcal{S}_j$ for each $u \in [N]$

return $\{\mathcal{R}_u\}_{u\in[N]}$

Algorithm 9.3 ITEMEXPLOIT$(\{\mathcal{R}_u\}_{u\in[N]})$

for remaining $t \le T$ **do**

  for $u \in [N]$ **do**

    if there is an item $i \in \mathcal{R}_u$ such that $u \notin \mathsf{rated}_t(i)$. **then**

      $a_{u,t} \leftarrow i$

    else $a_{u,t} \leftarrow$ a random item not yet rated by u.

The amount of time users spend in the exploration phase is denoted by T_0. We partition the set of recommendations made by the algorithm according to 1- whether or not recommendation was in the exploration phase; 2- whether or not the recommended item is in the list of exploitable items $\mathcal{R}_u$; and 3- whether or not the misclassification event $\mathcal{E}_{a_{u,t}}$ occurs for the recommended item:

$$\begin{aligned}
N\texttt{regret}(T) &= \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=1}^{T_0}\mathbb{1}[L_{u,a_{u,t}}=-1]\right] \\
&\quad + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\mathbb{1}[L_{u,a_{u,t}}=-1, a_{u,t}\notin\mathcal{R}_u]\right] \\
&\quad + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\mathbb{1}[L_{u,a_{u,t}}=-1, a_{u,t}\in\mathcal{R}_u, \mathcal{E}_{a_{u,t}}]\right] \\
&\quad + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\mathbb{1}[L_{u,a_{u,t}}=-1, a_{u,t}\in\mathcal{R}_u, (\mathcal{E}_{a_{u,t}})^c]\right] \\
&=: \texttt{A1}+\texttt{A2}+\texttt{A3}+\texttt{A4}\,. \qquad (9.2)
\end{aligned}$$

$\texttt{A1}$ is the regret from early time-steps before T_0. $\texttt{A2}$ is the regret due to not having enough items available for the exploitation phase, which is proved to be small with high probability for a good choice of M. $\texttt{A3}$ is the regret due to exploiting the misclassified items. It is small since few items are misclassified with the proper choice of ϵ and $\texttt{r}$. $\texttt{A4}$ is the regret due to exploiting the correctly classified items. We will see that $\texttt{A4}=0$.

Bounding $\texttt{A1}$. It takes at most $\lceil\frac{2M\texttt{r}}{N}\rceil$ units of time to rate each item in $\mathcal{M}_1$ and $\mathcal{M}_2$ by $\texttt{r}$ users from $[N]$ (since initially $|\mathcal{M}_1|=|\mathcal{M}_2|=M$). Then, ℓ representative items from different item types should be rated by every user which takes ℓ units of times. This gives

$$T_0 \le \left\lceil\frac{2M\texttt{r}}{N}\right\rceil + \ell \le \frac{2M\texttt{r}}{N}+\ell+1\,.$$

For $t\le T_0$, the algorithm recommends random items irrespective of feedback to all users. Since each item i has a uniformly distributed type $\tau_I(i)$ and any given user

likes type j with probability half,

$$\texttt{A1} = \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=1}^{T_0}\mathbb{1}[L_{u,a_{u,t}} = -1]\right] = \frac{1}{2}\sum_{u=1}^{N} T_0 \leq M\mathsf{r} + \frac{1}{2}(\ell+1)N\,. \tag{9.3}$$

Here, the last inequality uses the upper bound for T_0 given above.

Bounding `A2`**.** An item $a_{u,t} \notin \mathcal{R}_u$ is recommended to u in exploitation phase of the algorithm ($t > T_0$) only when all items in $\mathcal{R}_u$ have been recommended to this user before. So, the total number of times $a_{u,t} \notin \mathcal{R}_u$ is recommended at time interval $T_0 \leq t \leq T$, is at most $(T - |\mathcal{R}_u|)_+$. To observe that, note that if $|\mathcal{R}_u| \geq T$, then there are enough items in $\mathcal{R}_u$ up until time T, in which case $\sum_{t=T_0+1}^{T}\mathbb{1}[a_{u,t} \notin \mathcal{R}_u] = 0$. Alternatively, if $|\mathcal{R}_u| < T$, then the total number of times items in $|\mathcal{R}_u|$ are recommended to u in $t \leq T$, can be as large as $|\mathcal{R}_u|$ which implies that $\sum_{t=T_0+1}^{T}\mathbb{1}[a_{u,t} \notin \mathcal{R}_u] \leq (T - \mathcal{R}_u)_+$. Hence,

$$\begin{aligned}\texttt{A2} &= \mathbb{E}\left[\sum_{u\in[N]}\sum_{t=T_0+1}^{T}\mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \notin \mathcal{R}_u]\right] \leq \mathbb{E}\left[\sum_{u\in[N]}\sum_{t=T_0+1}^{T}\mathbb{1}[a_{u,t} \notin \mathcal{R}_u]\right] \\ &\leq \sum_{u\in[N]}\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] \leq 3TN\exp(-\ell/18) + 148T \end{aligned} \tag{9.4}$$

where the last inequality is from Lemma 9.4.

Bounding `A3`**.** Term `A3` in (9.2) is the expected number of mistakes made by the algorithm as a result of misclassification. Claim 9.2 upper bounds the expected number of "potential misclassifications" (defined in Equation (9.7)) in the algorithm to provide an upper bound for this value:

$$\begin{aligned}\texttt{A3} &= \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{R}_u, \mathcal{E}_{a_{u,t}}] \\ &\leq \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[a_{u,t} \in \mathcal{R}_u, \mathcal{E}_{a_{u,t}}] \overset{\text{claim9.2}}{\leq} 2NM\epsilon \leq \frac{64T}{\ell}\,. \end{aligned} \tag{9.5}$$

Bounding A4. By definition of event $\mathcal{E}_i$ given in Equation (9.1), if an item $i \in \mathcal{S}_j$ is correctly classified, then $\tau_I(i) = \tau_I(\mathrm{i}_j)$. According to the model introduced in Section 7.2, user preferences for an item depend only on the type of the item (since $L_{u,i} = \xi_{\tau_U(u),\tau_I(i)}$), so all users rate i the same as i_j. For an item $i \in \mathcal{R}_u$, there is some $j \in [\mathsf{q}_I]$ such that $i \in \mathcal{S}_j$ and u likes item i_j. Hence,

$$\begin{aligned}\mathbb{P}[L_{u,i} = -1 | i \in \mathcal{R}_u, (\mathcal{E}_i)^c] &= \mathbb{P}[L_{u,i} = -1 | \exists j : L_{u,\mathrm{i}_j} = +1, i \in \mathcal{S}_j, \tau_I(i) = \tau_I(\mathrm{i}_j)] \\ &= \mathbb{P}[L_{u,i} = -1 | \xi_{u,\tau_I(i)} = +1] = 0 \\ &\Rightarrow \mathtt{A4} = 0\,. \end{aligned} \tag{9.6}$$

Combining all the bounds. Plugging in Equation (9.3), (9.4), (9.5) and (9.6) into Equation (9.2) gives

$$N\mathtt{regret}(T) \leq M\mathsf{r} + \frac{1}{2}(\ell+1)N + 3TN\exp(-\frac{\ell}{18}) + 148T + \frac{64T}{\ell}\,.$$

Setting $M = \frac{64T\mathsf{q}_I}{\ell}$ gives

$$\begin{aligned}\mathtt{regret}(T) &\overset{(a)}{\leq} \frac{1}{2} + 165\frac{T\mathsf{q}_I\mathsf{r}}{\ell N} + \frac{1}{2}\ell + 3T\exp(-\ell/18) \\ &\overset{(b)}{\leq} 4 + 3T\exp(-\mathsf{q}_I/18) + 3\max\{18\log T, \sqrt{330\frac{\mathsf{q}_I\mathsf{r}}{N}T}, 330\frac{\mathsf{r}}{N}T\}\end{aligned}$$

where (a) is derived by $\ell > 4$, (b) is derived choosing the parameter ℓ to be

$$\ell = \max\left\{4, \min\left\{\left\lceil \max\left\{18\log T, \sqrt{330\frac{\mathsf{q}_I\mathsf{r}}{N}T}\ \right\}\right\rceil, \mathsf{q}_I\right\}\right\}.$$

Note that $\mathsf{r} = \log(N\mathsf{q}_I^2) \leq 3\log N$. For N large enough so that $\mathsf{q}_I > 18\log(3N)$, we have $3\exp(-\mathsf{q}_I/18) < 1/N$ and

$$\mathtt{regret}(T) < 4 + 57\max\left\{\log T\,,\ \sqrt{3\frac{\mathsf{q}_I\mathsf{r}}{N}T},\ 53\frac{\log N}{N}T\right\}\,. \qquad \square$$

Claim 9.2. *The expected number of total number of times a misclassified item is*

recommended in the exploitation step is upper bounded as

$$\mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[a_{u,t}=i, i\in\mathcal{R}_u, \mathcal{E}_i]\right] \leq NM\epsilon\,.$$

Proof. Event $\mathcal{E}_i$ (misclassifying item i by the algorithm) happens if preference of witnesses in classifying item i is similar to either of previous representative items with a different type. Given matrix Ξ, this event is a function of order of the choosing the representative items and the exact choice of witnesses. Analysis of probability of this event in its generality is not easy. Instead, we will define an event called "potential error event," $\bar{\mathcal{E}}_{i,U_i}$. We will show that by properly defining U_i, we have $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,U_i}$; misclassifying item i happens only if potential misclassification $\bar{\mathcal{E}}_{i,U_i}$ holds. Hence, an upper bound for $\mathbb{P}[\bar{\mathcal{E}}_{i,U_i}]$ (given in Claim 9.3) provides an upper bound for $\mathbb{P}[\mathcal{E}_i]$.

For item $i \in \mathcal{M}_2$ and subset of users $U_i \subseteq [N]$, define

$$\bar{\mathcal{E}}_{i,U_i} = \{\exists j \neq \tau_I(i) : L_{u,i} = \xi_{\tau_U(u),j}\,,\ \text{for all } u \in U_i\} \tag{9.7}$$

to be the event that the ratings of users in U_i for item i agree with a different type. For item $i \in \mathcal{S}_j$, let t be the time i was added to $\mathcal{S}_j$ in the exploration phase. Let $U_i = \mathsf{rated}_t(i) \cap \mathsf{rated}_t(\mathrm{i}_j)$ be the set of users whose ratings verified that i and i_j are of the same type. All users in U_i agree on i vs. i_j. Hence, if item $i \in \mathcal{S}_j$ is misclassified (which implies $\tau_I(i) \neq \tau_I(\mathrm{i}_j)$ using Equation (9.1)), then $\bar{\mathcal{E}}_{i,U_i}$ holds. So, $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,U_i}$.

$$\begin{aligned}
&\mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[a_{u,t}=i, i\in\mathcal{R}_u, \mathcal{E}_i] \\
&\qquad\overset{(a)}{\leq} \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0+1}^{T}\sum_{i\in\mathcal{M}_2}\mathbb{1}[a_{u,t}=i, i\in\mathcal{R}_u, \bar{\mathcal{E}}_{i,U_i}] \\
&\qquad\overset{(b)}{\leq} N\mathbb{E}\left[\sum_{i\in\mathcal{M}_2}\mathbb{1}[\bar{\mathcal{E}}_{i,U_i}]\right] \leq NM\max_{i\in\mathcal{M}_2}\mathbb{P}[\bar{\mathcal{E}}_{i,U_i}] \\
&\qquad\overset{(c)}{\leq} 2NM\epsilon\,.
\end{aligned}$$

(a) holds because for every user u, the set of items $\mathcal{R}_u$ is a subset of $\mathcal{M}_2$, also the

containment $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,U_i}$; (b) is true since each item i is recommended at most N times; (c) is proved in Claim 9.3. Also note that initially $|\mathcal{M}_2| = M$. □

Claim 9.3. *Consider the event $\bar{\mathcal{E}}_{i,U_i}$ defined in* (9.7) *with the set of users U_i defined immediately after. Then $\mathbb{P}[\bar{\mathcal{E}}_{i,U_i}] \leq 2\epsilon$ for all $i \in \mathcal{M}_2$.*

Proof. Each representative item studied in the exploration subroutine is rated by all users. Each item in $\mathcal{M}_1$ and $\mathcal{M}_2$ is rated by at least r users. Hence, there are at least r users who rate both item $i \in \mathcal{S}_j$ and i_j independent of feedback. Using Lemma A.4, given $\mathsf{q}_U > 3\mathsf{r}$, with probability at least $1 - \exp(-\mathsf{r}/2) \geq 1 - \epsilon/\mathsf{q}_I$ there are $\mathsf{r}/2$ users with distinct user types in U_i.

A set of $\mathsf{r}/2$ users with distinct types, chosen independently of the feedback, rate item i of type $\tau_I(i)$. Any two item types $j \neq j'$ have jointly independent user preferences $(\xi_{u,j})_{u\in[N]}, (\xi_{u,j'})_{u\in[N]}$, so two items of different types are rated in the same way by the $\mathsf{r}/2$ users of different types with probability at most $2^{-\mathsf{r}/2}$. The choice $\mathsf{r} \geq 2\log(\mathsf{q}_I/\epsilon)$ and a union bound over item types j completes the proof. □

Lemma 9.4. *For user $u \in [N]$, let R_u be defined as inline 10 of exploration algorithm. Then,*

$$\mathbb{P}[|\mathcal{R}_u| \leq T] \leq 3\exp(-\frac{\ell}{18}) + \frac{148}{N}$$

And,

$$\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] \leq 3T\exp(-\frac{\ell}{18}) + \frac{148}{N}T\,.$$

Proof. Roughly speaking, any user u likes about half of the ℓ item types learned by the algorithm. The total number of items in $\mathcal{M}_2$ is $\frac{64\mathsf{q}_I}{\ell}T$. Hence, there are about $\frac{64}{\ell}T$ items from each of the item types in $\mathcal{M}_2$. So, with high probability $\mathcal{R}_u$ being the items in $\mathcal{M}_2$ which user u is expected to like is more than T.

Making this argument rigorous requires careful reasoning for the following reasons: i) $\mathcal{R}_u$ is the union of $\mathcal{S}_j$'s with $L_{u,\mathrm{i}_j} = +1$. But, for a specific j such that $L_{u,\mathrm{i}_j} = +1$, due to misclassification of items, there might be some items of the same type as i_j which have been misclassified and removed from $\mathcal{S}_j$ before. ii) The choice of item types learned by the algorithm at each step depends on the number of remaining items in

each item type in $\mathcal{M}_1$. Again, due to the misclassification, this could be different from the actual number of items of each given type. Additionally, the probability of misclassifying an item of a given type depends on the ratings of a users for that item. Hence, the choice of next item type to be learned is *not* statistically independent of the rating of a given user u for a given item type. The effect of this dependency is addressed carefully in the proof of Claim 9.5.

For $u \in [N]$, we define

$$\widetilde{\mathcal{R}}_u = \{i \in \mathcal{M}_2 : \tau_I(i) = \tau_I(\mathrm{i}_j), \text{ for some } j \in [\ell] \text{ such that } L_{u,\mathrm{i}_j} = +1\} \tag{9.8}$$

to be the items in $\mathcal{M}_2$ whose types are similar to one of the i_j's which are liked by u. If there is no error in the classification of items, $\widetilde{\mathcal{R}}_u$ is equal to $\mathcal{R}_u$. It will be easier to bound $\widetilde{\mathcal{R}}_u$.

Note that if an item $i \in \widetilde{\mathcal{R}}_u$ is correctly classified by the algorithm, then $i \in \mathcal{R}_u$; so $i \in \widetilde{\mathcal{R}}_u \setminus \mathcal{R}_u$ implies $\mathcal{E}_i$ (misclassification defined in Equation (9.1)). Hence,

$$|\mathcal{R}_u| \geq |\widetilde{\mathcal{R}}_u| - \sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i]\,. \tag{9.9}$$

Observing immediately after Equation (9.9), $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,\mathcal{U}_i}$. Plugging $\mathbb{1}[\mathcal{E}_i] \leq \mathbb{1}[\bar{\mathcal{E}}_{i,U_i}]$ into Equation (9.9) gives

$$|\mathcal{R}_u| \geq |\widetilde{\mathcal{R}}_u| - \sum_{i \in \mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}]\,.$$

It follows that

$$\mathbb{P}\big[\,|\mathcal{R}_u| \leq T\,\big] \leq \mathbb{P}\big[\,|\widetilde{\mathcal{R}}_u| - \sum_{i \in \mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}] \leq T\,\big] \leq \mathbb{P}\big[\,|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T\,\big] + \mathbb{P}\big[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}] \geq \frac{T}{2}\,\big] \tag{9.10}$$

and

$$\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] \leq \mathbb{E}\left[(T - |\widetilde{\mathcal{R}}_u|)_+\right] + \mathbb{E}\left[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}]\right]. \tag{9.11}$$

We bound the second expectation of Equation (9.11) using Claim 9.3

$$\mathbb{E}\left[\sum_{i\in\mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}]\right] \leq 2|\mathcal{M}_2|\epsilon = \frac{64}{N\ell}T \tag{9.12}$$

and by Markov Inequality for any $s > 0$

$$\mathbb{P}\big[\sum_{i\in\mathcal{M}_2} \mathbb{1}[\bar{\mathcal{E}}_{i,\mathcal{U}_i}] \geq sT\big] \leq \frac{64}{s\ell N}\,. \tag{9.13}$$

We now bound tail probability of $|\widetilde{\mathcal{R}}_u|$. Remember $\widetilde{\mathcal{R}}_u$ is the set of items in $\mathcal{M}_2$ with types the same as representatives which are liked by user u. Let $\widetilde{\mathcal{R}}(j)$ be the set of items in $\mathcal{M}_2$ whose types are the same as i_j :

$$\widetilde{\mathcal{R}}(j) = \{i \in \mathcal{M}_2 : \tau_I(i) = \tau_I(\mathrm{i}_j)\}\,. \tag{9.14}$$

Note that the sets $\widetilde{\mathcal{R}}(j)$ are mutually exclusive. Let

$$\mathcal{L}_u^{(\ell)} = \{j \in [\ell] : L_{u,\mathrm{i}_j} = +1\} \tag{9.15}$$

be the set of learned item types in $[\ell]$ which are liked by user u. Note that the algorithm guarantees that each representative corresponds to a different item type. Using Equation (9.8),

$$\widetilde{\mathcal{R}}_u = \cup_{\substack{j\in[\ell] \\ L_{u,\mathrm{i}_j}=+1}} \widetilde{\mathcal{R}}(j)\,.$$

Hence,

$$|\widetilde{\mathcal{R}}_u| = \sum_{i\in\mathcal{M}_2} \mathbb{1}[\tau_I(i) \in \mathcal{L}_u^{(\ell)}].$$

Note that according to the definition of $\mathcal{L}_u^{(\ell)}$ given in Equation (9.15), $\mathcal{L}_u^{(\ell)}$ is determined by: i) u-th row of matrix Ξ (denoted by $\Xi_{u,\cdot}$) which determines whether $L_{u,\mathrm{i}_j} = \xi_{\tau_U(u),\tau_I(\mathrm{i}_j)} = +1$ or not; ii) the items in $\mathcal{M}_1$, their types and some randomness in the algorithm which determines the choice of i_j and $\tau_I(\mathrm{i}_j)$. Hence, the set $\mathcal{L}_u^{(\ell)}$ is statistically independent of all items $i \in \mathcal{M}_2$ and their item types. Hence, conditional

on any given $\mathcal{L}_u^{(\ell)}$, the types $\tau_I(i)$ for $i \in \mathcal{M}_2$ is independently uniformly distributed on $[\mathtt{q}_I]$. So, given $\mathcal{L}_u^{(\ell)}$ with size $\ell_u^{(\ell)} = |\mathcal{L}_u^{(\ell)}|$, $|\widetilde{\mathcal{R}}_u|$ is the sum of $|\mathcal{M}_2|$ i.i.d. binary random variables; it is a Binomial random variable with mean $M\frac{\ell_u}{\mathtt{q}_I} = 64T\frac{\ell_u}{\ell}$. Using Chernoff bound given in Lemma A.3, condition on $\ell_u \geq \frac{\ell}{30}$,

$$\mathbb{P}\left[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T \,\middle|\, \ell_u^{(\ell)} \geq \ell/30\right] \leq \exp(-T/11)\,.$$

In Claim 9.5, we will show that $\ell_u^{(\ell)} > \ell/30$ with high probability. The above inequality and the statement of Claim 9.5 give

$$\mathbb{P}[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T] \leq \exp(-T/11) + \exp(-\ell/5) + \exp(-\mathtt{q}_I/18) + \frac{20}{N}\,.$$

Plugging this and Equation (9.13) into Equation (9.10) gives

$$\mathbb{P}[|\mathcal{R}_u| < T] \leq \exp(-T/11) + \exp(-\ell/5) + \exp(-\mathtt{q}_I/18) + \frac{20}{N} + \frac{128}{\ell N}\,.$$

Since $\ell < T$ and $\ell < \mathtt{q}_I$,

$$\mathbb{P}[|\mathcal{R}_u| < T] \leq 3\exp(-\frac{\ell}{18}) + \frac{148}{N}\,. \qquad \square$$

Claim 9.5. *For any $u \in [N]$, and $\ell_u^{(\ell)} = |\mathcal{L}_u^{(\ell)}|$ with $\mathcal{L}_u^{(\ell)}$ defined in Equation* (9.15) *to be the set of learned item types which u likes, then*

$$\mathbb{P}\big[\,\ell_u^{(\ell)} < \ell/30\,\big] < \exp(-\ell/5) + \exp(-\mathtt{q}_I/18) + \frac{20}{N}\,.$$

Proof. For given user u, the parameter $\ell_u^{(\ell)}$ denotes the number of learned item types liked by user u. The set of item types learned by the algorithm is a function of the number of items of a given item type in $\mathcal{M}_1$ at each step; misclassifying items in previous steps changes the number of items of a given type at a given time. Also, the probability of misclassifying an item from a given type depends on the preference of users for that item type. Hence, whether or not user u likes an item type is statistically

dependent on whether or not that item type is chosen as one of the learned item types. We will show that the effect of this dependency is small; subsequently, we find upper bounds for probability of $\ell_u^{(\ell)} < \ell/30$.

Let $\mathcal{L}_u$ be the item types in $[\mathsf{q}_I]$ which are liked by user u.

$$\mathcal{L}_u = \{j \in [\mathsf{q}_I] : \xi_{u,j} = +1\} \tag{9.16}$$

Since the elements of matrix Ξ (determining $\mathcal{L}_u$) are independetly Bern(1/2) distributed, using Chernoff bound in Lemma A.3 gives

$$\mathbb{P}[|\mathcal{L}_u| < \mathsf{q}_I/3] < \exp(-\mathsf{q}_I/18)\,. \tag{9.17}$$

We will show that

$$\mathbb{P}\big[\ell_u^{(\ell)} \leq \ell/30 \,\big|\, |\mathcal{L}_u| > \mathsf{q}_I/3\big] \leq \exp(-\ell/5) + \frac{20}{N} \tag{9.18}$$

To do so, let the sequence of random variables $X_1 = \tau_I(\mathrm{i}_1), X_2 = \tau_I(\mathrm{i}_2), \cdots$ denote the item types chosen to be learned by the algorithm. Then, $|\mathcal{L}_u^{(\ell)}| = \sum_{j\in[\ell]} \mathbb{1}[X_j \in \mathcal{L}_u]$. Note that the set $\mathcal{L}_u$ is a one-to-one function of $u-$th row of matrix Ξ (denoted by $\Xi_{u,\cdot}$). Define $\bar{\mathcal{R}}(j)$ to be the set of items in $\mathcal{M}_1$ of type j

$$\bar{\mathcal{R}}(j) = \{i \in \mathcal{M}_1 : \tau_I(i) = j\}\,. \tag{9.19}$$

Define the error event $\widetilde{\mathcal{E}}_i$ for $i \in \mathcal{M}_1$ similar to the event $\mathcal{E}_i$ for $i \in \mathcal{M}_2$ given in Equation (9.1) so that

$$\widetilde{\mathcal{E}}_i = \{\exists j : i \in \widetilde{\mathcal{S}}_j, \tau_I(i) \neq \tau_I(\mathrm{i}_j)\}\,. \tag{9.20}$$

to be the event item i is miscategorized.

Definition 9.6. Let event Err be defined so that on this event, for any item type j,

at most $1/10-$th of the items in $\mathcal{M}_1$ of type j are misclassified

$$\mathsf{Err} = \left\{ \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i] \leq \frac{|\bar{\mathcal{R}}(j)|}{10}, \forall j \in [\mathsf{q}_I] \right\}.$$

Using the total probability lemma, for any fixed value of the $u-$th row of the matrix Ξ (denoted by $\Xi_{u,\cdot}$):

$$\begin{aligned}
\mathbb{P}\Bigg[\sum_{i=1}^{\ell} \mathbb{1}[X_i \in \mathcal{L}_u] < \ell/30 \Bigg| \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3 \Bigg] \\
&\leq \mathbb{P}\Bigg[\sum_{i=1}^{\ell} \mathbb{1}[X_i \in \mathcal{L}_u] < \ell/30 \Bigg| \mathsf{Err}, \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3 \Bigg] \\
&\quad + \mathbb{P}\big[\mathsf{Err}^c \,|\, \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3 \big]. \qquad (9.21)
\end{aligned}$$

Using tower property of expectation on the result of Claim 9.7 provides an upper bound for the second term. Lemma 9.8 provides the upper bound for the first term for all $\Xi_{u,\cdot}$ such that $|\mathcal{L}_u| \geq \mathsf{q}_I/3$ (remember that $\mathcal{L}_u$ is a deterministic function of the $u-$th row of matrix Ξ). Taking the expectation of the the above inequality with respect to $\Xi_{u,\cdot}$ condition on $|\mathcal{L}_u| \geq \mathsf{q}_I/3$ gives Equation (9.18). This and Equation (9.17) give the statement of the claim. □

Claim 9.7. *Let* Err *be the event defined in Definition 9.6. Then,*

$$\mathbb{P}[\mathsf{Err}^c \,|\, \Xi_{u,\cdot}] \leq \frac{20}{N}$$

Proof. Similar to the proof of Claim 9.2, we defined the potential error event for item $i \in \mathcal{M}_1$, $\bar{\mathcal{E}}_{i,U_i}$ to be the event that there exists $j \neq \tau_I(i)$ such that for all the users in U_i, we have $L_{u,i} = \xi_{u,j}$. As explained in the proof of Claim 9.2, for $i \in \widetilde{\mathcal{S}}_j$, we have $\widetilde{\mathcal{E}}_i \subseteq \bar{\mathcal{E}}_{i,U_i}$ for the $U_i = \mathsf{rated}(i) \cap \mathsf{rated}(\mathrm{i}_j)$ at the time the algorithm added i to $\widetilde{\mathcal{S}}_j$. Note that according to Algorithm ItemExplore, $|U_i| \geq \mathsf{r}$.

The remaining of the proof is very similar to the proof of Claim 9.3. A set U_i of r users, chosen independently of feedback, rate item i. Using Lemma A.4, given

$\mathsf{q}_I > 3\mathsf{r}$, there are users of at least $\mathsf{r}/2$ distinct user types in U_i with probability at least $1 - \exp(-\mathsf{r}/2) \geq 1 - \epsilon/\mathsf{q}_I$.

Given this property and conditional on the $u-$th row of matrix Ξ (denoted by $\Xi_{\cdot\setminus\{u,\cdot\}}$), any two item types $j \neq j'$ have jointly independent user preferences by the at least $\mathsf{r}/2 - 1$ users with distinct types in $U_i \setminus \{u\}$. So, Condition on $\Xi_{\cdot\setminus\{u,\cdot\}}$ two items of different types are rated in the same way by $\mathsf{r}/2 - 1$ users with probability at most $2^{-(\mathsf{r}/2-1)}$. The choice of $\mathsf{r} > 2\log(\mathsf{q}_I/\epsilon)$ and $\epsilon = \frac{1}{2\mathsf{q}_I N}$ and union bounding over item types give

$$\mathbb{P}[\widetilde{\mathcal{E}}_i | \bar{\mathcal{R}}(\cdot), \Xi_{u,\cdot}] \leq 3\epsilon \leq \frac{2}{\mathsf{q}_I N}\,.$$

Using Markov inequality gives

$$\mathbb{P}\Big[\sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i] > \frac{|\bar{\mathcal{R}}(j)|}{10} \,\Big|\, \bar{\mathcal{R}}(\cdot), \Xi_{u,\cdot}\Big] \leq \frac{20}{\mathsf{q}_I N}\,.$$

Union bounding over $j \in [\mathsf{q}_I]$ and tower property of expectation gives the statement of the claim. □

Lemma 9.8. *Let the sequence of random variables $X_1 = \tau_I(i_1), X_2 = \tau_I(i_2), \cdots$ denote the item types chosen to be learned by the algorithm.*

Conditional on the u-th row of matrix Ξ, given $X_1, \cdots, X_{m-1}$, the probability of choosing a representative item with any specific type $j \notin \{X_1, \cdots, X_{m-1}\}$ is close to uniform. Hence,

$$\mathbb{P}\Big[\sum_{i=1}^{\ell} \mathbb{1}[X_i \in \mathcal{L}_u] < \ell/30 \,\Big|\, \mathsf{Err}, \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3\Big] < \exp(-\ell/5)\,.$$

Proof. We will show that condition on the event Err, variables $\bar{\mathcal{R}}(\cdot)$, $\Xi_{u,\cdot}$ such that $|\mathcal{L}_u| \geq \mathsf{q}_I/4$ and conditional on $X_1, \cdots, X_{m-1}$, the type of $m-$th representative item, X_m, is almost uniform over all the item types not learned yet, *i.e.,* $[\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}$. Next, we will show that if the choice of learned items is almost uniform, then the statement of lemma holds.

A new representative item is chosen uniformly at random among the remaining items in $\mathcal{M}_1$. Conditional on the u-th row of matrix Ξ, given $X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots$ and $\bar{\mathcal{R}}(X_{m-1})$, we are going to find upper and lower bounds for probability of choosing the next representative item with a specific type $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}$:

$$\mathbb{P}[X_m = j \big| X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \mathsf{Err}, \Xi_{u,\cdot}] \tag{9.22}$$

Lower bound for (9.22)**.** For $m \leq \ell$, let t be the time the m-th item type is chosen randomly by the algorithm. For $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}$ the probability of choosing representative so that $X_m = j$ is equal to the proportion of items of type j in the remaining items in $\mathcal{M}_1$ at time $t-1$. Note that some items of type j might have been miscategorized by time $t-1$. So the total number of items of type j in the remaining items in $\mathcal{M}_1$ might be smaller than $\bar{\mathcal{R}}(j)$ (defined in Equation (9.19)).

The number of items of type j in $\mathcal{M}_1$ by time $t-1$ is at least $|\bar{\mathcal{R}}(j)| - \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i]$. The number of items removed from $\mathcal{M}_1$ by the time t when m-th representative is chosen is at least $\sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|$. The number of items removed might be larger as there could be some items with types not in $\{X_1, \cdots, X_m\}$ that were removed from $\mathcal{M}_1$ due to miscategorization. Hence, the total number of remaining items in $\mathcal{M}_1$ by time t is at most $M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|$. Let $\widetilde{\mathcal{E}}$ be the sequence of error events for all items in $\mathcal{M}_1$ (*i.e.,* $\widetilde{\mathcal{E}} = (\widetilde{\mathcal{E}}_i)_{i \in \mathcal{M}_1}$). Let $\bar{\mathcal{R}}(\cdot)$ be the set of items of all types in $\mathcal{M}_1$. Thus, for any $j \in [\mathsf{q}_I] \setminus \{X_1, X_2, \cdots, X_m\}$:

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \widetilde{\mathcal{E}}, \Xi_{u,\cdot}] \geq \frac{|\bar{\mathcal{R}}(j)| - \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i]}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|}$$

Using the tower property of expectation, for $j \notin \{X_1, \cdots, X_m\}$, we have

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \Xi_{u,\cdot}, \mathsf{Err}] \geq \frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|} \frac{9}{10} \,. \tag{9.23}$$

Upper bound for (9.22)**.** Remember that the probability of choosing $X_m = j$ for some $j \notin \{X_1, \cdots, X_{m-1}\}$ is proportional to the number of items with type j in the remaining items in $\mathcal{M}_1$. The minimum number of remaining items in $\mathcal{M}_1$ is the

number of items in types $[\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}$ minus the number of possible mistakes which removed some items from them.

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \widetilde{\mathcal{E}}, \Xi_{u,\cdot}] \leq \frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)| - \sum_{i \in \mathcal{M}_1 \setminus \cup_{l=1}^{m-1} \bar{\mathcal{R}}(X_l)} \mathbb{1}[\mathcal{E}_i]}$$

Using the tower property of expectation and definition of Err,

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \Xi_{u,\cdot}, \mathsf{Err}] \leq \frac{10}{9} \frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|} \tag{9.24}$$

Taking the expectation of $|\bar{\mathcal{R}}(j)|$ conditional on $X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{u,\cdot}$ for any $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_m\}$ is invariant to choice of j. So, there is a constant C

$$C = \mathbb{E}\left[\frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|} \middle| X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{u,\cdot} \right],$$

independent of j for $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_m\}$. Using tower property of expectation on Equation (9.23) and (9.24) along with the definition of C above gives

$$\frac{9}{10} C < \mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{u,\cdot}, \mathsf{Err}] < \frac{10}{9} C\,.$$

Since there are $\mathsf{q}_I - (m-1)$ types j such that $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}$, taking the summation over j's, the right inequality of the above display gives $C \geq \frac{9}{10} \frac{1}{\mathsf{q}_I - (m-1)}$. The left inequality of the above display gives for all $j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, , X_m\}$,

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{u,\cdot}, \mathsf{Err}] > \frac{2}{5} \frac{1}{\mathsf{q}_I - m + 1}\,.$$

Using tower property of expectation, conditional on $\Xi_{u,\cdot}$ such that $|\mathcal{L}_u| \geq \mathsf{q}_I/3$, for $m \leq \ell \leq \mathsf{q}_I$ and $s < \min\{m, \ell/30\}$ we have

$$\mathbb{P}\left[X_m \in \mathcal{L}_u \middle| X_1, \cdots, X_{m-1}, \sum_{m'=1}^{m-1} \mathbb{1}[X_{m'} \in \mathcal{L}_u] < s, \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3, \mathsf{Err} \right]$$
$$\geq \frac{2}{5} \frac{\frac{\mathsf{q}_I}{3} - s}{\mathsf{q}_I - m + 1} \geq \frac{1}{9}\,.$$

So, the random variable $\sum_{m'=1}^{\ell} \mathbb{1}[X_{m'} \in \mathcal{L}_u]$ conditional on $\Xi_{u,\cdot}, |\mathcal{L}_u| \geq \mathsf{q}_I/3, \mathsf{Err}$, stochastically dominates a Binomial random variable with mean $\ell/9$. Hence,

$$\mathbb{P}\left[\sum_{m'=1}^{\ell} \mathbb{1}[X_{m'} \in \mathcal{L}_u] < \frac{\ell}{30} \,\middle|\, \Xi_{u,\cdot}, |\mathcal{L}_u| \geq \frac{\mathsf{q}_I}{3}, \mathsf{Err}\right] \leq \exp(-\ell/5)\,. \qquad \square$$

9.2 Item structure only: lower bound

9.2.1 Proof strategy

Recommending item $a_{u,t}$ to user u at time t is considered *a good recommendation* when the probability of $L_{u,a_{u,t}} = +1$ given feedback from history is large (close to 1). Alternatively, the recommendations in which the probability of $L_{u,a_{u,t}} = +1$ is close to (or smaller than) 1/2 are considered *bad recommendations* (or uncertain recommendations).

The lower bound for regret of any online algorithm based on item structure only model is based on two main observations about good recommendations:

- *An accurate estimation of item type of the recommended item is necessary to make a good recommendation:*

  Similar to the lower bound of user structure only model, we define the concept of $(\mathsf{r}, \epsilon)-$ row regularity as in Definition 9.11. With proper choice of r and ϵ, the preference matrix is row regular with high probability. This implies the following observation: for any subset of r users, the item types are behaving almost uniformly and independently at random (within a factor of ϵ deviation). For a given item i, let the algorithm know the preference of any subset of $\mathsf{r} - 1$ users for i. The preference of any other user is almost independent and uniformly distributed on $\{-1, +1\}$.

  It is shown that it takes no less than $\log \mathsf{q}_I$ recommendation to learn the item type of a given item. This matches the information theoretic bound in source coding problem in which a random variable (type of an item) with equiprobable distribution on q_I values requires at least $\log \mathsf{q}_I$ binary integer valued code.

- *Recommending an item from an item type to a specific user for the first time is an uncertain recommendation*:

  In contrast, even if the algorithm is given the item types of each item, still the preference of a user u for a given item type j should be learned to make accurate prediction about the recommendation of any item from the item type j to u. This can only be done by recommending one item with type j (representative item) to u.

To provide a lower bound for regret defined in Equation (7.1), first we rigorize the above high level ideas to characterize the bad recommendations made by the algorithm. Next, we determine an upper bound for the number of good recommendations in terms of some parameters chosen by the algorithm. We use some counting techniques to show tradeoff in the choice of these parameter and prove that for any value of the parameters chosen by the algorithm, there is an upper bound for the number of good recommendation (translating into a lower bound for regret) in terms of parameters of the model (N and $\mathtt{q}_I$) and time horizon T.

9.2.2 Lower bound theorem statement

Theorem 9.9. *Let* $\mathtt{r} = \log \mathtt{q}_I - 6 \log\log N$ *and* $\epsilon = 1/\log N$. *In the item-structure model with* $N > 20$ *and* $\mathtt{q}_I < N$, *any recommendation algorithm must incur regret*

$$\mathtt{regret}(T) \geq \frac{1-4\epsilon}{2} \max\left\{1\,,\, \sqrt{\frac{\mathtt{q}_I}{6N}T}\,,\, \frac{\mathtt{r}}{N}T\right\} - \frac{1}{N}T\,.$$

The statement of the theorem follows immediately from the Lemmas 9.15 and 9.17. The approach in the proof is twofold: i) to show that for a preference matrix satisfying row regularity, items that are rated less than $\mathtt{r}$ times have probability roughly half of being liked by any user given the feedback obtained thus far, even if the type matrix is known; ii) if a user u has not rated any item with the same type as item i before, the probability that user u likes item i conditional on history is $1/2$. Below, we start by getting into more details of the first approach:

Definition 9.10. We define c_i^t to be the number of times item i has been rated by any user up until time $t-1$, $c_i^t := \sum_{u=1}^{N} \sum_{s=1}^{t-1} \mathbb{1}[a_{u,s} = i]$.

Definition 9.11. The matrix $A \in \{-1, +1\}^{n \times m}$ is said to be $(t, \epsilon)-$*row regular* if its transpose is $(t, \epsilon)-$ column regular.

Let

$$\mathsf{r} = \log \mathsf{q}_I - 6 \log \log N \quad \text{and} \quad \epsilon = \frac{1}{\log N} \,. \tag{9.25}$$

As an immediate application of Lemma 8.7 in Section 8.2, we know that the matrix Ξ is $(\mathsf{r}, \epsilon)-$row regular (denoted by $\Xi^T \in \Omega_{\mathsf{r},\epsilon}$) with high probability. (We will show this in more detail in Lemma 9.16.)

Next, in the following lemma, we show that if the number of times an item is recommended (c_i^t) is small and the preference matrix is row-regular, then the outcome of recommending that item to any user is uncertain.

Lemma 9.12. *Let* r *and* ϵ *be as in Equation* (9.25)*. For any user* $u \in [N]$ *and any item* $i \in \mathbb{N}$,

$$\mathbb{P}[L_{u,i} = +1 \,|\, a_{u,t} = i, c_i^t < \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \leq (1 + 4\epsilon)/2 \,.$$

Proof. We show that if an item has been rated by less than r users, its type is uncertain (many item types are consistent with the history). Given any row regular preference matrix, the uncertainty of the item type makes the accurate prediction of preference of any user for the item impossible.

We focus on a particular item i at time t. Let $w = \{u \in [N] : a_{u,s} = i, s < t\}$ be the set of users that were recommended item i up to time $t-1$, and let $b = \{L_{u,i}\}_{u \in w}$ be the vector of feedback from users in w about item i. Note that $c_i^t < \mathsf{r}$ implies $|w| < \mathsf{r}$. Using the notation in Definition 8.5, if M is the matrix formed by concatenating the rows of Ξ indexed by w, then $K_{b,w}(\Xi^T)$ would be the set of columns (corresponding to the item types) which are identical. We claim that conditional on the matrix Ξ, vector b and w, the type $\tau_I(i)$ of item i at the end of time instant $t-1$ is uniformly distributed over the set of item types $K_{b,w}(\Xi^T)$ consistent with this data. This can be derived by applying Bayes rule considering that the prior distribution of type of each item is uniform over $[\mathsf{q}_I]$.

Let $b^+ = [b \quad 1] \in \{-1, +1\}^{|w|+1}$ be obtained from b by appending $+1$. For a given user $u \notin w$, we have $L_{u,i} = +1$ precisely when $\tau_I(i) \in K_{b^+,\{w,u\}}(\Xi^T)$, which in words reads "item i is among those types that are consistent with the ratings of i up to time $t-1$ and have preference vector with '+1' for user u". It follows that for any preference matrix Ξ and any user u which has not rated i up to time $t-1$,

$$\begin{aligned}\mathbb{P}[L_{u,i} = +1 | H_{t-1}, \Xi] &= \mathbb{P}[\tau_I(i) \in K_{b^+,\{w,u\}}(\Xi^T) | H_{t-1}, \Xi] \\ &= \frac{k_{b^+,\{w,u\}}(\Xi^T)}{k_{b,w}(\Xi^T)} \end{aligned} \tag{9.26}$$

The second inequality is due to: i) w and b are functions of H_{t-1} and $u \notin w$ according to H_{t-1}; ii) for fixed w and b, the set $K_{b,w}(\Xi^T)$ is determined by Ξ^T; iii) $\tau_I(i)$ is uniformly distributed on $K_{b,w}(\Xi^T)$ right before making the recommendations at time t.

Recall the event that the preference matrix Ξ is $(\mathsf{r}, \epsilon)-$ row regular is denoted by $\Xi^T \in \Omega_{\mathsf{r},\epsilon}$. It now follows by tower property of conditional expectation

$$\begin{aligned}&\mathbb{P}[L_{u,i} = +1, a_{u,t} = i \,|\, c_i^t \leq \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \\ &\qquad \overset{\text{(a)}}{=} \mathbb{E}\left[\mathbb{P}[L_{u,i} = +1 \,|\, H_{t-1}, \Xi]\, \mathbb{P}[a_{u,t} = i | H_{t-1}] \,\middle|\, c_i^t \leq \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}\right] \\ &\qquad \overset{\text{(b)}}{=} \mathbb{E}\left[\frac{k_{b^+,\{w,u\}}(\Xi^T)}{k_{b,w}(\Xi^T)} \mathbb{1}[u \notin w]\, \mathbb{P}[a_{u,t} = i | H_{t-1}] \,\middle|\, c_i^t \leq \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}\right] \\ &\qquad \overset{\text{(c)}}{\leq} \frac{(1+\epsilon)\frac{\mathsf{q}_I}{2^{|w|+1}}}{(1-\epsilon)\frac{\mathsf{q}_I}{2^{|w|}}} \mathbb{P}[a_{u,t} = i | c_i^t \leq \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \\ &\qquad \overset{\text{(d)}}{\leq} \frac{1}{2}(1+4\epsilon)\mathbb{P}[a_{u,t} = i | c_i^t \leq \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}]\,.\end{aligned}$$

This immediately gives the statement of Lemma. Equality (a) is a result of i) tower property of conditional expectation; ii) conditional on H_{t-1}, the random variable $a_{u,t}$ is independent of all other random variables. Note that $\mathbb{P}[a_{u,t} = i | H_{t-1}]$ is nonzero only if u has not rated item i by time $t-1$ according to H_{t-1}. Hence, we can use Equation (9.26) to get Equality (b) knowing that this is nonzero only for the history H_{t-1} in which $u \notin w$ for item i. Inequality (c) is justified as follows: if $\Xi^T \in \Omega_{\mathsf{r},\epsilon}$ then by Claim 8.6, $\Xi^T \in \Omega_{\mathsf{r}-1,\epsilon}$. By Definition 8.5, this means that $k_{b,w}(\Xi^T) \geq (1-\epsilon)\mathsf{q}_I/2^{|w|}$

and $k_{b^+,\{w,u\}}(\Xi^T) \leq (1+\epsilon)\mathsf{q}_I/2^{|w|+1}$. Using the choice of ϵ given in Equation (9.25) with $N > 8$ gives $\epsilon < 1/2$ which guarantees inequality (d). $\square$

Next, we will observe that if a user u has not rated any item with the same type as item i before, the probability that u likes item i is $1/2$.

Definition 9.13. Let $\mathcal{B}^t_{u,\tau_I(i)}$ be the event that user u has rated an item of the same item type as item i by time $t-1$, *i.e.*,

$$\mathcal{B}^t_{u,\tau_I(i)} = \{\exists i' \in \mathbb{N} : a_{u,s} = i' \text{ for some } s < t \text{ with } \tau_I(i) = \tau_I(i')\}\,.$$

Lemma 9.14. *If user u has not rated any item with the same type as i by time $t-1$, the probability that user u likes item i conditional on history is $1/2$:*

$$\mathbb{P}[L_{u,i} = -1 | a_{u,t} = i, (\mathcal{B}^t_{u,\tau_I(i)})^c, c_i^t \geq \mathsf{r}] = \frac{1}{2}\,. \tag{9.27}$$

Proof. We make two observations: i) if user u has not rated any item with item type $\tau_I(i)$ before, the feedback in history H_{t-1} is statistically independent of the value of $\xi_{u,\tau_I(i)}$ given all other elements of matrix Ξ and the item types; ii) According to the model, the types of items (function $\tau_I(\cdot)$) are independent of the matrix Ξ. Also, the element of matrix Ξ are statistically independent of each other. Hence, conditional on $(\mathcal{B}^t_{u,\tau_I(i)})^c$, the posterior distribution of $L_{u,i}$ is uniform on $\{-1,+1\}$ at time t.

First, define $\Xi_{.\setminus\{u,j\}}$ to be the set of elements in Ξ except $\xi_{u,j}$, *i.e.*,

$$\Xi_{.\setminus\{u,j\}} = \{\xi_{u',j'},\ \text{for all } u' \in [N] \text{ and } j' \in [\mathsf{q}_I]\} \setminus \{\xi_{u,j}\}\,.$$

Note that $\Xi_{.\setminus\{u,j\}}$ is statistically independent of $\xi_{u,j}$. Also define $\tau_I(\cdot) = \{\tau_I(i) : i \in \mathbb{N}\}$ to be the item types of all items. Note that this set of random variables is also statistically independent of elements of matrix Ξ and specifically $\xi_{u,j}$. Also, recall that conditional on H_{t-1}, $a_{u,t}$ is statistically independent of the preference matrix Ξ and item types $\tau_I(\cdot)$.

Next, we show that conditional on $\Xi_{.\setminus\{u,j\}}$ and $\tau_I(\cdot)$, the event $\{a_{u,t} = i, \tau_I(i) =$

$j, (\mathcal{B}^t_{u,\tau_I(i)})^c, c^t_i \geq \mathsf{r}, H_{t-1} = H\}$ is independent of $\xi_{u,j}$:

$$\xi_{u,j} \perp\!\!\!\perp \{a_{u,t} = i, \tau_I(i) = j, (\mathcal{B}^t_{u,\tau_I(i)})^c, c^t_i \geq \mathsf{r}, H_{t-1} = H\} \Big| \{\Xi_{.\setminus\{u,j\}}, \tau_I(\cdot)\}. \tag{9.28}$$

The above follows from:

- $a_{u,t}$ is conditionally independent of Ξ and $\tau_I(\cdot)$ given H_{t-1}.
- The event $\{\tau_I(i) = j\}$ is contained in the specification of $\tau_I(.)$.
- $\mathcal{B}^t_{u,\tau_I(i)}$ as in Definition 9.13 is a deterministic function of H_{t-1} and $\tau_I(i)$.
- c^t_i as in Definition 9.10 is a deterministic function of H_{t-1}.
- The history H_{t-1} is defined to be the sequence of random variables

  $$\{(a_{u',s}, L_{u',a_{u',s}}), \text{ for all } u' \in [N] \text{ and } s < t\}.$$

  We show inductively that if $(\mathcal{B}^t_{u,\tau_I(i)})^c$ and $\tau_I(i) = j$ holds, conditional on $\{\Xi_{.\setminus\{u,j\}}, \tau_I(.)\}$, for $s < t$, H_s is independent of $\xi_{u,j}$. Let this be true for $s-1$. Recall that given H_{s-1}, for all $u' \in [N]$, the recommended items $a_{u',s}$ are independent of Ξ and $\tau_I(\cdot)$ and in particular independent of $\xi_{u,j}$. We also know that $\tau_I(a_{u,s}) \neq j$ as a result of $(\mathcal{B}^t_{u,\tau_I(i)})^c$ and $\tau_I(i) = j$. Thus, $L_{u',a_{u',s}}$ for all $u' \in [N]$ is a function of $\Xi_{.\setminus\{u,j\}}$ and $\tau_I(.)$ and is independent of $\xi_{u,j}$.

Since i) $\Xi_{.\setminus\{\xi_{u,j}\}}$ and $\tau_I(.)$, are independent of $\xi_{u,j}$ and ii) Conditional independence property in Equation (9.28) holds for every H and $j \in [\mathsf{q}_I]$:

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 \,|\, a_{u,t} = i, \tau_I(i) = j, (\mathcal{B}^t_{u,\tau_I(i)})^c, c^t_i \geq \mathsf{r}, H_{t-1} = H,\, \Xi_{.\setminus\{u,j\}}, \tau_I(.)] \\
&\quad = \mathbb{P}[\xi_{u,j} = +1 \,|\, a_{u,t} = i, \tau_I(i) = j, (\mathcal{B}^t_{u,\tau_I(i)})^c, c^t_i \geq \mathsf{r}, H_{t-1} = H,\, \Xi_{.\setminus\{u,j\}}, \tau_I(.)] \\
&\quad = \mathbb{P}[\xi_{u,j} = +1 \,|\, \Xi_{.\setminus\{u,j\}}, \tau_I(.)] = \frac{1}{2}
\end{aligned}$$

Using the tower property of expectation on the first term of above display gives the statement of the Lemma. □

In Lemma 9.12, we showed that the outcome of recommending items that have not been rated more than r are uncertain. In Lemma 9.14, we showed that the outcome of the first time a user is recommended an item from a specific type is also uncertain. In the following lemma, we make the connection to provide a lower bound for regret in terms of recommendations which are not proved to be uncertain (not belonging to either of above categories of recommendations).

Lemma 9.15. *For ϵ and r defined in Equation* (9.25)*,*

$$N\mathtt{regret}(T) \geq \frac{1-4\epsilon}{2}(TN - \sum_{u\in[N],t\in[T]} \mathbb{P}[c^t_{a_{u,t}} > \mathsf{r}, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c]) - T\,.$$

Proof. To get a lower bound for regret, we provide an upper bound for the number of good recommendations. We partition the recommendations based on the properties of interest. For $\mathsf{r} = \log \mathsf{q}_I - 6\log\log N$ and $\epsilon = \frac{1}{\log N}$,

$$\begin{aligned}
N(T - \mathtt{regret}(T)) &= \mathbb{E}[\sum_{t=1}^{T}\sum_{u=1}^{N} \mathbb{1}[L_{u,a_{u,t}} = +1]] \\
&= \mathbb{E}\sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{1}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} < \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \\
&\quad + \mathbb{E}\sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{1}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} < \mathsf{r}, \Xi^T \notin \Omega_{\mathsf{r},\epsilon}] \\
&\quad + \mathbb{E}\sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{1}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} \geq \mathsf{r}, \mathcal{B}^t_{u,\tau_I(a_{u,t})}] \\
&\quad + \mathbb{E}\sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{1}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} \geq \mathsf{r}, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] \\
&=: \mathtt{A1} + \mathtt{A2} + \mathtt{A3} + \mathtt{A4}\,,
\end{aligned} \tag{9.29}$$

here the term $\mathtt{A1}$ corresponds to the number of successful recommendations of items that have not been recommended to many users before (*i.e.,* such that $c^t_i < \mathsf{r}$) on the event that the matrix Ξ is $(\mathsf{r},\epsilon)-$row regular. The term $\mathtt{A2}$ corresponds to the same recommendation on the event that the matrix Ξ is not $(\mathsf{r},\epsilon)-$row regular. $\mathtt{A2}$

is upper bounded since this event happens with very small probability. The term $\mathtt{A3}$ corresponds to the recommendations of items with $c_i^t \geq \mathsf{r}$ to users u such that the user u has rated another item with the same item type by time $t-1$. The term $\mathtt{A4}$ is the number of correct recommendations of items with $c_i^t \geq \mathsf{r}$ to users u such that the user u has not rated another item with the same item type by time $t-1$. These recommendations are liked with probability $1/2$.

First, we look at the recommendation of items that have been recommended to less than r users before. For row-regular preference matrices, Lemma 9.12 shows that the outcome of these recommendations are uncertain. Note that the preference matrix is row regular with high probability. Recommending users an item from a specific item type for the first time gives uncertain outcome as well, proved in Lemma 9.14. Hence, the only possibly good recommendations are the ones in which an item that has been recommended more than r times and it is not the first time a user is rating an item from that item type.

Bounding $\mathtt{A1}$. For ϵ defined in Equation (9.25), multiplying the statement of Lemma 9.12 by $\mathbb{P}[a_{u,t} = i, c_i^t < \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}]$ and summing over i gives

$$\begin{aligned}\mathtt{A1} &= \sum_{\substack{t\in[T]\\u\in[N]}} \sum_{i\geq 1} \mathbb{P}[L_{u,i} = +, a_{u,t} = i, c_i^t < \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \\ &\leq \frac{1+4\epsilon}{2} \sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{P}[c_{a_{u,t}}^t < \mathsf{r}, \Xi^T \in \Omega_{\mathsf{r},\epsilon}] \leq \frac{1+4\epsilon}{2} \sum_{\substack{t\in[T]\\u\in[N]}} \mathbb{P}[c_{a_{u,t}}^t < \mathsf{r}]\,. \end{aligned} \tag{9.30}$$

Bounding $\mathtt{A2}$. The following lemma is an immediate corollary of Lemma 8.7 in Section 8.2.

Lemma 9.16. *Let matrix $A = [A_{i,j}] \in \{-1,+1\}^{n\times m}$ have i.i.d.* Bern(1/2) *entries. Then the matrix A is $(t,\epsilon)-$row regular with probability at least*

$$1 - 2(2n)^t \max\left\{ \exp\left(-\frac{\epsilon^2}{3}\frac{m}{2^t}\right), \exp\left(-\frac{\epsilon}{2}\frac{m}{2^t}\right) \right\}.$$

Using Lemma 9.16, for $\mathsf{r} = \log \mathsf{q}_I - 6\log\log N$ and $\epsilon = \frac{1}{\log N}$, consistent with

the definition given in Equation (9.25), Ξ is (r, ϵ) row-regular (*i.e.*, $\Xi^T \in \Omega_{\mathsf{r},\epsilon}$) with probability at least $1 - \frac{1}{N}$ given $N > 20$.

$$\begin{aligned}\texttt{A2} &= \sum_{\substack{t\in[T]\\ u\in[N]}} \mathbb{P}[L_{u,a_{u,t}} = +1, c_{a_{u,t},t} < \mathsf{r}, \Xi \notin \Omega_{\mathsf{r},\epsilon}] \\ &\leq \mathbb{P}[\Xi \notin \Omega_{\mathsf{r},\epsilon}] \sum_{\substack{t\in[T]\\ u\in[N]}} \mathbb{P}[L_{u,a_{u,t}} = +1, c_{a_{u,t},t} < \mathsf{r} | \Xi^T \notin \Omega_{\mathsf{r},\epsilon}] \leq T\,. \end{aligned} \tag{9.31}$$

Bounding `A3`**.** We use the following:

$$\texttt{A3} = \sum_{u\in[N],t\in[T]} \mathbb{P}[L_{u,a_{u,t}} = +1, c_{a_{u,t},t} \geq \mathsf{r}, \mathcal{B}^t_{u,\tau_I(i)}] \leq \sum_{u\in[N],t\in[T]} \mathbb{P}[c_{a_{u,t},t} \geq \mathsf{r}, \mathcal{B}^t_{u,\tau_I(i)}]\,. \tag{9.32}$$

Bounding `A4`**.** Lemma 9.14 gives

$$\frac{\mathbb{P}[L_{u,i} = +1, a_{u,t} = i, c^t_i \geq \mathsf{r}, (\mathcal{B}^t_{u,\tau_I(i)})^c]}{\mathbb{P}[a_{u,t} = i, c^t_i \geq \mathsf{r}, (\mathcal{B}^t_{u,\tau_I(i)})^c]} = \frac{1}{2}\,.$$

Multiplying this by $\mathbb{P}[a_{u,t} = i, (\mathcal{B}^t_{u,\tau_I(i)})^c, c^t_i \geq \mathsf{r}]$ and taking the summation over i gives

$$\texttt{A4} = \frac{1}{2} \sum_{u\in[N],t\in[T]} \mathbb{P}[c_{a_{u,t},t} > \mathsf{r}, (\mathcal{B}^t_{u,\tau_I(i)})^c]\,. \tag{9.33}$$

Plugging Equations (9.30), (9.31), (9.32) and (9.33) into Equation (9.29) gives the statement of lemma. □

Next, we find an upper bound on the expected number of "good" recommendations made by the algorithm in terms of parameters of the model.

Lemma 9.17. *The expected number of good recommendations made by any algorithm is upper bounded as:*

$$\sum_{t\in[T],u\in[N]} \mathbb{P}[c^t_{a_{u,t}} > \mathsf{r}, \mathcal{B}^t_{u,\tau_I(a_{u,t})}] \leq TN - \max\{N\,,\ \sqrt{T\,\mathsf{q}_I N/6}\,,\ T\mathsf{r}\}\,.$$

Proof. We will find upper bound for the expected number of good recommendations in terms of several parameters chosen by the algorithm (defined below). Then, we show that the expected number of good recommendations is upper bounded for any value of these parameters chosen by the algorithm. Let $\mathsf{good}(T)$ be the total number of good recommendations made by the algorithm by time T:

$$\mathsf{good}(T) = \sum_{\substack{t\in[T]\\ u\in[N]}} \sum_{i\geq 1} \mathbb{1}[a_{u,t} = i, c_i^t > \mathsf{r}, \mathcal{B}^t_{u,\tau_I(i)}]\,. \tag{9.34}$$

We introduce the following parameters: 1) the total number of items rated by more than r users by time T; and 2) the total number of items rated by less than r users by time T

Definition 9.18. Let $\mathcal{F}_T$ be the items that have been rated by more than r users by time T and $f_T := |\mathcal{F}_T|$;

$$f_T = \sum_{i\in\mathbb{N}} \mathbb{1}[c_i^T \geq \mathsf{r}].$$

Definition 9.19. Let $\mathcal{G}_T$ to be the items that have been rated by less than r users by time T and $g_T := |\mathcal{G}_T|$;

$$g_T = \sum_{i\in\mathbb{N}} \mathbb{1}[0 < c_i^T < \mathsf{r}].$$

In the following claim, we bound the number of good recommendations in terms of f_T and g_T.

Claim 9.20. *Let* $\mathsf{good}(T)$ *be defined as Equation* (9.34)*,*

$$\mathsf{good}(T) \leq TN - g_T - f_T\mathsf{r}\,.$$

Proof. Using Definition 9.18, any $i \in \mathcal{F}_T$ is recommended to more than r users by T. So, for any $i \in \mathcal{F}_T$, there are at least r recommendations in which i has been rated less than r times before that:

$$\sum_{t\in[T],u\in[N]} \mathbb{1}[c_i^t \leq \mathsf{r}, a_{u,t} = i] \geq \mathsf{r}\,.$$

Using the Definition 9.19, any $i \in \mathcal{G}_T$ has been recommended at least once by the algorithm:

$$\sum_{t\in[T],u\in[N]} \mathbb{1}[c_i^t \leq \mathsf{r}, a_{u,t} = i] \geq 1\,.$$

The items recommended by the algorithm by time T are either in $\mathcal{F}_T$ or $\mathcal{G}_T$. Hence, the total number of recommendations satisfy

$$\begin{aligned}
NT &= \sum_{t=1}^{T}\sum_{u=1}^{N}\sum_{i\geq 1} \mathbb{1}[a_{u,t} = i] \\
&= \sum_{t=1}^{T}\sum_{u=1}^{N}\sum_{i\geq 1} \mathbb{1}[c_i^t > \mathsf{r}, a_{u,t} = i] + \sum_{t=1}^{T}\sum_{u=1}^{N}\sum_{i\geq 1} \mathbb{1}[c_i^t \leq \mathsf{r}, a_{u,t} = i] \\
&\geq \mathsf{good}(T) + \sum_{\substack{t\in[T]\\u\in[N]}}\sum_{i\in\mathcal{F}_T} \mathbb{1}[c_i^t \leq \mathsf{r}, a_{u,t} = i] + \sum_{\substack{t\in[T]\\u\in[N]}}\sum_{i\in\mathcal{G}_T} \mathbb{1}[c_i^t \leq \mathsf{r}, a_{u,t} = i] \\
&\geq \mathsf{good}(T) + f_T\mathsf{r} + g_T\,.
\end{aligned}$$

where we used $\mathsf{good}(T) \leq \sum_{t=1}^{T}\sum_{u=1}^{N}\sum_{i\geq 1} \mathbb{1}[c_i^t > \mathsf{r}, a_{u,t} = i]$. □

Each item is recommended at most once to each user. Hence, each item (and specifically items in $\mathcal{F}_T$) are recommended at most N times by time T. According to Definition 9.19, items in $\mathcal{G}_T$ are recommended at most r times by time T. The total number of recommendations made by the algorithm by time T is upper bounded as

$$TN \leq f_T N + g_T(\mathsf{r} - 1)\,.$$

Hence,

$$g_T \geq (T - f_T)\frac{N}{\mathsf{r}}\,. \tag{9.35}$$

Plugging this into the statement of Claim 9.20, one would get

$$\mathsf{good}(T) \leq TN - g_T(1 - \frac{\mathsf{r}^2}{N}) - T\mathsf{r} \leq TN - T\mathsf{r} \tag{9.36}$$

where the second inequality is due to the assumptions $\mathsf{q}_I < N$ and $N > 20$.

Next, we will provide a different upper bound for the value of $\mathsf{good}(T)$ in terms of number of item types each user has rated by time T.

Definition 9.21. Let Γ_u^T be the set of item types that are rated by user u by time T.

$$\Gamma_u^T = \left\{ j \in [\mathsf{q}_I] : \sum_{t \in [T], i : \tau_I(i) = j} \mathbb{1}[a_{u,t} = i] \neq 0 \right\}$$

Let $\gamma_u^T = |\Gamma_u^T|$.

Claim 9.22. *Using Definition 9.21, for E defined in Equation* (9.34)*,*

$$\mathsf{good}(T) \leq N(T - \min_u \gamma_u^T)\,.$$

Proof. For user u, number of times an item type is rated for the first time by time T is equal to the number of item types rated by user u by time T. Using Definition 9.13, this is precisely the number of recommendations made to u in which $(\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c$ holds. Hence,

$$\sum_{t=1}^{T} \mathbb{1}[(\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] = \gamma_u^T\,.$$

The total number of recommendations made to user u by time T is equal to T. Hence,

$$\sum_{t=1}^{T} \mathbb{1}[\mathcal{B}^t_{u,\tau_I(a_{u,t})}] = T - \gamma_u^T\,.$$

Summing over u and using $\mathsf{good}(T) \leq \sum_{t=1}^{T} \sum_{u=1}^{N} \mathbb{1}[\mathcal{B}^t_{u,\tau_I(a_{u,t})}]$ gives the statement of the claim. $\square$

Definition 9.23. Let r_j^T be the number of items with item type j that have been rated by time T:

$$r_j^T = |\{i : \tau_I(i) = j, a_{u,t} = i \text{ for some } u \in [N] \text{ and } t \leq T\}|.$$

T items from $\mathcal{F}_T \cup \mathcal{G}_T$ with types in Γ_u^T (defined in Definition 9.21) are recom-

mended to user u by time T. Hence, $T \leq \sum_{j \in \Gamma_u^T} r_j^T \leq \gamma_u^T \max_j r_j^T$. This implies

$$\gamma_*^T := \min_u \gamma_u^T \geq \frac{T}{\max_j r_{j,T}} . \tag{9.37}$$

The total number of items recommended (ever seen by the algorithm) by time T is $f_T + g_T$. Note that the prior distribution of type of each item is uniform on $[\mathsf{q}_I]$ independently of other items. So, r_j^T is $\mathsf{Bin}(f_T + g_T, \frac{1}{\mathsf{q}_I})$. We will provide two different deviation bounds for the random variable $\max_j r_j^T$ as a function of $f_T + g_T$:

- When $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$, Chernoff bound in Lemma A.3 with $\epsilon = 1$ gives

  $$\mathbb{P}\big[\max_{j \in \mathsf{q}_I} r_j^T \geq 2\frac{f_T + g_T}{\mathsf{q}_I}\big] \leq \mathsf{q}_I \exp\left(-\frac{f_T + g_T}{3\mathsf{q}_I}\right) \leq \mathsf{q}_I \exp\big(-2\log \mathsf{q}_I\big) < \frac{1}{\mathsf{q}_I},$$

  where we used $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$. Plugging Equation (9.37) into this shows that when $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$, with probability at least 1/2, $\gamma_*^T \geq \frac{T\mathsf{q}_I}{2(f_T+g_T)}$ and otherwise, $\gamma_*^T \geq 1$. Plugging this observation into the statement of Claim 9.22 gives

  $$\mathbb{E}\left[\mathsf{good}(T)\right] \leq NT - N\frac{T\mathsf{q}_I}{4(f_T + g_T)}, \quad \text{if } f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I. \tag{9.38}$$

- When $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$, Chernoff bound in Lemma A.3 with

  $$\epsilon = \frac{18\mathsf{q}_I \log \mathsf{q}_I}{f_T + g_T} - 1 > 2$$

  gives

  $$\begin{aligned}\mathbb{P}\big[\max_{j \in \mathsf{q}_I} r_j^T \geq 18\log \mathsf{q}_I\big] &\leq \mathsf{q}_I \exp\left(-(\frac{18\mathsf{q}_I \log \mathsf{q}_I}{f_T + g_T} - 1)^2 \frac{f_T + g_T}{2\mathsf{q}_I}\right) \\ &\leq \mathsf{q}_I \exp\left(-9\log \mathsf{q}_I\right) < \frac{1}{\mathsf{q}_I},\end{aligned}$$

  where we used the assumption $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$.

  This result and Equation (9.37) shows that if $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$, then with probability at least 1/2, $\gamma_*^T \geq \frac{T}{18\log \mathsf{q}_I}$ and otherwise, $\gamma_*^T \geq 1$.

$$\mathbb{E}\left[\mathsf{good}(T)\right] \leq NT - N\frac{T}{36\log \mathsf{q}_I}\,, \quad \text{if } f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I. \tag{9.39}$$

Equations (9.38) and (9.39) show that

$$\mathbb{E}\left[\mathsf{good}(T)\right] \leq NT - N\min\{\frac{T}{36\log \mathsf{q}_I}, \frac{T}{6(f_T+g_T)}\}\,.$$

The statement of Claim 9.20, give

$$\begin{aligned}
\mathbb{E}\left[\mathsf{good}(T)\right] &\leq \max_{f_T,g_T}\left[NT - \max\left\{f_T\mathsf{r} + g_T, N\min\{\frac{T}{36\log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{4(f_T+g_T)}\}\right\}\right] \\
&\leq \max_{f_T,g_T}\left[NT - \max\left\{f_T + g_T, N\min\{\frac{T}{36\log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{6(f_T+g_T)}\}\right\}\right] \\
&\overset{(a)}{\leq} NT - \max\left\{f_T + g_T, \frac{TN\mathsf{q}_I}{6(f_T+g_T)}\}\right\} \\
&\overset{(b)}{\leq} NT - \sqrt{NT\mathsf{q}_I/6}\,,
\end{aligned}$$

Inequality (a) holds for N large enough so that $NT \geq 216\mathsf{q}_I(\log \mathsf{q}_I)^2$. For such N, in the regime in which $N\min\{\frac{T}{36\log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{6(f_T+g_T)}\} \leq f_T + g_T$, we have $\frac{T}{36\log \mathsf{q}_I} \geq \frac{T\mathsf{q}_I}{4(f_T+g_T)}$. And (b) holds for any choice of parameters f_T and $g_T > 0$.

Also, since $\gamma_*^T \geq 1$, we have $\mathbb{E}\left[\mathsf{good}(T)\right] \leq NT - N$.

The above two statements and Equation (9.36) give the result in the Lemma. □

Chapter 10

Hybrid algorithm

In this chapter, a hybrid algorithm is proposed which utilizes the structure in both user and item space. Then, information theoretic lower bound shows that the proposed algorithm achieves almost optimal performance in virtually all regimes of parameters of the model.

Utilizing both structure in user space and item space provides the benefits of both user-user and item-item algorithm. Similar to item-item algorithm, the hybrid algorithm can start making meaningful recommendation very early on (small cold-start time). The asymptotic slope of regret in hybrid algorithm is also as small as the one in item-item algorithm. Additionally, similar to user-user algorithm, the asymptotic slope is achieved much faster compared to item-item algorithm. The explanation for this observation is that the by learning the structure in user space, the effort of learning the item space is shared among the users of similar types.

10.1 Hybrid algorithm - upper bound

The algorithm we are proposing based on the joint structure in user space and item space is quite similar to the proposed item-item algorithm. For small T, the algorithm is oblivious to the structure in user space and performing exactly like the item-item algorithm.

By exploring part of the item space, the structure in user space is known to the

algorithm. After that, exploring the rest of item space is easier due to collaboration between users. When the number of learned item types in item-item algorithm (the parameter ℓ in Algorithm 9.1) is greater than $\mathsf{r}_U = \log N \mathsf{q}_U^2$, it is possible to partition users based on their type correctly with high probability. This happens when T is large enough so that $\ell > \mathsf{r}_U$. The algorithm learns the user structure through partitioning the users (similar to user-user model in Algorithm 8.1)

Learning user types reduces the cost of item exploration. Knowing the type of users, to learn an item type, instead of recommending it to all users, one needs to recommend it to one user from each user type. This reduces the cost of exploration by a factor of $\frac{\mathsf{q}_U}{N}$.

After all q_I item types are learned, the joint algorithm works similar to item-item algorithm (Algorithm 9.1) again. The user structure becomes irrelevant again. The intuition is that utilizing structure in user space gives asymptotic regret slope of $\frac{\mathsf{q}_U}{N}$ whereas the asymptotic regret slope using the structure in item space is $\frac{\log N}{N}$ which is much smaller for N being large.

The algorithm uses some notation, which we define next. For an item i and time $t > 0$, $\mathsf{rated}_t(i) = \{u \in [N] : a_{u,s} = i \text{ for some } s < t\}$ is the set of users that have rated the item before time t. We use the notation $\mathsf{rated}(i)$ in the algorithm for $\mathsf{rated}_t(i)$ at the time it is used. If $\ell \geq \mathsf{r}_U$ and a partitioning over users is formed, $\mathsf{ratedtype}_t(i)$ denotes the partitions from which a user has rated item i by time t, *i.e.,* $\mathsf{ratedtype}_t(i) = \{k \in [\mathsf{q}_U] : a_{u,s} = i \text{ for some } s < t, u \in \mathcal{P}_k\}$.

Algorithm 10.1 JOINT$(T, \mathsf{q}_I, \mathsf{q}_U, N)$ (fixed time horizon)

1: ℓ chosen as Equation (10.1)
2: $\mathsf{r}_U = 13 \log N \mathsf{q}_U^2$; $\mathsf{r}_I = 10 \log N \mathsf{q}_I^2$; $M = \lceil T \frac{64 \mathsf{q}_I}{\ell} \rceil$
3: $\mathcal{M}_1 = M$ random items; $\mathcal{M}_2 = M$ new random items;
4: $\mathcal{R}_u \leftarrow \varnothing, \quad u \in [N]$ (items exploitable by user u)
5: $\{\mathcal{R}_u\}_{u \in [N]}) \leftarrow$ ITEMEXPLORE$(\mathcal{M}_1, \mathcal{M}_2, \ell)$
6: ITEMEXPLOIT$(\{\mathcal{R}_u\}_{u \in [N]})$

Algorithm 10.2 ITEMEXPLORE($\mathcal{M}_1, \mathcal{M}_2, \ell$)

$\mathcal{S}_j \leftarrow \varnothing, \quad j \in [\mathsf{q}_I]$ (initialize sets of items of type j)

Rate items by a few users (preliminary exploration)

Recommend each item in $\mathcal{M}_1$ and $\mathcal{M}_2$ to r_I users (This is done over $\lceil(|\mathcal{M}_1| + |\mathcal{M}_2|)\mathsf{r}_I/N\rceil < 2M\mathsf{r}_I/N + 1$ time-steps, with extra recommendations being random new items.)

Partition items and learn item types (joint exploration, oblivious to user structure)

for $j = 1, \cdots, \min\{\ell, \mathsf{r}_U\}$ **if** $\mathcal{M}_1 \neq \varnothing$ **do**

  $\mathrm{i}_j \leftarrow$ a random item in $\mathcal{M}_1$ (representative item)

  $a_{u,t} \leftarrow \mathrm{i}_j$ if $u \notin \mathsf{rated}_t(\mathrm{i}_j)$, otherwise a random item not from $\mathcal{M}_1$.

  $\widetilde{\mathcal{S}}_j \leftarrow \{i \in \mathcal{M}_1 : U = \mathsf{rated}(i) \cap \mathsf{rated}(\mathrm{i}_j), L_{U,i} = L_{U,\mathrm{i}_j}\}$

  $\mathcal{M}_1 \leftarrow \mathcal{M}_1 \setminus \widetilde{\mathcal{S}}_j$

Partition users and learn item types (joint exploration with user structure)

if $\ell > \mathsf{r}_U$ **then**

  Partition users into fewest possible groups such that each group agrees on all items. Let $\widehat{\tau}_U(u) \in [\mathsf{q}_U]$ be the label of user u's partition.

  $\mathcal{P}_k = \{u \in [N] : \widehat{\tau}_U(u) = k\}, \quad \forall k \in [\mathsf{q}_U]$

  for $j = \mathsf{r}_U + 1, \cdots, \ell$ **if** $\mathcal{M}_1 \neq \varnothing$ **do**

    $\mathrm{i}_j \leftarrow$ a random item in $\mathcal{M}_1$ (representative item)

    $a_{u,t} \leftarrow \mathrm{i}_j$ if $\forall u' \in \mathcal{P}_{\widehat{\tau}_U(u)}, u' \notin \mathsf{rated}_t(\mathrm{i}_j)$. (no user in the same partiotion of u has rated i_j before)

    $\widetilde{\mathcal{S}}_j \leftarrow \{i \in \mathcal{M}_1 : V = \mathsf{ratedtype}(i) \cap \mathsf{ratedtype}(\mathrm{i}_j), L_{V,i} = L_{V,\mathrm{i}_j}\}$

    $\mathcal{M}_1 \leftarrow \mathcal{M}_1 \setminus \widetilde{\mathcal{S}}_j$

for $j = 1, \cdots, \ell$ **do**

  $\mathcal{S}_j \leftarrow \{i \in \mathcal{M}_2 : V = \mathsf{ratedtype}(i) \cap \mathsf{ratedtype}(\mathrm{i}_j), L_{V,i} = L_{V,\mathrm{i}_j}\}$

  $\mathcal{M}_2 \leftarrow \mathcal{M}_2 \setminus \mathcal{S}_j$

$\mathcal{R}_u = \bigcup_{j: L_{u,\mathrm{i}_j} = +1} \mathcal{S}_j$ for each $u \in [N]$

if $u \in \mathcal{P}_k$ for some k **then**

  $\mathcal{R}_u = \bigcup_{u' \in \mathcal{P}_k} \bigcup_{j: L_{u',\mathrm{i}_j} = +1} \mathcal{S}_j$ for each $u \in [N]$

return $\{\mathcal{R}_u^s\}_{u \in [N]}$

Algorithm 10.3 EXPLOIT$\{\mathcal{R}_u\}_{u \in [N]}$

for remaining $t \leq T$ **do**

  for $u \in U$ **do**

    if there is an item $i \in \mathcal{R}_u$ such that $u \notin \mathsf{rated}_t(i)$. **then**

      $a_{u,t} \leftarrow i$

    else $a_{u,t} \leftarrow$ a random item not yet rated by u.

10.2 Performance analysis

Theorem 10.1. *Choosing the parameter ℓ as follows:*

$$\ell = \begin{cases} \mathsf{q}_I, & \text{if } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} > \mathsf{q}_I \\ \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T}, & \text{if } \mathsf{r}_U > \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \text{ and } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \leq \mathsf{q}_I \\ \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}, & \text{if } 8\log T \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \leq \mathsf{r}_U \\ 8\log T, & \text{if } 8\log T > \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}, \end{cases} \tag{10.1}$$

let $\mathsf{r}_I = 10\log(N\mathsf{q}_I^2)$ and $\mathsf{r}_U = 13\log(N\mathsf{q}_U^2)$. Algorithm 10.1 obtains regret per user at time T upper bounded as

$$\texttt{regret}(T) \leq \begin{cases} 8 + \mathsf{r}_U + \frac{\mathsf{q}_U}{N}(\mathsf{q}_I - \mathsf{r}_U) + 200\frac{\mathsf{q}_I\mathsf{r}_I}{N}T, & \text{if } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} > \mathsf{q}_I \\ 8 + \mathsf{r}_U(1 - \mathsf{q}_U/N) + 2\sqrt{200\frac{\mathsf{q}_U\mathsf{q}_I\mathsf{r}_I}{N}T}, & \\ \qquad\qquad \text{if } \mathsf{r}_U > \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} & \text{and } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \leq \mathsf{q}_I \\ 8 + 2\sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}, & \text{if } 8\log T \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \leq \mathsf{r}_U \\ 8\log T, & \text{if } 8\log T > \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}. \end{cases}$$

Proof. The basic error event is misclassification of an item or some users.

Item misclassification: In Algorithm 10.2, $\mathcal{S}_j$ is the set of items which the algorithm posits are of type j (i.e., of the same type as the jth representative i_j). Let $\mathcal{E}_i$ be the event that item i was misclassified,

$$\mathcal{E}_i = \{\exists j : i \in \mathcal{S}_j, \tau_I(i) \neq \tau_I(\mathrm{i}_j)\}. \tag{10.2}$$

In Lemma 10.2, we will find an upper bound for probability of $\mathcal{E}_i$.To do so, one should carefully distinguish slightly different proof techniques for items classified before and after partitioning the user space, as any error in partitioning the user space increases the probability of errors in classification of items.

User misclassification We bound the probability that the partitioning over users created by the algorithm is incorrect: Let $\mathcal{B}$ (defined in Equation (10.17) and (10.18)) be the event that the partitioning over the users is correct according to the type of users. Note that if $\ell < \mathsf{r}_U$, the algorithm does not partition the users. In this case, we defined $\mathcal{B}$ in Equation (10.18) such that $\mathcal{B}$ holds. According to Lemma 10.7 $\mathbb{P}[\mathcal{B}^c] \leq 2/N^2$.

Also, according to the model introduced in Section 7.2, users of the same type rate all items similarly all the time. So, the only error that can happen in partitioning the users is when two or more distinct user types rate the first r_U representative items similarly. In this case, the users from these user types are not distinguishable from each other. Consequently, users of the same type are always in the same partition whether or not $\mathcal{B}$ holds.

Categorization of recommendations: The amount of time a user u spends in the exploration phase (Step 1 of the algorithm) is denoted by $T_0(u)$.

To bound regret (the expected number of bad recommendations), we partition the set of recommendations made by the algorithm according to: 1-whether or not recommendation was in the exploration phase; 2-whether or not the item is in the list of exploitable items $\mathcal{R}_u$; 3-whether or not the item misclassification event $\mathcal{E}_{a_{u,t}}$ occurs; and 4-whether or not the partitioning over the users is correct ($\mathcal{B}^c$ holds).

$$
\begin{aligned}
N\texttt{regret}(T) \leq\ & \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=1}^{T_0(u)} \mathbb{1}[L_{u,a_{u,t}} = -1]\right] \\
& + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \notin \mathcal{R}_u]\right] \\
& + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{R}_u, \mathcal{B}]\right] \\
& + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{R}_u, \mathcal{E}_{a_{u,t}}]\right] \\
& + \mathbb{E}\left[\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{R}_u, \mathcal{E}^c_{a_{u,t}}, \mathcal{B}^c]\right] \\
=:\ & \texttt{A1} + \texttt{A2} + \texttt{A3} + \texttt{A4} + \texttt{A5}\,. \qquad (10.3)
\end{aligned}
$$

- `A1` is the regret from early time-steps before $T_0(u)$. We control the value of $T_0(u)$ by studying the exploration exploitation tradeoff in the design and analysis of the algorithm. Proper choice of the parameter ℓ has an important role in bounding this term.

- `A2` is the regret due to not having enough items available for the exploitation phase, which is proved to be small with high probability for a good choice of M.

- `A3` is the regret due to the error in partitioning the users. This will be small with proper choice of r_U.

- `A4` is the regret due to exploiting the misclassified items. It will be small because few items are misclassified as a result of proper choice of r_I.

- `A5` is the regret due to exploiting correctly classified items under the correct partitioning of the users and we will see that $\texttt{A5} = 0$.

In Equation (10.3), we provide an upper bound for regret as we are double counting the recommendations in which not only the partitioning on user structure is incorrect,

but also the recommended item is also misclassfieid. The probability of this event is small so that using this upper bound does not make the analysis less relevant.

Bounding `A1`. It takes at most $\lceil \frac{2M\mathsf{r}_I}{N} \rceil$ units of time to rate each item in $\mathcal{M}_1$ and $\mathcal{M}_2$ by r_I users (since intially $|\mathcal{M}_1| = |\mathcal{M}_2| = M$ and later some items are removed from $\mathcal{M}_1$ and $\mathcal{M}_2$).

At each iteration ℓ representative items from different item types are rated by either all users (if $\ell < \mathsf{r}_U$) or one user from each partition of users (if $\ell \geq \mathsf{r}_U$). Let W_k be the set of users with type k and $w_k = |W_k|$

$$W_k = \{u : \tau_U(u) = k\}, \quad \text{and} \quad w_k = |\{u : \tau_U(u) = k\}| \,. \tag{10.4}$$

- So, if $\ell < \mathsf{r}_U$, user u rates at most $\lceil \frac{2M\mathsf{r}_I}{N} \rceil$ items in step 2 and ℓ items in step 5 of ITEMEXPLORE algorithm (Algorithm 10.2) .

$$T_0(u) \leq \left\lceil \frac{2M\mathsf{r}_I}{N} \right\rceil + \ell \leq \frac{2M\mathsf{r}_I}{N} + (\ell + 1) \,. \tag{10.5}$$

- If $\ell \geq \mathsf{r}_U$, for given user type $k \in [\mathsf{q}_U]$, all the users in W_k rate the first r_U representative items at each iteration they go through. The remaining $\ell - \mathsf{r}_U$ representative items are rated by at most one user in W_k. To see this, remember that we showed all the users of the same type end up in the same partition. But if the user partitioning is incorrect, there might be users from more than one user type in a partiotion. Since one user from each partition rate remaining $\ell - \mathsf{r}_U$ representative items, at most one user from W_k rates each of them. Hence,

$$\sum_{u \in W_k} T_0(u) \leq \sum_{u \in W_k} \left(\left\lceil \frac{2M\mathsf{r}_I}{N} \right\rceil + \mathsf{r}_U \right) + (\ell - \mathsf{r}_U) \leq w_k(\frac{2M\mathsf{r}_I}{N} + 1 + \mathsf{r}_U) + \ell - \mathsf{r}_U \,. \tag{10.6}$$

Hence, if $u \in \mathcal{P}_k$,

$$\mathcal{R}_u = \bigcup_{u' : u' \in \mathcal{P}_k} \bigcup_{\substack{j \in [\ell] \\ L_{u', \mathsf{i}_j} = +1}} \mathcal{S}_j \tag{10.7}$$

Note that if $\ell < \mathsf{r}_U$ and there is no partitioning, then $\mathcal{R}_u = \cup_{\substack{j\in[\ell] \\ L_{u,\mathrm{i}_j}=+1}} \mathcal{S}_j$.

At $t \leq T_0(u)$, the algorithm recommends random items to u. Since each item i has a uniformly distributed type $\tau_I(i)$, any given user likes item type j with probability half. For $\ell < \mathsf{r}_U$,

$$\begin{aligned}
\mathtt{2A1} &= 2\mathbb{E}\sum_{u=1}^{N}\sum_{t=1}^{T_0(u)} \mathbb{1}[L_{u,a_{u,t}} = -1] = \mathbb{E}\sum_{u=1}^{N} T_0(u) \\
&\leq \left(\left\lceil \frac{2M\mathsf{r}}{N} \right\rceil + \ell\right) N \leq 2M\mathsf{r} + (\ell+1)N\,, \qquad (10.8)
\end{aligned}$$

where we used Equation (10.5).

For $\ell \geq \mathsf{r}_U$,

$$\begin{aligned}
\mathtt{2A1} &= \mathbb{E}\sum_{u=1}^{N} T_0(u) = \mathbb{E}\sum_{k\in[\mathsf{q}_U]}\sum_{u\in W_k} T_0(u) \\
&\overset{(a)}{\leq} \mathbb{E}\sum_{k\in[\mathsf{q}_U]} [w_k(\frac{2M\mathsf{r}_I}{N} + 1 + \mathsf{r}_U) + \ell - \mathsf{r}_U] \\
&\overset{(b)}{\leq} N(\frac{2M\mathsf{r}_I}{N} + 1 + \mathsf{r}_U) + \mathsf{q}_U(\ell - \mathsf{r}_U) \qquad (10.9)
\end{aligned}$$

Here (a) is obtained by Equation (10.6). Definition of w_k in Equation (10.4) gives (b).

Bounding `A2`**.** According Algorithm 10.3, user u is recommended a new random item at $t \geq T_0(u)$ if there is no available items in $\mathcal{R}_u$ to recommend to it. Hence, for every $u \in [N]$ at these times $\mathbb{P}[L_{u,a_{u,t}} = -1 | a_{u,t} \notin \mathcal{R}_u] = \frac{1}{2}$. Also, an item $a_{u,t} \notin \mathcal{R}_u$ is recommended to u in Algorithm 10.3 only when all items in $\mathcal{R}_u$ have been recommended to this user before. So, the total number of times $a_{u,t} \notin \mathcal{R}_u$ is recommended at time interval $T_0(u) \leq t \leq T$, is at most $(T - |\mathcal{R}_u|)_+$. To observe that, note that if $|\mathcal{R}_u| \geq T$, then there are enough items in $\mathcal{R}_u$ up until time T, in which case $\sum_{t=T_0(u)+1}^{T} \mathbb{1}[a_{u,t} \notin \mathcal{R}_u] = 0$. Alternatively, if $|\mathcal{R}_u| < T$, then there will be at most $(T - |\mathcal{R}_u|)$ times that user u is recommended an item $a_{u,t} \notin \mathcal{R}_u$, which

implies that $\sum_{t=T_0(u)+1}^{T} \mathbb{1}[a_{u,t} \notin \mathcal{R}_u] \leq (T - |\mathcal{R}_u|)_+$. Hence,

$$\begin{aligned}
\texttt{A2} &= \sum_{u\in[N]} \sum_{t=T_0(u)+1}^{T} \mathbb{P}[L_{u,a_{u,t}} = -1, a_{u,t} \notin \mathcal{R}_u] \\
&= \sum_{u\in[N]} \sum_{t=T_0(u)+1}^{T} \frac{1}{2}\mathbb{P}[a_{u,t} \notin \mathcal{R}_u] \leq \sum_{u\in[N]} \frac{1}{2}\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] .
\end{aligned}$$

Using Lemma 10.12, we get

$$\texttt{A2} \leq 2TN\exp(-\frac{\ell}{8}) + \frac{32T}{\ell} \tag{10.10}$$

Bounding A3. Term A3 in Equation (10.3) is the regret due to the misclassification of users. Using Lemma 10.7, we know that the probability of having any error in partitioning the users is at most $2/N^2$. Hence,

$$\begin{aligned}
\texttt{A3} &= \mathbb{E}\sum_{u=1}^{N} \sum_{t=T_0u+1}^{T} \mathbb{1}[L_{u,a_{u,t}} = -1, a_{u,t} \in \mathcal{R}_u, \mathcal{B}] \\
&\leq \mathbb{E}\sum_{u=1}^{N} \sum_{t=T_0u+1}^{T} \mathbb{1}[a_{u,t} \in \mathcal{R}_u, \mathcal{B}] \leq \frac{2}{N}T
\end{aligned} \tag{10.11}$$

where the last inequality is proved in Lemma 10.7.

Bounding A4. Term A4 in Equation (10.3) is the expected number of mistakes made by the algorithm as a result of item misclassifications. Lemma 10.2 upper bounds the expected number of "potential item misclassifications" in the algorithm to provide an upper bound for this value:

$$\begin{aligned}
\texttt{A4} &= \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[L_{u,i}=-1, a_{u,t}=i, i\in\mathcal{R}_u, \mathcal{E}_i] \\
&\leq \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T}\sum_{i\in\mathbb{N}}\mathbb{1}[a_{u,t}=i, i\in\mathcal{R}_u, \mathcal{E}_i] \\
&\overset{(a)}{=} \mathbb{E}\sum_{u=1}^{N}\sum_{t=T_0(u)+1}^{T}\sum_{i\in\mathcal{M}_2}\mathbb{1}[a_{u,t}=i, i\in\mathcal{R}_u, \mathcal{E}_i] \\
&\overset{(b)}{\leq} N\mathbb{E}\left[\sum_{i\in\mathcal{M}_2}\mathbb{1}[\bar{\mathcal{E}}_{i,U_i}]\right] \leq NM\max_{i\in\mathcal{M}_2}\mathbb{P}[\mathcal{E}_i] \overset{(c)}{\leq} M\frac{3}{\mathtt{q}_I}\,. \qquad (10.12)
\end{aligned}$$

Here (a) follows because for every user u, the set of items $\mathcal{R}_u$ is a subset of $\mathcal{M}_2$; (b) is true since each item i is recommended at most N times; (c) is proved in Lemma 10.2 that for any $i\in\mathcal{M}_2, \mathbb{P}[\mathcal{E}_i] < 3/(\mathtt{q}_I N)$.

Bounding A5. By the definition of event $\mathcal{E}_i$ given in Equation (9.1), if an item $i\in\mathcal{S}_j$ is not misclassified, then $\tau_I(i)=\tau_I(\mathrm{i}_j)$. By definition of the model, user preferences for an item depend only on the type of the item (since $L_{u,i}=\xi_{\tau_U(u),\tau_I(i)}$), so all users rate i the same as i_j. If an item $i\in\mathcal{R}_u$ we know that there is some j such that $i\in\mathcal{S}_j$ and u likes item i_j.

$$\begin{aligned}
\mathbb{P}[L_{u,i}=-1|i\in\mathcal{R}_u,\mathcal{E}_i^c] &= \mathbb{P}[L_{u,i}=-1|\exists j: L_{u,\mathrm{i}_j}=+1, i\in\mathcal{S}_j, \tau_I(i)=\tau_I(\mathrm{i}_j)] \\
&= \mathbb{P}[L_{u,i}=-1|\xi_{u,\tau_I(i)}=+1]=0 \\
&\Rightarrow \texttt{A5}=0\,. \qquad (10.13)
\end{aligned}$$

Combining all the bounds. Plugging in Equation (9.3), (9.4), (9.5) and (9.6) into Equation (9.2) gives for $\ell \leq \mathsf{r}_U$

$$\begin{aligned}\texttt{regret}(T) &\leq \frac{128\mathsf{q}_I\mathsf{r}_I}{\ell N}T + 1 + \ell + 2T\exp(-\frac{\ell}{8}) + \frac{32}{\ell N}T + \frac{2}{N^2}T + \frac{19}{\ell N}T \\ &\leq \frac{200\mathsf{q}_I\mathsf{r}_I}{\ell N}T + 1 + \ell + 2T\exp(-\frac{\ell}{8})\,.\end{aligned}$$

and for $\ell > \mathsf{r}_U$

$$\begin{aligned}\texttt{regret}(T) &\leq \frac{128\mathsf{q}_I\mathsf{r}_I}{\ell N}T + 1 + \mathsf{r}_U + \frac{\mathsf{q}_U}{N}(\ell - \mathsf{r}_U) + 2T\exp(-\frac{\ell}{8}) + \frac{32}{\ell N}T + \frac{2}{N^2}T + \frac{19}{\ell N}T \\ &\leq \frac{200\mathsf{q}_I\mathsf{r}_I}{\ell N}T + \frac{\mathsf{q}_U}{N}\ell + 2T\exp(-\frac{\ell}{8}) + 1 + (1 - \frac{\mathsf{q}_U}{N})\mathsf{r}_I\,.\end{aligned}$$

The choice of ℓ in the algorithm gives the statement of the theorem.

$$\ell = \begin{cases} \mathsf{q}_I, & \text{if } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \geq \mathsf{q}_I \\ \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T}, & \text{if } \mathsf{r}_U \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \text{ and } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \leq \mathsf{q}_I \\ \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}, & \text{if } 8\log T \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \leq \mathsf{r}_U \\ 8\log T, & \text{if } 8\log T \geq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \end{cases}$$

gives

$$\texttt{regret}(T) \leq \begin{cases} 8 + \mathsf{r}_U + \frac{\mathsf{q}_U}{N}(\mathsf{q}_I - \mathsf{r}_U) + 200\frac{\mathsf{q}_I\mathsf{r}_I}{N}T, & \text{if } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \geq \mathsf{q}_I \\ 8 + \mathsf{r}_U(1 - \mathsf{q}_U/N) + 2\frac{\sqrt{200\mathsf{q}_U\mathsf{q}_I\mathsf{r}_I T}}{N}, & \text{if } \mathsf{r}_U \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \text{ and } \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{\mathsf{q}_U}T} \leq \mathsf{q}_I \\ 8 + 2\sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}, & \text{if } 8\log T \leq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T} \leq \mathsf{r}_U \\ 8\log T, & \text{if } 8\log T \geq \sqrt{\frac{200\mathsf{q}_I\mathsf{r}_I}{N}T}\,. \end{cases}$$

□

10.2.1 Error probability in item classification

For a sequence of labels of learned representative items $X_{1:m} = X_1, \cdots, X_m$, let

$$\mathcal{E}_i(X_{1:m}) = \{\exists j \in \{X_{1:m}\} : i \in \mathcal{S}_j, \tau_I(i) \neq \tau_I(\mathrm{i}_j)\} \tag{10.14}$$

to be the event that item i is misclassified after learning the first m representative items. According to the definition of $\mathcal{E}_i$ given in Equation (10.2), $\mathcal{E}_i = \mathcal{E}_i([\mathsf{q}_I])$.

Lemma 10.2. *Given N large enough so that $\mathsf{q}_U > 10\mathsf{r}_I$, for every item $i \in \mathcal{M}_2$, the probability of mis-classifying item i is upper bounded as $\mathbb{P}[\mathcal{E}_i] \leq \frac{3}{\mathsf{q}_I N}$.*

Proof. To bound the probability of misclassifying any item, we look at items that were classified using the partitioning over the users ($i \in \mathcal{S}_j$ with $j \geq \mathsf{r}_U$) and the ones that were classified not using the partitioning over the users ($i \in \mathcal{S}_j$ with $j < \mathsf{r}_U$) separately. Hence, we study the error events $\mathcal{E}_i([\mathsf{r}_I])$ and $\mathcal{E}_i([\mathsf{q}_I] \setminus [\mathsf{r}_I])$ separately.To do so, first we guarantee that there are enough distinct user types rating each item in the following claim:

Claim 10.3. *Given N large enough so that $\mathsf{q}_U > 10\mathsf{r}_I$, the r_I users rating any item in $\mathcal{M}_2$ belong to at least $\frac{\mathsf{r}_I}{2}$ distinct user types with probability at least $1 - \frac{1}{\mathsf{q}_I N}$.*

Proof. The choice of r_I users rating any item in $\mathcal{M}_2$ is independent of the feedback. For given item i and $m \in [\mathsf{r}_I]$ let X_m be a binary valued random variable which denotes whether the $m-$th user which rates item i is from a new user type or not. We are interested to find an upper bound for $\mathbb{P}[\sum_{m=1}^{\mathsf{r}_I} X_i < \frac{\mathsf{r}_I}{2}]$. To do so, note that $\mathbb{P}[X_m = 1 | \sum_{m'=1}^{m-1} X_{m'} = d] \geq 1 - \frac{d}{\mathsf{q}_U}$, this is a direct consequence of the uniform prior distribution of user types over users. Since we are interested in $\mathbb{P}[\sum_{m=1}^{\mathsf{r}_I} X_i < \frac{\mathsf{r}_I}{2}]$, we know that the sequence of random variables X_m stochastically dominates an i.i.d. binary-valued random variable with mean $\frac{19}{20}$. Hence, $\mathbb{P}[\sum_{m=1}^{\mathsf{r}_I} X_i < \frac{\mathsf{r}_I}{2}] \leq \exp(-\frac{\mathsf{r}_I}{10}) \leq \frac{1}{\mathsf{q}_I N}$. □

Claim 10.4. *For item $i \in \mathcal{S}_j$, such that $j < \mathsf{r}_U$, we have $\mathbb{P}[\mathcal{E}_i] \leq \frac{3}{\mathsf{q}_I N}$, i.e.,*

$$\mathbb{P}\big[\, \mathcal{E}_i([\mathsf{r}_U]) \,\big] \leq \frac{2}{\mathsf{q}_I N}\,.$$

Proof. For item $i \in \mathcal{M}_2$ and subset of users $U_i \subseteq [N]$, define

$$\bar{\mathcal{E}}_{i,U_i} = \{\exists j \neq \tau_I(i) : L_{u,i} = \xi_{\tau_U(u),j}, \forall u \in U_i\} \tag{10.15}$$

to be the event that the ratings of users in U_i for item i agree with a different type. For item $i \in \mathcal{S}_j$, such that $j < \mathsf{r}_U$, we know that item i was added to $\mathcal{S}_j$ since some groups of users of size r_I have rated i and i_j similarly. Let t be the time i was added to $\mathcal{S}_j$ in the exploration phase. Let $U_i = \mathsf{rated}_t(i) \cap \mathsf{rated}_t(\mathrm{i}_j)$ be the set of users whose ratings verified that i and i_j are of the same type. All users in U_i agree on i vs. i_j. Hence, if item $i \in \mathcal{S}_j$ is misclassified (which implies $\tau_I(i) \neq \tau_I(\mathrm{i}_j)$ using Equation (9.1)), then $\bar{\mathcal{E}}_{i,U_i}$ holds. So, $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,U_i}$.

Using the statement of Claim10.3, there are at least $\mathsf{r}_I/2$ users with distinct user types who rate both item $i \in \mathcal{S}_j$ and i_j. A set of $\mathsf{r}_I/2$ users with distinct types, chosen independently of the feedback, rate item i of type $\tau_I(i)$. Any two item types $j \neq j'$ have jointly independent user preferences $(\xi_{u,j})_{u\in[N]}, (\xi_{u,j'})_{u\in[N]}$, so two items of different types are rated in the same way by the $\mathsf{r}_I/2$ users of distinct types with probability at most $2^{-\mathsf{r}_I/2}$. The choice $\mathsf{r}_I = 10\log(N\mathsf{q}_I^2)$ and a union bound over item types j gives that conditional on the fact that U_i contains at least $\mathsf{r}_I/2$ distinct user types, $\mathbb{P}[\bar{\mathcal{E}}_{i,U_i}] \leq \frac{1}{\mathsf{q}_I N}$. Plugging in the statement of Claim 10.3 gives $\mathbb{P}\big[\,\mathcal{E}_i([\mathsf{r}_U])\,\big] \leq \frac{2}{\mathsf{q}_I N}$. □

Claim 10.5. *For item $i \in \mathcal{S}_j$, such that $j \geq \mathsf{r}_U$, we have $\mathbb{P}[\mathcal{E}_i] \leq \frac{3}{\mathsf{q}_I N}$, i.e.,*

$$\mathbb{P}\big[\,\mathcal{E}_i([\mathsf{q}_I] \setminus [\mathsf{r}_U])\,\big] \leq \frac{3}{\mathsf{q}_I N}$$

Proof. For an item $i \in \mathcal{M}_2$ and a subset of user types $V_i \subseteq [\mathsf{q}_U]$, define

$$\bar{\mathcal{E}}_{i,V_i} = \{\exists j \neq \tau_I(i) : \xi_{k,\tau_I(i)} = \xi_{k,j}, \forall k \in V_i\} \tag{10.16}$$

to be the event that the ratings of users with types in V_i for item i agrees with a different type. For item $i \in \mathcal{S}_j$, such that $j \geq \mathsf{r}_U$ we know that i was added to $\mathcal{S}_j$

because according to the partitioning made over the users, items i and i_j are rated similarly.

We will provide the upper bound on probability of error $\mathcal{E}_i$ conditional on $\mathcal{B}$ (the partitioning over the users is correct). Note that in Lemma 10.7 it is shown that $\mathbb{P}[\mathcal{B}^c] < 2/N^2$. But this bound is provided conditional on statement of Claim 10.4.

Let $\mathcal{B}^c$ holds. Let t be the time i was added to $\mathcal{S}_j$ in the exploration phase. Let $V_i = \mathsf{ratedtype}_t(i) \cap \mathsf{ratedtype}_t(\mathrm{i}_j)$ be the set of user types whose ratings verified that i and i_j are of the same type. Hence, if item $i \in \mathcal{S}_j$ is misclassified (which implies $\tau_I(i) \neq \tau_I(\mathrm{i}_j)$ using Equation (10.2)), then $\bar{\mathcal{E}}_{i,V_i}$ holds. So, $\mathcal{E}_i \subseteq \bar{\mathcal{E}}_{i,V_i}$ which implies $\mathbb{P}[\mathcal{E}_i] \leq \mathbb{P}[\bar{\mathcal{E}}_{i,V_i}]$.

Next, we will show that the probability of $\bar{\mathcal{E}}_{i,V_i}$ is small. To see that, we use Claim 10.3 that V_i contains at least $\mathsf{r}_I/2$ distinct user types with high probability.

The probability of the event $\bar{\mathcal{E}}_{i,V_i}$ is upper bounded as $\mathbb{P}[\bar{\mathcal{E}}_{i,V_i}] \leq$ Any two item types $j \neq j'$ have jointly independent user-type preferences $(\xi_{k,j})_{k\in[\mathsf{q}_U]}, (\xi_{k,j'})_{k\in[\mathsf{q}_U]}$, so two items of different types are rated in the same way by the $\mathsf{r}_I/2$ distinct user types with probability at most $2^{-\mathsf{r}/2}$. The choice $\mathsf{r}_I = 10\log(N\mathsf{q}_I^2)$ and a union bound over item types j, $\mathbb{P}[\bar{\mathcal{E}}_{i,V_i}|\mathcal{B}] \leq \mathsf{q}_I 2^{-\mathsf{r}_I/10} \leq \frac{1}{\mathsf{q}_I N}$. Plugging in Claim 10.3 gives

$$\mathbb{P}[\mathcal{E}_i] \leq \frac{2}{\mathsf{q}_I N} + \mathbb{P}[\mathcal{B}]\,.$$

Using the fact that $\mathbb{P}[\mathcal{B}] < 2/N^2$ and $\mathsf{q}_U = N^\alpha$, for $N > N_0(\alpha)$, we have $\mathbb{P}[\mathcal{E}_i] < \frac{3}{\mathsf{q}_I N}$. □

Claims 10.4 and 10.5 give the statement of lemma. □

10.2.2 Error probability in user classification

Definition 10.6. For user types k and k', let

$$\mathcal{B}_{k,k'} = \{\mathbb{1}[\widehat{\tau}_U(u) = \widehat{\tau}_U(v)] = \mathbb{1}[k = k'],\ \text{for all } u, v \text{ such that } \tau_U(u) = k \text{ and } \tau_U(v) = k'\} \tag{10.17}$$

be the event that user types k and k' are partitioned correctly with respect to each other in Algorithm 10.2. Let

$$\mathcal{B} = \cap \mathcal{B}_{k,k'} \tag{10.18}$$

be the event that the partitioning over the users match the partitioning induced by user types.

Lemma 10.7. *Probability of not distinguishing two user types k and k' in the partitioning over the users made by the algorithm is upper bounded as $\mathbb{P}[(\mathcal{B}_{k,k'})^c] \leq 2/(N\,\mathsf{q}_U^2)^2$. It follows that the partitioning over the users is correct with probability at least $\mathbb{P}[\mathcal{B}^c] \leq 2/N^2$.*

Proof. Distinct user types k and k' are distinguished in the partitioning made by the algorithm if there is at least one item among the first r_U representative items which is rated differently by user types k and k'. Knowing that the elements of matrix Ξ are i.i.d., it would be easy to find the probability of this event happening if the representative items were drawn uniformly at random.

But unfortunately, this is not the case: At each step, the choice of the next item type depends on the proportion of number of items of a given item type among the remaining items in $\mathcal{M}_1$. Hence, item of an item types which are more probable to be misclassified are less likely to be chosen as the representative items (as most items of these types are misclassified and already removed from $\mathcal{M}_1$). Hence, there is statistical dependence between matrix Ξ and the choice of representative items.

To solve this problem, let $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ be the k and $k'-$th rows of matrix Ξ. We bound the probability of choosing any specific sequence of distinct item types as the first r_U representatives conditional on fixed $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$. Then we show that with high probability, there are more than $\mathsf{q}_I/3$ item types which distinguish k and k'. Conditional on this property of matrix Ξ, with high probability there is at least one item among the first r_U learned representative items which distinguish k and k'.

Let $X_m = \tau_I(\mathrm{i}_m)$ be the item type of the $m-$th representative item. For given user types k and k', we want to find the probability that both user types k and k' rate all item types $X_1, \cdots, X_{\mathsf{r}_U}$ similarly. To do so, given $X_1, \cdots, X_{m-1}$ for $m < \mathsf{r}_U$, $\Xi_{k,\cdot}$

and $\Xi_{k',\cdot}$ we find the conditional probability of choosing item type X_m as the next type to learn.

Let $\mathcal{L}_{k,k'} = \{j : \xi_{k,j} \neq \xi_{k',j}\}$ be the set of item types who are rated differently by user types k and k'. Hence, the event $\mathcal{B}_{k,k'}$ holds if there is some $m \in [\mathsf{r}_U]$, so that $X_m \in \mathcal{L}_{k,k'}$ (there is at least one learned item type among the ones that are helpful in distinguishing k and k'). Hence, we want to show an upper bound for

$$\mathbb{P}[(\mathcal{B}_{k,k'})^c] = \mathbb{P}\big[\, X_m \notin \mathcal{L}_{k,k'}, \text{ for all } m \in [\mathsf{r}_U]\,\big]\,. \tag{10.19}$$

Since the elements of matrix Ξ are independently Bern(1/2) distributed, using Chernoff bound in Lemma A.3,

$$\mathbb{P}[|\mathcal{L}_{k,k'}| < \mathsf{q}_U/3] \leq \exp(-\mathsf{q}_U/18)\,. \tag{10.20}$$

Note that the set $\mathcal{L}_{k,k'}$ is a function of $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$. Define $\bar{\mathcal{R}}(j)$ to be the set of items in $\mathcal{M}_1$ of type j

$$\bar{\mathcal{R}}(j) = \{i \in \mathcal{M}_1 : \tau_I(i) = j\}\,. \tag{10.21}$$

Define the error event $\widetilde{\mathcal{E}}_i$ for $i \in \mathcal{M}_1$ similar to the event $\mathcal{E}_i$ for $i \in \mathcal{M}_2$ given in Equation (9.1) so that

$$\widetilde{\mathcal{E}}_i(X_1, \cdots, X_m) = \{\exists j \in \{X_1, \cdots, X_m\} : i \in \widetilde{\mathcal{S}}_j, \tau_I(i) \neq \tau_I(\mathrm{i}_j)\}\,. \tag{10.22}$$

to be the event item $i \in \mathcal{M}_1$ is miscategorized into one of the representative item types labeled by $X_1, \cdots, X_m$. Let $\widetilde{\mathcal{E}}_i = \widetilde{\mathcal{E}}_i([\mathsf{q}_I])$ to be the error event associated with item $i \in \mathcal{M}_1$.

Definition 10.8. Let $\mathsf{Err}(X_1, \cdots, X_m)$, be the event such that on this event, for any item type j, at most $1/10-$th fraction of items in $\mathcal{M}_1$ of type j are misclassified into partitions corresponding to representative items labeled by $X_1, \cdots, X_m$. We use the

notation introduced in Equation 10.22 to define:

$$\mathsf{Err}(X_1, \cdots, X_m) = \left\{ \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i(X_1, \cdots, X_m)] \leq \frac{|\bar{\mathcal{R}}(j)|}{10}, \text{ for all } j \in [\mathsf{q}_I] \right\}. \quad (10.23)$$

Let $\mathsf{Err} = \mathsf{Err}([\mathsf{q}_I])$.

Using the total probability lemma, for any fixed value of $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$:

$$\begin{aligned}
&\mathbb{P}\Bigg[\sum_{i=1}^{\mathsf{r}_U} \mathbb{1}\big[\, X_i \in \mathcal{L}_{k,k'} \,\big] < 1 \,\Bigg|\, \Xi_{k,\cdot}, \Xi_{k',\cdot}, |\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3 \Bigg] \\
&\qquad \leq \mathbb{P}\Bigg[\sum_{i=1}^{\mathsf{r}_U} \mathbb{1}[X_i \in \mathcal{L}_{k,k'}] < 1 \,\Bigg|\, \mathsf{Err}([\mathsf{r}_U]), \Xi_{k,\cdot}, \Xi_{k',\cdot}, |\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3 \Bigg] \\
&\qquad\quad + \mathbb{P}\big[\, \mathsf{Err}^c([\mathsf{r}_U]) \,\big|\, \Xi_{k,\cdot}, \Xi_{k',\cdot}, |\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3 \,\big]. \qquad (10.24)
\end{aligned}$$

Application of Claim 10.9 provides an upper bound for the second term on the right hand side of above display. Lemma 10.10 provides the upper bound for the first term for all $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ such that $|\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3$ (remember that $\mathcal{L}_{k,k'}$ is a deterministic function of the k and $k'-$th rows of matrix Ξ). Hence,

$$\mathbb{P}\Bigg[\sum_{i=1}^{\mathsf{r}_U} \mathbb{1}\big[\, X_i \in \mathcal{L}_{k,k'} \,\big] < 1 \,\Bigg|\, \Xi_{k,\cdot}, \Xi_{k',\cdot}, |\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3 \Bigg] \leq \frac{1}{N\mathsf{q}_U^2}$$

Taking the expectation of the the above inequality with respect to $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ condition on $|\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3$ gives

$$\mathbb{P}\Bigg[\sum_{i=1}^{\mathsf{r}_U} \mathbb{1}\big[\, X_i \in \mathcal{L}_{k,k'} \,\big] < 1 \,\Bigg|\, |\mathcal{L}_{k,k'}| \geq \mathsf{q}_I/3 \Bigg] \leq \frac{1}{N\mathsf{q}_U^2} + \frac{40}{\mathsf{q}_I^4 N^{10}}$$

Plugging Equation (10.20) gives an upper bound for the right term in Equation (10.19). Hence,

$$\mathbb{P}\big[(\mathcal{B}_{k,k'})^c\big] \leq \frac{1}{(N\mathsf{q}_U^2)^2} + \frac{40}{\mathsf{q}_I^4 N^{10}} \leq \frac{2}{(N\mathsf{q}_U^2)^2}.$$

Union bound over $k, k' \in [\mathsf{q}_U]$ gives $\mathbb{P}[\mathcal{B}^c] \leq \frac{2}{N^2}$. □

Claim 10.9. *Let* Err *be the error defined in Definition 10.8 be the event under which*

at most $1/10$*-th fraction of items in* $\mathcal{M}_1$ *from any item types are mislcassified. Then,*

$$\mathbb{P}\big[\, \mathsf{Err}^c([\mathsf{r}_U]) \,\big|\, \Xi_{k,\cdot},\, \Xi_{k',\cdot}\,\big] \le \frac{40}{\mathsf{q}_I^4 N^{10}}$$

Proof. Similar to the proof of Claim 10.4, we defined the potential error event for item $i \in \mathcal{M}_1$, $\widetilde{\mathcal{E}}_{i,U_i}$ to be the event that there exists j such that $\tau_I(\mathrm{i}_j) \neq \tau_I(i)$ such that for all the users in U_i agree on i vs. i_j. As explained in the proof of Claim 10.4, for $i \in \widetilde{\mathcal{S}}_j$ and $j \le \mathsf{r}_U$, we have $\widetilde{\mathcal{E}}_i([\mathsf{r}_U]) \subseteq \widetilde{\mathcal{E}}_{i,U_i}$ for the $U_i = \mathsf{rated}(i) \cap \mathsf{rated}(\mathrm{i}_j)$ at the time the algorithm added i to $\widetilde{\mathcal{S}}_j$. Note that according to Claim 10.3, there are users of at least $\mathsf{r}_I/2$ distinct user types in U_i.

The remaining of the proof is very similar to the proof of Claim 9.3. A set U_i of users of at least $\mathsf{r}_I/2$ user types, chosen independently of feedback, rate item i. Conditional on the k and $k'-$th row of matrix Ξ (denoted by $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$), any two item types $j \neq j'$ have jointly independent user preferences by at least $\mathsf{r}_I/2 - 2$ user types in U_i. So, Condition on $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ the probability of not distinguishing two distinct item types by users in U_i at most $2^{-(\mathsf{r}_I/2-2)}$. The choice of $\mathsf{r}_I = 10\log(\mathsf{q}_I N^2)$ and union bounding over item types give

$$\mathbb{P}\big[\, \widetilde{\mathcal{E}}_i([\mathsf{r}_U]) \,\big|\, \Xi_{k,\cdot}, \Xi_{k',\cdot}\,\big] \le \frac{4}{\mathsf{q}_I^4 N^{10}}\,.$$

Using Markov inequality gives

$$\mathbb{P}\Bigg[\sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}[\widetilde{\mathcal{E}}_i([\mathsf{r}_U])] > \frac{|\bar{\mathcal{R}}(j)|}{10} \,\Big|\, \Xi_{k,\cdot}, \Xi_{k',\cdot}\Bigg] \le \frac{40}{\mathsf{q}_I^4 N^{10}}\,.$$

Union bounding over $j \in [\mathsf{q}_I]$ and tower property of expectation gives the statement of the claim. □

Claim 10.10. *For given user types* k *and* k'*, let* $\mathcal{L}_{k,k'} = \{j \in [\mathsf{q}_I] : \xi_{k,j} \neq \xi_{k',j}\}$*. let* $X_m = \tau_I(i_m)$ *and event* $\mathsf{Err}(X_1, \cdots, X_m)$ *as in Definition 10.8. Then,*

$$\mathbb{P}\Bigg[\sum_{i=1}^{\mathsf{r}_U} \mathbb{1}[X_i \in \mathcal{L}_{k,k'}] < 1 \,\Bigg|\, \mathsf{Err}([\mathsf{r}_U]), \Xi_{k,\cdot}, \Xi_{k',\cdot}, |\mathcal{L}_{k,k'}| \ge \mathsf{q}_I/3\Bigg] \le \frac{1}{(N\,\mathsf{q}_U^2)^2}$$

Proof. To prove this, we use the statement of Lemma 10.11 which gives a lower bound for the type of next representative item conditional on the type of previous item types, $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ and the event $\mathsf{Err}([\mathsf{r}_U])$. Note that $\mathsf{Err}([\mathsf{r}_U])$ implies $\mathsf{Err}([m])$ for all $m \leq \mathsf{r}_U$. Assuming N being large enough so that $\mathsf{r}_U/\mathsf{q}_U < 5/6$, for $m \leq \mathsf{r}_U$ we have

$$\mathbb{P}\left[X_m \notin \mathcal{L}_{k,k'} \middle| X_1, \cdots, X_{m-1}, \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1]), |\mathcal{L}_{k,k'}| > \frac{\mathsf{q}_I}{3} \right] < \frac{81}{100} \frac{\frac{2\mathsf{q}_I}{3}}{\mathsf{q}_I - (m-1)} \leq \frac{7}{10}.$$

Hence, for $m \leq \mathsf{r}_U$, condition on $\Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1]), |\mathcal{L}_{k,k'}| > \frac{\mathsf{q}_I}{3}$, the random variable $\mathbb{1}[X_m \in \mathcal{L}_{k,k'}]$ stochastically dominates a Bern(7/10) random variable. Hence,

$$\mathbb{P}\left[\sum_{m=1}^{\mathsf{r}_U} \mathbb{1}[X_m \in \mathcal{L}_{k,k'}] = 0 \middle| \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1]), |\mathcal{L}_{k,k'}| > \frac{\mathsf{q}_I}{3} \right] \leq \left(\frac{7}{10}\right)^{\mathsf{r}_U} \leq \frac{1}{(N\mathsf{q}_U^2)^2}.$$

where the choice of $\mathsf{r}_U = 13 \log N\mathsf{q}_U^2$ gives the last inequality. □

In the following lemma, we study the probability distribution and dependence of choice of representative items from distinct items types. Note that the item types chosen to be learned are not i.i.d. over time: i) No item type can be chosen twice to be learned; ii) If a given item type has more items in $\mathcal{M}_1$, it has a higher probability of being picked earlier to be learned; iii)There is a slight dependence between the preference matrix Ξ and the distribution of item types being chosen to learn: If two item types j and j' are rated similarly by many users and one of them, j, was chosen to be learned before, there is a high probability that many items of type j' is $\mathcal{M}_1$ were misclassfied by the algorithm to be of type j. This reduces the probability of choosing j' in the next steps to be learned.

In the following lemma, we will study the effect of all these parameters and show that fixing at most two rows of matrix Ξ, at each step, the choice of item types to be learned is close to uniform over the types that have not been learned before. This lemma will be used to bound the probability of error of partitioning the users. It will

also be used in providing a lower bound of the number of learned item types which are liked by a given user type.

Lemma 10.11. *Let the sequence of random variables $X_1 = \tau_I(i_1), \cdots, X_m = \tau_I(i_m)$ denote the item types of the first m representative items to be learned by the algorithm.*

For given distinct user types k and k', let $\Xi_{k,\cdot}$ and $\Xi_{k',\cdot}$ to be the $k-$th and $k'-$th rows of matrix Ξ.

Conditional on the k and $k'-$th row of matrix Ξ, given $X_1, \cdots, X_{m-1}$, we are going to find upper and lower bounds for probability of choosing a representative item with a specific type $j \notin \{X_1, \cdots, X_{m-1}\}$ is almost uniform (up to a constant) over the set of item types not learned yet by the algorithm.

$$\mathbb{P}[X_m = j \big| X_1, \cdots, X_{m-1}, \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])] > \frac{81}{100} \frac{1}{q_I - (m-1)} \tag{10.25}$$

Proof. To bound the probability of choosing a representative item with a specific type at the next step, we look at the proportion of items the the given type compared to the total number of remaining items in $\mathcal{M}_1$.

Let $\bar{\mathcal{R}}(j)$ to be the set of items in $\mathcal{M}_1$ of type j

$$\bar{\mathcal{R}}(j) = \{i \in \mathcal{M}_1 : \tau_I(i) = j\}. \tag{10.26}$$

A new item type is chosen randomly among the remaining items in $\mathcal{M}_1$. Conditional on the k and $k'-$th row of matrix Ξ and the event $\mathsf{Err}([m-1])$, given $X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots$ and $\bar{\mathcal{R}}(X_{m-1})$, we are going to find upper and lower bounds for probability of choosing a representative item with a specific type $j \notin \{X_1, \cdots, X_{m-1}\}$:

$$\mathbb{P}[\, X_m = j \,\big|\, X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])\,] \tag{10.27}$$

Then, using tower property of expectation and some calculations gives the statement of the lemma.

For m, let t be the time the m-th representative item is chosen randomly by the algorithm. For $j \notin \{X_1, \cdots, X_{m-1}\}$ the probability of choosing representative so that

$X_m = j$ is equal to the proportion of items of type j in the remaining items in $\mathcal{M}_1$ at time $t-1$. Note that some items of type j might have been already miscategorized by time $t-1$. This corresponds to items $i \in \mathcal{M}_1$ such that $i \in \bar{\mathcal{R}}(j)$, but $i \in \widetilde{\mathcal{S}}_{m'}$ for some $m' < m$. It implies $\widetilde{\mathcal{E}}_i([m-1]) = 1$. So the total number of items of type j in the remaining items in $\mathcal{M}_1$ might be smaller than $\bar{\mathcal{R}}(j)$.

Define the error event $\widetilde{\mathcal{E}}_i(X_1, \cdots, X_m)$ for $i \in \mathcal{M}_1$ as in Equation (10.22). Remember that we use Definition 10.8 for the event $\mathsf{Err}([m-1])$.

The number of items of type j in $\mathcal{M}_1$ by time $t-1$ is

$$|\bar{\mathcal{R}}(j)| - \sum_{i \in \bar{\mathcal{R}}(j) \cap (\bigcup_{r \in [m-1]} \widetilde{\mathcal{S}}_{X_r})} \mathbb{1}\left[\, \widetilde{\mathcal{E}}_i \,\right] = |\bar{\mathcal{R}}(j)| - \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}\left[\, \widetilde{\mathcal{E}}_i([m-1]) \,\right].$$

which is lower bounded by $|\bar{\mathcal{R}}(j)|$. Conditioning on event $\mathsf{Err}([m-1])$, the number of items of type j in $\mathcal{M}_1$ by time $t-1$ is upper bounded by $\frac{9}{10}|\bar{\mathcal{R}}(j)|$.

The number of items removed from $\mathcal{M}_1$ by the time t when X_m is chosen is

$$\sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)| + \sum_{j \in [\mathsf{q}_I] \setminus \{X_1, \cdots, X_{m-1}\}} \sum_{i \in \bar{\mathcal{R}}(j)} \mathbb{1}\left[\, \widetilde{\mathcal{E}}_i([m-1]) \,\right],$$

which is upper bounded by $\sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|$. Conditioning on event $\mathsf{Err}([m-1])$, the total number of remaining items in $\mathcal{M}_1$ is upper bounded by $\frac{9}{10}(M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|)$.

Lower bound for (10.27)**.** For $j \notin \{X_1, \cdots, X_m\}$, we have

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])] \geq \frac{9}{10} \frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|}. \tag{10.28}$$

Upper bound for (10.27)**.** For $j \notin \{X_1, \cdots, X_m\}$, we have

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(\cdot), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])] \geq \frac{10}{9} \frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|}. \tag{10.29}$$

Taking the expectation of $|\bar{\mathcal{R}}(j)|$ conditional on

$$X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])$$

for any $j \notin \{X_1, \cdots, , X_m\}$ is invariant to choice of j. So, there exists a constant C

$$C = \mathbb{E}\left[\frac{|\bar{\mathcal{R}}(j)|}{M - \sum_{l=1}^{m-1} |\bar{\mathcal{R}}(X_l)|} \middle| X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])\right]$$

which is independent of j as long as $j \notin \{X_1, \cdots, X_m\}$. Using tower property of expectation on Equation (10.28) and (10.29) along with the definition of C above gives

$$\frac{9}{10}C < \mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])] < \frac{10}{9}C\,.$$

Since there are $\mathsf{q}_I - (m-1)$ types j such that $j \notin \{X_1, \cdots, X_{m-1}\}$ taking the summation over j's, the second inequality of the above display gives $C \geq \frac{9}{10}\frac{1}{\mathsf{q}_I - (m-1)}$. The left inequality of the above display gives for all $j \notin \{X_1, \cdots, , X_m\}$,

$$\mathbb{P}[X_m = j | X_1, \cdots, X_{m-1}, \bar{\mathcal{R}}(X_1), \cdots, \bar{\mathcal{R}}(X_{m-1}), \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])] > \frac{81}{100}\frac{1}{\mathsf{q}_I - (m-1)}\,.$$

Using $\epsilon = \frac{1}{\mathsf{q}_I N}$ and $N > 32$,

$$\mathbb{P}\big[\, X_m = j | X_1, \cdots, X_{m-1}, \Xi_{k,\cdot}, \Xi_{k',\cdot}, \mathsf{Err}([m-1])\,\big] \geq \frac{3}{8(\mathsf{q}_I - (m-1))}\,.$$

□

10.2.3 Probability of not finding enough exploitable items

Lemma 10.12. *For user $u \in [N]$,*

$$\mathbb{P}[|\mathcal{R}_u| \leq T] \leq 3\exp(-\frac{\ell}{18}) + \frac{64}{\ell N}$$

And,

$$\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] \leq 3T\exp(-\frac{\ell}{8}) + \frac{64T}{\ell N}\,.$$

Proof. Roughly speaking, any user u likes about half of the ℓ item types learned by the algorithm. The total number of items in $\mathcal{M}_2$ is $T\frac{64\mathsf{q}_I}{\ell}$. Hence, there are about $T\frac{64}{\ell}$ items from each of the item types in $\mathcal{M}_2$. So, with high probability $\mathcal{R}_u$ being the items in $\mathcal{M}_2$ which user u is expected is more than T.

Making this argument rigorous requires careful reasoning for the following reasons: i) $\mathcal{R}_u$ is the union of $\mathcal{S}_j$'s with $L_{u,\mathrm{i}_j} = +1$. But, for a specific j such that $L_{u,\mathrm{i}_j} = +1$, due to misclassification of items, there might be some items of the same type as i_j which have been misclassified and removed from $\mathcal{S}_j$. This issue is addressed carefully in the remaining of this proof. ii) The choice of item types learned by the algorithm at each step is statistically dependent of number of remaining items in each item type in $\mathcal{M}_1$. Again, due to the misclassification, this could be different from the actual number of items of each given type. Additionally, the probability of misclassifying an item of a given type depends on the ratings of a users for that item. Hence, the choice of next item type to be learned is not statistically independent of the rating of a given user u for a given item type. The effect of this dependency should be studied carefully. This phenomenon is addressed carefully in the proof of Claim 10.13.

For $u \in \mathcal{P}_k$, we define

$$\widetilde{\mathcal{R}}_u = \{i \in \mathcal{M}_2 : \tau_I(i) = \tau_I(\mathrm{i}_j),\ \text{for some } j \in [\ell] \text{ such that } L_{v,\mathrm{i}_j} = +1,\ \text{for some } v \in \mathcal{P}_k\} \tag{10.30}$$

to be the items in $\mathcal{M}_2$ whose types are similar to one of the i_j's which are liked by u. If there is no error in the classification of items, $\widetilde{\mathcal{R}}_u$ is equal to $\mathcal{R}_u$. It will be easier to bound $\widetilde{\mathcal{R}}_u$.

Note that if an item $i \in \widetilde{\mathcal{R}}_u$ is correctly classified by the algorithm, then $i \in \mathcal{R}_u$; so $i \in \widetilde{\mathcal{R}}_u \setminus \mathcal{R}_u$ implies $\mathcal{E}_i$ (misclassification defined in Equation (10.2)). Hence,

$$|\mathcal{R}_u| \geq |\widetilde{\mathcal{R}}_u| - \sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i]\,. \tag{10.31}$$

Observing from Claim 10.4, $\mathbb{P}[\mathcal{E}_i] \leq 2/(\mathtt{q}_I N)$. Hence,

$$\begin{aligned}\mathbb{P}[|\mathcal{R}_u| \leq T] &\leq \mathbb{P}[|\widetilde{\mathcal{R}}_u| - \sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i] \leq T] \\ &\leq \mathbb{P}[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T] + \mathbb{P}[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i] \geq \frac{T}{2}]\end{aligned} \tag{10.32}$$

and

$$\mathbb{E}\left[(T - |\mathcal{R}_u|)_+\right] \leq \mathbb{E}\left[(T - |\widetilde{\mathcal{R}}_u|)_+\right] + \mathbb{E}\left[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i]\right] . \tag{10.33}$$

We bound the second expectation of Equation (10.33) using Claim 10.4

$$\mathbb{E}\left[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i]\right] \leq |\mathcal{M}_2|\epsilon = \frac{64T}{N\ell} \tag{10.34}$$

and by Markov Inequality for any $s > 0$

$$\mathbb{P}[\sum_{i \in \mathcal{M}_2} \mathbb{1}[\mathcal{E}_i] \geq sT] \leq \frac{64}{s\ell N} . \tag{10.35}$$

We now bound tail probability of $|\widetilde{\mathcal{R}}_u|$. Remember $\widetilde{\mathcal{R}}_u$ is the set of items in $\mathcal{M}_2$ with types the same as representatives which are liked by some user in the same partition as user u.

Let $\widetilde{\mathcal{R}}(j)$ be the set of items in $\mathcal{M}_2$ whose types are the same as i_j :

$$\widetilde{\mathcal{R}}(j) = \{i \in \mathcal{M}_2 : \tau_I(i) = \tau_I(\mathrm{i}_j)\} . \tag{10.36}$$

Note that the sets $\widetilde{\mathcal{R}}(j)$ are mutually exclusive. Let

$$\mathcal{L}_k^{(\ell)} = \{\tau_I(\mathrm{i}_j) : j \in [\ell], \xi_{k,\tau_I(\mathrm{i}_j)} = +1\} \tag{10.37}$$

be the set of learned item types in $[\ell]$ which are liked by user u. Note that the algorithm guarantees that each representative corresponds to a different item type.

Hence, if the partitioning over the users is correct,

$$|\widetilde{\mathcal{R}}_u| = \sum_{i \in \mathcal{M}_2} \mathbb{1}\big[\, \tau_I(i) \in \mathcal{L}^{(\ell)}_{\tau_U(u)} \,\big].$$

Note that according to the definition of $\mathcal{L}^{(\ell)}_k$ given in Equation (10.37), $\mathcal{L}^{(\ell)}_k$ is determined by: i)k-th row of matrix Ξ (denoted by $\Xi_{k,\cdot}$) which determines whether for all v such that $\tau_U(v) = \tau_U(u)$, $L_{v,\mathrm{i}_j} = \xi_{k,\tau_I(\mathrm{i}_j)} = +1$ or not; ii) the items in $\mathcal{M}_1$, their types and some randomness in the algorithm which determines the choice of i_j and $\tau_I(\mathrm{i}_j)$. Hence, the set $\mathcal{L}^{(\ell)}_k$ is statistically independent of all items $i \in \mathcal{M}_2$ and their item type. Hence, conditional on any given $\mathcal{L}^{(\ell)}_k$, the types $\tau_I(i)$ for $i \in \mathcal{M}_2$ is uniformly and independently distributed over $[\mathtt{q}_I]$. So, given $\mathcal{L}^{(\ell)}_k$ with size $\ell^{(\ell)}_k = |\mathcal{L}^{(\ell)}_k|$, condition on the partiotioning over the users being correct, $|\widetilde{\mathcal{R}}_u|$ is the sum of $|\mathcal{M}_2|$ i.i.d. binary random variables; it is a Binomial random variable with mean $M\frac{\ell_k}{\mathtt{q}_I} = 64T\frac{\ell_k}{\ell}$. Using Chernoff bound given in Lemma A.3, condition on $\ell_k \geq \frac{\ell}{30}$,

$$\mathbb{P}\left[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T \big| \ell^{(\ell)}_k \geq \ell/30, \mathcal{B}\right] \leq \exp(-T/11)\,.$$

The above inequality and the statement of Claim 10.13 and Lemma 10.7 give

$$\begin{aligned}
\mathbb{P}[|\widetilde{\mathcal{R}}_u| \leq \frac{3}{2}T] &< \mathbb{P}\left[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T \big| \ell^{(\ell)}_k \geq \ell/30, \mathcal{B}\right] + \mathbb{P}\left[\ell^{(\ell)}_k < \ell/30, \mathcal{B}\right] + \mathbb{P}[(\mathcal{B})^c] \\
&< \mathbb{P}\left[|\widetilde{\mathcal{R}}_u| < \frac{3}{2}T \big| \ell^{(\ell)}_k \geq \ell/30, \mathcal{B}\right] + \mathbb{P}\left[\ell^{(\ell)}_k < \ell/30\right] + \mathbb{P}[(\mathcal{B})^c] \\
&\leq \exp(-T/11) + \exp(-\ell/15) + \exp(-\mathtt{q}_I/18) + \frac{20}{N}\,.
\end{aligned}$$

Plugging this and Equation (10.35) into Equation (10.32) gives

$$\mathbb{P}[|\mathcal{R}_u| < T] \leq \exp(-T/11) + \exp(-\ell/15) + \exp(-\mathtt{q}_I/18) + \frac{20}{N} + \frac{128}{\ell N}\,.$$

Since $\ell < T$ and $\ell < \mathtt{q}_I$,

$$\mathbb{P}[|\mathcal{R}_u| < T] \leq 3\exp(-\frac{\ell}{18}) + \frac{148}{N}\,. \qquad \square$$

Claim 10.13. *For any $u \in [N]$, and $\ell_k^{(\ell)} = |\mathcal{L}_k^{(\ell)}|$ with $\mathcal{L}_k^{(\ell)}$ defined in Equation* (10.37) *to be the set of learned item types which u likes,*

$$\mathbb{P}[\ell_k^{(\ell)} < \ell/30] < \exp(-\ell/15) + \exp(-\mathsf{q}_I/18) + \frac{20}{N}\,.$$

Proof. For given user type k, the parameter $\ell_k^{(\ell)}$ denotes the number of learned item types liked by user type k. The set of item types learned by the algorithm is a function of the number of items of a given item type in $\mathcal{M}_1$ at each step; misclassifying items in previous steps changes the number of items of a given type at a given time. Also, the probability of misclassifying an item from a given type depends on the preference of users for that item type. Hence, whether or not user type u likes an item type is statistically dependent on whether or not that item type is chosen as one of the learned item types. We will show that the effect of this dependency is small; subsequently, we find upper bounds for probability of $\ell_k^{(\ell)} < \ell/30$.

Let $\mathcal{L}_k$ be the item types in $[\mathsf{q}_I]$ which are liked by user type k.

$$\mathcal{L}_k = \{j \in [\mathsf{q}_I] : \xi_{k,j} = +1\} \tag{10.38}$$

Since the elements of matrix Ξ (determining $\mathcal{L}_k$) are independetly Bern(1/2) distributed, using Chernoff bound in Lemma A.3 gives

$$\mathbb{P}[|\mathcal{L}_k| < \mathsf{q}_I/3] < \exp(-\mathsf{q}_I/18)\,. \tag{10.39}$$

We will show that

$$\mathbb{P}\big[\ell_k^{(\ell)} \leq \ell/30 \,\big|\, |\mathcal{L}_k| > \mathsf{q}_I/3\big] \leq \exp(-\ell/5) + \frac{20}{N} \tag{10.40}$$

To do so, let the sequence of random variables $X_1 = \tau_I(\mathrm{i}_1), X_2 = \tau_I(\mathrm{i}_2), \cdots$ denote the item types chosen to be learned by the algorithm. Then, $|\mathcal{L}_k^{(\ell)}| = \sum_{j\in[\ell]} \mathbb{1}[X_j \in \mathcal{L}_k]$. Note that the set $\mathcal{L}_k$ is a one-to-one function of $k-$th row of matrix Ξ (denoted by

$\Xi_{k,\cdot}$). Define $\bar{\mathcal{R}}(j)$ to be the set of items in $\mathcal{M}_1$ of type j

$$\bar{\mathcal{R}}(j) = \{i \in \mathcal{M}_1 : \tau_I(i) = j\}\,. \tag{10.41}$$

Define the error event $\widetilde{\mathcal{E}}_i$ for $i \in \mathcal{M}_1$, defined in Equation (10.22), be the event item $i \in \mathcal{M}_1$ is miscategorized into one of the representative item types labeled by $X_1, \cdots, X_m$. Also, let $\widetilde{\mathcal{E}}_i = \widetilde{\mathcal{E}}_i([\mathsf{q}_I])$ be the error event associated with item $i \in \mathcal{M}_1$ similar to before.

Using Definition 10.8 for event $\mathsf{Err}(X_1, \cdots, X_m)$, applying the total probability lemma, for any fixed value of $\Xi_{k,\cdot}$:

$$\begin{aligned}
&\mathbb{P}\Big[\sum_{i=1}^{\ell} \mathbb{1}\big[\,X_i \in \mathcal{L}_k\,\big] < \ell/30 \,\Big|\, \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3\Big] \\
&\qquad\qquad \leq \mathbb{P}\Big[\sum_{i=1}^{\ell} \mathbb{1}\big[\,X_i \in \mathcal{L}_k\,\big] < \ell/30 \,\Big|\, \mathsf{Err}([\ell]), \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3\Big] \\
&\qquad\qquad\quad + \mathbb{P}\big[\,(\mathsf{Err}([\ell]))^c \,\big|\, \mathsf{Err}([\ell])\Xi_{k,\cdot}|\mathcal{L}_k| \geq \mathsf{q}_I/3\,\big]\,.
\end{aligned} \tag{10.42}$$

Considering $\mathbb{P}\big[\,(\mathsf{Err}([\ell]))^c \,\big|\, \Xi_{k,\cdot}|\mathcal{L}_k| \geq \mathsf{q}_I/3\,\big] \leq \mathbb{P}\big[\,\mathsf{Err}^c \,\big|\, \Xi_{k,\cdot}|\mathcal{L}_k| \geq \mathsf{q}_I/3\,\big]$, using the tower property of expectation on the result of Claim 10.14 provides an upper bound for the second term on the right hand side of above display:

$$\mathbb{P}\big[\,(\mathsf{Err}([\ell]))^c \,\big|\, \Xi_{k,\cdot}|\mathcal{L}_k| \geq \mathsf{q}_I/3\,\big] \leq \frac{20}{N}\,. \tag{10.43}$$

Lemma 10.15 provides the upper bound for the first term for all $\Xi_{k,\cdot}$ such that $|\mathcal{L}_k| \geq \mathsf{q}_I/3$ (remember that $\mathcal{L}_k$ is a deterministic function of the $k-$th rows of matrix Ξ). Hence,

$$\mathbb{P}\Big[\sum_{i=1}^{\ell} \mathbb{1}\big[\,X_i \in \mathcal{L}_k\,\big] < \ell/30 \,\Big|\, \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3\Big] \leq \exp(-\ell/15)\,.$$

Plugging this into Equation (10.42) gives

$$\mathbb{P}\Big[\sum_{i=1}^{\ell} \mathbb{1}\big[\,X_i \in \mathcal{L}_k\,\big] < \ell/30 \,\Big|\, \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3\Big] \leq \exp(-\ell/15) + \frac{20}{N}$$

Taking the expectation of this with respect to $\Xi_{k,\cdot}$ condition on $|\mathcal{L}_k| \geq \mathsf{q}_I/3$ and lugging Equation (10.39) gives the statement of claim. $\square$

Claim 10.14. *Let* Err *be the error defined in Definition 10.8 be the event under which at most a fraction of* $1/10$*-th of items in* $\mathcal{M}_1$ *from any item types are mislcassified. Then,*

$$\mathbb{P}\big[\, \mathsf{Err}^c \,\big|\, \Xi_{k,\cdot} \,\big] \leq \frac{21}{N}\,.$$

Proof. We use

$$\mathbb{P}\big[\, \mathsf{Err}^c \,\big|\, \Xi_{k,\cdot} \,\big] \leq \mathbb{P}\big[\, \mathsf{Err}^c \,\big|\, \Xi_{k,\cdot}, \mathcal{B} \,\big] + \mathbb{P}[\mathcal{B}^c \,|\, \Xi_{k,\cdot}]\,.$$

The exact same steps in the proof of Lemma 10.7 shows that $\mathbb{P}[\mathcal{B}^c \,|\, \Xi_{k,\cdot}] \leq 2/N^2$.

To bound the first term in the above display, for any $i \in \mathcal{M}_1$ define

$$\widetilde{\mathcal{E}}_{i,U_i} = \{\exists j \neq \tau_I(i) : L_{u,i} = \xi_{\tau_U(u),j}\,, \forall u \in U_i\} \tag{10.44}$$

Condition on the partitioning over users being correct, the event $\widetilde{\mathcal{E}}_i \subseteq \widetilde{\mathcal{E}}_{i,U_i}$ for the set of users U_i which are used in classifying item i. Exactly similar to the proof of Claim 10.3, Given N large enough so that $\mathsf{q}_U > 10\mathsf{r}_I$, the r_I users rating any item in $\mathcal{M}_1$ belong to at least $\frac{\mathsf{r}_I}{2}$ distinct user types with probability at least $1 - \frac{1}{\mathsf{q}_I N}$. Conditioning on $\Xi_{k,\cdot}$ implies that there are at least $\mathsf{r}_I/2 - 1$ distinct user types (chosen independent of feedback and matrix Ξ) are used as witnesses in categorizing item i. Hence, the probability that there is another item type $j \neq \tau_I(i)$ which is rated the same by these $\mathsf{r}_I/2 - 1$ user types is $2^{-(\mathsf{r}_I/2-1)} \leq 2/(N\mathsf{q}_I^2)^5$. Hence,

$$\mathbb{P}\big[\, \widetilde{\mathcal{E}}_i \,\big|\, \Xi_{k,\cdot}, \mathcal{B} \,\big] \leq \frac{1}{N\mathsf{q}_I} + \frac{2}{(N\mathsf{q}_I^2)^5} \leq \frac{2}{N\mathsf{q}_I}\,.$$

Using Definition 10.8 for Err gives

$$\mathbb{P}\big[\, \mathsf{Err}^c \,\big|\, \Xi_{k,\cdot}, \mathcal{B} \,\big] \leq \frac{20}{N}\,.$$

The above two inequalities give the statement of the Claim. $\square$

Lemma 10.15. *Let the sequence of random variables $X_1 = \tau_I(i_1), X_2 = \tau_I(i_2), \cdots$ denote the item types chosen to be learned by the algorithm. Then,*

$$\mathbb{P}\left[\sum_{i=1}^{\ell} \mathbb{1}[X_i \in \mathcal{L}_k] < \ell/30 \,\middle|\, \mathsf{Err}([\ell]), \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3\right] < \exp(-\ell/15)\,.$$

Proof. We use the statement of Lemma 10.11 to show that for any $k \in [\mathsf{q}_U]$ and any $j \notin \{X_1, \cdots, X_{m-1}\}$

$$\mathbb{P}[X_m = j \big| X_1, \cdots, X_{m-1}, \Xi_{k,\cdot}, \mathsf{Err}([m-1])] > \frac{81}{100}\frac{1}{\mathsf{q}_I - (m-1)} \tag{10.45}$$

Hence, for $\Xi_{k,\cdot}$ such that $|\mathcal{L}_k| \geq \mathsf{q}_I/3$,

$$\mathbb{P}[X_m \in \mathcal{L}_k \,\big|\, \sum_{m'=1}^{m-1} \mathbb{1}[X_{m'} \in \mathcal{L}_k] < s, \Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3, \mathsf{Err}([m-1])] > \frac{81}{100}\frac{\frac{\mathsf{q}_I}{3} - s}{\mathsf{q}_I - (m-1)} > \frac{1}{5}\,.$$

where the last inequality holds for $s \leq \ell/30$ and $m \leq \ell \leq \mathsf{q}_I$. So the random variable $\sum_{m'=1}^{\ell} \mathbb{1}[X_{m'} \in \mathcal{L}_k]$ conditional on $\Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3$ and $\mathsf{Err}([m-1])$, stochastically dominates a $\mathrm{Bin}(\ell, 1/5)$ random variable. Hence,

$$\mathbb{P}[\sum_{m'=1}^{\ell} \mathbb{1}[X_{m'} \in \mathcal{L}_k] < \frac{\ell}{30} \,\big|\Xi_{k,\cdot}, |\mathcal{L}_k| \geq \mathsf{q}_I/3, \mathsf{Err}] > \exp(-\ell/15)\,.$$

□

10.3 General lower bound

In this section, we provide the lower bound for the joint model introduced in Section 7.2.

Theorem 10.16. *In the joint structure model, given $N > 3\mathtt{q}_U \log(2\mathtt{q}_U)$, there are numerical constants c_1, c_2, c_3 and C such that any recommendation algorithm must incur regret*

$$\mathtt{regret}(T) \geq -1 + C \begin{cases} \sqrt{\frac{T\mathtt{q}_I}{N}}, & \text{if } T \leq c_1 \frac{N(\log \mathtt{q}_U)^2}{\mathtt{q}_I} \\ \frac{\sqrt{T\mathtt{q}_I\mathtt{q}_U}}{N}, & \text{if } c_1 \frac{N(\log \mathtt{q}_U)^2}{\mathtt{q}_I} \leq T \leq c_2 \frac{\mathtt{q}_I\mathtt{q}_U}{(\log \mathtt{q}_U)^2} \\ T\frac{\mathtt{r}_I}{N}, & \text{if } c_2 \frac{\mathtt{q}_I\mathtt{q}_U}{(\log \mathtt{q}_U)^2} \leq T\,. \end{cases}$$

The strategy of the proof is very similar to the lower bounds of regret for user-user model and item-item model. First, we will show certain recommendations (defined in Equation (10.53)) are *bad* recommendations in the sense that the probability of them being successful is at best close to half. We will represent regret defined in Equation (7.1) as the expected number of bad recommendation. Then, using a careful counting argument, we can find a lower bound for the number of bad recommendations in any algorithm which gives the lower bound for regret. The proof of this statement is a straightforward consequence of Lemmas 10.30 and 10.42.

In this section we use the following notations: $\tau_I(\cdot) = \{\tau_I(i') : i' \in \mathbb{N}\}$ denotes the sequence of item types for all items. $\tau_I(\cdot \setminus \{i\}) = \{\tau_I(i') : i' \in \mathbb{N}, i' \neq i\}$ denotes the sequence of items types for all items excluding item i. The notations $\tau_U(\cdot)$ and $\tau_U(\cdot \setminus \{u\})$ are defined similarly.

$\Xi_{\cdot\setminus\{k,j\}} = \{\xi_{k',j'} : (k',j') \in [\mathtt{q}_U] \times [\mathtt{q}_I], (k',j') \neq (k,j)\}$ denotes all the elements in matrix Ξ except $\xi_{k,j}$.

Throughout the proof we use Definition 8.5 for (t,ϵ)-column regularity repeated here: Let $A \in \{-1,+1\}^{m\times n}$. For ordered tuple of distinct (column) indices $w = (i_1,\ldots,i_t) \in [n]^t$, let $M = (A_{\cdot i})_{i\in w} \in \{-1,+1\}^{m\times t}$ be the matrix formed by concatenating columns of A indexed by w. For given row vector $b \in \{-1,+1\}^t$, let

$K_{b,w}(A) \subseteq [m]$ be the set of rows in $M = (A_{\cdot i})_{i \in w}$ which are identical to the row b . The cardinality of $K_{b,w}(A)$ is denoted by $k_{b,w}(A)$. A is said to be (t, ϵ)-*column regular* if

$$\max_{w,b} \left| k_{b,w}(A) - \frac{m}{2^t} \right| \leq \epsilon \frac{m}{2^t} ,$$

where the maximum is over tuples w of t columns and ± 1 vectors b of size t. We define $\Omega_{t,\epsilon}$ to be the set of (t, ϵ)- column regular matrices.

According to the Definition 9.11, matrix A is (t, ϵ)-row regular (denoted by $A^T \in \Omega_{t,\epsilon}$) if its transpose is $(t, \epsilon)-$ column regular.

Claim 8.6 shows that if a matrix $A \in \{-1, +1\}^{m \times n}$ is (t, ϵ)-*column regular*, then it is also (s, ϵ)-*column regular* for all $s < t$.

In Lemma 8.7 it is proved that for matrix $A \in \{-1, +1\}^{m \times n}$ with i.i.d. Bern(1/2) entries is (t, ϵ)-column regular with probability at least

$$1 - 2(2n)^t \max \left\{ \exp\left(-\frac{\epsilon^2}{3} \frac{m}{2^t} \right), \exp\left(-\frac{\epsilon}{2} \frac{m}{2^t} \right) \right\} .$$

Definition 10.17. Let c_i^t be the number of user types from which a user has rated item i by time $t - 1$:

$$c_i^t := |\{k \in [\mathsf{q}_U] : \text{there exists user } u \in [N] \text{with user type } \tau_U(u) = k, \text{ s.t. } a_{u,s} = i \text{ for some } s < t\}| .$$

For any $t \leq T$ and any constant r_I,

$$\mathbb{1}[c_i^t \geq \mathsf{r}_I] \leq \mathbb{1}[c_i^T \geq \mathsf{r}_I] . \tag{10.46}$$

Definition 10.18. Let d_u^t be the number of distinct item types that were rated by user u up to time $t - 1$:

$$d_u^t := |\{j \in [\mathsf{q}_I] : \text{there exists item } i \in \mathbb{N} \text{with item type } \tau_I(i) = j, \text{ s.t. } a_{u,s} = i \text{ for some } s < t\}| .$$

For any $t \leq T$ and any constant r_U,

$$\mathbb{1}[d_u^t \geq \mathsf{r}_U] \leq \mathbb{1}[d_u^T \geq \mathsf{r}_U]\,. \tag{10.47}$$

Throughout this section, let

$$\epsilon = \frac{1}{\log N}, \quad \mathsf{r}_I = \log \mathsf{q}_I - 5 \log\log N, \quad \mathsf{r}_U = \log \mathsf{q}_U - 5 \log\log N\,. \tag{10.48}$$

Claim 10.19. *For any user $u \in [N]$, and item i recommended to u at time t,*

$$\mathbb{P}[L_{u,i} = +1 \,\big|\, a_{u,t} = i, c_i^t < \mathsf{r}_I, d_u^t < \mathsf{r}_U, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\,, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \leq \frac{1+4\epsilon}{2}.$$

Proof. We are going to show that given a preference matrix, there are many item types that are consistent with the history of an item i which has been rated by less than r_I user types. Similarly, there are many user types consistent with the history of a user that has rated less than r_U item types. The row regularity and the column regularity of the preference matrix implies that uncertainty in types of item and user results in uncertainty of outcome of recommending i to u.

Given Ξ, $\tau_I(. \setminus \{i\})$ $\tau_U(. \setminus \{u\})$ and H_{t-1} let $w_i = \{k \in [\mathsf{q}_U] : a_{u',s} = i, \tau_U(u') = k, s < t\}$ be the set of user types that were recommended item i up to time $t-1$, and let b_i be the vector of feedback from users of types in w_i about item i (for user type $k \in w_i$, the corresponding element in b_i is $L_{u,i}$ for some user u' with $\tau_U(u') = k$). Note that since $a_{u,t} = i$ and each user is recommended an item only once, we know that user u was not recommended item i by time $t-1$. Hence conditioning on $\tau_U(. \setminus \{u\})$ (not knowing the type of user u) does not make the definition of w_i ambiguous.

Using the notation in Definition 8.5, if M is the matrix formed by concatenating the rows of Ξ indexed by w_i, then $K_{b_i,w_i}(\Xi^T)$ would be the set of columns (corresponding to the item types) which are identical. We claim that conditional on matrix Ξ, vectors b_i and w_i, the type $\tau_I(i)$ of item i at the end of time instant $t-1$ is uniformly distributed over the set of item types $K_{b_i,w_i}(\Xi^T)$ consistent with this data. This can be derived by applying Bayes rule considering that the prior distribution of type of

each item is uniform over $[\mathsf{q}_I]$.

Very similarly, given Ξ, $\tau_I(. \setminus \{i\})$ $\tau_U(. \setminus \{u\})$ and H_{t-1} let $w_u = \{j \in [\mathsf{q}_I] : a_{u,s} = i', \tau_I(i') = j, s < t\}$ be the set of item types that were recommended to user u up to time $t-1$, and let b_u be the vector of feedback from recommending items of types in w_u to user u (for item type $j \in w_u$, the corresponding element in b_u is $L_{u,i'}$ for some user i' with $\tau_I(i') = j$). By a similar argument as above, conditioning on $\tau_I(. \setminus \{i\})$ does not make the definition of w_u ambiguous.

As before, using the notation in Definition 8.5, conditional on matrix Ξ, vector b_u and w_u, the type $\tau_U(u)$ of user u at the end of time instant $t-1$ is uniformly distributed over the set of user types $K_{b_u,w_u}(\Xi)$.

So, conditioning on $\Xi, \tau_I(. \setminus \{i\}), \tau_U(. \setminus \{u\})$ and H_{t-1} implies conditioning on

$$\{\tau_U(u) \in K_{b_u,w_u}(\Xi), \tau_I(i) \in K_{b_i,w_i}(\Xi^T)\}$$

and Ξ for the corresponding b_u, w_u, b_i and w_i. On the other hand conditioning on $c_i^t < \mathsf{r}_I$ and $d_u^t < \mathsf{r}_U$ implies $w_i < \mathsf{r}_I$ and $w_u < \mathsf{r}_U$.

The event $\{L_{u,i} = +1\}$ is equivalent to the event

$$\{\exists j \in [\mathsf{q}_I], k \in [\mathsf{q}_U], \xi_{k,j} = +1, \tau_U(u) = k, \tau_I(i) = j\}.$$

Using total probability lemma,

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 | \Xi, \tau_I(. \setminus \{i\}), \tau_U(. \setminus \{u\}), H_{t-1} = H] \\
&\quad \overset{(a)}{=} \sum_{j \in K_{b_i,w_i}(\Xi^T)} \mathbb{P}[\tau_I(i) = j,\ \tau_U(u) \in K_{b_u^+,\{w_u,j\}}(\Xi) | \tau_U(u) \in K_{b_u,w_u}(\Xi),\ \tau_I(i) \in K_{b_i,w_i}(\Xi^T)] \\
&\quad \overset{(b)}{=} \sum_{j \in K_{b_i,w_i}(\Xi^T)} \frac{1}{k_{b_i,w_i}(\Xi^T)} \frac{k_{b_u^+,\{w_u,j\}}(\Xi)}{k_{b_u,w_u}(\Xi)}
\end{aligned}$$

here equality (a) is application of total probability lemma on the type of item i. Equality (b) uses the fact that, condition on variables specified above at time $t-1$:

i)type of item i is uniform over $K_{b_i,w_i}(\Xi^T)$; ii)type of user u is uniform over $K_{b_u,w_u}(\Xi)$; iii) For given $j \in K_{b_i,w_i}(\Xi^T)$, the probability of type of user u being in $K_{b_u^+,\{w_u,j\}}(\Xi)$ is uniform proportional to the size of this set.

Using tower property of expectation on the above gives:

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 \,\big|\, a_{u,t} = i, c_i^t < \mathsf{r}_I, d_u^t < \mathsf{r}_U, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\,, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \\
&\overset{(b)}{=} \sum_{j \in K_{b_i,w_i}(\Xi^T)} \frac{1}{k_{b_i,w_i}(\Xi^T)} \frac{k_{b_u^+,\{w_u,j\}}(\Xi)}{k_{b_u,w_u}(\Xi)} \\
&\overset{(c)}{\leq} \frac{\frac{\mathsf{q}_U}{2^{|w_u|+1}}(1+\epsilon)}{\frac{\mathsf{q}_U}{2^{|w_u|}}(1-\epsilon)} \overset{(d)}{\leq} \frac{1}{2}(1+4\epsilon)
\end{aligned}$$

The last inequality is the result of the fact that: i)$k_{b_i,w_i}(\Xi^T) > 1$ since $|w_i| < \mathsf{r}_I$; ii)$|w_u| \leq \mathsf{r}_U$; iii)The preference matrix Ξ in $(\mathsf{r}_U, \epsilon)-$ column regular: for $|w_u| < \mathsf{r}_U$, , $k_{b_u^+,\{w_u,j\}}(\Xi) < \frac{\mathsf{q}_U}{2^{|w_u|+1}}(1+\epsilon)$ and $k_{b_u,w_u}(\Xi) > \frac{\mathsf{q}_U}{2^{|w_u|}}(1-\epsilon)$. Finally, inequality (d) uses $\epsilon < \frac{1}{\log N}$. □

Corollary 10.20. *Given (r_U, ϵ)-column regular and (r_I, ϵ)-row regular preference matrix Ξ, recommending an item that has been rated by less than r_I user types to a user that has rated less than r_U item types gives uncertain outcome (these are among bad recommendations made by the algorithm):*

$$\begin{aligned}
\mathtt{A1} &:= \mathbb{P}[L_{u,a_{u,t}} = +1, c_{a_{u,t}}^t < \mathsf{r}_I, d_u^t < \mathsf{r}_U, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\,, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \\
&\leq \frac{1+4\epsilon}{2}\, \mathbb{P}[c_{a_{u,t}}^t < \mathsf{r}_I, d_u^t < \mathsf{r}_U, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\,, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}]\,.
\end{aligned}$$

Proof. Multiplying the statement of Claim 10.19 by $\mathbb{P}[a_{u,t} = i, c_i^t < \mathsf{r}_I, d_u^t < \mathsf{r}_U, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\,, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}]$ and taking summation over i gives the corollary. □

Definition 10.21. Let $\mathcal{B}_{\tau_U(u),i}^t$ be the event that a user with type $\tau_U(u)$ has rated item i by time $t-1$, i.e.,

$$\mathcal{B}_{\tau_U(u),i}^t = \{\exists u' \in [N] : a_{u',s} = i \text{ for some } s < t \text{ with } \tau_U(u) = \tau_U(u')\}\,.$$

Claim 10.22. *Given (r_I, ϵ)-row regular preference matrix Ξ, if the number of user*

types that have rated item i is less than r_I, and no user of type $\tau_U(u)$ has rated i, then the outcome of recommending item i to user u is uncertain:

$$\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c_i^t < \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}] \leq \frac{1+4\epsilon}{2}\,.$$

Proof. Conditional on $\Xi, \tau_U(.), \tau_I(. \setminus \{i\})$ and H_{t-1} similar to the proof of Claim 10.19, define w_i, b_i and $K_{b_i,w_i}(\Xi^T)$. We will show that if the event $(\mathcal{B}^t_{\tau_U(u),i})^c$ holds, conditional on $\Xi, \tau_U(.), \tau_I(. \setminus \{i\})$ and vectors b_i and w_i, the random variable $L_{u,i}$ is statistically independent of H_{t-1} :

$$L_{u,i} \perp\!\!\!\perp H_{t-1} \big| (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi, \tau_U(\cdot), \tau_I(\cdot \setminus \{i\}), \tau_I(i) \in K_{b_i,w_i}(\Xi^T)\,.$$

To see this, note that for all the history which $(\mathcal{B}^t_{\tau_U(u),i})^c$ holds, all the past recommendations of item i are a function of Ξ, b_i and w_i. Past recommendations of other items are also function of $\tau_I(\cdot)$ and Ξ.

Using this and the fact that i) $\mathcal{B}^t_{\tau_U(u),i}$ is a deterministic function of H_{t-1} and $\tau_U(.)$; ii) b_i and w_i are deterministic functions of $H_{t-1}, \tau_U(.)$ and preference matrix Ξ, one can show that

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, H_{t-1}, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi, \tau_U(\cdot), \tau_I(\cdot \setminus \{i\}), \tau_I(i) \in K_{b_i,w_i}(\Xi^T)] \\
&= \mathbb{P}[L_{u,i} = +1 | \Xi, \tau_U(\cdot), \tau_I(\cdot \setminus \{i\}), \tau_I(i) \in K_{b_i,w_i}(\Xi^T)] \\
&\overset{(a)}{=} \mathbb{P}[\tau_I(i) \in K_{b_i^+,\{w_i,\tau_U(u)\}}(\Xi^T) | \Xi, \tau_U(\cdot), \tau_I(\cdot \setminus \{i\}), \tau_I(i) \in K_{b_i,w_i}(\Xi^T)] \\
&\leq \frac{k_{b_i^+,\{w_i,\tau_U(u)\}}(\Xi^T)}{k_{b_i,w_i}(\Xi^T)}
\end{aligned}$$

where in (a) we are using the fact that condition on the event $(\mathcal{B}^t_{\tau_U(u),i})^c$, it is guaranteed that $\tau_U(u) \notin w_i$.

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c_i^t < \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}] \\
&\qquad \leq \mathbb{E}\left[\frac{k_{b_i^+,\{w_i,\tau_U(u)\}}(\Xi^T)}{k_{b_i,w_i}(\Xi^T)} | \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}\right] \leq \frac{1}{2}\frac{1+\epsilon}{1-\epsilon} \leq \frac{1+4\epsilon}{2}
\end{aligned}$$

where we used the definition of row regularity and the assumption $\epsilon < 1/2$. $\square$

Corollary 10.23. *For* A2 *defined below, the following upper bound holds:*

$$\begin{aligned}\texttt{A2} &:= \mathbb{P}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}] \\ &\leq \frac{1+4\epsilon}{2}\,\mathbb{P}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}]\,.\end{aligned}$$

Proof. Multiplying the statement of Claim 10.22 by $\mathbb{P}[a_{u,t} = i, c^t_i < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c, \Xi^T \in \Omega_{\mathsf{r}_I,\epsilon}]$ and summing over i gives the statement of the corollary. $\square$

Definition 10.24. Let $\mathcal{B}^t_{u,\tau_I(i)}$ be the event that user u has rated an item of the same item type as item i by time $t-1$, i.e.,

$$\mathcal{B}^t_{u,\tau_I(i)} = \{\exists i' \in \mathbb{N} : a_{u,s} = i' \text{ for some } s < t \text{ with } \tau_I(i) = \tau_I(i')\}\,.$$

Claim 10.25. *If user u has not rated any item with the same type as i by time $t-1$, and user u has rated less than r_U distinct item types by time t, then the outcome of recommending i to u at time t is uncertain:*

$$\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c^t_i \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \leq \frac{1+4\epsilon}{2}$$

.

Proof. The proof is very similar to the proof of Claim 10.22. Conditional on $\Xi, \tau_I(\cdot), \tau_U(\cdot, \setminus\{u\})$ and H_{t-1}, define w_u, b_u and $K_{b_u,w_u}(\Xi)$. Note that if event $(\mathcal{B}^t_{u,\tau_I(i)})^c$ holds, then $\tau_I(i) \notin w_u$. We will show that

$$L_{u,i} \perp\!\!\!\perp H_{t-1} \big| (\mathcal{B}^t_{u,\tau_I(i)})^c, \Xi, \tau_U(\cdot \setminus \{u\}), \tau_I(\cdot), \tau_I(i) \in K_{b_u,w_u}(\Xi)\,.$$

To see this, note that for all the history which $(\mathcal{B}^t_{u,\tau_I(i)})^c$ holds, the outcome of all the past recommendations to user u are a function of $\Xi, \tau_I(\cdot)$, b_u and w_u. Past

recommendations to other users are also function $\tau_U(\cdot \setminus \{u\}), \tau_I(\cdot)$ and Ξ.

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 \,\big|\, a_{u,t} = i, H_{t-1}, (\mathcal{B}^t_{u,\tau_I(i)})^c, \Xi, \tau_U(\cdot \setminus \{u\}), \tau_I(\cdot), \tau_I(i) \in K_{b_u,w_u}(\Xi)] \\
&\qquad = \mathbb{P}[L_{u,i} = +1 \,\big|\, \Xi, \tau_U(\cdot \setminus \{u\}), \tau_I(\cdot), \tau_I(i) \in K_{b_u,w_u}(\Xi)] \\
&\qquad \overset{(a)}{=} \mathbb{P}[\tau_U(u) \in K_{b_u^+,\{w_u,\tau_I(i)\}}(\Xi) | \Xi, \tau_I(\cdot), \tau_U(\cdot \setminus \{u\}), \tau_U(u) \in K_{b_u,w_u}(\Xi)] \\
&\qquad \leq \frac{k_{b_u^+,\{w_u,\tau_I(i)\}}(\Xi)}{k_{b_u,w_u}(\Xi)}
\end{aligned}$$

where in (a) we are using the fact that condition on the event $(\mathcal{B}^t_{u,\tau_I(i)})^c$, it is guaranteed that $\tau_I(i) \notin w_u$.

$$\begin{aligned}
\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c_i^t \geq \mathsf{r}_I, d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \\
\leq \mathbb{E}\left[\frac{k_{b_u^+,\{w_u,\tau_I(i)\}}(\Xi)}{k_{b_u,w_u}(\Xi)} | \Xi \in \Omega_{\mathsf{r}_U,\epsilon}\right] \\
\leq \frac{1}{2}\frac{1+\epsilon}{1-\epsilon} \leq \frac{1+4\epsilon}{2},
\end{aligned}$$

where we used the definition of column regularity and the assumption $\epsilon < 1/2$. □

Corollary 10.26. *For* A3 *defined below, the following upper bound holds:*

$$\begin{aligned}
\mathsf{A3} &:= \mathbb{P}[L_{u,a_{u,t}} = +1, c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}] \\
&\leq \frac{1+4\epsilon}{2}\mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}].
\end{aligned}$$

Proof. Multiplying the statement of Claim 10.25 by $\mathbb{P}[a_{u,t} = i, c_i^t \geq \mathsf{r}_I, d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c, \Xi \in \Omega_{\mathsf{r}_U,\epsilon}]$ and summing over i gives the statement of the corollary. □

Definition 10.27. Let $\mathcal{B}^t_{\tau_U(u),\tau_I(i)}$ be the event that a user with type $\tau_U(u)$ has rated an item with type $\tau_I(i)$ by time $t-1$, i.e.,

$$\mathcal{B}^t_{\tau_U(u),\tau_I(i)} = \{\exists i' \in \mathbb{N}, u' \in [N] : a_{u',s} = i' \text{ for some } s < t \text{ with } \tau_I(i') = \tau_I(i), \tau_U(u') = \tau_U(u)\}.$$

Claim 10.28. *Recommending item i to user u gives uncertain outcome if no user*

with the same type as u has rated any item with the same type as i:

$$\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c_i^t \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c] = \frac{1}{2}$$

Proof. We can show that

$$L_{u,i} \perp\!\!\!\perp H_{t-1} \big| (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c, \Xi_{\cdot \setminus \{\tau_U(u),\tau_I(i)\}}, \tau_U(\cdot), \tau_I(\cdot)\,.$$

The above property holds because for all history H_{t-1} such that $(\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c$ holds, the outcome of all the past recommendations are function of $\Xi_{\cdot \setminus \{\tau_U(u),\tau_I(i)\}}$ and $\tau_U(\cdot)$. Also, condition on $\tau_U(\cdot)$ and $\tau_I(\cdot)$, we have $L_{u,i} = \xi_{\tau_U(u),\tau_I(i)}$. Hence

$$\begin{aligned}
&\mathbb{P}[L_{u,i} = +1 \,\big|\, a_{u,t} = i, H_{t-1}, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c, \Xi_{\cdot \setminus \{\tau_U(u),\tau_I(i)\}}, \tau_U(\cdot), \tau_I(\cdot)] \\
&\qquad = \mathbb{P}[\xi_{\tau_U(u),\tau_I(i)} = +1 \,\big|\, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c, \Xi_{\cdot \setminus \{\tau_U(u),\tau_I(i)\}}, \tau_U(\cdot), \tau_I(\cdot)] \\
&\qquad = \frac{1}{2}
\end{aligned}$$

Hence, using tower property of expectation,

$$\mathbb{P}[L_{u,i} = +1 | a_{u,t} = i, c_i^t \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c] = \frac{1}{2}$$

□

Corollary 10.29. *For* `A4` *defined below, the following upper bound holds:*

$$\begin{aligned}
\texttt{A4} := \mathbb{P}[L_{u,a_{u,t}} = +1, &c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c] \\
&= \frac{1}{2}\,\mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c]\,.
\end{aligned}$$

Proof. Multiplying the statement of Claim 10.28 by $\mathbb{P}[a_{u,t} = i, c_i^t \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c]$ and summing over i gives the statement of the corollary. □

Let

$$\mathsf{B1}(u,t) := \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U] \tag{10.49}$$

$$\mathsf{B2}(u,t) := \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c] \tag{10.50}$$

$$\mathsf{B3}(u,t) := \mathbb{1}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] \tag{10.51}$$

$$\mathsf{B4}(u,t) := \mathbb{1}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c]\,. \tag{10.52}$$

Next, define the set of bad recommendations up until time T to be:

$$\mathsf{bad}(T) := \sum_{t=1}^{T}\sum_{u=1}^{N} \mathbb{E}\left[\mathsf{B1}(u,t) + \mathsf{B2}(u,t) + \mathsf{B3}(u,t) + \mathsf{B4}(u,t)\right]\,. \tag{10.53}$$

Lemma 10.30. *For the set of bad recommendations defined in Equation* (10.53)*, the following lower bound for regret holds:*

$$N\mathtt{regret}(T) \geq \frac{1-4\epsilon}{2}\mathsf{bad}(T) - 4\,.$$

Proof. We decompose the set of recommendations made up until time T according to: i) the preference matrix Ξ is $(\mathsf{r}_I, \epsilon)-$row regular or not; ii) the preference matrix Ξ is $(\mathsf{r}_U, \epsilon)-$column regular or not; iii) the user been has recommended items from more than r_U distinct item types; iv) The recommended item has been recommended to users from more than r_I distinct user types or not; v) user u or users with similar type has rated item i or any item with the similar item type before.

We use the Definitions 10.17, 10.18, 10.21,10.24 and 10.27.

$$
\begin{aligned}
N(T - \texttt{regret}(T)) &= \sum_{t=1}^{T}\sum_{u=1}^{N} \mathbb{1}[L_{u,a_{u,t}} = +1] \\
&\leq \texttt{A1} + \texttt{A2} + \texttt{A3} + \texttt{A4} \quad + 2\mathbb{P}[\Xi \notin \Omega_{\mathsf{r}_U,\epsilon}] + 2\mathbb{P}[\Xi^T \notin \Omega_{\mathsf{r}_I,\epsilon}] \\
&\quad + \mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, \mathcal{B}^t_{u,\tau_I(i)}] + \mathbb{P}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, \mathcal{B}^t_{\tau_U(u),i}] \\
&\quad + \mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, \mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})}] \\
&\overset{\text{(a)}}{\leq} \frac{1+4\epsilon}{2}\,\mathbb{P}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U] \\
&\quad + \frac{1+4\epsilon}{2}\,\mathbb{P}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c] \\
&\quad + \frac{1+4\epsilon}{2}\mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] \\
&\quad + \frac{1}{2}\,\mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c] \\
&\quad + \frac{4}{N} + \mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, \mathcal{B}^t_{u,\tau_I(i)}] \\
&\quad + \mathbb{P}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, \mathcal{B}^t_{\tau_U(u),i}] \\
&\quad + \mathbb{P}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, \mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})}] \\
&\leq NT - \frac{1-4\epsilon}{2}\mathsf{bad}(T) + \frac{4}{N} \qquad (10.54)
\end{aligned}
$$

where in inequality (a) we used the statement of Claims ,10.19, 10.22, 10.25 and 10.28.

Using statement of Lemma 8.7, $\mathbb{P}[\Xi \notin \Omega_{\mathsf{r}_U,\epsilon}] < \frac{1}{N}$ and $\mathbb{P}[\Xi^T \notin \Omega_{\mathsf{r}_I,\epsilon}] < \frac{1}{N}$ with the choice of parameters given in Equation (10.48). □

Next, we will represent find several lower bounds for the total number of bad recommendations defined in Equation (10.53) as a function of parameters of the algorithm. Later, we will show that for any choice of these parameters, there is some lower bound for $\mathsf{bad}(T)$.

Definition 10.31. Let $\mathcal{F}_T$ be the items that have been rated by more than r_I user types by time T and

$$f_T := |\mathcal{F}_T| = \sum_{i\in\mathbb{N}} \mathbb{1}[c_i^T \geq \mathsf{r}_I]. \qquad (10.55)$$

Definition 10.32. Let $\mathcal{G}_T$ to be the items that have been rated by less than r_I user types by time T and

$$g_T := |\mathcal{G}_T| = \sum_{i \in \mathbb{N}} \mathbb{1}[0 < c_i^T < \mathsf{r}_I]. \tag{10.56}$$

Claim 10.33. *The expected number of bad recommendations defined in Equation* (10.53) *satisfies the following lower bound:*

$$\mathsf{bad}(T) \geq f_T(\mathsf{r}_I - 1) + g_T\,. \tag{10.57}$$

Proof. Using Equations (10.49) and (10.50):

$$\begin{aligned}
\mathsf{B1}(u,t) + \mathsf{B2}(u,t) &= \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U, \mathcal{B}^t_{\tau_U(u),a_{u,t}}] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c] \\
&\geq \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, (\mathcal{B}^t_{\tau_U(u),a_{u,t}})^c]\,.
\end{aligned}$$

According to Definition 10.31 the items in $\mathcal{F}_T$ are rated by at least r_I distinct user types by time T. So, each item $i \in \mathcal{F}_T$, there are recommended at least $\mathsf{r}_I - 1$ times in which $\mathbb{1}[c^t_i < \mathsf{r}_I, (\mathcal{B}^t_{\tau_U(u),i})^c]$ holds. The items in $\mathcal{G}_T$ are recommended to at least one user by time T. Hence, all items $i \in \mathcal{G}_T$ are recommended at least once when $\mathbb{1}[c^t_i < \mathsf{r}_I, (\mathcal{B}^t_{\tau_U(u),i})^c]$ holds. Hence,

$$\sum_{t=1}^{T} \sum_{u=1} \mathsf{B1}(u,t) + \mathsf{B2}(u,t) \geq f_T(\mathsf{r}_I - 1) + g_T\,.$$

□

Claim 10.34. *The expected number of bad recommendations defined in Equation* (10.53) *satisfies the following lower bound:*

$$\mathsf{bad}(T) \geq T(\mathsf{r}_I - 1)\,. \tag{10.58}$$

Proof. The total number of recommendations made to all the users by the end of

time T is TN. The items in $\mathcal{F}_T$ can be recommended at most N times. According to the definition given in Equation (10.56), the items in $\mathcal{G}_T$ are recommended at most r_I times. Hence,

$$TN \leq f_T N + g_T(\mathsf{r}_I - 1)\,.$$

Plugging this into Equation (10.57) gives

$$\mathsf{bad}(T) > T(\mathsf{r}_I - 1) + g_T(1 - \frac{(\mathsf{r}_I - 1)^2}{N}).$$

Note that for N large enough, $(\mathsf{r}_I - 1)^2 < N$ which gives the statement of the claim.

□

Definition 10.35. Let $\mathcal{W}_T \subseteq [N]$ be the set of users that have rated more than r_U distinct item types by time T. Let

$$w_T := \frac{1}{N}|\mathcal{W}_T| = \frac{1}{N}\sum_{u\in[N]} \mathbb{1}[d_u^T \geq \mathsf{r}_U]\,. \tag{10.59}$$

Note $u \notin \mathcal{W}_T$, for all $t \in [T]$, we have $d_u^t < \mathsf{r}_I$.

Claim 10.36. *The expected number of bad recommendations defined in Equation* (10.53) *satisfies the following lower bound:*

$$\mathsf{bad}(T) \geq w_T N(\mathsf{r}_U - 1)\,. \tag{10.60}$$

Proof. Using Equations (10.49) and (10.51):

$$\begin{aligned}
\mathsf{B1}(u,t) + \mathsf{B3}(u,t) &= \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U, \mathcal{B}^t_{u,\tau_I(a_{u,t})}] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(a_{u,t})})^c] \\
&\geq \mathbb{1}[d^t_u < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c]
\end{aligned} \tag{10.61}$$

According to Definition 10.35 the users in $\mathcal{W}_T$ have rated at least r_U distinct item types by time T. So, each user $u \in \mathcal{W}_T$, are recommended at least $\mathsf{r}_U - 1$ times in

which $\mathbb{1}[d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c]$ holds. The users in $[N] \setminus \mathcal{W}_T$ are recommended at least one item by time T. Hence, all users $u \in [N] \setminus \mathcal{W}_T$ are recommended at least once when $\mathbb{1}[d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(i),i})^c]$ holds. Hence,

$$\sum_{t=1}^{T} \sum_{u \in \mathcal{W}_T} \mathsf{B1}(u,t) + \mathsf{B3}(u,t) \geq w_T N(\mathsf{r}_U - 1)$$

□

Definition 10.37. Let Γ_u^T be the set of item types that are rated by user u by time T.

$$\Gamma_u^T = \{j \in [\mathsf{q}_I] : \sum_{t \in [T], i: \tau_I(i) = j} \mathbb{1}[a_{u,t} = i] \neq 0\}$$

Let $\gamma_u^T = |\Gamma_u^T|$.

Note that for according to the definition of $\mathcal{W}^T$

$$\forall u \notin \mathcal{W}_T, \quad \gamma_u^T < \mathsf{r}_U. \tag{10.62}$$

Definition 10.38. Let r_j^T be the number of items with item type j that have been rated by time T:

$$r_j^T = |\{i : \tau_I(i) = j, a_{u,t} = i, u \in [N], t \leq T\}|.$$

The total number of items recommended (ever seen by the algorithm) by time T is $f_T + g_T$. Also, for any user u, T items from $\mathcal{F}_T \cup \mathcal{G}_T$ with types in Γ_u^T (defined in Definition 10.37) are recommended by time T. Hence, $T \leq \sum_{j \in \Gamma_u^T} r_j^T$. This implies $\gamma_*^T := \min_{u \notin \mathcal{W}_T} \gamma_u^T \geq \frac{T}{\max_j r_{j,T}}$.

Note that the prior distribution of type of each item is uniform on $[\mathsf{q}_I]$ independently of other items. So, r_j^T is $\mathsf{Bin}(f_T + g_T, \frac{1}{\mathsf{q}_I})$. We provide two different deviation bounds for the random variable $\max_{j \in \mathsf{q}_I} r_j^T$ as a function of $f_T + g_T$:

- When $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$, using the Chernoff bound in Lemma A.3 with $\epsilon = 1$,

$$\mathbb{P}[\max_{j \in \mathsf{q}_I} r_j^T \geq 2\frac{f_T + g_T}{\mathsf{q}_I}] \leq \mathsf{q}_I \exp\left(-\frac{f_T + g_T}{3\mathsf{q}_I}\right) \leq \frac{1}{\mathsf{q}_I}, \tag{10.63}$$

  where the last inequality is due to the fact that $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$. This implies that

$$\mathbb{P}[\gamma_*^T \geq \frac{T\mathsf{q}_I}{2(f_T + g_T)}] \geq 1/2 \tag{10.64}$$

  and we always have $\gamma_*^T \geq 1$.

- When $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$, using the Chernoff bound in Lemma A.3 with

$$\epsilon = \frac{18\mathsf{q}_I \log \mathsf{q}_I}{f_T + g_T} - 1 > 2,$$

  gives

$$\mathbb{P}[\max_{j \in \mathsf{q}_I} r_j^T \geq 18 \log \mathsf{q}_I] \leq \mathsf{q}_I \exp\left(-(\frac{18\mathsf{q}_I \log \mathsf{q}_I}{f_T + g_T} - 1)^2 \frac{f_T + g_T}{2\mathsf{q}_I}\right) \leq \frac{1}{\mathsf{q}_I}, \tag{10.65}$$

  where the last inequality is due to the fact that $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$. This implies that

$$\mathbb{P}[\gamma_*^T \geq \frac{T}{18 \log \mathsf{q}_I}] \geq 1/2 \tag{10.66}$$

  and we always have $\gamma_*^T \geq 1$.

The above two cases show that

$$\mathbb{P}[\gamma_*^T \geq \min\{\frac{T}{36 \log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{6(f_T + g_T)}\}] \geq 1/2 \tag{10.67}$$

and we always have $\gamma_*^T \geq 1$.

Claim 10.39. *If $w_T \neq 1$, for $\delta = \mathsf{q}_I \exp(-\frac{T}{3\mathsf{q}_I})$ we have*

$$\mathsf{bad}(T) \geq (1 - \delta)\frac{T\mathsf{q}_I}{3\mathsf{r}_U} \tag{10.68}$$

Proof. Plugging Equation (10.64) into Equation (10.62), if $w_T \neq 1$, then $\gamma_*^T \leq \mathsf{r}_U$ and

$$g_T \geq \frac{T\mathsf{q}_I}{2\mathsf{r}_U} - f_T$$

Plugging this into Equation (10.57) shows that if $w_T \neq 0$, with probability $1-\delta$,

$$\sum_{t=1}^{T}\sum_{u=1} \mathsf{B1}(u,t) + \mathsf{B2}(u,t) \geq \frac{T\mathsf{q}_I}{2\mathsf{r}_U} + f_T(\mathsf{r}_I - 2) > \frac{T\mathsf{q}_I}{2\mathsf{r}_U}\,. \tag{10.69}$$

Using definition of $\mathsf{bad}(T)$ given in Equation (10.53) gives the statement of the claim.

□

Definition 10.40. Let Θ_k^T be the set of item types that are rated by any user of type k by time T.

$$\Theta_k^T = \{j \in [\mathsf{q}_I] : \sum_{\substack{t\in[T]\\ i:\tau_I(i)=j}} \mathbb{1}[\sum_{u:\tau_U(u)=k} a_{u,t} = i] \neq 0\}$$

Let $\theta_k^T = |\Theta_k^T|$.

Similarly, for any user type k, T items from $\mathcal{F}_T \cup \mathcal{G}_T$ with types in Θ_u^T (defined in Definition 10.40) are recommended by time T. Hence, $T \leq \sum_{j\in\Theta_k^T} r_j^T$. This implies $\theta_*^T := \min_k \theta_k^T \geq \frac{T}{\max_j r_{j,T}}$. Similar to the Equation (10.63) and (10.65):

- When $f_T + g_T \geq 6\mathsf{q}_I \log \mathsf{q}_I$,

$$\mathbb{P}[\theta_*^T \geq \frac{T\mathsf{q}_I}{2(f_T+g_T)}] \geq 1/2\,. \tag{10.70}$$

  we always have $\theta_*^T \geq 1$.

- When $f_T + g_T < 6\mathsf{q}_I \log \mathsf{q}_I$

$$\mathbb{P}[\theta_*^T \geq \frac{T}{18\log \mathsf{q}_I}] \geq 1/2\,. \tag{10.71}$$

  we always have $\theta_*^T \geq 1$.

The above two cases show that

$$\mathbb{P}[\theta_*^T \geq \min\{\frac{T}{36\log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{6(f_T+g_T)}\}] \geq 1/2 \tag{10.72}$$

and we always have $\theta_*^T \geq 1$.

Claim 10.41. *The number of bad recommendations up to time T satisfies the lower bound:*

$$\mathsf{bad}(T) \geq \frac{w_T \mathsf{q}_U}{4}\theta_*^T + (1-w_T)N\gamma_*^T\,. \tag{10.73}$$

Proof. As shown in Equation (10.61)

$$\mathsf{B1}(u,t) + \mathsf{B3}(u,t) \geq \mathbb{1}[d_u^t < \mathsf{r}_U, (\mathcal{B}^t_{u,\tau_I(i)})^c]$$

Taking the summation for $u \notin \mathcal{W}_T$ gives

$$\sum_{t=1}^{T}\sum_{u\notin\mathcal{W}_T} \mathsf{B1}(u,t) + \mathsf{B3}(u,t) \geq (1-w_T)N\gamma_*^T \tag{10.74}$$

Note that if no user from user type $\tau_U(u)$ has rated any item from item type $\tau_I(i)$, then we are sure that no user from user type $\tau_U(i)$ has rated item i: $\mathbb{1}[(\mathcal{B}^t_{\tau_U(u),i})^c] \geq \mathbb{1}[(\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c]$. We also know that $\mathbb{1}[d_u^t > \mathsf{r}_U] \geq \mathbb{1}[d_u^T > \mathsf{r}_U]$ Hence,

$$\begin{aligned}
\mathsf{B2}(u,t) + \mathsf{B4}(u,t) &= \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),i})^c] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^t \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c] \\
&\geq \mathbb{1}[c^t_{a_{u,t}} < \mathsf{r}_I, d_u^T \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(i)})^c] \\
&\quad + \mathbb{1}[c^t_{a_{u,t}} \geq \mathsf{r}_I, d_u^T \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c] \\
&\geq \mathbb{1}[d_u^T \geq \mathsf{r}_U, (\mathcal{B}^t_{\tau_U(u),\tau_I(a_{u,t})})^c]
\end{aligned}$$

Using the Chernoff bound, with probability $1 - \mathsf{q}_U \exp(-\frac{N}{3\mathsf{q}_U})$, the maximum number of users with any given user type is less than $2\frac{N}{\mathsf{q}_U}$. So, with probability $1 - \mathsf{q}_U \exp(-\frac{N}{3\mathsf{q}_U})$, there are at least $\frac{w_T N}{2\frac{N}{\mathsf{q}_U}} = w_T\mathsf{q}_U/2$ distinct user types among the users in $\mathcal{W}_T$. Counting the total number of item types any given user type with users

in $\mathcal{W}_T$ has rated (and given Definition 10.40), with probability $1 - \mathsf{q}_U \exp(-\frac{N}{3\mathsf{q}_U})$,

$$\sum_{t=1}^{T} \sum_{u \in \mathcal{W}_T} \mathsf{B2}(u,t) + \mathsf{B4}(u,t) \geq \frac{w_T \mathsf{q}_U \theta_*^T}{2} \,. \tag{10.75}$$

Merging this with Equation (10.74) gives

$$\mathsf{bad}(T) \geq [1 - \mathsf{q}_U \exp(-\frac{N}{3\mathsf{q}_U})] \frac{w_T \mathsf{q}_U \theta_*^T}{2} + (1 - w_T) N \gamma_*^T$$

We assume N is large enough so that $N > 3\mathsf{q}_U \log(2\mathsf{q}_U)$. which gives the statement of claim. □

Lemma 10.42. *For some constant $c > 0$, the expected number of bad recommendations defined in Equation* (10.53) *satisfies the following lower bound:*

$$\mathsf{bad}(T) \geq \begin{cases} c\sqrt{T\,\mathsf{q}_I N}, & \text{if } 3\,\mathsf{q}_I \log(2\,\mathsf{q}_I) \leq T \leq \frac{N \mathsf{r}_U^2}{\mathsf{q}_I} \\ c\sqrt{T\,\mathsf{q}_I\,\mathsf{q}_U}, & \text{if } \frac{N \mathsf{r}_U^2}{\mathsf{q}_I} \leq T \leq \frac{\mathsf{q}_I \mathsf{q}_U}{\mathsf{r}_U^2} \\ cT\mathsf{r}_I, & \text{if } \frac{\mathsf{q}_I \mathsf{q}_U}{\mathsf{r}_U^2} \leq T \,. \end{cases}$$

Proof. Equation (10.57) and plugging Equation (10.72) into Equation (10.73) gives:

$$2\mathsf{bad}(T) \geq \max\left\{ f_T + g_T \,,\ \frac{w_T \mathsf{q}_U}{4} \min\{\frac{T}{36 \log \mathsf{q}_I}, \frac{T\mathsf{q}_I}{6(f_T + g_T)}\} + (1 - w_T) N \gamma_*^T \right\}$$

If $T w_T \mathsf{q}_U \geq 144 \mathsf{q}_I (\log \mathsf{q}_I)^2$, then

$$\begin{aligned} 2\mathsf{bad}(T) &\geq \max\left\{ f_T + g_T \,,\ \frac{w_T \mathsf{q}_U}{4} \frac{T\mathsf{q}_I}{6(f_T + g_T)} + (1 - w_T) N \gamma_*^T \right\} \\ &\geq \sqrt{\frac{w_T \mathsf{q}_U \mathsf{q}_I}{8} T} + (1 - w_T) N \gamma_*^T \end{aligned}$$

otherwise,

$$2\mathsf{bad}(T) \geq \frac{w_T \mathsf{q}_U}{4} \frac{T}{36 \log \mathsf{q}_I} + (1 - w_T) N \gamma_*^T$$

Similarly, Equation (10.57) and plugging Equation (10.67) into above gives:

- If $Tw_T\mathsf{q}_U \geq 144\mathsf{q}_I(\log \mathsf{q}_I)^2$ and $T(1-w_T)N \geq 216\mathsf{q}_I(\log \mathsf{q}_I)^2$

$$2\mathsf{bad}(T) \geq \sqrt{\frac{w_T\mathsf{q}_U\mathsf{q}_I}{32}T} + \sqrt{\frac{(1-w_T)N\mathsf{q}_I}{6}T}$$

- If $Tw_T\mathsf{q}_U \geq 144\mathsf{q}_I(\log \mathsf{q}_I)^2$ and $T(1-w_T)N < 216\mathsf{q}_I(\log \mathsf{q}_I)^2$

$$2\mathsf{bad}(T) \geq \sqrt{\frac{w_T\mathsf{q}_U\mathsf{q}_I}{32}T} + \frac{(1-w_T)N}{36\log \mathsf{q}_I}T$$

- If $Tw_T\mathsf{q}_U < 144\mathsf{q}_I(\log \mathsf{q}_I)^2$ and $T(1-w_T)N \geq 216\mathsf{q}_I(\log \mathsf{q}_I)^2$

$$2\mathsf{bad}(T) \geq \frac{w_T\mathsf{q}_U}{4}\frac{T}{36\log \mathsf{q}_I} + \sqrt{\frac{(1-w_T)N\mathsf{q}_I}{6}T}$$

- If $Tw_T\mathsf{q}_U < 144\mathsf{q}_I(\log \mathsf{q}_I)^2$ and $T(1-w_T)N < 216\mathsf{q}_I(\log \mathsf{q}_I)^2$

$$2\mathsf{bad}(T) \geq \frac{w_T\mathsf{q}_U}{4}\frac{T}{36\log \mathsf{q}_I} + \frac{(1-w_T)N}{36\log \mathsf{q}_I}T$$

Some algebraic manipulations shows that for any choice of the parameter w_T by the algorithm, the following lower bound holds:

$$\mathsf{bad}(T) \geq c\begin{cases} \sqrt{T\mathsf{q}_I N}, & \text{if } T \leq \frac{N\mathsf{r}_U^2}{\mathsf{q}_I} \\ \sqrt{T\mathsf{q}_I\mathsf{q}_U}, & \text{if } \frac{N\mathsf{r}_U^2}{\mathsf{q}_I} \leq T \leq \frac{\mathsf{q}_I\mathsf{q}_U}{\mathsf{r}_U^2} \\ T\mathsf{r}_I, & \text{if } \frac{\mathsf{q}_I\mathsf{q}_U}{\mathsf{r}_U^2} \leq T \end{cases}$$

□

Chapter 11

Discussion

We analyzed the performance of online collaborative filtering in a latent variable model in which users of the same type rate all items similarly and items of the same type are rated by every user similarly. We studied two extreme regimes of interest corresponding to model only based on structure in user space (called user-user) and model based on only structure in item space (called item-item). We provided optimal algorithms and compared their performance in different regimes.

We showed that the user-user algorithm has a large cold-start time, but immediately afterwards it achieves its asymptotic regime, *i.e.,* all the recommendations for a period of time are essentially random, but then after learning the user types, the performance stays constant.

In comparison, the item-item algorithms can make meaningful recommendations very early on. But it takes much longer time to reach the asymptotic state. The asymptotic performance of item-item algorithm is also much better than the user-user case.

The joint algorithm proposed here has the benefits of both worlds. It starts off by making meaningful recommendations pretty soon and it takes much smaller time compared to item-item to reach the asymptotic performance. The rate of regret in the asymptotic regime of the joint algorithm is also equal to item item (which is the better of two models).

We provide information theoretic lower bounds for the performance of any algo-

rithm in all three regimes. To the best of our knowledge, this is one of the first body of works studying the optimality of online algorithms in the context of collaborative filtering problems.

Appendix A

Deviation bounds

Lemma A.1 (Hoeffding's inequality)**.** *Let $Z^{(1)}, \cdots, Z^{(n)}$ be n i.i.d. random variables taking values in the interval $[-1, 1]$ and let $S_n = \frac{1}{n}\left(Z^{(1)} + \cdots + Z^{(n)}\right)$. Then,*

$$\mathsf{P}\left[|S_n - \mathbb{E}Z^{(1)}| > t\right] \leq 2\exp(-nt^2/2)\,.$$

Lemma A.2 (Bernstein's inequality)**.** *Let $Z^{(1)}, ..., Z^{(n)}$ be i.i.d. random variables. Suppose that for all i, $|Z^{(i)} - \mathbb{E}[Z^{(i)}]| \leq M$ almost surely. Then, for all positive t,*

$$\mathsf{P}\left[\left|\sum_{i=1}^{n} Z^{(i)} - n\mathbb{E}[Z^{(1)}]\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2n\mathsf{VAR}(Z) + \frac{2}{3}Mt}\right)\,.$$

Lemma A.3 (Chernoff bound)**.** *Let $X_1, \cdots, X_n \in [0, 1]$ be independent random variables. Let $X = \sum X_i$ and $\bar{X} = \sum \mathbb{E}X_i$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(X \geq (1+\epsilon)\bar{X}) \leq \exp\left(-\frac{\epsilon^2}{2+\epsilon}\bar{X}\right) \leq \max\{\exp\left(-\frac{\epsilon^2}{3}\bar{X}\right), \exp\left(-\frac{\epsilon}{2}\bar{X}\right)\}$$

$$\mathbb{P}(X \leq (1-\epsilon)\bar{X}) \leq \exp\left(-\frac{\epsilon^2}{2}\bar{X}\right).$$

$$\mathbb{P}(|X - \bar{X}| \geq \epsilon\bar{X}) \leq 2\max\{\exp\left(-\frac{\epsilon^2}{3}\bar{X}\right), \exp\left(-\frac{\epsilon}{2}\bar{X}\right)\}$$

Lemma A.4 (Balls and bins)**.** *Given $m \leq n/3$, if we throw m balls into n bins independently uniformly at random, then probability that less than $m/2$ bins are nonempty is smaller than* $\exp(-m/2)$.

Proof. Let k be the number of nonempty bins. There are $\binom{n}{k}$ possible choices for nonempty bins. The probability that at most $k \leq m/2$ bins are nonempty is

$$\binom{n}{k}\left(\frac{k}{n}\right)^m \leq \left(\frac{n \cdot e}{k}\right)^k \left(\frac{k}{n}\right)^m \leq \left(\frac{k \cdot e}{n}\right)^{m/2} \leq \exp(-m/2)\,,$$

where we used $k \leq m/2$ and $k \leq n/6$. □

Bibliography

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

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

[3] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, pp. 1436–1462, 2006.

[4] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," in *APPROX*, 2008.

[5] M. Vuffray, S. Misra, A. Y. Lokhov, and M. Chertkov, "Interaction screening: Efficient and sample-optimal learning of ising models," *arXiv preprint arXiv:1605.07252*, 2016.

[6] P. Abbeel, D. Koller, and A. Ng, "Learning factor graphs in polynomial time and sample complexity," *JMLR*, 2006.

[7] P. Ravikumar, M. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[8] M. Meila and M. I. Jordan, "Learning with mixtures of trees," *The Journal of Machine Learning Research*, vol. 1, pp. 1–48, 2001.

[9] U. Heinemann and A. Globerson, "Inferning with high girth graphical models," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1260–1268, 2014.

[10] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. on Info. Theory*, vol. 58, no. 7, pp. 4117–4134, 2012.

[11] J. Bento and A. Montanari, "Which graphical models are difficult to learn?," in *NIPS*, 2009.

[12] R. Tandon, K. Shanmugam, P. K. Ravikumar, and A. G. Dimakis, "On the information theoretic limits of learning Ising models," in *Advances in Neural Information Processing Systems*, pp. 2303–2311, 2014.

[13] P. Rebeschini, R. Van Handel, *et al.*, "Can local particle filters beat the curse of dimensionality?," *The Annals of Applied Probability*, vol. 25, no. 5, pp. 2809–2866, 2015.

[14] H.-O. Georgii, *Gibbs measures and phase transitions*, vol. 9. Walter de Gruyter, 2011.

[15] S. Lauritzen, *Graphical models*. Oxford University Press, 1996.

[16] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[17] M. J. Wainwright, "Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching," *IN AISTATS*, 2003.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

[20] G. Bresler, "Efficiently learning ising models on arbitrary graphs," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 771–782, ACM, 2015.

[21] R. Wu, R. Srikant, and J. Ni, "Learning loosely connected Markov random fields," *Stochastic Systems*, vol. 3, no. 2, pp. 362–404, 2013.

[22] P.-L. Loh, M. J. Wainwright, *et al.*, "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses," *The Annals of Statistics*, vol. 41, no. 6, pp. 3022–3049, 2013.

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

[24] V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning high-dimensional Markov forest distributions: Analysis of error rates," *The Journal of Machine Learning Research*, vol. 12, pp. 1617–1653, 2011.

[25] H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman, "Forest density estimation," *The Journal of Machine Learning Research*, vol. 12, pp. 907–951, 2011.

[26] S. Dasgupta, "Learning polytrees," in *UAI*, 1999.

[27] N. Srebro, "Maximum likelihood bounded tree-width Markov networks," in *UAI*, 2001.

[28] M. Narasimhan and J. Bilmes, "PAC-learning bounded tree-width graphical models," *in Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference*, pp. 410–417, 2004.

[29] A. Anandkumar and R. Valluvan, "Learning loopy graphical models with latent variables: Efficient methods and guarantees," *The Annals of Statistics*, vol. 41, no. 2, pp. 401–435, 2013.

[30] A. Anandkumar, V. Tan, F. Huang, and A. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *Annals of Stat.*, vol. 40, no. 3, pp. 1346–1375, 2012.

[31] A. Anandkumar, F. Huang, D. Hsu, and S. Kakade, "Learning mixtures of tree graphical models," in *NIPS*, 2012.

[32] P. L. Erdös, M. A. Steel, L. Székely, and T. J. Warnow, "A few logs suffice to build (almost) all trees: Part ii," *Theoretical Computer Science*, vol. 221, no. 1, pp. 77–118, 1999.

[33] C. Daskalakis, M. Mossel, and S. Roch, "Phylogenies without branch bounds: Contracting the short, pruning the deep," *Research in Computational Molecular Biology: 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009. Proceedings*, pp. 451 – 465, 2009.

[34] E. Mossel, "Distorted metrics on trees and phylogenetic forests," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, 2007.

[35] R. Agarwala, V. Bafna, M. Farach, M. Paterson, and M. Thorup, "On the approximability of numerical taxonomy (fitting distances by tree metrics)," *SIAM Journal on Computing*, vol. 28, no. 3, pp. 1073–1085, 1998.

[36] A. Ambainis, R. Desper, M. Farach, and S. Kannan, "Nearly tight bounds on the learnability of evolution," in *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pp. 524–533, IEEE, 1997.

[37] M. Wainwright, "Estimating the "wrong" graphical model: Benefits in the computation-limited setting," *The Journal of Machine Learning Research*, vol. 7, pp. 1829–1857, 2006.

[38] T. H. Cormen, *Introduction to algorithms*. MIT press, 2009.

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

[40] V. Jog and P.-L. Loh, "On model misspecification and KL separation for Gaussian graphical models," in *Information Theory (ISIT), 2015 IEEE International Symposium on*, pp. 1174–1178, IEEE, 2015.

[41] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer, 2008.

[42] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[43] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, vol. 6. MIT press Cambridge, 2001.

[44] Y. Babichenko, S. Barman, and R. Peretz, "Empirical distribution of equilibrium play and its testing application," *arXiv preprint arXiv:1310.7654*, 2013.

[45] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[46] S. Aditya, O. Dabeer, and B. K. Dey, "A channel coding perspective of collaborative filtering," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2327–2341, 2011.

[47] G. Biau, B. Cadre, L. Rouviere, *et al.*, "Statistical analysis of k-nearest neighbor collaborative recommendation," *The Annals of Statistics*, vol. 38, no. 3, pp. 1568–1592, 2010.

[48] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[49] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[50] N. Srebro, N. Alon, and T. S. Jaakkola, "Generalization error bounds for collaborative prediction with low-rank matrices," in *Advances In Neural Information Processing Systems*, pp. 1321–1328, 2005.

[51] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[52] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, ACM, 2013.

[53] A. Rohde, A. B. Tsybakov, *et al.*, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[54] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[55] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," in *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2014.

[56] S. Bubeck, N. Cesa-Bianchi, *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[57] G. Bresler, G. H. Chen, and D. Shah, "A latent source model for online collaborative filtering," in *Advances in Neural Information Processing Systems*, pp. 3347–3355, 2014.

[58] A. Bellogin and J. Parapar, "Using graph partitioning techniques for neighbour selection in user-based collaborative filtering," in *Proceedings of the sixth ACM conference on Recommender systems*, pp. 213–216, ACM, 2012.

[59] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*, pp. 271–280, ACM, 2007.

[60] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[61] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, ACM, 2001.

[62] G. Bresler, D. Shah, and L. F. Voloch, "Collaborative filtering with low regret," *SIGMETRICS Perform. Eval. Rev.*, vol. 44, pp. 207–220, June 2016.

[63] J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying, "Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, pp. 29–41, ACM, 2014.

[64] B.-H. Shen, S. Ji, and J. Ye, "Mining discrete patterns via binary matrix factorization," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 757–766, ACM, 2009.

[65] J. Wang, A. P. De Vries, and M. J. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 501–508, ACM, 2006.

[66] D. Song, C. E. Lee, Y. Li, and D. Shah, "Blind regression: Nonparametric regression for latent variable models via collaborative filtering," in *Advances in Neural Information Processing Systems*, pp. 2155–2163, 2016.

[67] O. Dabeer, "Adaptive collaborating filtering: The low noise regime," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pp. 1197–1201, IEEE, 2013.

[68] K. Barman and O. Dabeer, "Analysis of a collaborative filter based on popularity amongst neighbors," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7110–7134, 2012.

[69] B.-H. Kim, A. Yedla, and H. D. Pfister, "Imp: A message-passing algorithm for matrix completion," in *Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on*, pp. 462–466, IEEE, 2010.