The London School of Economics and Political Science

Relational learning for set value functions

Yiliu Wang

A thesis submitted to the Department of Statistics of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, September 2022

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 45224 words.

Statement of co-authored work

I confirm that Chapter 3, 4 and 5 was jointly co-authored with Professor Milan Vojnovic. Chapter 6 was jointly co-authored with Professor Milan Vojnovic and Professor Wei Chen.

Abstract

Relational learning is learning in a context where we have a set of items with relationships. For example, in a recommender system or advertising platform, items are grouped into lists to attract user attention, and some items may be more popular than others. We are often interested in learning individual abilities, approximating group performances and making best set selection. However, it could be challenging as we have limited feedback and various uncertainties. We might only observe noisy aggregate feedback at the set level (set level randomness), and each item could be a random variable following some distributions (item level randomness). To tackle the problem, we model the group performance using a set value function, defined as a function of item values within the group of interest.

We first study the beta model for hypergraphs. The model treats relational data as hypergraphs where nodes represent items and hyper-edges group items into sets. The goal is to estimate individual beta values from the group outcomes. We study the inference problem under different settings using maximum likelihood estimation (MLE).

We move on to consider more general set value functions and the second source of randomness at the item level. The goal is to find good item representations (sketches) for approximation of stochastic valuation functions, defined as the expectation of set value functions of independent random variables. We present an approximation everywhere guarantee for a wide range of stochastic valuation functions.

Finally, we study an online variant where an agent can draw samples sequentially. At each time step, the agent chooses a group of items subject to constraints and receives some form of feedbacks. The goal is to select a set of items with maximum performances according to some stochastic valuation functions. We consider the regret minimization setting and address the problem under value-index feedback.

Acknowledgements

I would like to thank all my family members, mentors and friends that gave me enormous support these years. I am grateful to Professor Milan Vojnovic, my supervisor, for supporting and inspiring me throughout the doctoral programme. There are countless times that I emailed him after our weekly meetings when I encountered troubles or have new thoughts. And I remembered that there are countless times that Milan replied to my emails and inspired me of solutions. He helped me check technical correctness of proofs, and never ignored any vague points. This greatly helped me improve my writing skills and logical thinking. I would say I could not reach this milestone without his help.

I am also grateful to my husband, my parents and my grandparents for their love and support. This is especially important for my first year when I felt hopeless about completing the PhD, and the third year when I returned home due to the COVID. I definitely could not pass the upgrade and complete the thesis without their encouragement.

I would like to thank my office mates on the 5th and the 7th floor of our statistics department. I can always find someone to talk to when I felt tired of writing my thesis. I also would like to thank all the teachers and administrative staffs in our department. I benefited a lot from being a graduate teaching assistant (GTA) in our department.

Finally, I was lucky to be supported by the Lee Family scholarship. Thanks to the scholarship, I could focus on my research at this challenging time. It largely reduced my financial burden. My research will definitely be affected without the scholarship.

Contents

List of Figures

Chapter 1

Introduction

Relational data is widely used in our lives. Think of our daily routine. Nowadays, people can shop online for almost every aspects of their lives. For example, we can order meals on Deliveroo, buy groceries on Amazon, book taxis on Uber, work on online labour platforms such as UpWork, study on knowledge exchange platforms such as StackExchange, play various games such as League of Legends and etc. Many of us may be surprised by the intelligence of the online system. When we order deliveries, we see lists of goods attracting our interests. We can click individual pages to check details and select the one we prefer. In the case we don't find what we need, we can refresh the webpage and a new list will appear. A convenient and interesting feature is that the more often we use the platform, it knows better of our tastes. Another example is the knowledge exchange platform. When we want to find answers or help others, we see a list of questions for relevant topics. We can always find a thread of interest and study in an efficient way.

Relational learning is concerned with domain models that exhibit both uncertainty and complex, relational structure. It enables effective and robust reasoning about richly structured systems and data building on ideas from probability theory and statistics (Koller et al. (2007)). The online system is intelligent as it can learns from relational data. In the examples given above, the relational data comes from meal and grocery orders, bookings for taxis, information of online workers and players and etc.

1.1 Motivations and objectives

We elaborate more in details with three problems we are going to study in this thesis. We will relate the applications with models and briefly outline the research questions for each problem.

1.1.1 β-model for random hypergraphs

Many relational data can be represented by hypergraphs, where nodes represent items or individuals, while hyperedges, defined as subsets of nodes, represent relationships among entities. Graphs or networks are special cases of hypergraphs where entities are grouped pair-wisely. Extensive empirical research has been done on social and economic networks (Aral (2016); Breza (2016)).

Take the online labor platforms for example. We consider items to be experts and they are grouped into teams to work on projects. Clearly, some teams may have higher chances of success. In most cases, we are not able to directly measure individual abilities. Instead, we may only observe past project outcomes for a given collection of teams. We are interested in learning individual worker abilities that explain the group outcomes. Then we can use these individual scores to select workers for future projects.

To model group performance, we introduce the concept of set value functions, defined as function of item values within the group of interest. We start with statistical inference on a simple model of group performance called the beta model, which is well-known for graphs (Chatterjee et al. (2011)). We will study a generalized hypergraph variant. This is motivated by the fact that complete graph data might be expensive to collect and we may only have observations at the set level in real-world applications. The beta model assigns each node a strength parameter and models the hypergraph according to some rules.

We are interested in estimating the individual strength parameters, which naturally leads us to the framework of maximum likelihood estimation (MLE). We will derive conditions

for MLE existence and uniqueness, and give bounds for MLE accuracy. Importantly, we would like to link the MLE properties with graph-theoretic properties and see how the experimental design would affect our learning. Realized the importance of experimental design, we will further study the beta model under random design of experiments. This is motivated by real-life settings where we have limited resources for experiments and our designs may not be regular or complete. We are interested in MLE conditions and threshold number of experiments that guarantee the MLE accuracy.

1.1.2 Sketching for stochastic valuation functions

In some application scenarios, a simple model may not suffice for our task. We may need to take randomness in individual abilities into account. For example, in the case of online gaming platforms where items are players and the platform assigns players to teams for matches. In a competitive situation such as gaming, the performance of any individual is not deterministic and varies greatly according to personalities (Minka et al. (2018)). In particular, high-risk high-reward individuals may outperform stable-value individuals even if the later has higher expected value. This means assigning a simple score to each item may not be appropriate for such task. To tackle with this issue, we introduce the notion of stochastic valuation function, defined as the expectation of set value functions of independent random variables.

At the same time, we may need to consider more complex set value functions. The set outcome may not depend linearly on the individual item values. On the other hand, we may assume general properties such that the group performances grow with group size, but grows more and more slowly as the size increases due to coordination inefficiency.

Randomness and complexity of set function structures make it hard for many optimization problems. A natural way to solve this is to use approximations. In our second problem, we consider general set value functions and try to understand how can we approximate such functions everywhere using simple item representations in a computationally

efficient manner. We look at one approach called sketching (Cohavi and Dobzinski (2017)) to achieve this for a class of stochastic valuation functions.

Our problem is challenging as we would like to have computationally efficient algorithms that allows us to approximate a set valuation function everywhere, and controls over the sketch size at the same time. We will propose one such algorithm based on the concept of exponential binning, a method that discretizes item distribution into exponentially many histograms. We characterize the distribution of each item as histogram, then we store and compute set values based on these histograms. We will prove that the algorithm provides a constant-factor approximation for a wide range of stochastic valuation functions.

1.1.3 The k-max problem with value-index feedback

We also consider the case when data comes in streams. This happens in many real-world applications such as online shopping, advertising and question-answering platforms as mentioned above. Consider an advertising platform where users interact with the platform and their preferences are revealed by the choices they make through clicks. After receiving user feedback, the platform learns their preferences and updates the list of items presented to the users for better user experiences. It is important for the platform to accurately learn the user preferences and update the list of items that best suit the user interest. We call the problem k-max when we restrict the size of list to some constant k.

The problem is challenging since in most cases, we only observe the aggregate feedback for a list of items. For the online advertising example, we may observe the most popular item which receives the user click and rating. On the other hand, it is impossible to collect information on those items not selected by the user. Moreover, there are considerable uncertainties with the user click and item values. These all make it hard to select the best set of items.

The final part of the thesis is concerned with a class of online combinatorial optimization problem where an agent draws samples sequentially and receives the aggregate rewards

and index values as feedback. We call it the k-max problem with value-index feedback. Our goal is to maximize the expected cumulative reward over the time horizon. It is challenging as we observe limited feedback, and the aggregate reward is nonlinear in the individual rewards of constituent items. We will propose a new algorithm based on the combinatorial UCB-style algorithm (Slivkins et al. (2019)) to solve this problem and will show that the algorithm achieves satisfactory performances.

1.2 Summary of contributions

Our main contributions are summarized as follows.

Firstly, we study the maximum likelihood estimation (MLE) for the beta model of random hypergraphs under different settings. The beta model (Chatterjee et al. (2011)) assumes set level randomness such that a beta parameter is assigned to each item and the group outcome is modelled as a binary random variable with success probability according to a logistic function of the sum of item parameters in the group. In Chapter 3, we look at the setting of fixed general design of experiments, which allow for different number of experiments over candidate edges. We show easy-to-interpret conditions for the MLE existence and uniqueness. We provide bounds for the MLE error that crucially depend on the smallest eigenvalue of a signless Laplacian matrix, which corresponds to the correlation matrix of vertex-experiment incidence vectors. This eigenvalue is related to some parameters reflecting a graph non-bipartiteness, providing an intuitive interpretation of the algebraic conditions.

In Chapter 4, we further consider the beta model of random hypergraphs with random design matrices, defined by sampling candidate edges independently with replacement from the set of all combinations of k vertices from the set of n vertices. We present a sufficient and a necessary condition for a random design matrix to have full rank almost surely and give conjecture of a tight condition that empirically holds. This requires the number of edge experiments to be at least $ckn\log(n)$, for a fixed constant $c > 2$. We

also show a sufficient condition for the MLE existence and uniqueness to hold with high probability.

The second research paper is presented in Chapter 5, titled 'Sketching stochastic valuation functions'. There we consider the problem of sketching a stochastic valuation function. We show that for monotone subadditive or submodular valuation functions that satisfy a weak homogeneity condition, or certain other conditions, there exist discretized distributions of item values with $O(k\log(k))$ support sizes that yield a sketch valuation function which is a constant-factor approximation, for any value query for a set of items of cardinality less than or equal to k. The sketches are computed by using an algorithm based on the well-known concept of exponential binning. Besides being of interest in their own right, the obtained sketch results are of interest for finding approximate solutions for various optimization problems such as best set selection and welfare maximization problems.

Finally, in Chapter 6, we move to the online case where an agent can draw samples sequentially. At each time, the agent chooses a subset of k items and observes the maximum value and the item which takes the maximum value. We call it the k-max problem with value-index feedback. The goal is to select the set with maximum performances according to the expected max reward while minimizing the total regret. Our problem can be put into the general framework of combinatorial multi-armed bandits (Cesa-Bianchi and Lugosi (2012); Chen et al. (2013)), with a setting in the middle ground of semi-bandit and full-bandit. We propose two algorithms to solve the k-max problem, a UCB-style algorithm and a new algorithm that is modified based on the UCB-style algorithm. We show that the regret bound for UCB-style algorithm contains an undesirable factor. Our new algorithm removes the factor and achieves comparable regret bound as standard combinatorial multi-armed bandit problems.

Chapter 2

Background Theory

2.1 Convex analysis and regression

This section provides background materials for Chapter 3 and Chapter 4.

Basic concepts in convex analysis

For any convex function $f : \mathbb{R}^n \to \mathbb{R}$, the *level sets* of f are defined by

$$\text{lev}_\alpha f = \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \leq \alpha\}, \text{ for } \alpha \in \mathbb{R}.$$

The level sets $\text{lev}_\alpha f$ are closed and convex. The union of $\text{lev}_\alpha f$ for $\alpha \in \mathbb{R}$ is the *effective domain* of f, which is denoted as $\text{dom} f$. The level set for which $\alpha = \inf f$ is called the *minimum set* of f. Function f has a unique minimizer if its minimum set is a singleton set.

Given a non-empty set C, a vector $\boldsymbol{d}$ is a *direction of recession* if starting at any $\boldsymbol{x} \in C$ and going indefinitely along $\boldsymbol{d}$, we never cross the relative boundary of C to points outside C, i.e.

$$\boldsymbol{x} + \lambda \boldsymbol{d} \in C, \text{ for all } \boldsymbol{x} \in C \text{ and all } \lambda \geq 0.$$

The following two theorems are from Rockafellar (1997).

Theorem 2.1.1. *For any closed proper convex function f, the minimum set of f is a non-empty bounded set if, and only if, $int(dom f^*)$. This holds if, and only if, f has no direction of recession.*

Theorem 2.1.2. *If f is strictly convex on dom f, then the minimum set of f contains no more than one point.*

For a closed proper convex function f, the *recession cone* R_f of the non-empty level sets is called the recession cone of f. The *linearity space* of the recession cone R_f is denoted by L_f such that $L_f = R_f \cap (-R_f)$.

Equivalently, $d \in L_f$ if, and only if, both d and $-d$ are directions of recession of each of the non-empty level sets. This happens if, and only if, the entire line $\{\boldsymbol{x} + \lambda \boldsymbol{d} : \lambda \in \mathbb{R}\}$ is contained in the same level set that contains $\boldsymbol{x}$, for all $\boldsymbol{x} \in$ dom f. Thus, any $\boldsymbol{d} \in L_f$ is a direction in which f stays constant, and L_f is also called the *constance space* of f.

The following theorem is from Bertsekas (2009).

Theorem 2.1.3. *For any closed convex function f, the minimum set of f over* dom$f$ *is non-empty if $R_f = L_f$. Under this condition, the minimum set $X^\star$ of f can be expressed as $X^\star = \bar{X} + L_f$, where $\bar{X}$ is some non-empty and compact set.*

MLE for logistic regression models

The logistic regression model belongs to the exponential family of models.

It is well-known that for any exponential-family distribution, the negative log-likelihood function is convex. Strict convexity of the negative log-likelihood function ensures the existence and uniqueness of an optimal point over a compact convex set. However, in the unconstrained case, an MLE may not exist, and if it exists may not be unique. Therefore, it is not a sufficient condition to guarantee MLE existence over $\mathbb{R}^n$.

A necessary an sufficient condition for MLE existence and uniqueness is given by Barndorff-Nielsen (1978) for general exponential-family distributions.

Theorem 2.1.4. *For any exponential-family distribution with sufficient statistic t, the log-likelihood function has a unique maximum if, and only if, $t \in \text{int}(C)$ where C is the convex support of the exponential family distribution.*

This result follows from the theorems for general convex functions by Rockafellar (1997), in particular from the necessary and sufficient conditions for the minimum set of a convex function to be non-empty and bounded, and the strict convexity condition ensuring that the minimum set is a singleton.

An alternative formulation of the necessary and sufficient condition for the MLE existence and uniqueness uses the concept of overlap which was developed for logistic regression models. $(\mathbf{X}, \mathbf{y})$ is said to satisfy the *overlapping condition*, if there exists no $\boldsymbol{\alpha} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that for all $j \in [m]$,

$$\boldsymbol{x}_j^\top \boldsymbol{\alpha} \geq 0 \text{ if } y_j = 1 \text{ and } \boldsymbol{x}_j^\top \boldsymbol{\alpha} \leq 0 \text{ if } y_j = 0.$$

In other words, we cannot separate the two classes of points using a hyperplane passing through the origin. The sign of the half-spaces is not important as we can always exchange the two classes by replacing $\boldsymbol{\alpha}$ with $-\boldsymbol{\alpha}$.

2.2 Properties of set functions

This section provides background materials for Chapter 5.

Set value functions map item values within the set of interest to set outcomes. Mathematically, they are defined on subsets of $\mathbb{R}^n$, $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where each $\mathcal{X}_i$ is a compact subset of $\mathbb{R}$.

We review some known properties for this class of functions.

Convexity and concavity

A function f is said to be convex on $\mathcal{X}$ if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for all $x, y \in \mathcal{X}$ and $\lambda \in [0,1]$. Concavity requires the inequality to hold in the reverse direction.

A function f is said to be *coordinate-wise concave* if for every $x \in \mathcal{X}$, $i \in [n]$, and $u, v \in \mathbb{R}_+$ such that $x + ue_i \in \mathcal{X}$, $x + ve_i \in \mathcal{X}$, and $x + (u+v)e_i \in \mathcal{X}$, it holds

$$f(x + ue_i) - f(x) \geq f(x + (u+v)e_i) - f(x + ve_i). \tag{2.1}$$

If f is twice-differentiable, by Bian et al. (2017) the coordinate-wise concave property is equivalent to $\partial^2 f(x)/\partial x_i^2 \leq 0$, for all $x \in \mathcal{X}$ and $i \in [n]$. Hence, if f is twice-differentiable, the coordinate-wise concave property corresponds to the standard notion of concave functions holding for each coordinate.

There exist functions that are coordinate-wise concave but are not coordinate-wise concave according to the classical notion of concave functions. An example is the max value function $f(x) = \max\{x_1, \ldots, x_n\}$ for $n > 1$. Example of a function that is concave according to classical notion of concave functions is $f(x) = g(\sum_{i=1}^{n} x_i)$ where g is a concave function.

Submodularity

A function f is *submodular* if for every $x, y \in \mathcal{X}$,

$$f(x \wedge y) + f(x \vee y) \leq f(x) + f(y) \tag{2.2}$$

where $\wedge$ and $\vee$ denote the coordinate-wise minimum and maximum operations, respectively. This concept is an extension of the standard notion of a submodular set function to vectors. If $\mathcal{X} = \{0, 1\}$, then f is a submodular set function satisfying the well-known diminishing returns property. If $\mathcal{X} = \mathbb{Z}$, then f is said to be a *lattice submodular* function.

By Topkis (1978), if f is twice-differentiable on its domain, then f is submodular if, and only if, all off-diagonal elements of the Hessian matrix of f are nonpositive, i.e. $\partial^2 f(x)/\partial x_i \partial x_j \leq 0$, for all $i \neq j$, for every x in the domain of f. Submodular functions may be concave, convex, or neither.

Equivalent definitions An equivalent definition of a submodular function is as follows: a function f is submodular if for every $x \in \mathcal{X}$, two distinct basis vectors e_i, $e_j \in \mathbb{R}^n$, and two non-negative real numbers z_i and z_j such that $x + z_i e_i \in \mathcal{X}$ and $x + z_j e_j \in \mathcal{X}$,

$$f(x + z_i e_i) + f(x + z_j e_j) \geq f(x) + f(x + z_i e_i + z_j e_j) \tag{2.3}$$

A function f is said to satisfy *the weak DR (diminishing returns) property* if for every $x, y \in \mathcal{X}$ such that $x \leq y$, $i \in [n]$ such that $x_i = y_i$, $z \in \mathbb{R}_+$ such that $x + z e_i \in \mathcal{X}$ and $y + z e_i \in \mathcal{X}$,

$$f(x + z e_i) - f(x) \geq f(y + z e_i) - f(y) \tag{2.4}$$

where e_i is a standard basis vector. By Bian et al. (2017), a function f is submodular if, and only if, it satisfies the weak DR property.

DR-submodularity A subclass of submodular functions are DR (diminishing returns)-submodular functions (Bian et al. (2017); Soma and Yoshida (2015)). A function f is said to be *DR-submodular*, if for all $x, y \in \mathcal{X}$ such that $x \leq y$ and any e_i and a non-negative number z such that $x + z e_i \in \mathcal{X}$ and $y + z e_i \in \mathcal{X}$, the diminishing returns property (3.15) holds. By Bian et al. (2017), a function f is DR-submodular if, and only if, it is submodular and coordinate-wise concave.

Subadditivity

A function f is said to be *subadditive* if $f(x+y) \leq f(x) + f(y)$. A set function u is subadditive if $u(S \cup T) \leq u(S) + u(T)$, for every $S, T \subseteq \Omega$. Clearly, any non-negative submodular set function is subadditive.

We can show the following relationship between DR-submodular functions and subadditive functions from definition.

Lemma 2.2.1. *If a function f is DR-submodular on $\mathcal{X} \subseteq \mathbb{R}_+^n$, $0 \in \mathcal{X}$, and $f(0) \geq 0$, then f is subadditive on $\mathcal{X}$.*

Proof. For any $x, y \in \mathcal{X}$, we have

$$\begin{aligned}
f(x+y) - f(x) &= f\left(x + \sum_{i=1}^{n} y_i e_i\right) - f\left(x + \sum_{i=2}^{n} y_i e_i\right) \\
&\quad + f\left(x + \sum_{i=2}^{n} y_i e_i\right) - f\left(x + \sum_{i=3}^{n} y_i e_i\right) \\
&\quad \vdots \\
&\quad + f\left(x + e_n y_n\right) - f\left(x\right) \\
&\leq f\left(\sum_{i=1}^{n} y_i e_i\right) - f\left(\sum_{i=2}^{n} y_i e_i\right) \\
&\quad + f\left(\sum_{i=2}^{n} y_i e_i\right) - f\left(\sum_{i=3}^{n} y_i e_i\right) \\
&\quad \vdots \\
&\quad + f\left(y_n e_n\right) - f\left(0\right) \\
&= f(y) - f(0)
\end{aligned}$$

where the inequalities hold by the DR-submodular property. Combining with $f(0) \geq 0$, we have $f(x+y) - f(x) \leq f(y)$, which is equivalent to saying that f is subadditive on $\mathcal{X}$. □

However, such relationship does not hold for general submodular functions.

Not all submodular functions are subadditive. Consider the success-probability value function

$$f(x) = 1 - \prod_{i=1}^{n}(1 - p(x_i))$$

where $p : \mathbb{R} \to [0, 1]$ is an increasing function. This function is submodular. This can be verified by checking that it satisfies the weak DR property as follows. Consider any $x, y \in \mathbb{R}^n$ such that $x \leq y$. Since p is an increasing function, $f(x) \leq f(y)$. To check the weak DR property, consider adding z to the j-th basis direction to x and y such that $x \leq y$ and $x_j = y_j$. Then, the weak-DR condition is equivalent to

$$\prod_{i \neq j}(1 - p(x_i))(1 - p(x_j + z)) \geq \prod_{i \neq j}(1 - p(y_i))(1 - p(y_j + z))$$

which clearly holds since $x_j = y_j$ and $f(x) \leq f(y)$. However, for some choices of function p, function f is not subadditive. Consider, for example, the case when $n = 1$, then f is subadditive if, and only if, p is subadditive.

Extended diminishing returns

A function f is said to satisfy the *extended diminishing returns property* Sekar et al. (2021) if for any $i \in [n]$ and $v \geq 0$ that has a non-empty preimage under f, there exists $y \in \mathbb{R}^n_+$ with $y_i = 0$ such that (a) $f(y) = v$ and (b) $f(x + ze_i) - f(x) \geq f(y + ze_i) - f(y)$ for any $z \in \mathbb{R}$ and x such that $f(x) \leq f(y) = v$ and $x_i = 0$. A simpler but stronger property is that f is such that $f(x + ze_i) - f(x) \geq f(y + ze_i) - f(y)$ for every $z \in \mathbb{R}$ and x, y such that $f(x) \leq f(y)$ and $x_i = y_i = 0$.

Function f satisfies the extended diminishing returns property as shown in Sekar et al. (2021). There are functions that satisfy the extended diminishing returns property but that are not DR-submodular. Consider, for example, $f(x) = (\sum_{i=1}^{n} x_i^r)^{1/r}$, for $r > 1$. However, f is not DR-submodular. To see this note that f is twice-differentiable and is a convex function, hence it is coordinate-wise convex according to standard notion of convex functions. On the other hand, twice-differentiable DR-submodular functions are

coordinate-wise concave according to the standard notion of concave functions.

Weakly homogeneity

A function f is homogeneous of degree d over a set $\Theta \subseteq \mathbb{R}$, if $f(\theta x) = \theta^d f(x)$ for all x in the domain of f, and all $\theta \in \Theta$.

Weakly homogeneity is a relaxed notion of homogeneity: we say that a function f is *weakly homogeneous of degree d and tolerance η over a set $\Theta \subseteq \mathbb{R}$* if

$$(1/\eta)\, \theta f(x) \leq f(\theta x) \leq \theta^d f(x)$$

for every x in the domain of f and all $\theta \in \Theta$.

Weakly homogeneous with constant degree and tolerance 1. Clearly, any homogeneous function f of degree 1 over Θ is weakly homogeneous of degree 1 and tolerance $\eta = 1$ over Θ. For example, $f(x) = \max\{x_1, \ldots, x_n\}$ and $f(x) = (\sum_{i=1}^n x_i^r)^{1/r}$ are homogeneous functions of degree 1 over $\mathbb{R}$. Note that any function that is convex on a domain that includes 0 and is such that $f(0) \leq 0$ is weakly homogeneous of degree 1 over $[0, 1]$. Some concave functions are weakly homogeneous with a strictly positive degree. For example, $f(x) = (\sum_{i=1}^n x_i)^r$ with domain $\mathbb{R}_+^n$, for $r \in (0, 1]$, is weakly homogeneous of degree r over $\mathbb{R}_+$. A differentiable function f is weakly homogeneous of degree d over $[0, 1]$ if, and only if,

$$x^\top \nabla f(x) \geq d\, f(x) \text{ for every } x \in \text{dom}(f). \tag{2.5}$$

For example, consider $f(x) = g(\sum_{i=1}^n x_i)$ where g is an increasing, differentiable and concave function on $\mathbb{R}_+$. Then, the inequality in (2.5) is equivalent to $\eta(z) \geq d$ for all $z \in \mathbb{R}_+$, where $\eta(z)$ is the *elasticity* of function g, defined as $\eta(z) = zg'(z)/g(z)$, which is always less than or equal to 1 for any increasing, differentiable and concave function g.

Function g has a constant elasticity r if, and only if, $g(z) = cz^r$ for an arbitrary constant $c > 0$. Some concave functions have zero minimum elasticity, e.g. $g(z) = 1 - e^{-\lambda z}$, for parameter $\lambda > 0$, has decreasing elasticity from value 1 at $z = 0$ to value 0 as z goes to infinity.

Weakly homogeneous with constant tolerance. Many functions are weakly homogeneous over $[0, 1]$ with a constant tolerance parameter η, which we discuss next.

Any monotone subadditive function $f : \mathbb{R}^n \to \mathbb{R}_+$ is weakly homogeneous over $[0, 1]$ with tolerance $\eta = 2$. To see this, we note that if f is a monotone, subadditive function, with $\mathcal{X} \subseteq \mathbb{R}^n_+$, then for every $\lambda \in (0, 1]$, and $x \in \mathcal{X}$,

$$f(x) \leq \lceil 1/\lambda \rceil \, f(\lambda x). \tag{2.6}$$

Therefore, for any monotone subadditive function $f(\theta x) \geq (1/\lceil 1/\theta \rceil) f(x)$. Note that $1/\lceil 1/\theta \rceil \geq 1/(1/\theta + 1) \geq \theta/2$. This implies $(1/2)\lambda f(x) \leq f(\lambda x)$, which may be interpreted as a weak homogeneity condition.

Any function f that is subadditive and convex on a domain that includes 0, and is such that $f(0) = 0$, is weakly homogeneous over $[0, 1]$ with tolerance $\eta = 1$. If f is a subadditive and convex function on a domain that includes 0 and $f(0) \leq 0$, then it is weakly homogeneous with tolerance $\eta = 1$. This follows from

$$\begin{aligned} f(\theta x) &\geq f(x) - f((1-\theta)x) \\ &\geq f(x) - (1-\theta) f(x) \\ &= \theta f(x) \end{aligned}$$

where the first inequality is by subadditivity and the second inequality is by convexity.

Finally, note that any concave function on a domain that includes 0 such that $f(0) \geq 0$ is weakly homogeneous with tolerance $\eta = 1$. This follows straightforwardly from the

definition of concave functions.

Properties for stochastic valuation functions

In the following chapters, we will mainly work with stochastic valuation functions, defined as expectation of set value functions of independent random variables. The following is a known relation between a set value function f and the stochastic valuation function u s.t. $u(S) = \mathbb{E}[f((X_i, i \in S))]$, where $X_1, \ldots, X_n$ are some independent random variables.

Lemma 2.2.2 (Lemma 3 Asadpour and Nazerzadeh (2016))**.** *Assume that f is a monotone submodular function, then u is a monotone submodular set function.*

We can generalize this relationship to subadditive functions.

Lemma 2.2.3. *Assume that f is a monotone function that is either subadditive or submodular, then u is a monotone subadditive set function.*

Proof. If f is a monotone submodular function, then by Lemma 2.2.2, u is a monotone submodular set function, hence, it is a monotone subadditive function. Consider now the case when f is a monotone subadditive function. For any $S, T \subseteq \Omega$,

$$\begin{aligned} u(S) + u(T) &= \mathbb{E}[f((X_i, i \in S))] + \mathbb{E}[f((X_i, i \in T))] \\ &= \mathbb{E}\left[f\left(\sum_{i \in S} X_i e_i\right)\right] + \mathbb{E}\left[f\left(\sum_{i \in T} X_i e_i\right)\right]. \end{aligned}$$

By monotonicity and subadditivity of f, for every x in the domain of f, we have

$$\begin{aligned} f\left(\sum_{i \in S} x_i e_i\right) + f\left(\sum_{i \in T} x_i e_i\right) &\geq f\left(\sum_{i \in S} x_i e_i + \sum_{i \in T} x_i e_i\right) \\ &= f\left(\sum_{i \in S \cup T} x_i e_i + \sum_{i \in S \cap T} x_i e_i\right) \\ &\geq f\left(\sum_{i \in S \cup T} x_i e_i\right). \end{aligned}$$

Thus, it follows

$$u(S) + u(T) \geq \mathbb{E}[f((X_i, i \in S \cup T))] = u(S \cup T).$$

□

2.3 Basic inequalities

Concentration inequalities

In this thesis, we will use the following well-known tail bounds for our analysis.

Lemma 2.3.1 (Hoeffding's Inequality Hoeffding (1994))**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with common support $[0, 1]$ and mean μ. Let $Y = \sum_{j=1}^{n} \boldsymbol{X}_j$. Then for all $\delta > 0$,*

$$\Pr[|Y - n\mu| \geq \delta] \leq 2e^{-2\delta^2/n}.$$

Lemma 2.3.2 (Multiplicative Chernoff bound Mitzenmacher and Upfal (2017))**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables taking values in $\{0, 1\}$ with mean μ. Let $Y = \sum_{j=1}^{n} \boldsymbol{X}_j$. Then for all $\delta > 0$,*

$$\Pr[Y \leq (1 - \delta)n\mu] \leq e^{-\delta^2 n\mu/2}.$$

We will treat these two lemmas as facts. The following lemma provides a vector version of Azuma-Hoeffding probability of deviation bound, which is from Hayes (2005).

Lemma 2.3.3. *Let $\boldsymbol{S}_m = \sum_{j=1}^{m} \boldsymbol{X}_j$ be a martingale where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are random variables taking values in $\mathbb{R}^n$ and satisfying $\mathbb{E}[\boldsymbol{X}_j] = \mathbf{0}$ and $||\boldsymbol{X}_j|| \leq \sigma$, for $\sigma > 0$. Then, for every $x \geq 0$,*

$$\Pr[||\boldsymbol{S}_m|| \geq x] \leq 2e^2 e^{-\frac{x^2}{m\sigma^2}}.$$

The next lemma states a matrix version of Chernoff type bound for the smallest eigenvalue of certain random matrices. The lemma follows from a more general result in Theorem 5.1.1 in Tropp (2015).

Lemma 2.3.4. *Let $\boldsymbol{S}_m = \sum_{j=1}^{m} \boldsymbol{X}_j$ where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are random, independent real symmetric matrices in $\mathbb{R}^{n\times n}$ such that $\lambda_1(\boldsymbol{X}_j) \geq 0$ and $||\boldsymbol{X}_j||_2 \leq \sigma$ for all $j \in [m]$. Then, for every $\epsilon \in (0, 1]$, we have*

$$\Pr[\lambda_1(\boldsymbol{S}_m) \leq (1-\epsilon)\lambda_1(\mathbb{E}[\boldsymbol{S}_m])] \leq n e^{-\frac{\epsilon^2 \lambda_1(\mathbb{E}[\boldsymbol{S}_m])}{2\sigma}}.$$

where $\lambda_1(\boldsymbol{M})$ denotes the smallest eigenvalue of a square real symmetric matrix $\boldsymbol{M}$.

We also introduce the famous Cauchy-Schwartz inequality and its applications.

Lemma 2.3.5. *For all $\boldsymbol{x}$ and $\boldsymbol{y}$ of an inner product space,*

$$|\boldsymbol{x}||\boldsymbol{y}| \geq \boldsymbol{x}^\intercal \boldsymbol{y}$$

The Cauchy-Schwartz inequality is a special case of Hölder's inequality with $p = q = 2$. The following lemma by Costello and Vu (2008) is an application of the Cauchy-Schwartz inequality.

Lemma 2.3.6. *Let X and Y be random variables, and let $E(X, Y)$ be an event depending on X and Y. Let X' be an independent copy of X. Then*

$$\Pr(E(X, Y)) \leq (\Pr(E(X, Y) \wedge E(X', Y)))^{1/2}$$

Negative association

As we can see from the previous section, independent random variables allow many powerful theorems to apply. However, in real life examples, we cannot always expect random variables we observe to be independent. Nonetheless, these random variables

may satisfy other special dependence properties. In this thesis, we will need one special dependence structure called negative association.

Definition 2.3.7. *A collection of random variables $Y = (Y_1, Y_2, \ldots, Y_n)$ is said to be negatively associated if for disjoint index sets I, $J \subseteq [n]$ and two functions f and g both monotone increasing or both monotone decreasing,*

$$\text{Cov}\left(f(X_i, i \in I), g(X_j, j \in J)\right) \leq 0$$

Negative association allows many useful properties of independence to carry over. Next, we list some of the useful properties that will be needed for the thesis. Proofs and more discussions can be found in the original paper by Joag-Dev and Proschan (1983).

The following properties are called closure of negative association (NA). It allows the NA property to be transferred to another set of random variables without calculation from definition. Joag-Dev and Proschan (1983) showed that the second property is unique to NA among a wide range of negative correlation structures.

Lemma 2.3.8. *The union of independent sets of NA random variables is negatively associated.*

Lemma 2.3.9. *Concordant monotone functions defined on disjoint subsets of a set of NA random variables are negatively associated.*

Many standard distributions possess the NA property. In particular, we point out the class of permutation distributions.

Lemma 2.3.10. *Let $x_1 \leq x_2 \leq \ldots \leq x_n$ and $X_1, X_2, \ldots, X_n$ be random variables such that $\{X_1, X_2, \ldots, X_n\} = \{x_1, x_2, \ldots, x_n\}$ always, with all possible assignments equally likely. Then $X_1, X_2, \ldots, X_n$ are NA.*

Chapter 3

The β-model for hypergraphs

3.1 Overview

Let G be an undirected simple graph on n vertices and $d_1, \ldots, d_n$ be the degrees of vertices of G. The study of degree distributions of networks is a classical topic in network analysis. The surveys (Newman (2003); Goldenberg et al. (2010)) contain many references for existing studies. The β-model for random graphs, originally introduced by Chatterjee et al. (2011), is the simplest instance of statistical network model that based exclusively on node degrees. Each vertex v is associated with a parameter β_v and edge (u, v) is present with probability $p_{u,v}(\boldsymbol{\beta}) = \sigma(\beta_u + \beta_v)$, where σ is the logistic function, independently of other edges. This model is an undirected version of the p_1 directed exponential random model originally proposed by Holland and Leinhardt (1981), and can be seen as a generalization of the classic random graph model by Erdös and Rényi (1959).

In real-world applications, collecting complete network data is often expensive and time-consuming. As discussed in the introduction chapter, in some cases we may only have limited observations at the set level. In this chapter, we study the β-model of random hypergraphs, defined for a given number $n \geq 2$ of vertices and parameter vector $\boldsymbol{\beta} \in \mathbb{R}^n$ such that edge $S \subseteq V := \{1, \ldots, n\}$ is present, independently of other edges, with

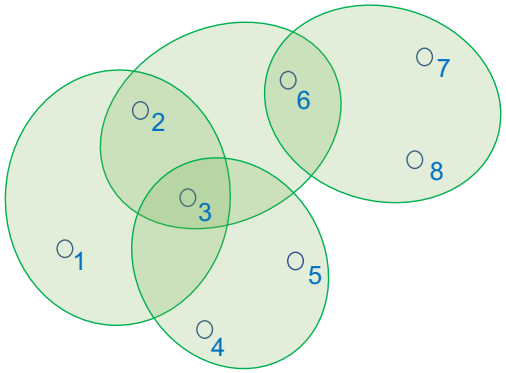

Figure 3.1: An example of partial design of experiments for a 3-uniform random hypergraph: not all combinations of 3 vertices are experimented but only those indicated in the figure.

probability

$$p_S(\boldsymbol{\beta}) := \sigma\left(\sum_{i\in S}\beta_i\right).$$

Each edge S can consist of two more vertices. In particular, a k-uniform hypergraph is a hypergraph such that each edge has has cardinality $k \geq 2$. The β-model for random graphs is a special case where each edge S consists of two vertices. Our setting differs from previous works in that we allow for a partial design of experiments that does not require all possible combinations of vertices of given cardinality to be experimented. See Figure 3.1 for an example of a partial design of experiments for k-uniform random hypergraph. In real examples, it happens when we have limited resources for experiments and we only observe group outcomes for a given collection of sets.

We are interested in understanding what are the fundamental statistical inference limits for inferring parameters of vertices from observed outcomes of edge experiments for the β-model of random hypergraphs. This naturally falls in the framework of maximum likelihood estimation. Specifically, we are interested in understanding conditions for existence and uniqueness of the MLE parameter estimator, and the MLE parameter estimation error bounds when an MLE exists and is unique. We consider this for random graphs and more general case of random hypergraphs according to the β-model, for both full and partial design of experiments.

3.1.1 Related work

The β-model of random graphs has attracted a substantial research interest, with much work devoted to finding conditions for the existence and uniqueness of the maximum likelihood estimation (MLE) of parameter vector $\boldsymbol{\beta}$ from given observation data. In the original paper, Chatterjee et al. (2011) proved uniform consistency of the MLE in the limit when the number of parameters goes to infinity. Subsequently, Yan and Xu (2013) established its asymptotic normality. Rinaldo et al. (2013) found a necessary and sufficient condition for the MLE existence and uniqueness for a given sample of observations. This condition requires the expected degree sequence to be in the interior of a polytope of degree sequences. Hillar and Wibisono (2013) used the general theory of exponential family distributions to derive the existence and uniqueness of the MLE estimator, and proved consistency of the MLE from a single sample in the limit of large graphs. Yan et al. (2016) established consistency and asymptotic normality of a moment estimator for a model of undirected random graphs parametrized by the strength of vertices, which includes the β-model as a special case. Mukherjee et al. (2016) identified sharp detection thresholds for the hypothesis testing problem asking to detect whether the parameter vector $\boldsymbol{\beta}$ of the β-model random graph is a null vector, given observations of edge experiment outcomes for all distinct pairs of vertices. A sparse β-model was studied by Chen et al. (2021).

We note that most of the above-mentioned previous studies considered beta model for graphs with complete design. Our setting differs from previous work in that we allow for a partial design of experiments that does not require all possible combinations of vertices of given cardinality to be experimented. The β-model of random graphs is a special case where each edge S consists of two vertices.

The beta model is a simple model of group outcomes. Other related work include the log-linear model of random graphs by Chung and Lu (2002). Alaoui and Montanari (2019) studied the question of estimating discrete vertex variables from noisy edge observations, showing that linear-time algorithms can achieve a reconstruction accuracy arbitrarily near to the information-theoretic optimum, for graph sequences converging to

so-called amenable graphs. Another related work is on inference for statistical ranking models, e.g. Huang et al. (2006a,b), where the goal is to estimate parameters representing strengths of items from noisy observations of group comparisons.

3.1.2 Summary of contributions

Our results can be summarized in the following points.

- For the β-model of random graphs with a fixed partial design matrix, we found a succinct and easy to interpret condition for the MLE existence and uniqueness. This condition requires that the expected degree sequence is sufficiently bounded away from facets of the polytope of degree sequences. Specifically, for any $c > 1/2$, the MLE exists and is unique with probability at least $1 - 2/n^{2c-1}$, under condition

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \sqrt{\frac{c}{H(\boldsymbol{M})} \frac{\log(n)}{n-1}}$$

  where $\mathcal{E}(\mathbb{E}[\boldsymbol{d}])$ is a "distance" of the expected degree sequence $\mathbb{E}[\boldsymbol{d}]$ from a facet of the polytope of degree sequences, and $H(\boldsymbol{M})$ is a "norm" of the correlation matrix $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{X}$ where $\boldsymbol{X}$ is the $m \times n$ design matrix. Note that $\boldsymbol{M}$ admits an intuitive interpretation as $M_{u,v}$ is the number of (u, v) edge experiments.

  In particular, for the β-model of random graphs with the full design matrix and parameter $\boldsymbol{\beta}$ such that $\epsilon \leq p_{u,v}(\boldsymbol{\beta}) \leq 1 - \epsilon$, for all $u, v \in V$, for some $\epsilon \in (0, 1)$, the condition boils down to the condition on the number of experiments m of the following simple form

$$m \geq c \frac{1}{\epsilon^2} n \log(n)$$

  for some constant $c > 0$.

  These results are obtained by lower bounding the distance between the expected degree vector and facets of the polytope of degree sequences. The proof is based

on the Erdös-Gallai necessary and sufficient condition for graph degree sequences Erdös and Gallai (1960) and concentration of measure.

- We identified sufficient conditions for the MLE existence and uniqueness for the β-model of k-uniform hypergraphs. These conditions are derived from a necessary Erdös-Gallai type condition for k-uniform hypergraphs. The conditions are on the expected degree sequence and the expected density of edges for all sufficiently large sets of vertices. Specifically, for the β-model with parameter $\boldsymbol{\beta}$ such that $\epsilon \leq p_S(\boldsymbol{\beta}) \leq 1-\epsilon$, for all $S \subseteq V$ with $|S| = k$, the MLE exists with high probability provided that $\epsilon = \Omega(1/n^{(k-1)/(k+2)})$.

- We derived bounds on the MLE error $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||$, where a key role has the smallest eigenvalue of matrix $\boldsymbol{M}$, we denote with $\lambda_1(\boldsymbol{M})$. The bound allows for arbitrary design matrices $\boldsymbol{X}$ as long as $\lambda_1(\boldsymbol{M}) > 0$, i.e. $\text{rank}(\boldsymbol{X}) = n$. The eigenvalue $\lambda_1(\boldsymbol{M})$ is related to graph property known as graph non-bipartiteness. This connection provides an intuitive interpretation of the algebraic condition $\lambda_1(\boldsymbol{M}) > 0$.

Organization of chapter The chapter is organized as follows. In section 3.2 we define the model formally. We present three different MLE existence and uniqueness conditions in section 3.3. Section 3.4 contains our results on the MLE error bounds and the relation between the full rank condition for the design matrix and the graph non-bipartiteness.

3.2 Problem formulation

3.2.1 Model specification

Let $V = \{1, \ldots, n\}$ be a set of vertices with $n \geq 2$. For any given collection of non-empty sets $S_1, \ldots, S_m \subseteq V$, let $y_1, \ldots, y_m$ be binary variables taking values in $\{0, 1\}$. Under the

β-model, $y_1, \ldots, y_m$ are independent random variables with distribution

$$\Pr[y_j = 1] = 1 - \Pr[y_j = 0] = \sigma\left(\sum_{i \in S_j} \beta_i\right)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$ is the parameter vector in $\mathbb{R}^n$.

The general case, defined above, allows us to model random hypergraphs, where each edge consists of two or more vertices. In particular, it allows us to model k-uniform hypergraphs.

The β-model is a logistic regression model with binary-valued covariate vectors. Let $\mathbf{X} \in \{0,1\}^{m \times n}$ be the *design matrix* with row (covariate) vectors $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_m^\top$. Then, we can write

$$\Pr[y_j = 1] = 1 - \Pr[y_j = 0] = \sigma(\boldsymbol{x}_j^\top \boldsymbol{\beta}).$$

Using standard graph theory terminology, we may refer to $\boldsymbol{B} := \boldsymbol{X}^\top$ as a graph *incidence matrix*, where $B_{v,e} = 1$ if, and only if, vertex v is an element of edge e. Note that, in general, we allow for hypergraphs with multiple edges, i.e. we allow for $\boldsymbol{x}_e = \boldsymbol{x}_{e'}$ for some $1 \leq e < e' \leq m$. We also define the correlation matrix $\boldsymbol{M}$ s.t.

$$\boldsymbol{M} = \sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^\top = \boldsymbol{X}^\top \boldsymbol{X}.$$

The $(u.v)$ entry of the correlation matrix $\boldsymbol{M}$ denotes the number of experiments involving vertex u and v.

3.2.2 The log-likelihood function

We study the maximum likelihood estimation (MLE) for the β-model of random hypergraphs. We are interested in understanding to which extent we can estimate parameters associated with individual vertices from group (edge) experiments. This naturally leads

us to use the framework of maximum likelihood estimation.

The log-likelihood function of the β-model can be written as

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left(y_j \log(\sigma(\boldsymbol{x}_j^\top \boldsymbol{\beta})) + (1 - y_j) \log(1 - \sigma(\boldsymbol{x}_j^\top \boldsymbol{\beta})) \right). \tag{3.1}$$

An MLE parameter vector $\hat{\boldsymbol{\beta}}$ in a given convex set $\Theta \subseteq \mathbb{R}^n$ is a point $\hat{\boldsymbol{\beta}} \in \Theta$ satisfying

$$\hat{\boldsymbol{\beta}} \in \arg\max_{\boldsymbol{\beta} \in \Theta} \ell(\boldsymbol{\beta}).$$

We are primarily focused on the unconstrained case when $\Theta = \mathbb{R}^n$. In this case, an MLE may not exist, and if it exists may not be unique. If Θ is a bounded convex set, then an MLE $\hat{\boldsymbol{\beta}}$ in Θ always exists because $-\ell(\boldsymbol{\beta})$ is a convex function.

The logistic regression model belongs to the exponential family of models, hence, we can express the log-likelihood function as follows

$$\ell(\boldsymbol{\beta}) = \boldsymbol{t}(\boldsymbol{y})^\top \boldsymbol{\beta} - \kappa(\boldsymbol{\beta}) \tag{3.2}$$

where $\boldsymbol{t}(\boldsymbol{y})$ is the minimal sufficient statistic given by

$$\boldsymbol{t}(\boldsymbol{y}) = \sum_{j=1}^{m} y_j \boldsymbol{x}_j$$

and $\kappa(\boldsymbol{\beta})$ is the log-partition function given by

$$\kappa(\boldsymbol{\beta}) = \sum_{j=1}^{m} \log(1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}).$$

Note that the sufficient statistic $\boldsymbol{t}(\boldsymbol{y})$ has an intuitive interpretation as each $\boldsymbol{t}_v(\boldsymbol{y})$ is the number of successful experiments that involve item v.

We will also use a different but equivalent formulation that is defined as follows. Let $\tilde{\boldsymbol{X}} \in \{0,1\}^{\tilde{m} \times n}$ be a design matrix with distinct row vectors $\tilde{\boldsymbol{x}}_1^\top, \ldots, \tilde{\boldsymbol{x}}_{\tilde{m}}^\top$. Here we may

interpret $\tilde{\boldsymbol{X}}$ to specify all possible distinct covariate vectors. Let m_j denote the number of occurrences of $\tilde{\boldsymbol{x}}_j$ in the observed data. Note that $\sum_{j=1}^{\tilde{m}} m_j = m$. Let $\tilde{y}_j$ be the number of successful experiments with the covariate vector $\tilde{\boldsymbol{x}}_j$.

The log-likelihood function (3.1) can be written as (up to a constant additive term that we can ignore),

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{\tilde{m}} \left(\tilde{y}_j \log(\sigma(\tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta})) + (m_j - \tilde{y}_j) \log(1 - \sigma(\tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta})) \right).$$

Using the exponential-family parametrization, the log-likelihood function can be further expressed as

$$\ell(\boldsymbol{\beta}) = \tilde{\boldsymbol{t}}(\boldsymbol{y})^\top \boldsymbol{\beta} - \tilde{\kappa}(\boldsymbol{\beta})$$

where $\tilde{\boldsymbol{t}}(\boldsymbol{y}) = \sum_{j=1}^{\tilde{m}} \tilde{y}_j \tilde{\boldsymbol{x}}_j$ and $\tilde{\kappa}(\boldsymbol{\beta}) = \sum_{j=1}^{\tilde{m}} m_j \log(1 + e^{\tilde{\boldsymbol{x}}_j^\top \boldsymbol{\beta}})$.

3.3 MLE existence and uniqueness condition

The necessary and sufficient conditions for the MLE existence and uniqueness for the β-model can be expressed in different forms by drawing from the literature on statistical inference for exponential-family models and logistic regression models.

3.3.1 Overlapping condition

The necessary and sufficient condition for MLE existence and uniqueness can be derived from the overlapping condition developed for logistic models. As introduced in section 2.1, $(\mathbf{X}, \mathbf{y})$ is said to satisfy the *overlapping condition*, if there exists no $\boldsymbol{\alpha} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that for all $j \in [m]$,

$$\boldsymbol{x}_j^\top \boldsymbol{\alpha} \geq 0 \text{ if } y_j = 1 \text{ and } \boldsymbol{x}_j^\top \boldsymbol{\alpha} \leq 0 \text{ if } y_j = 0.$$

A theorem by Albert and Anderson (1984) implies that under condition that $\boldsymbol{X}$ has full

rank, an MLE exists for the β-model if, and only if, $(\boldsymbol{X}, \boldsymbol{y})$ satisfies the overlapping condition.

Silvapulle (1981) has also derived a necessary and sufficient condition, under condition that $\boldsymbol{X}$ has full rank, using convex cones

$$S = \left\{ \sum_{j \in [m]: y_j = 1} w_j \boldsymbol{x}_j : \boldsymbol{w} \in \mathbb{R}_+^m \right\} \text{ and } F = \left\{ \sum_{j \in [m]: y_j = 0} w_j \boldsymbol{x}_j : \boldsymbol{w} \in \mathbb{R}_+^m \right\}$$

which reads us

$$S \cap F = \emptyset \text{ or one of } S, F \text{ is } \mathbb{R}^n. \tag{3.3}$$

The overlapping condition is equivalent to condition (3.3) as stated next.

Lemma 3.3.1. *$(\boldsymbol{X}, \boldsymbol{y})$ satisfies the overlapping condition if, and only if, the convex cones S and F satisfy $S \cap F = \emptyset$.*

We first note that it is necessary to have a full rank design matrix.

Lemma 3.3.2. *The negative log-likelihood function of the β-model is strictly convex if, and only if, the design matrix $\boldsymbol{X}$ has linearly independent columns, i.e. $\boldsymbol{X}$ has full rank, $\text{rank}(\boldsymbol{X}) = n$.*

The strict convexity of the log-likelihood function ensures the uniqueness of an optimal point if it exists. The overlapping condition further ensures that the minimum set is non-empty and bounded, therefore the MLE exists and is unique.

We further note that for the β-model, the overlapping condition and the rank of matrix $\boldsymbol{X}$ satisfy the following relation.

Lemma 3.3.3. *If $(\boldsymbol{X}, \boldsymbol{y})$ satisfies the overlapping condition, then $\boldsymbol{X}$ has full rank.*

The theorem by Albert and Anderson (1984) and Lemma 3.3.3 imply the following fact for the β-model.

Proposition 3.3.4. *For any given $(\boldsymbol{X}, \boldsymbol{y})$, there exists a unique MLE for the β-model if, and only if, $(\boldsymbol{X}, \boldsymbol{y})$ satisfies the overlapping condition.*

We are also interested in the case when the MLE exists but is not unique. By Proposition 3.3.4, this could happen when $(\boldsymbol{X}, \boldsymbol{y})$ does not satisfy the overlapping condition and $\boldsymbol{X}$ does not have full rank. The following result provides a necessary and sufficient condition for the MLE existence in the case when $\boldsymbol{X}$ does not have full rank.

Proposition 3.3.5. *If $(\boldsymbol{X}, \boldsymbol{y})$ is such that $\boldsymbol{X}$ is not of full rank, then non-unique MLE exist for the β-model if, and only if,*

$$0 < \tilde{y}_j < m_j \text{ for all } j \in \{1, \ldots, \tilde{m}\}.$$

The condition in Proposition 3.3.5 means that for each distinct set of vertices that participate in an experiment, the fraction of successful experiments involving these vertices is bounded away from 0 and 1.

An intuitive example comes from the binary-valued beta models. Suppose the experiments are drawn according to a bipartite graph such that all left nodes are of a low type and all right nodes are of high type. Then given any data, we cannot identify between the left and right nodes. In this case, we can have MLE solutions but there cannot exist a unique MLE $\hat{\boldsymbol{\beta}}$. For contradiction, assume that $\hat{\boldsymbol{\beta}}$ is a unique MLE parameter vector. Then, by changing the sign for all entries $\hat{\beta}_v$ for $v \in S \cup T$, the resulting parameter vector is also a MLE parameter vector. This contradicts the assumption that $\hat{\boldsymbol{\beta}}$ is a unique MLE parameter vector.

3.3.2 Polytope-typed condition

An alternative formulation of the necessary and sufficient condition for the MLE existence and uniqueness can be derived from the statistical inference theory for exponential-family distributions. By Theorem 2.1.4 in section 2.1, for any exponential-family distribution with sufficient statistic $\boldsymbol{t}$, the log-likelihood function has a unique maximum if, and only if, $\boldsymbol{t} \in \text{int}(C)$ where C is the convex support of the exponential family distribution.

For the β-model, the sufficient statistic is $\boldsymbol{t}(\boldsymbol{y}) = \sum_{j=1}^{m} y_j \boldsymbol{x}_j = \boldsymbol{X}^\top \boldsymbol{y}$, where $\boldsymbol{y}$ is the vector of experiment outcomes. This can be interpreted as a graph degree sequence. The support of an exponential-family distribution is the set of all possible values of sufficient statistic $\boldsymbol{t}(\boldsymbol{y})$. Thus the condition in the theorem by Barndorff-Nielsen (1978) corresponds to the following for the MLE existence and uniqueness for the β-model: for any given experiment outcomes $\boldsymbol{y} \in \{0,1\}^m$, there is a unique MLE for the β-model if, and only if,

$$\boldsymbol{t}(\boldsymbol{y}) \in \text{int conv}(\{\boldsymbol{t}(\boldsymbol{z}) : \boldsymbol{z} \in \{0,1\}^m\})$$

where $\text{conv}(S)$ denotes the convex-hull of a set S and $\text{int conv}(S)$ denotes the interior of this set. In other words, a unique MLE exists if, and only if, the sufficient statistic $\boldsymbol{t}(\boldsymbol{y})$ is in the interior of a polytope of graph degree sequences. We call this the polypote-type condition.

However, this condition is hard to interpret intuitively and to test computationally. Chatterjee et al. (2011) provided more explicit conditions for the MLE existence and uniqueness, under assumption that for each pair of distinct vertices there is exactly one experiment (hence, graph is simple). Rinaldo et al. (2013) provided a condition that allows for random graphs with one or more experiments for distinct pairs of vertices. This condition involves a normalized degree sequence $d(\boldsymbol{y})$ with entries defined as

$$d_u(\boldsymbol{y}) = \sum_{v \in V \setminus \{u\} : M_{u,v} > 0} \frac{\tilde{y}_{u,v}}{M_{u,v}}$$

which for each vertex corresponds to the sum of empirical success frequencies of edge experiments incident to this vertex.

Both results are based on the well-known Erdös-Gallai characterization of graph degree sequences. By Erdös and Gallai (1960), a sequence of non-negative integers $d_1, \ldots, d_n$ is a degree sequence of a finite simple graph $G = (V, E)$ on n vertices if, and only if,

$d_1 + \cdots + d_n$ is even and for every non-empty set $S \subseteq [n]$,

$$\sum_{v \in S} d_v \leq |S|(|S| - 1) + \sum_{v \in V \setminus S} \min\{d_v, |S|\}. \tag{3.4}$$

It has been shown that the Erdös-Gallai inequalities determines the polytope P_n of graph degree sequences. Peled and Srinivasan (1989) explicitly showed that the facets of P_n are defined by the following linear inequalities, for any $n \geq 4$,

$$d_v \geq 0 \text{ for all } v \in [n] \tag{3.5}$$

$$d_v \leq n - 1 \text{ for all } v \in [n] \tag{3.6}$$

$$f(S, T, \boldsymbol{d}, n) \geq 0 \text{ for all } (S, T) \in \Omega \tag{3.7}$$

where

$$f(S, T, \boldsymbol{d}, n) := |S|(n - 1 - |T|) - \left(\sum_{v \in S} d_v - \sum_{v \in T} d_v \right) \tag{3.8}$$

and

$$\Omega := \{(S, T) \subseteq [n] : S \cap T = \emptyset, |S \cup T| \in \{2, \ldots, n - 3\} \cup \{n\}\} \tag{3.9}$$

and, for $n = 3$, the facets are only as given by (3.7).

Therefore, for the β-model of random graphs, a necessary and sufficient condition for the MLE existence is that the degree sequence $\boldsymbol{t}(\boldsymbol{y})$ satisfies the linear inequalities (3.5)-(3.7) with strict inequalities.

3.3.3 Interpretable MLE condition

We next present our main results for MLE existence and uniqueness. These conditions are expressed using two key parameters, one quantifying a "distance" of the expected normalized degree sequence to a facet of polytope P_n and other quantifying "connectivity" of a graph associated with $\boldsymbol{M}$.

The first parameter is $\mathcal{E}(\boldsymbol{d})$ defined as, for $\boldsymbol{d} \in [0, n-1]^n$,

$$\mathcal{E}(\boldsymbol{d}) = \frac{1}{n-1} \min \left\{ \min_{v \in V} \{d_v\}, \min_{v \in V} \{n-1-d_v\}, \min_{(S,T) \in \Omega} \left\{ \frac{f(S,T,\boldsymbol{d},n)}{|S \cup T|} \right\} \right\}$$

where f and Ω are defined in (3.8) and (3.9). Intuitively, we can interpret $(n-1)\mathcal{E}(\boldsymbol{d})$ as a minimum slack for the linear inequalities in (3.5)-(3.7) at point $\boldsymbol{d}$ with respect to the constraints defining the facets of P_n. Condition that $\boldsymbol{d}$ is in the interior of P_n is equivalent to minimum slack being strictly positive. We scale the minimum slack with factor $n-1$ as this is a natural normalization.

The second parameter is $H(\boldsymbol{M})$, defined as $H(\boldsymbol{M}) = \min_{u \in V} H_u(\boldsymbol{M})$, where for $u \in V$,

$$H_u(\boldsymbol{M}) = \frac{n-1}{\sum_{v \in V \setminus \{u\} : M_{u,v} > 0} \frac{1}{M_{u,v}}}.$$

Theorem 3.3.6. *For any β-model of random graphs with correlation matrix $\boldsymbol{M}$ with $n \geq 3$ and no null rows, for any $c > 1/2$, there exists a unique MLE with probability at least $1 - 2/n^{2c-1}$, under condition*

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \sqrt{\frac{c}{H(\boldsymbol{M})} \frac{\log(n)}{n-1}}.$$

Furthermore, if $m \leq H(\boldsymbol{M})\binom{n}{2}$, then the condition can be written as

$$m \geq \frac{c}{2} \frac{1}{\mathcal{E}(\mathbb{E}[\boldsymbol{d}])^2} n \log(n).$$

In particular, the theorem applies to the complete graph case, where $M_{u,v} = r \geq 1$ for every $u \neq v$, as follows. In this case, $H(\boldsymbol{M}) = r$ and $m = r\binom{n}{2}$, and the condition is equivalent to the condition on the number of experiments per distinct pair of vertices given as follows

$$r \geq \max \left\{ c \frac{1}{\mathcal{E}(\mathbb{E}[\boldsymbol{d}])^2} \frac{\log(n)}{n-1}, 1 \right\}.$$

Next, we give a bound for the function of expected degree sequence $\mathcal{E}(\mathbb{E}[\boldsymbol{d}])$ and give a

corollary of the main theorem that explicitly writes the condition in terms of minimum number of experiments.

Corollary 3.3.7 (of Theorem 3.3.6)**.** *Consider the case where $n \geq 3$ and for each distinct pair of vertices there is at least one experiment. Assume that $\boldsymbol{\beta}$ is such that $\epsilon \leq p_{u,v}(\boldsymbol{\beta}) \leq 1-\epsilon$ for all $u, v \in V$, $u \neq v$, for some $\epsilon \in (0,1)$, then we have $\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \frac{1}{4}\epsilon$. For any $c > 1/2$, an MLE exists and is unique with probability at least $1 - 2/n^{2c-1}$, provided that the number of edge experiments m satisfies*

$$m \geq 12c\frac{1}{\epsilon^2} n \log(n).$$

Note that $m \geq \binom{n}{2}$ under the assumptions of the corollary, so the asserted condition for the number of edge experiments is non-trivial only for sufficiently small ϵ, in particular when $\epsilon = O(\sqrt{\log(n)/n})$.

The condition in Theorem 3.3.6 is a generalization of a condition in Rinaldo et al. (2013). Compared to their condition, we relax the assumption of testing each edge the same number of times. It is interpretable as we expressed it in terms of two specific parameters with graph-theoretical meanings. Moreover, we define and give a bound for the function of expected degree sequence $\mathcal{E}(\mathbb{E}[\boldsymbol{d}])$. This enables us to explicitly convert the condition to the requirement on the number of experiments.

The proof for Theorem 3.3.6 is given at the end of the chapter. It is based on the Erdös-Gallai condition for graph degree sequences and concentration of measure. We define a bad event such that the linear inequalities (3.5)-(3.7) holds with equality. Then we bound the probability of this bad event by a concentration of measure for the normalized degree sequences $\boldsymbol{d}$. Finally, we apply the Hoeffding's bound on $\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]$ and take a union bound to arrive at the final condition.

Lower bound results We next discuss a lower bound for the number of edge experiments m. Suppose that for each distinct pair of items there are $r \geq 1$ experiments and that

$\boldsymbol{\beta}$ is such there exists u such that $p_{u,v}(\boldsymbol{\beta}) = \epsilon$, for all $v \neq u$, for some $\epsilon \in (0,1)$. Note that the probability that the degree sequence $\boldsymbol{d}$ is on a facet of the degree sequence polytope is greater than or equal to the probability of the event $\{d_u = 0\}$, and

$$\Pr[d_u = 0] = \prod_{v \in V \setminus \{u\}} (1 - p_{u,v}(\boldsymbol{\beta}))^r.$$

Since $m = r\binom{n}{2}$, we have $\Pr[d_u = 0] = (1-\epsilon)^{\frac{2m}{n}}$. From this it follows that for $\Pr[d_u = 0] \leq 1/n^a$ to hold, for some $a > 0$, it is *necessary* that

$$m \geq \frac{a}{2} \frac{1}{\log(\frac{1}{1-\epsilon})} n \log(n).$$

This establishes the lower bound $\Omega(\frac{1}{\epsilon} n \log(n))$ for the number of edge experiments for the normalized degree sequence d to be in the interior of the polytope with probability at least $1 - 1/n^a$. This matches the bound in Corollary 3.3.7 up to a factor $1/\epsilon$. The factor $1/\epsilon^2$ in Corollary 3.3.7 comes from using a concentration bound for the deviation of the normalized degree sequence from the expected normalized degree sequence in the proof of Corollary 3.3.7.

3.3.4 Results for hypergraphs

As mentioned in the overview, collecting complete network data is often expensive and infeasible. Therefore, it is important to consider the case of β-model of random hypergraphs. We consider the case of k-uniform random hypergraphs with full design matrix. It is challenging to extend the results for random graphs to random hypergraphs. Note that the previous analysis relies heavily on the Erdös-Gallai characterization of graph degree sequences and the facet definition of the degree sequences polytope. However, the Erdös-Gallai condition only holds for graphs and facet definition for hypergraph P_n has not been established yet.

To claim a sufficient condition on the number of experiments such that MLE almost surely

exists, we first provide an Erdös-Gallai typed necessary condition for a sequence to be a degree sequence of a k-uniform hypergraph.

Lemma 3.3.8. *If $d = (d_1, \ldots, d_n)$ is a degree sequence of a k-uniform hypergraph $H = (V, E)$ with $|V| = n$, then, for every $S \subseteq V$ such that $|S| \geq k$,*

$$\sum_{v \in S} d_v \leq k\binom{|S|}{k} + \sum_{v \in V \setminus S} \min\left\{(k-1)d_v, \frac{|S|}{n-|S|}\left(\binom{n-1}{k-1} - \binom{|S|-1}{k-1}\right)\right\}.$$

For the graph case $G = (V, E)$, the claim of the lemma boils down to the Erdös-Gallai conditions (3.4).

Based on the Erdös-Gallai typed characterization of hypregraph degree sequences, we next introduce a set of conditions that are sufficient to guarantee MLE existence.

Suppose $d = (d_1, \ldots, d_n)$ is a point in the set of *expected* degree sequences of a k-uniform random hypergraph with, for some $\hat{\boldsymbol{\beta}} \in (\mathbb{R} \cup \{\infty\})^n$,

$$d_v = \sum_{S \subseteq V : |S| = k, v \in S} p_S(\hat{\boldsymbol{\beta}}), \text{ for all } v \in V.$$

Assume there exist constants α_1, α_2, α_3, and $\alpha_4 \in (0, 1)$ such that

(C1) For all $v \in V$,

$$\alpha_1\binom{n-1}{k-1} \leq d_v \leq (1-\alpha_2)\binom{n-1}{k-1} \tag{3.10}$$

and

(C2) For all $S \subseteq V$ such that $|S| \geq \alpha_1 n$,

$$\alpha_3\binom{|S|}{k} \leq \sum_{S' \subseteq S : |S'| = k} p_{S'}(\hat{\boldsymbol{\beta}}) \leq (1-\alpha_4)\binom{|S|}{k}. \tag{3.11}$$

Condition (C1) requires that the expected degree of each vertex is within specified factors of the maximum possible degree of a vertex. Condition (C2) requires that the expected

number of edges contained in every sufficiently large set of vertices is within specified factors of the maximum possible number of edges.

Lemma 3.3.9. *Under conditions (C1) and (C2), $||\hat{\boldsymbol{\beta}}||_\infty \leq c$, where c is a positive constant depending on $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and k.*

The proof is given at the end of the chapter. It based on a contradiction argument similar to the proof in Chatterjee et al. (2011) for β-model of random graphs. If we further have the design matrix $\boldsymbol{X}$ to be of full rank, then the uniqueness of MLE follows from Theorem 1.5 in Chatterjee et al. (2011).

We next present the following theorem which provides a sufficient condition for the existence of MLE for the β-model of a k-uniform random hypergraph with high probability.

Theorem 3.3.10. *Suppose $H = (V, E)$ is a k-uniform random hypergraph drawn from the β-model with parameter $\boldsymbol{\beta}$ such that $\epsilon \leq p_S(\boldsymbol{\beta}) \leq 1 - \epsilon$, for all $S \subseteq V$ with $|S| = k$, for some $\epsilon \in (0, 1)$. Then, for any $c > 0$, conditions (3.10) and (3.11) hold with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \epsilon/2$, with probability at least $1 - 1/n^c$, provided that*

$$\epsilon \geq \max\{f_{n,k}, g_{n,k}\} \tag{3.12}$$

where

$$\begin{aligned} f_{n,k} &= \sqrt{2} k^{\frac{k-1}{2}} \sqrt{\frac{(c+1)\log(n) + \log(4)}{n^{k-1}}} \\ g_{n,k} &= 2^{\frac{k+1}{k+2}} k^{\frac{k}{k+2}} \left(\frac{\log(2)}{n^{k-1}} + \frac{c\log(n) + \log(4)}{n^k} \right)^{\frac{1}{k+2}}. \end{aligned}$$

The key idea of the proof is similar as the random graph case. We define bad events such that conditions (3.10) and (3.11) do not hold. We bound the probability of bad events in terms of concentration measure of expected degree sequences. We arrive at the final result by giving a bound for the expected degree sequences in terms of ϵ.

Note that the right-hand side in (3.12) scales with n as $1/n^{(k-1)/(k+2)}$, asymptotically for large n. In particular, for $k = 2$, it scales as $1/n^{1/4}$. By Theorem 3.3.6, we know that for $k = 2$, it suffices that $\epsilon \geq \Omega(\sqrt{\log(n)/n})$. The extra $n^{1/4}/\sqrt{\log(n)}$ factor is due to using union bound for the events over sets $S \subseteq V$ such that $|S| \geq \alpha_1 n$ to ensure (3.11) holds with high probability. Note, however, that the right-hand side in (3.12) is $O(1/n^{2/5})$ for $k = 3$, and $O(1/\sqrt{n})$, for all $k \geq 4$. Hence, in the latter case, the condition for the existence of MLE is weaker for a k-uniform random graph than for the graph case, for any sufficiently large n.

3.4 MLE error bounds

In this section, we consider the MLE error under condition that MLE exists and is unique. We consider the estimation error measured by the L_2-norm of the difference of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ and the true parameter vector $\boldsymbol{\beta}$, i.e. $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||$. We will show that the key parameter that determines the parameter estimation error is the smallest eigenvalue $\lambda_1(\boldsymbol{M})$ of the the correlation matrix $\boldsymbol{M}$. Recall that

$$\boldsymbol{M} = \sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^\top = \boldsymbol{X}^\top \boldsymbol{X}.$$

We will then show how $\lambda_1(\boldsymbol{M})$ is related to to some parameters reflecting non-bipartiteness of the graph with adjacency matrix $\boldsymbol{M}$.

3.4.1 Parameter estimation error bounds

We first show the following parameter estimation bound for any β-model with edges of cardinality $k \geq 2$.

Proposition 3.4.1. *Suppose $\boldsymbol{X}$ is the design matrix with row vectors $\boldsymbol{x}_j^\top$ satisfying $||\vec{x}_j||_1 = k$ for all $j \in [m]$, $\lambda_1(\boldsymbol{M}) > 0$, and that experiment outcomes $\boldsymbol{y}$ are according to the β-model with*

parameter vector $\boldsymbol{\beta}$ *such that* $||\boldsymbol{\beta}||_\infty \leq b$*, for some* $b > 0$*. Then, under condition* $||\hat{\boldsymbol{\beta}}||_\infty \leq b$*,*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \leq c_{kb} \frac{\sqrt{mk(\log(n) + 2)}}{\lambda_1(\boldsymbol{M})} \tag{3.13}$$

with probability at least $1 - 2/(n \Pr[||\hat{\boldsymbol{\beta}}||_\infty \leq b])$*, where* c_{kb} *is a positive constant depending only on the product kb. In particular, we can take* $c_{kb} = 2\sqrt{2}(1 + e^{kb})$*.*

The proof mainly relies on the Taylor expansion. We can see that $\lambda_1(\boldsymbol{M}) > 0$ is necessary for a finite mean square error bound. Condition $\lambda_1(\boldsymbol{M}) > 0$ is equivalent to matrix $\boldsymbol{X}$ having full rank. Moreover, as shown in section 3.3.4, if the design matrix $\boldsymbol{X}$ is of full rank and $||\hat{\boldsymbol{\beta}}||_\infty \leq b$, $(\boldsymbol{X}, \boldsymbol{y})$ satisfies the MLE existence and uniqueness condition. If the MLE parameter vector $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of the negative log-likelihood function over a bounded convex set, then $\Pr[||\hat{\boldsymbol{\beta}}||_\infty \leq b] = 1$.

It is insightful to consider the parameter estimation bound in Proposition 3.4.1, for the case of a complete k-uniform hypergraph, i.e. when rows of $\boldsymbol{X}$ consist of all distinct vectors in $\{0,1\}^n$ with k entries equal to 1 and the remaining entries equal to 0. In this case, we have

$$m = \binom{n}{k}, M_{u,v} = \binom{n-2}{k-2} \text{ for } u \neq v \text{ and } M_{u,u} = \binom{n-1}{k-1}.$$

It can be readily shown that

$$\lambda_1(\boldsymbol{M}) = \frac{k^2}{n}\binom{n}{k}.$$

For the complete k-uniform hypergraph, from (3.13), we have

$$\frac{1}{\sqrt{n}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \leq c_{kb}\sqrt{\frac{n(\log(n) + 2)}{k^3\binom{n}{k}}} \leq c_{kb} 2k^{\frac{k-3}{2}}\sqrt{\frac{\log(n)}{n^{k-1}}}.$$

Thus, for every fixed $k \geq 2$, $\frac{1}{\sqrt{n}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| = O(\sqrt{\log(n)/n^{k-1}})$. Hence, we observe that the parameter estimation error bound decreases faster with n for larger values of k.

For the complete graph case, we have

$$\frac{1}{\sqrt{n}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \leq c_b\sqrt{\frac{\log(n)}{n}}$$

where c_b is a positive constant. This bound is of the same form as the bound for the L_∞ norm in Theorem 1.3 of Chatterjee et al. (2011). Specifically, their theorem says that if $||\boldsymbol{\beta}||_\infty \leq b$, then there exists $c_b > 0$ such that with probability at least $1 - c_b/n^2$, there exists a unique MLE $\hat{\boldsymbol{\beta}}$, which satisfies

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_\infty \leq c_b\sqrt{\frac{\log(n)}{n}}.$$

3.4.2 Key property of graph bipartiteness

We further investigate if the eigenvalue $\lambda_1(\boldsymbol{M})$ relates to any graph-theoretic properties of the inputs for beta model.

We first consider the graph case. The key property is known as *graph non-bipartiteness*. For a given graph $G = (V, E)$, let $G(S)$ be the subgraph of G with the set of vertices restricted to set $S \subseteq V$. For any non-empty $S \subseteq V$, let $\text{cut}(S)$ be the set of edges between vertices in S and $V \setminus S$ and $e_{\min}(S)$ be the minimum number of edges that need to be removed from $G(S)$ so that the resulting subgraph is bipartite. Let

$$\psi = \min_{S \subseteq V} \frac{|\text{cut}(S)| + e_{\min}(S)}{|S|}.$$

The quantity ψ is a natural measure of a graph non-bipartiteness. If $\psi = 0$, then clearly the graph has a bipartite component. Intuitively, the larger the ψ for a graph, in some sense the further away is the graph from a graph with a bipartite component. We make note of the following basic fact:

$$\psi > 0 \text{ if, and only if } \lambda_1(\boldsymbol{M}) > 0. \tag{3.14}$$

By definition, $\psi = 0$ if, and only if, the underlying graph has a bipartite component. It is well known that the rank of the incidence matrix $\boldsymbol{X}^\top$ is related to the number of bipartite components in the associated graph. By Theorem 8.2.1 in Godsil and Royle (2001), if a graph has n vertices and c bipartite components with incidence matrix $\boldsymbol{X}^\top$, then $\text{rank}(\boldsymbol{X}^\top) = n - c$. Therefore, $\psi = 0$ if, and only if, the graph has no bipartite components.

It is readily observed that if $\psi = 0$, then there cannot exist a unique MLE $\hat{\boldsymbol{\beta}}$. To see this, note that $\psi = 0$ implies that there exists an isolated bipartite component, i.e. there exist two non-empty disjoint sets $S, T \subseteq N$ such that no edge exists with both vertices in S, or T, and no edge exists with exactly one vertex contained in $S \cup T$. Recall our intuitive example given at the end of section 3.3.1. For contradiction, assume that $\hat{\boldsymbol{\beta}}$ is a unique MLE. Then, we can obtain another MLE parameter vector by changing the sign for all entries $\hat{\beta}_v$ for $v \in S \cup T$. This contradicts the assumption of unique MLE.

The eigenvalue $\lambda_1(\boldsymbol{M})$ and the non-bipartiteness measure ψ satisfy a stronger relation than (3.14). By Desai and Rao (1994), for a graph G with incidence matrix $\boldsymbol{X}^\top$, the smallest eigenvalue of $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{X}$ satisfies

$$\frac{1}{4d^*}\psi^2 \leq \lambda_1(\boldsymbol{M}) \leq 4\psi \tag{3.15}$$

where d^* is the largest degree of a vertex in G (or, equivalently, the maximum column sum of $\boldsymbol{X}$). Indeed, (3.15) implies (3.14). Further relationships can be found in the more recent work by Fallat and Fan (2012).

Our analysis reveals a connection between the maximum likelihood error bound and the non-bipartiteness property of a graph associated with the design matrix. For paired comparisons and ranking models, it is well-known that the bound depends on the algebraic connectivity of the matrix of paired comparison counts, which is captured by the smallest eigenvalue of the Laplacian matrix Shah et al. (2016); Hajek et al. (2014); Vojnovic and Yun (2016). To the best of our knowledge, such connection has not been established previously

for β-model of hypergraphs.

We next show that a similar relation between $\lambda_1(\boldsymbol{M})$ and graph non-bipartitness measure ψ holds for the more general case of hypergraphs, which allow for design matrices $\boldsymbol{X}$ to have binary-valued elements with one or more unit-valued elements per row.

The key is to project a hypergraph to a weighted graph. For the graph case, $\boldsymbol{M}$ is the signed Laplacian matrix as it can be decomposed as the sum $\boldsymbol{M} = \boldsymbol{D} + \boldsymbol{A}$ where $\boldsymbol{D}$ is the degree matrix and $\boldsymbol{A}$ is the adjacency matrix. We can decompose $\boldsymbol{M}$ in a similar way for the more general case of a hypergraph. Note that $\boldsymbol{M}$ has elements given by

$$M_{u,v} = \sum_{j=1}^{m} x_{j,u} x_{j,v} \text{ for } u \neq v \text{ and } M_u := M_{u,u} = \sum_{j=1}^{m} x_{j,u}.$$

Let $\boldsymbol{A}$ be the adjacency matrix defined as

$$A_{u,v} = \begin{cases} M_{u,v} & \text{if } u \neq v \\ 0 & \text{if } u = v \end{cases}$$

and let $\boldsymbol{N}$ and $\boldsymbol{D}$ be two diagonal matrices defined as

$$D_{u,v} = \begin{cases} \sum_{w \neq u} M_{u,w} & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases} \text{ and } N_{u,v} = \begin{cases} M_u & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases}.$$

In this way, we can write $\boldsymbol{M} = \boldsymbol{N} + \boldsymbol{A}$. Note that $\boldsymbol{A}$ and $\boldsymbol{D}$ can be treated as the adjacency matrix and degree matrix defined on the weighted graph projected from the hypergraph, such that $\boldsymbol{A}_{u,v}$ denotes the number of experiments vertex u co-participate with vertex v and $\boldsymbol{D}_u$ sums up the total number of times vertex u co-participate with another vertex in the experiments. On the other hand, $\boldsymbol{D}$ is the diagonal matrix of hypergraph degree and denotes the number of experiments in which vertex u takes part. It is clear that $D_{u,u} \geq N_{u,u}$, for all $u \in V$, with equality holds for the graph case. For the uniform case, when each experiment involves exactly k items, we have $D_{u,u} = (k-1)N_{u,u}$.

Proposition 3.4.2. *Assume $G = (V, E)$ is a hypergraph with matrices $\boldsymbol{M}$, $\boldsymbol{A}$, $\boldsymbol{D}$ and $\boldsymbol{N}$ and ψ is*

the graph non-bipartiteness measure of a graph with adjacency matrix $\boldsymbol{A}$. Then, we have

$$\frac{1}{4d^*}\psi^2 - \max_{u\in V} d_u \le \lambda_1(\boldsymbol{M}) \le 4\psi - \min_{u\in V} d_u$$

where $d_u = (\boldsymbol{D}-\boldsymbol{N})_u$ and $d^* = \max_u D_{u,u}$.

For the graph case, the inequalities of Proposition 3.4.2 correspond to those in (3.15). For the k-uniform hypergraph case, when each experiment involves exactly k vertices, we have $d_u = (k-2)M_u$ and $D_{u,u} = (k-1)M_u$. In this case, the inequalities in Theorem 3.4.2 can be written as

$$\frac{1}{4(k-1)\max_{u\in V} M_u}\psi^2 - (k-2)\max_{u\in V} M_u \le \lambda_1(\boldsymbol{M}) \le 4\psi - (k-2)\min_{u\in V} M_u.$$

Note that

$$\begin{aligned} \psi > A_k \max_{u\in V} M_u \quad &\Rightarrow \quad \lambda_1(\boldsymbol{M}) > 0 \\ \lambda_1(\boldsymbol{M}) > 0 \quad &\Rightarrow \quad \psi > B_k \min_{u\in V} M_u \end{aligned}$$

where $A_k = 2\sqrt{(k-1)(k-2)}$ and $B_k = (k-2)/4$. Both A_k and B_k are with constants factors of k.

3.5 Conclusion

In this chapter, we study the maximum likelihood estimation for the β-model of random hypergraphs under general design of experiments, which allow for different number of experiments over candidate edges. We reviewed the overlapping and polytope-typed MLE conditions. We derived easy-to-interpret conditions for the MLE existence and uniqueness based on the polytope-typed condition. We also provided bounds on the MLE accuracy in terms of mean-square error and related it to a graph-theoretic property known as graph non-bipartiteness.

We derived a matching lower bound for the graph case when $k = 2$. We noted in the main text that our MLE condition for the k-uniform hypergraph is weaker than for the graph case. This is mainly because the Erdös-Gallai condition only holds for graphs and facet definition for hypergraphs has not been established yet. Proving a tight MLE condition for the hypergraph case remains an open problem.

3.6 Proofs

Proof of Lemma 3.3.1

Using the separating hyperplane theorem for cones, $S \cap F = \emptyset$ if, and only if, there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that

$$\forall \boldsymbol{s} \in S, \boldsymbol{f} \in F,\ \boldsymbol{s}^\top \boldsymbol{\alpha} \leq 0 \leq \boldsymbol{f}^\top \boldsymbol{\alpha}$$

which is equivalent to the overlapping condition:

$$\exists \boldsymbol{\alpha} \in \mathbb{R}^n \text{ such that } \forall j \in \{1, \ldots, m\} : (2y_j - 1)\boldsymbol{x}_j^\top \boldsymbol{\alpha} \geq 0.$$

Proof of Lemma 3.3.2

Let $g_j(\boldsymbol{\beta}) = \log(1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}})$. Then, we can write

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left[y_j g_j(\boldsymbol{\beta}) + (1 - y_j) g_j(-\boldsymbol{\beta}) \right].$$

If $g_j(\boldsymbol{\beta})$ is strictly convex for all $j \in \{1, \ldots, m\}$, then so is $\ell(\boldsymbol{\beta})$.

For every $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^n$ and $0 \leq \lambda \leq 1$, we have

$$\begin{aligned} \lambda g_j(\boldsymbol{\beta}) + (1 - \lambda) g_j(\boldsymbol{\beta}') &= \log\left((1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}})^\lambda (1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}'})^{1-\lambda} \right) \\ &\geq \log\left(1 + e^{-(\lambda \boldsymbol{x}_j^\top \boldsymbol{\beta} + (1-\lambda)\boldsymbol{x}_j^\top \boldsymbol{\beta}')} \right) \\ &= g_j(\lambda \boldsymbol{\beta} + (1 - \lambda)\boldsymbol{\beta}'). \end{aligned}$$

If $g_j(\boldsymbol{\beta})$ is strictly convex for all $j \in \{1, \ldots, m\}$, then the equality holds only when $\boldsymbol{x}_j^\top \boldsymbol{\beta} = \boldsymbol{x}_j^\top \boldsymbol{\beta}'$ for all $j \in \{1, \ldots, m\}$, i.e.

$$\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}') = 0$$

and vice versa. Hence, f is strictly convex if, and only if, $\text{null}(\boldsymbol{X}) = 0$. This is equivalent to $\text{rank}(\boldsymbol{X}) = n$.

Proof of Lemma 3.3.3

Suppose that $\boldsymbol{X}$ is not of full rank. Then there exists $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{X}^\top \boldsymbol{\alpha} = 0$, i.e. there exists $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} = 0$ for all $j \in \{1, \ldots, m\}$. Thus, there exists no overlap in the dataset $(\boldsymbol{X}, \boldsymbol{y})$.

It follows that if $(\boldsymbol{X}, \boldsymbol{y})$ satisfies the overlapping condition, then this implies that $\boldsymbol{X}$ has full rank.

Proof of Proposition 3.3.4

Let f denote the negative log-likelihood function of the β-model.

Necessity Suppose that $(\boldsymbol{X}, \boldsymbol{y})$ does not satisfy the overlapping condition and f attains its minimum at $\hat{\boldsymbol{\beta}}$. Then, there exists $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} \geq 0$ if $y_j = 1$ and $\boldsymbol{x}_j^\top \boldsymbol{\alpha} \leq 0$ if $y_j = 0$ for all $j \in \{1, \ldots, m\}$. Without loss generality, we assume that the first r points have value 1 and the remaining points have value 0. Then, we can write

$$f(\boldsymbol{\beta}) = \sum_{j=1}^{r} \log\left(1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}}\right) + \sum_{j=r+1}^{m} \log\left(1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}\right).$$

From the last equation, it is easy to observe that $f(\hat{\boldsymbol{\beta}} + c\boldsymbol{\alpha}) \leq f(\hat{\boldsymbol{\beta}})$ for any $c \geq 0$. This means that the MLE cannot be unique.

Sufficiency If there is an overlap in $(\boldsymbol{X}, \boldsymbol{y})$, then for all $\boldsymbol{\alpha} \in \mathbb{R}^n$, there exists some $1 \leq j \leq r$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} < 0$ or there exists some $r + 1 \leq j \leq m$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} > 0$. We can see that for all $\boldsymbol{\beta}' \in \text{dom } f$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$, $f(\boldsymbol{\beta}' + c\boldsymbol{\alpha}) \to +\infty$, which implies that f has no directions

of recession. By Theorem 2.1.1, the minimum set of f is non-empty and bounded. Then, the uniqueness follows from Theorem 2.1.2 and Lemma 3.3.3.

Proof of Proposition 3.3.5

Necessity If $\boldsymbol{X}$ is not of full rank but we have extreme observations such that $\tilde{y}_j \in \{0, m_j\}$ for some $j \in \{1, \ldots, \tilde{m}\}$, then there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} \geq 0$ for $1 \leq j \leq r$ and $\boldsymbol{x}_j^\top \boldsymbol{\alpha} \leq 0$ for $r+1 \leq j \leq m$ and $\boldsymbol{x}_j^\top \boldsymbol{\alpha} \neq 0$ for some $j \in \{1, \ldots, \tilde{m}\}$.

For any $\boldsymbol{\beta}' \in \operatorname{dom} f$, we consider the sequence of vectors $\boldsymbol{\beta}^{(l)} = \boldsymbol{\beta}' + l\boldsymbol{\alpha}$. The negative log-likelihood function at $\boldsymbol{\beta}^{(l)}$ is of value

$$f(\boldsymbol{\beta}^{(l)}) = \sum_{j=1}^{r} \log\left(1 + e^{-(\boldsymbol{x}_j^\top \boldsymbol{\beta}' + l\boldsymbol{x}_j^\top \boldsymbol{\alpha})}\right) + \sum_{j=r+1}^{m} \log\left(1 + e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}' + l\boldsymbol{x}_j^\top \boldsymbol{\alpha}}\right).$$

Since there is at least one i such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} > 0$ for $1 \leq i \leq r$, or $\boldsymbol{x}_j^\top \boldsymbol{\alpha} < 0$ for $r+1 \leq i \leq m$, $f(\boldsymbol{\beta}^{(l)})$ is strictly decreasing with l. Hence, the MLE is at infinity on the boundary of dom f, i.e. MLE does not exist.

Sufficiency Assume that $\boldsymbol{X}$ is not of full rank, i.e. $\operatorname{rank}(\boldsymbol{X}) < n$, and $0 < \tilde{y}_j < m_j$ for all $j \in \{1, \ldots, \tilde{m}\}$. Assumption $\operatorname{rank}(\boldsymbol{X}) < n$ implies that there exists $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{x}_j^\top \boldsymbol{\alpha} = 0$ for all $j \in \{1, \ldots, \tilde{m}\}$. These directions $\boldsymbol{\alpha}$ belong to the constancy space of f.

Since $0 < \tilde{y}_j < m_j$ for all $j \in \{1, \ldots, \tilde{m}\}$ for every vector $\boldsymbol{\beta}$ not in the null space of $\boldsymbol{X}$, there exists some $1 \leq i \leq r$ such that $\boldsymbol{x}_j^\top \boldsymbol{\beta} < 0$ or some $r+1 \leq i \leq m$ such that $\boldsymbol{x}_j^\top \boldsymbol{\beta} > 0$. From Theorem 2.1.1, we know that these vectors are not directions of recession of f. Therefore, vectors in the null space of $\boldsymbol{X}$ are the only recession directions which also belong to the constancy space of f. By Theorem 2.1.3, the minimum set of f is non-empty. Since the overlapping condition does not hold in this case, the minimum set of f is unbounded. Thus, we conclude that non-unique MLE exist.

Proof of Theorem 3.3.6

We first show a lemma that allows us to turn the problem of showing that the normalized degree sequence $\boldsymbol{d}$ is in the interior of the polytope of degree sequences with high probability to a problem of concentration of measure for the normalized degree sequence $\boldsymbol{d}$. Let B denote the event that $\boldsymbol{d}$ is on a facet of the polytope of degree sequences.

Lemma 3.6.1. *The following inequality holds:*

$$\Pr[B] \leq \Pr\left[||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq (n-1)\mathcal{E}(\mathbb{E}[\boldsymbol{d}])\right].$$

Proof. For any given $m_{u,v}$, for $u, v \in V$ with $u \neq v$, $\tilde{y}_{u,v}$ are independent random variables with $\tilde{y}_{u,v}$ having binomial distribution with parameters $m_{u,v}$ and $p_{u,v}(\boldsymbol{\beta}) = \sigma(\beta_u + \beta_v)$.

Recall the facet defining inequalities (3.5), (3.6), and (3.7).

For every $S, T \subseteq [n]$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$|f(S, T, \boldsymbol{x}, n) - f(S, T, \boldsymbol{y}, n)| \leq |S \cup T|||\boldsymbol{x} - \boldsymbol{y}||_\infty.$$

This yields the following inequality

$$f(S, T, \boldsymbol{d}, n) \geq f(S, T, \mathbb{E}[\boldsymbol{d}], n) - |S \cup T|||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty.$$

Let $B = B_1 \cup B_2 \cup B_3$ where

$$\begin{aligned} B_1 &= \cup_{v \in V}\{d_v = 0\} \\ B_2 &= \cup_{v \in V}\{d_v = n-1\} \\ B_3 &= \cup_{(S,T) \in \Omega}\{f(S, T, \boldsymbol{d}, n) = 0\}. \end{aligned}$$

Note that we have

$$\begin{aligned} B_1 &= \cup_{v\in V}\{d_v = 0\} \\ &\subseteq \cup_{v\in V}\{d_v \leq 0\} \\ &= \cup_{v\in V}\{\mathbb{E}[d_v] - d_v \geq \mathbb{E}[d_v]\} \\ &\subseteq \left\{||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq \min_{v\in V}\mathbb{E}[d_v]\right\}, \end{aligned}$$

$$\begin{aligned} B_2 &= \cup_{v\in V}\{d_v = n-1\} \\ &\subseteq \cup_{v\in V}\{d_v \geq n-1\} \\ &= \cup_{i\in V}\{d_v - \mathbb{E}[d_i] \geq n-1-\mathbb{E}[d_v]\} \\ &\subseteq \left\{||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq \min_{v\in V}\{n-1-\mathbb{E}[d_v]\}\right\}, \end{aligned}$$

and

$$\begin{aligned} B_3 &= \cup_{(S,T)\in\Omega}\{f(S,T,\boldsymbol{d},n) = 0\} \\ &\subseteq \cup_{(S,T)\in\Omega}\{f(S,T,\boldsymbol{d},n) \leq 0\} \\ &\subseteq \cup_{(S,T)\in\Omega}\left\{||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq \frac{f(S,T,\mathbb{E}[\boldsymbol{d}],n)}{|S\cup T|}\right\} \\ &\subseteq \left\{||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq \min_{(S,T)\in\Omega}\left\{\frac{f(S,T,\mathbb{E}[\boldsymbol{d}],n)}{|S\cup T|}\right\}\right\}. \end{aligned}$$

It follows that

$$\Pr[B] \leq \Pr\left[||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq (n-1)\mathcal{E}(\mathbb{E}[\boldsymbol{d}])\right].$$

□

We proceed with the proof of the theorem. By using union bound and Hoeffding's bound,

for every $x \geq 0$,

$$\Pr\left[||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq x\right] \leq 2n \exp\left(-\frac{2x^2}{n-1} H(\boldsymbol{M})\right).$$

From this, with probability at least $1 - 2/n^{2c-1}$,

$$||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \leq \sqrt{\frac{c}{H(\boldsymbol{M})} \frac{\log(n)}{n-1}}.$$

Combining with Lemma 3.6.1, it follows that $\Pr[B] \leq 2/n^{2c-1}$ under condition

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \sqrt{\frac{c}{H(\boldsymbol{M})} \frac{\log(n)}{n-1}}. \tag{3.16}$$

By assumption, we have

$$m \leq H(\boldsymbol{M}) \binom{n}{2},$$

hence, (3.16) can be rewritten as,

$$m \geq \frac{c}{2} \frac{1}{\mathcal{E}(\mathbb{E}[\boldsymbol{d}])^2} n \log(n).$$

Proof of Corollary 3.3.7

First, note that

$$\min_{u \in V} \mathbb{E}[d_u] = \min_{u \in V} \sum_{v \in V \setminus \{u\}} p_{u,v}(\boldsymbol{\beta}) \geq \epsilon(n-1). \tag{3.17}$$

Second, note that

$$\min_{u \in V} \{n - 1 - \mathbb{E}[d_u]\} \geq \epsilon(n-1). \tag{3.18}$$

Third, note that

$$\begin{aligned}\sum_{u\in S}\mathbb{E}[d_u]-\sum_{u\in T}\mathbb{E}[d_u] \;\leq\; & \sum_{u,v\in S:u\neq v}p_{u,v}(\boldsymbol{\beta})+\sum_{u\in S,v\in\overline{S\cup T}}p_{u,v}(\boldsymbol{\beta})\\ & -\sum_{u,v\in T:u\neq v}p_{u,v}(\boldsymbol{\beta})-\sum_{u\in T,v\in\overline{S\cup T}}p_{u,v}(\boldsymbol{\beta}).\end{aligned}$$

Hence, we have

$$\begin{aligned}f(S,T,\mathbb{E}[\boldsymbol{d}],n) \;\geq\; & \sum_{u,v\in S:u\neq v}(1-p_{u,v}(\boldsymbol{\beta}))+\sum_{u\in S,v\in\overline{S\cup T}}(1-p_{u,v}(\boldsymbol{\beta}))\\ & +\sum_{u,v\in T:u\neq v}p_{u,v}(\boldsymbol{\beta})+\sum_{u\in T,v\in\overline{S\cup T}}p_{u,v}(\boldsymbol{\beta})\end{aligned}$$

and, thus

$$\begin{aligned}f(S,T,\mathbb{E}[\boldsymbol{d}],n) &\geq \epsilon|S|(n-1-|T|)+\epsilon|T|(n-1-|S|)\\ &= \epsilon[(n-1)(|S|+|T|)-2|S||T|].\end{aligned}$$

Combining this with S and T being disjoint sets, we have

$$\frac{f(S,T,\mathbb{E}[\boldsymbol{d}],n)}{|S\cup T|}\geq\epsilon\left(n-1-2\frac{|S||T|}{|S|+|T|}\right).$$

Now, since $|S|,|T|\geq 1$ and $|S|+|T|\leq n$, we have

$$\frac{|S||T|}{|S|+|T|}=\frac{1}{\frac{1}{|S|}+\frac{1}{|T|}}\leq\frac{1}{\frac{1}{|S|}+\frac{1}{n-|S|}}\leq\frac{n}{4}.$$

Thus, we have

$$f(S,T,\mathbb{E}[\boldsymbol{d}],n)\geq\epsilon\left(\frac{1}{2}n-1\right). \tag{3.19}$$

The right-hand sides in (3.17)-(3.19) are greater than or equal to $\epsilon(n-1)/4$, for all $n\geq 3$. Hence, for all $n\geq 3$,

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}])\geq\frac{1}{4}\epsilon.$$

Using this, condition in Theorem 3.3.6 holds for all $n \geq 3$, under condition

$$m \geq \frac{12c}{\epsilon^2} n \log(n).$$

Proof of Lemma 3.3.8

Fix an arbitrary set of vertices $S \subseteq V$ such that $|S| \geq k$. Note that

$$\sum_{u \in S} d_u = Z_1 + Z_2$$

where

$$Z_1 = \sum_{u \in S} \sum_{\tilde{S} \in E} \mathbb{1}_{\{\tilde{S} \subseteq S\}} \mathbb{1}_{\{u \in \tilde{S}\}} \text{ and } Z_2 = \sum_{u \in S} \sum_{\tilde{S} \in E} \mathbb{1}_{\{\tilde{S} \not\subseteq S\}} \mathbb{1}_{\{u \in \tilde{S}\}}.$$

We obviously have

$$Z_1 \leq |S| \binom{|S| - 1}{k - 1} = k \binom{|S|}{k}.$$

We next upper bound Z_2. Note that

$$Z_2 = \sum_{u \in V \setminus S} \sum_{\tilde{S} \in E} \frac{|S \cap \tilde{S}|}{|\tilde{S} \setminus S|} \mathbb{1}_{\{\tilde{S} \not\subseteq S\}} \mathbb{1}_{\{u \in \tilde{S}\}}.$$

Under $\tilde{S} \not\subseteq S$, we have

$$\frac{|S \cap \tilde{S}|}{|\tilde{S} \setminus S|} \leq k - 1.$$

Hence,

$$\sum_{\tilde{S} \in E} \frac{|S \cap \tilde{S}|}{|\tilde{S} \setminus S|} \mathbb{1}_{\{\tilde{S} \not\subseteq S\}} \mathbb{1}_{\{u \in \tilde{S}\}} \leq (k-1) \sum_{\tilde{S} \in E} \mathbb{1}_{\{\tilde{S} \not\subseteq S\}} \mathbb{1}_{\{u \in \tilde{S}\}} \leq (k-1) d_u.$$

We use the notation $\mathcal{S}_{n,k} = \{S \subseteq V : |S| = k\}$. We have

$$\begin{aligned}
\sum_{\tilde{S}\in E} \frac{|S\cap\tilde{S}|}{|\tilde{S}\setminus S|}\mathbb{1}_{\{\tilde{S}\not\subseteq S\}}\mathbb{1}_{\{u\in\tilde{S}\}} &\leq \sum_{\tilde{S}\in\mathcal{S}_{n,k}} \frac{|S\cap\tilde{S}|}{|\tilde{S}\setminus S|}\mathbb{1}_{\{\tilde{S}\not\subseteq S\}}\mathbb{1}_{\{u\in\tilde{S}\}} \\
&= \sum_{s=1}^{k-1} \frac{s}{k-s} \sum_{\tilde{S}\in\mathcal{S}_{n,k}:|S\cap\tilde{S}|=s} \mathbb{1}_{\{u\in\tilde{S}\}} \\
&= \sum_{s=1}^{k-1} \frac{s}{k-s}\binom{|S|}{s}\binom{n-|S|-1}{k-s-1} \\
&= \frac{|S|}{n-|S|}\sum_{s=1}^{k-1}\binom{|S|-1}{s-1}\binom{n-|S|}{k-s} \\
&= \frac{|S|}{n-|S|}\sum_{s=0}^{k-2}\binom{|S|-1}{s}\binom{n-|S|}{k-1-s} \\
&= \frac{|S|}{n-|S|}\left(\binom{n-1}{k-1} - \binom{|S|-1}{k-1}\right).
\end{aligned}$$

Hence,

$$Z_2 \leq \sum_{u\in V\setminus S} \min\left\{(k-1)d_u, \frac{|S|}{n-|S|}\left(\binom{n-1}{k-1} - \binom{|S|-1}{k-1}\right)\right\}.$$

Proof of Lemma 3.3.9

We first show a key step for proving the lemma. The goal is to show the existence of a sufficiently large set that contains vertices with sufficiently large values of parameters, as stated in the following lemma.

Lemma 3.6.2. *Assume $d = (d_1, \ldots, d_n)$ is the expected degree sequence of a k-uniform random hypergraph $H = (V, E)$ with $|V| = n$ satisfying (3.10) and $k < (1-\alpha_1)n$, for constants $\alpha_1, \alpha_2 \in (0,1)$. Then, if $||\hat{\boldsymbol{\beta}}||_\infty \geq C(\alpha_1, \alpha_2, k)$, there exists a set $S \subseteq V$ such that $|S| \geq \alpha_1 n$ and $\hat{\beta}_v \geq ||\hat{\boldsymbol{\beta}}||_\infty/k^2$ for all $v \in S$.*

Proof. Without loss of generality, assume that $\hat{\beta}_{max} := \max_{v\in V}\hat{\beta}_v > 0$. Consider the set $\bar{S} = \{v \in V : \hat{\beta}_v > -\hat{\beta}_{max}/k\}$. Let m be the cardinality of $\bar{S}$. We will next show that $m < n$. If $m < k$, then $m < n$ obviously holds. Hence, we assume $m \geq k$. By the moment

equations, we have

$$d_{max} := \max_{v \in V} d_v \geq d_{v^*} \geq \binom{m-1}{k-1} \sigma(\hat{\beta}_{max}/k)$$

where v^* denotes the index maximizing d_v. As $d_{max} \leq (1-\alpha_2)\binom{n-1}{k-1}$, this implies

$$\binom{n-1}{k-1} - \binom{m-1}{k-1} > \binom{n-1}{k-1} \left[1 - (1-\alpha_2)(1+e^{-\hat{\beta}_{max}/k})\right].$$

Thus, if $\hat{\beta}_{max} > C(\alpha_2, k)$, then we have $m < n$.

Under $m < n$, the set $V \setminus \bar{S}$ is non empty. Fix $u \in V \setminus \bar{S}$. Consider the set $\underline{S}_u = \{v \in V : \hat{\beta}_v < -\hat{\beta}_i/k\}$. Let m_u denote the cardinality of $\underline{S}_u$. We next show that $m_u < n$. Since we assumed $k < (1-\alpha_1)n$, if $m_u < k-1$ then $m_u < (1-\alpha_1)n$ obviously holds. Hence, we assume $m_u \geq k-1$.

By the moment equations, we have

$$d_{min} := \min_{v \in V} d_v \leq d_u < \binom{m_u}{k-1} \sigma(\hat{\beta}_{max}/k) + \binom{n-1}{k-1} - \binom{m_u}{k-1}.$$

As $d_{min} \geq \alpha_1 \binom{n-1}{k-1}$, this implies

$$\binom{m_u}{k-1} < \binom{n-1}{k-1} (1-\alpha_1) \left(1 + e^{-\hat{\beta}_{max}/k}\right).$$

Using the bounds $\left(\frac{n}{i}\right)^i \leq \binom{n}{i} \leq \left(\frac{en}{i}\right)^i$, it follows

$$m_u < e[(1-\alpha_1)(1+e^{-\hat{\beta}_{max}/k})]^{\frac{1}{k-1}} n.$$

If $\hat{\beta}_{max} > C(\alpha_1, k)$, then we have $m_u < (1-\alpha_1)n$.

By assumption $u \in V/\bar{S}$ and definition of $\underline{S}_u$, there are at least $n - m_u$ vertices $v \in V$ such that $\hat{\beta}_v > \hat{\beta}_{max}/k^2$. Hence, if $\hat{\beta}_{max} > C(\alpha_1, \alpha_2, k)$, then there are at least $\alpha_1 n$ vertices $v \in V$ such that $\hat{\beta}_v > \hat{\beta}_{max}/k^2$. □

Assume $\hat{\beta}_{max} := \max_{v \in V} \hat{\beta}_v > 0$. By Lemma 3.6.2, if $\hat{\beta}_{max} > C(\alpha_1, \alpha_2, k)$, there exists a set $S^\star$ of cardinality $|S^\star| \geq \alpha_1 n$ such that $\hat{\beta}_v \geq \hat{\beta}_{max}/k^2$ for all $v \in S^\star$. Hence,

$$\sum_{S' \subseteq S^\star : |S'| = k} p_{S'}(\hat{\beta}) \geq \binom{|S^\star|}{k} \sigma(\hat{\beta}_{max}/k).$$

By taking $\hat{\beta}_{max}$ large enough, we obtain a contradiction with (3.11). Hence, $\hat{\boldsymbol{\beta}}$ must be such that $||\hat{\boldsymbol{\beta}}||_\infty \leq C(\alpha_1, \alpha_2, \alpha_3, k)$.

The case when $\hat{\beta}_{max} \leq 0$ follows by the same arguments by considering complements of experiment outcomes and the reparametrization $\hat{\boldsymbol{\beta}}' := -\hat{\boldsymbol{\beta}}$.

Proof of Theorem 3.3.10

Let us define the following events:

$$\begin{aligned}
B_1 &= \cup_{v \in V} \left\{ d_v \leq \alpha_1 \binom{n-1}{k-1} \right\} \\
B_2 &= \cup_{v \in V} \left\{ d_v \geq (1-\alpha_2) \binom{n-1}{k-1} \right\} \\
B_3 &= \cup_{S \subseteq V : |S| \geq \alpha_1 n} \left\{ \sum_{S' \subseteq S : |S'| = k} y_{S'} \leq \alpha_3 \binom{|S|}{k} \right\} \\
B_4 &= \cup_{S \subseteq V : |S| \geq \alpha_1 n} \left\{ \sum_{S' \subseteq S : |S'| = k} y_{S'} \geq (1-\alpha_4) \binom{|S|}{k} \right\}.
\end{aligned}$$

Assume that for all $S \subseteq V$ such that $|S| \geq \alpha_1 n$,

$$\max\{\alpha_1, \alpha_3\} \leq \frac{1}{\binom{|S|}{k}} \sum_{S' \subseteq S : |S'| = k} p_{S'}(\boldsymbol{\beta}) \leq \min\{1-\alpha_2, 1-\alpha_4\}.$$

By Hoeffding's bound, we have:

$$\begin{aligned}
\Pr[B_1] &\leq n \exp\left(-2\binom{n-1}{k-1}\left(\min_{i\in n}\left\{\frac{1}{\binom{|S|}{k}}\sum_{S'\subseteq V:|S'|=k,i\in S'} p_{S'}(\boldsymbol{\beta}) - \alpha_1\right\}\right)^2\right) \\
\Pr[B_2] &\leq n \exp\left(-2\binom{n-1}{k-1}\left(\min_{i\in n}\left\{1-\alpha_2 - \frac{1}{\binom{|S|}{k}}\sum_{S'\subseteq V:|S'|=k,i\in S'} p_{S'}(\boldsymbol{\beta})\right\}\right)^2\right) \\
\Pr[B_3] &\leq 2^n \max_{S\subseteq V:\alpha_1 n\leq|S|\leq n} \exp\left(-2\binom{|S|}{k}\left(\alpha_3 - \frac{1}{\binom{|S|}{k}}\sum_{S'\subseteq S:|S'|=k} p_{S'}(\boldsymbol{\beta})\right)^2\right) \\
\Pr[B_4] &\leq 2^n \max_{S\subseteq V:\alpha_1 n\leq|S|\leq n} \exp\left(-2\binom{|S|}{k}\left(1-\alpha_4 - \frac{1}{\binom{|S|}{k}}\sum_{S'\subseteq S:|S'|=k} p_{S'}(\boldsymbol{\beta})\right)^2\right).
\end{aligned}$$

Assuming $\epsilon \leq p_S(\boldsymbol{\beta}) \leq 1-\epsilon$, for all $S \subseteq V$ with $|S| = k$, we have

$$\begin{aligned}
\Pr[B_1] &\leq n \exp\left(-2(\epsilon-\alpha_1)^2\left(\frac{n}{k}\right)^{k-1}\right) \\
\Pr[B_2] &\leq n \exp\left(-2(\epsilon-\alpha_2)^2\left(\frac{n}{k}\right)^{k-1}\right) \\
\Pr[B_3] &\leq 2^n \exp\left(-2(\epsilon-\alpha_3)^2\left(\frac{\alpha_1 n}{k}\right)^{k}\right) \\
\Pr[B_4] &\leq 2^n \exp\left(-2(\epsilon-\alpha_4)^2\left(\frac{\alpha_1 n}{k}\right)^{k}\right)
\end{aligned}$$

under condition $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \leq \epsilon$.

Taking $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \epsilon/2$, we have

$$\begin{aligned}
\Pr[B_1] &\leq n \exp\left(-\frac{1}{2}\epsilon^2\left(\frac{n}{k}\right)^{k-1}\right) \\
\Pr[B_2] &\leq n \exp\left(-\frac{1}{2}\epsilon^2\left(\frac{n}{k}\right)^{k-1}\right) \\
\Pr[B_3] &\leq 2^n \exp\left(-\frac{1}{2}\epsilon^2\left(\frac{n\epsilon}{2k}\right)^{k}\right) \\
\Pr[B_4] &\leq 2^n \exp\left(-\frac{1}{2}\epsilon^2\left(\frac{n\epsilon}{2k}\right)^{k}\right).
\end{aligned}$$

By union bound, for event B defined by $B = B_1 \cup B_2 \cup B_3 \cup B_4$, we have $\Pr[B] \leq \Pr[B_1] + \Pr[B_2] + \Pr[B_3] + \Pr[B_4]$. For $\Pr[B] \leq 1/n^c$ to hold, for a positive constant $c > 0$, it suffices that

$$n \exp\left(-\frac{1}{2}\epsilon^2 \left(\frac{n}{k}\right)^{k-1}\right) \leq \frac{1}{4n^c} \tag{3.20}$$

and

$$2^n \exp\left(-\frac{1}{2}\epsilon^2 \left(\frac{n\epsilon}{2k}\right)^{k}\right) \leq \frac{1}{4n^c}. \tag{3.21}$$

Equation (3.20) is equivalent to

$$\epsilon \geq \sqrt{2} k^{\frac{k-1}{2}} \sqrt{\frac{(c+1)\log(n) + \log(4)}{n^{k-1}}}.$$

Equation (3.21) is equivalent to

$$\epsilon \geq 2^{\frac{k+1}{k+2}} k^{\frac{k}{k+2}} \left(\frac{\log(2)}{n^{k-1}} + \frac{c\log(n) + \log(4)}{n^k}\right)^{\frac{1}{k+2}}.$$

Proof of Proposition 3.4.1

We first establish the following lemma:

Lemma 3.6.3. *If* $\min_{\beta' \in [\hat{\beta}, \beta]} \lambda_1(\nabla^2(-\ell(\beta'))) > 0$, *then*

$$||\hat{\beta} - \beta|| \leq \frac{2||\nabla \ell(\beta)||}{\min_{\beta' \in [\hat{\beta}, \beta]} \lambda_1(\nabla^2(-\ell(\beta')))}.$$

Proof. Let f denote the negative log-likelihood function and $\Delta = \hat{\beta} - \beta$. By limited Taylor expansion, we have

$$f(\hat{\beta}) \geq f(\beta) + \nabla f(\beta)^\top \Delta + \frac{1}{2} \min_{\lambda \in [0,1]} \Delta^\top \nabla^2 f(\beta + \lambda\Delta)\Delta.$$

Combining with $f(\hat{\boldsymbol{\beta}}) \leq f(\boldsymbol{\beta})$, we have

$$\min_{\lambda \in [0,1]} \boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{\beta} + \lambda \boldsymbol{\Delta}) \boldsymbol{\Delta} \leq -2 \nabla f(\boldsymbol{\beta})^\top \boldsymbol{\Delta}.$$

Next, note

$$\min_{\lambda \in [0,1]} \boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{\beta} + \lambda \boldsymbol{\Delta}) \boldsymbol{\Delta} \geq \min_{\lambda \in [0,1]} \lambda_1(f(\boldsymbol{\beta} + \lambda \boldsymbol{\Delta}))||\boldsymbol{\Delta}||^2.$$

Hence,

$$\min_{\lambda \in [0,1]} \lambda_1(f(\boldsymbol{\beta} + \lambda \boldsymbol{\Delta}))||\boldsymbol{\Delta}||^2 \leq -2 \nabla f(\boldsymbol{\beta})^\top \boldsymbol{\Delta}.$$

By Cauchy-Schwartz inequality, $|-\nabla f(\boldsymbol{\beta})^\top \boldsymbol{\Delta}| \leq ||\nabla f(\boldsymbol{\beta})||||\boldsymbol{\Delta}||$, hence,

$$\min_{\lambda \in [0,1]} \lambda_1(f(\boldsymbol{\beta} + \lambda \boldsymbol{\Delta}))||\boldsymbol{\Delta}|| \leq 2||\nabla f(\boldsymbol{\beta})||.$$

□

For every $\boldsymbol{\beta}' \in \mathbb{R}^n$, the gradient vector of the log-likelihood function is given by

$$\nabla \ell(\boldsymbol{\beta}') = \sum_{j=1}^{m} \left(y_j - \frac{1}{1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}'}} \right) \boldsymbol{x}_j \tag{3.22}$$

and the Hessian matrix of the log-likelihood function is given by

$$\nabla^2(-\ell(\boldsymbol{\beta}')) = \sum_{j=1}^{m} \frac{e^{\boldsymbol{x}_j \boldsymbol{\beta}'}}{(1 + e^{\boldsymbol{x}_j \boldsymbol{\beta}'})^2} \boldsymbol{x}_j \boldsymbol{x}_j^\top. \tag{3.23}$$

We next show that under the overlapping condition,

$$||\nabla \ell(\boldsymbol{\beta})|| \leq \frac{1}{1 + e^{-bk}} \sqrt{2mk(\log(n) + 2)} \tag{3.24}$$

with probability larger than or equal to $1 - 2/(n \Pr[||\hat{\boldsymbol{\beta}}||_\infty \leq b])$.

From (3.22), note that $\nabla \ell(\boldsymbol{\beta})$ is the sum of independent random vectors $\boldsymbol{z}_j = (y_j -$

$\sigma(\boldsymbol{x}_j^\top \boldsymbol{\beta}))\boldsymbol{x}_j$ for $j = 1, \ldots, m$ satisfying

$$\mathbb{E}[\boldsymbol{z}_j] = 0 \text{ and } ||\boldsymbol{z}_j|| \leq \frac{1}{1+e^{-bk}}\sqrt{k}.$$

Using this in Azuma-Hoeffding's inequality (Lemma 2.3.3), we have

$$\Pr\left[||\nabla \ell(\boldsymbol{\beta})|| > \frac{1}{1+e^{-bk}}\sqrt{2mk(\log(n)+2)}\right] \leq \frac{2}{n}.$$

Combining this with

$$\begin{aligned} & \Pr\left[||\nabla \ell(\boldsymbol{\beta})|| > \frac{1}{1+e^{-bk}}\sqrt{2mk(\log(n)+2)} \mid ||\hat{\boldsymbol{\beta}}||_\infty \leq b\right] \\ \leq & \frac{\Pr\left[||\nabla \ell(\boldsymbol{\beta})|| > \frac{1}{1+e^{-bk}}\sqrt{2mk(\log(n)+2)}\right]}{\Pr[||\hat{\boldsymbol{\beta}}||_\infty \leq b]} \end{aligned}$$

we prove the statement in (3.24).

We next show that

$$\min_{\boldsymbol{\beta}' \in [\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}]} \lambda_1(\nabla^2(-\ell(\boldsymbol{\beta}))) \geq \frac{e^{kb}}{(1+e^{kb})^2}\lambda_1(\boldsymbol{M}). \tag{3.25}$$

Since $||\hat{\boldsymbol{\beta}}||_\infty \leq b$ and $||\boldsymbol{\beta}||_\infty \leq b$, for every $\boldsymbol{\beta}' \in [\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}]$, $|\boldsymbol{x}_j^\top \boldsymbol{\beta}'| \leq kb$ for all $j \in \{1, \ldots, m\}$, and

$$\frac{e^{\boldsymbol{x}_j \boldsymbol{\beta}'}}{(1+e^{\boldsymbol{x}_j \boldsymbol{\beta}'})^2} \geq \frac{e^{kb}}{(1+e^{kb})^2} \text{ for all } \boldsymbol{\beta}' \in [\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}] \text{ and } j \in \{1, \ldots, m\}.$$

It follows that

$$\nabla^2(-\ell(\boldsymbol{\beta}')) \succeq \frac{e^{kb}}{(1+e^{kb})^2}\boldsymbol{M} \text{ for all } \boldsymbol{\beta}' \in [\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}]$$

from which (3.25) follows.

The statement of Lemma 3.4.1 follows from Lemma 3.6.3, (3.24) and (3.25).

Proof of Theorem 3.4.2

Let $\boldsymbol{x}$ be the normalized eigenvector ($||\boldsymbol{x}|| = 1$) corresponding to the smallest eigenvalue of $\boldsymbol{M}$. We can write

$$\lambda_1(\boldsymbol{M}) = \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{x} - \boldsymbol{x}^\top(\boldsymbol{D}-\boldsymbol{N})\boldsymbol{x}.$$

By definition the smallest eigenvalue, $\boldsymbol{x}^\top(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{x} \geq \lambda_1(\boldsymbol{D}+\boldsymbol{A})$. By (3.15), $\lambda_1(\boldsymbol{D}+\boldsymbol{A}) \geq \psi^2/(4d^*)$. Since $\boldsymbol{D}-\boldsymbol{N}$ is a diagonal matrix, we have $\boldsymbol{x}^\top(\boldsymbol{D}+\boldsymbol{A})\boldsymbol{x} \leq \max_u D_{u,u}$. This proves the lower bound.

Similarly, we consider $\lambda_1(\boldsymbol{D}+\boldsymbol{A})$ and bound it as follows:

$$\lambda_1(\boldsymbol{D}+\boldsymbol{A}) = \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^\top(\boldsymbol{D}-\boldsymbol{N})\boldsymbol{x} \geq \lambda_1(\boldsymbol{M}) + \min_{u\in V} D_{u,u}.$$

Again, by (3.15), $\lambda_1(\boldsymbol{D}+\boldsymbol{A}) \leq 4\psi$. This proves the upper bound.

Chapter 4

The β-model with random design

4.1 Overview

The experimental design is important to our analysis. In the previous chapter, we have discussed about the accuracy of MLE and how the experiment design affects it. So far, our results are for β-model with fixed design of experiments, i.e. the design matrix X is assumed to be fixed. However, in real life settings, we may have limited resources for experiments and our designs may not be regular or complete. Consider the testing phase of a new game. We have limited number of volunteers and we do not allow too many rounds of play. In order to gather as much information as possible, we may consider grouping players randomly and uniformly into subsets of fixed size. It is important to know the threshold number of experiments that guarantee the estimation accuracy for model parameters under such random design cases.

In this chapter, we consider the beta model of random hypergraphs with random design matrix $\boldsymbol{X}$. We add another level of randomness and consider the setting where the underlying design matrix corresponds to a k-uniform random hypergraph, where experiments are conducted for edges drawn by sampling with replacement from the set of all combinations of k vertices of n vertices.

Our model formulation is the similar to the fixed design case in the previous chapter. Let $V = \{1, \ldots, n\}$ be a set of vertices with $n \geq 2$. The beta model assigns individual $\beta_i \in \mathbb{R}$ for each vertex i and constructs a random hypergraph by putting a hyperedge y_j independently for a group of nodes $S_j \subseteq V$ with probability

$$\Pr[y_j = 1] = 1 - \Pr[y_j = 0] = \sigma\left(\sum_{i \in S_j} \beta_i\right)$$

We say that $\boldsymbol{X} \in \{0,1\}^{m \times n}$ is the design matrix. We also define the correlation matrix $\boldsymbol{M}$ s.t.

$$\boldsymbol{M} = \sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^\top = \boldsymbol{X}^\top \boldsymbol{X}.$$

Previously, we assumed that $\boldsymbol{X}$ is fixed. Now we consider random design matrix $\boldsymbol{X}$ with independent rows sampled with replacement from the set of vectors $\{\boldsymbol{x} \in \{0,1\}^n : ||\boldsymbol{x}||_1 = k\}$. We are interested in the conditions that are necessary and sufficient for $\boldsymbol{X}$ to be of full rank. As for the fixed design case, we also study the MLE conditions and MLE error bounds in this random design setting. The learning problem is more challenging due to the randomness. We note that the correlation matrix $\boldsymbol{M}$ does not conform to the classical definition of adjacency matrix, as its diagonal elements are dependent on its off-diagonal elements. Due to this complicated dependency structure, existing results on random matrices cannot be directly applied to our setting.

4.1.1 Related work

The β-model of random hypergraphs is a logistic regression model with covariate vectors $\boldsymbol{x}$ such that $\boldsymbol{x} \in \{0,1\}^n$ and $||\boldsymbol{x}||_1 = k$. In this view, it is worth mentioning some recent work on statistical inference for high-dimensional logistic regression models. Candés and Sur (2020) established a sharp phase transition threshold for the MLE of a logistic regression model with Gaussian covariates, using the framework of convex geometry (Amelunxen et al. (2014)). Candés and Sur (2020) showed that for logistic regression with

independent covariate vectors with dimension that is a constant-factor of the number of observations, the MLE is biased and has greater variability than suggested by classical estimation theory. Salehi et al. (2019) extended these results to high-dimensional logistic regression models with regularization. Note that the randomness results from the Gaussian covariates. An important property used in their analysis is the rotational invariance of Gaussian random variables, which does not apply in our setting.

The necessary condition is that $\boldsymbol{X}$ does not have a null-column almost surely, which can be reduced to the *coupon subset selection problem* (Stadje (1990); Adler and Ross (2001)). It is much more challenging to derive the sufficient condition. There are few existing work (Costello and Vu (2008, 2010); Cooley et al. (2018, 2019); Karoński and Łuczak (2002)) on deriving full rank condition for adjacent matrices of random graphs. In particular, Costello and Vu (2008) considers the adjacency matrix of Erdő-Rényi $G(n, p)$ random graphs and further extend it to a class of symmetric sparse matrices. They have developed a generalized framework for identifying full rank conditions which we will follow for our analysis. However, our setting does not conform to the definition of adjacency matrices and requires new proofs and techniques in the analysis.

4.1.2 Summary of contributions

Our results can be summarized in the following points.

- We established new results for the β-model of random hypergraphs with a random design matrix $\boldsymbol{X}$, which has independent rows sampled with replacement from the set of vectors $\{\boldsymbol{x} \in \{0,1\}^n : ||\boldsymbol{x}||_1 = k\}$. We prove a *necessary* condition for the $\boldsymbol{X}$ to have full rank almost surely, i.e. $\text{rank}(\boldsymbol{X}) = n$. This is a sharp threshold for $\boldsymbol{X}$ to not have a null-column almost surely, which we established by a reduction to the *coupon subset selection problem*. We also conjectured the following *sufficient* condition

for $\boldsymbol{X}$ to have full rank almost surely:

$$m \geq \max\left\{c\frac{1}{k}n\log(n), n\right\}$$

for any fixed constant $c > 2$, under assumption $2 \leq k = o(n/\log(n))$. This sufficient condition is tight in the sense of being within a factor two of the *necessary* condition $m \geq c(n/k)\log(n)$, for a fixed constant $c > 1$. We have tested numerically to show that this condition empirically holds. We give a partial proof for the sufficient condition, which is established by a framework for deriving full rank conditions for adjacency matrices of random graphs Costello and Vu (2008, 2010). The proof required new results to accommodate the class of random matrices we consider in our case. Our results may be of an independent interest for the line of work on the rank of random matrices.

- We found a sufficient condition for the MLE existence and uniqueness for the β-model of random graphs with random design matrices. Specifically, for any β-model with parameter $\boldsymbol{\beta}$ such that $\epsilon \leq p_{u,v}(\boldsymbol{\beta}) \leq 1-\epsilon$ for all $u, v \in V$ and $u \neq v$, for some $\epsilon \in (0,1)$, there exists a unique MLE with high probability provided that the number of experiments is $\Omega(\frac{1}{\epsilon}n^{5/4}(\log(n))^{1/4})$.

- On the applications side, we believe that our results provide useful insights into statistical inference of β-model of random hypergraphs. Many relational data can be represented by hypergraphs, where data entities are represented by vertices and their group responses are represented by edges (sets of vertices). Our results provide theoretical guarantees for the MLE estimation under real-life settings where experiment resources are limited.

Organization of chapter The chapter is organized as follows. We explore the condition for the random design matrix to have full rank in section 4.2. The full rank condition is necessary for the MLE uniqueness when the MLE exists and of interest for bounding the parameter estimation error. Then, in section 4.3 we consider the MLE conditions and the

MLE error bounds for the random design.

4.2 Rank of the design matrices

Since $\boldsymbol{X}$ is of full rank if, and only if, $\lambda_1(\boldsymbol{M}) > 0$, we focus our attention to finding conditions under which $\lambda_1(\boldsymbol{M}) > 0$ with high probability. For every pair of vertices (u, v) such that $u \neq v$, $M_{u,v}$ is the number of experiments involving both u and v. Hence, for a random design matrix, we have

$$\mathbb{E}[M_{u,v}] = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} m = \frac{k(k-1)}{n(n-1)} m.$$

Similarly, for every vertex u, $M_{u,u}$ is the number of experiments involving vertex u. Hence, we have

$$\mathbb{E}[M_{u,u}] = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} m = \frac{k}{n} m.$$

It can be readily shown that

$$\lambda_1(\mathbb{E}[\boldsymbol{M}]) = \frac{k(n-k)}{n(n-1)} m. \tag{4.1}$$

4.2.1 Necessary condition

The first necessary condition asserted in the following proposition is derived by establishing a condition for $\lambda_1(\boldsymbol{M}) > 0$ to hold with high probability, using a concentration of measure for sums of random matrices.

Proposition 4.2.1. *Assume $k = o(n)$. For every $a > 0$ and $\epsilon \in (0, 1]$, there exists a constant $c_{a,\epsilon} > 0$ such that if*

$$m > c_{a,\epsilon} n \log(n) \tag{4.2}$$

then, with probability at least $1 - 1/n^a$,

$$\lambda_1(\boldsymbol{M}) \geq (1-\epsilon)\lambda_1(\mathbb{E}[\boldsymbol{M}]).$$

In particular, we can take $c_{a,\epsilon} = 2(a+1)/\epsilon^2$.

Proposition 4.2.1 has the following two implications. First, it implies that $\boldsymbol{X}$ has full rank with probability at least $1 - 1/n^a$ provided that $k \leq cn$ for some fixed $c \in (0,1)$ and the number of experiments is $\Omega(n\log(n))$. Second, for the MLE error bound in Proposition 3.4.1 when $||\hat{\boldsymbol{\beta}}||_\infty \leq b$ with probability 1, if $m > \frac{4}{\epsilon^2} n\log(n)$, then with probability at least $1 - 3/n$,

$$\frac{1}{\sqrt{n}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \leq \frac{c_{kb}}{1-\epsilon}\sqrt{\frac{n(\log(n)+2)}{km}}. \tag{4.3}$$

This implies $\frac{1}{n}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 = O(1)$ with high probability, provided that $k \leq cn$ for some fixed $c \in (0,1)$ and the number of experiments is $\Omega(n\log(n)/k)$.

We next present a sharper necessary condition for the correlation matrix $\boldsymbol{M}$ of a random design matrix $\boldsymbol{X}$ to have full rank, in terms of the constant factor and its dependency on the size of experiments k. A necessary condition for $\boldsymbol{M}$ to have full rank is that $\boldsymbol{X}$ does not have a null column. We first provide a tight condition for $\boldsymbol{X}$ to have non-null columns with high probability.

Theorem 4.2.2. *For any* $a > 0$, *a random design matrix* $\boldsymbol{X}$ *with* m *rows and* n *columns, with each row having* k *elements equal to* 1 *and other elements equal to* 0, *has no null column with probability at least* $1 - 1/n^a$, *if*

$$m \geq (1+a)\frac{1}{k} n\log(n). \tag{4.4}$$

Moreover, this condition is tight in the sense that if $k = o(n/\log(n))$, *then there exists a sequence* c_n *such that* $c_n = O(1/\log(n))$, *so that for* $m = (1+c_n)\frac{1}{k} n\log(n)$, $\boldsymbol{X}$ *has a null column almost surely.*

The sufficiency of (4.4) is straightforward to establish by using union bound and the probability of the event that an arbitrarily fixed column is null. The necessity of (4.4) is more

intricate and follows from the solution of the coupon set selection problem, where each edge experiment is seen as drawing a set of k coupons uniformly at random with replacement from the set of n distinct coupons.

We illustrate the condition of Theorem 4.2.1 by the following numerical example.

Example 4.2.1. *We randomly sample design matrices $\boldsymbol{X}$ by drawing m independent rows from from the set of vectors with $\{0,1\}$-valued entries with exactly k entries equal to 1. For each such random design matrix $\boldsymbol{X}$, we check whether $\boldsymbol{X}$ has a null-column. We repeat this for a set number of independent repetitions to evaluate the fraction of instances for which $\boldsymbol{X}$ does not have a null-column. We report numerical results for $n = 100$ and the number of repetitions equal to 1000.*

In Figure 4.1 we show the fraction of realizations of matrix $\boldsymbol{X}$ with no null-column. The solid line shows the probability that $\boldsymbol{X}$ has a null-column that follows from the coupon collector problem. The dots are results of empirical evaluations.

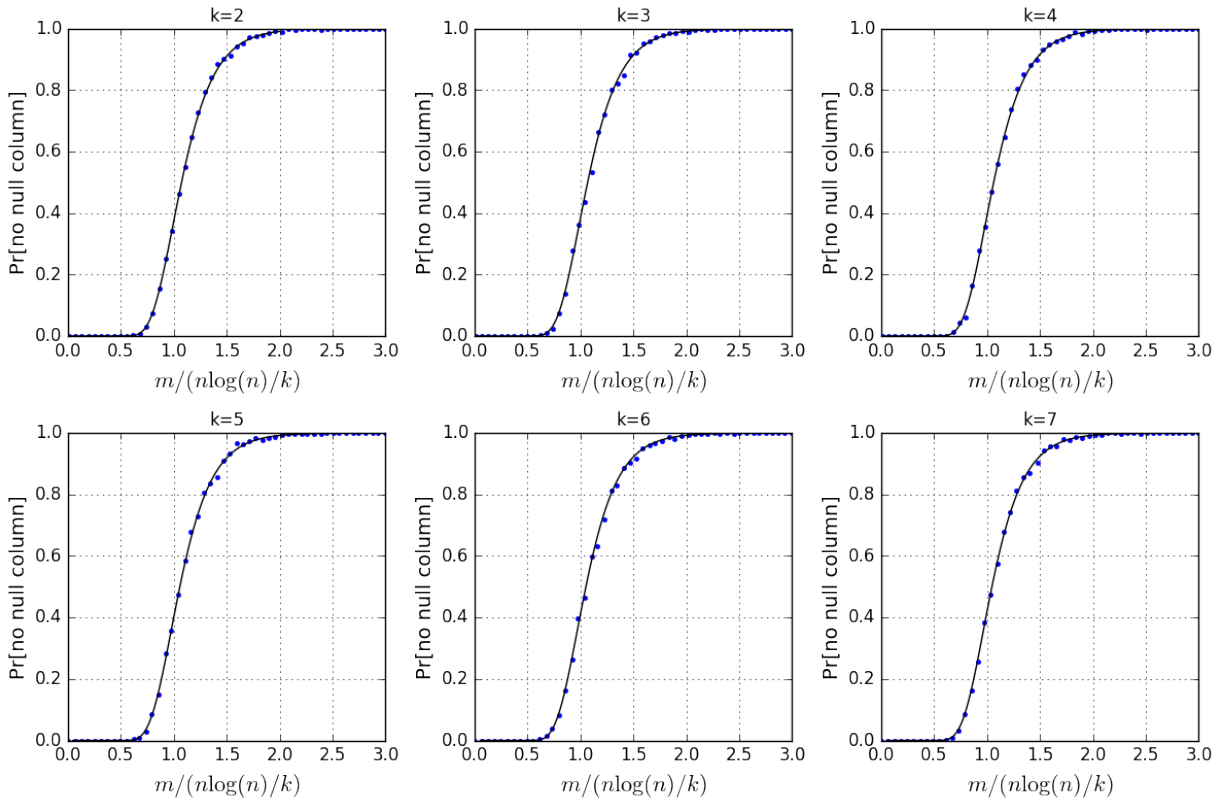

Figure 4.1: Estimated probability for matrix $\boldsymbol{X}$ having a null column versus the normalized number of experiments, for different values of parameter k.

4.2.2 Sufficient condition

We next present a conjecture that identifies a *sufficient* number of experiments for $\boldsymbol{X}$ to have full rank. Proving this conjecture is of interest as it provides a tight sufficient condition, which is within a factor of two of the necessary condition in Theorem 4.2.2.

Conjecture 4.2.3. *Consider a random design matrix $\boldsymbol{X}$ with m rows and n columns such that $m \geq n$, with each row drawn independently with replacement from the set of vectors $\{\boldsymbol{x} \in \{0,1\}^n : ||\boldsymbol{x}||_1 = k\}$, with $k \geq 2$ and $k = o(n/\log(n))$. Under the given assumptions, if the number of rows m is such that*

$$m \geq c\frac{1}{k}n\log(n)$$

for any fixed constant $c > 2$, then $\boldsymbol{X}$ (and $\boldsymbol{M}$ equivalently) has full rank with probability at least $1 - O(1/(\log(\log(n)))^{1/4})$.

We tested numerically by the following example.

Example 4.2.2. *We tested whether $\boldsymbol{X}$ has full rank for independent samples of $\boldsymbol{X}$. We fixed $n =$ 100 and the number of samples to* 1000*, and varied the number of experiments m. In Figure 4.2 we show the fraction of instances for which $\boldsymbol{X}$ has full rank for different values of m.*

We conclude that the results agree with our conjecture from Figure 4.2, which indicate that the threshold number of experiments m for $\boldsymbol{X}$ to have full rank is $\max\{c(n/k)\log(n), n\}$, for a constant c greater than 1. Moreover, the results indicate that $\boldsymbol{X}$ has full rank almost surely, if $m \geq \max\{c(n/k)\log(n), n\}$, for any fixed constant $c > 2$. All these results conform to the necessary condition in Theorem 4.2.2 and the sufficient condition in Conjecture 4.2.3. Note that for larger values of k there is a sharp phase transition. The reason is that for $\boldsymbol{X}$ to have rank n, it is necessary that $m \geq n$. Hence, if $m = cn\log(n)/k$, then it must hold $c \geq k/\log(n)$, which fails to hold for large enough values of k in our numerical example.

We propose to prove by following the same framework as in the work by Costello and Vu (2008). Specifically, it is based on analysis of a graph growth process defined by adding

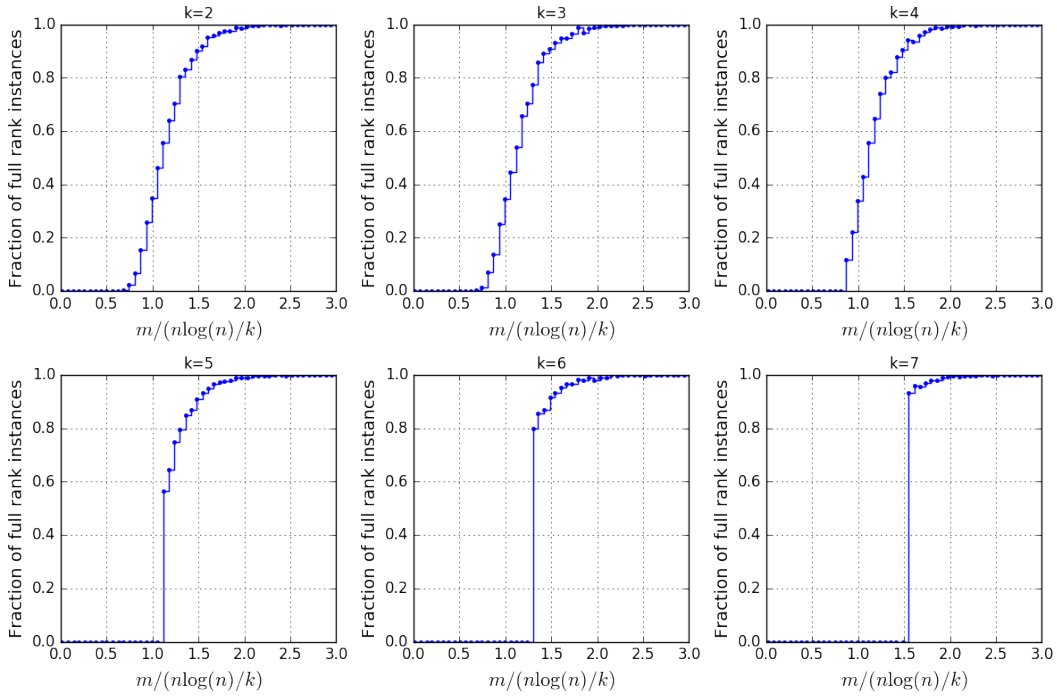

Figure 4.2: Estimated probability of $\boldsymbol{X}$ having full rank versus the normalized number of experiments, for different values of parameter k.

vertices one at a time, following a similar approach used for identifying full rank conditions for adjacency matrices of random graphs. This amounts to studying a sequence of correlation matrices that converges to correlation matrix $\boldsymbol{M}$. The major difference between our setting and the setting in Costello and Vu (2008) is that the correlation matrix $\boldsymbol{M}$ does not conform to the definition of adjacency matrices, as the diagonal elements of $\boldsymbol{M}$ are dependent on its off-diagonal elements. Specifically, $\boldsymbol{M}$ is a signless Laplacian matrix, with diagonal element of a row equal to the sum of off-diagonal elements of this row.

However, proving the conjecture appears to be tricky due to this extra dependency structure of the correlation matrix $\boldsymbol{M}$. Unfortunately, as of yet we were unable to prove it exactly. In the following paragraphs, we will list key properties of interest and outline the major steps of the general framework for proving the conjecture. Detailed proofs and discussions can be found in section 4.5.

Key properties The key property we discovered for the matrix $\boldsymbol{M}$ and its corresponding graph $\mathcal{G}$ is negative association, which is a special form of dependence that allows many useful properties of independence to carry over.

Recall the following definition of negative association mentioned in section 2.3.

Definition 4.2.4. *A collection of random variables $Y = (Y_1, Y_2, \ldots, Y_n)$ is said to be negatively associated if for disjoint index sets $I, J \subseteq [n]$ and two functions f and g both monotone increasing or both monotone decreasing,*

$$\mathrm{Cov}\left(f(X_i, i \in I), g(X_j, j \in J)\right) \leq 0$$

We first note a simple fact of the random design matrix $\boldsymbol{X}$.

Lemma 4.2.5. *For each experiment t, RVs $\{X_{tu}, u \in V\}$ are NA.*

Note that we can write $M_{uv} = \sum_t X_{tu}X_{tv}$ where $(u, v) \in E$. Intuitively, if some edges are chosen, each of the others is less likely to be chosen. The next lemma proves the negative dependency structure of the RVs $\{M_{uv}, (u, v) \in E\}$.

Lemma 4.2.6. *The RVs $\{M_{uv}, (u, v) \in E\}$ are NA.*

The key point of the proof uses a common property first derived by Feder and Mihail (1992) for balanced matroid. Specifically, any monotone property m over the variables in a set $S \backslash \{e\}$ is negatively correlated with e.

Next, we show that the NA property can be transferred to the corresponding graph $\mathcal{G}$. Take a set $S \subset V$ and denote $d_u(S) = \sum_{v \in S} \mathbb{1}\{M_{uv} > 0\}$ as the number of distinct neighbors of vertex u in the set S for all $u \in V \setminus S$.

Lemma 4.2.7. *The RVs $\{d_u(S), u \in V \setminus S\}$ are NA.*

We also specify the following property of the set of vertices.

Definition 4.2.8 (nice set). *For any graph $G = (V, E)$, set over vertices $S \subseteq V$ is said to be* nice *if there are at least two vertices u, v in V such that $d_u(S) = d_v(S) = 1$.*

Let us define, for $0 < c < 1$,

$$\gamma_n = c\frac{\log(\log(n))}{(k-1)\log(n)}. \tag{4.5}$$

Definition 4.2.9 (good). *A graph $G = (V, E)$ is said to be* good *if every set $S \subseteq V$ that is not nice has cardinality $|S| > \gamma_n n$. A symmetric matrix $\boldsymbol{A}$ is said to be* good *if the graph G with adjacency matrix equal to the support of $\boldsymbol{A}$ is good.*

Outline of the framework We follow the graph growth process exposing $\boldsymbol{M}$ minor by minor. The framework consists of two major steps.

First, we find $\boldsymbol{M}_r$ such that its rank is close to r, where $\boldsymbol{M}_r$ is the upper $r \times r$ minor of $\boldsymbol{M}$. Let δn denote such an r, for $0 < \delta \leq 1$. Such a value of δ can be chosen by the following lemma. The proof relies on Lemma 4.2.6.

Lemma 4.2.10. *Suppose $p \geq c_1(k-1)\log(n)/n$ and $c_1(1-1/k)(\delta k)^2 > 2$. Then, for any $\epsilon > 0$,*

$$\Pr[\operatorname{rank}(\boldsymbol{M}_{\delta n}) < (1-\epsilon)\delta n] = O(e^{-\epsilon^2 \frac{1}{k} n \log(n)}).$$

Then, for $\delta n \leq r < n$, we augment $\boldsymbol{M}_r$ with a new row and a new column to obtain $\boldsymbol{M}_{r+1}$ and show that the number of such augmentations is sufficient to remove any row and column dependencies in $\boldsymbol{M}_n$. This can be proved by showing that the augmentation process runs into good matrices with high probability.

Lemma 4.2.11. *For any $r \in \{\delta n, \ldots, n\}$, we have that $\boldsymbol{M}_{\delta n}, \ldots, \boldsymbol{M}_r$ is good with probability $1 - O(1/n^{c_1\delta - 2 - \epsilon})$, for any fixed $\epsilon > 0$.*

We note that this lemma follows from the following conjecture on nicety.

Conjecture 4.2.12. *For any $r \in \{\delta n, \ldots, n\}$, $\delta \in [1/k, 1)$ and $p \geq c_1(k-1)\log(n)/n$ with $c_1 > 1$, $G_r = (V_r, E_r)$ is such that every set $S \subseteq V_r$ such that $|S| \leq \gamma_n n$ is nice, with probability $1 - O(1/n^{2c_1\delta - 1 - \epsilon})$, for any fixed $\epsilon > 0$.*

Fix an arbitrary $r \in \{\delta n, \dots, n\}$. Recall that we defined $d_u(S) = \sum_{v \in S} \mathbb{1}\{M_{uv} > 0\}$ as the number of distinct neighbors of vertex u. Let $Z_u = 1$ if $d_u(S) = 1$ and $Z_u = 0$ otherwise. Denote the sum of all Z_u as $N(V \setminus S, S)$. Since a set is nice if there are at least two vertices u, v such that $d_u(S) = d_v(S) = 1$, our goal is to upper bound the event that $N(V \setminus S, S)$ is smaller or equal to 1. Note that we can simply apply the Chebyshev's inequality, but the resulting bound is not tight. It will be a tight bound if we can prove that the set of random variables $\{Z_u, u \in V \setminus S\}$ are negatively associated. However, proving this is tricky as Z_u is not an increasing function of $d_u(S)$. Perhaps it is possible to prove it directly by definition. We leave the analysis of this approach as a future work. We will assume that this lemma holds in the full proof of the main conjecture.

Let $\Delta_r := \text{rank}(\boldsymbol{M}_{r+1}) - \text{rank}(\boldsymbol{M}_r)$. Next, we prove that good matrices have the following good properties.

Lemma 4.2.13. *For every $\delta n \leq r < n$, and real $r \times r$ matrix $\boldsymbol{A}$, we have*

1. *If* $\text{rank}(\boldsymbol{A}) < r$ *and* $\boldsymbol{A}$ *is good,*

$$\Pr[\Delta_r < 2 \mid \boldsymbol{M}_r = \boldsymbol{A}] = O\left(\frac{1}{\sqrt{\log(\log(n))}}\right)$$

2. *If* $\text{rank}(\boldsymbol{A}) = r$ *and* $\boldsymbol{A}$ *is good,*

$$\Pr[\Delta_r < 1 \mid \boldsymbol{M}_r = \boldsymbol{A}] = O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

Note that

$$\text{rank}(\boldsymbol{M}_r) \leq \text{rank}(\boldsymbol{M}_{r+1}) \leq \min\{\text{rank}(\boldsymbol{M}_r) + 2, r + 1\}. \tag{4.6}$$

It is obvious that $\text{rank}(\boldsymbol{M}_{r+1}) \leq r + 1$. Since in each step we only add one new row and one new column, the rank increases at most by two. Lemma 4.2.13 shows that the second inequality in (4.6) holds with equality for all r such that $\delta n \leq r < n$ with high probability. Thus sufficient number of augmentation steps can remove any row and column dependencies in the matrix $\boldsymbol{M}$.

Under Lemma 4.2.10 and Lemma 4.2.13, we can prove the main conjecture. The full proof is given at the end of the chapter.

4.3 MLE conditions and MLE accuracy

As for the fixed design case, we are interested in the conditions that guarantee MLE existence and uniqueness in the random design case, as well as the MLE error under the condition that MLE exists and is unique.

4.3.1 MLE existence and uniqueness

We present sufficient conditions for $(\boldsymbol{X}, \boldsymbol{y})$ to satisfy the overlapping condition, when $\boldsymbol{X}$ is a random design matrix.

Theorem 4.3.1. *Assume that design matrix $\boldsymbol{X}$ is random with m rows and n columns and observations are according to the β-model with parameter $\boldsymbol{\beta}$ such that $\epsilon \leq p_{u,v}(\boldsymbol{\beta}) \leq 1-\epsilon$ for all $1 \leq u < v \leq n$, for some $\epsilon \in (0,1)$. Then, for any $c > 1/2$, an MLE exists and is unique with probability at least $1 - 2/n^{2c-1}$, provided that the number of experiments m satisfies*

$$m \geq 2c^{3/4}\frac{1}{\epsilon}n^{5/4}(\log(n))^{1/4}.$$

The proof follows similar ideas we used to establish results in Section 3.3.3 but also makes additional steps to bound the probability of the degree sequence being on a facet of the degree sequence polytope due to random design matrix.

From Theorem 4.3.1, we observe that the sufficient number of edge experiments m is sublinear in n^2, for any sufficiently large ϵ.

In Candés and Sur (2020), the authors established a sharp phase transition threshold for the MLE for logistic regression with Gaussian covariates. The design matrix thus has a different structure compared to our case. Our result only gives a sufficient condition for the existence of a unique MLE. This condition is interpretable in the sense that it explicitly describes the condition in terms of the number of experiments. Establishing a sharp condition for the MLE existence for the type of design matrices considered in this paper is an open problem.

Similarly as in the discussion of Theorem 3.3.7, we can derive a lower bound for the number of edge experiments needed for the normalized degree sequence to be in the interior of the degree sequence polytope. Assume that $\boldsymbol{\beta}$ is such that there exists $u \in V$ such that $p_{u,v}(\boldsymbol{\beta}) = \epsilon$ for all $v \neq u$. Note that

$$\Pr[d_u = 0] = \Pr[(1-\epsilon)^{M_u}]$$

where M_u is the number of edge experiments in which vertex u takes part in. Note that M_u is a random variable with binomial distribution with parameters m and $2/n$. It follows that

$$\mathbb{E}[(1-\epsilon)^{M_u}] = \left(1 - \frac{2\epsilon}{n}\right)^m.$$

Hence, for the normalized degree sequence $\boldsymbol{d}$ to be in the interior of the degree sequence polytope with probability at least $1 - 1/n^a$, for some $a > 0$, it is necessary that the number of edge experiments m is such that

$$m \geq \frac{a}{2}\frac{1}{\epsilon} n \log(n)(1 + o(1)). \tag{4.7}$$

This matches the upper bound in Theorem 4.3.1 with respect to $1/\epsilon$ while the upper bound has an extra factor of $n^{1/4}/\log(n)^{3/4}$ with respect to n.

The lower bound (4.7) for the MLE existence and the sufficient number of edge experiments for the full rank condition in Theorem 4.2.3 imply that there can be an arbitrarily large gap between the two by taking an instance for which the expected normalized de-

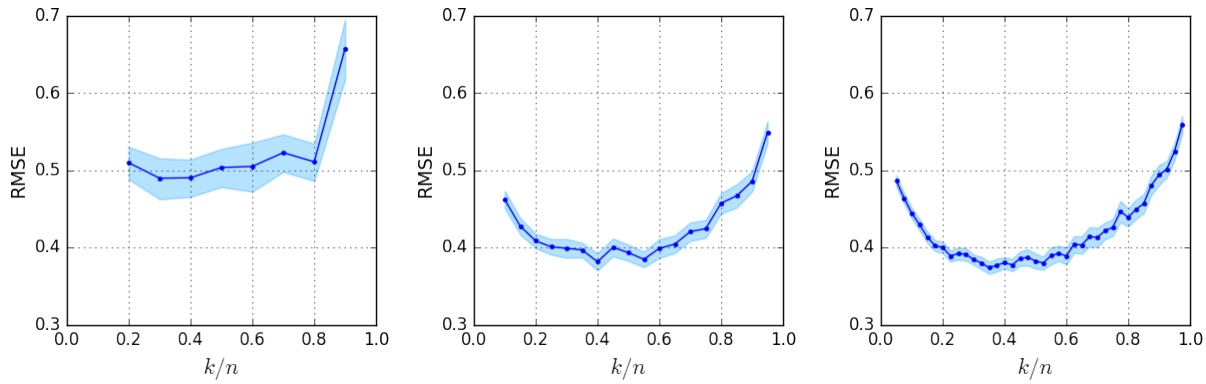

Figure 4.3: Root-mean-square error versus the normalized parameter k, for different values of n: (left) $n = 10$, (middle) $n = 20$ and (right) $n = 40$.

gree sequence is sufficiently close to a facet of the degree sequence polytope.

4.3.2 MLE error bounds

By Proposition 3.4.1 and (4.1), we have

$$\frac{1}{\sqrt{n}}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| \lessapprox c_{kb}\sqrt{\frac{\log(n)+2}{m}}\frac{1}{\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}}. \tag{4.8}$$

The upper bound on the MLE parameter estimation error increases as k goes to the boundary points 2 and $n-1$, for sufficiently small b. We validate that this holds the MLE parameter estimation error by experiments.

Example 4.3.1. *We consider k-uniform hypergraph instances with even number n of items according to the β-model with parameter vector $\boldsymbol{\beta}$ such that a half of items have parameter of value $-b$ and the other half have parameter of value b, for some $b > 0$. We fix the number of items n and the number of experiments m and evaluate the root-mean-square-error of the MLE $\hat{\boldsymbol{\beta}}$ for a sample of independently drawn random design matrices. In our experiments, we set $b = 1$, $m = n(n-1)/2$, and the number of repeated experiments to* 100.

The MLE $\boldsymbol{\beta}$ is estimated using gradient descent algorithm, which produces a sequence $\boldsymbol{\beta}^{(t)}$, with initial vector $\boldsymbol{\beta}^{(0)}$ having independent entries, sampled from uniform distribution on $[-b, b]$. The

gradient descent algorithm returns the estimate $\boldsymbol{\beta}^{(t^\star)}$ *where* $t^\star$ *is the smallest integer* t *such that* $||\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}||_\infty < \delta$, *and we set* $\delta = 0.0001$. *In Figure 4.3 we show the root-mean-square error* $||\boldsymbol{\beta}^{(t^\star)} - \boldsymbol{\beta}||/\sqrt{n}$ *versus the normalized parameter* k/n *for three different values of* n.

From the figure, we observe that the root-mean-square error follows a "U" shaped curve for large enough n, as suggested by equation (4.8).

4.4 Conclusion

In this chapter, we studied the β-model of random hypergraphs with random design matrices, defined by sampling candidate edges independently with replacement from the set of all combinations of k vertices from the set of n vertices. We showed conditions for the random design matrix to have full rank almost surely. We conjectured the condition is tight, which requires the number of edge experiments to be at least $c\frac{1}{k}n\log(n)$, for a fixed constant $c > 2$. Similarly as the fixed design case, we also derived a sufficient condition for MLE existence and uniqueness.

Note that we tested numerically that our conjecture holds. However, it is challenging to prove it exactly. As discussed in the sketch proof, this would require another conjecture on the nicety property to hold. We leave this as future work.

4.5 Proofs

Proof of Proposition 4.2.1

We consider random matrix $\boldsymbol{M}$, defined as the sum of independent random matrices, as given by

$$\boldsymbol{M} = \sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^\top.$$

Note that $\lambda_1(\boldsymbol{x}_j \boldsymbol{x}_j^\top) \geq 0$ and $||\boldsymbol{x}_j \boldsymbol{x}_j^\top||_2 \leq k$ for all $j \in \{1, \ldots, m\}$. Note, also, that $\lambda_1(\mathbb{E}[\boldsymbol{M}])$ is given in (4.1). The statement of the lemma follows by applying the matrix Chernoff bound in Lemma 2.3.4.

Proof of Theorem 4.2.2

Sufficiency We first establish the first claim of the theorem, which states a sufficient number of experiments for $\boldsymbol{X}$ to have all columns being non-null vectors with probability at least $1 - 1/n^a$. To prove this, we bound the probability that an arbitrarily fixed column is null. For every $v \in V$,

$$\Pr[v\text{-th column of } \boldsymbol{X} \text{ is null}] = \left(1 - \frac{k}{n}\right)^m.$$

By union bound, we have

$$\begin{aligned}
\Pr[\boldsymbol{X} \text{ has a null column}] &= \Pr[\cup_{v=1}^{n} \{v\text{-th column of } \boldsymbol{X} \text{ is null}\}] \\
&\leq n \Pr[\text{1-st column of } \boldsymbol{X} \text{ is null}] \\
&= n \left(1 - \frac{k}{n}\right)^m \\
&\leq n e^{-\frac{km}{n}}.
\end{aligned}$$

Hence, for $\Pr[\boldsymbol{X} \text{ has a null column}] \leq 1/n^a$ to hold it suffices that

$$m \geq (1+a)\frac{1}{k} n \log(n).$$

Necessity We next prove the second claim of the theorem, by using the solution of the *coupon subset selection problem* that is defined as follows. Let $N = \{1, \ldots, n\}$ be a ground set of coupons. Let $S_1, \ldots, S_m$ be subsets of coupons, each of cardinality k, drawn independently, uniformly at random with replacement from N. For any given set $A \subseteq N$, let $Y(A)$ be the number of coupons in A each contained in at least one set $S_1, S_2, \ldots, S_m$. The coupon subset selection problem asks to solve for the distribution of $Y(A)$. The classical coupon collector problem, where the goal is to evaluate the probability of collecting all distinct coupons by sampling one coupon at a time uniformly at random with replacement, is accommodated as a special case where each subset selection is of cardinality 1.

By Theorem 1 in Stadje (1990), the distribution of $Y(A)$ is given by

$$\Pr[Y(A) < x] = \sum_{i=0}^{x-1} (-1)^{x-i+1} \binom{|A|}{i} \binom{|A|-i-1}{|A|-x} \left(\frac{\binom{n-|A|+i}{k}}{\binom{n}{k}} \right)^m$$

where $x = 0, 1, \ldots, |A|$.

Note that the probability of the event that $\boldsymbol{X}$ has a null column is equal to the probability of the event $\{Y(N) < n\}$. Therefore, it follows

$$\Pr[\boldsymbol{X} \text{ has a null column}] = \sum_{i=1}^{n} (-1)^{i-1} \binom{n}{i} \left(\frac{\binom{n-i}{k}}{\binom{n}{k}} \right)^m.$$

This can be equivalently written as

$$\Pr[\boldsymbol{X} \text{ has a null column}] = \sum_{i=1}^{n-k} a_i \tag{4.9}$$

where

$$a_i := (-1)^{i-1} \binom{n}{i} \left(\left(1 - \frac{k}{n}\right) \cdots \left(1 - \frac{k}{n-i+1}\right) \right)^m.$$

Now, note that $a_i \geq 0$ for all odd $1 \leq i \leq n-k$ and $a_i + a_{i+1} \geq 0$ for all odd $1 \leq i < n-k$ if, and only if,

$$\frac{i+1}{n-i} - \left(1 - \frac{k}{n-i}\right)^m \geq 0 \text{ for all odd } 1 \leq i \leq n-k.$$

The left-hand side in the last inequality is increasing in i. Hence, the condition is equivalent to

$$m \geq \frac{\log\left(\frac{n-1}{2}\right)}{\log\left(1 + \frac{k}{n-1-k}\right)}. \tag{4.10}$$

Note that the right-hand side in the last inequality is equal to $\frac{1}{k} n \log(n)(1 - a_n)$ when $k = o(n/\log(n))$ for any sequence a_n such that $a_n = O(1/\log(n))$.

From (4.9), under condition (4.10), we have

$$\Pr[\boldsymbol{X} \text{ has a null column}] \quad \geq \quad a_1 + a_2.$$

Note that

$$a_1 + a_2 = n\left(1 - \frac{k}{n}\right)^m \left(1 - \frac{n-1}{2}\left(1 - \frac{k}{n-1}\right)^m\right).$$

Now, assume that $m = (1 + c_n)\frac{k}{n}\log(n)$ where c_n is a positive valued sequence. Then, we obtain

$$a_1 + a_2 \quad = \quad \frac{1}{n^{c_n}}\left(1 - \frac{1}{2}\frac{1}{n^{c_n}}\right)(1 + o(1)).$$

Under condition $c_n = O(1/\log(n))$, it follows that $a_1 + a_2 = \Omega(1)$, and hence, we have $\Pr[\boldsymbol{X} \text{ has a null column}] = \Omega(1)$.

Proof of Lemma 4.2.5

Define $\mathbb{E}_{uv} = \mathbb{1}\{M_{uv} > 0\}$. Note that this is an increasing function on M_{uv}. Thus RVs $\{E_{uv}, (u,v) \in E\}$ are negatively associated. We can write $d_u(S) = \sum_{v \in S} E_{uv}$ for all $u \in$

$V \setminus S$. Since $d_u(S)$ is an increasing function on disjoint subset of RVs $\{E_{uv}, (u,v) \in E\}$, we can conclude that RVs $\{d_u(S), u \in V \setminus S\}$ are also negatively associated.

Proof of Lemma 4.2.6

Let $\eta_{tuv} = X_{tu}X_{tv}$ where $(u,v) \in E$ for each experiment t. We omit t for a while as we fix a particular experiment t and simply write η_e where $e \in E$ when we do not refer to a particular pair of elements.

Lemma 4.5.1. *The RVs $\{\eta_{uv}, (u,v) \in E\}$ are non-positively correlated.*

Proof. Consider a pair of edges (u,v) and (u',v'). If $u \neq u'$ and $v \neq v'$, then η_{uv} and $\eta_{u'v'}$ are increasing functions defined on disjoint subsets of the set of random variables $\{X_{tu}, u \in V\}$. By Lemma 4.2.5, $\{X_{tu}, u \in V\}$ are NA, thus η_{uv} and $\eta_{u'v'}$ are NA, implying $\mathrm{Cov}(\eta_{uv}, \eta_{u'v'}) \leq 0$.

If $u = u'$, we consider the covariance conditioning on X_u,

$$\mathrm{Cov}(\eta_{uv}, \eta_{u'v'}) = \mathbb{E}\, Cov(\eta_{uv}, \eta_{u'v'} \mid X_u) + Cov(\mathbb{E}(\eta_{uv} \mid X_u), \mathbb{E}(\eta_{u'v'} \mid X_u))$$

Note that the second term is zero. By Lemma 4.2.5 the first term is nonpositive. Thus we can conclude that $\mathrm{Cov}(\eta_{uv}, \eta_{u'v'}) \leq 0$. The same argument works for the case when $v = v'$. □

From Lemma 4.5.1, we can derive an important property on the negative correlation between η_k and any monotone increasing property m over the variables η_e where $e \in E \setminus k$. This is a common property firstly derived by Feder and Mihail (1992) for the balanced matroid.

Lemma 4.5.2. *For any set $U \in E$, monotone increasing function f defined over the random variables η_e where $e \in U$, and all $k \in E \setminus U$, we have*

$$\mathrm{Cov}(\eta_k, f(\eta_e, e \in U)) \leq 0$$

Proof. The proof in Feder and Mihail (1992) can be adapt to our case, To simplify the notation, denote by m the random variable $f(\eta_e, e \in U)$. The goal is to show that $\Pr(m \mid \eta_k) \leq \Pr(m)$. We use induction on $|E|$. The case $|E| = 1$ is true as f is monotone increasing. For the general case, note that

$$\Pr(m \mid \eta_k) = \Pr(\eta_e \mid \eta_k) \Pr(m \mid \eta_e \eta_k) + \Pr(\bar{\eta_e} \mid \eta_k) \Pr(m \mid \bar{\eta_e} \eta_k)$$

$$\Pr(m) = \Pr(\eta_e) \Pr(m \mid \eta_e) + \Pr(\bar{\eta_e}) \Pr(m \mid \bar{\eta_e})$$

By Lemma 4.5.1, $\Pr(\eta_e \mid \eta_k) \leq \Pr(\eta_e)$. By the induction hypothesis applied to graph G with deleted edge k, $\Pr(m \mid \eta_e \eta_k) \leq \Pr(m \mid \eta_e)$. If in addition we have $\Pr(m \mid \eta_e \eta_k) \leq \Pr(m \mid \bar{\eta_e} \eta_k)$, then the lemma would follow by averaging principles. But such e can always be chosen. Note that

$$\sum_{e \neq k} \Pr(\eta_e \mid m \eta_k) = \binom{k}{2} - 1 = \sum_{e \neq k} \Pr(\eta_e \mid \eta_k)$$

Hence there exist some e such that $\Pr(\eta_e \mid m \eta_k) \geq \Pr(\eta_e \mid \eta_k)$, which is equivalent to the condition $\Pr(m \mid \eta_e \eta_k) \leq \Pr(m \mid \bar{\eta_e} \eta_k)$ as required. □

Using Lemma 4.5.1 and 4.5.2, we prove that the set of random variables $\{\eta_{uv}, (u, v) \in E\}$ satisfies a strong negative dependency structure, the negative association.

Lemma 4.5.3. *The RVs $\{\eta_{uv}, (u, v) \in E\}$ are NA.*

Proof. We use induction on $|E|$. The case $|E| = 1$ is trivial. For the general case, let η_1, η_2 be an arbitrary partition of η and f, g be binary increasing functions. We want to show that

$$\mathrm{Cov}\{f(\eta_1), g(\eta_2)\} \leq 0$$

Since $\sum_{e \in E} \eta_e = \binom{k}{2}$, we have

$$0 = \mathrm{Cov}\{f(\eta_1), \binom{k}{2}\} = \sum_{e \in E} \mathrm{Cov}\{f(\eta_1), \eta_e\}$$

This means there exists some $k \in E$ such that

$$\mathrm{Cov}\{f(\eta_1), \eta_k\} \geq 0 \tag{4.11}$$

By conditional covariance formula,

$$\mathrm{Cov}\{f(\eta_1), g(\eta_2)\} = \mathbb{E}\,\mathrm{Cov}\{f(\eta_1), g(\eta_2) \mid \eta_k\} + \mathrm{Cov}\{\mathbb{E}(f(\eta_1) \mid \eta_k), \mathbb{E}(g(\eta_2) \mid \eta_k)\} \tag{4.12}$$

The first term of eqn.(4.12) is nonpositive by induction hypothesis applied to graph G deleting edge k. Note that $\mathbb{E}(f(\eta_1) \mid \eta_k)$ and $\mathbb{E}(g(\eta_2) \mid \eta_k)$ inside the second term are binary random variables. If the covariance (4.11) is zero, then η_k and $f(\eta_1)$ are independent and the second term is zero. If the covariance (4.11) is positive, by Lemma 4.5.2, we know that $\mathbb{E}(f(\eta_1) \mid \eta_k)$ and $\mathbb{E}(g(\eta_2) \mid \eta_k)$ are discordant functions of η_k. Then, by the Chebyshev inequality we can conclude that the second term is nonpositive. Thus the whole expression is nonpositive. □

Finally, we consider m independent experiments $t = 1, \ldots, m$. Since union of independent sets of negatively associated random variables are NA, the set of RVs $\{\eta_{tuv}, t = 1, \ldots, m, (u, v) \in E\}$ are NA.

Note that we can write $M_{uv} = \sum_{t=1}^{m} \eta_{tuv}$. We immediately conclude the lemma as increasing functions defined on disjoint subsets of the set of RVs are NA.

Proof of Lemma 4.2.7

Define $\mathbb{E}_{uv} = \mathbb{1}\{M_{uv} > 0\}$. Note that this is an increasing function on M_{uv}. Thus RVs $\{E_{uv}, (u, v) \in E\}$ are negatively associated. We can write $d_u(S) = \sum_{v \in S} E_{uv}$ for all $u \in V \setminus S$. Since $d_u(S)$ is an increasing function on disjoint subset of RVs $\{E_{uv}, (u, v) \in E\}$, we

can conclude that RVs $\{d_u(S), u \in V \setminus S\}$ are also negatively associated.

Proof of Lemma 4.2.10

To ease the proof, we introduce some notation. Let p be the probability that $M_{u,v}$ takes a non-zero value,

$$p = 1 - \Pr[M_{u,v} = 0] = 1 - \left(1 - \frac{k(k-1)}{n(n-1)}\right)^m. \tag{4.13}$$

In what follows, we assume that

$$c_1 \frac{1}{k} n \log(n) \leq m \leq c_2 \frac{1}{k} n \log(n) \tag{4.14}$$

for some constants $c_2 > c_1 > 1$. From (4.13) and (4.14), it follows

$$p \geq c_1(k-1)\frac{\log(n)}{n}\left(1 - \frac{1}{2}c_1(k-1)\frac{\log(n)}{n}\right)$$

and

$$p \leq c_2(k-1)\frac{\log(n)}{n}\left(1 + \frac{1}{n-1}\right).$$

By assumption, $k = o(n/\log(n))$, hence

$$c_1(k-1)\frac{\log(n)}{n}(1 - o(1)) \leq p \leq c_2(k-1)\frac{\log(n)}{n}(1 + o(1)).$$

We need to upper bound the probability of the event $\{\text{rank}(M_{\delta n}) < (1-\epsilon)\delta n\}$ for any fixed $0 < \delta < 1$ and $0 < \epsilon < 1$. Let E denote the event that the last $\epsilon\delta n$ columns of $M_{\delta n}$ are in the span of remaining columns of $M_{\delta n}$. By symmetry and union bound, we have

$$\Pr[\text{rank}(M_{\delta n}) < (1-\epsilon)\delta n] \leq \binom{\delta n}{\epsilon \delta n} \Pr[E].$$

We decompose $M_{\delta n}$ as follows

$$M_{\delta n} = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

where A is a $(1-\epsilon)\delta n \times (1-\epsilon)\delta n$ matrix, B is a $(1-\epsilon)\delta n \times \epsilon\delta n$ matrix, and C is a $\epsilon\delta n \times \epsilon\delta n$ matrix.

We condition on the values of A and B. Then, under event E, $B = AF$ for a $(1-\epsilon)\delta n \times \epsilon\delta n$ matrix F and $C = B^\top F$. Note that F is not necessarily unique but $B^\top F$ is unique. Therefore, to bound $\Pr[E]$ it suffices to find a uniform upper bound for $\Pr[C = B^\top F^* \mid A = A', B = B']$ where F^* is a solution to $A'F^* = B'$ that holds for all A' and B'.

We note that C is still random and the entries of C are not independent. We further explore the dependency structure of the upper diagonal elements of C. Conditioning on the values of A and B, the randomness of C is due to experiments containing only the last $\delta\epsilon n$ vertices. Let V denote the set of last $\delta\epsilon n$ vertices and E the edge set. By Lemma 4.5.3, RVs $\{C_{uv}, (u,v) \in E\}$ are negatively associated.

Now we find a uniform bound for $\Pr[C_{uv} = Z_{uv}, \forall u, v]$ with arbitrary values of Z. Let $I = \{(u,v) \mid Z_{uv} > 0\}$ and $J = \{(u,v) \mid Z_{uv} = 0\}$. Note that conditioning on Z, I and J are fixed sets s.t. $|I| + |J| = (\delta\epsilon n - 1)\delta\epsilon n/2$. We define $C_I = \min\{C_{u,v} \mid (u,v) \in I\}$ and $C_J = \max\{C_{u,v} \mid (u,v) \in J\}$.

$$\begin{aligned} \Pr(C_{uv} = Z_{uv}, \forall u, v) &\leq \Pr(C_I > 0, C_J = 0) \\ &\leq \Pr[C_J = 0] \Pr(C_I > 0 | C_J = 0) \\ &\leq (1-p)^{|J|} \Pr(C_I > 0 | C_J = 0) \end{aligned} \tag{4.15}$$

The last inequality is by the negative association property of $\{C_{uv}, (u,v) \in J\}$.

When the set of zero elements is larger, $|J| \geq (\delta\epsilon n - 1)\delta\epsilon n/4$. In this case, we have

$$\begin{aligned}
\Pr[\text{rank}(\boldsymbol{M}_{\delta n}) \leq (1-\epsilon)\delta n] &\leq \binom{\delta n}{\epsilon\delta n} e^{-p\frac{(\epsilon\delta n-1)\epsilon\delta n}{4}} \\
&\leq \left(\frac{\delta n e}{\epsilon\delta n}\right)^{\epsilon\delta n} e^{-p\frac{(\epsilon\delta n-1)\epsilon\delta n}{4}} \\
&\leq e^{-\left(\frac{pk(\epsilon\delta n-1)\epsilon\delta n}{4\epsilon^2 n\log(n)} - \frac{\delta k}{\epsilon}\log\left(\frac{e}{\epsilon}\right)\frac{1}{\log(n)}\right)\epsilon^2\frac{1}{k}n\log(n)} \\
&\leq e^{-\left(c_1\frac{1-1/k}{4}(\delta k)^2\left(1-\frac{1}{\epsilon\delta n}\right) - \frac{\delta k}{\epsilon}\log\left(\frac{e}{\epsilon}\right)\frac{1}{\log(n)}\right)\epsilon^2\frac{1}{k}n\log(n)}.
\end{aligned}$$

Under assumed condition $c_1\frac{1-1/k}{4}(\delta k)^2 > 1$, we have

$$\Pr[\text{rank}(\boldsymbol{M}_{\delta n}) \leq (1-\epsilon)\delta n] = O(e^{-\epsilon^2\frac{1}{k}n\log(n)}).$$

On the other hand, when the set of non-zero elements is larger, we have $|I| \geq (\delta\epsilon n - 1)\delta\epsilon n/4$. In this case, we condition the other way and we have,

$$\begin{aligned}
\Pr(C_{uv} = Z_{uv}) &\leq \Pr(C_I > 0, C_J = 0) \\
&\leq \Pr[C_I > 0]\Pr(C_J = 0 | C_I > 0) \\
&\leq (1-p)^{|I|}\Pr(C_J = 0 | C_I > 0) \qquad (4.16)
\end{aligned}$$

We can obtain the bound using the same argument as the case above. Combining the results for both cases, we can prove the lemma.

Proof of Lemma 4.2.13

Fix an arbitrary $r \in \{\delta n, \ldots, n\}$. To simplify the notation, let $\boldsymbol{A}$ be the upper left $r \times r$ minor of $\boldsymbol{M}$ and $\boldsymbol{A}'$ be the augmentation of $\boldsymbol{A}$ by adding a new column $\boldsymbol{x} = (x_1, \ldots, x_{m+1})^\top$ and its transpose as a new row.

We prove the lemma by separately considering the following two cases: Case 1: $\text{rank}(\boldsymbol{A}) < r$ and Case 2: $\text{rank}(\boldsymbol{A}) = r$.

Case 1 Consider the null space of $\boldsymbol{A}$. Since $\boldsymbol{A}$ is singular, the dimension of its null space is non zero. Hence, there exists $\boldsymbol{\xi} \neq 0$ such that $\boldsymbol{A\xi} = \boldsymbol{0}$.

Let s_n be the number of nonzero elements of $\boldsymbol{\xi}$. If $s_n \leq \gamma_n n$, then because $\boldsymbol{A}$ is good, there exists some vertex v that has only one neighbor in the support of $\boldsymbol{\xi}$. This implies that the product of the v-th row of $\boldsymbol{A}$ and $\boldsymbol{\xi}$ is non-zero, which contradicts $\boldsymbol{A\xi} = \boldsymbol{0}$. Thus, we have $s_n > \gamma_n n$.

Consider the new column $\boldsymbol{x}$. By definition, the new column $\boldsymbol{x}$ specifies the number of experiments containing both the new vertex (say, vertex $r+1$) and each of the old vertices u, $1 \leq u \leq r$ such that

$$x_u = \sum_{t=1}^{m} X_{t,u} X_{t,r+1}.$$

Let us define the notation $z_{t,u} = X_{t,u} X_{t,r+1}$.

If $\boldsymbol{x}$ does not satisfy

$$\sum_{u=1}^{r} \xi_u x_u = 0 \tag{4.17}$$

then the new column is independent from the columns of $\boldsymbol{A}$, and augmenting $\boldsymbol{A}$ by $\boldsymbol{x}$ increases its rank by 2. Therefore, we bound the probability of the event $\{\sum_{u=1}^{r} \xi_u x_u = 0\}$.

We rewrite (4.17) as $\boldsymbol{\xi}^\top \boldsymbol{x} = \sum_{t=1}^{m} \boldsymbol{\xi}^\top \boldsymbol{z}_t$. Let $\chi_t = \boldsymbol{\xi}^\top \boldsymbol{z}_t$. Since the experiments are independent for all $t = 1, \ldots, m$, χ_t are independent and identically distributed random variables satisfying

$$\chi_t = \sum_{u=1}^{r} \xi_u z_{t,u} = X_{t,r+1} \sum_{u=1}^{r} \xi_u X_{t,u}.$$

By the Littlewood-Offord theorem for sum of random variables Esseen (1968), if

$$2\alpha \leq \Pr[\chi_t = 0] \leq 1 - 2\alpha \tag{4.18}$$

with $\alpha = \Omega((\log(\log(n))/\log(n))(k/n))$, then

$$\Pr\left[\sum_{t=1}^{m} \chi_t = 0\right] = O\left(\frac{1}{\sqrt{mk/n}}\right) = O\left(\frac{1}{\sqrt{\log(\log(n))}}\right)$$

which is our desired result. This requires the number of non-zero elements of $\boldsymbol{\xi}$ to be sufficiently large.

Lemma 4.5.4. *If $s_n \geq \gamma_n n$ with γ_n defined in (4.5), then*

$$\alpha = \Omega\left(\frac{\log(\log(n))}{\log(n)}\frac{k}{n}\right).$$

Proof. Let s_n^+ and s_n^- denote the number of positive and negative elements of $\boldsymbol{\xi}$. Without loss of generality, we can assume $s_n^+ \geq s_n^-$. If this inequality does not hold, then we can simply consider the vector $-\boldsymbol{\xi}$. For the random variable χ_1, we have

$$\Pr[\chi_1 = 0] = 1 - \frac{k}{n} + \frac{k}{n}\Pr[\chi_1' = 0 \mid X_{1,r+1} = 1] \tag{4.19}$$

where $\chi_1' = \sum_{u=1}^{r} \xi_u X_{1,u}$. From (4.19), $\Pr[\chi_1 = 0] \geq 1 - k/n$ and hence $\Pr[\chi_1 = 0] \geq 2\alpha$, for $0 < \alpha \leq (1/2)(1 - k/n)$. Furthermore, if $\Pr[\chi_1' = 0 \mid X_{1,r+1} = 1] \leq 1 - \beta$, then $\Pr[\chi_1 = 0] \leq 1 - 2\alpha$ for $0 < \alpha \leq (\beta/2)k/n$.

Now, let E' denote the event that by sampling without replacement $k-1$ balls from an urn of $n-1$ balls, none of the sampled balls are from S_- and at least one of the sampled balls is from S_+, where S_- and S_+ are two disjoint sets of balls of cardinalities s_n^+ and s_n^-, respectively. Then, we have

$$\Pr[\chi_1' > 0 \mid X_{1,r+1} = 1] \geq \Pr[E']$$

and

$$\begin{aligned}
\Pr[E'] &= \frac{\binom{n-1-s_n^-}{k-1}}{\binom{n-1}{k-1}}\left(1 - \frac{\binom{n-1-s_n}{k-1}}{\binom{n-1-s_n^-}{k-1}}\right) \\
&= \left(1 - \frac{s_n^-}{n-1}\right)\cdots\left(1 - \frac{s_n^-}{n-(k-1)}\right) \\
&\quad \left(1 - \left(1 - \frac{s_n^+}{n-s_n^- - 1}\right)\cdots\left(1 - \frac{s_n^+}{n-s_n^- - (k-1)}\right)\right) \\
&\geq \left(1 - \frac{1}{2}\frac{s_n}{n-(k-1)}\right)^{k-1}\left(1 - \left(1 - \frac{1}{2}\frac{s_n}{n}\right)^{k-1}\right).
\end{aligned}$$

Hence, $\Pr[\chi_1' = 0] \leq 1 - \beta(k, n, s_n)$ where

$$\beta(k, n, s_n) = \left(1 - \frac{1}{2}\frac{s_n}{n-(k-1)}\right)^{k-1}\left(1 - \left(1 - \frac{1}{2}\frac{s_n}{n}\right)^{k-1}\right).$$

Now, note

$$\beta(k, n, \gamma_n n) = \frac{c}{2}\frac{\log(\log(n))}{\log(n)}(1 + o(1)).$$

Hence,

$$\alpha = \Omega\left(\frac{\log(\log(n))}{\log(n)}\frac{k}{n}\right).$$

□

Case 2 In this case $\boldsymbol{A}$ is a non-singular $r \times r$ matrix. The determinant of $\boldsymbol{A}'$ can be expressed as

$$\det(\boldsymbol{A}') = \sum_{u=1}^{r}\sum_{v=1}^{r} c_{u,v}x_u x_v + \det(\boldsymbol{A})x_{r+1} \tag{4.20}$$

where $c_{u,v}$ is the (u, v) cofactor matrix of $\boldsymbol{A}$. Our goal is to bound the probability of the event $\{\det(\boldsymbol{A}') = 0\}$.

As in Case 1, we define $z_{t,u} = X_{t,u}X_{t,r+1}$. Then we can rewrite the first term in the right-hand side of equation (4.20) as follows

$$\begin{aligned}\sum_{u=1}^{r}\sum_{v=1}^{r} c_{u,v}x_u x_v &= \sum_{u=1}^{r}\sum_{v=1}^{r} c_{u,v}\left(\sum_{t=1}^{m} z_{t,u}\right)\left(\sum_{s=1}^{m} z_{s,v}\right) \\ &= \sum_{t=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{r}\sum_{v=1}^{r} c_{u,v}z_{t,u}z_{s,v} = \sum_{t=1}^{m}\sum_{s=1}^{m} Q_{t,s}\end{aligned} \tag{4.21}$$

where

$$Q_{t,s} := \sum_{u=1}^{r}\sum_{v=1}^{r} c_{u,v}z_{t,u}z_{s,v}.$$

Define sets $X = \{z_{u,\cdot} : u \leq m/2\}$ and $Y = \{z_{u,\cdot} : u > m/2\}$. Then,

$$\sum_{t=1}^{m}\sum_{s=1}^{m} Q_{t,s} = Q(X, Y)$$

where $Q(X,Y)$ is the quadratic form of X and Y satisfying

$$Q(X,Y) = Q_{X,X} + 2Q_{X,Y} + Q_{Y,Y}$$

where

$$\begin{aligned} Q_{X,X} &:= \sum_{t=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{r}\sum_{v=1}^{r} x_{t,u}x_{s,v} \\ Q_{X,Y} &:= \sum_{t=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{r}\sum_{v=1}^{r} x_{t,u}y_{s,v} \\ Q_{Y,Y} &:= \sum_{t=1}^{m}\sum_{s=1}^{m}\sum_{u=1}^{r}\sum_{v=1}^{r} y_{t,u}y_{s,v}. \end{aligned}$$

Let $c = -\det(\boldsymbol{A})x_{r+1}$ be a fixed constant. By an application of Cauchy-Schwartz inequality, we bound the event $\{\det(\boldsymbol{A}') = 0\}$ by considering

$$\Pr[Q(X,Y) = c]^2 \le \Pr[Q(X,Y) = Q(\tilde{X},Y) = c] \le \Pr[Q(X,Y) - Q(\tilde{X},Y) = 0] \quad (4.22)$$

where $\tilde{X}$ is an independent copy of X.

We next bound the difference between $Q(X,Y)$ and $Q(\tilde{X},Y)$. Note that

$$\begin{aligned} Q(X,Y) - Q(\tilde{X},Y) &= 2(Q_{X,Y} - Q_{\tilde{X},Y}) + Q_{X,X} - Q_{\tilde{X},\tilde{X}} \\ &= 2\sum_{s>m/2}\sum_{v=1}^{r}\left(\sum_{t\le m/2}\sum_{u=1}^{r} c_{u,v}(z_{t,u} - \tilde{z}_{t,u})\right) z_{s,v} + f(X,\tilde{X}) \end{aligned}$$

where $f(X,\tilde{X})$ is independent of set Y. Let ξ_v be the term inside the large bracket. Then, we can write

$$\Pr[Q(X,Y) - Q(\tilde{X},Y) = 0] = \Pr\left[\sum_{s>m/2}\sum_{v=1}^{r}\xi_v z_{s,v} = -\frac{1}{2}f(X,\tilde{X})\right].$$

As in Case 1, if the number of non-zero elements of $\boldsymbol{\xi}$ is sufficiently large, then we can bound the term using the result from Case 1. Conditioning on the number of zero ele-

ments of $\boldsymbol{\xi}$ we have,

$$\begin{aligned} & \Pr[Q(X,Y) - Q(\tilde{X},Y) = 0] \\ \leq\ & \Pr\left[\sum_{s>m/2}\sum_{v=1}^{r} \xi_v z_{s,v} = -\frac{1}{2}f(X,\tilde{X}) \mid \sum_{v=1}^{r} \mathbb{1}_{\{\xi_v=0\}} < \gamma_n n\right] \qquad (4.23) \\ & + \Pr\left[\sum_{v=1}^{r} \mathbb{1}_{\{\xi_v=0\}} \geq \gamma_n n\right]. \qquad (4.24) \end{aligned}$$

Further condition the first term (4.23) on the possible values of $\boldsymbol{\xi}$ such that

$$\sum_{\boldsymbol{x}} \Pr\left[\sum_{s>m/2}\sum_{v=1}^{r} \boldsymbol{\xi}_v z_{s,v} = -\frac{1}{2}f(X,\tilde{X}) \mid \sum_{v=1}^{r} \mathbb{1}_{\{\boldsymbol{\xi}_v=0\}} < \gamma_n n, \boldsymbol{\xi} = \boldsymbol{x}\right] \Pr[\boldsymbol{\xi} = \boldsymbol{x}]$$

Since each conditional probability term is bounded by $O(1/\sqrt{\log(\log(n))})$ for any fixed $\boldsymbol{\xi}$, the whole sum is bounded by $O(1/\sqrt{\log(\log(n))})$.

For the second term (4.24), we first consider the probability that each ξ_v takes value zero. Since z and $\tilde{z}$ are independent copies, we have

$$\begin{aligned} & \Pr\left[\sum_{t\leq m/2}\sum_{u=1}^{r} c_{u,v}(z_{t,u} - \tilde{z}_{t,u}) = 0\right] \\ =\ & \sum_{x} \Pr\left[\sum_{t\leq m/2}\sum_{u=1}^{r} c_{u,v} z_{t,u} = x\right] \Pr\left[\sum_{t\leq m/2}\sum_{u=1}^{r} c_{u,v} \tilde{z}_{t,u} = x\right] \end{aligned}$$

where the sum is over all possible values the second copy can take. If we can show that for each v there are sufficiently many indices u for which $c_{u,v} \neq 0$, then we can use the result from Case 1 to bound the term (4.24).

Consider the cofactors of matrix $\boldsymbol{A}$. Since $\boldsymbol{A}$ is non-singular, dropping any columns of $\boldsymbol{A}$ will lead to a matrix $\tilde{\boldsymbol{A}}$ with exactly one linear combination of its rows equal to zero. The cofactor is non-zero when we drop any of the rows in this combination. With high probability, this combination of rows has size greater than $\gamma_n n$. This can be proved by contradiction. Take a set S of rows in the column-deleted matrix $\tilde{\boldsymbol{A}}$ with size at most $\gamma_n n$. If S is not independent, there exists $\boldsymbol{\xi} \neq \mathbf{0}$ such that $\boldsymbol{A}_S\boldsymbol{\xi} = 0$ where $\boldsymbol{A}_S$ is the transpose of the matrix $\tilde{\boldsymbol{A}}$ with columns restricted to set S. Since $|S| \leq \gamma_n n$, then by condition that

$\boldsymbol{A}$ is good there exist at least two vertices with only one neighbor in the set S. Thus, after column deletion, there exists at least one row in $\boldsymbol{A}_S$, say $\boldsymbol{A}_{S_u}$ with exactly one non-zero element. This means $\boldsymbol{A}_{S_u}\boldsymbol{\xi} \neq \mathbf{0}$, which contradicts $\boldsymbol{A}\boldsymbol{\xi} = 0$. Therefore, it follows that for each index v there are at least $\gamma_n n$ indices u such that $c_{u,v} \neq 0$.

These together give $\Pr[\xi_v = 0] = O(1/\sqrt{\log(\log(n))})$ for all v. Thus,

$$\mathbb{E}\left[\sum_{v=1}^{r} \mathbb{1}_{\{\xi_v=0\}}\right] = O\left(\frac{n}{\sqrt{\log(\log(n))}}\right).$$

By Markov's inequality, the second term in (4.24) is $O(1/\sqrt{\log(\log(n))})$.

We have shown that

$$\Pr[Q(X,Y) - Q(\tilde{X},Y) = 0] = O\left(\frac{1}{\sqrt{\log(\log(n))}}\right).$$

Combining with (4.22) we obtain

$$\Pr[\det(\boldsymbol{A}') = 0] = O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

Proof of Conjecture 4.2.3

Assuming the conjecture 4.2.12 holds, we have the full proof for the main conjecture.

Let $\mathcal{G}_r$ denote the event that $\boldsymbol{M}_{\delta n}, \ldots, \boldsymbol{M}_r$ are good, for $r \in \{\delta n, \ldots, n\}$. It can be readily seen that

$$\Pr[\operatorname{rank}(\boldsymbol{M}_n) < n] \leq A + B + C \tag{4.25}$$

where

$$A := \Pr\left[\{\operatorname{rank}(\boldsymbol{M}_n) < n\} \cap \mathcal{G}_n \mid \operatorname{rank}(\boldsymbol{M}_{\delta n}) \geq (1-\epsilon)\delta n\right],$$

$$B := \Pr\left[\operatorname{rank}(\boldsymbol{M}_{\delta n}) < (1-\epsilon)\delta n\right]$$

and

$$C := \Pr[\overline{\mathcal{G}}_n].$$

By lemma 4.2.10, $B = O(e^{-(\epsilon^2/k)n\log(n)})$. Assume that the conjecture 4.2.12 holds, we have $C = O(1/n^{c_1\delta-2-\epsilon})$. In the following, we bound A. Let $Y_r = r - \text{rank}(\boldsymbol{M}_r)$ and X_r be defined as

$$X_r = 4^{Y_r}\mathbb{1}_{\{Y_r>0\}}\mathbb{1}_{\mathcal{G}_r}.$$

Matrix $\boldsymbol{M}_n$ is not of full rank if, and only if, $Y_n \geq 1$, or $X_n \geq 4$ equivalently. We bound the event that $X_n \geq 4$ as follows.

Lemma 4.5.5. *For every $\delta n \leq r < n$,*

$$\mathbb{E}[X_{r+1} \mid \boldsymbol{M}_{\delta n}, \ldots, \boldsymbol{M}_r] \leq \frac{3}{5}X_r + O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

Proof. If $Y_r = y > 0$, by Lemma 4.2.13, $Y_{r+1} = y-1$ with probability $1 - O(1/\sqrt{\log(\log(n))})$, otherwise Y_{r+1} is at most $y+1$. Thus,

$$\begin{aligned}\mathbb{E}[X_{r+1} \mid \boldsymbol{M}_{\delta n}, \ldots, \boldsymbol{M}_n, Y_r = y] &\leq 4^{y-1}\left(1 - O\left(\frac{1}{\sqrt{\log(\log(n))}}\right)\right)\\ &\quad + 4^{y+1}O\left(\frac{1}{\sqrt{\log(\log(n))}}\right)\\ &\leq \frac{3}{5}4^y.\end{aligned}$$

If $Y_r = 0$, by Lemma 4.2.13, $Y_{r+1} = 0$ with probability $1 - O(1/(\log(\log(n)))^{1/4})$, otherwise $Y_{r+1} = 1$. Thus,

$$\mathbb{E}[X_{r+1} \mid \boldsymbol{M}_{\delta n}, \ldots, \boldsymbol{M}_n, Y_r = 0] = O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

The proof of the lemma follows from the bounds shown above. □

By Lemma 4.5.5 and the law of total expectation,

$$\mathbb{E}[X_{r+1} \mid \boldsymbol{M}_{\delta n}] \leq \frac{3}{5}\,\mathbb{E}[X_r \mid \boldsymbol{M}_{\delta n}] + O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

By induction from δn to n, we have

$$\mathbb{E}[X_n \mid \boldsymbol{M}_{\delta n}] \leq \left(\frac{3}{5}\right)^{n-\delta n} X_{\delta n} + O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

For $\boldsymbol{M}_{\delta n}$ satisfying the condition $\operatorname{rank}(\boldsymbol{M}_{\delta n}) \geq (1-\epsilon)\delta n$ where $\epsilon = (1-\delta)/(4\delta)$, we have $X_{\delta n} \leq (\sqrt{2})^{n-\delta n}$, and

$$\begin{aligned}\mathbb{E}[X_n \mid \operatorname{rank}(\boldsymbol{M}_{\delta n}) \geq \delta n(1-\epsilon)] &\leq \left(\frac{3\sqrt{2}}{5}\right)^{n-\delta n} + O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right) \\ &= O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).\end{aligned}$$

By Markov inequality, it follows that

$$\Pr\left[X_n \geq 4 \mid \operatorname{rank}(\boldsymbol{M}_{\delta n}) \geq (1-\epsilon)\delta n\right] = O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right).$$

This implies $A = O(1/(\log(\log(n)))^{1/4})$.

From (4.25) and the established bounds on A, B and C, we have that for any c_1 and δ such that $1/5 < \delta < 1$ and $c_1\delta > 2$, we have

$$\Pr[\operatorname{rank}(\boldsymbol{M}_n) < n] = O\left(\frac{1}{(\log(\log(n)))^{1/4}}\right)$$

which completes the proof of the theorem.

Proof of Theorem 4.3.1

Note that $(M_{u,v}, 1 \leq u < v \leq n)$ is a random vector with multinomial distribution with parameter m and uniform probability parameters equal to $1/\binom{n}{2}$.

Let us condition on that each $u \in V$ takes part in m_u experiments. Then, we have

$$\begin{aligned}
\mathbb{E}\left[d_u \mid \sum_{w\neq u} M_{u,w} = m_u\right] &= \sum_{v\neq u} \mathbb{E}\left[\frac{\tilde{y}_{u,v}}{M_{u,v}} \mathbb{1}_{\{M_{u,v}>0\}} \mid \sum_{w\neq u} M_{u,w} = m_u\right] \\
&= \sum_{v\neq u} p_{u,v}(\boldsymbol{\beta}) \Pr\left[M_{u,v} > 0 \mid \sum_{w\neq u} M_{u,w} = m_u\right] \\
&= \left(1 - \left(1 - \frac{1}{n-1}\right)^{m_u}\right) \sum_{v\neq u} p_{u,v}(\boldsymbol{\beta}).
\end{aligned}$$

Now, note that the number of experiments in which vertex $u \in V$ participates, m_u, is a random variable that has binomial distribution with parameters m and $2/n$. Hence, it follows

$$\begin{aligned}
\mathbb{E}[d_u] &= \left(1 - \mathbb{E}\left[\left(1 - \frac{1}{n-1}\right)^{m_u}\right]\right) \sum_{v\neq u} p_{u,v}(\boldsymbol{\beta}) \\
&= \left(1 - \left(1 - \frac{2}{n(n-1)}\right)^{m}\right) \sum_{v\neq u} p_{u,v}(\boldsymbol{\beta}).
\end{aligned}$$

We can write

$$\mathbb{E}[d_u] = c_{m,n} \sum_{v\neq u} p_{u,v}(\boldsymbol{\beta})$$

where

$$c_{m,n} := 1 - \left(1 - \frac{2}{n(n-1)}\right)^{m}.$$

Under $m = o(n^2)$, we have

$$c_{m,n} = \frac{2m}{n^2}(1 + o(1)).$$

We show the following lemma.

Lemma 4.5.6. *For any $\gamma > 0$ and $c > 1/2$, under condition $\gamma^2 c_{n,m}^2 (n-1) \geq c\log(n)$, with probability at least $1 - 2/n^{2c-1}$,*

$$||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \leq \sqrt{c(1+\gamma)c_{n,m}(n-1)\log(n)}.$$

Proof. By using union bound, we have

$$\Pr[||\boldsymbol{d} - \mathbb{E}[\boldsymbol{d}]||_\infty \geq x] \leq n \max_{v \in V} \Pr[|d_v - \mathbb{E}[d_v]| \geq x].$$

Fix an arbitrary $u \in V$. Conditional on $\boldsymbol{M}_u := (M_{u,v}, v \neq u) = \boldsymbol{m}_u$, by Hoeffding's inequality, we have

$$\begin{aligned} & \Pr[|d_u - \mathbb{E}[d_u]| \geq x \mid \boldsymbol{M}_u = \boldsymbol{m}_u] \\ = \; & \Pr\left[\left|\sum_{v \neq u} \frac{\tilde{y}_{u,v}}{m_{u,v}} \mathbb{1}_{\{m_{u,v}>0\}} - \mathbb{E}\left[\sum_{j \neq i} \frac{\tilde{y}_{u,v}}{m_{u,v}} \mathbb{1}_{\{m_{u,v}>0\}}\right]\right| \geq x \mid \boldsymbol{M}_u = \boldsymbol{m}_u\right] \\ \leq \; & 2\exp\left(-\frac{2x^2}{\sum_{v \neq u} \frac{1}{m_{u,v}} \mathbb{1}_{\{m_{u,v}>0\}}}\right). \end{aligned}$$

Hence, we have

$$\Pr[|d_u - \mathbb{E}[d_u]| \geq x] \leq 2\,\mathbb{E}\left[\exp\left(-\frac{2x^2}{\sum_{v \neq u} \frac{1}{M_{u,v}} \mathbb{1}_{\{M_{u,v}>0\}}}\right)\right].$$

We next use the obvious fact

$$\sum_{v \neq u} \frac{1}{m_{u,v}} \mathbb{1}_{\{m_{u,v}>0\}} \leq \sum_{v \neq u} \mathbb{1}_{\{m_{u,v}>0\}}$$

which yields

$$\Pr[|d_u - \mathbb{E}[d_u]| \geq x] \leq 2\,\mathbb{E}\left[\exp\left(-\frac{2x^2}{\sum_{v \neq u} \mathbb{1}_{\{M_{u,v}>0\}}}\right)\right].$$

Now, for $\gamma > 0$, we have

$$\begin{aligned} & \mathbb{E}\left[\exp\left(-\frac{2x^2}{\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}}\right)\right] \\ \leq\ & \exp\left(-\frac{2x^2}{(1+\gamma)\,\mathbb{E}\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}\right]}\right) \\ & + \Pr\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}} \geq (1+\gamma)\,\mathbb{E}\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}\right]\right]. \end{aligned}$$

Random variables $\mathbb{1}_{M_{u,v}>0}$, $v \neq u$, are negatively associated, hence by Dubhashi and Ranjan (1998), we can apply Hoeffding's bound,

$$\Pr\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}} \geq (1+\gamma)\,\mathbb{E}\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}\right]\right] \leq e^{-\frac{2\gamma^2\,\mathbb{E}\left[\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}\right]^2}{n-1}}.$$

Therefore, we have

$$\begin{aligned} & \frac{1}{2}\Pr[|d_u - \mathbb{E}[d_u]| \geq x] \\ \leq\ & \exp\left(-\frac{2x^2}{(1+\gamma)c_{n,m}(n-1)}\right) + \exp\left(-\frac{2\gamma^2[c_{n,m}(n-1)]^2}{n-1}\right). \end{aligned}$$

We have $\Pr[|d_u - \mathbb{E}[d_u]| \geq x] \leq 2/n^{2c}$, under the condition that each term in the right-hand side of the last inequality is bounded by $1/n^{2c}$, which is equivalent to

$$x \geq \sqrt{c(1+\gamma)c_{n,m}(n-1)\log(n)} \tag{4.26}$$

and

$$\gamma^2 c_{n,m}^2(n-1) \geq c\log(n). \tag{4.27}$$

□

We next show the following lemma.

Lemma 4.5.7. *For all $n \geq 3$,*

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \frac{1}{4} c_{m,n} \epsilon.$$

Proof. First, note that

$$\min_{u \in V} \mathbb{E}[d_u] \geq c_{m,n} \epsilon,$$

and

$$\min_{u \in V} \{n - 1 - \mathbb{E}[d_u]\} \geq (n-1)\left(1 - c_{m,n}(1-\epsilon)\right) \geq c_{m,n}\epsilon(n-1).$$

For every $(S, T) \in \Omega$, we have

$$\begin{aligned} \sum_{u \in S} \mathbb{E}[d_u] - \sum_{u \in T} \mathbb{E}[d_u] \;\leq\; & \sum_{u,v \in S: u \neq v} c_{m,n} p_{u,v}(\boldsymbol{\beta}) + \sum_{u \in S, v \in \overline{S \cup T}} c_{m,n} p_{u,v}(\boldsymbol{\beta}) \\ & - \sum_{u,v \in T: u \neq v} c_{m,n} p_{u,v}(\boldsymbol{\beta}) - \sum_{u \in T, v \in \overline{S \cup T}} c_{m,n} p_{u,v}(\boldsymbol{\beta}). \end{aligned}$$

Hence, we have

$$\begin{aligned} f(S, T, \mathbb{E}[\boldsymbol{d}], n) \;\geq\; & \sum_{u,v \in S: u \neq v} (1 - c_{m,n} p_{u,v}(\boldsymbol{\beta})) + \sum_{u \in S, v \in \overline{S \cup T}} (1 - c_{m,n} p_{u,v}(\boldsymbol{\beta})) \\ & + \sum_{u,v \in T: u \neq v} c_{m,n} p_{u,v}(\boldsymbol{\beta}) + \sum_{u \in T, v \in \overline{S \cup T}} c_{m,n} p_{u,v}(\boldsymbol{\beta}) \end{aligned}$$

and, thus

$$\begin{aligned} & f(S, T, \mathbb{E}[\boldsymbol{d}], n) \\ \geq\; & (1 - c_{m,n}(1-\epsilon))|S|(n - 1 - |T|) + c_{m,n}\epsilon|T|(n - 1 - |S|) \\ =\; & (1 - c_{m,n})|S|(n - 1 - |T|) + c_{m,n}\epsilon[(n-1)(|S| + |T|) - 2|S||T|]. \end{aligned}$$

Combining this with S and T being disjoint sets, we have

$$\frac{f(S, T, \mathbb{E}[\boldsymbol{d}], n)}{|S \cup T|} \geq (1 - c_{m,n}) \frac{|S|(n - 1 - |T|)}{|S| + |T|} + c_{m,n}\epsilon \left(n - 1 - 2\frac{|S||T|}{|S| + |T|} \right).$$

Using this with (3.19), we have

$$\frac{f(S,T,\mathbb{E}[\boldsymbol{d}],n)}{|S\cup T|} \geq c_{m,n}\epsilon\left(\frac{1}{2}n-1\right).$$

It follows that for $n \geq 3$,

$$\mathcal{E}(\mathbb{E}[\boldsymbol{d}]) \geq \frac{1}{4}c_{m,n}\epsilon.$$

□

Now, take x such that

$$x = \frac{1}{4}c_{m,n}\epsilon \leq \mathcal{E}(\mathbb{E}[\boldsymbol{d}]).$$

For this choice of x, (4.26) reads as

$$c_{m-1} \geq \frac{16c(1+\gamma)}{\epsilon^2}\log(n).$$

Since $c_{m,n} = \frac{2m}{n^2}(1+o(1))$, we have

$$m \geq \frac{8c(1+\gamma)}{\epsilon^2}n\log(n)(1+o(1))$$

and

$$m \geq \frac{\sqrt{c}}{2\gamma}n^{3/2}\sqrt{\log(n)}(1+o(1)).$$

From this it follows that for $\Pr[B] \leq \frac{2}{n^{2c-1}}$, for every fixed $\epsilon \in (0,1]$, it suffices

$$m \geq \max\left\{\frac{8c(1+\gamma)}{\epsilon^2}\sqrt{\frac{\log(n)}{n}}, \frac{\sqrt{c}}{2\gamma}\right\}n^{3/2}\sqrt{\log(n)}(1+o(1)).$$

Thus, m can be chosen such that $m = O(n^{3/2}\sqrt{\log(n)})$ which indeed is sublinear in n^2 but grows faster with n than $n\log(n)$. This scaling comes from the concentration bound for $\sum_{v\neq u}\mathbb{1}_{\{M_{u,v}>0\}}$.

The value of the parameter γ can be optimized by choosing $\gamma^\star$ such that

$$\frac{8c(1+\gamma^\star)}{\epsilon^2}\sqrt{\frac{\log(n)}{n}} = \frac{\sqrt{c}}{2\gamma^\star}.$$

It follows that

$$\gamma^\star = \frac{\epsilon}{4c^{1/4}}\left(\frac{n}{\log(n)}\right)^{1/4}.$$

This yields the sufficient number of experiments:

$$m \geq \frac{2c^{3/4}}{\epsilon} n^{5/4}(\log(n))^{1/4}.$$

Chapter 5

Sketching stochastic valuation functions

5.1 Overview

Evaluation of sets of items arises in various applications such as for ranking and selecting items in assortment optimization, team selection in online gaming, freelancing platforms, web search, and other online platforms. We have mentioned some of the typical examples in the introduction chapter.

In these applications, we often model set outcomes using set valuation functions, defined as function of item values within the group of interest. It is important to enable computing a set valuation function accurately and in a computation cost-efficient manner. A general approach to achieve this is to use compact summaries of items and use these summaries to approximate the underlying valuation function with a sketch valuation function. We call such approximation as sketching. Formally, given a set function u, function v is an α-sketch for u if for every set $S \subseteq N$ we have

$$\alpha v(S) \leq u(S) \leq v(S)$$

for every $S \subseteq \Omega$, for some $\alpha \in (0, 1]$.

Our goal is to find a γ-sketch v for a set function u that allows us to approximate u *every-*

where. At the same time, we are interested in finding good representations $Q_1, \ldots, Q_n$ of respective item value distributions $P_1, \ldots, P_n$ such that we have control of their representation sizes. Importantly, we require that the sketch v can be evaluated by only having access to summaries $Q_1, \ldots, Q_n$. It is desired for these summaries to be compact while allowing for (a) the sketch function to be an α-approximation and (b) efficient evaluation of value queries for sketch function v.

Having an α-approximate sketch valuation function is useful for different optimization problems. For example, consider the best set selection problem that asks to find a set S^* that maximizes $u(S)$ over $S \subseteq \Omega$ subject to the cardinality constraint $|S| = k$. If there exists an algorithm that provides a c-approximation for the best selection problem with respect to a α-sketch function v, then using the output of this algorithm is a αc-approximation for the original best set selection problem. Another example is the welfare maximization problem, where the goal is to find disjoint sets of items that maximize a welfare function defined as the sum of expected group values subject to cardinality constraints. It is of interest in online platforms where individuals are assigned to multiple disjoint groups.

The key problem is how to construct such summaries and how to use them to evaluate the sketch valuation function so that the sketch function provides a good approximation and can be evaluated efficiently for any queried set of items. We note that compact item summaries can be constructed in many different ways. However, our goal is not only to approximate the stochastic valuation function with high accuracy, but also to find good representations for item distributions such that we have control of their representation sizes. It is challenging to attain both requirements at the same time. Moreover, we require the approximation to hold everywhere, not only for the best set, while most of the existing works focus solely on the optimization problem.

5.1.1 Related work

Goemans et al. (2009) were first to formulate the problem of *approximating a submodular*

function everywhere, i.e. approximating its value for points of the domain. Given a value oracle access to a function u on a ground set of size n, the goal is to design an algorithm that performs a polynomial number in n of value queries to the oracle, and then construct an oracle for a function v such that for every set S, $v(S)$ approximates $u(S)$ to within a factor α. The authors have shown that there exists an algorithm that for any non-negative, monotone, submodular function u, achieves approximation factor $\alpha = O(\sqrt{n}\log n)$. It was also shown that no algorithm can achieve a factor better than $\Omega(\sqrt{n}/\log n)$. The approximation function is the *root-linear* function $v(S) = \sqrt{\sum_{i\in S} c_i}$ for some coefficients $c_1, \ldots, c_n$ in $\mathbb{R}_+$. Balcan and Harvey (2011) showed that for some matroid rank functions, a subclass of submodular set functions, every sketch fails to provide an approximation ratio better than $n^{1/3}$. Badanidiyuru et al. (2012) showed that every subadditive set function u has an α-sketch where $\alpha = O(\sqrt{n}\text{polylog}(n))$, and that there is an algorithm that can achieve this with a polynomial number of demand queries. They have also shown that every deterministic algorithm that only has access to a value oracle cannot guarantee a sketching ratio better than $n^{1-\epsilon}$.

The sketches in references discussed so far used geometric constructions, by finding an ellipsoid that approximates well the polymatroid that is associated with u. Cohavi and Dobzinski (2017) showed how to obtain faster and simpler sketches for valuation functions, using an algorithm that finds a $\tilde{O}(\sqrt{n})$ sketch of a submodular set function with only $\tilde{O}(n^{3/2})$ value queries, and an algorithm that finds a $\tilde{O}(\sqrt{n})$ sketch of a subadditive function with $O(n)$ value queries.

The problem of approximating the expected value of a function of independent random variables was studied as early as by Klass (1981), focused on approximating expected value of a function of a sum of independent random variables, by a function that involves expectations only with respect to univariate marginal distributions.

In this chapter, we consider a different class of set valuation functions called stochastic valuation function, defined as the expectation of a valuation function of independent random item values. Various instances of stochastic valuation functions have been con-

sidered in previous works with different aims. Most of them focused on valuation maximization problems subject to some constraints, such as best set selection subject to a cardinality or more general budget constraints, or more general welfare maximization problems.

Asadpour and Nazerzadeh (2016) studied the problem of maximizing a monotone submodular function, defined as the expected value of a monotone submodular value function, subject to a matroid constraint. Kleinberg and Raghu (2018) studied this problem for the special case of cardinality constraints, for the class of test score algorithms, which use one-dimensional representations of item value distributions. They showed that for a sum of top-order statistics objective function, there exist test scores that guarantee a constant-factor approximation. They also showed that a constant-factor approximation is the best achievable. Using a framework based on sketch functions, Sekar et al. (2021) showed that there exist test scores that guarantee a constant-factor approximation for a subset of monotone submodular functions that satisfy an extended diminishing returns property. In particular, they found a $O(\log n)$-approximate sketch function using a k-dimensional test score. To the best of our knowledge, this is the best previously-known sketch for stochastic valuation functions, for monotone submodular functions that satisfy the extended diminishing returns property. Compared to this work, our work achieves a better approximation guarantee (constant factor), at the cost of an extra factor of $\log(n)$ in the representation size. We work with comparable function classes. The current work applies to monotone subadditive or submodular set functions defined for valuation functions satisfying a weak homogeneity or an extendable concavity property, while Sekar et al. (2021) is about monotone submodular set functions for valuation functions satisfying an extended diminishing returns property. Lee et al. (2021) further extended the framework of test scores for stochastic valuation maximization subject to more general budget constraints.

Mehta et al. (2020) showed a PTAS (Polynomial Time-Approximation Scheme) for a stochastic valuation maximization problem with the maximum valuation function subject to a cardinality constraint with budget k, by representing each item's distribution with a

$O(k\log(k))$-size histogram. Our discretization algorithm uses a similar binning strategy as in Mehta et al. (2020) (exponential binning). An important difference is that our algorithm computes a discretized distribution for each item separately without any computations involving multiple items, while Mehta et al. (2020) require using the same binning boundaries for all discretized distributions which are computed by a computation involving all items. This means that our algorithm is more practical and can be computed more efficiently without relying on joint item distributions. Importantly, our guarantees are different as they hold for the problem of approximating a stochastic valuation function everywere, while Mehta et al. (2020) is focused on best set selection problem only for a specific valuation function. We note that neither Lee et al. (2021) nor Mehta et al. (2020) provided results on sketching for approximating a stochastic valuation function everywhere.

Our work is also related to the concept of tensor estimation. Note that we can construct a k-order tensor where the value of each entry corresponds to utility of a set S of size k. Existing works in high-order tensor estimation (Gandy et al. (2011); Tomioka et al. (2010); Shah et al. (2016)) assume independent additive noises for tensor entries, and the goal is find algorithms that achieve consistent estimation for all tensor entries with minimum sample complexity. Compared to the line of works in high-order tensor estimation, we consider a different framework and goal, but share the same idea of approximation everywhere. Our work can be seen as exploring different approaches for tensor estimation, instead of using direct sampling.

Finally, we point to the line of work on data summaries, which considered sketching of various properties of sets, multisets, ordered data, vectors and matrices, and graph data. We refer the reader to the book Cormode and Yi (2020) and the references therein.

5.1.2 Summary of contributions

Our results can be summarized in the following points.

- We present an approximation everywhere guarantee for different classes of stochastic valuation functions. Our work is the first step towards understanding *approximation of stochastic valuation functions everywhere*. The existing related work focused instead on optimization problems only, or approximation schemes using one-dimensional item value distribution representations.

- We show that for weakly homogeneous valuation functions f with degree d and tolerance η, a constant-factor approximation can be guaranteed for approximating a stochastic valuation function everywhere, for any cardinality of a set of items, with a summary of each item's distribution of size $O(n\log(n))$. More specifically, an α-approximate sketch function can be found where α is arbitrarily close to $1/(4\eta)$, with the support size of each discretized distribution $s = O((1/d)k\log(k))$ for sets of cardinality less than or equal to k. Several commonly used valuation functions are weakly homogeneous with degree $d = 1$ and tolerance $\eta = 1$. Hence, for these valuation functions, we have an α-sketch with α arbitrarily close to $1/4$ and $s = O(k\log(k))$. By extending the approximation guarantee to other conditions and using univariate transformations, we are able to cover a wide range of stochastic valuation functions. Note that most work on best set selection problem using score representations of item value distributions is focused on specific functions, such as maximum value, concave function of the sum of values, and other similar functions. Our work goes much beyond this in providing results for a class of functions, which include all these special functions.

- The discretized distributions are computed by using an algorithm based on the well-known concept of exponential binning, as, for example, used in Mehta et al. (2020). This algorithm uses two input parameters, ϵ and a, that allow us to control the size of the support of the output discretized distribution. For an item value distribution, the algorithm outputs a discretized distribution with support of size $s = O((1/\epsilon)\log(1/a))$. Our work shows that this disretization algorithm can provide a constant-factor approximation for the problem of approximating a stochastic valuation set function everywhere. Previous work Mehta et al. (2020) showed

that similar discretization algorithm can achieve an approximation guarantee for a different problem, maximizing the expected value of the maximum of random item values subject to a cardinality constraint. Our results are established by using different proof techniques as we consider the problem of approximating a stochastic valuation function everywhere, and allow for a more general class of valuation functions. The disretization algorithm uses value oracle access to item value distribution. This is a natural and mild assumption in applications where an item value distribution is the empirical distribution of item values in a dataset. In any case, the algorithm uses simple properties of a distribution such as quantile values, which can be efficiently estimated from samples when item value distributions are unknown.

- Our numerical results, obtained by using both randomly generated and real-world data, validate various hypotheses suggested by our theoretical analysis and demonstrate that accurate function approximations can be obtained by our proposed method.

5.2 Problem formulation

Let $\Omega = \{1, \ldots, n\}$ be a ground set of items. Each item $i \in \Omega$ has a value according to a random variable X_i with distribution P_i, and item values $X_1, \ldots, X_n$ are assumed to be independent. We consider the class of stochastic valuation functions, which for a ground set of items Ω, are defined as

$$u(S) = \mathbb{E}[f(X_S)] \text{ for } S \subseteq \Omega$$

where $f : \mathbb{R}_+^n \to \mathbb{R}_+$ is a monotone function and X_S is a n-dimensional vector with the i-th component equal to X_i if $i \in S$ and is equal to 0, otherwise.

Alternatively, we can write the stochastic valuation function as

$$u(S) = \mathbb{E}[f((X_i, i \in S))] \tag{5.1}$$

where for any $S \subseteq \Omega$ and $x \in \mathbb{R}^n$, we write $(x_i, i \in S)$ to denote $z \in \mathbb{R}^n$ such that $z_i = x_i$ if $i \in S$, and $z_i = 0$, otherwise. We will use these two forms interchangeably.

We have two specific goals in mind.

First goal We aim to find good representations of items for approximating the stochastic valuation function $u(S)$. We compute $Q_1, \ldots, Q_n$ as representations of $P_1, \ldots, P_n$ which correspond to distributions of some new discrete random variables $Y_1, \ldots, Y_n$. Then we compute a sketch set function $v(S)$ defined as the expected value of function f with respect to item value distributions $Q_1, \ldots, Q_n$. The sketch function should approximate the stochastic valuation function $u(S)$ for any set of items S within a multiplicative approximation error tolerance, i.e., for some $\beta \geq \alpha > 0$,

$$\alpha v(S) \leq u(S) \leq \beta v(S).$$

When this guarantee holds we say that v is an (α, β)-*approximation* of u that allows us to approximate u everywhere. Note that $\alpha v(S) \leq u(S) \leq \beta v(S)$ can always be interpreted as an approximation guarantee $\gamma \tilde{v}(S) \leq u(S) \leq \tilde{v}(S)$ where $\gamma = \alpha/\beta$ and $\tilde{v}(S) = \beta v(S)$. In this case we say that $\tilde{v}$ is a γ-approximation of u.

Second goal We are also interested in two optimization problems: best set selection and welfare maximization. The best set selection problem asks to find a set S^* that maximizes $u(S)$ over $S \subseteq \Omega$ subject to cardinality constraint $|S| = k$. A set S is said to provide a c-approximation for the best selection problem of function u if for some $c > 0$,

$$u(S) \geq c \cdot \max\{u(S) : |S| = k\}$$

where sketch v is evaluated by only having access to summaries $Q_1, \ldots, Q_n$. Preferably, the approximation factor should not depend on the cardinality of set k and we have control of the representation size.

The welfare maximization problem is a strict generalization of the best set selection problem. Specifically, we are given a positive integer m and cardinality constraints $k_1, \ldots, k_m$, the goal is to find disjoint sets $S_1, \ldots, S_m \subseteq \Omega$ of cardinalities $k_1, \ldots, k_m$ that maximize $\sum_{j=1}^{m} u_j(S_j)$, where $u_1, \ldots, u_m$ are monotone submodular set functions.

5.3 Approximation everywhere guarantees

In this section, we first present a discretization algorithm and then study its approximation guarantees for approximating the set function $u(S) = \mathbb{E}[f((X_i, i \in S))]$. We assume that valuation function f and items' value distributions $P_1, \ldots, P_n$ satisfy the following condition: $\mathbb{E}[f(X_i) \mid X_i > \tau]$ is finite, for all $i \in \Omega$ and $\tau \in \mathbb{R}_+$. This condition is used to summarise the tail of an item's value distribution. For some valuations functions, such as maximum value ($f(x) = \max\{x_1, \ldots, x_n\}$) and CES[1] function ($f(x) = (x_1^r + \cdots + x_n^r)^{1/r}$, for $r > 0$), this condition is equivalent to $\mathbb{E}[X_i]$ being finite for all $i \in \Omega$.

We also assume that distributions $P_1, \ldots, P_n$ have a mass on any atom bounded by $\Delta \in [0, 1)$, i.e. for all $i \in \Omega$,

$$P_i(x) - \lim_{z \uparrow x} P_i(z) \leq \Delta, \text{ for all } x \in \mathbb{R}. \tag{5.2}$$

In fact, it suffices that (5.2) holds only for $x = \tau_i$ where τ_i is the $(1-\epsilon)$-quantile of P_i. If for each $i \in \Omega$, P_i is continuous and strictly increasing on its support, then $\Delta = 0$.

Under this assumption, for all $x \in \mathbb{R}$, we have for all $i \in \Omega$,

$$1 - \epsilon \leq \mathbb{P}(X_i \leq \tau_i) \leq 1 - \epsilon + \Delta. \tag{5.3}$$

[1]CES refers to *constant elasticity of substitution*, which is a terminology used in economic theory literature.

5.3.1 Discretization algorithm

We consider a discretization algorithm that transforms input distributions $P_1, \ldots, P_n$ to discrete distributions $Q_1, \ldots, Q_n$ with finite supports. For each item's distribution, this discretization algorithm first constructs a random variable with distribution that has a bounded support, and then uses an exponential binning to obtain final discretized distribution. For each random variable X_i we define τ_i to be the $(1-\epsilon)$-quantile of its distribution P_i, i.e. $\tau_i = \inf\{x \in \mathbb{R} : P_i(x) \geq 1-\epsilon\}$, where ϵ is a parameter in $(0,1]$. The method first limits the upper end of the support of each item's value distribution. This is done by defining $H_i = \mathbb{E}[f(X_i) \mid X_i > \tau_i]$ for each $i \in \Omega$, and letting $\hat{X}_i$ be a new random variable that is equal to X_i if $X_i \leq \tau_i$ and, is equal to $f^{-1}(H_i)$, otherwise. Here f^{-1} denotes the inverse of function $f(x, 0, \ldots, 0)$ with respect to x. Note that $\hat{X}_i$ has distribution with support contained in $[0, \tau_i] \cup \{f^{-1}(H_i)\}$. The method then limits the lower end of the support by assigning values of $\hat{X}_i$ smaller than $a\tau_i$ to 0, where $a \in (0,1)$ is a parameter. This results in a new random variable $\tilde{X}_i = \hat{X}_i \mathbb{1}_{\{\hat{X}_i > a\tau_i\}}$, whose support is contained in $[a\tau_i, \tau_i] \cup \{f^{-1}(H_i)\}$.[2] Finally, each random variable $\tilde{X}_i$ is transformed by using an exponential binning of the interval $[a\tau_i, \tau_i]$ and mapping each value in a bin to the lower boundary of this bin. Formally, let q be the *quantization function* defined as $q(x; \tau, \epsilon, a) = a\tau/(1-\epsilon)^{j-1}$, for $x \in I_j(\tau, \epsilon, a)$ and $1 \leq j \leq l$, where l is the largest integer j' such that $j' \leq \log_{1/(1-\epsilon)}(1/a)$ and $I_j(\tau, \epsilon, a) = (a\tau/(1-\epsilon)^{j-1}, a\tau/(1-\epsilon)^{j}]$. Then, $Y_i = q(\tilde{X}_i; \tau_i, \epsilon, a)$. Note that Y_i is a random variable with discrete distribution Q_i with finite support of size

$$s = O\left(\frac{1}{\epsilon}\log(1/a)\right).$$

It is noteworthy that the discretization algorithm only uses two properties of an item's value distribution to compute the discretized distribution, specifically, for each item $i \in \Omega$, it uses (a) the value of the $(1-\epsilon)$-quantile of the input distribution P_i and (b) the value of $H_i = \mathbb{E}[f(X_i) \mid X_i > \tau_i]$. The discretized distribution for each item can be computed independently, not requiring any joint computation over the ground set of items. This is

[2] Hereinafter, $\mathbb{1}_A$, for some event A, is equal to 1 if A is true, and is equal 0, otherwise.

a desirable property for practical applications, especially in cases when the ground set of items can change over time.

The discretization algorithm can be efficiently implemented in distributed systems when an item's value distribution corresponds to the empirical distribution of values in a multiset partitioned across nodes in a distributed system. Computing a discretized distribution requires to compute a quantile value and evaluate range queries for a multiset of values, both of which can be efficiently computed.

5.3.2 Guarantees for weakly homogeneous functions

We will show approximation guarantees for a class of functions that satisfy a weak homogeneity condition. Recall that a function f is *weakly homogeneous of degree d and tolerance η over a set* $\Theta \subseteq \mathbb{R}$ if

$$(1/\eta)\, \theta f(x) \leq f(\theta x) \leq \theta^d f(x)$$

for every x in the domain of f and all $\theta \in \Theta$. Many functions are weakly homogeneous with a positive degree and tolerance equal to 1. In Table 5.1 we show properties of some functions f. In the table, elasticity of a differentiable function $g : \mathbb{R} \to \mathbb{R}$ at a point z is defined as $zg'(z)/g(z)$. Most of the functions in the table are introduced earlier in section 2.2. The last one is called the success probability function. We will prove its properties in section 5.7.

Table 5.1: Properties of some functions f.

$f(x)$	subadditive	submodular	convex	concave	d	η
$\max\{x_1, \ldots, x_n\}$	✓	✓	✓		1	1
$f(x) = x_{(1)} + \cdots + x_{(h)}$ *	✓	✓	✓		1	1
$f(x) = (\sum_{i=1}^n x_i^r)^{1/r}, r \geq 1$	✓	✓	✓		1	1
$f(x) = g(\sum_{i=1}^n x_i)$, concave g	✓	✓		✓	g min elasticity	1
$f(x) = 1 - \prod_{i=1}^n (1 - x_i)$	✓	✓			$\leq 1/2$, for $n \geq 2$	1

* $x_{(i)}$ denotes the i-th element of a sequence corresponding to values $x_1, \ldots, x_n$ sorted in decreasing order

Next, we show the approximation guarantee obtained for the class of weakly homogeneous functions. We will also provide key lemmas highlighting some of the main points

of the proof, with the full proof provided in section 5.7.

The main theorem is provided as follows.

Theorem 5.3.1. *Assume that f is a monotone subadditive or submodular function, and is weakly homogeneous with degree d and tolerance η over $[0,1]$, and $\epsilon \in (\Delta, 1)$. Then, the discretization algorithm guarantees that for every set $S \subseteq \Omega$ such that $|S| \leq k$, we have*

$$\frac{1}{2}(1-\epsilon)^{k-1}(1-\Delta/\epsilon)v(S) \leq u(S) \leq 2\eta \frac{1+a^d k/(\epsilon-\Delta)}{(1-\epsilon)^k(1-\Delta/\epsilon)} v(S).$$

The approximation factors in Theorem 5.3.1 depend on parameters d and η specifying a subset of monotone functions that are either subadditive or submodular, to which the theorem applies. The approximation factors also depend on the parameters of the algorithm, namely a and ϵ, as well as on the set cardinality k, and parameter Δ.

Theorem 5.3.1 implies a constant-factor approximation guarantee for any function f that is weakly homogeneous with a constant tolerance η and $\Delta = o(1/k)$ by choosing the algorithm's parameters a and ϵ appropriately.

Corollary 5.3.2. *Assume that $\Delta k < 1$. Under the same conditions as in Theorem 5.3.1 and taking $a = [\epsilon(\epsilon-\Delta)]^{1/d}$ and $\epsilon = c/k$, for some constant $c \in (\Delta k, 1)$, for every set $S \subseteq \Omega$ such that $|S| \leq k$,*

$$\frac{1}{2}\psi(c, \Delta k/c)v(S) \leq u(S) \leq 2\eta \frac{1}{\psi(c, \Delta k/c)}(1+c)v(S)$$

where $\psi(c,\delta) := e^{-\frac{c}{1-c}}(1-\delta)$.

Note that when $\Delta = 0$, the approximation factors α and β can be made arbitrarily close to $1/2$ and 2η, respectively, by taking c small enough. If $\Delta = o(1/k)$, each discretized distribution has the support of size $s = O\left((1/d)k\log(k)\right)$. For weakly homogeneous valuation functions with the degree lower bounded by a positive constant, we have the support sizes $O(k\log(k))$. As discussed previously, classes of weakly homogeneous functions with a constant degree include homogeneous and convex functions.

We argue that in general, the dependence on the degree parameter d is unavoidable due to assigning all values smaller than $a\tau_i$ to 0 for an item $i \in \Omega$. Intuitively, this can cause an excessive loss of approximation accuracy for functions with a small degree of weak homogeneity, when distributions of item values have a sufficient mass near to zero. To demonstrate this, we consider the simple example of the scalar function $f(x) = x^r$ on $\mathbb{R}_+$, for a parameter $r \in (0, 1]$. Let X be a random variable with cumulative distribution $P(x)$ with support in $\mathbb{R}_+$, and let $P(\tau) = 1 - \epsilon$. For any $a \in [0, 1]$, we can write

$$\mathbb{E}[(X\mathbb{1}_{\{X \geq a\tau\}})^r] \;\; = \;\; \rho\, \mathbb{E}[X^r]$$

where

$$\rho = \frac{\mathbb{E}[X^r \mathbb{1}_{\{X \geq a\tau\}}]}{\mathbb{E}[X^r]} = \frac{\int_{a\tau}^{\infty} x^r dP(x)}{\int_0^{\infty} x^r dP(x)}.$$

Consider an instance where $P(x) = x^d$ for $x \in [0, 1]$, for a parameter $d > 0$. Note that $\tau^d = 1 - \epsilon$. By simple calculus, we have

$$\rho = 1 - (a(1 - \epsilon)^{1/d})^{r+d}.$$

Assuming $a = \epsilon^c$ for some fixed $c > 0$, and $r = d = \epsilon$, we have

$$\rho = 1 - \epsilon^{2c\epsilon}(1 - \epsilon)^2 \downarrow 0 \;\text{ as }\; \epsilon \downarrow 0.$$

Next, we present a sketch proof highlighting key lemmas for the theorem. Recall that the algorithm consists of three main steps: limiting the upper end, removing the lower end and exponential binning of the middle part. The main idea of the proof is that for each step, we compare the set value of the new random variables with the original one and show that we are not too far from it. At different steps in our proof, we leverage properties of the class of valuation functions that we consider to derive desired bounds.

Upper end For each $i \in \Omega$, by definition of $\hat{X}_i$, we have

$$\mathbb{E}[f(\hat{X}_i) \mid \hat{X}_i > \tau_i] = \mathbb{E}[f(X_i) \mid X_i > \tau_i].$$

Let $w(S) = \mathbb{E}\left[f\left((X_i \mathbb{1}_{\{X_i \leq \tau_i\}}, i \in S)\right)\right]$. We have the following lemma.

Lemma 5.3.3. *Assume that f is a monotone function that is either subadditive or submodular on its domain. Then, for every $S \subseteq \Omega$ such that $|S| \leq k$,*

$$u(S) \geq (1-\epsilon)^{k-1}(1-\Delta/\epsilon)\max\left\{\epsilon \sum_{i \in S} H_i, w(S)\right\}$$

and

$$u(S) \leq 2\max\left\{\epsilon \sum_{i \in S} H_i, w(S)\right\}.$$

Note that the upper and lower bounds both apply to the truncated random variable $\hat{X}_i$. We can compare the set function $v_1(S) = \mathbb{E}[f((\hat{X}_i, i \in S))]$ to the set function $u(S) = \mathbb{E}[f((X_i, i \in S))]$ and obtain the following lemma.

Lemma 5.3.4. *Assume that f is a monotone function that is either subadditive or submodular on its domain. Then, for every $S \subseteq \Omega$ such that $|S| \leq k$,*

$$\frac{1}{2}(1-\epsilon)^{k-1}(1-\Delta/\epsilon)v_1(S) \leq u(S) \leq 2\frac{1}{(1-\epsilon)^{k-1}}(1-\Delta/\epsilon)^{-1}v_1(S).$$

Lower end We next consider random variables defined as $\tilde{X}_i := \hat{X}_i \mathbb{1}_{\{\hat{X}_i \geq a\tau_i\}}$, for some $a \in [0,1]$. We compare the set function $v_2(S) = \mathbb{E}[f((\tilde{X}_i, i \in S))]$ and the set function $v_1(S) = \mathbb{E}[f((\hat{X}_i, i \in S))]$ in the following lemma.

Lemma 5.3.5. *Assume that f is a monotone function that is either subadditive or submodular, and is weakly homogeneous of degree d over $[0,1]$. Then, for every set $S \subseteq \Omega$ such that $|S| \leq k$, we have*

$$v_2(S) \geq \frac{1}{1 + a^d k/(\epsilon - \Delta)} v_1(S).$$

A key point in the proof is that we use binning boundaries that can be different for different item value distributions and distinguish each item based on whether or not its value exceeds an item-specific value. Note that for the maximum value function, the tail value dominates and it suffices to use a common tail binning boundary to achieve sufficiently large set value.

Middle part We next consider the approximation error due to the last step of the algorithm. Recall that the exponential binning partitions the range into l intervals and each random variable $\tilde{X}_i$ is transformed in a way that each value in a bin is mapped to the lower boundary of the bin, i.e., for each $i \in \Omega$, $Y_i = q(\tilde{X}_i; \tau_i, \epsilon)$.

We compare the set functions $v_2(S) = \mathbb{E}[f((\tilde{X}_i, i \in S))]$ and $v(S) = \mathbb{E}[f((Y_i, i \in S))]$.

Lemma 5.3.6. *Assume that f is monotone and weakly homogeneous with tolerance η. Then, we have*

$$v(S) \leq \frac{1-\epsilon}{\eta} v_2(S).$$

Putting the pieces together Combining the steps above, we obtain a constant factor approximation guarantee for the discretization strategy. Specifically, the lower bound in the theorem follows from Lemma 5.3.4. The upper bound in the theorem follows by combining Lemmas 5.3.5 and 5.3.6.

5.3.3 Extension to other function classes

As discussed in table 5.1, the class of weakly homogenous functions already covers a number of common valuation functions. Next, we show that it is possible to extend the approximation guarantee to other conditions and random variables under univariate transformation. In this way, we are able to cover a even wider range of stochastic valuation functions.

Extendable concave functions The weak homogeneity condition restricts approximation guarantees to functions with a strictly positive degree of homogeneity. We here show that approximation guarantees can be provided for some concave functions that have zero degree of homogeneity.

A monotone subadditive and concave function f on $\mathbb{R}^n_+$ is said to have an *extension on* $\mathbb{R}^n$ if there exists a function f^* that is monotone subadditive and concave on $\mathbb{R}^n$ and $f^*(x) = f(x)$ for all $x \in \mathbb{R}^n_+$. In the next theorem we show an approximation guarantee for functions f that have extensions on $\mathbb{R}^n$.

Theorem 5.3.7. *Assume that f is a monotone subadditive, concave function on $\mathbb{R}^n_+$ that has an extension on $\mathbb{R}^n$, and $\epsilon \in (\Delta, 1)$. Then, the discretization algorithm guarantees that for every set $S \subseteq \Omega$ such that $|S| \leq k$, we have*

$$\frac{1}{2}(1-\epsilon)^{k-1}(1-\Delta/\epsilon)v(S) \leq u(S) \leq 2\frac{1+ak/(\epsilon-\Delta)}{(1-\epsilon)^k(1-\Delta/\epsilon)}v(S).$$

The proof strategy is similar to the case of weakly homogeneous functions. The full proof is shown in section 5.7. We also have the following corollary.

Corollary 5.3.8. *Assume that $\Delta k < 1$. Under same conditions as in Theorem 5.3.7 and taking $a = \epsilon(\epsilon - \Delta)$ and $\epsilon = c/k$, for some constant $c \in (\Delta k, 1)$, for every set $S \subseteq \Omega$ such that $|S| \leq k$,*

$$\frac{1}{2}\psi(c, \Delta k/c)v(S) \leq u(S) \leq 2\frac{1}{\psi(c, \Delta k/c)}(1+c)v(S).$$

Theorem 5.3.7 alleviates the need for the weak homogeneity condition for some concave functions, and covers some concave functions which do not satisfy this condition with a positive degree. For example, consider again $f(x) = g(\sum_{i=1}^n x_i)$ on $\mathbb{R}_+$ with $g(z) = 1 - e^{-\lambda z}$ for parameter $\lambda > 0$, and $z \in \mathbb{R}_+$. Recall that this function has the weak homogeneity degree of value 0 and hence the results in previous sections cannot be applied. However,

note that f has an extension on $\mathbb{R}^n$, e.g. given as $f^*(x) = g^*(\sum_{i=1}^n x_i)$ where $g^*(z) = 1 - e^{-\lambda z}$ for $z \geq 0$ and $g^*(z) = \lambda z$, otherwise.

Not all concave functions have an extension on $\mathbb{R}^n$. Consider the last example when $g(z)$ has a vertical tangent at $z = 0$. If $g(z)$ is differentiable at $z = 0$, then this is equivalent to $\lim_{z \downarrow 0} dg(z)/dz = \infty$. In this case, f does not have an extension on $\mathbb{R}^n$. An example when this is the case is when g is a power function $g(z) = z^r$, with $r \in (0, 1)$.

Coordinate-wise conditions In Section 5.3.2, we have shown an approximation guarantee for functions f that satisfy a weak homogeneous condition. In this section we show that similar approximation guarantees can be established for functions that satisfy the weakly homogeneous property only coordinate-wise.

A function f is said to be *coordinate-wise weakly homogeneous of degree d and tolerance η over a set $\Theta \in \mathbb{R}$* if for every $i \in [n]$,

$$(1/\eta)\, \theta f(x) \leq f\left(\sum_{j \neq i} x_j e_j + \theta x_i e_i\right) \leq \theta^d f(x),$$

for every x in the domain of f and all $\theta \in \Theta$.

Theorem 5.3.9. *Assume that f is a monotone subadditive or submodular function, and is coordinate-wise weakly homogeneous with degree d and tolerance η over $[0, 1]$ and $\epsilon \in (\Delta, 1)$. Then, the discretization algorithm guarantees that for every set $S \subseteq \Omega$ such that $|S| \leq k$, we have*

$$u(S) \geq \frac{1}{2}(1 - \epsilon)^{k-1}(1 - \Delta/\epsilon)v(S)$$

and

$$u(S) \leq 2\eta^k \frac{1 + a^d k/(\epsilon - \Delta)}{(1 - \epsilon)^{2k}} v(S).$$

From this theorem, we have the following corollary.

Corollary 5.3.10. *Assume that $\Delta k < 1$. Under same conditions as in Theorem 5.3.9 and such that $\eta = 1$, by taking $a = [\epsilon(\epsilon - \Delta)]^{1/d}$ and $\epsilon = c/k$, for some constant $c \in (\Delta k, 1)$, for every*

set $S \subseteq \Omega$ *such that* $|S| \leq k$,

$$\frac{1}{2}\psi(c, \Delta k/c)v(S) \leq u(S) \leq 2\frac{1}{\psi(c, \Delta k/c)}(1+c)v(S).$$

Note that any function f that is subadditive and coordinate-wise convex on a domain that includes 0 and is such that $f(0) = 0$, is weakly homogeneous over $[0, 1]$ with tolerance $\eta = 1$. And any function f that is coordinate-wise concave on a domain that includes 0 and is such that $f(0) \geq 0$ is weakly homogeneous over $[0, 1]$ with tolerance $\eta = 1$.

Univariate transformations For any given function f, we may establish approximation guarantees by validating conditions of the theorems in previous sections for a function f^* such that $f^*(x_1, \ldots, x_n) = f(\phi_1(x_1), \ldots, \phi_n(x_n))$ for some continuous and strictly increasing functions $\phi_1, \ldots, \phi_n$. The univariate transformations $\phi_1, \ldots, \phi_n$ correspond to a change of variables that only affects the input distributions. Using univariate transformations can be useful in some cases. We illustrate this by two examples.

First, let us consider the case when $f(x) = (\sum_{i=1}^n x_i)^r$ with $r \in (0, 1)$. This function is weakly homogeneous over $[0, 1]$ with degree r. We can apply Corollary 5.3.2 to obtain a constant-factor approximation, with discretized distributions having supports of size $O((1/r)k\log(k))$. We can avoid having this dependence on r by using univariate transformations $\phi_i(z) = z^{1/r}$. We thus need to validate conditions for $f^*(x) = (\sum_{i=1}^n x_i^{1/r})^r$, with $r \in (0, 1)$. Function f^* is subadditive, submodular, convex, and weakly homogeneous over $[0, 1]$ with degree 1 and tolerance 1. Thus, by Corollary 5.3.2, we have a constant-factor approximation with discretized distributions having supports of size $O(k\log(k))$.

Second, consider the case when $f(x) = 1 - \prod_{i=1}^n (1 - x_i)$ on $[0, 1]^n$. This function is submodular and is weakly homogeneous over $[0, 1]$ with degree $d \leq 1/2$ and tolerance 1. We elaborate on these properties in Section 5.7. Again, we can apply Corollary 5.3.2 which gives a constant-factor approximation with $O((1/d)k\log(k))$ support size of discretized

distributions. We can remove this dependence on d, by considering the transformations $\phi_i(z) = 1 - e^{-z}$. Hence, we have $f^*(x) = 1 - e^{-\sum_{i=1}^{n} x_i}$. We can apply Corollary 5.3.8 to show that a constant-factor approximation holds with $O(k\log(k))$ support size of discretized distributions.

5.4 Sketching for optimization problems

In this section we discuss application of our function approximation results to best set selection and submodular welfare maximization problems, which are defined as follows. The *best set selection* problem asks to find a set $S^* \subseteq \Omega$ such that $S^* \in \arg\max_{S\subseteq\Omega:|S|=k} u(S)$, for given cardinality constraint parameter k. A set S is said to be a ρ-approximate solution for the best set selection problem if $u(S) \geq \rho u(S^*)$. By Sekar et al. (2021), if v is a (α, β)-approximation of u and S is a ρ-approximation solution for the best set selection problem with objective function v, then S is a $\rho\alpha/\beta$-approximate solution for the best set selection problem with objective function u. This guarantee holds, even more generally, for the *submodular welfare maximization* problem, where given a positive integer m and cardinality constraint parameters $k_1, \ldots, k_m$, the goal is to find disjoint sets $S_1, \ldots, S_m \subseteq \Omega$ of cardinalities $k_1, \ldots, k_m$ that maximize $\sum_{j=1}^{m} u_j(S_j)$, where $u_1, \ldots, u_m$ are monotone submodular set functions.

It is well known that a greedy algorithm provides a $(1 - 1/e)$-guarantee for the best set selection problem for any monotone submodular objective function Nemhauser et al. (1978b). This greedy algorithm starts with an empty set and adds one item per step to this set, in each step choosing an item that maximizes the marginal value gain. A similar greedy algorithm provides a $1/2$-guarantee for the submodular welfare maximization problem Lehmann et al. (2006).

For a set function of the form (5.1), with probability distributions of item values having a finite support, each of size at most s, evaluating $u(S)$ for a set S of cardinality k has s^k computation complexity. The computation complexity of the greedy algorithm using

value oracle calls for the set function of the form (5.1) with distributions of item values having supports of size at most s is $O(ns^k)$. This is easily seen as follows. In each step $t \in \{1, \ldots, k\}$, the algorithm needs to compute values of $n - (t-1)$ set functions, each for a set of cardinality t. Hence, the total computation complexity is $O(ns^k)$. Clearly, if $s = O(1)$ and $k = O(1)$, then the greedy algorithm has $O(n)$ complexity. The greedy algorithm has a polynomial complexity $O(n^{1+\epsilon})$ for some positive constant ϵ, if, and only if, $s^k = O(n^\epsilon)$. For example, this holds if $s = O(k \log(k))$ and $k \leq \epsilon \log(n) / \log(\log(n))$.

We have the following implication of Corollary 5.3.2.

Corollary 5.4.1. *Assume that $\Delta k < 1$. For the class of functions satisfying conditions of Theorem 5.3.1, and by taking $a = [\epsilon(\epsilon - \Delta)]^{1/d}$ and $\epsilon = c/k$, for some $c \in (\Delta k, 1)$, greedy algorithms for best set selection and submodular welfare maximization problems guarantee the approximation ratio*

$$\frac{\psi(c, \Delta k/c)^2}{1+c} \frac{\rho}{4\eta}$$

where ρ is a constant, which for best set selection problem is equal to $1 - 1/e$, and for submodular welfare maximization problem is equal to $1/2$.

If $\Delta = 0$, this approximation ratio can be made arbitrarily close to $\rho/(4\eta)$ by taking c small enough.

5.5 Numerical results

In this section, we present results of our numerical experiments. The goal is to assess the performance of the sketch under various assumptions on item value distributions, set utility functions, set size, and parameters of the discretization algorithm. We also compare the performance against the baseline method based on test scores proposed in Sekar et al. (2021) and demonstrate that our sketch outperforms this baseline in terms of approximation accuracy.

We have performed our experiments on both synthetic and real-world data sets. The results for synthetic data is reported in section 5.5.1 and results for real-world data set is reported in section 5.5.2. The code we use is available on GitHub: `https://github.com/Sketch-EXP/Sketch`.

5.5.1 Synthetic data

We fix a ground set of n elements. For each element, we generate $N = 500$ training samples of its random performance values and estimate the value of the set utility function $u(S)$. Then we choose parameter ϵ of our discretization algorithm and compute the value of the sketch $v(S)$. To assess the performance of our algorithm, we randomly generate 50 sets of size k from the ground set of n elements and estimate the ratio $v(S)/u(S)$.

Our experimental setups are given as follows.

Set utility functions We examine three types of set utility functions, the maximum value, the CES function with degree 2, and the square root function of sum.

Item value distributions We consider two parametric families of distributions, the exponential and Pareto distribution. For exponential distribution, we sample the mean value of each item uniformly from the unit interval $[0, 1]$. For Pareto distribution, we sample the shape parameter of each item uniformly from the interval $[1.1, 3]$ and fix the scale parameter to 1.5.

Set size For each setting of set utility function and item value distribution, we tested various values from 1 to 20 for the set size k.

Threshold value The threshold value ϵ is an important parameter for the discretization algorithm. The lower the threshold value, the better the approximation ratio. For each value of set size k, we set $\epsilon = c/k$ where c is some constant between 0.1 and 10. From the main theorem, we expect good approximation ratios for those $c < 1$.

Results We first show a box plot (figure 5.1) aggregating the results from all settings of set size k for different set utility functions and item value distributions fixing $c = 0.1$. We can observe that the ratio values are concentrated around 1, thus our sketch approximates the original set utility function well for most instances.

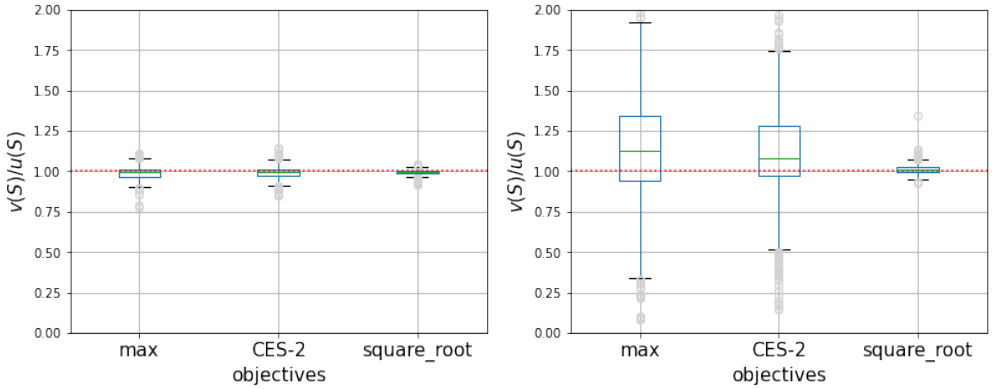

Figure 5.1: Performance ratio for various objective functions and item value distributions: (left) exponential distributions and (right) Pareto distributions.

Dependence on threshold value ϵ Figure 5.2 shows the results under different values of ϵ aggregated over all settings of set size k. Overall, we can see that the ratio starts to deteriorate from around 1, i.e. the value $\epsilon = 1/k$, regardless of the set size k.

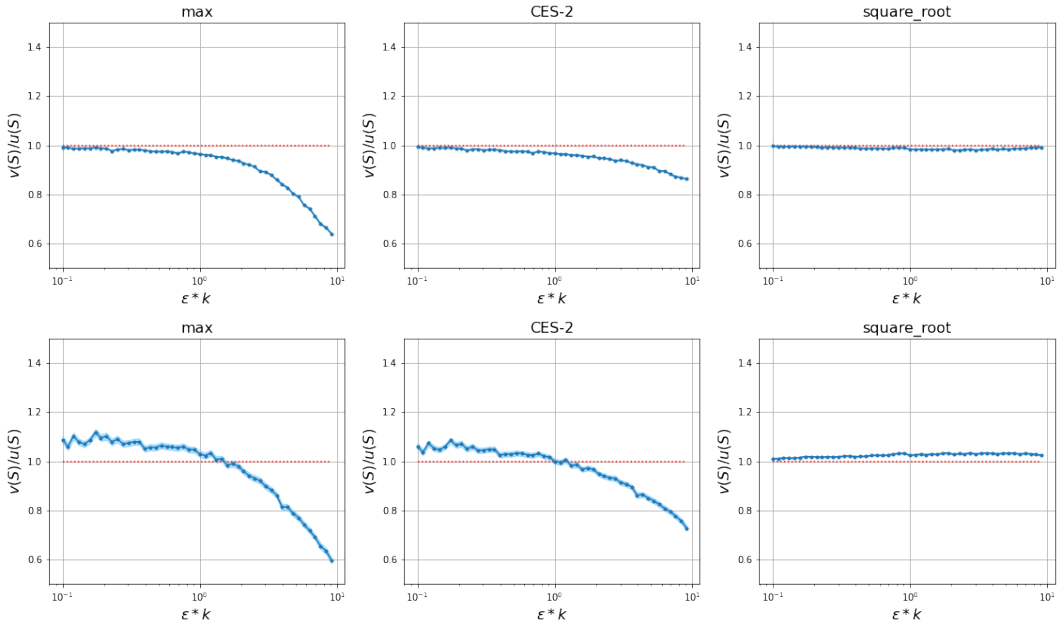

Figure 5.2: Results showing effect of different values of ϵ: (top) exponential distribution and (bottom) Pareto distribution.

Comparison with the test-score sketch To compare the results with the benchmark, we pro-

vide the bar plot (Figure 5.3) which shows the averaged ratio values for both methods. We can conclude from the plot that our sketch outperforms the test score sketch in terms of approximation accuracy.

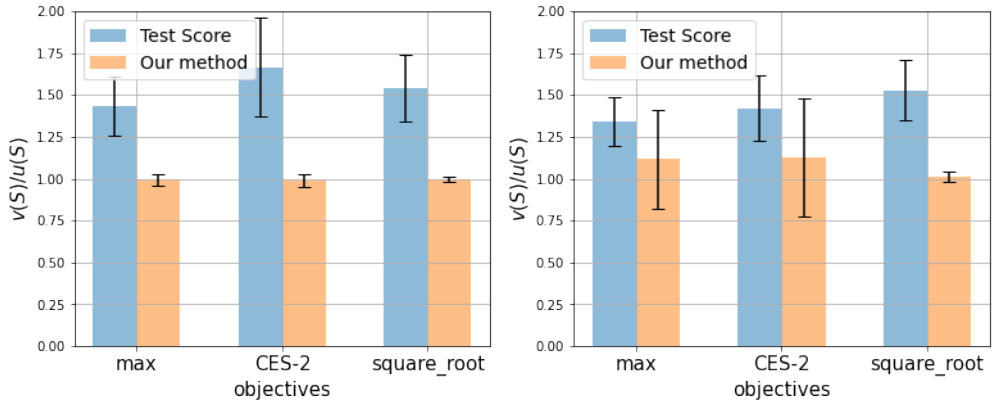

Figure 5.3: Performance of discretization v.s. test score: (left) exponential distributions and (right) Pareto distributions.

5.5.2 Real data

We tested our method on three real-world datasets: YouTube, StackExchange, and New York Times data. For the YouTube data, we consider items to be content publishers and their performance to be the number of views of their content pieces. For the StackExchange data, we consider items to be experts and performance to be the rate of up-votes of their answers to questions. For the New York Times data, we consider items to be news sections and performance to be the number of comments per news piece. All datasets that we use are available in the public domain. Detailed information about the three datasets are provided below.

YouTube data The YouTube dataset Kaggle.com (2021) contains information about 37422 unique videos, including publication date, view counts, number of likes and dislikes, for the period from August 2020 to December 2021, for the USA, Canada and Great Britain. For our experiments, we filtered out YouTubers with fewer than 50 uploads. In the main

experiment, we took the view counts per day as the measure for video performance. We also tested other metrics and the results can be found in section 5.8.1.

StackExchange data The StackExchange dataset contains information about 35218 questions and 88584 answers on the Academia.StackExchange platform. The dataset is retrieved on Jan 20, 2022 from the official StackExchange data dump. Each answer receives up-votes and down-votes from users of the platform, indicating quality of the answer. For our experiments, we took only the users who have submitted as least 100 answers. If an answer a to question q receives $u(a,q)$ up-votes and $d(a,q)$ down-votes, then we define

$$s(a,q) = \frac{u(a,q) + c_1}{u(a,q) + d(a,q) + c_2}$$

as the quality value of the answer, where c_1 and c_2 are positive-valued parameters. This metric is motivated by Bayesian estimation, and was used in Sekar et al. (2021). The ratio increases with the number of up-votes and decreases with the number of down-votes. It is called balanced when $c_1/c_2 = 1/2$, conservative if $c_1/c_2 < 1/2$. We took the conservative choice $(c_1, c_2) = (2, 8)$ for the main experiment. Results for other value pairs can be found in Section 5.8.2.

New York Times data The New York Times dataset Kaggle.com (2020) contains information about 16570 articles and comments on New York Times from January 2020 to December 2020. Each article belongs to one section. We took all articles and their comment numbers for our experiment.

We show the empirical CDFs for performance values aggregated over all data points in a dataset, for all three datasets in Figure 5.4. We observe that these three datasets have very different distributions, which implies that our method works well for different data distributions.

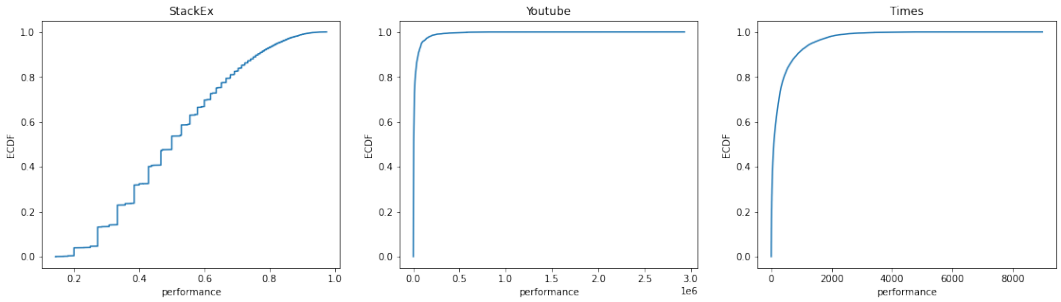

Figure 5.4: Empirical CDFs for performance values of three datasets.

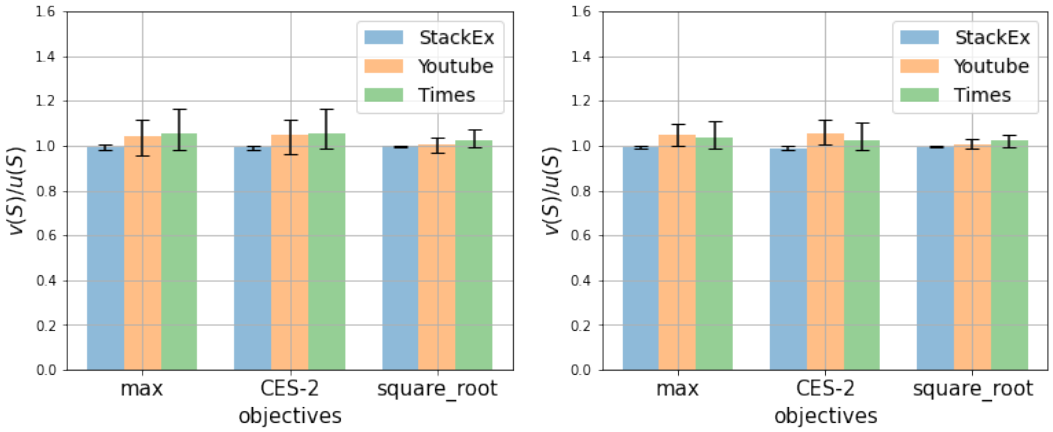

Figure 5.5: The approximation ratio of our method for various objective functions on three datasets: (left) $k = 5$ and (right) $k = 10$.

Results As for the synthetic data case, we test three types of set utility functions, the max, the CES utility function of degree 2 and the square root function of sum on the three datasets. For each item, we compute the empirical distribution of performance of this item from the given data. We then generate $N = 100$ training samples of each item's performance to estimate the set utility functions. We set $\epsilon = 0.1$ and set size k to 5 and 10.

The performance ratios for the three different objectives on three datasets are presented in Figure 5.5. We observe that our sketch provides a good approximation in most cases.

5.6 Conclusion

In this chapter, we looked at the problem of finding good sketch (representation) of item distributions for approximation of stochastic set utility function, defined as the expecta-

tion of a valuation function of independent random item values. We proposed an efficient discretization algorithm base on exponential binning strategy. The algorithm yields discretized distributions for each item with $O(k \log k)$ support size. We have shown that for a wide class of monotone subadditive or submodular valuation functions, our algorithm provides a constant-factor approximation for any value query for a set of items of size less than or equal to k.

Our work provides first positive results on function approximation for a class functions accommodating a wide-range of valuation functions studied in existing literature. The results are also of interest for application to best set selection and welfare maximization problems. It may be of interest to think of other systematic discretionary strategies and explore the trade-off between approximation accuracy and complexity of the representation. We leave this for future work.

5.7 Proofs

Properties of success probability function

We consider the function $f(x) = 1 - \prod_{i=1}^{n}(1 - x_i)$ on $[0,1]^n$. This function is clearly submodular as it it is twice-differentiable and $\partial^2 f(x)/\partial x_i \partial x_j$ is equal to $-\prod_{l \in [n]\setminus\{i,j\}}(1 - x_l) \leq 0$ when $i \neq j$, and is equal to 0 when $i = j$.

We will show some properties of f by induction over the sequence of functions $f_1, \ldots, f_n$, where $f_j(x) = 1 - \prod_{i=1}^{j}(1 - x_j)$, for $1 \leq j \leq n$. Note that, for $1 \leq j < n$,

$$f_{j+1}(x) = x_{j+1} + f_j(x) - x_{j+1} f_j(x).$$

We show that f is subadditive by induction as follows. Let $x, y \in [0,1]^n$ be such that $x + y \in [0,1]^n$. For the base case $j = 1$, function $f_1(x) = x_1$ is clearly subadditive. For the induction step, assume that f_j is subadditive, for an arbitrary $1 \leq j < n$. Then, we have

$$\begin{aligned}
f_{j+1}(x+y) &= x_{j+1} + y_{j+1} + f_j(x+y) - (x_{j+1} + y_{j+1}) f_j(x+y) \\
&= x_{j+1} + y_{j+1} + (1 - x_{j+1} - y_{j+1}) f_j(x+y) \\
&\leq x_{j+1} + y_{j+1} + (1 - x_{j+1} - y_{j+1})(f_j(x) + f_j(y)) \\
&= f_{j+1}(x) + f_{j+1}(y) - x_{j+1} f_j(y) - y_{j+1} f_j(x) \\
&\leq f_{j+1}(x) + f_{j+1}(y)
\end{aligned}$$

which shows that f_{j+1} is subadditive.

We next show that f is weakly homogeneous over $[0,1]$ with tolerance $\eta = 1$ by induction as follows. For the base case $j = 1$, $f_1(x) = x_1$, so clearly it holds $f_1(\theta x) \geq \theta f_1(x)$. For the

induction step, assume that $f_j(\theta x) \geq \theta f_j(x)$, for an arbitrary $1 \leq j < n$. Then, we have

$$\begin{aligned} f_{j+1}(\theta x) &= \theta x_{j+1} + f_j(\theta x) - \theta x_{j+1} f_j(\theta x) \\ &= \theta x_{j+1} + (1 - \theta x_{j+1}) f_j(\theta x) \\ &\geq \theta x_{j+1} + (1 - \theta x_{j+1}) \theta f_j(x) \\ &= \theta x_{j+1} + \theta f_j(x) - \theta^2 x_{j+1} f_j(x) \\ &\geq \theta x_{j+1} + \theta f_j(x) - \theta x_{j+1} f_j(x) \\ &= \theta f_{j+1}(x) \end{aligned}$$

which shows that f_{j+1} is weakly homogeneous over $[0,1]$ with tolerance $\eta = 1$.

We next show that f is weakly homogeneous over $[0,1]$ with degree $d \leq 1/2$. To show this, let us consider the case when $n = 2$. We then have $f(x) = x_1 + x_2 - x_1 x_2$. The condition $f(\theta x) \leq \theta^d f(x)$ can be written as follows

$$(1 - \theta^{2-d}) x_1 x_2 \leq (1 - \theta^{1-d})(x_1 + x_2)$$

for all $x_1, x_2 \in [0,1]$. Clearly the last inequality holds when either $x_1 = 0$ or $x_2 = 0$. Hence, the condition is equivalent to

$$1 - \theta^{2-d} \leq (1 - \theta^{1-d}) \left(\frac{1}{x_1} + \frac{1}{x_2} \right)$$

for all $x_1, x_2 \in (0,1]$. This is clearly equivalent to $1 - \theta^{2-d} \leq 2(1 - \theta^{1-d})$ which can be written as

$$\theta^{1-d}(2 - \theta) \leq 1. \tag{5.4}$$

The left-hand side is increasing in d and achieves the maximum value at $\theta^* = 1/(2(1 - d))$. Hence, equality in (5.4) is achieved at θ^* when $d = 1/2$.

Proof of Lemma 5.3.3

Upper end We compare the set function $w(S) = \mathbb{E}\left[f\left((X_i \mathbb{1}_{\{X_i \leq \tau_i\}}, i \in S)\right)\right]$ and the set function $u(S) = \mathbb{E}[f((X_i, i \in S))]$.

We first prove the upper bound. Let T be the subset of S containing those X_i exceeding the threshold τ_i, i.e. $T = \{i \in S \mid X_i > \tau_i\}$.

By Lemma 2.2.3, under condition that f is monotone and either subadditive or submodular, u is a monotone subadditive function. Hence, we have

$$\mathbb{E}[f((X_i, i \in S))] \leq \mathbb{E}[f((X_i, i \in T))] + \mathbb{E}[f((X_i, i \in S \setminus T))].$$

Now, note

$$\begin{aligned}\mathbb{E}[f((X_i, i \in S))] &\leq 2\max\left\{\mathbb{E}[f((X_i, i \in T))], \mathbb{E}[f((X_i, i \in S \setminus T))]\right\} \\ &\leq 2\max\left\{\mathbb{E}[f((X_i, i \in T))], w(S)\right\}.\end{aligned}$$

Again, by subadditivity of the set function u, we have

$$\mathbb{E}[f((X_i, i \in T))] \leq \mathbb{E}\left[\sum_{i \in T} f(X_i)\right] \leq \epsilon \sum_{i \in S} H_i \tag{5.5}$$

where recall $H_i = \mathbb{E}[f(X_i) \mid X_i > \tau_i]$.

Thus, it follows

$$\mathbb{E}[f((X_i, i \in S))] \leq 2\max\left\{\epsilon \sum_{i \in S} H_i, w(S)\right\}$$

which proves the upper bound in the lemma.

We next prove the lower bound. Since f is a monotone function,

$$\mathbb{E}[f((X_i, i \in S))] \geq \max\left\{\mathbb{E}[f((X_i, i \in T))], \mathbb{E}[f((X_i, i \in S \setminus T))]\right\}.$$

Thus, we have

$$u(S) \geq \max\left\{\mathbb{E}\left[f\left((X_i, i \in T)\right)\right], w(S)\right\}. \tag{5.6}$$

Now, note

$$\begin{aligned}
\mathbb{E}[f((X_i, i \in T))] &= \sum_{U \subseteq S} \Pr(T = U)\, \mathbb{E}[f((X_i, i \in T)) \mid T = U] \\
&\geq \sum_{U \subseteq S : |U| = 1} \Pr(T = U)\, \mathbb{E}[f((X_i, i \in T)) \mid T = U] \\
&= \sum_{i \in S} \Pr(X_i > \tau_i) \Pr(X_j \leq \tau_j, \forall j \neq i)\, \mathbb{E}[f(X_i) \mid X_i > \tau_i] \\
&\geq (\epsilon - \Delta)(1 - \epsilon)^{k-1} \sum_{i \in S} H_i
\end{aligned}$$

where we used the facts $\mathbb{P}(X_j > \tau_j) \geq \epsilon - \Delta$ and $\mathbb{P}(X_j \leq \tau_j) \geq 1 - \epsilon$ for all $j \in \Omega$ that follow from (5.3), and assumption that set S is such that $|S| \leq k$.

Combining with (5.6), we have

$$u(S) \geq \max\left\{(\epsilon - \Delta)(1 - \epsilon)^{k-1} \sum_{i \in S} H_i, w(S)\right\}$$

from which the lower bound in the lemma follows.

Proof of Lemma 5.3.4

Note that

$$\begin{aligned}
u(S) &\leq 2 \max\left\{\epsilon \sum_{i \in S} H_i, w(S)\right\} \\
&= \frac{2}{(1 - \epsilon)^{k-1}} (1 - \epsilon)^{k-1} \max\left\{\epsilon \sum_{i \in S} H_i, w(S)\right\} \\
&\leq \frac{2}{(1 - \epsilon)^{k-1}(1 - \Delta/\epsilon)} v_1(S)
\end{aligned}$$

where the first inequality is by the upper bound in Lemma 5.3.3 and the last inequality is by the lower bound in Lemma 5.3.3.

This shows the upper bound in the statement of the lemma. The lower bound in the statement of the lemma follows by similar arguments as below,

$$\begin{aligned} u(S) &\geq (1-\epsilon)^{k-1}(1-\Delta/\epsilon)\max\left\{\epsilon\sum_{i\in S}H_i, w(S)\right\} \\ &= \frac{1}{2}(1-\epsilon)^{k-1}(1-\Delta/\epsilon)\cdot 2\max\left\{\epsilon\sum_{i\in S}H_i, w(S)\right\} \\ &\geq \frac{1}{2}(1-\epsilon)^{k-1}(1-\Delta/\epsilon)v_1(S) \end{aligned}$$

where the first inequality is by the lower bound and the last inequality is by the upper bound in Lemma 5.3.3.

Proof of Lemma 5.3.5

Lower end We compare the set function $v_2(S) = \mathbb{E}[f((\tilde{X}_i, i \in S))]$ and the set function $v_1(S) = \mathbb{E}[f((\hat{X}_i, i \in S))]$.

Let $\tilde{X}_i = \hat{X}_i \mathbb{1}_{\{\hat{X}_i > a\tau_i\}}$, for some $a \in [0,1]$. For any monotone submodular function f and any monotone subadditive function f, it holds, for any $a \in [0,1]$,

$$\begin{aligned} v_1(S) &= \mathbb{E}[f((\hat{X}_i, i \in S))] \\ &= \mathbb{E}[f((\hat{X}_i \mathbb{1}_{\{\hat{X}_i \leq a\tau_i\}} + \hat{X}_i \mathbb{1}_{\{\hat{X}_i > a\tau_i\}}, i \in S))] \\ &\leq \mathbb{E}[f((\tilde{X}_i \mathbb{1}_{\{\hat{X}_i \leq a\tau_i\}}, i \in S))] + \mathbb{E}[f((\tilde{X}_i \mathbb{1}_{\{\hat{X}_i > a\tau_i\}}, i \in S))] \\ &\leq f((a\tau_i, i \in S)) + \mathbb{E}[f((\tilde{X}_i, i \in S))] \\ &= f((a\tau_i, i \in S)) + v_2(S). \end{aligned}$$

Combining with the condition that f is weakly homogeneous of degree d over $[0,1]$, we have

$$v_1(S) \leq a^d f((\tau_i, i \in S)) + v_2(S).$$

Now, note that for any monotone, subadditive or submodular function f,

$$\mathbb{E}[f((\tilde{X}_i, i \in S))] \geq \frac{\epsilon - \Delta}{k} \mathbb{E}[f((\tau_i, i \in S))].$$

This can be shown as follows. Let $j \in \arg\max_{i \in S} \tau_i$. Then, we have

$$\begin{aligned}
\mathbb{E}[f((\tilde{X}_i, i \in S))] &\geq \mathbb{P}(\tilde{X}_j > \tau_j) f(\tau_j e_j) \\
&\geq (\epsilon - \Delta) f(\tau_j e_j) \\
&\geq \frac{\epsilon - \Delta}{k} f\left(\sum_{i \in S} \tau_j e_i\right) \\
&\geq \frac{\epsilon - \Delta}{k} f\left(\sum_{i \in S} \tau_i e_i\right) \\
&= \frac{\epsilon - \Delta}{k} f((\tau_i, i \in S))
\end{aligned}$$

where we used the fact $\mathbb{P}(\tilde{X}_j > \tau_j) = \mathbb{P}(X_j > \tau_j) \geq \epsilon - \Delta$, with the last inequality following from (5.3).

Putting the pieces together, we have

$$v_2(S) \geq \frac{1}{1 + a^d k/(\epsilon - \Delta)} v_1(S).$$

Proof of Lemma 5.3.6

Millde part Recall that for each $i \in \Omega$, $Y_i = q(\tilde{X}_i; \tau_i, \epsilon)$. We compare the set functions $v_2(S) = \mathbb{E}[f((\tilde{X}_i, i \in S))]$ and $v(S) = \mathbb{E}[f((Y_i, i \in S))]$.

Note that q is such that, for every $\tau > 0$,

$$q(x; \tau, \epsilon, a) \geq (1 - \epsilon)x, \text{ for all } x \in [a\tau, \tau].$$

This combined with monotonicity of f immediately yields

$$v(S) \geq \mathbb{E}[f(((1-\epsilon)\tilde{X}_i, i \in S))].$$

Combining with the condition that f is weakly homogeneous with tolerance η yields $v(S) \geq ((1-\epsilon)/\eta)v_2(S)$.

Proof of Corollary 5.3.2

Note that $1 - x \geq e^{-\theta x}$, for all $x \leq 1 - 1/\theta$ and $\theta \geq 1$. Hence, for $\epsilon = c/k$, where c is some positive constant in $(0,1)$, we have $(1-\epsilon)^k \geq e^{-\theta c}$, for $c/k \leq 1 - 1/\theta$ and $\theta \geq 1$. By taking $\theta = 1/(1-c)$, we have $(1-\epsilon)^k \geq e^{-c/(1-c)}$. Using this, we can establish the statement of the corollary.

Proof of Theorem 5.3.7

The proof for the upper end remains the same as in the proof of Theorem 5.3.1. We thus only need to address the lower end and middle part of the proof.

Lower end Let f^* be a concave extension of f. Since $f^*(x) = f(x)$ for all $\mathbb{R}^n_+$ and we consider item value distributions with positive supports, we can consider $v_1(S) = \mathbb{E}[f^*((\hat{X}_i, i \in S))]$ and $v_2(S) = \mathbb{E}[f^*((\tilde{X}_i, i \in S))]$.

Recall that it holds

$$\mathbb{E}[f^*((\tilde{X}_i, i \in S))] \geq \frac{\epsilon - \Delta}{k} \mathbb{E}[f^*((\tau_i, i \in S))]. \tag{5.7}$$

Let $Z_i = \hat{X}_i - a\tau_i$ and note that $\tilde{X}_i = \hat{X}_i \mathbb{1}_{\{\hat{X}_i > a\tau_i\}} \geq Z_i$. Note that we can write,

$$Z_i = (1-a)\,\hat{X}_i + a\left(\hat{X}_i - \tau_i\right).$$

Since f^* is monotone, concave and subadditive, we have the following inequalities.

$$\begin{aligned} v_2(S) &\geq \mathbb{E}[f^*((Z_i, i \in S))] \\ &\geq (1-a)\,\mathbb{E}[f^*((\hat{X}_i, i \in S))] + a\,\mathbb{E}[f^*((\hat{X}_i - \tau_i, i \in S))] \\ &\geq (1-a)\,\mathbb{E}[f^*((\hat{X}_i, i \in S))] + a\,\mathbb{E}[f^*((\hat{X}_i, i \in S))] - a \cdot f^*((\tau_i, i \in S)) \\ &\geq v_1(S) - (ak/(\epsilon - \Delta))v_2(S) \end{aligned}$$

where the first inequality is by monotonicity, the second inequality is by concavity, the third inequality is by subadditivity, and the last inequality is by the definition of $v_1(S)$ and the inequality in (5.7).

Middle part This follows by the same arguments as in the proof of Theorem 5.3.1 and making use of the fact that any concave function is weakly homogeneous with tolerance 1.

Proof of Theorem 5.3.9

The proof of the upper end remains the same as in the proof of Theorem 5.3.1. In what follows, we show the proof for the lower end part and the middle part of the proof.

Lower end This part is shown by the following lemma.

Lemma 5.7.1. *Assume that f is a monotone function that is either subadditive or submodular and is coordinate-wise weakly homogeneous of degree d over $[0,1]$. Then, we have*

$$v_2(S) \geq \frac{1}{1 + a^d k/(\epsilon - \Delta)} v_1(S).$$

Proof. The proof can be established by similar steps as in the proof of Lemma 5.3.5 and making use of the following simple fact: under coordinate-wise weakly homogeneous condition, $f((a\tau_i, i \in S)) \leq a^{dk} f((\tau_i, i \in S)) \leq a^d f((\tau_i, i \in S))$. □

Middle part This follows by the same arguments as in the proof of Theorem 5.3.1 combined with repeated application of the weak homogeneity property that holds coordinate-wise which yields

$$\mathbb{E}[f(((1-\epsilon)\tilde{X}_i, i \in S))] \geq (1/\eta)^k (1-\epsilon)^k \, \mathbb{E}[f((\tilde{X}_i, i \in S))].$$

5.8 Supplementary numerical results

5.8.1 YouTube dataset: other performance metrics

In this section, we illustrate and compare the results using different measures for the performance value of YouTube video uploads. Specifically, if a video j uploaded τ days ago received $n(j)$ views, $l(j)$ likes and $d(j)$ dislikes, we calculate six measures for its performance as follows.

- View counts per day (view ratio): $n(j)/\tau$
- Log of the view counts: $\log(n(j)+1)$
- Standard like ratio: $l(j)/(l(j)+d(j))$
- Video Power Index (VPI): view ratio $\times$ standard like ratio
- Bayesian like ratio: $(l(j)+c_1)/(l(j)+d(j)+c_2)$
  - Conservative case where $c_1 = 0.01n(j)$ and $c_2 = 0.1n(j)$
  - Balanced case where $c_1 = 0.05n(j)$ and $c_2 = 0.1n(j)$

The Bayesian like ratio has a similar interpretation as for the StackExchange dataset. The difference is that in this setting the ratio factors in the effect of view counts. Note that c_1 can be seen as a threshold value needed for the number of likes to have an effect on the performance value.

We computed the performance values for all videos submitted by qualified YouTubers with more than 50 uploads.

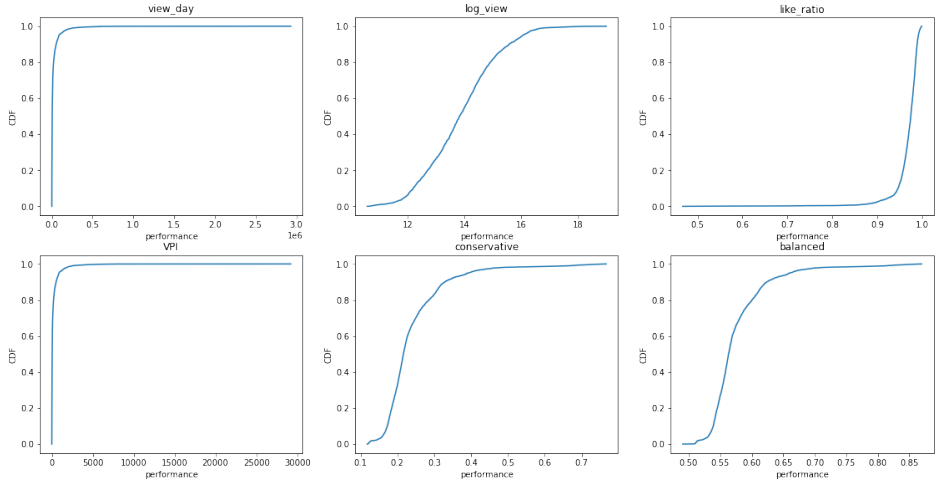

Figure 5.6: Empirical CDFs of performance values for the Youtube dataset, for six different performance metrics.

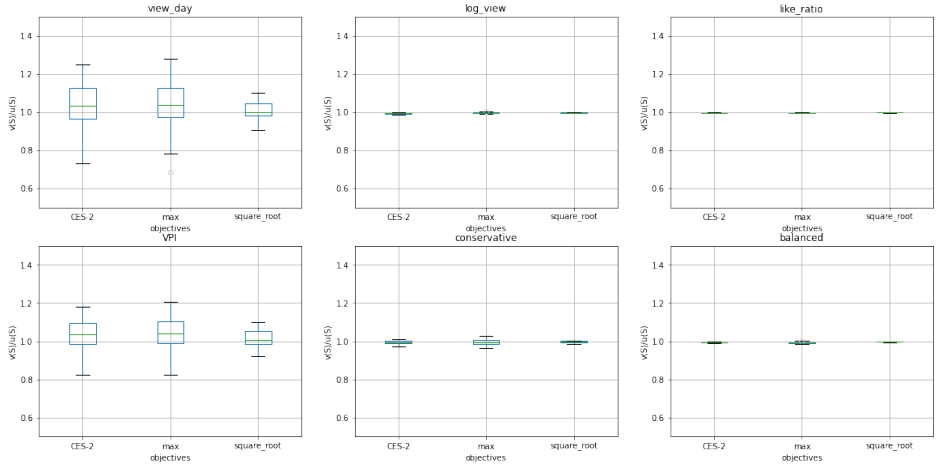

Figure 5.7: The approximation ratio for different valuation functions for the Youtube dataset, for six different performance metrics.

Figure 5.6 shows how the values are distributed for all six measures. Note that the distributions under the measures view counts per day, VPI and conservative Bayesian like ratio are more heavy-tailed compared to others.

Figure 5.7 shows the results for three objective functions (max, CES of degree 2 and square root function of sum) under the six measures. We can observe that all the approximation ratios for the measures log of the view counts, standard like ratio, and balanced Bayesian like ratio are more concentrated around 1 compared to the results using view counts per

day and VPI. This result is not surprising, as we noted that the value distributions under these two measures are centered and light-tailed.

5.8.2 StackExchange dataset: other (c_1, c_2) parameter settings

We tested the following values of (c_1, c_2): $(2, 8)$, $(8, 32)$ and $(10, 10)$. As explained previously, the first and second cases correspond to a conservative choice, and the last one is a balanced ratio.

Figure 5.8 shows the empirical CDF for performance values, for the three settings of (c_1, c_2) parameters. We can see that as the values of c_1 and c_2 increase, the values are more concentrated around c_1/c_2.

Figure 5.9 shows the corresponding results under the different value pairs for c_1 and c_2. From the plots, we observe that all the approximation ratios are highly concentrated around 1. Thus we can conclude that the choices of c_1 and c_2 have no particular effects on the approximation ratio.

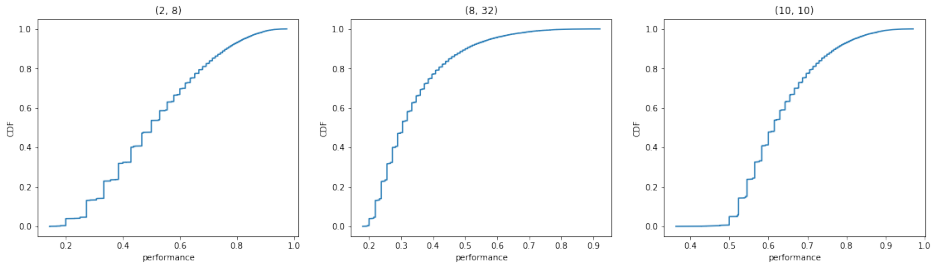

Figure 5.8: Empirical CDFs of performance values for the StackExchange dataset, for three different parameter settings.

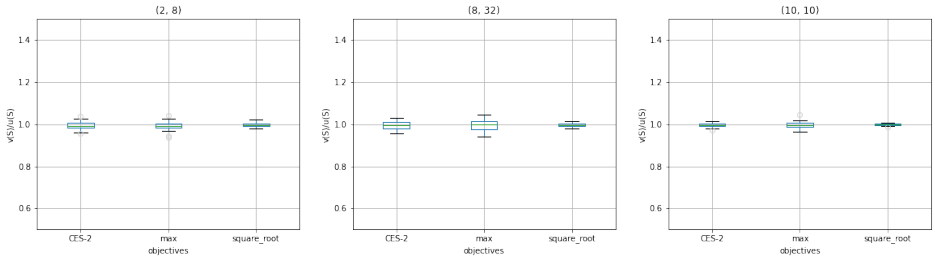

Figure 5.9: The approximation ratio for different valuation functions for the StackExchange dataset, for different parameter settings.

Chapter 6

The k-max problem with value-index feedback

6.1 Overview

In many real-world examples, users interact with an online system and data comes in streams, which motivates the need for sequential experimentation and online learning. In this chapter, we consider a class of online combinatorial optimization problem where an agent chooses samples sequentially. We assume a ground set of n items that are binary-valued. In each round, the agent chooses a set of items of size k from the ground set and receives the maximum value of the set and the index of the item taking the maximum value as feedback. We call the problem as k-max problem with value-index feedback. The problem is new as it assumes a special feedback structure. As discussed in the introduction chapter, this type of feedback arises naturally in real-world applications such as online advertising, where the agent observes the most popular item which receives the click and its value. The binary-valued assumption is justified in this example as the item either receives a click and reveals its value or does not receive a click. We will take this simplified assumption for analysis, while we note that by allowing general distributions we can model more complicated real-world applications.

Our goal is to maximize the expected cumulative reward for a learning agent over the time horizon. The problem is challenging mainly for two reasons. Firstly, the reward function is the max function, which is nonlinear and depends not only on the expected value of the constituent base arms. The uncertainty of binary-valued items makes the problem more challenging under the max reward. As we will show in the numerical section, high-risk high-reward items may outperform stable-value items in this case. The second challenge is due to the limited feedback. The agent only observes the maximum value and the winner index. These all make it hard to estimate the distributions of the individual base arms.

6.1.1 Related work

Stochastic multi-armed bandit (MAB), first studied by Robbins (1952), is a classical online learning framework motivated by such applications. It is typically formulated as a learning problem between an agent and a system of n arms with unknown distributions $P_1, \ldots, P_n$. The agent chooses some arm and receives a random reward from the environment following its distribution. The goal is to collect cumulative rewards as much as possible over the time horizon. The performance of a MAB algorithm is measured by its cumulative regret, which is defined as the difference in cumulative reward compared to always playing the best arm.

Our problem can be put into the general framework of combinatorial multi-armed bandits (CMAB)(Cesa-Bianchi and Lugosi (2012); Chen et al. (2013, 2016)), an extension to the MAB problem. A decision set of actions is given where each action is a subset of arms, not just one arm. In each round, the agent chooses an action from the decision set and receives feedback for the chosen subset of arms. The k-max problem is a special case of CMAB where the decision set includes all subsets of cardinality k.

There are two typical settings for the CMAB problem. In the semi-bandit setting, the outcomes of the selected arms are observed as feedback. In the full bandit setting, only

aggregate reward of selected arms is observed as the feedback, which is modeled using reward functions.

Most of the existing works of CMAB focused on semi-bandit setting (Chen et al. (2013); Kveton et al. (2015b)). The full-bandit CMAB is harder than the semi-bandit problem, due to lack of information on individual arms. However, in real life applications, it is often expensive or even infeasible to obtain per-item information and we may only have observations at the set-level. Consider the online advertising example given at the introduction section. Th use platform presents a list of items to a user, who selects and gives rating to selected item of interest. In most cases, it is impossible to collect information on those items not selected by the user. Therefore, it is important to consider the full-bandit setting. In most works on full-bandit CMAB, restrictions are placed on the reward function. Auer et al. (2002) first studied the problem under linear reward and provides a linear UCB algorithm. Dani et al. (2008) fully analyzed the linear UCB algorithm and gave a nearly optimal regret bound. However, the method is computationally intractable for combinatorial decision set. Rejwan and Mansour (2020) considered a special full-bandit setting where the reward is defined as the sum of individual arms. They proposed an algorithm based on the successive accepts and rejects (SAR) algorithm that iteratively estimates expected rewards of arms within increasing level of accuracy. The estimates for individual arms is obtained through solving a linear system of equations. Thus this method cannot be generalized to full bandit setting with non-linear reward functions.

Only a few algorithms are proposed for full bandit CMAB problem with non-linear reward. Katariya et al. (2017) considered a minimum function for $\{0,1\}$-valued base arms and proposed an elimination algorithm to find the best set without explicit estimation of individual arm's expected reward. However, their analysis largely depend on the binary nature of base arms. Gopalan et al. (2014) studied the full-bandit CMAB with general reward using Thompson sampling method. However, it is computationally hard to compute the posteriors in the algorithm and the regret bound involves a large exponential constant. A recent work by Agarwal et al. (2021) proposes a merge and sort algorithm for the CMAB problem with non-linear reward without any extra feedback. Instead of

estimating individual arms, the authors use the theory of stochastic dominance to obtain ordering of individual arms. However, this requires a strong assumption that first-order stochastic dominance exists between any two arms. Clearly, for binary-valued random variables, this assumption does not always hold.

We realized the difficulty in solving the full-bandit CMAB problem with non-linear reward. Our work can seen as taking a middle ground between the semi-bandit and full-bandit settings. Given a selection of set $S \subseteq [n]$, our feedback is (I, X_I) where I is the index of the item with maximum value in set S. As discussed above, our work is related with semi- and full-bandit CMAB problems. A closely related work is combinatorial cascading bandits (Kveton et al. (2015a,c)). In this problem, base arms are Bernoulli random variables. An agent will choose an ordered subsequence from the set of base arms and reveal the outcome of the base arms one by one until a stopping criteria is met. In the disjunctive form, the agent stops when the first one is observed. In the conjunctive case, the agent stops when the first zero is observed. Chen et al. (2016) generalizes the problem to the framework of combinatorial semi-bandits with probabilistically triggered arms (CMAB-T). We note that the main difference is that this line of work assumes more information than our problem and is inherently semi-bandit. By revealing the outcome of base arms one by one, the agent is able to observe individual rewards for all arms selected before the one meeting the criteria. Another difference in our work is that we assume that the base arms are binary valued. This would cause the reward function not only depend on the expected value, but the whole distributions of the constituent base arms, making the learning problem more challenging.

Another line of works related to our study is the dueling bandit problem (Ailon et al. (2014)) where the agent plays two arms at each time and observes the outcome of the duel. The goal is to find the best item in the sense of Condorcet winner under relative feedback of the dueling outcomes. Sui et al. (2017) extends the setting to multiple dueling bandits problem by simultaneously playing k arms instead of two arms. Compared to this line of work, we assume additional absolute value feedback X_I. We note that our goal is different, as we would like to select a set of items with maximum performance measured

by some non-linear utility function. A key assumption made in the dueling bandit problem is that approximate linearity holds for the stochastic preference relationship, but this assumption does not hold in our setting.

6.1.2 Summary of contributions

Our results can be summarized in the following points.

- We consider the k-max problem with expected max reward and winner index feedback. This is a new problem setting that stands at the middle ground of semi-bandit and full-bandit. Compared to the full-bandit setting, we assume additional information of winner index, which is a natural assumption to be made in real-world applications. On the other hand, we do not assume per-item value feedback, which differentiates with the semi-bandit problem. Our work is one step towards solving the full-bandit CMAB problem with non-linear reward under mild assumptions.
- We rephrase the problem in an interpretable way by introducing two sets of base arms. In the simpler case when the ordering of values is known within each action, the problem boils down into two separate standard CMAB-T problems. In the general case, the problem differs from CMAB-T as the triggered subset of the base arm set given an action depends on whether the item values are observed or not. We tackle with this difficulty by introducing the concept of item equivalence, such that we can restore the CMAB-T framework by using the replacement items.
- We present a CUCB algorithm to solve the simpler case of the k-max problem. The CUCB algorithm achieves comparable regret bound as standard CMAB problems. However, for the general case, it yields a sub-optimal regret bound that contains an undesirable factor of $1/p^*$ where p^* is the minimum value that an arm takes its positive value. To remove the extra factor, we propose a modified algorithm based on the CUCB algorithm. By using the concept of item equivalence, we show that the modified algorithm achieves comparable regret as the simpler case.

6.2 Problem formulation

6.2.1 Model specification

We consider an online learning problem between an agent and a system of n items or arms, denoted as $E = [n] = \{1, 2, \ldots, n\}$. The arms produce stochastic outcomes $\boldsymbol{X} = (X_1, \ldots, X_n)$. We assume that the random variables are binary-valued, i.e., $X_1, \ldots, X_n$ taking strictly positive values $v_1, \ldots, v_n$ with probability $p_1, \ldots, p_n$ respectively. The values $v_1, \ldots, v_n$ and the probabilities $p_1, \ldots, p_n$ are unknown.

We define $\mathcal{F} = \{S | S \in 2^E, |S| = k\}$ as the set of super arms of cardinality k. At each time step t, the agent takes an action to play a super arm $S_t \in \mathcal{F}$. The agent observes the maximum value of the selected arms and the index of the item taking the maximum value. The goal is to select a set of random variables with maximum performance according to the expected maximum objective.

We will start from a simpler case where we assume that ordering of the values v_i is known within each action. Then we move to the general case where both values and ordering of v_i are unknown.

We can rephrase our problem in a more intuitive way by adopting the notation of triggered arms (Wang and Chen (2017)). We introduce two set of base arms decomposed from the random variables $X_1, \ldots, X_n$. The first set of base arms $\mathcal{Z}$ consists of n independent Bernoulli random variables $Z_1, \ldots, Z_n$ with mean values $p_1, \ldots, p_n$. The second set of base arms $\mathcal{V} = \{V_1, \ldots, V_n\}$ are deterministic with mean values $v_1, \ldots, v_n$. We also define an extended set of base arms $\mathcal{B}$ containing both sets of base arms. Note that we have $X_i = V_i \cdot Z_i$. Each time an action S_t is played, we obtained information on some of the base arms Z_i and V_i in $\mathcal{B}$. We call these arms as being triggered, and we observe their values as feedbacks.

Recall that in the simpler case we know the ordering of arms in decreasing value of v_i. Therefore, when action S_t is played and the maximum value of the set is v_i, we immedi-

ately conclude that the base arm V_i takes value v_i, Z_i takes non-zero value and Z_j such that its corresponding arm j is ordered before i takes value zero. For the set $\mathcal{Z}$, we say the base arms $Z_1, \ldots, Z_i$ are triggered arms, and we observe their values as feedbacks. For the set $\mathcal{V}$, only one base arm V_i is triggered and we observe its deterministic value. We note that for the simpler case, our problem can be interpreted as a conjunctive cascading bandit with binary-valued arms. The ordering of arms within each action enables us to observe values of all arms ordered before the winner, which makes the problem easier to solve.

The general case is conceptually harder for the set of base arms $\mathcal{Z}$. Since we don't know the ordering of values within an action, we have no information about Z_i if v_i has not yet been observed. We can only conclude that Z_j takes value zero if v_j has been observed and $v_j > v_i$. The triggered arm is the same for the set of base arms $V_1, \ldots, V_n$.

Note that p_i and v_i are expectations of the base arms Z_i and V_i respectively. Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ and $\boldsymbol{v} = (v_1, \ldots, v_n)$ be the expectation vectors of the base arms. When an action is played, the agent obtains a non-negative reward of the maximum value, which is fully determined by the triggered arms. We denote the expected reward as $r_S(\boldsymbol{p}, \boldsymbol{v}) = \mathbb{E}[\max(X_i, i \in S)]$, which is a function of action S and expectation vectors p and v. Importantly, we note that if $S = [k]$ and arms are ordered in decreasing order of their values, then we can write the expected reward explicitly as

$$r_S(\boldsymbol{p}, \boldsymbol{v}) = p_1 v_1 + (1 - p_1) p_2 v_2 + \ldots + (1 - p_1) \ldots (1 - p_{k-1}) p_k v_k \tag{6.1}$$

The performance of a learning algorithm is measured by its cumulative regret, which is defined as the difference in expected cumulative reward by playing the best action and playing actions suggested by the algorithm. Denote $\text{opt}_{\boldsymbol{p},\boldsymbol{v}} = \sup_S r_S(\boldsymbol{p}, \boldsymbol{v})$. The expected regret can be written as

$$R(T) = T \cdot \text{opt}_{\boldsymbol{p},\boldsymbol{v}} - \mathbb{E}\left[\sum_{t=1}^{T} r_{S_t}(\boldsymbol{p}, \boldsymbol{v})\right]$$

We may assume an (α, β) approximation oracle, which takes $(\boldsymbol{p}, \boldsymbol{v})$ as input and outputs an action S such that

$$\Pr(r_{\boldsymbol{p},\boldsymbol{v}}(S) \geq \alpha \cdot \mathrm{opt}_{\boldsymbol{p},\boldsymbol{v}}) \geq \beta$$

where α is the approximation ratio and β is the success probability. Under the approximation oracle, the (α, β) regret can be written as

$$R(T) = T \cdot \alpha \cdot \beta \cdot \mathrm{opt}_{\boldsymbol{p},\boldsymbol{v}} - \mathbb{E}\left[\sum_{t=1}^{T} r_{S_t}(\boldsymbol{p}, \boldsymbol{v})\right]$$

Note the major difference with the classical combinatorial bandits is that we need to estimate the expectation vectors of two sets of base arms.

6.2.2 Properties of the reward functions

There are two key properties of the regret function that will be needed to guarantee the theoretical regret upper bound.

Monotonicity The first property is monotonicity.

Lemma 6.2.1. *$r_S(\boldsymbol{p}, \boldsymbol{v})$ is monotonic increasing in every p_i and v_i.*

Recall that for a given set of random variables, we can explicitly write $r_S(\boldsymbol{p}, \boldsymbol{v})$ as in equation (6.1). It is clear from the expression that $r_S(\boldsymbol{p}, \boldsymbol{v})$ is monotonic increasing in v_i. We can prove that it is monotonic increasing in p_i by taking first derivative with respect to p_i and showing that the differential is greater than zero.

Smoothness The second condition is called relative triggering probability modulated (TPM) smoothness. This notion was originally defined in an appendix of the work by Wang and Chen (2017) for a different purpose. Here we redefine the notion in a general framework.

Definition 6.2.2. *Denote the triggering probability of a base arm i in a set of base arms $\mathcal{B}$ with expectation $\boldsymbol{\mu}$ for action S as p_i^S. We say the problem satisfies 1-norm relative bounded smoothness with respect to the base arm set $\mathcal{B}$ if, for any two distributions with different expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$, and any action S we have*

$$|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu'})| \leq \sum_{i \in S} p_i^S b_i |\mu_i - \mu_i'|$$

where b_i is some per-arm weight coefficient.

Note that when $r_S(\boldsymbol{\mu})$ is monotonic increasing in $\boldsymbol{\mu}$ and $\boldsymbol{\mu} > \boldsymbol{\mu'}$, we can remove the absolute sign.

Note that we add the triggering probability p_i^S and a weight coefficient b_i to *modulate* the standard 1-norm condition. The intuition is that we underweight the importance of items with small triggering probability or weight in expected reward. Even if for some item i we cannot estimate its expected value accurately, we lose very little in the expected reward. This will be a very important concept in the regret analysis that follows.

Let the triggering probability of Z_i by action S be q_i^S and the triggering probability of V_i by action S be $\tilde{q}_i^S$. Note that if $S = [k]$ and arms are ordered in decreasing order of their values, then the triggering probability for Z_i by action S is

$$q_i^S = (1 - p_1)(1 - p_2) \dots (1 - p_{i-1}). \tag{6.2}$$

And the triggering probability for V_i by action S is

$$\tilde{q}_i^S = (1 - p_1)(1 - p_2) \dots (1 - p_{i-1}) p_i. \tag{6.3}$$

Note that $\tilde{q}_i^S = q_i^S \cdot p_i$.

Now we claim that the following property holds for our problem.

Lemma 6.2.3. *If $\boldsymbol{v} > \boldsymbol{v'}$ and $r_S(\boldsymbol{p}, \boldsymbol{v})$ is monotonic increasing in $\boldsymbol{p}$ and $\boldsymbol{v}$, the k-max problem with value-index feedback satisfies the 1-norm relative bounded smoothness condition with respect*

to the extended base arm set $\mathcal{B}$,

$$|r_S(\boldsymbol{p}, \boldsymbol{v}) - r_S(\boldsymbol{p'}, \boldsymbol{v'})| \leq 2\sum_i q_i^S v_i' |p_i - p_i'| + \sum_i \tilde{q}_i^S |v_i - v_i'|$$

We note that when we further have $\boldsymbol{p} > \boldsymbol{p'}$, then we can remove the factor of 2, i.e.,

$$r_S(\boldsymbol{p}, \boldsymbol{v}) - r_S(\boldsymbol{p'}, \boldsymbol{v'}) \leq \sum_i q_i^S v_i' (p_i - p_i') + \sum_i \tilde{q}_i^S (v_i - v_i') \tag{6.4}$$

The proof of lemma 6.2.3 uses a technique called bottom-up modification. We consider a sequence of vectors changing from $(\boldsymbol{p}, \boldsymbol{v})$ to $(\boldsymbol{p'}, \boldsymbol{v'})$ and add up the changes in expected rewards. The full proof is provided at the end of the chapter.

6.3 Algorithms and regret bounds

We first review the classical CMAB problem with triggered arms considered by Wang and Chen (2017). In this problem, the expected reward is a function of action S and expectation vector μ of base arms. It is assumed that in each round the value of triggered arms are observed by the agent. Standard CUCB algorithm is used to estimate the expectation vector μ directly from samples.

Regret bound for standard CMAB-T The following is a known result for standard CMAB problem with triggered arms.

Theorem 6.3.1. *For the CUCB algorithm that satisfies monotonicity and 1-norm TPM bounded smoothness with smoothness constant $b_i = B$ for all i, we have the following distribution-dependent bound,*

$$R(T) \leq \sum_{i \in E} \frac{576 v_i^2 k \ln T}{\Delta_{min}^i} + \sum_{i \in E} \left(\lceil \log_2 \frac{2Bk}{\Delta_{min}^i} \rceil + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{max} + 4Bn$$

where $\Delta_{min} > 0$ is a per-arm gap that will be defined later.

For our problem, we will start from the simpler case with an extra assumption that the ordering of v_i values are known. For this case, we will see that the k-max problem can be thought of as two separate CMAB-T problems. Then we move to the general case. Without the ordering assumption, standard CUCB algorithm will not provide us a satisfactory bound. We will propose a modified algorithm and take extra care to show that the regret bound is comparable to that of the standard CMAB problem.

Notations We define $T_{i,t}$ as the number of triggering times for Z_i and $\tilde{T}_{i,t}$ as the number of triggering times for V_i. We also define $N_{i,j,t}$ as the counter of times i in TP group $S_{i,j}$ is selected in the actions. We reset the counter to zero once V_i is triggered and v_i is observed.

For each action S, we define the gap $\Delta_S = \max(0, \alpha \cdot \mathrm{opt}_{\boldsymbol{p},\boldsymbol{v}} - r_S(\boldsymbol{p}, \boldsymbol{v}))$. We call an action *bad* if its gap is positive. For arms that are contained in at least one bad action, we define,

$$\Delta^i_{min} = \inf_{S: q^S_i, \tilde{q}^S_i > 0} \Delta_S, \quad \Delta^i_{max} = \sup_{S: q^S_i, \tilde{q}^S_i > 0} \Delta_S.$$

where $q^S_i, \tilde{q}^S_i > 0$ require that Z_i, V_i are triggered by action S with non-zero probabilities.

For other arms, we define $\Delta^i_{min} = \infty$ and $\Delta^i_{max} = 0$. Then we define $\Delta_{min} = \min_{i \in E} \Delta^i_{min}$ and $\Delta_{max} = \max_{i \in E} \Delta^i_{max}$.

We define the *event-filtered regret* as

$$R(T, \{\mathcal{E}_t\}) = T \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{p},\boldsymbol{v}} - \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(\mathcal{E}_t) r_{S_t}(\boldsymbol{p}, \boldsymbol{v})]$$

which means we count the regret in round t only if event $\mathcal{E}_t$ happens in round t.

We also define a series of good events (E1)-(E4) for the regret analysis. For space reason, they are listed in Section 6.6.

6.3.1 CUCB algorithm for the simpler case

We use a similar CUCB algorithm as standard CMAB problem to estimate p_i and v_i. Estimates of both sets of parameters are initialized to one at the beginning. Each time we observe v_j as the maximum value of the set, we update the corresponding estimates for v_j and the estimates for p_i, for items ordered before j. The algorithm maintains an upper confidence bound (UCB) for both parameters and feeds the UCB to the approximation oracle to obtain the next action. We will use the well-known greedy algorithm (Nemhauser et al. (1978a)) as the offline oracle. The Greedy k-max algorithm attains $(1-1/e)$ approximation guarantee in our case, as the expected maximum function is submodular.

Algorithm 6.3.1 CUCB algorithm for the simpler case with computation oracle

For each arm $i \in E$, $T_i \leftarrow 0$ ▷ Total number of triggering time for Z_i
For each arm $i \in E$, $\hat{p}_i \leftarrow 1$, $\hat{v}_i \leftarrow 1$ ▷ Empirical estimates of parameters
for $t = 1, 2, \ldots$ **do**
  For each arm $i \in E$, $\rho_i \leftarrow \sqrt{\frac{3 \log t}{2T_i}}$ ▷ Confidence radius of parameter p_i
  For each arm $i \in E$, $\bar{p}_i = \min\{\hat{p}_i + \rho_i, 1\}$, $\bar{v}_i \leftarrow \hat{v}_i$ ▷ UCB of parameters
  $S \leftarrow \text{Oracle}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}})$ ▷ Offline oracle decides the next action
  Play S and observe winner index j and value v_j as feedback.
  Update $\hat{v}_j$ for winner item j: $\hat{v}_j \leftarrow v_j$
  For each $i \in E$ such that $i \leq j$: $T_i \leftarrow T_i + 1$
  For each $i \in E$ such that $i < j$: $\hat{p}_i \leftarrow (1 - 1/T_i)\hat{p}_i$
  $\hat{p}_j \leftarrow (1 - 1/T_j)\hat{p}_j + 1/T_j$
end for

The regret bound for the CUCB algorithm is provided as follows.

Theorem 6.3.2. *If $\Delta_{min} > 0$, the CUCB algorithm defined above has the following distribution-dependent bound*

$$R(T) \leq \sum_{i \in E} \left(\frac{2304 v_i^2 k}{\Delta_{min}^i} + 6 \log_2 \frac{4k}{\Delta_{min}^i} \right) \ln T + \sum_{i \in E} \left(\lceil \log_2 \frac{4 v_i k}{\Delta_{min}^i} \rceil + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{max} + 4 v_i n$$

This regret bound achieves $O(\frac{nk}{\Delta}\log T)$ and is comparable to the standard CMAB problem. It is tight with respect to T up to logarithmic factor. To see how the algorithm can be boiled down into two CMAB-Ts, we consider the contribution of each action to regret, i.e, $\Delta_{S_t} = \max(0, \alpha \cdot \text{opt}_{\boldsymbol{p},\boldsymbol{v}} - r_S(\boldsymbol{p},\boldsymbol{v}))$. By the smoothness condition, we have

$$\Delta_{S_t} \leq r_{S_t}(\bar{\boldsymbol{p}}_{\boldsymbol{t}}, \bar{\boldsymbol{v}}_{\boldsymbol{t}}) - r_{S_t}(\boldsymbol{p},\boldsymbol{v}) \leq \sum_{i\in S_t} q_i^S v_i(\bar{p}_{i,t} - p_i) + \sum_{i\in S_t} \tilde{q}_i^S(\bar{v}_{i,t} - v_i) \tag{6.5}$$

Clearly, the first term corresponds to regrets from the set of base arms $\{Z_1, \ldots, Z_n\}$, and the second term corresponds to regrets from the set of base arms $\{V_1, \ldots, V_n\}$. We bound Δ_{S_t} by bounding the two summation terms individually.

For the first term, we can directly apply Theorem 6.3.1 to obtain the regret bound. Note that the second term is non-standard as our estimates for v_i will not be more and more accurate as the number of selected times increase. The UCB of v_i remains at the upper bound value 1, but becomes exact once we trigger V_i once and know the exact value of v_i. The contribution to regret by arm V_i is zero afterwards. We take extra steps to bound the second summation term, as shown in the full proof at the end of the chapter.

6.3.2 CUCB algorithm for the general case

In the general case, the agent does not know the ordering of v_i within each action. This greatly decreases the information we have.

To see this, we consider each arm i in two stages, before and after its value v_i is observed. For the first stage when v_i is unknown, the corresponding base arm Z_i is never triggered. Note the in the simpler case, Z_i is triggered whenever arm i is ordered before the winner. However, since the ordering is unknown in the general case, we do not have information about the value Z_i takes before v_i is observed. Consider in one round we observe the winner value v_j of some other arm j. Arm i could have smaller value v_i and takes a non-zero value thus not being observed, or arm i could have larger value v_i and takes zero value at the game. Importantly, we note that the problem differs from the standard

CMAB-T framework, as the triggered subset of base arm set $Z_1, \ldots, Z_n$ depends on the base arm set $V_1, \ldots, V_n$. The triggering distribution is not fixed, but depends on whether v_i is observed or not.

On the other hand, when V_i is triggered once and v_i becomes known, then the corresponding random variable Z_i is triggered whenever the winner value is smaller than v_i. We can immediately conclude that Z_i takes value zero. Thus the analysis for second stage is the same as the simpler case.

Algorithm 6.3.2 CUCB algorithm for the general case with computation oracle

For each $i \in E$, $T_i \leftarrow 0$, $\tilde{T}_i \leftarrow 0$ ▷ Total number of triggering time for Z_i and V_i
For each $i \in E$, $\hat{p}_i \leftarrow 1$, $\hat{v}_i \leftarrow 1$ ▷ Initialization of empirical estimates of parameters
for $t = 1, 2, \ldots$ **do**
    For each $i \in E$, $\rho_i \leftarrow \sqrt{\frac{3 \log t}{2 T_i}}$, $\tilde{\rho}_i \leftarrow \mathbf{1}\{\tilde{T}_i = 0\}$ ▷ Confidence radius of parameters
    Note that $\rho_i = \infty$ if $T_i = 0$. If $\tilde{T}_i = 0$ then $T_i = 0$.
    For each $i \in E$, $\bar{p}_i = \min\{\hat{p}_i + \rho_i, 1\}$, $\bar{v}_i \leftarrow \min\{\hat{v}_i + \tilde{\rho}_i, 1\}$ ▷ UCB of parameters
    $S \leftarrow \text{Oracle}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}})$ ▷ Offline oracle decides the next action
    Play S and observe winner index j and value v_j as feedback.
    if $\tilde{T}_j = 0$ **then**
        $\tilde{T}_j \leftarrow \tilde{T}_j + 1, \quad \hat{v}_j \leftarrow v_j$
    end if
    For each i such that $\hat{v}_i \geq v_j$ and $\tilde{T}_i \neq 0$, update: $T_i \leftarrow T_i + 1$
    For each i such that $\hat{v}_i > v_j$ and $\tilde{T}_i \neq 0$: $\hat{p}_i \leftarrow (1 - 1/T_i)\hat{p}_i$
    For each i such that $\hat{v}_i = v_j$ and $\tilde{T}_i \neq 0$: $\hat{p}_i \leftarrow (1 - 1/T_i)\hat{p}_i + 1/T_i$
end for

A naive approach is to adopt the CUCB algorithm for the simpler case and introduce $\tilde{T}_i$ as the triggering time for V_i. We update parameters of item i only when $\tilde{T}_i \neq 0$.

The regret bound for the CUCB algorithm is provided as follows.

Theorem 6.3.3. *If $\Delta_{min} > 0$, the CUCB algorithm defined above has the following distribution-*

dependent bound

$$R(T) \leq \sum_{i \in E} \left(\frac{2304 v_i^2 k}{\Delta_{min}^i} + 6 \frac{1}{p^*} \log_2 \frac{4k}{\Delta_{min}^i} \right) \ln T + \sum_{i \in E} \left(\lceil \log_2 \frac{4 v_i k}{\Delta_{min}^i} \rceil + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{max} + 4 v_i n$$

where $p^* = \min_{i \in E} p_i$.

We can see that this bound is not satisfactory as it contains an undesirable factor of $1/p^*$. This is due to the analysis of the first stage of the first summation term. Consider an item i with large v_i but small p_i. Recall that estimates of p_i will not be updated until v_i is observed. For a given action S, since V_i is triggered with probability $\tilde{q}_i^S$ defined in equation 6.3, action S needs to be played $\Theta(\log T / p_i)$ times. The upper bound of 1 for p_i is clearly an overestimate for this type of items, which means this type of items would cause large regrets during the period when its value is not observed. This is reflected in the bound as the term containing the factor $1/p^*$ can be arbitrarily large if some p_i value is arbitrarily small. For completeness, we will show the proof at the end of the chapter.

6.3.3 Modified algorithm for the general case

To remove the extra factor, we propose a variant of the CUCB algorithm (Algorithm 6.3.3).

The main difference to the previous algorithm is with respect to the estimates for p_i. Previously, we initiate the estimates $\hat{p}_i = 1$, which acts as an upper bound for the parameter. We start to update the estimate for p_i once we observe v_i. As discussed above, this may not be the best algorithm for items with large v_i and small p_i.

In the modified algorithm, we do not wait to update p_i until we observe v_i. On the other hand, we use the estimates $\hat{v}_i$ and pretend that Z_i is triggered and takes value zero when v_i is not observed. This intuitively makes sense as even if v_i takes value 1, the above-mentioned type of items will not be important to our regret analysis as their probability parameters remains at zero until v_i is observed.

Algorithm 6.3.3 Modified algorithm for the general case with computation oracle

For each $i \in E$, $T_i \leftarrow 0$, $\tilde{T}_i \leftarrow 0$ ▷ Total number of triggering time for Z_i and V_i

For each $i \in E$, $\hat{p}_i \leftarrow 1$, $\hat{v}_i \leftarrow 1$ ▷ Initialization of empirical estimates of parameters

for $t = 1, 2, \ldots$ **do**

  For each $i \in E$, $\rho_i \leftarrow \sqrt{\frac{3 \log t}{2T_i}}$, $\tilde{\rho}_i \leftarrow \mathbf{1}\{\tilde{T}_i = 0\}$ ▷ Confidence radius of parameters

  For each $i \in E$, $\bar{p}_i = \min\{\hat{p}_i + \rho_i, 1\}$, $\bar{v}_i \leftarrow \min\{\hat{v}_i + \tilde{\rho}_i, 1\}$ ▷ UCB of parameters

  $S \leftarrow \text{Oracle}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}})$ ▷ Offline oracle decides the next action

  Play S and observe winner index j and value v_j as feedback.

  if $\tilde{T}_j = 0$ **then**

    Reset $T_j \leftarrow 0, \quad \tilde{T}_j \leftarrow \tilde{T}_j + 1, \quad \hat{v}_j \leftarrow v_j$

  end if

  For each i such that $\hat{v}_i \geq v_j$ update: $T_i \leftarrow T_i + 1$

  For each i such that $\hat{v}_i > v_j$: $\hat{p}_i \leftarrow (1 - 1/T_i)\hat{p}_i$

  For each i such that $\hat{v}_i = v_j$: $\hat{p}_i \leftarrow (1 - 1/T_i)\hat{p}_i + 1/T_i$

end for

The regret bound for the modified algorithm is provided as follows.

Theorem 6.3.4. *If $\Delta_{min} > 0$, the modified algorithm defined above has the following distribution-dependent bound*

$$R(T) \leq \sum_{i \in E} \left(\frac{4608k}{\Delta^i_{min}} + 18 \log_2 \frac{8k}{\Delta^i_{min}} \right) \ln T + \sum_{i \in E} \left(\lceil \log_2 \frac{4v_i k}{\Delta^i_{min}} \rceil + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{max} + 4n$$

This regret bound achieves $O(\frac{nk}{\Delta} \log T)$ and is comparable to the standard CMAB problem. Compared to the simpler case, it has the same scaling up to constant factors.

We note that our problem still does not fit into the standard CMAB-T framework. As discussed above, we are *pretending* that Z_i is triggered and takes value zero. This may not be the ground truth in the case when v_i is actually less than the winner value. Therefore, using the observatation $Z_i^{(t)} = 0$ will make the estimate biased, not following the standard CMAB-T framework. In particular, for items with small v_i and large p_i, we clearly

underestimated its p_i values, since this type of items could take non-zero value but not observed due to small v_i. On the other hand, intuitively these items are not important due to small value of v_i.

To tackle with this difficulty in our analysis, we introduce the concept of *item equivalence*. In each round t, for those item i with parameters (p_i, v_i) and $\tilde{T}_{i,t} = 0$, we replace them with equivalent items i' of parameters (p'_i, v'_i) where $p'_i = p_i v_i$ and $v'_i = 1$. Note that items with small v_i and large p_i are mapped to equivalent items with large v_i and small v_i, for which our improved algorithm can estimate accurately. We will formally justify this equivalence in the following regret analysis.

Proof sketch Next, we give a sketch for the proof of Theorem 6.3.4. The full proof can be found at the end of the chapter.

We use a similar framework for regret analysis as the CUCB method. However, we note that one of the key assumption for CUCB algorithm fails to hold in the improved method, i.e., we don't always have upper confidence bounds for parameters p_i. Thus we need to make extensive modifications to the proof.

Firstly, we notice the following fact when replacing item i with (p_i, v_i) by its equivalent item i' with (p'_i, v'_i).

Lemma 6.3.5. *For any set S, $r_S(\boldsymbol{p}, \boldsymbol{v}) \leq r_S(\boldsymbol{p'}, \boldsymbol{v'})$.*

Then we consider the contribution of each action to regret Δ_t. Under the good event (E1) that the approximation oracle works well,

$$r_{S_t}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}}) \geq \alpha \cdot \mathrm{opt}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}})$$

By Lemma 6.3.5, for each t such that $1 \leq t \leq T$ we have,

$$\alpha \cdot \mathrm{opt}(\boldsymbol{p'_t}, \boldsymbol{v'_t}) \geq \alpha \cdot \mathrm{opt}(\boldsymbol{p}, \boldsymbol{v}) \tag{6.6}$$

Thus

$$\begin{aligned}\Delta_{S_t} &\leq \alpha \cdot \mathrm{opt}(\boldsymbol{p}'_t, \boldsymbol{v}'_t) - r_{S_t}(\boldsymbol{p}, \boldsymbol{v}) \\ &\leq \alpha \cdot \mathrm{opt}(\boldsymbol{p}'_t, \boldsymbol{v}'_t) - r_{S_t}(\boldsymbol{p}, \boldsymbol{v}) + r_{S_t}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}}) - \alpha \cdot \mathrm{opt}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}}) \\ &\leq r_{S_t}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}}) - r_{S_t}(\boldsymbol{p}, \boldsymbol{v}) \\ &= (r_{S_t}(\bar{\boldsymbol{p}}, \bar{\boldsymbol{v}}) - r_{S_t}(\boldsymbol{p}'_t, \boldsymbol{v}'_t)) + (r_{S_t}(\boldsymbol{p}'_t, \boldsymbol{v}'_t) - r_{S_t}(\boldsymbol{p}, \boldsymbol{v}))\end{aligned}$$

where the first inequality is due to condition 6.6, the second inequality is due to the approximation oracle, and the third inequality is due to monotonicity of r_S in $\boldsymbol{p}$ and $\boldsymbol{v}$. We call the term inside first bracket as regrets caused by estimation error $\Delta^e_{S_t}$, and the term inside the second bracket as regrets caused by replacement error $\Delta^r_{S_t}$. To obtain a tight regret upper bound, we require that the regret caused by replacement error over the time horizon T is not greater than the that by estimation error.

By the general smoothness condition 6.2.3, we have

$$\Delta^e_{S_t} \leq \sum_{i \in S_t} q^S_i v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) \tag{6.7}$$

Note that we don't need to include the v_i term as $v'_{i,t} = \bar{v}_i = 1$ for all i when v_i is not observed, and $v'_{i,t} = \bar{v}_i = v_i$ after v_i is observed. In both cases, there is no estimation error for v_i.

We also apply the general smoothness condition 6.2.3 to the second summation term and we have,

$$\Delta^r_{S_t} \leq 2 \sum_{i \in S_t} q^S_i v_i(p_i - p'_{i,t}) + \sum_{i \in S_t} \tilde{q}^S_i(v'_{i,t} - v_i) \tag{6.8}$$

To sum up, we have

$$\Delta_{S_t} \leq \sum_{i \in S_t} q^S_i v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) + 2 \sum_{i \in S_t} q^S_i v_i(p_i - p'_{i,t}) + \sum_{i \in S_t} \tilde{q}^S_i(v'_{i,t} - v_i)$$

We are going to bound Δ_{S_t} by bounding these error terms in different cases.

We can bound the first term by following the proof of Theorem 6.3.1. To see this, recall that we have reset the counts T_i and $N_{i,j,t}$ at the time v_i is observed. This is because $p'_{i,t} = p_i v_i$ when v_i is unknown and $p'_{i,t} = p_i$ afterwards, we need to reset our estimates for p_i and the confidence intervals. However, for both stages our estimates are accurate in the sense that p'_i always lies within the confidence interval which decreases as the counter number increases.

For the second term, we note that $p'_{i,t} = p_i v_i$ in the first stage, and $p'_i = p_i$ after v_i is observed. Therefore, the contribution to regret by the second term is zero in the second stage. For the first stage, this term can be analyzed in the similar way as the last term. The key observation is that $p_i - p'_{i,t} = p_i(1 - v_i) \leq p_i$. This will be the key for removing the factor of $1/p^*$ in Theorem 6.3.3.

Finally, we note that the analysis for the last summation term is the same as the simpler case, since there is no change to the triggering process of V_is.

Summing up the bounds over time horizon T, we can prove the main theorem. We can also derive the following results.

Lemma 6.3.6. *Take $M_i = \Delta^i_{min}$. Assume that all the good events $(E1)$-$(E4)$ hold, and $\Delta_{S_t} \geq M_{S_t}$ where $M_S = \max_{i \in S} M_i$, we have*

$$\sum_t^T \Delta^r_{S_t} \leq \sum_t^T \Delta^e_{S_t}$$

This justifies the intuition of using replacement items. Note that by using replacement items, our estimates for p_i are always accurate and lies within the confidence bound. Thus the total expected regret is comparable to the simpler case.

6.4 Numerical results

We perform experiments to test the results presented in the previous section. Our goal is to check how the cumulative regret depend on T, under different item value distributions.

We set $n = 9$ and $k = 3$, i.e., the ground set consists of 9 arms and each time we choose 3 arms from the ground set. We consider the following distributions for arms $i = 1, 2, \ldots, 9$. For all of these cases, the optimal super arm is $S^* = \{7, 8, 9\}$.

- Distribution 1: For the first set of distributions, we assume the support of arm i is $\{0, 0.i\}$, i.e., $v_i = 0.i$. For $i = 1, 2, \ldots, 6$, $p_i = 0.2$ and for $i = 7, 8, 9$, $p_i = 0.5$. It is a relatively simple case as distributions of optimal arms $7, 8, 9$ are far away from the suboptimal arms.

- Distribution 2: Compared to the first case, we change the distribution of the first arm such that it takes non-zero value 0.1 with probability 0.9, i.e, $v_1 = 0.1$ and $p_1 = 0.9$. In this way, we introduce an arm i with small v_i but large p_i. As discussed in the main text, this type of items causes key challenges for our algorithm. Due to small v_i, they can hardly win, thus it is very hard to observe their values. To tackle with this difficulty, in our algorithm we pretend that the arm was triggered before its value v_i is observed.

- Distribution 3: Compared to the first case, we change the distribution of the last arm such that it takes non-zero value 0.9 with probability 0.1, i.e, $v_9 = 0.9$ and $p_9 = 0.1$. Contrary to the previous case, we introduce an arm i with large v_i but small p_i. This type of high-risk high-reward items are unique to our problem and their existence make the setting hard. As discussed in the introduction, this type of items may outperform stable-value items under the expected max objective.

These distributions represent different scenarios. Distribution 1 corresponds to the case where optimal arms are easy to distinguish from suboptimal arms. In Distribution 2, there exists an item whose value is hard to observe. In Distribution 3, we have high-risk high-reward item which greatly affects the group performances. As a good algorithm for the k-max problem with value-index feedback, we expect the algorithm to select the best set under all three settings.

Figure 6.4 shows the regrets of the modified algorithm for three cases. We plot the 1-approximation regrets instead of $(1-1/e)$-approximation regret as the offline greedy oracle usually performs much better than $(1-1/e)$-approximation. From the plot, we see that T-step regret flattens as time T increases. We also plot the number of selection times for all items in figure 6.4. From the plot, we can see that the algorithm stops selecting sub-optimal arms. We conclude that our modified algorithm achieves good performances in all three cases.

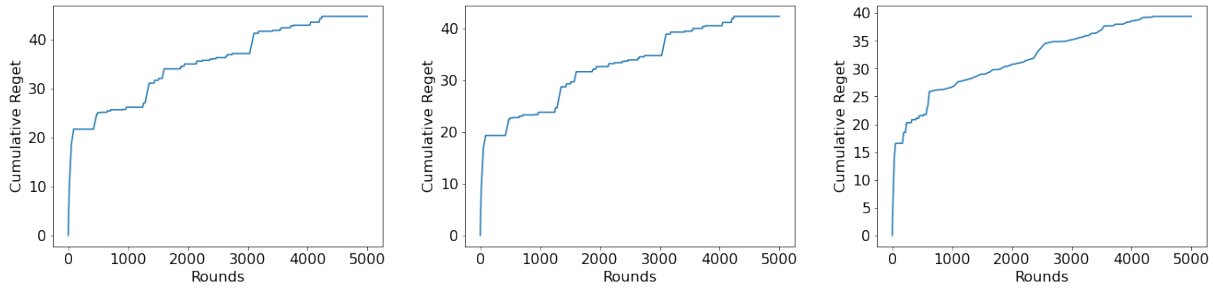

Figure 6.1: Regrets of the modified algorithm on the k-max problem with value-index feedback for Distributions 1,2,3 listed from left to right correspondingly.

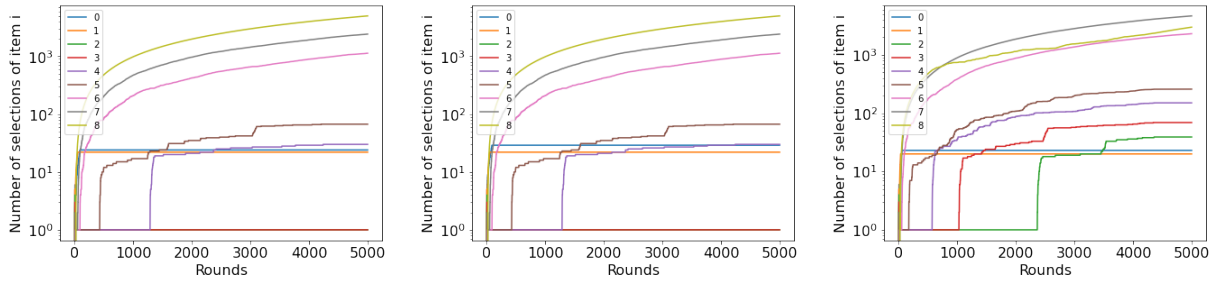

Figure 6.2: Number of selection times for all items for Distributions 1,2,3 listed from left to right correspondingly.

6.5 Conclusion

In this chapter, we studied a new class of online combinatorial optimization problem with value-index feedback, which is motivated by real-world examples in online advertising and recommender systems. This problem is inherently full-bandit, with extra feedbacks on the winner index. We proposed a CUCB algorithm to solve a special case of the problem when ordering of items are known. For the general case, we proposed a new algo-

rithm based on the concept of item equivalence. We proved its regret upper bound and showed that the algorithm performs well by simulation examples.

We note that our algorithm has a matching regret lower bound up to a $O(\sqrt{\log T})$ factor in T. It is also of interest to explore whether the bound is tight in the factor v, the item-specific values. Moreover, as mentioned at the beginning of the chapter, we restricted our studies to binary-valued items for analysis purpose. We can possibly model more complicated real-life examples by relaxing this assumptions to items with general distributions. We leave this for a future work.

6.6 Proofs

Proof of lemma 6.2.1

Recall that we can write

$$r_S(\boldsymbol{p}, \boldsymbol{v}) = p_1 v_1 + (1-p_1)p_2 v_2 + \ldots + (1-p_1)\ldots(1-p_{k-1})p_k v_k$$

assuming WLOG that $v_1 \geq v_2 \geq \ldots \geq v_k$. It is clear from the expression that $r_S(p, v)$ is monotonic increasing in v_i.

Take differential with respect to p_i we have

$$dr_S(p)/dp_i = (1-p_1)\ldots(1-p_{i-1})\left[v_i - p_{i+1}v_{i+1} - \left(\sum_{j>i}(1-p_{i+1})\ldots(1-p_j)p_{j+1}v_{j+1}\right)\right]$$

We claim that the terms inside the bracket is greater than zero. Specifically, it can be lower bounded as follows,

$$\begin{aligned}
&v_i - p_{i+1}v_{i+1} - \left(\sum_{j>i}(1-p_{i+1})\ldots(1-p_j)p_{j+1}v_{j+1}\right) \\
\geq &v_i(1-p_{i+1}) - (1-p_{i+1})p_{i+2}v_{i+2} - (1-p_{i+1})(1-p_{i+2})p_{i+3}v_{i+3} - \ldots \\
\geq &(1-p_{i+1})(1-p_{i+2})\ldots(1-p_{k-1})(v_i - p_k v_k) \\
\geq &(1-p_{i+1})(1-p_{i+2})\ldots(1-p_k)v_i \qquad (6.9)
\end{aligned}$$

Thus the reward function is monotonic increasing in p_i.

Proof of lemma 6.2.3

Let $\boldsymbol{p} = (p_1, \ldots, p_k)$ and $\boldsymbol{p'} = (p'_1, \ldots, p'_k)$. Assume that the items are ordered in descending values. For every $j = 0, 1, \ldots, k$, let

$$\boldsymbol{p}^{(j)} = (p_1, \ldots, p_j, p'_{j+1}, \ldots, p'_k)$$

Similarly for $\boldsymbol{v}$ and $\boldsymbol{v'}$.

Since $\boldsymbol{v} > \boldsymbol{v'}$, the item ordering is preserved and we have,

$$r_S(\boldsymbol{p}^{(j)}, \boldsymbol{v}^{(j)}) = p_1 v_1 + \ldots + (1-p_1)\ldots(1-p_{j-1})p_j v_j + (1-p_1)\ldots(1-p_j)p'_{j+1}v'_{j+1} + \ldots$$

$$r_S(\boldsymbol{p}^{(\boldsymbol{j-1})}, \boldsymbol{v}^{(\boldsymbol{j-1})}) = p_1 v_1 + \ldots + (1-p_1)\ldots(1-p_{j-1})p'_j v'_j + (1-p_1)\ldots(1-p'_j)p'_{j+1}v'_{j+1} + \ldots$$

Note that the only difference is caused by position j. By definition of triggering probabilities q_i^S and $\tilde{q}_i^S$ we can write,

$$\begin{aligned}
|r_S(\boldsymbol{p}^{(j)}, \boldsymbol{v}^{(j)}) - r_S(\boldsymbol{p}^{(\boldsymbol{j-1})}, \boldsymbol{v}^{(\boldsymbol{j-1})})| =& |q_j^S(p_j v_j - p'_j v'_j - \sum_{i>j}(1-p'_{j+1})\ldots(1-p'_{i-1})p'_i v'_i(p_j - p'_j)| \\
\leq& q_j^S p_j |v_j - v'_j| + q_j^S v'_j |p_j - p'_j| \\
&+ q_j^S(p'_{j+1}v'_{j+1} + (1-p'_{j+1})v'_{j+2} + \ldots)|p_j - p'_j| \\
\leq& 2q_j^S v'_j |p_j - p'_j| + \tilde{q}_j^S |v_j - v'_j|
\end{aligned}$$

where the first inequality is due to triangle inequality and the second inequality is due to equation (6.9).

Summing up over j we can obtain the desired result.

Good events for regret analysis

We define the concept of *triggering probability group* s.t. if the triggering probability of arm i in a set of base arms is p_i^S and j is a positive natural number, then the triggering probability group (i, j) is

$$\mathcal{S}_{i,j} = \{S | 2^{-j} < p_i^S \leq 2^{-j+1}\}$$

Note we have two sets of TP groups corresponding to two sets of base arms with different triggering probabilities. We call it $\mathcal{S}_{i,j}$ for q_i^S and $\tilde{\mathcal{S}}_{i,j}$ for $\tilde{q}_i^S$.

Next, we define a series of good event as follows.

E1 Event that approximation oracle works well.

$$\mathcal{F}_t = \{r_{S_t}(\bar{p}) \geq \alpha \cdot \text{opt}(\bar{p}))\}$$

Note the event-filtered regret for $\neg\mathcal{F}_t$ is bounded as

$$R(T, \neg\mathcal{F}_t) \leq (1 - \beta)T \cdot \Delta_{max}$$

E2 Event that we estimate p well.

$$\mathcal{N}_t^s = \{|\hat{p}_{i,t-1} - p_i| < \rho_{i,t}\}$$

With ρ_i defined in the algorithm as $\rho_{i,t} = \sqrt{\frac{3 \ln t}{2T_i}}$, we have

$$\Pr(\neg\mathcal{N}_t^s) \leq 2nt^{-2}$$

Thus the event-filtered regret is bounded as

$$R(T, \neg\mathcal{N}_t^s) \leq \sum_{t=1}^{T} 2nt^{-2}\Delta_{max} \leq \pi^2 n\Delta_{max}/3$$

E3 Event that triggering is nice for Z_i.

Assume arm i is in TP group $S_{i,j}$ and $\tilde{T}_i \neq 0$, i.e., its value v_i is observed. Under the condition $\sqrt{\frac{6 \ln t}{1/3 N_{i,j,t-1} \cdot 2^{-j}}} \leq 1$,

$$\mathcal{N}_t^t = \{T_{i,t-1} \geq \frac{1}{3} N_{i,j,t-1} \cdot 2^{-j}\}$$

It is known that for a series of integers $\{j_{max}^i\}$ we have

$$\Pr(\neg \mathcal{N}_t^t) \leq \sum_i j_{max}^i t^{-2}$$

Thus the event-filtered regret is bounded as

$$R(T, \neg \mathcal{N}_t^t) \leq \pi^2 \sum_i j_{max}^i \cdot \Delta_{max}/6$$

E4 Event that triggering is nice for V_i when $N_{i,j,t}$ is large.

Assume arm i is in TP group $S_{i,j}$. Under the condition $N_{i,j,t} \geq 3 p_i^{-1} \ln t \cdot 2^j$,

$$\tilde{\mathcal{N}}_t^t = \{\tilde{T}_{i,t} \neq 0\}$$

Equivalently, we can define this event in terms of TP group $\tilde{S}_{i,j}$. We remove the factor of p_i^{-1} if arm i is in TP group $\tilde{S}_{i,j}$.

Using the same proof technique as for the event that triggering is nice for Z_i, we can show the following bound for the last event.

Lemma 6.6.1. *For a series of $\{j_{max}^i\}$ and for every TP group identified by arm i and $1 \leq j \leq j_{max}^i$, we have*

$$\Pr(\neg \tilde{\mathcal{N}}_t^t) \leq \sum_i j_{max}^i t^{-2}$$

Proof of Theorem 6.3.2

We consider the contribution of each action to regret Δ_t. We introduce a positive real number $M_i = \Delta^i_{min}$. Assume that $\Delta_{S_t} \geq M_{S_t}$ where $M_S = \max_{i \in S} M_i$.

By the smoothness condition, we have

$$\Delta_{S_t} \leq r_{S_t}(\bar{\boldsymbol{p}}_{\boldsymbol{t}}, \bar{\boldsymbol{v}}_{\boldsymbol{t}}) - r_{S_t}(\boldsymbol{p}, \boldsymbol{v}) \leq \sum_{i \in S_t} q_i^S v_i (\bar{p}_{i,t} - p_i) + \sum_{i \in S_t} \tilde{q}_i^S (\bar{v}_{i,t} - v_i)$$

We use the reverse amortization trick and do the transformation such that

$$\begin{aligned}\Delta_{S_t} &\leq -M_{S_t} + 2\left(\sum_{i \in S_t} q_i^S v_i (\bar{p}_{i,t} - p_i) + \sum_{i \in S_t} \tilde{q}_i^S (\bar{v}_{i,t} - v_i)\right) \\ &\leq 2\left(\sum_{i \in S_t} q_i^S v_i (\bar{p}_{i,t} - p_i) - \frac{M_i}{4k}\right) + 2\left(\sum_{i \in S_t} \tilde{q}_i^S (\bar{v}_{i,t} - v_i) - \frac{M_i}{4k}\right)\end{aligned}$$

We may call the first term $\Delta^p_{S_t}$ and the second term $\Delta^v_{S_t}$. We bound Δ_{S_t} by bounding the two summation terms individually.

Note that for $\Delta^p_{S_t}$ we can bound following the same procedure as the proof for 6.3.1. However, we cannot use the same procedure for $\Delta^v_{S_t}$. The key difference is that our estimate for v_i will not be more and more accurate as the number of selected times increase. We know the exact value of v_i as soon as we trigger it once. We assume that the arm i is in TP group $\tilde{S}_{ij}$. Let j_i be the index of the TP group with $S_t \in \tilde{S}_{i,j_i}$. We take $j^i_{max} = \log_2 \frac{4k}{M_i}$.

- Case 1: $1 \leq j_i \leq j^i_{max}$. Then $\tilde{q}_i^S \leq 2 \cdot 2^{j_i}$,

  $$\tilde{q}_i^S (\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i} \cdot \mathbb{1}\{\tilde{T}_{i,t} = 0\}$$

  Under the good event $\tilde{\mathcal{N}}^t_t$, we know that when $N_{i,j_i,t-1} \geq 3 \ln t \cdot 2^j$, the contribution to regret is zero. Otherwise, it is bounded by

  $$\tilde{q}_i^S (\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i}$$

- Case 2: $j_i > j^i_{max} = \log_2 \frac{4k}{M_i}$. In this case,

$$\tilde{q}^S_i(\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i} \leq \frac{M_i}{4k}$$

  Thus the contribution to regret is non-positive in this case.

Then we calculate the filtered regret under the above mentioned good events and the event that $\Delta_{S_t} \geq M_{S_t}$. Note that

$$R(T, \{\Delta_{S_t} \geq M_{S_t}\}, \mathcal{F}_t, \mathcal{N}^s_t, \mathcal{N}^t_t, \tilde{\mathcal{N}}^t_t) \leq \sum_{t=1}^{T} \Delta^p_{S_t} + \sum_{t=1}^{T} \Delta^v_{S_t}$$

By Theorem 6.3.1, we know the first term is bounded by

$$\sum_{t=1}^{T} \Delta^p_{S_t} \leq \sum_i \frac{2304 k v_i^2 \ln T}{M_i} + 4 v_i n$$

We focus on bounding the second term. Note that

$$\sum_{t=1}^{T} \Delta^v_{S_t} = \sum_i \sum_j \sum_{s=0}^{N_{i,j,T-1}} \kappa_{j_i,T}(M_i, s)$$

where

$$\kappa_{j,T}(M, s) = \begin{cases} 2 \cdot 2^{-j} & \text{if } s < 3 \ln t \cdot 2^j \\ 0 & \text{if } s \geq 3 \ln t \cdot 2^j \end{cases}$$

For every i and j, we have

$$\sum_{s=0}^{N_{i,j,T-1}} \kappa_{j_i,T}(M_i, s) \leq \sum_{s=0}^{3 \ln T \cdot 2^{j_i}} \kappa_{j_i,T}(M_i, s) = 6 \ln T$$

Hence the second term is bounded by

$$\sum_{t=1}^{T} \sum_{i \in \tilde{S}_t} \tilde{\kappa}_{j_i,T}(M_i, N_{i,j_i,t-1}) \leq \sum_i 6 \ln T \cdot \log_2 \frac{4k}{M_i}$$

To calculate the total regret, we recall the filtered regrets for the case when good events fail to hold.

$$\begin{aligned} R(T, \neg\mathcal{F}_t) &\leq (1-\beta)T \cdot \Delta_{max} \\ R(T, \neg\mathcal{N}_t^s) &\leq \pi^2 n \Delta_{max}/3 \\ R(T, \neg\mathcal{N}_t^t) &\leq \pi^2 \sum_i j_{max}^i \cdot \Delta_{max}/6 \\ R(T, \neg\tilde{\mathcal{N}}_t^t) &\leq \pi^2 \sum_i j_{max}^i \cdot \Delta_{max}/6 \end{aligned} \tag{6.10}$$

Adding up the filtered regrets shown above, we can prove the theorem.

Proof of lemma 6.3.3

As discussed in the main text, we only need to show the bound for the first summation term $\Delta_{S_t}^p$. We assume that the arm i is in TP group S_{ij}. Let j_i be the index of the TP group with $S_t \in S_{i,j_i}$. We take $j_{max}^i = \log_2 \frac{4k}{M_i}$.

We consider each item in two stages, before and after its value v_i is observed. When v_i has not been observed, p_i is upper bounded by 1 as v_i. We can derive similar bounds as for v_i in this stage. Under the good event $\tilde{\mathcal{N}}_t^t$, we know that v_i is observed with high probability in $3p_i^{-1} \ln t \cdot 2^j$ time steps.

- Case 1: $1 \leq j_i \leq j_{max}^i$. Then $q_i^S \leq 2 \cdot 2^{-j_i}$,

$$q_i^S(\bar{p}_{i,t} - p_i) \leq 2 \cdot 2^{-j_i}$$

- Case 2: $j_i > j_{max}^i$. In this case, the contribution to regret is non-positive.

Once v_i is observed, we can bound the regret in similar way as standard CMAB-T problems. Recall that we reset the counter $N_{i,j,t}$ as soon as v_i is observed.

- Case 1: $1 \leq j_i \leq j^i_{max}$. Then $q^S_i \leq 2 \cdot 2^{-j_i}$,

$$q^S_i(\bar{p}_{i,t} - p_i) \leq 2 \cdot 2^{-j_i} \cdot 2\rho_i$$

The contribution to regret will be non-positive if $N_{i,j,t} \geq l_{j_i,T}(M_i)$ where

$$l_{j,T}(M) = \lfloor \frac{1152 \cdot 2^{-j} v_i^2 K^2 \ln T}{M^2} \rfloor$$

- Case 2: $j_i > j^i_{max}$. Similarly, the contribution to regret is non-positive.

Next, we sum up the bounds over the time horizon T. Recall that we split the counter into two stages. For notation convenience, we use $N_{i,j,1}$ to denote the counts for the first stage, and $N_{i,j,2}$ to denote the counts for the second stage. Note that $N_{i,j,T} = N_{i,j,1} + N_{i,j,2}$.

$$\sum_{t=1}^{T} \Delta^p_{S_t} = \sum_i \sum_j \left(\sum_{s=0}^{N_{i,j,1}} \kappa_{j_i,T}(M_i, s) + \sum_{s=0}^{N_{i,j,2}} \tilde{\kappa}_{j_i,T}(M_i, s) \right)$$

where

$$\kappa_{j,T}(M, s) = \begin{cases} 2 \cdot 2^{-j} & \text{if } s < 3p_i^{-1} \ln t \cdot 2^j \\ 0 & \text{if } s \geq 3p_i^{-1} \ln t \cdot 2^j \end{cases}$$

and

$$\tilde{\kappa}_{j,T}(M, s) = \begin{cases} 2 \cdot 2^{-j} & \text{if } s = 0 \\ 4 \cdot 2^{-j} \rho_i & \text{if } s \leq l_{j,T}(M) \\ 0 & \text{if } s \geq l_{j,T}(M) \end{cases} .$$

The first stage can be analyzed similarly as for v_i. For every i and j, we have

$$\sum_{s=0}^{N_{i,j,1}} \kappa_{j_i,T}(M_i, s) \leq 3p_i^{-1} \ln T \cdot 2^{j_i} \cdot 2 \cdot 2^{-j_i} = 6p_i^{-1} \ln T$$

Hence in first stage we have,

$$\sum_{t=1}^{T_1} \sum_{i \in S_t} \tilde{\kappa}_{j_i,T}(M_i, N_{i,j_i,t-1}) \leq \sum_i 6p_i^{-1} \ln T \cdot \log_2 \frac{4k}{M_i} \tag{6.11}$$

Note that this is where the extra p_i factor comes from.

The second stage can be analyzed following the same procedure as the proof for Theorem 6.3.1. We omit the proof and gives the results as follows.

$$\sum_{t=1}^{T_2} \sum_{i \in S_t} \kappa_{j_i,T}(M_i, N_{i,j_i,t-1}) \leq \sum_i \frac{2304 v_i^2 k \ln T}{M_i} + 4 v_i n$$

Since the bound for the second summation term $\Delta_{S_t}^v$ is the same as the simpler case, we conclude the theorem. Note that the extra term in the regret bound is due to equation (6.11).

Proof of lemma 6.3.5

Assume WLOG that $S = [k]$ and $v_1 \geq v_2 \geq \ldots \geq v_k$. Recall that we can write

$$r_S(\boldsymbol{p}, \boldsymbol{v}) = p_1 v_1 + (1 - p_1) p_2 v_2 + \ldots + (1 - p_1) \ldots (1 - p_{k-1}) p_k v_k$$

Now $\boldsymbol{p} = (p_1, \ldots, p_k)$ and $\boldsymbol{p'} = (p'_1, \ldots, p'_k)$; similarly for $\boldsymbol{v}$ and $\boldsymbol{v'}$. Let

$$\boldsymbol{p}^{(j)} = (p'_1, \ldots, p'_j, p_{j+1}, \ldots, p_k)$$

After changing p_1 to $p'_1 = p_1 v_1$ and v_1 to $v'_1 = 1$,

$$r_S(\boldsymbol{p^{(1)}}, \boldsymbol{v^{(1)}}) = p_1 v_1 + (1 - p_1 v_1) p_2 v_2 + \ldots + (1 - p_1 v_1) \ldots (1 - p_{k-1}) p_k v_k$$

Clearly we have $r_S(\boldsymbol{p^{(1)}}, \boldsymbol{v^{(1)}}) \geq r_S(\boldsymbol{p}, \boldsymbol{v})$. Following the same argument, we can see that $r_S(\boldsymbol{p^{(2)}}, \boldsymbol{v^{(2)}}) \geq r_S(\boldsymbol{p^{(1)}}, \boldsymbol{v^{(1)}})$. Continue this way to $r_S(\boldsymbol{p^{(k)}}, \boldsymbol{v^{(k)}})$ we can prove the lemma.

Proof of Theorem 6.3.4

By the general smoothness condition, we have

$$\Delta_{S_t} \leq \sum_{i \in S_t} q_i^S v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) + 2 \sum_{i \in S_t} q_i^S v_i (p_i - p'_i) + \sum_{i \in S_t} \tilde{q}_i^S (v'_{i,t} - v_i)$$

Key step: Bound contribution of each action to regret Firstly, we use the reverse amortization trick to perform a transformation. Take $M_i = \Delta^i_{min}$. Assume that all the good events mentioned above hold, and $\Delta_{S_t} \geq M_{S_t}$ where $M_S = \max_{i \in S} M_i$.

$$\begin{aligned}
\Delta_{S_t} &\leq -M_{S_t} + 2 \left(\sum_{i \in S_t} q_i^S v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) + 2 \sum_{i \in S_t} q_i^S v_i (p_i - p'_i) + \sum_{i \in S_t} \tilde{q}_i^S (v'_{i,t} - v_i) \right) \\
&\leq 2 \left[\left(\sum_{i \in S_t} q_i^S v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) - \frac{M_i}{8k} \right) + 2 \left(\sum_{i \in S_t} q_i^S v_i (p_i - p'_i) - \frac{M_i}{8k} \right) + \left(\sum_{i \in S_t} \tilde{q}_i^S (v'_{i,t} - v_i) - \frac{M_i}{4k} \right) \right]
\end{aligned} \tag{6.12}$$

Let j_i be the index of the TP group with $S_t \in S_{i,j_i}$. Take $j^i_{max} = \log_2 \frac{8k}{M_i}$. In the case $j_i > j^i_{max}$, we note that the contribution to regret is non-positive for all three terms. This is because

$$\tilde{q}_i^S (\bar{v}_{i,t} - v_i) \leq q_i^S (\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i} \leq \frac{M_i}{8k}$$

Then, for the case $1 \leq j_i \leq j^i_{max}$, we consider each term individually.

- The first term $q_i^S v'_{i,t}(\bar{p}_{i,t} - p'_{i,t})$.

  Recall that we have reset the counts T_i and $N_{i,j,t}$ at the time v_i is observed. This is because $p'_{i,t} = p_i v_i$ when v_i is unknown and $p'_{i,t} = p_i$ afterwards, we need to reset our estimates for p_i and the confidence intervals. A key observation is that within both stages our estimates are accurate in the sense that the approximation error decreases as the counter number increases in the following way.

$$\bar{p}_{i,t} - p'_{i,t} \leq 2\rho_i$$

Thus in both stages the contribution to regrets will be non-positive if $N_{i,j,t} \geq l_{j_i,T}(M_i)$ where

$$l_{j,T}(M) = \lfloor \frac{288 \cdot 16 \cdot 2^{-j} k^2 \ln T}{M^2} \rfloor$$

Otherwise

$$q_i^S v'_{i,t}(\bar{p}_{i,t} - p'_{i,t}) \leq 2 \cdot 2^{-j_i} \cdot 2\rho_i$$

- The second term $q_i^S v_i(p_i - p'_{i,t}) - \frac{M_i}{8k}$.
  As $p'_{i,t} = p_i v_i$ and $1 \leq j_i \leq j^i_{max}$, we have $q_i^S \leq 2 \cdot 2^{-j_i}$,

$$q_i^S v_i(p_i - p'_{i,t}) \leq 2 \cdot 2^{-j_i} p_i v_i(1 - v_i)$$

  Under the good event $\tilde{\mathcal{N}}_t^t$, we know that v_i is observed with high probability in $3p_i^{-1} \ln t \cdot 2^j$ time steps and $p'_{i,t} = p_i$. Thus the term is upper bounded by zero when $N_{i,j_i,t-1} \geq 3p_i^{-1} \ln t \cdot 2^j$, and by $2 \cdot 2^{-j_i} p_i$ otherwise.

- The third term $\tilde{q}_i^S(v'_{i,t} - v_i) - \frac{M_i}{4k}$.
  Since $1 \leq j_i \leq j^i_{max}$, we have $\tilde{q}_i^S \leq 2 \cdot 2^{-j_i} p_i$,

$$\tilde{q}_i^S(\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i} p_i \tilde{\rho}_i \leq 2 \cdot 2^{-j_i} p_i \cdot \mathbb{1}\{\tilde{T}_{i,t} = 0\}$$

  Under the good event $\tilde{\mathcal{N}}_t^t$, we know that when $N_{i,j_i,t-1} \geq 3p_i^{-1} \ln t \cdot 2^j$, the contribution to regret is zero. Otherwise, it is bounded by

$$\tilde{q}_i^S(\bar{v}_{i,t} - v_i) \leq 2 \cdot 2^{-j_i} p_i$$

  We note that this upper bound is the same as the second term.

Summing over the time horizon Next, we sum up Δ_{S_t} over time T and calculate the filtered regret under the above mentioned good events and the event that $\Delta_{S_t} \geq M_{S_t}$, i.e, $R(\{\Delta_{S_t} \geq M_{S_t}\}, \mathcal{F}_t, \mathcal{N}_t^s, \mathcal{N}_t^t, \tilde{\mathcal{N}}_t^t)$.

By equation (6.12), we know that the filtered regret can be upper bounded by sum of three terms over the time horizon T.

By Theorem 6.3.1, we know that

$$\sum_{t=1}^{T}\left(\sum_{i\in S_t} q_i^S v'_{i,t}(\bar{p}_{i,t}-p'_{i,t})-\frac{M_i}{8k}\right)\leq 2\sum_i \frac{4\cdot 576k\ln T}{M_i}+4n$$

Note that the extra factor of two considers both stages.

By the analysis for the previous CUCB algorithm, we know that

$$\sum_{t=1}^{T}\left(\sum_{i\in S_t} \tilde{q}_i^S (v'_{i,t}-v_i)-\frac{M_i}{4k}\right)\leq \sum_i 6\log_2\frac{8k}{M_i}\cdot\ln T$$

Similarly we can bound

$$2\sum_{t=1}^{T}\left(\sum_{i\in S_t} q_i^S v_i(p_i-p'_{i,t})-\frac{M_i}{8k}\right)\leq 2\sum_i 6\log_2\frac{8k}{M_i}\cdot\ln T$$

Add up the filtered regrets in equation (6.10) for the case when good events (E1)-(E4) fail to hold, we prove the theorem.

Bibliography

I. Adler and S. M. Ross. The coupon subset collection problem. *Journal of Applied Probability*, 38(3):737–746, 2001.

M. Agarwal, V. Aggarwal, C. J. Quinn, and A. K. Umrawal. Stochastic top-*k* subset bandits with linear space and non-linear feedback. In *Algorithmic Learning Theory*, pages 306–339. PMLR, 2021.

N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

A. E. Alaoui and A. Montanari. On the computational tractability of statistical estimation on amenable graphs, 2019.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *Information and Inference: A Journal of the IMA*, 3:224–294, 2014.

S. Aral. Networked experiments. *The Oxford handbook of the economics of networks*, pages 376–411, 2016.

A. Asadpour and H. Nazerzadeh. Maximizing stochastic monotone submodular functions. *Management Science*, 62(8):2374–2391, 2016.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

A. Badanidiyuru, S. Dobzinski, H. Fu, R. Kleinberg, N. Nisan, and T. Roughgarden. Sketching valuation functions. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, page 1025–1035, USA, 2012. Society for Industrial and Applied Mathematics.

M.-F. Balcan and N. J. Harvey. Learning submodular functions. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, page 793–802, New York, NY, USA, 2011. Association for Computing Machinery.

O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley series in probability and mathematical statistics, 1978.

D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

A. A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause. Guaranteed Non-convex Optimization: Submodular Maximization over Continuous Domains. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 111–120, 20–22 Apr 2017.

E. Breza. Field experiments, social networks, and development. *The Oxford handbook of the economics of networks*, 4, 2016.

E. J. Candés and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, 48(1):27–42, 02 2020.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Ann. Appl. Probab.*, 21(4):1400–1435, 08 2011.

M. Chen, K. Kato, and C. Leng. Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a), 2021.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29, 2016.

F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

K. Cohavi and S. Dobzinski. Faster and simpler sketches of valuation functions. *ACM Trans. Algorithms*, 13(3), Mar. 2017.

O. Cooley, M. Kang, and C. Koch. The size of the giant high-order component in random hypergraphs. *Random Structures & Algorithms*, 53(2):238–288, 2018.

O. Cooley, M. Kang, and C. Koch. The size of the giant component in random hypergraphs: a short proof. In *The Electronic Journal of Combinatorics*, volume 26, 2019.

G. Cormode and K. Yi. *Small Summaries for Big Data*. Cambridge University Press, 2020.

K. P. Costello and V. Vu. On the rank of random sparse matrices. *Combinatorics, Probability and Computing*, 19(3):321–342, 2010.

K. P. Costello and V. H. Vu. The rank of random graphs. *Random Structures & Algorithms*, 33(3):269–285, 2008.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. 2008.

M. Desai and V. Rao. A characterization of the smallest eigenvalue of a graph. *Journal of Graph Theory*, 18(2):181–194, 1994.

D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2), 1998.

P. Erdös and T. Gallai. Gráfok elöirt fokszámú pontokkal. *Matematikai Lapok*, 11(264-274), 1960.

P. Erdös and A. Rényi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

C. G. Esseen. On the concentration function of a sum of independent random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 9(4):290–308, 1968.

S. Fallat and Y.-Z. Fan. Bipartiteness and the least eigenvalue of signless laplacian of graphs. *Linear Algebra and its Applications*, 436(9):3254 – 3267, 2012.

T. Feder and M. Mihail. Balanced matroids. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 26–38, 1992.

S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems*, 27(2):025010, 2011.

C. Godsil and G. Royle. *Algebraic Connectivity of Graphs*. Springer, 2001.

M. X. Goemans, N. J. Harvey, S. Iwata, and V. Mirrokni. Approximating submodular functions everywhere. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 535–544. SIAM, 2009.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *International conference on machine learning*, pages 100–108. PMLR, 2014.

B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems 27*, pages 1475–1483. Curran Associates, Inc., 2014.

T. P. Hayes. A large-deviation inequality for vector-valued martingales, 2005.

C. Hillar and A. Wibisono. Maximum entropy distributions on graphs, 2013.

W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

T.-K. Huang, C.-J. Lin, and R. C. Weng. Ranking individuals by group comparisons. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 425–432, New York, NY, USA, 2006a. ACM.

T.-K. Huang, R. C. Weng, and C.-J. Lin. Generalized bradley-terry models and multi-class probability estimates. *J. Mach. Learn. Res.*, 7:85–115, Dec. 2006b.

K. Joag-Dev and F. Proschan. Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295, 1983. ISSN 00905364. URL `http://www.jstor.org/stable/2240482`.

Kaggle.com. New york times articles & comments (2020), 2020. URL `https://www.kaggle.com/benjaminawd/new-york-times-articles-comments-2020`.

Kaggle.com. Youtube dislikes dataset, 2021. URL `https://www.kaggle.com/dmitrynikolaev/youtube-dislikes-dataset`.

M. Karoński and T. Łuczak. The phase transition in a random hypergraph. *Journal of Computational and Applied Mathematics*, 142(1):125 – 135, 2002. Probabilistic Methods in Combinatorics and Combinatorial Optimization.

S. Katariya, B. Kveton, C. Szepesvari, C. Vernade, and Z. Wen. Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, pages 392–401. PMLR, 2017.

M. J. Klass. A Method of Approximating Expectations of Functions of Sums of Independent Random Variables. *The Annals of Probability*, 9(3):413 – 428, 1981.

J. Kleinberg and M. Raghu. Team performance with test scores. *ACM Trans. Econ. Comput.*, 6(3–4), Oct. 2018. ISSN 2167-8375.

D. Koller, N. Friedman, S. Džeroski, C. Sutton, A. McCallum, A. Pfeffer, P. Abbeel, M.-F. Wong, C. Meek, J. Neville, et al. *Introduction to statistical relational learning*. MIT press, 2007.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015a.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015b.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28, 2015c.

D. Lee, M. Vojnovic, and S.-Y. Yun. Test score algorithms for budgeted stochastic utility maximization, 2021.

B. Lehmann, D. Lehmann, and N. Nisan. Combinatorial auctions with decreasing marginal utilities. *Games and Economic Behavior*, 55(2):270–296, 2006.

A. Mehta, U. Nadav, A. Psomas, and A. Rubinstein. Hitting the high notes: Subset selection for maximizing expected order statistics. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15800–15810. Curran Associates, Inc., 2020.

T. Minka, R. Cleven, and Y. Zaykov. Trueskill 2: An improved bayesian skill rating system. *Technical Report*, 2018.

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

R. Mukherjee, S. Mukherjee, and S. Sen. Detection Thresholds for the β-Model on Sparse Graphs. *ArXiv e-prints*, Aug. 2016.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978a.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978b.

M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2): 167–256, 2003.

U. N. Peled and M. K. Srinivasan. The polytope of degree sequences. *Linear Algebra and its Applications*, 114-115:349 – 377, 1989.

I. Rejwan and Y. Mansour. Top-k combinatorial bandits with full-bandit feedback. In *Algorithmic Learning Theory*, pages 752–776. PMLR, 2020.

A. Rinaldo, S. Petrovic, and S. E. Fienberg. Maximum likelihood estimation in the β-model. *Ann. Statist.*, 41(3):1085–1110, 06 2013.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

F. Salehi, E. Abbasi, and B. Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems 32*, pages 11982–11992. Curran Associates, Inc., 2019.

S. Sekar, M. Vojnovic, and S. Yun. A test score-based approach to stochastic submodular optimization. *Manag. Sci.*, 67(2):1075–1092, 2021.

N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

M. J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3): 310–313, 1981.

A. Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

T. Soma and Y. Yoshida. A generalization of submodular cover via the diminishing return property on the integer lattice. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

W. Stadje. The collector's problem with group drawings. *Advances in Applied Probability*, 22(4):866–882, 1990. doi: 10.2307/1427566.

Y. Sui, V. Zhuang, J. W. Burdick, and Y. Yue. Multi-dueling bandits with dependent arms. *arXiv preprint arXiv:1705.00253*, 2017.

R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

D. M. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2): 305–321, 1978.

J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

M. Vojnovic and S.-Y. Yun. Parameter estimation for generalized thurstone choice models. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 498–506, 2016.

Q. Wang and W. Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.

T. Yan and J. Xu. A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika*, 100(2):519–524, 2013.

T. Yan, H. Qin, and H. Wang. Asymptotics in undirected random graph models parameterized by the strengths of vertices. *Statistica Sinica*, 26(1):273–293, 2016.