Learning From People

Nihar Shah

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2017-133
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-133.html

July 28, 2017

Learning From People

by

Nihar Bhadresh Shah

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin J. Wainwright, Chair
Professor Kannan Ramchandran
Professor Thomas L. Griffiths
Professor Christos Papadimitriou

Summer 2017

Learning From People

Abstract

Learning From People

by

Nihar Bhadresh Shah

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Martin J. Wainwright, Chair

Learning from people represents a new and expanding frontier for data science. Crowdsourcing, where data is collected from non-experts online, is now extensively employed in academic research, industry, and also for many societal causes. Two critical challenges in crowdsourcing and learning form people are that of (i) developing algorithms for maximally accurate learning and estimation that operate under minimal modeling assumptions, and (ii) designing incentive mechanisms to elicit high-quality data from people. In this thesis, we addresses these fundamental challenges in the context of several canonical problem settings that arise in learning from people.

For the challenge of estimation, there are various algorithms proposed in past literature, but their reliance on strong parameter-based assumptions is severely limiting. In this thesis, we introduce a class of "permutation-based" models that are considerably richer than classical parameter-based models. We present algorithms for estimation which we show are both statistically optimal and significantly more robust than prior state-of-the-art methods. We also prove that our estimators automatically adapt and are simultaneously optimal over the classical parameter-based models as well, thereby enjoying a surprising win-win in the statistical bias-variance tradeoff.

As for the second challenge of incentivizing people, we design a class of payment mechanisms that take a "multiplicative" form. For several common interfaces in crowdsourcing, we show that these multiplicative mechanisms are surprisingly the only mechanisms that can guarantee honest responses and satisfy a mild and natural requirement which we call no-free-lunch. We show that our mechanisms have several additional desirable qualities. The simplicity of our mechanisms imparts them with an additional practical appeal.

Contents

Acknowledgments

I must begin with a big thank you to my advisors Martin J. Wainwright and Kannan Ramchandran.

I was very fortunate to be advised by Martin. When I started out at Berkeley, I had no background in statistics and machine learning. It is largely from Martin, through his classes as well as his guidance, that I understood these topics and developed the tools for research in this area. Martin also helped a lot in improving my writing – he used to edit the tex files of the papers and leave very useful suggestions regarding my writing. Martin also gave me the freedom to pursue the research directions of my choice. I will look towards various aspects of his advising style as I myself move to a faculty position.

It was great to have Kannan as my co-advisor. He is always full of energy. In fact, this energy is also reflected in our group's dynamics and in group meetings where there is generally tremendous amount of discussion. For my research, I was looking for independence on what to work on, and Kannan was more than happy to give me a free hand. His broad research interests meant that our group was also working on many different topics, giving me a chance to learn and explore various directions. Kannan is always super enthusiastic about research and his enthusiasm is quite infectious.

Not only was it a great experience working with Martin and Kannan, it was also tremendous fun. Their sportiveness is evidenced by the fact that about 15% of my dissertation talk was dedicated exclusively to a roast of my advisors. Thanks for putting up with all of my pranks through these years![1]

During my PhD, I also spent two summers in wonderful internships with Denny Zhou at Microsoft Research Redmond. In fact, Denny was the one who introduced me to crowdsourcing. I am thankful to Denny for the wonderful collaborations we have had during the internships and after.

I am also thankful to my committee members Christos Papadimitriou and Tom Griffiths, who provided me with very timely and helpful feedback during my qualifying examination and in all of our subsequent interactions.

I had the pleasure to have several fantastic collaborations with Abhay Parekh on various topics, where Abhay brought the "real world" perspective from industry and startup-land. I also had wonderful collaborations with Sivaraman Balakrishnan on topics in learning theory, and also a lot of fun chatting about various things. Starting with a chance meeting at the NIPS conference and then in the CS189 course, I also had the awesome experience of working with Isabelle Guyon on crowdsourcing in the wild. Before coming to Berkeley, I worked with P. Vijay Kumar at the Indian Institute of Science for my masters. I learnt my first steps of research from him and these foundations have been exceedingly helpful.

I also had the great fortune of being able to collaborate with Jean Walrand, Yuval Peres, Adityanand Guntuboyina, John Platt, Ulrike von Luxburg, Anant Sahai, Chris Meek, Jianwei Huang, Kangwook Lee, Reinhard Heckel, Yuan Luo, David Marn, Vijay Kamble, Joseph

[1] Dear reader, please feel free to ask Martin and Kannan about "match-making" and "updates" respectively the next time you meet them.

Bradley, Longbo Huang, Jingyan Wang, Preetum Nakkiran, Qiang Liu, Sergio Escalera, Behzad Tabibian, Krikamol Muandet, and Ka Kit Lam. Thank you all for the wonderful collaborations.

I was also lucky to have some amazing teachers in the courses that I took at Berkeley – Jean Walrand, Martin J. Wainwright, Michael Jordan, Noureddine El Karoui, Suvrit Sra, Elchanan Mossel, Yun Song, Tsu Jae King Liu and Vivek Subramanian – and I would like to thank them for teaching me the tools and skills that have been and will continue to be extremely useful in my career.

I also GSIed two courses: I was a GSI for EE226A (Random processes in systems) with Thomas Courtade, and the head GSI for CS 189 (Introduction to machine learning) with Isabelle Guyon and Alyosha Efros. I would like to thank Tom, Isabelle, and Alyosha for teaching me so much about teaching, grading, managing a course. These experiences are going to be very helpful in the next phase of my career as a faculty.

I am also grateful to all the staff in the EECS department, and in particular Shirley Salanio and Kim Kail, for always being so extremely helpful.

And also thank you to all of WiFo and BLISS for the amazing time that I spent here, and to other members in Martins group and many other friends whom I interacted with during my time at Berkeley – David Tse, Venkat Ananthram, Sahand Negahban, Sameeer Pawar, Venkatesan Ekambaram, Pulkit Grover, Kris Woyach, Guy Bresler, Sudeep Kamath, Sreeram Kannan, Ramtin Pedarsani, Dimitris Papailiopoulos, Po-Ling Loh, Varun Jog, Gireeja Ranade, Kate Harrison, Giulia Fanti, Ilan Shomorony, Yudong Chen, Longbo Huang, Nima Noorshams, Nebojsa Milosavljevic, Salim El Rouayheb, Se Yong Park, Hao Zhang, John Duchi, Alekh Agarwal, Bao Zhang, Abolfazl Motahari, Eren Sasoglu, Baris Nakiboglu, Soheil Mohajer, Omur Ozel, Vasuki Narasimha Swamy, Stephan Adams, Fanny Yang, Reza Abbasi Asl, Yuting Wei, Vidya Muthukumar, Ashwin Pananjady, Simon Li, Matthieu Finiasz, Payam Delgosha, Orhan Ocal, Dong Yin, Raaz Dwivedi, Soham Phade, Sang Min Han, Vipul Gupta, Avishek Ghosh, Zihan Liu, Govinda Kamath, Rachid Elazouzi, Aaditya Ramdas, Yun Yang, Yuchen Zhang, Mahdi Soltanolkotabi, Mert Pilanci, Aymeric Dieuleveut, Leah Dickstein, Kabir Chandrasekher, Cheng Mao, Geoffrey Schiebinger, Elina Robeva, Nishant Totla, and Sahaana Suri. It was also great to have Venky, Gireeja and Vasuki as my BLISSful neighbours with whom I had plenty of fun conversations. On a related note, one of the contributions I made during my PhD that I am very happy about is solving the incredibly hard problem getting all six BLISS faculty (and all the students) together at the same place at the same time for a group picture! I also thank the powers that be in the lab for choosing my contribution "Berkeley Laboratory for Information Systems and Sciences" or "BLISS" as the name for the new incarnation of our lab. Also, a special thanks to all the folks joining in for our ultimate frisbee games every Friday.

Lastly, and most importantly, I am deeply thankful to my family for their constant support and guidance. There is no way I could have done any of this without the enormous support of my mother Ragini Shah. I am also grateful to Rashmi Vinayak for all the collaborations—both academic [175, 176, 194–204, 217, 226–233] and non-academic—we have had over the past many years. Rashmi, have I ever told you how awesomely awesome you are? ;) And finally, many thanks to our eight-month old Aanya. Rashmi and I were in the academic job market right after she was born: we submitted our applications at around her birth and interviewed starting when she was just over two months old. She happily flew for about a dozen interviews across the US and put up with all the crazy interview schedules without any fuss. Aanya, you may not remember any of these events by the time you are old enough to read, so let me tell you how immensely helpful you were even at such a young age! (And enjoy all those frequent flyer miles that you have accumulated.)

Chapter 1

Introduction

Data from people arises in a variety of different domains. For instance, “crowdsourcing” – where (non-expert) people online are asked to perform tasks that are too hard for machines and too expensive and time consuming for experts – has gained tremendous popularity. Crowdsourcing is used to collect scientific data in various fields of science and engineering such as bioinformatics [29, 72, 93, 96, 124, 138, 142, 147, 162, 173, 262], astronomy [87, 128, 150, 165], psychology [97, 123, 152, 170], epidemiology [127], medicine [155, 242, 269], radiology [179], ontology [214], environmental modeling [81], history [139], marketing and business [55, 266], fashion [268], and computer science [101, 134, 148, 211, 273]. For instance, the paper [142] employs crowdsourcing for cancer detection; the paper [134] surveys the use of crowdsourcing to collect annotations for training machine learning algorithms to perform computer vision tasks.

Crowdsourcing is also vastly popular in the industry. For instance, a recent survey [210] estimates that 85 of the 100 “best global brands” employ crowdsourcing. Crowdsourcing and learning from people helps in many societal causes, such as healthcare [167], detecting cyberbullying [208], helping the blind [16], search and rescue [183], crisis mapping [180], improving nutrition [181], and others. It is also the crowd that has been the key driver for many innovative and disruptive technologies in the sharing economy such as Uber, Lyft, and Airbnb, and in sharing experiences such as Yelp, TripAdvisor, and IMDB. Even in our day to day academic lives, we need to learn from people in the form of personalized teaching as well as peer grading in both conventional classrooms and massive open online courses (MOOCs) [187, 225].

The prevalence of data from people can be attributed to the proliferation of the Internet across the world and to the many platforms that have recently emerged to harness this opportunity. For instance, there are many online crowdsourcing platforms such as Amazon Mechanical Turk (`mturk.com`) and others (`crowdflower.com`, `microtask.com`, `upwork.com`) on which any entity (called a “requester”) may put up a task along with a promised payment, and then any person online (called a “worker”) may complete the task in exchange for the promised payment. These crowdsourcing platforms have gained tremendous popularity because they are usually much cheaper and the tasks are completed much faster as compared to enlisting experts, thereby making it a highly scalable process [243]. Furthermore, with the current platforms for crowdsourcing, the initial overhead of setting up a crowdsourcing task is minimal.

Data obtained from people is usually quite noisy. For instance, when workers on crowdsourcing platforms are asked objective binary-choice questions, it is not uncommon to have an error as large as 40%; when people are asked for their personal preferences over a set of items, the data obtained typically has many inconsistencies. There are various reasons for this noise, ranging from the lack of expertise in objective tasks, differences in preferences across people, absence of proper incentives, and sometimes due to inadequacies of the interfaces. These issues give rise to two fundamental challenges in crowdsourcing and learning from people: (i) accurate estimation from noisy data obtained from people, and (ii) incentivizing people to provide better data. These two challenges form the two main parts of this thesis.

Except for this introductory chapter and the concluding chapter, each of the two parts and all seven chapters are written in an (almost) independent fashion, allowing the reader to jump to any chapter of interest with negligible loss in context.

Estimation: Permutation-based models and algorithms

A central challenge in learning from people is to draw inferences from the highly noisy data from people, for instance, to estimate the correct answers to objective questions from workers' erroneous responses or to estimate the preferences of the population from disparate individual responses. These estimation tasks must be performed in a statistically and computationally efficient manner while making minimal assumptions about how people behave.

In this thesis, we address this estimation challenge for several different settings such as ranking and preference prediction, labeling and classification, and recommender systems and matrix completion. We approach these problems from the lens of statistical learning theory where we assume that the noise is stochastic and is governed by some underlying unknown probabilities. Unsurprisingly, there is a substantial body of past literature for each of these problems [23, 34, 35, 44, 46, 56, 84, 85, 88, 98, 112, 116, 117, 121, 122, 125, 133, 151, 177, 206, 220, 245, 248, 274]. These prior works operate largely under what we call parameter-based modeling assumptions.

Definition 1. Parameter-based models (informal). *Every entity is governed by one (or few) unknown parameters. The probability of any event is a specific, known function of the parameters associated to the entities involved in the event.*

Let us illustrate the concept of parameter-based models with an example. Consider a collection of items such as a set of movies, cars, or sports teams. In many applications, it is required to perform a ranking or estimation task pertaining to these items based on noisy comparisons between various pairs of items. The comparisons are stochastic, that is, for each pair of items (i, j) there is some probability $\mathbb{P}(i \text{ beats } j)$ that item i beats item j in any comparison between them. A “model” in this context is then a set of assumptions on these probabilities. In this setting, a popular class of parameter-based models assumes that every item i is associated to some real-valued parameter, say w_i. The model then assumes that there exists an increasing function, say F, such that the probability of any item i beating any item j is exactly

$$\mathbb{P}(i \text{ beats } j) = F(w_i - w_j).$$

The function F is also assumed to be known. For instance, two popular choices of this function F are the Gaussian CDF which leads to the Thurstone model [253] and the sigmoid function which leads to the Bradley-Terry-Luce model [20, 154].

In various applications, parameter-based models are a popular modeling choice, sometimes because they are more intuitive to write down, and sometimes because they are more amenable for analysis and algorithm design. However, for the applications of our interest, we find such parameter-based models to be quite restrictive. For instance, the parameter-based models assume that the entire behavior of any person or any item is governed by a single (or a few) number. Moreover, the assumption that the data is governed by some specific function of these parameters, and further that this function is known, forms a severely restrictive assumption.

In this thesis, we instead consider what we call permutation-based models that are *strictly and significantly more general* than parameter-based models.

Definition 2. Permutation-based models (informal)**.** *The set of entities in the system has some unknown total ordering and the probabilities of events are monotonic with respect to this ordering.*

For instance, the permutation-based model that we consider for the pairwise-comparison setting assumes that the items have some underlying total ordering, and for any triplet of items (i, j, ℓ) such that i is higher than j in the ordering, it assumes that

$$\mathbb{P}(i \text{ beats } \ell) \geq \mathbb{P}(j \text{ beats } \ell).$$

Observe that the permutation-based model includes all parameter-based models as special cases and is considerably more general. Under this permutation-based model, the items are not governed by any parameters, and the probabilities do not have to obey any restrictive functional form, thereby imparting this model significant generality.

Our permutation-based models are inspired from empirical evidence in psychology and economics (see, for instance, the papers [11, 59, 163, 255]). Experimental results in this line of literature reveal that parameter-based models form a poor fit to the data, whereas assumptions of the permutation-based form are much more representative of the data in the applications considered therein.

In this thesis, we consider several problems pertaining to estimation from data from people. In each of these problems, we design estimation algorithms and perform an associated statistical analysis. Under standard metrics of measuring the performance of estimation algorithms, we establish upper bounds on the error incurred by our estimators under the permutation-based models. We complement these results with matching information-theoretic lower bounds showing that our results are sharp, that is, no other estimation algorithm can perform better (up to logarithmic factors) under the permutation-based models.

With these results one may naturally have the following question: The aforementioned estimation algorithms and bounds address the quite general permutation-based model. Now suppose that the data is guaranteed to be drawn from some parameter-based model. Then if one were to design an estimator tailored specifically to this parameter-based model, then can this estimator exploit the restrictive assumptions of this model to yield a (much) better performance than the estimators based on permutation-based models? We show that, somewhat surprisingly, the answer is *no* – even if an estimator was handcrafted to incorporate the restrictive assumptions of parameter-based models and if the data was always drawn from this model, this estimator cannot perform any better (up to logarithmic factors) than the much more generally applicable permutation-based estimators. This phenomenon is a recurring theme of Part I of this thesis.

Permutation-based models enjoy a win-win in terms of the statistical bias-variance tradeoff as compared to parameter-based models.

Please see Figure 1.1 for a pictorial illustration.

We now describe the contributions of Part I of this thesis in some more detail.

In Chapter 2 we consider data in the form of pairwise comparisons – there is a collection of items and we have data of the form "item i is better than item j" for various pairs (i, j) of these items. This form of data is motivated by the well known [13, 247] phenomenon that for human beings, choosing one of two options requires far lesser time and effort as compared to giving a cardinal score (numeric rating) for each item. Moreover empirical evidence [220, Section 5] also suggests pairwise-comparison data is typically subject to a lower noise than cardinal scores. In this chapter, we address the problem of estimating the underlying probabilities of the pairwise comparisons – for any given pair of items (i, j), what is the probability that item i will beat item j if they are compared? In contrast to prior works [23, 46, 98, 112, 121, 177, 220] that consider restrictive parameter-based models for analyzing pairwise-comparison data, we consider a significantly more general permutation-based model which is also called the strong stochastic transitivity model in the literature. We provide various algorithms for estimation under this model, establish associated statistical guarantees, and show that making parameter-based assumptions offer little help. We also show that our assumed model is a remarkable sweet spot: On one hand, our aforementioned results imply that moving to more restrictive models do not help, and on the other hand we also show that more general models are not very useful as they result in an estimation error that is almost as high as when making no assumptions at all. This chapter is based on joint works with S. Balakrishnan, M. J. Wainwright, and A. Guntuboyina [221, 223].

In Chapter 3 we continue with the setting of pairwise comparisons, with the goal of identifying the top k items for some value of k, or alternatively, recovering a ranking of all

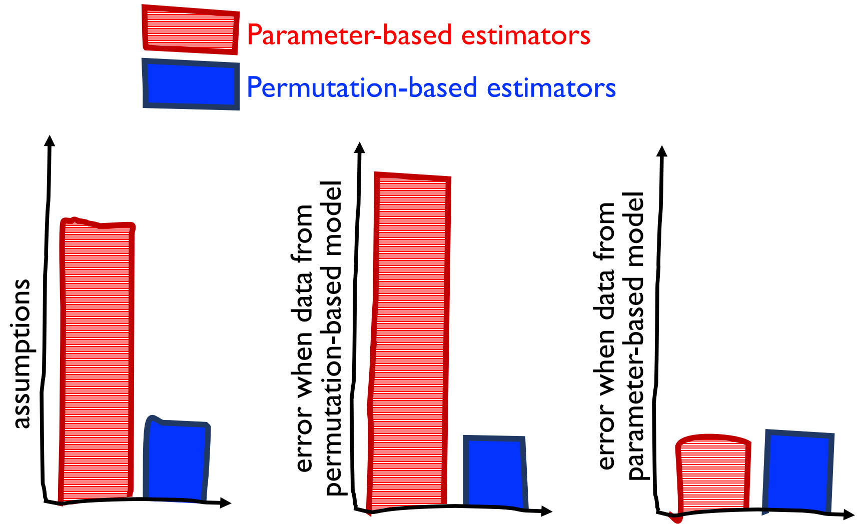


Figure 1.1: A comparison of the performance of permutation-based estimators with parameter-based estimators: (a) permutation-based estimators rely on strictly and significantly fewer assumptions; (b) when data follows a permutation-based model, the error incurred by permutation-based estimators is an order of magnitude lower; and (c) when data follows a parameter-based model, the error incurred by permutation-based estimators is no more (up to logarithmic factors) than the error incurred by the respective parameter-based estimators.

the items. We consider requirements of both exact and approximate recovery. For the latter requirement, we propose an abstract class of approximation metrics that is based on a simple and natural motonicity condition and encapsulates many popular approximation metrics such as the Hamming error. We analyze a simple counting algorithm that ranks the items in order of the number of pairwise comparisons won, and show it has three attractive features: (i) its computational efficiency leads to speed-ups of several orders of magnitude in computation time as compared to prior work; (ii) it is robust in that theoretical guarantees impose very mild (permutation-based) conditions on the underlying pairwise-comparison probabilities; and (iii) it is an optimal method up to constant factors for all of the aforementioned metrics. In contrast, prior works on this topic such as [46, 248] address only a few specific metrics of recovery, and furthermore, restrict attention to the parameter-based BTL model. This chapter is based on joint work with M. J. Wainwright [234].

In Chapter 4, we move to the problem of labeling (or classification) from the crowd. The aggregation and denoising of crowd labeled data is a task that has gained increased significance with the advent of crowdsourcing platforms and massive datasets. Here one has access to the (noisy) answers of multiple workers to several binary-choice questions and the goal is to estimate the true answers to each of the questions. In this chapter, we propose a permutation-based model for crowd labeled data that is a significant generalization of the parameter-based models of the Dawid-Skene [60] type, which are the focus of prior works [56, 84, 85, 88, 116, 117, 125, 151, 274]. We also introduce a new error metric by which to compare different estimators which is more suited for this problem as compared to the Hamming error metric considered in all prior works. We design estimation algorithms for this problem and show several associated strong statistical guarantees with respect to its performance on both the permutation-based and the parameter-based models. This chapter is based on joint work with S. Balakrishnan and M. J. Wainwright [222].

In Chapter 5, we consider the problem of noisy matrix completion, in which the goal is to reconstruct a structured matrix whose entries are partially observed in noise. Standard approaches to this problem are based on assuming that the underlying matrix has low rank, or is well-approximated by a low-rank matrix. In this chapter, we advocate a rethinking of this low-rank assumption, and propose a richer model based on what we term the "permutation-rank" of a matrix. We describe how the classical non-negative rank model can be seen to enforce restrictive and often undesirable parameter-based assumptions, and the richer permutation-rank model avoids these strong assumptions. We again present estimation algorithms and various associated statistical guarantees. These guarantees also include sharp oracle inequalities – bounds on the error incurred by our estimator when data does not follow the assumed models – that are also applicable to the problem settings of earlier chapters. We also provide various structural results characterizing the uniqueness of the permutation-rank decomposition, and characterizing convex approximations of the permutation-rank polytope. This chapter is based on joint work with S. Balakrishnan and M. J. Wainwright [224].

We also mention in passing some other works on closely related topics that the author contributed to during his PhD, but are excluded from this thesis for the purposes of brevity and cohesiveness. We briefly outline these works here and provide relevant references for the interested reader:

- Cardinal data is noisier in practice as compared to ordinal data [220, Section 5].
- Ranking from active pairwise comparisons: an efficient algorithm, sharp guarantees under a general permutation-based model, and futility of parameter-based models [100].
- A case for ordinal (comparison-based) peer-grading in massive open online courses (MOOCs) and an aggregation algorithm [225].
- Estimation from comparisons between arbitrary pairs of items, and error-bounds based on on this comparison graph for parameter-based models [220].
- Adaptivity to local underlying smoothness when estimating pairwise comparison matrices: sharp statistical and computational guarantees, and a surprising negative result about the popular least squares estimator [223].
- Impossibility of reasonable models for crowdsourced-labeling that guarantee convex estimation [236].
- An algorithm based on a regularized entropy approach for aggregating crowdsourced labels and associated empirical evaluations [275].
- An analysis of the review data from the Neural Information Processing Systems (NIPS) 2016 conference [218].

Incentives: Multiplicative payment mechanisms

The second fundamental challenge in learning from people is to collect higher-quality data by incentivizing people to respond in a suitable manner. We consider incentives in the form of performance-based monetary payments, which are applicable to the extensively used crowdsoucing platforms such as Amazon Mechanical Turk (`mturk.com`) and others (`crowdflower.com`,`microtask.com`, `upwork.com`). Alongside, we also consider the naturally associated requirement of providing appropriate interfaces to people so that they can express their knowledge in the best possible manner.

We consider a crowdsourcing setting where the requester possesses a collection of objective questions and wishes to elicit the answers to these questions from the crowd. We further focus on a standard setting in crowdsourcing where the requester has access to the correct answers to a (small) subset of the questions; these questions are called *gold standard* questions and are used to evaluate the workers' performance and to make payments to them. The gold standard questions are mixed uniformly at random among the actual questions, and the worker does not know which of the questions are gold standard. On receiving any worker's responses to all the questions, for the purpose of determining the payment to the worker, we retain only the responses to the gold standard questions.

Any payment mechanism employed must ensure that workers are incentivized to report honestly, and that the system cannot be gamed. In order to formalize this requirement, we employ the standard notion of incentive compatibility from game theory and decision theory.

Definition 3. Incentive compatibility (informal)**.** *The expected payment, from the worker's point of view, must be strictly maximized when the worker responds honestly.*

As a toy example to understand the concept of incentive compatibility, consider a binary-choice question with options A and B to which the requester knows the answer. The requester wishes to ask this question to a worker on a crowdsourcing platform and incentivize the worker to provide the answer that the worker believes is most likely to be correct. Now suppose that the payment scheme is set as follows: pay an amount $x > 0$ when the worker's answer is correct and pay an amount y ($x > y \geq 0$) when it is incorrect. The payment scheme is made known to the worker. Now suppose that the worker believes that option A is the correct answer with probability p_A and that option B is the correct answer with probability $p_B = 1 - p_A$. Then from the worker's perspective, choosing option A will yield a payment of x if A turns out to be correct (which happens with probability p_A from the worker's perspective) and a payment of y if B is correct (with probability p_B), thereby yielding a payment of $(xp_A + yp_B)$ in expectation. Similarly, choosing option B will yield an expected payment of $(yp_A + xp_B)$. Under the assumption that the worker wishes to maximize his/her expected payment, one can verify that the worker is incentivized to report option A if $p_A > p_B$ and option B if $p_B > p_A$. The mechanism is thus incentive compatible.

The goal is to design payment mechanisms for crowdsourcing that are incentive compatible. To this end, the framework of "strictly proper scoring rules" [24, 94, 216] provides a general theory for eliciting information for settings where the responses can subsequently be

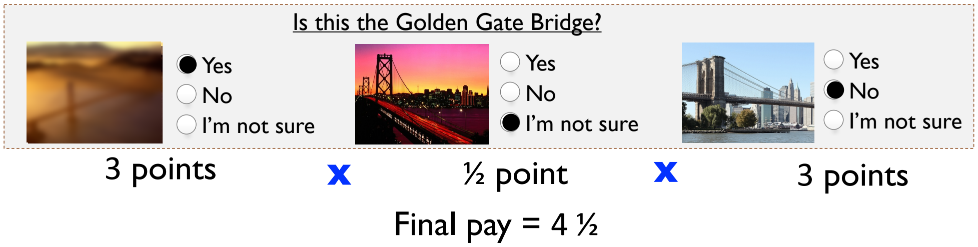


Figure 1.2: An illustration of a multiplicative payment mechanism. The three questions shown in the figure are the gold standard questions. The depicted payment is the multiplicative part which depends on the worker's answers, and the fixed part of the payment is set as zero in this example.

verified by the requester. (Consequently, our mechanisms can also be called strictly proper scoring rules for crowdsourcing.) Importantly, the framework of strictly proper scoring rules, however, provides a large collection of possible mechanisms and *does not* guide the choice of a specific mechanism from this collection [94].

Our approach towards this problem is to design mechanisms that can be used in practical crowdsourcing applications, and hence we must somehow choose one mechanism for deployment. As a first step in this direction, in order to narrow down the possible choices for mechanisms in a principled manner, we begin by imposing a very simple and natural requirement that we call no-free-lunch.

Definition 4. No-free-lunch (informal)**.** *The payment should be minimum if all attempted (gold standard) questions are answered incorrectly.*

Observe that no-free-lunch is a very weak requirement. For instance, consider a task where each question is of a binary-choice format. If a worker chooses answers uniformly at random for every question, at least one answer will be correct with high probability. On the other hand, the no-free-lunch requirement is invoked only when all attempted questions are answered incorrectly, and hence there is an exponentially small (in the number of gold standard questions) probability that the no-free-lunch requirement will be invoked.

In this thesis, we design mechanisms for various settings that satisfy the two natural requirements of incentive compatibility and no-free-lunch. Interestingly, a common theme across all our mechanisms is that they have a multiplicative form.

Definition 5. Multiplicative payment mechanism (informal)**.** *The payment mechanism gives a certain number of points separately for each (gold standard) question depending on the worker's answer, and the final payment is a product of all of these points. This answer-dependent payment may be augmented by adding a fixed (independent of worker's answers) payment to the computed product.*

An example of such a multiplicative payment mechanism is illustrated in Figure 1.2.

We now return to our goal of choosing a mechanism using a principled approach, where we had imposed the no-free-lunch requirement to reduce the number of possible mechanisms from a massively large class to a possibly somewhat smaller set. It turns out that our no-free-lunch requirement provides remarkable assistance on that front – we show that, surprisingly, there is no other incentive-compatible mechanism that satisfies no-free-lunch.

> Multiplicative incentive mechanisms are the only mechanisms that are incentive compatible and satisfy no-free-lunch.

We also show that multiplicative mechanisms have additional appealing properties in the context of crowdsourcing. The simplicity of our mechanisms is an added benefit ithat imparts a significant practical appeal.

With this background, we now discuss the individual chapters of Part II of this thesis in more detail.

In Chapter 6, we consider a crowdsourced data-collection setting with a goal of incentivizing workers to answer only the questions that they are sure about, and skip the questions for which the worker is not confident enough. An interface of this form is illustrated in Figure 1.2 (note that our work is not restricted to only multiple choice questions). We design multiplicative incentive mechanisms for this setting and show that they are are the only incentive compatible mechanisms to satisfy the no-free-lunch requirement. We also show that among all incentive compatible mechanisms that may or may not satisfy no-free-lunch, our mechanisms make the strictly smallest expected payment when a worker answers randomly. This chapter is based on joint work with D. Zhou [235].

In Chapter 7, we consider an interface that additionally elicits a quantized confidence of the worker. For example, for each attempted question the worker may be asked to indicate whether he/she has a “low”, “moderate”, or “high” confidence. We consider an even weaker notion of no-free-lunch that applies only when all attempted questions (in the gold standard) are answered incorrectly and are indicated as the highest confidence level by the worker. We then design a multiplicative payment mechanism and prove its uniqueness. This chapter is based on joint work with D. Zhou [235].

In Chapter 8 we consider an “approval voting” interface [21, 51, 185] in which the worker is allowed to select any number of options that he/she thinks could possibly be the correct answer. We first show an impossibility result that no mechanism can be incentive compatible in this setting. Then under an additional assumption on the granularity of peoples’ responses, we design a multiplicative mechanism for which we prove strong guarantees of uniqueness and optimality. This chapter is based on joint work with D. Zhou and Y. Peres [237].

Finally, we also mention in passing some other closely related works by the author during his PhD that are omitted from this thesis. We also provide relevant references for the interested reader.

- A new “self-correction” interface for crowdsourcing and associated incentive mechanisms [219].
- Mechanisms that operate without gold standard questions, and are simpler than all incentive-compatible mechanisms in past literature [115].
- Mechanisms for parametric prediction from parametric agents [156].

Notation

We now describe some notation that we employ throughout this thesis.

We use the notation c along with any subscripts or superscripts, such as c_1, c' etc., to denote positive universal constants.

We use $\mathbb{R}$ to denote the set of all real numbers, and $\mathbb{R}_+$ to denote all non-negative real numbers. For any vector or matrix, we use the superscript ^{T} to denote its transpose. For any positive integer k, we use the notation $[k]$ to represent the set $\{1, \ldots, k\}$. The indicator function is denoted by $\mathbf{1}$, that is, $\mathbf{1}\{z\} = 1$ if z is true, and 0 otherwise.

For any vector $v \in \mathbb{R}^n$ and any value $p \geq 1$, we use $\|v\|_p$ to denote the ℓ_p-norm of vector v, that is, $\|v\|_p = (\sum_{i=1}^n |v_i|^p)^{1/p}$. We also use $\|v\|_\infty = \max_{i=1}^n |v_i|$. For any square matrix $M \in \mathbb{R}^{n \times n}$, we use the notation $\mathrm{trace}(M)$ to denote the trace of matrix M, that is, $\mathrm{trace}(M) = \sum_{i=1}^n M_{ii}$. For any pair of matrices $M_1 \in \mathbb{R}^{n \times d_1}$ and $M_2 \in \mathbb{R}^{n \times d_2}$, we use the notation $\langle\!\langle M_1,\ M_2 \rangle\!\rangle$ to denote the trace inner product between the two matrices, that is, $\langle\!\langle M_1,\ M_2 \rangle\!\rangle = \mathrm{trace}(M_1^T M_2)$. For any matrix $M \in \mathbb{R}^{n \times d}$, we use $|\!|\!| M |\!|\!|_{\mathrm{F}}$ to denote its Frobenius norm and $|\!|\!| M |\!|\!|_{\mathrm{op}}$ to denote its ℓ_2 operator norm, that is, $|\!|\!| M |\!|\!|_{\mathrm{F}} = \sqrt{\langle\!\langle M,\ M \rangle\!\rangle} = \sqrt{\sum_{i=1}^n \sum_{j=1}^d M_{ij}^2}$ and $|\!|\!| M |\!|\!|_{\mathrm{op}} = \sup_{v \in \mathbb{R}^d \setminus \{0\}} \|Mv\|_2 / \|v\|_2$. For any matrix $M \in \mathbb{R}^{n \times d}$, we let $\sigma_1(M), \ldots, \sigma_{\max\{n,d\}}(M)$ denote its singular values (ordered from largest to smallest); if the rank of M equals r, then we must have $\sigma_{r+1}(M) = \cdots = \sigma_{\max\{n,d\}}(M) = 0$. For any symmetric matrix $M \in \mathbb{R}^{n \times n}$, we let $\lambda_1(M), \ldots, \lambda_n(M)$ denote its ordered eigenvalues.

We use the standard Landau order notation for asymptotics: we write $a_n = \mathcal{O}(b_n)$ to mean that there are universal constants $C \geq 0$ and $N \geq 1$ such that $a_n \leq C b_n$ for every $n \geq N$. Similarly, we write $a_n = \Omega(b_n)$ to mean that $a_n \geq c b_n$ for every $n \geq N$, where $c > 0$ and $N \geq 1$ are some universal constants. We write $a_n = \Theta(b_n)$ when $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. We write $a_n = o(b_n)$ to mean that for every $c > 0$ there exists some $N \geq 1$ (which may depend on c) such that $a_n \leq c b_n$ for all $n \geq N$. We augment the Landau order notation with a tilde to mean the corresponding notation up to logarithmic factors.

Part I

Statistical Learning: Permutation-based Models and Estimation Algorithms

Chapter 2

Estimating Pairwise Comparison Probabilities

"Compare yourself with the best not to feel jealous but to strive to become at least as good."

– Albert Einstein

2.1 Introduction

Pairwise comparison data is ubiquitous and arises naturally in a variety of applications, including tournament play, voting, online search rankings, and advertisement placement problems. In rough terms, given a set of n objects along with a collection of possibly inconsistent comparisons between pairs of these objects, the goal is to aggregate these comparisons in order to estimate underlying properties of the population. One property of interest is the underlying matrix of pairwise comparison probabilities—that is, the matrix in which entry (i, j) corresponds to the probability that object i is preferred to object j in a pairwise comparison. The Bradley-Terry-Luce [20, 154] and Thurstone [253] models are mainstays in analyzing this type of pairwise comparison data. These models are parameter-based in nature: more specifically, they assume the existence of an n-dimensional weight vector that measures the quality or strength of each item. The pairwise comparison probabilities are then determined via some fixed function of the qualities of the pair of objects. Estimation in these models reduces to estimating the underlying weight vector, and a large body of prior work has focused on these models (e.g., see the papers [98, 177, 220]). However, such models enforce strong relationships on the pairwise comparison probabilities that often fail to hold in real applications. Various papers [11, 59, 163, 255] have provided experimental results in which these parameter-based modeling assumptions fail to hold.

Our focus in this chapter is on models that have their roots in social science and psychology (e.g., see Fishburn [78] for an overview), in which the only coherence assumption imposed on the pairwise comparison probabilities is that of a permutation-based form and is known as *strong stochastic transitivity*, or SST for short. These models include the parameter-based models as special cases but are considerably more general. The permutation-based SST model has been validated by several empirical analyses, including those in a long line of work [11, 59, 163, 255]. The conclusion of Ballinger et al. [11] is especially strongly worded:

> *All of these parametric c.d.f.s are soundly rejected by our data. However, SST usually survives scrutiny.*

We are thus provided with strong empirical motivation for studying the fundamental properties of pairwise comparison probabilities satisfying the permutation-based SST assumptions.

In this chapter, we focus on the problem of estimating the matrix of pairwise comparison probabilities—that is, the probability that an item i will beat a second item j in any given comparison. Estimates of these comparison probabilities are useful in various applications. For instance, when the items correspond to players or teams in a sport, the predicted odds of one team beating the other are central to betting and bookmaking operations. In a supermarket or an ad display, an accurate estimate of the probability of a customer preferring one item over another, along with the respective profits for each item, can effectively guide the choice of which product to display. Accurate estimates of the pairwise comparison probabilities can also be used to infer partial or full rankings of the underlying items.

Our contributions: We begin by studying the performance of optimal methods for estimating matrices in the permutation-based SST class: our first main result (Theorem 1) characterizes the minimax rate in squared Frobenius norm up to logarithmic factors. This result reveals that even though the permutation-based SST class of matrices is considerably larger than the classical parameter-based class, surprisingly, it is possible to estimate any permutation-based SST matrix at nearly the same rate as the classical parameter-based family. On the other hand, our achievability result is based on an estimator involving prohibitive computation, as a brute force approach entails an exhaustive search over permutations. Accordingly, we turn to studying computationally tractable estimation procedures. Our second main result (Theorem 2) applies to a polynomial-time estimator based on soft-thresholding the singular values of the data matrix. An estimator based on hard-thresholding was studied previously in this context by Chatterjee [44]. We sharpen and generalize this previous analysis, and give a tight characterization of the rate achieved by both hard and soft-thresholding estimators. Our third contribution is a polynomial-time computable estimator which we term the CRL estimator that we show is consistent and is guaranteed to output a matrix in the permutation-based SST class, and is also optimal over the parameter-based class (Theorem 3). Our fourth contribution, formalized in Theorems 4 and 5, is to show how for certain interesting subsets of the full permutation-based SST class, a combination of parameter-based maximum likelihood [220] and noisy sorting algorithms [23] leads to a tractable two-stage method that achieves the minimax rate. Our fifth contribution is to supplement our minimax lower bound with lower bounds for various known estimators, including those based on thresholding singular values [44], noisy sorting [23], as well as parameter-based estimators [98, 177, 220]. These lower bounds show that none of these tractable estimators achieve the minimax rate uniformly over the entire class. The lower bounds also show that the minimax rates for any of these subclasses is no better than the full permutation-based SST class (up to logarithmic factors). Finally we show (Proposition 2) that the permutation-based SST class is a sweet spot: in addition to our aforementioned results showing that restricting to smaller parameter-based classes does not help, we show that richer classes that are studied in psychology and economics are so general that they prohibit any meaningful estimation.

Related work: The literature on ranking and estimation from pairwise comparisons is vast and we refer the reader to various surveys [41, 79, 158] for a more detailed overview. Here we focus our literature review on those papers that are most closely related to our contributions. Some recent work [98, 177, 220] studies procedures and minimax rates for estimating the latent quality vector that underlie parameter-based models. Theorem 5 in this chapter provides an extension of these results, in particular by showing that an optimal estimate of the latent quality vector can be used to construct an optimal estimate of the pairwise comparison probabilities. Chatterjee [44] analyzed matrix estimation based on singular value thresholding, and obtained results for the class of permutation-based SST matrices. In Theorem 2, we provide a sharper analysis of this estimator, and show that our upper bound is—in fact—unimprovable.

In past work, various authors [23, 121] have considered the noisy sorting problem, in which the goal is to infer the underlying order under a so-called high signal-to-noise ratio (SNR) condition. The high SNR condition means that each pairwise comparison has a probability of agreeing with the underlying order that is bounded away from $\frac{1}{2}$ by a fixed constant. Under this high SNR condition, these authors provide polynomial-time algorithms that, with high probability, return an estimate of true underlying order with a prescribed accuracy. Part of our analysis leverages an algorithm from the paper [23]; in particular, we extend their analysis in order to provide guarantees for estimating pairwise comparison probabilities as opposed to estimating the underlying order.

As will be clarified in the sequel, the assumption of strong stochastic transitivity has close connections to the statistical literature on shape constrained inference (e.g., [241]), particularly to the problem of bivariate isotonic regression. In our analysis of the least-squares estimator, we leverage metric entropy bounds from past work in this area (e.g., [42, 86]).

In Section 2.3.7, we study estimation under two popular models that are closely related to the permutation-based SST class, and make even weaker assumptions. We show that under these moderate stochastic transitivity (MST) and weak stochastic transitivity (WST) models, the Frobenius norm error of any estimator, measured in a uniform sense over the class, must be almost as bad as that incurred by making no assumptions whatsoever. Consequently, from a statistical point of view, these assumptions are not strong enough to yield reductions in estimation error.

Organization: The remainder of the chapter is organized as follows. We begin by providing a background and a formal description of the problem in Section 2.2. In Section 2.3, we present the main theoretical results of the chapter. We then present results from numerical simulations in Section 2.4. We present a concluding discussion in Section 2.5. Finally, we present proofs of our main results in Section 2.6.

2.2 Problem setting

Given a collection of $n \geq 2$ items, suppose that we have access to noisy comparisons between any pair $i \neq j$ of distinct items. The full set of all possible pairwise comparisons can be described by a probability matrix $M^* \in [0,1]^{n \times n}$, in which M^*_{ij} is the probability that item i is preferred to item j. The upper and lower halves of the probability matrix M^* are related by the *shifted-skew-symmetry condition*[1] $M^*_{ji} = 1 - M^*_{ij}$ for all $i, j \in [n]$. For concreteness, we set $M^*_{ii} = 1/2$ for all $i \in [n]$.

2.2.1 Estimation of pairwise comparison probabilities

For any matrix $M^* \in [0,1]^{n \times n}$ with $M^*_{ij} = 1 - M^*_{ji}$ for every (i,j), suppose that we observe a random matrix $Y \in \{0,1\}^{n \times n}$ with (upper-triangular) independent Bernoulli entries, in particular, with $\mathbb{P}[Y_{ij} = 1] = M^*_{ij}$ for every $1 \leq i \leq j \leq n$ and $Y_{ji} = 1 - Y_{ij}$. Based on observing Y, our goal in this chapter is to recover an accurate estimate, in the squared Frobenius norm, of the full matrix M^*.

Our primary focus in this chapter will be on the setting where for n items we observe the outcome of a single pairwise comparison for each pair. We will subsequently (in Section 2.3.6) also address the more general case when we have partial observations, that is, when each pairwise comparison is observed with a fixed probability.

For future reference, note that we can always write the Bernoulli observation model in the linear form

$$Y = M^* + W, \tag{2.1}$$

where $W \in [-1,1]^{n \times n}$ is a random matrix with independent zero-mean entries for every $i \geq j$ given by

$$W_{ij} \sim \begin{cases} 1 - M^*_{ij} & \text{with probability } M^*_{ij} \\ -M^*_{ij} & \text{with probability } 1 - M^*_{ij}, \end{cases} \tag{2.2}$$

and $W_{ji} = -W_{ij}$ for every $i < j$. This linearized form of the observation model is convenient for subsequent analysis.

[1] In other words, the shifted matrix $M^* - \frac{1}{2}$ is skew-symmetric.

2.2.2 Strong stochastic transitivity

Beyond the previously mentioned constraints on the matrix M^*—namely that $M^*_{ij} \in [0,1]$ and $M^*_{ij} = 1 - M^*_{ij}$—more structured and interesting models are obtained by imposing further restrictions on the entries of M^*. We now turn to one such condition, known as *strong stochastic transitivity* (SST), which is a permutation-based model that reflects the natural transitivity of any complete ordering. Formally, suppose that the full collection of items $[n]$ is endowed with a complete ordering π^*. We use the notation $\pi^*(i) < \pi^*(j)$ to convey that item i is preferred to item j in the total ordering π^*. Consider some triple (i,j,k) such that $\pi^*(i) < \pi^*(j)$. A matrix M^* satisfies the permutation-based SST condition if the inequality $M^*_{ik} \geq M^*_{jk}$ holds for every such triple. The intuition underlying this constraint is the following: since i dominates j in the true underlying order, when we make noisy comparisons, the probability that i is preferred to k should be at least as large as the probability that j is preferred to k. The SST condition was first described[2] in the psychology literature (e.g., [59, 78]).

The permutation-based SST condition is characterized by the existence of a permutation such that the permuted matrix has entries that increase across rows and decrease down columns. More precisely, for a given permutation π^*, let us say that a matrix M is π^*-faithful if for every pair (i,j) such that $\pi^*(i) < \pi^*(j)$, we have $M_{ik} \geq M_{jk}$ for all $k \in [n]$. With this notion, the set of permutation-based SST matrices is given by

$$\mathbb{C}_{\text{SST}} = \big\{ M \in [0,1]^{n \times n} \;\mid\; M_{ba} = 1 - M_{ab} \;\forall\, (a,b), \text{ and } \exists \text{ perm. } \pi^* \text{ s.t. } M \text{ is } \pi^*\text{-faithful} \big\}. \tag{2.3}$$

Note that the stated inequalities also guarantee that for any pair with $\pi^*(i) < \pi^*(j)$, we have $M_{ki} \leq M_{kj}$ for all k, which corresponds to a form of column ordering. The class $\mathbb{C}_{\text{SST}}$ is our primary focus in this chapter.

2.2.3 Classical parameter-based models

Let us now describe a family of classical parameter-based models, one which includes Bradley-Terry-Luce and Thurstone (Case V) models [20, 154, 253]. In the sequel, we show that these parameter-based models induce a relatively small subclass of the permutation-based SST matrices $\mathbb{C}_{\text{SST}}$.

In more detail, parameter-based models are described by a weight vector $w^* \in \mathbb{R}^n$ that corresponds to the notional qualities of the n items. Moreover, consider any non-decreasing function $F : \mathbb{R} \mapsto [0,1]$ such that $F(t) = 1 - F(-t)$ for all $t \in \mathbb{R}$; we refer to any such function F as being *valid*. Any such pair (F, w^*) induces a particular pairwise comparison model in which

$$M^*_{ij} = F(w^*_i - w^*_j) \qquad \text{for all pairs } (i,j). \tag{2.4}$$

[2] We note that the psychology literature has also considered what are known as weak and moderate stochastic transitivity conditions. From a statistical standpoint, pairwise comparison probabilities cannot be consistently estimated in a minimax sense under these conditions. We establish this formally in Section 2.3.7.

For each valid choice of F, we define

$$\mathbb{C}_{\text{PAR}}(F) = \Big\{ M \in [0,1]^{n \times n} \;\mid M \text{ induced by Equation (2.4) for some } w^* \in \mathbb{R}^n \Big\}. \tag{2.5a}$$

For any choice of F, it is straightforward to verify that $\mathbb{C}_{\text{PAR}}(F)$ is a subset of $\mathbb{C}_{\text{SST}}$, meaning that any matrix M induced by the relation (2.4) satisfies all the constraints defining the set $\mathbb{C}_{\text{SST}}$. As particular important examples, we recover the Thurstone parameter-based model by setting $F(t) = \Phi(t)$ where Φ is the Gaussian CDF, and the Bradley-Terry-Luce model by setting $F(t) = \frac{e^t}{1+e^t}$, corresponding to the sigmoid function.

Remark: Since the pairwise probabilities depend only on the differences $w_i^* - w_j^*$, we can assume without loss of generality that $\langle w^*, 1 \rangle = 0$. Moreover, since the choice of F can include rescaling its argument, we can also assume that $\|w^*\|_\infty \leq 1$. Accordingly, we assume in our subsequent analysis that w^* belongs to the set

$$M^* \in \big\{ w \in \mathbb{R}^n \;\mid\; \text{such that } \langle w, 1 \rangle = 0 \text{ and } \|w\|_\infty \leq 1 \big\}. \tag{2.5b}$$

2.2.4 Inadequacies of parameter-based models

As noted in the introduction, a large body of past work (e.g., [11, 59, 163, 255]) has shown that parameter-based models, of the form (2.5a) for some choice of F, often provide poor fits to real-world data. What might be a reason for this phenomenon? Roughly, parameter-based models impose the very restrictive assumption that the choice of an item depends on the value of a single latent factor (as indexed by w^*)—e.g., that our preference for cars depends only on their fuel economy, or that the probability that one hockey team beats another depends only on the skills of the goalkeepers.

This intuition can be formalized to construct matrices $M^* \in \mathbb{C}_{\text{SST}}$ that are far away from *every valid parameter-based approximation* as summarized in the following result:

Proposition 1. *There exists a universal constant $c > 0$ such that for every $n \geq 4$, there exist matrices $M^* \in \mathbb{C}_{SST}$ for which*

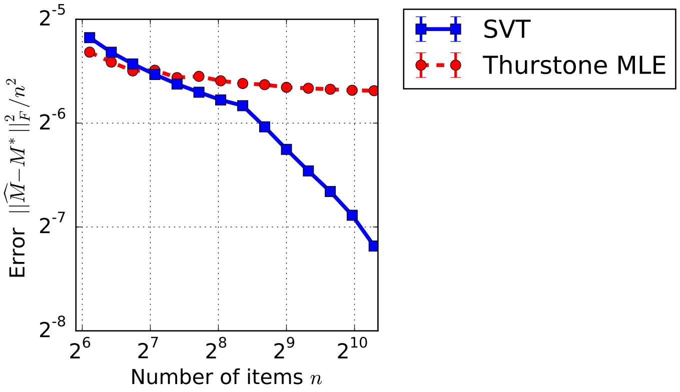
$$\frac{1}{n^2} \inf_{\text{valid } F} \inf_{M \in \mathbb{C}_{PAR}(F)} |\!|\!| M - M^* |\!|\!|_F^2 \geq c. \tag{2.6}$$

Given that every entry of matrices in $\mathbb{C}_{\text{SST}}$ lies in the interval $[0, 1]$, the Frobenius norm diameter of the class $\mathbb{C}_{\text{SST}}$ is at most n^2, so that the scaling of the lower bound (2.6) cannot be sharpened.

What sort of matrices M^* are "bad" in the sense of satisfying a lower bound of the form (2.6)? Panel (a) of Figure 2.1 describes the construction of one "bad" matrix M^*. In order to provide some intuition, let us return to the analogy of rating cars. A key property of any parameter-based model is that if we prefer car 1 to car 2 more than we prefer car 3

$$M^* := \frac{1}{8}\begin{bmatrix} 4 & 6 & 7 & 8 \\ 2 & 4 & 7 & 8 \\ 1 & 1 & 4 & 5 \\ 0 & 0 & 3 & 4 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

(a)



(b)

Figure 2.1: (a) Construction of a "bad" matrix: for n divisible by 4, form the matrix $M^* \in \mathbb{R}^{n\times n}$ shown, where each block has dimensions $n/4 \times n/4$. (b) Estimation for a class of permutation-based SST matrices that are far from the parameter-based models. The parameter-based model (Thurstone MLE) yields a poor fit, whereas fitting using the singular value thresholding (SVT) estimator, which allows for estimates over the full permutation-based SST class, leads to consistent estimation.

to car 4, then we must also prefer car 1 to car 3 more than we prefer car 2 to car 4.[3] This condition is potentially satisfied if there is a single determining factor across all cars—for instance, their fuel economy.

This ordering condition is, however, violated by the pairwise comparison matrix M^* from Figure 2.1(a). In this example, we have $M^*_{12} = \frac{6}{8} > \frac{5}{8} = M^*_{34}$ and $M^*_{13} = \frac{7}{8} < \frac{8}{8} = M^*_{24}$. Such an occurrence can be explained by a simple two-factor model: suppose the fuel economies of cars $1, 2, 3$ and 4 are 20, 18, 12 and 6 kilometers per liter respectively, and the comfort levels of the four cars are also ordered $1 \succ 2 \succ 3 \succ 4$, with $i \succ j$ meaning that i is more comfortable than j. Suppose that in a pairwise comparison of two cars, if one car is more fuel efficient by at least 10 kilometers per liter, it is always chosen. Otherwise the choice is governed by a parameter-based choice model acting on the respective comfort levels of the two cars. Observe that while the comparisons between the pairs $(1, 2)$, $(3, 4)$ and $(1, 3)$ of cars can be explained by this parameter-based model acting on their respective comfort levels, the preference between cars 1 and 4, as well as between cars 2 and 4, is governed by their fuel economies. This two-factor model accounts for the said values of M^*_{12}, M^*_{34}, M^*_{24} and M^*_{13}, which violate parameter-based requirements.

While this was a simple hypothetical example, there is a more ubiquitous phenomenon underlying our example. It is often the case that our preferences are decided by comparing items on a multitude of dimensions. In any situation where a single (latent) parameter per

[3] This condition follows from the proof of Proposition 1.

item does not adequately explain our preferences, one can expect that the class of parameter-based models to provide a poor fit to the pairwise preference probabilities.

The lower bound on approximation quality guaranteed by Proposition 1 means that any parameter-based estimator of the matrix M^* should perform poorly. This expectation is confirmed by the simulation results in panel (b) of Figure 2.1. After generating observations from a "bad matrix" over a range of n, we fit the data set using either the maximum likelihood estimate in the Thurstone parameter-based model, or the singular value thresholding (SVT) estimator, to be discussed in Section 2.3.2. For each estimator $\widehat{M}$, we plot the rescaled Frobenius norm error $\frac{|\!|\!|\widehat{M}-M^*|\!|\!|_{\mathrm{F}}^2}{n^2}$ versus the sample size. Consistent with the lower bound (2.6), the error in the Thurstone-based estimator stays bounded above a universal constant. In contrast, the SVT error goes to zero with n, and as our theory in the sequel shows, the rate at which the error decays is at least as fast as $1/\sqrt{n}$.

2.3 Main results

Thus far, we have introduced two classes of models for matrices of pairwise comparison probabilities. Our main results characterize the rates of estimation associated with different subsets of these classes, using either optimal estimators (that we suspect are not polynomial-time computable in certain cases), or more computationally efficient estimators that can be computed in polynomial-time.

2.3.1 Characterization of the minimax risk

We begin by providing a result that characterizes the minimax risk in squared Frobenius norm over the class $\mathbb{C}_{\text{SST}}$ of permutation-based SST matrices. The minimax risk is defined by taking an infimum over the set of all possible estimators, which are measurable functions $Y \mapsto \widetilde{M} \in [0,1]^{n\times n}$. Here the data matrix $Y \in \{0,1\}^{n\times n}$ is drawn from the observation model (2.1).

Theorem 1. *There are universal constants* $0 < c_2 < c_1$ *such that*

$$\frac{c_2}{n} \;\leq\; \inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2}\mathbb{E}[|\!|\!|\widetilde{M} - M^*|\!|\!|_F^2] \;\leq\; c_1 \frac{\log^2(n)}{n}, \tag{2.7}$$

where the infimum ranges over all measurable functions $\widetilde{M}$ *of the observed matrix* Y.

We prove this theorem in Section 2.6.2. The proof of the lower bound is based on extracting a particular subset of the class $\mathbb{C}_{\text{SST}}$ such that any matrix in this subset has at least n positions that are unconstrained, apart from having to belong to the interval $[\frac{1}{2}, 1]$. We can thus conclude that estimation of the full matrix is at least as hard as estimating n Bernoulli parameters belonging to the interval $[\frac{1}{2}, 1]$ based on a single observation per number. This reduction leads to an $\Omega(n^{-1})$ lower bound, as stated.

Proving an upper bound requires substantially more effort. In particular, we establish it via careful analysis of the constrained least-squares estimator

$$\widehat{M} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} |||Y - M|||_{\text{F}}^2. \tag{2.8a}$$

In particular, we prove that there are universal constants (c_0, c_1, c_2) such that, for any matrix $M^* \in \mathbb{C}_{\text{SST}}$, this estimator satisfies the high probability bound

$$\mathbb{P}\Big[\frac{1}{n^2} |||\widehat{M} - M^*|||_{\text{F}}^2 \geq c_0 \frac{\log^2(n)}{n}\Big] \leq c_1 e^{-c_2 n}. \tag{2.8b}$$

Since the entries of $\widehat{M}$ and M^* all lie in the interval $[0, 1]$, integrating this tail bound leads to the stated upper bound (2.7) on the expected mean-squared error. Proving the high probability bound (2.8b) requires sharp control on a quantity known as the localized Gaussian complexity of the class $\mathbb{C}_{\text{SST}}$. We use Dudley's entropy integral (e.g., [256, Corollary 2.2.8]) in order to derive an upper bound that is sharp up to a logarithmic factor; doing so in turn requires deriving upper bounds on the metric entropy of the class $\mathbb{C}_{\text{SST}}$ for which we leverage the prior work of Gao and Wellner [86].

We do not know whether the constrained least-squares estimator (2.8a) is computable in time polynomial in n, but we suspect not. This complexity is a consequence of the fact that the set $\mathbb{C}_{\text{SST}}$ is not convex, but is a union of $n!$ convex sets. Given this issue, it becomes interesting to consider the performance of alternative estimators that can be computed in polynomial-time.

2.3.2 Sharp analysis of singular value thresholding (SVT)

The first polynomial-time estimator that we consider is a simple estimator based on thresholding singular values of the observed matrix Y, and reconstructing its truncated singular value decomposition. For the full class $\mathbb{C}_{\text{SST}}$, Chatterjee [44] analyzed the performance of such an estimator and proved that the squared Frobenius error decays as $\mathcal{O}(n^{-\frac{1}{4}})$ uniformly over $\mathbb{C}_{\text{SST}}$. In this section, we prove that its error decays as $\mathcal{O}(n^{-\frac{1}{2}})$, again uniformly over $\mathbb{C}_{\text{SST}}$, and moreover, that this upper bound is unimprovable.

Let us begin by describing the estimator. Given the observation matrix $Y \in \mathbb{R}^{n \times n}$, we can write its singular value decomposition as $Y = UDV^T$, where the $(n \times n)$ matrix D is diagonal, whereas the $(n \times n)$ matrices U and V are orthonormal. Given a threshold level $\lambda_n > 0$, the soft-thresholded version of a diagonal matrix D is the diagonal matrix $T_{\lambda_n}(D)$ with values

$$[T_{\lambda_n}(D)]_{jj} = \max\{0, D_{jj} - \lambda_n\} \quad \text{for every integer } j \in [n]. \tag{2.9}$$

With this notation, the soft singular-value-thresholded (soft-SVT) version of Y is given by $T_{\lambda_n}(Y) = UT_{\lambda_n}(D)V^T$. The following theorem provides a bound on its squared Frobenius error:

Theorem 2. *There are universal positive constants (c_1, c_0, c_1) such that the soft-SVT estimator $\widehat{M}_{\lambda_n} = T_{\lambda_n}(Y)$ with $\lambda_n = 2.1\sqrt{n}$ satisfies the bound*

$$\mathbb{P}\Big[\frac{1}{n^2}|\!|\!|\widehat{M}_{\lambda_n} - M^*|\!|\!|_F^2 \geq \frac{c_1}{\sqrt{n}}\Big] \leq c_0 e^{-c_1 n} \tag{2.10a}$$

for any $M^* \in \mathbb{C}_{SST}$. Moreover, there is a universal constant $c_2 > 0$ such that for any choice of λ_n*, we have*

$$\sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2}|\!|\!|\widehat{M}_{\lambda_n} - M^*|\!|\!|_F^2 \geq \frac{c_2}{\sqrt{n}}. \tag{2.10b}$$

A few comments on this result are in order. Since the matrices $\widehat{M}_{\lambda_n}$ and M^* have entries in the unit interval $[0, 1]$, the normalized squared error $\frac{1}{n^2}|\!|\!|\widehat{M}_{\lambda_n} - M^*|\!|\!|_{\mathrm{F}}^2$ is at most 1. Consequently, by integrating the the tail bound (2.10a), we find that

$$\sup_{M^* \in \mathbb{C}_{\mathrm{SST}}} \mathbb{E}[\frac{1}{n^2}|\!|\!|\widehat{M}_{\lambda_n} - M^*|\!|\!|_{\mathrm{F}}^2] \leq \frac{c_1}{\sqrt{n}} + c_0 e^{-c_1 n} \leq \frac{c_1'}{\sqrt{n}}.$$

On the other hand, the matching lower bound (2.10b) holds with probability one, meaning that the soft-SVT estimator has squared error of the order $1/\sqrt{n}$, irrespective of the realization of the noise.

To be clear, Chatterjee [44] actually analyzed the hard-SVT estimator, which is based on the hard-thresholding operator

$$[H_{\lambda_n}(D)]_{jj} = D_{jj}\ \mathbf{1}\{D_{jj} \geq \lambda_n\}.$$

Here $\mathbf{1}\{\cdot\}$ denotes the 0-1-valued indicator function. In this setting, the hard-SVT estimator is simply, $H_{\lambda_n}(Y) = UH_{\lambda_n}(D)V^T$. With essentially the same choice of λ_n as above, Chatterjee showed that the estimate $H_{\lambda_n}(Y)$ has a mean-squared error of $\mathcal{O}(n^{-1/4})$. One can verify that the proof of Theorem 2 in this chapter goes through for the hard-SVT estimator as well. Consequently the performance of the hard-SVT estimator is of the order $\Theta(n^{-1/2})$, and is identical to that of the soft-thresholded version up to universal constants.

Together the upper and lower bounds of Theorem 2 provide a sharp characterization of the behavior of the soft/hard SVT estimators. On the positive side, these are easily implementable estimators that achieve a mean-squared error bounded by $\mathcal{O}(1/\sqrt{n})$ uniformly over the entire class $\mathbb{C}_{\mathrm{SST}}$. On the negative side, this rate is slower than the $\mathcal{O}(\log^2 n/n)$ rate achieved by the least-squares estimator, as in Theorem 1.

Note that the hard and soft-SVT estimators return matrices that may not lie in the permutation-based SST class $\mathbb{C}_{\mathrm{SST}}$. In Section 2.3.3, we provide an alternate polynomial-time computable estimator with similar statistical guarantees that is guaranteed to return a matrix in the permutation-based SST class.

2.3.3 A polynomial-time computable proper learning estimator

In this section, we propose an estimator that is computable in polynomial time, which we term the *Count-Randomize-Least-Squares (CRL)* estimator, and show that it incurs an error of at most $\widetilde{\mathcal{O}}(1/\sqrt{n})$. This upper bound on the error of the CRL estimator matches (up to logarithmic factors) the error of the SVT estimator discussed earlier, thereby imparting CRL with the best known error guarantees for polynomial-time estimation of permutation-based SST matrices. Moreover, unlike the SVT estimator, the CRL estimator is guaranteed to output a matrix that lies in the permutation-based SST class. This property is known as 'proper learning'. It follows from our results that the CRL estimator is the first known polynomial-time, proper learning estimator that is consistent over the permutation-based SST class. Moreover, we also show that the CRL estimator yields a minimax-optimal estimate for every parameter-based classes.

In order to define the CRL estimator, we require some additional notation. For any permutation π on n items, let $\mathbb{C}_{\text{SST}}(\pi) \subseteq \mathbb{C}_{\text{SST}}$ denote the set of all permutation-based SST matrices that are faithful to the permutation π—that is

$$\mathbb{C}_{\text{SST}}(\pi) := \big\{ M \in [0,1]^{n\times n} \;\mid\; M_{ba} = 1 - M_{ab} \;\forall\; (a,b), M_{ik} \geq M_{jk} \;\forall\; i,j,k \in [n] \text{ s.t. } \pi(i) < \pi(j) \big\}. \tag{2.11}$$

One can verify that the sets $\{\mathbb{C}_{\text{SST}}(\pi)\}$ for all permutations π on n items together comprise the permutation-based SST class $\mathbb{C}_{\text{SST}}$.

The CRL estimator acts on the observed matrix Y and outputs an estimate $\widehat{M}_{\text{CRL}} \in \mathbb{C}_{\text{SST}}$ via a three-step procedure:

Step 1 (Count): For each $i \in [n]$, compute the total number $N_i = \sum_{j=1}^{n} Y_{ij}$ of pairwise comparisons that it wins. Order the n items in terms of $\{N_i\}_{i=1}^{n}$, with ties broken arbitrarily.

Step 2 (Randomize): Find the largest subset of items S such that $|N_i - N_j| \leq \sqrt{n}\log n$ for all $i, j \in S$. Taking the ordering computed in Step 1, permute this (contiguous) subset of items uniformly at random within the subset. Denote the resulting permutation as π_{CRL}.

Step 3 (Least squares): Compute the least squares estimate assuming that the permutation π_{CRL} is the true permutation of the items:

$$\widehat{M}_{\text{CRL}} \in \underset{M \in \mathbb{C}_{\text{SST}}(\pi_{\text{CRL}})}{\arg\min} \; |\!|\!| Y - M |\!|\!|_{\text{F}}^2. \tag{2.12}$$

It is not hard to see that computing the first two steps of the algorithm requires at most an order n^2 computational complexity. The optimization problem (2.12) in the third step corresponds to a projection onto the polytope of bi-isotone matrices contained within the hypercube $[0,1]^{n\times n}$, along with skew symmetry constraints. Problems of the form (2.12) have been studied in past work [25, 44, 135, 209], and the estimator $\widehat{M}_{\text{CRL}}$ is indeed computable in polynomial time.

The second step involving randomization serves to discard "non-robust" information from the ordering computed in Step 1. To clarify our choice of threshold $T = \sqrt{n}\log n$, the factor $\sqrt{n}$ corresponds to the standard deviation of a typical win count N_i (as a sum of Bernoulli variables), whereas the $\log n$ serves to control fluctuations in a union bound. An ordering of the items whose counts are within this threshold is likely to arise from the noise due to the Bernoulli sampling process, as opposed to structural information about the matrix. If we do not perform this second step—effectively retaining considerable bias from Step 1—then isotonic regression procedure in Step 3 may amplify it, leading to a poorly performing estimator. In particular, the randomization step helps the estimator adapt to the situation when there is a large indifference set of size at least $\frac{n}{2}$. Such situations arise in various practical applications, for instance, in depth recognition via crowdsourcing. In this application, the n items are pixels of an image, and workers in crowdsourcing compare pairs of points (pixels) and choose the one that seems closer.

The following theorem provides upper bounds on the risk of the CRL estimator for the permutation-based SST model as well as for parameter-based models.

Theorem 3. *(a) For every M^* in the permutation-based SST model $\mathbb{C}_{\text{SST}}$, the CRL estimator has mean-squared Frobenius error at most*

$$\frac{1}{n^2}\mathbb{E}[|\!|\!|\widehat{M}_{CRL} - M^*|\!|\!|_F^2] \leq c_1 \frac{(\log n)^3}{\sqrt{n}}, \tag{2.13}$$

where c_1 is a universal constant.
(b) For every matrix M^* in any parameter-based model (2.5)*, the risk of the CRL estimator $\widehat{M}_{CRL}$ is upper bounded as*

$$\frac{1}{n^2}\mathbb{E}[|\!|\!|\widehat{M}_{CRL} - M^*|\!|\!|_F^2] \leq c_F \frac{1}{n} \log^2 n,$$

where the c_F is a constant that depends only on F.

A few remarks are in order. In comparison to any estimator tailored to any parameter-based model, there are two key benefits offered by the CRL estimator. First, unlike the parameter-based estimators, the CRL estimator does not need to know the function F. Second, the CRL estimator is more robust to model misspecification, with an error at most $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})$ over the richer permutation-based SST model. This guarantee is significantly superior to the $\Omega(1)$ error incurred by the estimators that fit some parameter-based model (shown in Theorem 5 to follow).

We show in a companion paper [223] that the CRL estimator has an additional appealing property that it can automatically adapt to smoothness in the true matrix M^* and provides better rates of estimation. The primary purpose of the randomization step in the CRL estimator is to facilitate is adaptivity and improved rates. If one is concerned only about attaining the upper bound (2.13) on the worst case error, then the randomization step in the CRL estimator is unnecessary and the count and the l east-squares steps alone suffice to achieve this bound.

An implication of part (b) of this theorem is that for parameter-based models, up to logarithmic factors in n, the CRL estimator is minimax optimal and matches the lower bounds for parameter-based models derived subsequently in Section 2.3.5 of this chapter. In an independent piece of work concurrent with our paper [223] on the CRL estimator, Chatterjee and Mukherjee [43] also investigate adaptivity of a similar estimator to parameter-based models, and show that it attains an error of order $\widetilde{\mathcal{O}}(\frac{1}{n})$. The proof techniques employed in the paper [43] are however markedly different from our proof techniques.

In conjunction, Theorem 1, Theorem 2, and Theorem 3 raise a natural open question: is there a polynomial-time estimator that achieves the minimax rate uniformly over the family $\mathbb{C}_{\text{SST}}$? We do not know the answer to this question, but the following subsections provide some partial answers by analyzing some polynomial-time estimators that (up to logarithmic factors) achieve the optimal $\tilde{\mathcal{O}}(1/n)$-rate over some interesting sub-classes of $\mathbb{C}_{\text{SST}}$. In the next two sections, we turn to results of this type.

2.3.4 Optimal rates for high SNR subclass

In this section, we describe a multi-step polynomial-time estimator that (up to logarithmic factors) can achieve the optimal $\tilde{\mathcal{O}}(1/n)$ rate over an interesting subclass of $\mathbb{C}_{\text{SST}}$. This subset corresponds to matrices M that have a relatively high signal-to-noise ratio (SNR), meaning that no entries of M fall within a certain window of $1/2$. More formally, for some $\gamma \in (0, \frac{1}{2})$, we define the class

$$\mathbb{C}_{\text{HIGH}}(\gamma) = \big\{M \in \mathbb{C}_{\text{SST}} \mid \max(M_{ij}, M_{ji}) \geq 1/2 + \gamma \;\; \forall \;\; i \neq j\big\}. \tag{2.14}$$

By construction, for any matrix $\mathbb{C}_{\text{HIGH}}(\gamma)$, the amount of information contained in each observation is bounded away from zero uniformly in n, as opposed to matrices in which some large subset of entries have values equal (or arbitrarily close) to $\frac{1}{2}$. In terms of estimation difficulty, this SNR restriction does not make the problem substantially easier: as the following theorem shows, the minimax mean-squared error remains lower bounded by a constant multiple of $1/n$. Moreover, we can demonstrate a polynomial-time algorithm that achieves this optimal mean squared error up to logarithmic factors.

The following theorem applies to any fixed $\gamma \in (0, \frac{1}{2}]$ independent of n, and involves constants (c_2, c_1, c) that may depend on γ but are independent of n.

Theorem 4. *There is a constant $c_2 > 0$ such that*

$$\inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{HIGH}(\gamma)} \frac{1}{n^2} \mathbb{E}\big[|\!|\!|\widetilde{M} - M^*|\!|\!|_F^2\big] \geq \frac{c_2}{n}, \tag{2.15a}$$

where the infimum ranges over all estimators. Moreover, there is a two-stage estimator $\widehat{M}$, computable in polynomial-time, for which

$$\mathbb{P}\Big[\frac{1}{n^2}|\!|\!|\widehat{M} - M^*|\!|\!|_F^2 \geq \frac{c_1 \log^2(n)}{n}\Big] \leq \frac{c}{n^2}, \tag{2.15b}$$

valid for any $M^* \in \mathbb{C}_{HIGH}(\gamma)$.

As before, since the ratio $\frac{1}{n^2}|||\widehat{M} - M^*|||_{\text{F}}^2$ is at most 1, so the tail bound (2.15b) implies that

$$\sup_{M^* \in \mathbb{C}_{\text{HIGH}}(\gamma)} \frac{1}{n^2}\mathbb{E}[|||\widehat{M} - M^*|||_{\text{F}}^2] \leq \frac{c_1 \log^2(n)}{n} + \frac{c}{n^2} \leq \frac{c_1' \log^2(n)}{n}. \tag{2.16}$$

As with our proof of the lower bound in Theorem 1, we prove the lower bound by considering the sub-class of matrices that are free only on the two diagonals just above and below the main diagonal. We now provide a brief sketch for the proof of the upper bound (2.15b). It is based on analyzing the following two-step procedure:

1. In the first step of algorithm, we find a permutation $\widehat{\pi}_{\text{FAS}}$ of the n items that minimizes the total number of disagreements with the observations. (For a given ordering, we say that any pair of items (i, j) are in disagreement with the observation if either i is rated higher than j in the ordering and $Y_{ij} = 0$, or if i is rated lower than j in the ordering and $Y_{ij} = 1$.) The problem of finding such a disagreement-minimizing permutation $\widehat{\pi}_{\text{FAS}}$ is commonly known as the minimum feedback arc set (FAS) problem. It is known to be NP-hard in the worst-case [1, 2], but our set-up has additional probabilistic structure that allows for polynomial-time solutions with high probability. In particular, we call upon a polynomial-time algorithm due to Braverman and Mossel [23] that, under the model (2.14), is guaranteed to find the exact solution to the FAS problem with high probability. Viewing the FAS permutation $\widehat{\pi}_{\text{FAS}}$ as an approximation to the true permutation π^*, the novel technical work in this first step is show that $\widehat{\pi}_{\text{FAS}}$ is "good enough" for Frobenius norm estimation, in the sense that for any matrix $M^* \in \mathbb{C}_{\text{HIGH}}(\gamma)$, it satisfies the bound

$$\frac{1}{n^2}|||\pi^*(M^*) - \widehat{\pi}_{\text{FAS}}(M^*)|||_{\text{F}}^2 \leq \frac{c \log n}{n} \tag{2.17a}$$

   with high probability. In this statement, for any given permutation π, we have used $\pi(M^*)$ to denote the matrix obtained by permuting the rows and columns of M^* by π. The term $\frac{1}{n^2}|||\pi^*(M^*) - \widehat{\pi}_{\text{FAS}}(M^*)|||_{\text{F}}^2$ can be viewed in some sense as the *bias* in estimation incurred from using $\widehat{\pi}_{\text{FAS}}$ in place of π^*.

2. Next we define $\mathbb{C}_{\text{BISO}}$ as the class of "bivariate isotonic" matrices, that is, matrices $M \in [0, 1]^{n \times n}$ that satisfy the linear constraints $M_{ij} = 1 - M_{ji}$ for all $(i, j) \in [n]^2$, and $M_{k\ell} \geq M_{ij}$ whenever $k \leq i$ and $\ell \geq j$. This class corresponds to the subset of matrices $\mathbb{C}_{\text{SST}}$ that are faithful with respect to the identity permutation. Letting $\widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}}) = \{\widehat{\pi}_{\text{FAS}}(M),\ M \in \mathbb{C}_{\text{BISO}}\}$ denote the image of this set under $\widehat{\pi}_{\text{FAS}}$, the second step involves computing the constrained least-squares estimate

$$\widehat{M} \in \underset{M \in \widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}})}{\arg\min} |||Y - M|||_{\text{F}}^2. \tag{2.17b}$$

   Since the set $\widehat{\pi}_{\text{FAS}}(\mathbb{C}_{\text{BISO}})$ is a convex polytope, with a number of facets that grows polynomially in n, the constrained quadratic program (2.17b) can be solved in polynomial-time.

The final step in the proof of Theorem 4 is to show that the estimator $\widehat{M}$ also has mean-squared error that is upper bounded by a constant multiple of $\frac{\log^2(n)}{n}$.

Our analysis in Theorem 4 shows that for any fixed $\gamma \in (0, \frac{1}{2}]$, the proposed two-step estimator works well for any matrix $M^* \in \mathbb{C}_{\text{HIGH}}(\gamma)$. Since this two-step estimator is based on finding a minimum feedback arc set (FAS) in the first step, it is natural to wonder whether an FAS-based estimator works well over the full class $\mathbb{C}_{\text{SST}}$. Somewhat surprisingly the answer to this question turns out to be negative.

Minimizing feedback arc set over entire permutation-based SST class The two-step estimator analyzed in Theorem 4 for the High-SNR subclass, $\mathbb{C}_{\text{HIGH}}(\gamma) \subseteq \mathbb{C}_{\text{SST}}$ for a fixed γ, is based on finding a minimum feedback arc set (FAS) in the first step. We now show that minimizing the FAS, however, does not work well over the full class $\mathbb{C}_{\text{SST}}$. The intuition is that although minimizing the feedback arc set appears to minimize disagreements at a global scale, it makes only local decisions: if it is known that items i and j are in adjacent positions, the order among these two items is decided based solely on the outcome of the comparison between items i and j, and is independent of the outcome of the comparisons of i and j with all other items.

Here is a concrete example to illustrate this property. Suppose n is divisible by 3, and consider the following $(n \times n)$ block matrix $M \in \mathbb{C}_{\text{SST}}$:

$$M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix},$$

where each block is of size $(\frac{n}{3} \times \frac{n}{3})$. Let π^1 be the identity permutation, and let π^2 be the permutation $[\frac{n}{3}+1, \ldots, \frac{2n}{3}, 1, \ldots, \frac{n}{3}, \frac{2n}{3}+1, \ldots, n]$, that is, π^2 swaps the second block of $\frac{n}{3}$ items with the first block. For any permutation π of the n items and any $M \in \mathbb{C}_{\text{SST}}$, let $\pi(M)$ denote the $(n \times n)$ matrix resulting from permuting both the rows and the columns by π.

One can verify that $|\!|\!|\pi^1(M) - \pi^2(M)|\!|\!|_{\text{F}}^2 \geq cn^2$, for some universal constant $c > 0$. Now suppose an observation Y is generated from $\pi^1(M)$ as per the model (2.1). Then the probability distribution of the size of the feedback arc set of π^1 is identical to the probability distribution of the size of the feedback arc set of π^2. Minimizing FAS cannot distinguish between $\pi^1(M)$ and $\pi^2(M)$ at least 50% of the time, and consequently, any estimator based on the minimum FAS output cannot perform well over the permutation-based SST class.

2.3.5 Optimal rates for parameter-based subclasses

Let us now return to the class of parameter-based models $\mathbb{C}_{\text{PAR}}(F)$ introduced earlier in Section 2.2.3. As shown previously in Proposition 1, this class is much smaller than the class $\mathbb{C}_{\text{SST}}$, in the sense that there are models in $\mathbb{C}_{\text{SST}}$ that cannot be well-approximated by any parameter-based model. Nonetheless, in terms of minimax rates of estimation, these classes differ only by logarithmic factors. An advantage of the parameter-based class is that it is possible to achieve the $1/n$ minimax rate by using a simple, polynomial-time estimator. In particular, for any log concave function F, the maximum likelihood estimate $\widehat{w}_{\mathrm{ML}}$ can be obtained by solving a convex program. This MLE induces a matrix estimate $M(\widehat{w}_{\mathrm{ML}})$ via Equation (2.4), and the following result shows that this estimator is minimax-optimal up to constant factors.

Theorem 5. *Suppose that F is strictly increasing, strongly log-concave and twice differentiable. Then there is a constant $c_2 > 0$, depending only on F, such that the minimax risk over $\mathbb{C}_{PAR}(F)$ is lower bounded as*

$$\inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} \mathbb{E}[\vert\!\vert\!\vert \widetilde{M} - M^* \vert\!\vert\!\vert_F^2] \geq \frac{c_2}{n}, \tag{2.18a}$$

Conversely, there is a constant $c_1 \geq c_2$, depending only on F, such that the matrix estimate $M(\widehat{w}_{ML})$ induced by the MLE satisfies the bound

$$\sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} \mathbb{E}[\vert\!\vert\!\vert M(\widehat{w}_{ML}) - M^* \vert\!\vert\!\vert_F^2] \leq \frac{c_1}{n}. \tag{2.18b}$$

To be clear, the constants (c_2, c_1) in this theorem are independent of n, but they do depend on the specific properties of the given function F. We note that the stated conditions on F are true for many popular parameter-based models, including (for instance) the Thurstone and BTL models.

The lower bound (2.18a) is, in fact, stronger than the the lower bound in Theorem 1, since the supremum is taken over a smaller class. The proof of the lower bound in Theorem 1 relies on matrices that cannot be realized by any parameter-based model, so that we pursue a different route to establish the bound (2.18a). On the other hand, in order to prove the upper bound (2.18b), we make use of bounds on the accuracy of the MLE $\widehat{w}_{\mathrm{ML}}$ from our own past work [220].

2.3.6 Extension to partial observations

We now consider the extension of our results to the setting in which not all entries of Y are observed. Suppose instead that every entry of Y is observed independently with probability p_{obs}. In other words, the set of pairs compared is the set of edges of an Erdős-Rényi graph $\mathcal{G}(n, p_{\mathrm{obs}})$ that has the n items as its vertices.

In this setting, we obtain an upper bound on the minimax risk of estimation by first setting $Y_{ij} = \frac{1}{2}$ whenever the pair (i, j) is not compared, then forming a new $(n \times n)$ matrix Y' as

$$Y' := \frac{1}{p_{\mathrm{obs}}} Y - \frac{1 - p_{\mathrm{obs}}}{2 p_{\mathrm{obs}}} 11^T, \tag{2.19a}$$

and finally computing the least squares solution

$$\widehat{M} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} |||Y' - M|||_{\text{F}}^2. \tag{2.19b}$$

The intuition behind the transformation (2.19a) is provided following the statement of Theorem 6.

Likewise, the computationally-efficient singular value thresholding estimator is also obtained by thresholding the singular values of Y' with a threshold $\lambda_n = 3\sqrt{\frac{n}{p_{\text{obs}}}}$. See our discussion following Theorem 6 for the motivation underlying the transformed matrix Y'.

The parameter-based estimators continue to operate on the original (partial) observations, first computing a maximum likelihood estimate $\widehat{w}_{\text{ML}}$ of M^* using the observed data, and then computing the associated pairwise-comparison-probability matrix $M(\widehat{w}_{\text{ML}})$ via (2.4).

Theorem 6. *In the setting where each pair is observed with a probability* p_{obs}*, there are positive universal constants* c_2*,* c_1 *and* c_4 *such that:*

(a) The minimax risk is sandwiched as

$$\frac{c_2}{p_{\text{obs}} n} \leq \inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2} \mathbb{E}[|||\widetilde{M} - M^*|||_F^2] \leq \frac{c_1 (\log n)^2}{p_{\text{obs}} n}, \tag{2.20a}$$

when $p_{\text{obs}} \geq \frac{c_4}{n}$.

(b) The soft-SVT estimator, $\widehat{M}_{\lambda_n}$ *with* $\lambda_n = 3\sqrt{\frac{n}{p_{\text{obs}}}}$*, satisfies the bound*

$$\sup_{M^* \in \mathbb{C}_{SST}} \frac{1}{n^2} \mathbb{E}[|||\widehat{M}_{\lambda_n} - M^*|||_F^2] \leq \frac{c_1}{\sqrt{n p_{\text{obs}}}}, \tag{2.20b}$$

when $p_{\text{obs}} \geq \frac{c_4 \log^7 n}{n}$.

(c) For a parameter-based sub-class based on a strongly log-concave and smooth F*, the estimator* $M(\widehat{w}_{ML})$ *induced by the maximum likelihood estimate* $\widehat{w}_{ML}$ *of the parameter vector* w^* *has mean-squared error upper bounded as*

$$\sup_{M^* \in \mathbb{C}_{PAR}(F)} \frac{1}{n^2} \mathbb{E}[|||M(\widehat{w}_{ML}) - M^*|||_F^2] \leq \frac{c_1}{p_{\text{obs}} n}, \tag{2.20c}$$

when $p_{\text{obs}} \geq \frac{c_4 \log^2 n}{n}$.

The intuition behind the transformation (2.19a) is that the matrix Y' can equivalently be written in a linearized form as

$$Y' = M^* + \frac{1}{p_{\text{obs}}} W', \tag{2.21a}$$

where W' has entries that are independent on and above the diagonal, satisfy skew-symmetry, and are distributed as

$$[W']_{ij} = \begin{cases} p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) + \frac{1}{2} & \text{with probability } p_{\text{obs}}[M^*]_{ij} \\ p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) - \frac{1}{2} & \text{with probability } p_{\text{obs}}(1 - [M^*]_{ij}) \\ p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) & \text{with probability } 1 - p_{\text{obs}}. \end{cases} \tag{2.21b}$$

The proofs of the upper bounds exploit the specific relation (2.21a) between the observations Y' and the true matrix M^*, and the specific form of the additive noise (2.21b).

The result of Theorem 6(b) yields an affirmative answer to the question, originally posed by Chatterjee [44], of whether or not the singular value thresholding estimator can yield a vanishing error when $p_{\text{obs}} \leq \frac{1}{\sqrt{n}}$.

We note that we do not have an analogue of the high-SNR result in the partial observations case since having partial observations reduces the SNR. In general, we are interested in scalings of p_{obs} which allow $p_{\text{obs}} \to 0$ as $n \to \infty$. The noisy-sorting algorithm of Braverman and Mossel [23] for the high-SNR case has computational complexity scaling as $e^{\gamma^{-4}}$, and hence is not computable in time polynomial in n when $\gamma < (\log n)^{-\frac{1}{4}}$. This restriction disallows most interesting scalings of p_{obs} with n.

2.3.7 Relation to other models

In this section, we put the permutation-based SST model and the parameter-based models that we studied earlier in perspective to other models considered in the literature. We begin with two weaker versions of stochastic transitivity that are also investigated in the literature on psychology and social science.

Moderate and weak stochastic transitivity

The model $\mathbb{C}_{\text{SST}}$ that we consider is called strong stochastic transitivity in the literature on psychology and social science [59, 78]. The two other popular (and weaker) models are those of *moderate stochastic transitivity* $\mathbb{C}_{\text{MST}}$ defined as

$$\mathbb{C}_{\text{MST}} := \big\{M \in [0,1]^{n\times n} \mid M_{ik} \geq \min\{M_{ij}, M_{jk}\} \text{ for every } (i,j,k) \text{ satisfying } \\ M_{ij} \geq \tfrac{1}{2} \text{ and } M_{jk} \geq \tfrac{1}{2}\big\},$$

and *weak stochastic transitivity* $\mathbb{C}_{\text{WST}}$ defined as

$$\mathbb{C}_{\text{WST}} := \big\{M \in [0,1]^{n\times n} \mid M_{ik} \geq \tfrac{1}{2} \text{ for every } (i,j,k) \text{ satisfying } M_{ij} \geq \tfrac{1}{2} \text{ and } M_{jk} \geq \tfrac{1}{2}\big\}.$$

Clearly, we have the inclusions $\mathbb{C}_{\text{SST}} \subseteq \mathbb{C}_{\text{MST}} \subseteq \mathbb{C}_{\text{WST}}$.

In Theorem 1, we prove that the minimax rates of estimation under the strong stochastic transitivity assumption are $\tilde{\Theta}(n^{-1})$. It turns out, however, that the two weaker transitivity conditions do not permit meaningful estimation.

Proposition 2. *There exists a universal constant $c > 0$ such that under the moderate $\mathbb{C}_{\scriptscriptstyle MST}$ stochastic transitivity model,*

$$\inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{MST}} \frac{1}{n^2} \mathbb{E}[\lvert\!\lvert\!\lvert \widetilde{M} - M^* \rvert\!\rvert\!\rvert_F^2] > c.$$

where the infimum is taken over all measurable mappings from the observations Y to $[0,1]^{n \times n}$. Consequently, for the weak stochastic transitivity model $\mathbb{C}_{\scriptscriptstyle WST}$, we also have

$$\inf_{\widetilde{M}} \sup_{M^* \in \mathbb{C}_{WST}} \frac{1}{n^2} \mathbb{E}[\lvert\!\lvert\!\lvert \widetilde{M} - M^* \rvert\!\rvert\!\rvert_F^2] > c,$$

The minimax risk over these two classes is clearly the worst possible (up to a universal constant) since for any two arbitrary matrices M and M' in $[0,1]^{n \times n}$, we have $\frac{1}{n^2} \lvert\!\lvert\!\lvert M - M' \rvert\!\rvert\!\rvert_{\mathrm{F}}^2 \leq 1$. For this reason, in the chapter we restrict our analysis to the strong stochastic transitivity condition.

Comparison with statistical models

Let us now investigate relationship of the strong stochastic transitivity model considered in this chapter with two other popular models in the literature on statistical learning from comparative data. Perhaps the most popular model in this regard is the class of parameter-based models $\mathbb{C}_{\scriptscriptstyle \mathrm{PAR}}$: recall that this class is defined as

$$\begin{aligned} \mathbb{C}_{\scriptscriptstyle \mathrm{PAR}} := \{ M \mid M_{ij} = F(w_i^* - w_j^*) & \text{ for some non-decreasing function } F : \mathbb{R} \to [0,1], \\ & \text{and vector } w^* \in \mathbb{R}^n \}. \end{aligned}$$

The parameter-based class of models assumes that the function F is fixed and known. Statistical estimation under the parameter-based class is studied in several recent papers [98, 177, 220]. The setting where the function F is fixed, but unknown leads to a semi-parameter-based variant. The results presented in this section also readily apply to the semi-parameter-based class.

The second class is that generated from distributions over complete rankings [63, 65, 76]. Specifically, every element in this class is generated as the pairwise marginal of an arbitrary probability distribution over all possible permutations of the n items. We denote this class as $\mathbb{C}_{\scriptscriptstyle \mathrm{FULL}}$.

The following result characterizes the relation between the classes.

Proposition 3. *Consider any value of $n > 10$. The parameter-based class $\mathbb{C}_{\scriptscriptstyle PAR}$ is a strict subset of the strong stochastic transitivity class $\mathbb{C}_{\scriptscriptstyle SST}$. The class $\mathbb{C}_{\scriptscriptstyle FULL}$ of marginals of a distribution on total rankings is neither a subset nor a superset of either of the classes $\mathbb{C}_{\scriptscriptstyle SST}$, $\mathbb{C}_{\scriptscriptstyle PAR}$, and $\mathbb{C}_{\scriptscriptstyle SST} \backslash \mathbb{C}_{\scriptscriptstyle PAR}$.*

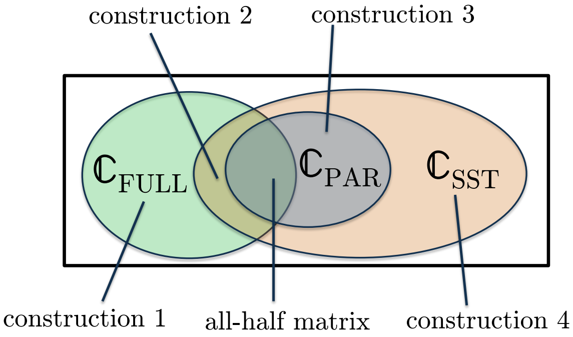

construction 2
construction 3
$\mathbb{C}_{\text{FULL}}$
$\mathbb{C}_{\text{PAR}}$
$\mathbb{C}_{\text{SST}}$
construction 1
all-half matrix
construction 4

Figure 2.2: Relations between various models of pairwise comparisons. The constructions proving these relations are presented as a part of the proof of Proposition 3.

The various relationships in Proposition 3 are depicted pictorially in Figure 2.2. These relations are derived by first establishing certain conditions that matrices in the classes considered must satisfy, and then constructing matrices that satisfy or violate one or more of these conditions. The conditions on $\mathbb{C}_{\text{FULL}}$ arise from the observation that the class is the convex hull of all permutation-based SST matrices that have their non-diagonal elements in $\{0, 1\}$; we derive conditions on this convex hull that leads to properties of the $\mathbb{C}_{\text{FULL}}$ class. To handle the parameter-based class $\mathbb{C}_{\text{PAR}}$, we employ a necessary condition discussed earlier in Section 2.2.4 and defined formally in Lemma 1. The permutation-based SST class $\mathbb{C}_{\text{SST}}$ is characterized using the insights derived throughout this chapter.

2.4 Simulations

In this section, we present results from simulations to gain a further understanding of the problem at hand, in particular to understand the rates of estimation under specific generative models. We investigate the performance of the soft-SVT estimator (Section 2.3.2) and the maximum likelihood estimator under the Thurstone parameter-based model (Section 2.2.3) which is optimal for the Thurstone parameter-based model.[4] The output of the SVT estimator need not lie in the set $[0, 1]^{n\times n}$ of matrices; in our implementation, we take a projection of the output of the SVT estimator on this set, which gives a constant factor reduction in the error.

[4] We could not compare the algorithm that underlies Theorem 4, since it is not easily implementable. In particular, it relies on the algorithm due to Braverman and Mossel [23] to compute the feedback arc set minimizer. The computational complexity of this algorithm, though polynomial in n, has a large polynomial degree which precludes it from being implemented for matrices of any reasonable size.

The simulations in this section add to the simulation results of Section 2.2.4 (Figure 2.1) demonstrating a large class of matrices in the permutation-based SST class that cannot be represented by any parameter-based class.

In our simulations, we generate the ground truth M^* in the following five ways:

- Uniform: The matrix M^* is generated by drawing $\binom{n}{2}$ values independently and uniformly at random in $[\frac{1}{2}, 1]$ and sorting them in descending order. The values are then inserted above the diagonal of an $(n \times n)$ matrix such that the entries decrease down a column or left along a row. We then make the matrix skew-symmetric and permute the rows and columns.

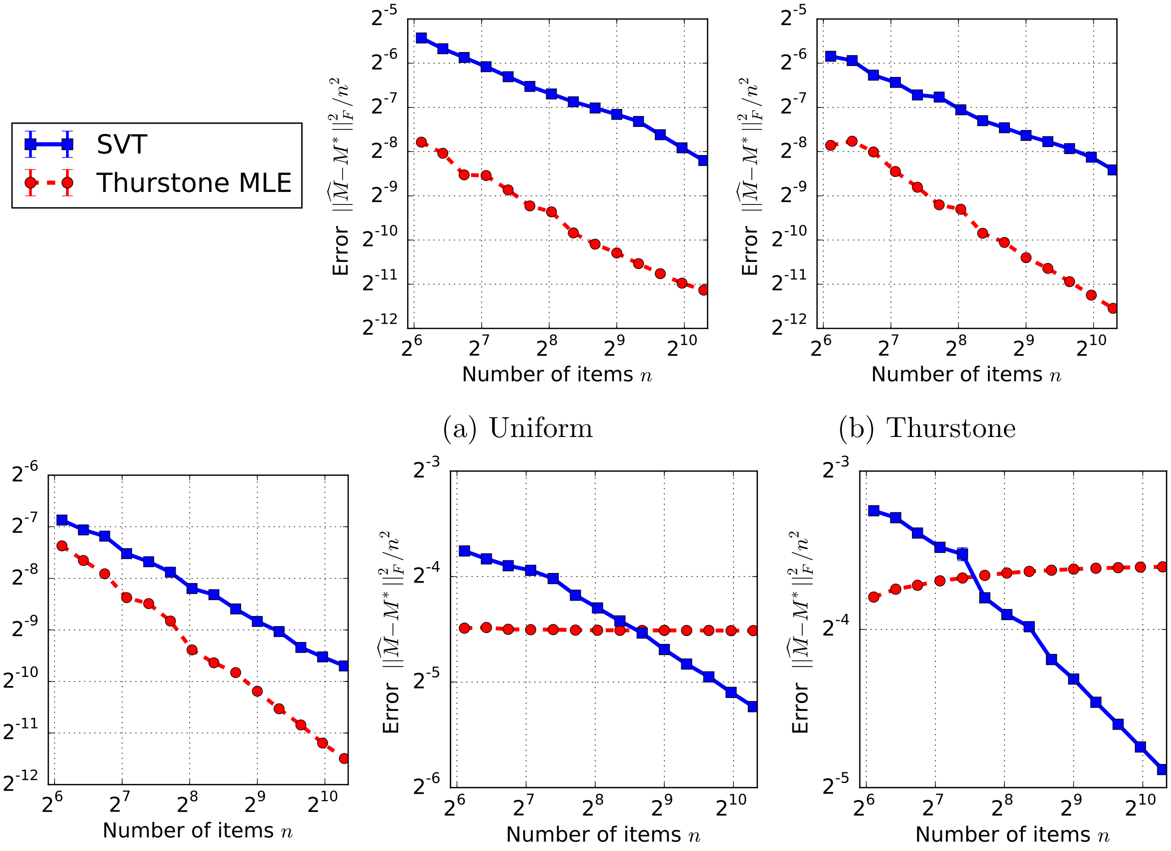
- Thurstone parameter-based model: The matrix $M^* \in [-1, 1]^n$ is generated by first choosing w^* uniformly at random from the set satisfying $\langle w^*, 1\rangle = 0$. The matrix M^* is then generated from w^* via Equation (2.4) with F chosen as the CDF of the standard normal distribution.

- Bradley-Terry-Luce (BTL) parameter-based model: Identical to the Thurstone case, except that F is given by the sigmoid function.

- High SNR: A setting studied previously by Braverman and Mossel [23], in which the noise is independent of the items being compared. Some global order is fixed over the n items, and the comparison matrix M^* takes the values $M^*_{ij} = 0.9 = 1 - M^*_{ji}$ for every pair (i, j) where i is ranked above j in the underlying ordering. The entries on the diagonal are 0.5.

- Independent bands: The matrix M^* is chosen with diagonal entries all equal to $\frac{1}{2}$. Entries on diagonal band immediately above the diagonal itself are chosen i.i.d. and uniformly at random from $[\frac{1}{2}, 1]$. The band above is then chosen uniformly at random from the allowable set, and so on. The choice of any entry in this process is only constrained to be upper bounded by 1 and lower bounded by the entries to its left and below. The entries below the diagonal are chosen to make the matrix skew-symmetric.

Figure 2.3 depicts the results of the simulations based on observations of the entire matrix Y. Each point is an average across 20 trials. The error bars in most cases are too small and hence not visible. We see that the uniform case (Figure 2.3a) is favorable for both estimators, with the error scaling as $\mathcal{O}(\frac{1}{\sqrt{n}})$. With data generated from the Thurstone parameter-based model, both estimators continue to perform well, and the Thurstone MLE yields an error of the order $\frac{1}{n}$ (Figure 2.3b). Interestingly, the Thurstone parameter-based model also fits relatively well when data is generated via the BTL parameter-based model (Figure 2.3c). This behavior is likely a result of operating in the near-linear regime of the logistic and the Gaussian CDF where the two curves are similar. In these two parameter-based settings, the SVT estimator has squared error strictly worse than order $\frac{1}{n}$ but better than $\frac{1}{\sqrt{n}}$. The Thurstone parameter-based model, however, yields a poor fit for the model in the high-SNR (Figure 2.3d) and the independent bands (Figure 2.3e) cases, incurring a constant error as compared to an error scaling as $\mathcal{O}(\frac{1}{\sqrt{n}})$ for the SVT estimator. We



(a) Uniform

(b) Thurstone

(c) BTL

(d) High SNR

(e) Independent bands

Figure 2.3: Errors of singular value thresholding (SVT) estimator and the optimal (maximum likelihood) estimator for the Thurstone parameter-based model under different methods to generate M^*. The error incurred by the CRL estimator is of the same order as that of the SVT.

recall that the poor performance of the Thurstone estimator was also described previously in Proposition 1 and Figure 2.1.

In summary, we see that while the Thurstone maximum likelihood estimator gives minimax optimal rates of estimation when the underlying model is parameter-based, it can be inconsistent when the parameter-based assumptions are violated. On the other hand, the SVT estimator is robust to violations of parameter-based assumptions, and while it does not necessarily give minimax-optimal rates, it remains consistent across the entire permutation-based SST class. Finally, we remark that our theory predicts that the least squares estimator, if implementable, would outperform both these estimators in terms of statistical error.

2.5 Discussion

In this chapter, we analyzed a flexible model for pairwise comparison data that includes various parameter-based models, including the BTL and Thurstone models, as special cases. We analyzed various estimators for this broader matrix family, ranging from optimal estimators to various polynomial-time estimators, including forms of singular value thresholding, the CRL estimator, as well as a multi-stage method based on a noisy sorting routine. We show that this permutation-based SST model permits far more robust estimation as compared to popular parameter-based models, while surprisingly, incurring little penalty for this significant generality. We also show that under weaker notions of stochastic transitivity, the pairwise-comparison probabilities are unestimable. Our results thus present a strong motivation towards the use of such general permutation-based models.

In some applications, choices can be systematically intransitive, for instance when objects have multiple features and different features dominate different pairwise comparisons. In these situations, the permutation-based SST assumption may be weakened to one where the underlying pairwise comparison matrix is a mixture of a small number of permutation-based SST matrices. Later in Chapter 5 of this thesis, we analyze a seting that is equivalent to such a general setting.

All of the results in this chapter focused on estimation of the matrix of pairwise comparison probabilities in the Frobenius norm. Estimation of probabilities in other metrics, such as the KL divergence or estimation of the ranking in the Spearman's footrule or Kemeny distance, follow as corollaries of our results; see Appendix 2.A. Establishing the best possible rates for polynomial-time algorithms over the full class $\mathbb{C}_{\text{SST}}$ is a challenging open problem.

2.6 Proofs

This section is devoted to the proofs of our theoretical results. Throughout these and other proofs, we use the notation $\{c, c', c_0, C, C'\}$ and so on to denote positive constants whose values may change from line to line. In addition, we assume throughout that n is lower bounded by a universal constant so as to avoid degeneracies. For any square matrix $A \in \mathbb{R}^{n \times n}$, we let $\{\sigma_1(A), \ldots, \sigma_n(A)\}$ denote its singular values (ordered from largest to smallest), and similarly, for any symmetric matrix $M \in \mathbb{R}^{n \times n}$, we let $\{\lambda_1(M), \ldots, \lambda_n(M)\}$ denote its ordered eigenvalues. The identity permutation is one where item i is the i^{th} most preferred item, for every $i \in [n]$.

Our lower bounds are based on a standard form of Fano's inequality [54, 254] for lower bounding the probability of error in an L-ary hypothesis testing problem. We state a version here for future reference. For some integer $L \geq 2$, fix some collection of distributions $\{\mathbb{P}^1, \ldots, \mathbb{P}^L\}$. Suppose that we observe a random variable Y that is obtained by first sampling an index A uniformly at random from $[L] = \{1, \ldots, L\}$, and then drawing $Y \sim \mathbb{P}^A$. (As a result, the variable Y is marginally distributed according to the mixture distribution $\overline{\mathbb{P}} = \frac{1}{L}\sum_{a=1}^{L} \mathbb{P}^a$.) Given the observation Y, our goal is to "decode" the value of A, corresponding to the index of the underlying mixture component. Using $\mathcal{Y}$ to denote the sample space associated with the observation Y, Fano's inequality asserts that any test function $\phi : \mathcal{Y} \rightarrow [L]$ for this problem has error probability lower bounded as

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{I(Y;A) + \log 2}{\log L},$$

where $I(Y;A)$ denotes the mutual information between Y and A. A standard convexity argument for the mutual information yields the weaker bound

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{\max\limits_{a,b \in [L]} D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) + \log 2}{\log L}, \tag{2.22}$$

We make use of this weakened form of Fano's inequality in several proofs.

2.6.1 Proof of Proposition 1: Parameter-based models are very restrictive

We show that the matrix M^* specified in Figure 2.1a satisfies the conditions required by the proposition. It is easy to verify that $M^* \in \mathbb{C}_{\text{SST}}$, so that it remains to prove the approximation-theoretic lower bound (2.6). In order to do so, we require the following auxiliary result:

Lemma 1. *Consider any matrix M that belongs to $\mathbb{C}_{\text{PAR}}(F)$ for a valid function F. Suppose for some collection of four distinct items $\{i_1, \ldots, i_4\}$, the matrix M satisfies the inequality $M_{i_1 i_2} > M_{i_3 i_4}$. Then it must also satisfy the inequality $M_{i_1 i_3} \geq M_{i_2 i_4}$.*

We return to prove this lemma at the end of this section. Taking it as given, let us now proceed to prove the lower bound (2.6). For any valid F, fix an arbitrary member M of a class $\mathbb{C}_{\text{PAR}}(F)$, and let $w \in \mathbb{R}^n$ be the underlying weight vector (see the definition (2.4)).

Pick any item in the set of first $\frac{n}{4}$ items (corresponding to the first $\frac{n}{4}$ rows of M^*) and call this item as "1"; pick an item from the next set of $\frac{n}{4}$ items (rows) and call it item "2"; item "3" from the next set and item "4" from the final set. Our analysis proceeds by developing some relations between the pairwise comparison probabilities for these four items that must hold for every parametric model, that are strongly violated by M^*. We divide our analysis into two possible relations between the entries of M.

Case I: First suppose that $M_{12} \leq M_{34}$. Since $M^*_{12} = 6/8$ and $M^*_{34} = 5/8$ in our construction, it follows that

$$(M_{12} - M^*_{12})^2 + (M_{34} - M^*_{34})^2 \geq \frac{1}{256}.$$

Case II: Otherwise, we may assume that $M_{12} > M_{34}$. Then Lemma 1 implies that $M_{13} \geq M_{24}$. Moreover, since $M^*_{13} = 7/8$ and $M^*_{24} = 1$ in our construction, it follows that

$$(M_{13} - M^*_{13})^2 + (M_{24} - M^*_{24})^2 \geq \frac{1}{256}.$$

Aggregating across these two exhaustive cases, we find that

$$\sum_{(u,v)\in\{1,2,3,4\}} (M_{uv} - M^*_{uv})^2 \geq \frac{1}{256}.$$

Since this bound holds for any arbitrary selection of items from the four sets, we conclude that $\frac{1}{n^2}|\!|\!|M - M^*|\!|\!|_{\mathrm{F}}^2$ is lower bounded by a universal constant $c > 0$ as claimed.

Finally, it is easy to see that upon perturbation of any of the entries of M^* by at most $\frac{1}{32}$—while still ensuring that the resulting matrix lies in $\mathbb{C}_{\text{SST}}$—the aforementioned results continue to hold, albeit with a worse constant. Every matrix in this class satisfies the claim of this proposition.

Proof of Lemma 1: It remains to prove Lemma 1. Since M belongs to the parametric family, there must exist some valid function F and some vector w that induce M (see Equation (2.4)). Since F is non-decreasing, the condition $M_{i_1 i_2} > M_{i_3 i_4}$ implies that

$$w_{i_1} - w_{i_2} > w_{i_3} - w_{i_4}.$$

Adding $w_{i_2} - w_{i_3}$ to both sides of this inequality yields $w_{i_1} - w_{i_3} > w_{i_2} - w_{i_4}$. Finally, applying the non-decreasing function F to both sides of this inequality gives yields $M_{i_1 i_3} \geq M_{i_2 i_4}$ as claimed, thereby completing the proof.

2.6.2 Proof of Theorem 1: Minimax risk

This section is devoted to the proof of Theorem 1, including both the upper and lower bounds on the minimax risk in squared Frobenius norm.

Proof of upper bound

Define the difference matrix $\widehat{\Delta} := \widehat{M}$- M^* between M^* and the optimal solution $\widehat{M}$ to the constrained least-squares problem. Since $\widehat{M}$ is optimal and M^* is feasible, we must have $|\!|\!|Y - \widehat{M}|\!|\!|_{\mathrm{F}}^2 \leq |\!|\!|Y - M^*|\!|\!|_{\mathrm{F}}^2$, and hence following some algebra, we arrive at the *basic inequality*

$$\frac{1}{2}|\!|\!|\widehat{\Delta}|\!|\!|_{\mathrm{F}}^2 \leq \langle\!\langle \widehat{\Delta},\; W \rangle\!\rangle, \tag{2.23}$$

where $W \in \mathbb{R}^{n\times n}$ is the noise matrix in the observation model (2.1), and $\langle\!\langle A,\; B\rangle\!\rangle := \text{trace}(A^T B)$ denotes the trace inner product.

We introduce some additional objects that are useful in our analysis. Recall from the main text (2.11) that the class of bivariate isotonic matrices $\mathbb{C}_{\text{BISO}}$ is defined as

$$\mathbb{C}_{\text{BISO}} := \big\{ M \in [0,1]^{n\times n} \mid M_{k\ell} \geq M_{ij} \text{ whenever } k \leq i \text{ and } \ell \geq j \big\}. \tag{2.24}$$

For a given permutation π and matrix M, we let $\pi(M)$ denote the matrix obtained by applying π to its rows and columns. We then define the set

$$\mathbb{C}_{\text{DIFF}} := \Big\{\pi_1(M_1) - \pi_2(M_2) \mid \text{for some } M_1,\ M_2 \in \mathbb{C}_{\text{BISO}}, \text{ and perm. } \pi_1 \text{ and } \pi_2\Big\}. \tag{2.25}$$

corresponding to the set of difference matrices. Note that $\mathbb{C}_{\text{DIFF}} \subset [-1,1]^{n\times n}$ by construction. One can verify that for any $M^* \in \mathbb{C}_{\text{SST}}$, we are guaranteed the inclusion

$$\{M - M^* \mid M \in \mathbb{C}_{\text{SST}},\ |\!|\!| M - M^* |\!|\!|_{\text{F}} \leq t\} \subset \mathbb{C}_{\text{DIFF}} \cap \{|\!|\!| D |\!|\!|_{\text{F}} \leq t\}.$$

Consequently, the error matrix $\widehat{\Delta}$ must belong to $\mathbb{C}_{\text{DIFF}}$, and so must satisfy the properties defining this set. Moreover, as we discuss below, the set $\mathbb{C}_{\text{DIFF}}$ is star-shaped, and this property plays an important role in our analysis.

For each choice of radius $t > 0$, define the random variable

$$Z(t) := \sup_{D \in \mathbb{C}_{\text{DIFF}}, |\!|\!| D |\!|\!|_{\text{F}} \leq t} \langle\!\langle D,\ W \rangle\!\rangle. \tag{2.26}$$

Using our earlier basic inequality (2.23), the Frobenius norm error $|\!|\!| \widehat{\Delta} |\!|\!|_{\text{F}}$ then satisfies the bound

$$\frac{1}{2} |\!|\!| \widehat{\Delta} |\!|\!|_{\text{F}}^2 \leq \langle\!\langle \widehat{\Delta},\ W \rangle\!\rangle \ \leq \ Z\big(|\!|\!| \widehat{\Delta} |\!|\!|_{\text{F}}\big). \tag{2.27}$$

Thus, in order to obtain a high probability bound, we need to understand the behavior of the random quantity $Z(\delta)$.

One can verify that the set $\mathbb{C}_{\text{DIFF}}$ is star-shaped, meaning that $\alpha D \in \mathbb{C}_{\text{DIFF}}$ for every $\alpha \in [0,1]$ and every $D \in \mathbb{C}_{\text{DIFF}}$. Using this star-shaped property, we are guaranteed $Z(t)$ grows at most linearly in t, and hence there is a non-empty set of scalars $\delta_0 > 0$ satisfying the critical inequality

$$\mathbb{E}[Z(\delta_0)] \leq \frac{\delta_0^2}{2}. \tag{2.28}$$

Our interest is in the smallest strictly positive solution δ_0 to the critical inequality (2.28), and moreover, our goal is to show that for every $t \geq \delta_0$, we have $|\!|\!| \widehat{\Delta} |\!|\!|_{\text{F}} \leq c\sqrt{t\delta_0}$ with probability at least $1 - c_1 e^{-c_2 n t \delta_0}$.

For each $t > 0$, define the "bad" event $\mathcal{A}_t$ as

$$\mathcal{A}_t = \big\{\exists \Delta \in \mathbb{C}_{\text{DIFF}} \mid |\!|\!| \Delta |\!|\!|_{\text{F}} \geq \sqrt{t\delta_0} \quad \text{and} \quad \langle\!\langle \Delta,\ W \rangle\!\rangle \geq 2 |\!|\!| \Delta |\!|\!|_{\text{F}} \sqrt{t\delta_0}\big\}. \tag{2.29}$$

Now suppose the event $\mathcal{A}_t$ is true for some $t \geq \delta_0$, and let $\Delta_0 \in \mathbb{C}_{\text{DIFF}}$ be a matrix that satisfies the two conditions required for $\mathcal{A}_t$ to occur. Then using the fact that $Z(t)$ grows at most linearly in t, and that $|\!|\!| \Delta_0 |\!|\!|_{\text{F}} \geq \delta_0$, we have that whenever event $\mathcal{A}_t$ is true,

$$Z(\delta_0) \geq \frac{\delta_0}{|\!|\!| \Delta_0 |\!|\!|_{\text{F}}} Z(|\!|\!| \Delta_0 |\!|\!|_{\text{F}}) \geq \frac{\delta_0}{|\!|\!| \Delta_0 |\!|\!|_{\text{F}}} \langle\!\langle \Delta_0,\ W \rangle\!\rangle \geq 2\delta_0 \sqrt{t\delta_0},$$

where the final inequality uses the second condition in the definition of event $\mathcal{A}_t$. As a consequence, we obtain the following bound on the probabilities of the associated events

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}] \qquad \text{for all } t \geq \delta_0.$$

The entries of W lie in $[-1,1]$, are i.i.d. on and above the diagonal, are zero-mean, and satisfy skew-symmetry. Moreover, the function $W \mapsto Z(t)$ is convex and Lipschitz with parameter t. Consequently, from known concentration bounds (e.g., [144, Theorem 5.9], [213]) for convex Lipschitz functions, we have

$$\mathbb{P}\big[Z(\delta_0) \geq \mathbb{E}[Z(\delta_0)] + \sqrt{t\delta_0}\delta_0\big] \leq 2e^{-c_1 t\delta_0} \qquad \text{for all } t \geq \delta_0. \tag{2.30}$$

By the definition of δ_0, we have $\mathbb{E}[Z(\delta_0)] \leq \delta_0^2 \leq \delta_0\sqrt{t\delta_0}$ for any $t \geq \delta_0$, and consequently

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}\big] \;\leq\; 2e^{-c_1 t\delta_0} \quad \text{for all } t \geq \delta_0.$$

Consequently, either $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq \sqrt{t\delta_0}$, or we have $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} > \sqrt{t\delta_0}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, our basic inequality implies that $\frac{1}{2}|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2 \leq 2|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}\sqrt{t\delta_0}$, and hence $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_0}$ with probability at least $1 - 2e^{-c_1 t\delta_0}$. Putting together the pieces yields that

$$|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq c_0\sqrt{t\delta_0} \tag{2.31}$$

with probability at least $1 - 2e^{-c_1 t\delta_0}$ for every $t \geq \delta_0$.

In order to determine a feasible δ_0 satisfying the critical inequality (2.28), we need to bound the expectation $\mathbb{E}[Z(\delta_0)]$. We do using Dudley's entropy integral and bounding the metric entropies of certain sub-classes of matrices. In particular, the remainder of this section is devoted to proving the following claim:

Lemma 2. *There is a universal constant C such that*

$$\mathbb{E}[Z(t)] \leq C\Big\{n\log^2(n) + t\,\sqrt{n\log n}\Big\}, \tag{2.32}$$

for all $t \in [0, 2n]$.

See the end of this section for a proof of this lemma.

Given this lemma, we see that the critical inequality (2.28) is satisfied with $\delta_0 = C'\sqrt{n}\log n$. Consequently, from our bound (2.31), there are universal positive constants C'' and c_1 such that

$$\frac{|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2}{n^2} \leq C''\frac{\log^2(n)}{n},$$

with probability at least $1 - 2e^{-c_1 n(\log n)^2}$, which completes the proof.

Proof of Lemma 2: It remains to prove Lemma 2, and we do so by using Dudley's entropy integral, as well as some auxiliary results on metric entropy. We use the notation $\log N(\epsilon, \mathbb{C}, \rho)$ to denote the ϵ metric entropy of the class $\mathbb{C}$ in the metric ρ. Our proof requires the following auxiliary lemma:

Lemma 3. *For every $\epsilon > 0$, we have the metric entropy bound*

$$\log N(\epsilon, \mathbb{C}_{\textit{DIFF}}, |||.|||_F) \leq 9\frac{n^2}{\epsilon^2}\big(\log\frac{n}{\epsilon}\big)^2 + 9n\log n.$$

See the end of this section for the proof of this claim. Letting $\mathbb{B}_F(t)$ denote the Frobenius norm ball of radius t, the truncated form of Dudley's entropy integral inequality (e.g., [256, Corollary 2.2.8]) yields that the mean $\mathbb{E}[Z(t)]$ is upper bounded as

$$\begin{aligned}\mathbb{E}[Z(t)]] &\leq c \inf_{\delta\in[0,n]} \Big\{n\delta + \int_{\frac{\delta}{2}}^{t} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}} \cap \mathbb{B}_F(t), |||.|||_{\text{F}})} d\epsilon\Big\} \\ &\leq c\,\Big\{n^{-8} + \int_{\frac{1}{2}n^{-9}}^{t} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}, |||.|||_{\text{F}})} d\epsilon\Big\},\end{aligned} \tag{2.33}$$

where the second step follows by setting $\delta = n^{-9}$, and making use of the set inclusion $(\mathbb{C}_{\text{DIFF}} \cap \mathbb{B}_F(t)) \subseteq \mathbb{C}_{\text{DIFF}}$. For any $\epsilon \geq \frac{1}{2}n^{-9}$, applying Lemma 3 yields the upper bound

$$\sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}, |||.|||_{\text{F}})} \leq c\Big\{\frac{n}{\epsilon}\log\frac{n}{\epsilon} + \sqrt{n\log n}\Big\}.$$

Over the range $\epsilon \geq n^{-9}/2$, we have $\log\frac{n}{\epsilon} \leq c\log n$, and hence

$$\sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}, |||.|||_{\text{F}})} \leq c\Big\{\frac{n}{\epsilon}\log n + \sqrt{n\log n}\Big\}.$$

Substituting this bound into our earlier inequality (2.33) yields

$$\begin{aligned}\mathbb{E}[Z(t)] &\leq c\Big\{n^{-8} + \big(n\log n\big)\log(nt) + t\sqrt{n\log n}\Big\} \\ &\overset{(i)}{\leq} c\Big\{\big(n\log n\big)\log(n^2) + t\sqrt{n\log n}\Big\} \\ &\leq c\Big\{n\log^2(n) + t\sqrt{n\log n}\Big\},\end{aligned}$$

where step (i) uses the upper bound $t \leq 2n$.

The only remaining detail is to prove Lemma 3.

Proof of Lemma 3: We first derive an upper bound on the metric entropy of the class $\mathbb{C}_{\text{BISO}}$ defined previously in equation (2.24). In particular, we do so by relating it to the set of all bivariate monotonic functions on the square $[0,1]\times[0,1]$. Denoting this function class by $\mathcal{F}$, for any matrix $M \in \mathbb{C}_{\text{BISO}}$, we define a function $g_M \in \mathcal{F}$ via

$$g_M(x,y) = M_{\lceil n(1-x)\rceil, \lceil ny\rceil}.$$

In order to handle corner conditions, we set $M_{0,i} = M_{1,i}$ and $M_{i,0} = M_{i,1}$ for all i. With this definition, we have

$$\|g_M\|_2^2 = \int_{x=0}^{1} \int_{y=0}^{1} (g_M(x,y))^2 dxdy = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{i,j}^2 = \frac{1}{n^2} \|\!|M|\!\|_{\mathrm{F}}^2.$$

As a consequence, the metric entropy can be upper bounded as

$$\begin{aligned} \log N(\epsilon, \mathbb{C}_{\text{BISO}}, \|\!|.|\!\|_{\mathrm{F}}) \leq \;& \log N\Big(\frac{\epsilon}{n}, \mathcal{F}, \|.\|_2\Big) \\ \overset{(i)}{\leq} \;& \frac{n^2}{\epsilon^2}\big(\log \frac{n}{\epsilon}\big)^2, \end{aligned} \tag{2.34}$$

where inequality (i) follows from Theorem 1.1 of Gao and Wellner [86].

We now bound the metric entropy of $\mathbb{C}_{\text{DIFF}}$ in terms of the metric entropy of $\mathbb{C}_{\text{BISO}}$. For any $\epsilon > 0$, let $\mathbb{C}_{\text{BISO}}^{\epsilon}$ denote an ϵ-covering set in $\mathbb{C}_{\text{BISO}}$ that satisfies the inequality (2.34). Consider the set

$$\mathbb{C}_{\text{DIFF}}^{\epsilon} := \{\pi_1(M_1) - \pi_2(M_2) \mid \text{for some permutations } \pi_1,\, \pi_2 \text{ and some } M_1, M_2 \in \mathbb{C}_{\text{BISO}}^{\epsilon/2}\}.$$

For any $D \in \mathbb{C}_{\text{DIFF}}$, we can write $D = \pi_1(M_1') - \pi_2(M_2')$ for some permutations π_1 and π_2 and some matrices M_1' and $M_2' \in \mathbb{C}_{\text{BISO}}$. We know there exist matrices $M_1, M_2 \in \mathbb{C}_{\text{BISO}}^{\epsilon/2}$ such that $\|\!|M_1' - M_1|\!\|_{\mathrm{F}} \leq \epsilon/2$ and $\|\!|M_2' - M_2|\!\|_{\mathrm{F}} \leq \epsilon/2$. With these choices, we have $\pi_1(M_1) - \pi_2(M_2) \in \mathbb{C}_{\text{DIFF}}^{\epsilon}$, and moreover

$$\begin{aligned} \|\!|D - (\pi_1(M_1) - \pi_2(M_2))|\!\|_{\mathrm{F}}^2 &\leq 2\|\!|\pi_1(M_1) - \pi_1(M_1')|\!\|_{\mathrm{F}}^2 + 2\|\!|\pi_2(M_2) - \pi_1(M_2')|\!\|_{\mathrm{F}}^2 \\ &\leq \epsilon^2. \end{aligned}$$

Thus the set $\mathbb{C}_{\text{DIFF}}^{\epsilon}$ forms an ϵ-covering set for the class $\mathbb{C}_{\text{DIFF}}$. One can now count the number of elements in this set to find that

$$N(\epsilon, \mathbb{C}_{\text{DIFF}}, \|\!|.|\!\|_{\mathrm{F}}) \leq \big(n! N(\epsilon/2, \mathbb{C}_{\text{BISO}}, \|\!|.|\!\|_{\mathrm{F}})\big)^2.$$

Some straightforward algebraic manipulations yield the claimed result.

Proof of lower bound

We now turn to the proof of the lower bound in Theorem 1. We may assume that the correct row/column ordering is fixed and known to be the identity permutation. Here we are using the fact that revealing the knowledge of this ordering cannot make the estimation problem any harder. Recalling the definition (2.24) of the bivariate isotonic class $\mathbb{C}_{\text{BISO}}$, consider the subclass

$$\mathbb{C}_{\text{SST}}' := \{M \in \mathbb{C}_{\text{BISO}} \mid M_{i,j} = 1 \text{ when } j > i + 1 \text{ and } M_{i,j} = 1 - M_{j,i} \text{ when } j \leq i\}$$

Any matrix M is this subclass can be identified with the vector $q = q(M) \in \mathbb{R}^{n-1}$ with elements $q_i := M_{i,i+1}$. The only constraint imposed on $q(M)$ by the inclusion $M \in \mathbb{C}_{\text{SST}}$ is that $q_i \in [\frac{1}{2}, 1]$ for all $i = 1, \ldots, n-1$.

In this way, we have shown that the difficulty of estimating $M^* \in \mathbb{C}'_{\text{SST}}$ is at least as hard as that of estimating a vector $q \in [\frac{1}{2}, 1]^{n-1}$ based on observing the random vector $Y = \{Y_{1,2}, \ldots, Y_{n-1,n}\}$ with independent coordinates, and such that each $Y_{i,i+1} \sim \text{Ber}(q_i)$. For this problem, it is easy to show that there is a universal constant $c_2 > 0$ such that

$$\inf_{\widehat{q}} \sup_{q \in [\frac{1}{2},1]^{n-1}} \mathbb{E}\Big[\|\widehat{q} - q\|_2^2\Big] \geq \frac{c_2}{2} n,$$

where the infimum is taken over all measurable functions $Y \mapsto \widehat{q}$. Putting together the pieces, we have shown that

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}_{\text{SST}}} \frac{1}{n^2} \mathbb{E}[|\!|\!|\hat{M} - M^*|\!|\!|_{\text{F}}^2] \geq \frac{2}{n^2} \inf_{\widehat{q}} \sup_{q \in [0.5,1]^{n-1}} \mathbb{E}[\|\widehat{q} - q\|_2^2] \geq \frac{c_2}{n},$$

as claimed.

2.6.3 Proof of Theorem 2: Singular Value Thresholding

Recall from equation (2.1) that we can write our observation model as $Y = M^* + W$, where $W \in \mathbb{R}^{n \times n}$ is a zero-mean matrix with entries that are drawn independently (except for the skew-symmetry condition) from the interval $[-1, 1]$.

Proof of upper bound

Our proof of the upper bound hinges upon the following two lemmas.

Lemma 4. *If $\lambda_n \geq 1.01 |\!|\!|W|\!|\!|_{op}$, then*

$$|\!|\!|T_{\lambda_n}(Y) - M^*|\!|\!|_F^2 \leq c \sum_{j=1}^{n} \min\big\{\lambda_n^2, \sigma_j^2(M^*)\big\},$$

where c is a positive universal constant.

Our second lemma is an approximation-theoretic result:

Lemma 5. *For any matrix $M^* \in \mathbb{C}_{SST}$ and any $s \in \{1, 2, \ldots, n-1\}$, we have*

$$\frac{1}{n^2} \sum_{j=s+1}^{n} \sigma_j^2(M^*) \leq \frac{1}{s}.$$

See the end of this section for the proofs of these two auxiliary results.

Based on these two lemmas, it is easy to complete the proof of the theorem. The entries of W are zero-mean with entries in the interval $[-1, 1]$, are i.i.d. on and above the diagonal, and satisfy skew-symmetry. Consequently, we may apply Theorem 3.4 of Chatterjee [44], which guarantees that

$$\mathbb{P}\Big[|\!|\!| W |\!|\!|_{\text{op}} > (2+t)\sqrt{n}\Big] \leq c e^{-f(t) n},$$

where c is a universal constant, and the quantity $f(t)$ is strictly positive for each $t > 0$. Thus, the choice $\lambda_n = 2.1\sqrt{n}$ guarantees that $\lambda_n \geq 1.01 |\!|\!| W |\!|\!|_{\text{op}}$ with probability at least $1 - ce^{-cn}$, as is required for applying Lemma 4. Applying this lemma guarantees that the upper bound

$$|\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2 \leq c \sum_{j=1}^n \min\big\{n, \sigma_j^2(M^*)\big\}$$

hold with probability at least $1 - c_1 e^{-c_2 n}$. From Lemma 5, with probability at least $1 - c_1 e^{-c_2 n}$, we have

$$\frac{1}{n^2} |\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2 \leq c \Big\{\frac{s}{n} + \frac{1}{s}\Big\}$$

for all $s \in \{1, \ldots, n\}$. Setting $s = \lceil \sqrt{n} \rceil$ and performing some algebra shows that

$$\mathbb{P}\Big[\frac{1}{n^2} |\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2 > \frac{c_1}{\sqrt{n}}\Big] \leq c_1 e^{-c_2 n},$$

as claimed. Since $\frac{1}{n^2} |\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2 \leq 1$, we are also guaranteed that

$$\frac{1}{n^2} \mathbb{E}[|\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2] \leq \frac{c_1}{\sqrt{n}} + c_1 e^{-c_2 n} \ \leq \frac{c_1'}{\sqrt{n}}.$$

Proof of Lemma 4 Fix $\delta = 0.01$. Let b be the number of singular values of M^* above $\frac{\delta}{1+\delta}\lambda_n$, and let M_b^* be the version of M^* truncated to its top b singular values. We then have

$$\begin{aligned} |\!|\!| T_{\lambda_n}(Y) - M^* |\!|\!|_{\text{F}}^2 &\leq 2 |\!|\!| T_{\lambda_n}(Y) - M_b^* |\!|\!|_{\text{F}}^2 + 2 |\!|\!| M_b^* - M^* |\!|\!|_{\text{F}}^2 \\ &\leq 2 \operatorname{rank}(T_{\lambda_n}(Y) - M_b^*) |\!|\!| T_{\lambda_n}(Y) - M_b^* |\!|\!|_{\text{op}}^2 + 2 \sum_{j=b+1}^n \sigma_j^2(M^*). \end{aligned}$$

We claim that $T_{\lambda_n}(Y)$ has rank at most b. Indeed, for any $j \geq b+1$, we have

$$\sigma_j(Y) \leq \sigma_j(M^*) + |\!|\!| W |\!|\!|_{\text{op}} \ \leq \ \lambda_n,$$

where we have used the facts that $\sigma_j(M^*) \leq \frac{\delta}{1+\delta}\lambda_n$ for every $j \geq b+1$ and $\lambda_n \geq (1+\delta)|\!|\!|W|\!|\!|_{\text{op}}$. As a consequence we have $\sigma_j(T_{\lambda_n}(Y)) = 0$, and hence $\text{rank}(T_{\lambda_n}(Y) - M^*_b) \leq 2b$. Moreover, we have

$$\begin{aligned} |\!|\!|T_{\lambda_n}(Y) - M^*_b|\!|\!|_{\text{op}} &\leq |\!|\!|T_{\lambda_n}(Y) - Y|\!|\!|_{\text{op}} + |\!|\!|Y - M^*|\!|\!|_{\text{op}} + |\!|\!|M^* - M^*_b|\!|\!|_{\text{op}} \\ &\leq \lambda_n + |\!|\!|W|\!|\!|_{\text{op}} + \frac{\delta}{1+\delta}\lambda_n \\ &\leq 2\lambda_n. \end{aligned}$$

Putting together the pieces, we conclude that

$$|\!|\!|T_{\lambda_n}(Y) - M^*|\!|\!|_{\text{F}}^2 \leq 16b\lambda_n^2 + 2\sum_{j=b+1}^{n} \sigma_j^2(M^*) \overset{(i)}{\leq} C\sum_{j=1}^{n} \min\{\sigma_j^2(M^*), \lambda_n^2\},$$

for some constant[5] C. Here inequality (i) follows since $\sigma_j(M^*) \leq \frac{\delta}{1+\delta}\lambda_n$ whenever $j \geq b+1$ and $\sigma_j(M^*) > \frac{\delta}{1+\delta}\lambda_n$ whenever $j \leq b$.

Proof of Lemma 5 In this proof, we make use of a construction due to Chatterjee [44]. For a given matrix M^*, we can define the vector $t \in \mathbb{R}^n$ of row sums—namely, with entries $t_i = \sum_{j=1}^n M^*_{ij}$ for $i \in [n]$. Using this vector, we can define a rank s approximation M to the original matrix M^* by grouping the rows according to the vector t according to the following procedure:

- Observing that each $t_i \in [0, n]$, let us divide the full interval $[0, n]$ into s groups—say of the form $[0, n/s), [n/s, 2n/s), \ldots [(s-1)n/s, n]$. If t_i falls into the interval α for some $\alpha \in [s]$, we then map row i to the group G_α of indices.
- For each group G_α, we choose a particular row index $k = k(\alpha) \in G_\alpha$ in an arbitrary fashion. For every other row index $i \in G_\alpha$, we set $M_{ij} = M_{kj}$ for all $j \in [n]$.

By construction, the matrix M has at most s distinct rows, and hence rank at most s. Let us now bound the Frobenius norm error in this rank s approximation. Fixing an arbitrary group index $\alpha \in [s]$ and an arbitrary row in $i \in G_\alpha$, we then have

$$\sum_{j=1}^{n}(M^*_{ij} - M_{ij})^2 \leq \sum_{j=1}^{n} |M^*_{ij} - M_{ij}|.$$

By construction, we either have $M^*_{ij} \geq M_{ij}$ for every $j \in [n]$, or $M^*_{ij} \leq M_{ij}$ for every $j \in [n]$. Thus, letting $k \in G_\alpha$ denote the chosen row, we are guaranteed that

$$\sum_{j=1}^{n} |M^*_{ij} - M_{ij}| \leq |t_i - t_k| \leq \frac{n}{s},$$

where we have used the fact the pair (t_i, t_k) must lie in an interval of length at most n/s. Putting together the pieces yields the claim.

[5] To be clear, the precise value of the constant C is determined by δ, which has been fixed as $\delta = 0.01$.

Proof of lower bound

We now turn to the proof of the lower bound in Theorem 2. We split our analysis into two cases, depending on the magnitude of λ_n.

Case 1: First suppose that $\lambda_n \leq \frac{\sqrt{n}}{3}$. In this case, we consider the matrix $M^* := \frac{1}{2}11^T$ in which all items are equally good, so any comparison is simply a fair coin flip. Let the observation matrix $Y \in \{0,1\}^{n\times n}$ be arbitrary. By definition of the singular value thresholding operation, we have $|\!|\!| Y - T_{\lambda_n}(Y) |\!|\!|_{\text{op}} \leq \lambda_n$, and hence the SVT estimator $\widehat{M}_{\lambda_n} = T_{\lambda_n}(Y)$ has Frobenius norm at most

$$|\!|\!| Y - \widehat{M}_{\lambda_n} |\!|\!|_{\text{F}}^2 \leq n\lambda_n^2 \leq \frac{n^2}{9}.$$

Since $M^* \in \{\frac{1}{2}\}^{n\times n}$ and $Y \in \{0,1\}^{n\times n}$, we are guaranteed that $|\!|\!| M^* - Y |\!|\!|_{\text{F}} = \frac{n}{2}$. Applying the triangle inequality yields the lower bound

$$|\!|\!| \widehat{M}_{\lambda_n} - M^* |\!|\!|_{\text{F}} \; \geq \; |\!|\!| M^* - Y |\!|\!|_{\text{F}} - |\!|\!| \widehat{M}_{\lambda_n} - Y |\!|\!|_{\text{F}} \; \geq \frac{n}{2} - \frac{n}{3} \; = \; \frac{n}{6}.$$

Case 2: Otherwise, we may assume that $\lambda_n > \frac{\sqrt{n}}{3}$. Consider the matrix $M^* \in \mathbb{R}^{n\times n}$ with entries

$$[M^*]_{ij} = \begin{cases} 1 & \text{if } i > j \\ \frac{1}{2} & \text{if } i = j \\ 0 & \text{if } i < j. \end{cases} \tag{2.35}$$

By construction, the matrix M^* corresponds to the degenerate case of noiseless comparisons.

Consider the matrix $Y \in \mathbb{R}^{n\times n}$ generated according to the observation model (2.1). (To be clear, all of its off-diagonal entries are deterministic, whereas the diagonal is population with i.i.d. Bernoulli variates.) Our proof requires the following auxiliary result regarding the singular values of Y:

Lemma 6. *The singular values of the observation matrix $Y \in \mathbb{R}^{n\times n}$ generated by the noiseless comparison matrix M^* satisfy the bounds*

$$\frac{n}{4\pi(i+1)} - \frac{1}{2} \leq \sigma_{n-i-1}(Y) \leq \frac{n}{\pi(i-1)} + \frac{1}{2} \qquad \textit{for all integers } i \in [1, \tfrac{n}{6} - 1].$$

We prove this lemma at the end of this section.

Taking it as given, we get that $\sigma_{n-i-1}(Y) \leq \frac{\sqrt{n}}{3}$ for every integer $i \geq 2\sqrt{n}$, and $\sigma_{n-i}(Y) \geq \frac{n}{50i}$ for every integer $i \in [1, \frac{n}{25}]$. It follows that

$$\sum_{i=1}^{n} (\sigma_i(Y))^2 \mathbf{1}\{\sigma_i(Y) \leq \frac{\sqrt{n}}{3}\} \geq \frac{n^2}{2500} \sum_{i=2\sqrt{n}}^{\frac{n}{25}} \frac{1}{i^2} \; \geq \; cn^{\frac{3}{2}},$$

for some universal constant $c > 0$. Recalling that $\lambda_n \geq \frac{\sqrt{n}}{3}$, we have the lower bound $|\!|\!| Y - \widehat{M}_{\lambda_n} |\!|\!|_{\mathrm{F}}^2 \geq cn^{\frac{3}{2}}$. Furthermore, since the observations (apart from the diagonal entries) are noiseless, we have $|\!|\!| Y - M^* |\!|\!|_{\mathrm{F}}^2 \leq \frac{n}{4}$. Putting the pieces together yields the lower bound

$$|\!|\!| \widehat{M}_{\lambda_n} - M^* |\!|\!|_{\mathrm{F}} \; \geq |\!|\!| \widehat{M}_{\lambda_n} - Y |\!|\!|_{\mathrm{F}} - |\!|\!| M^* - Y |\!|\!|_{\mathrm{F}} \geq \; cn^{\frac{3}{4}} - \frac{\sqrt{n}}{2} \; \geq \; c'n^{\frac{3}{4}},$$

where the final step holds when n is large enough (i.e., larger than a universal constant).

Proof of Lemma 6: Instead of working with the original observation matrix Y, it is convenient to work with a transformed version. Define the matrix $\bar{Y} := Y - \text{diag}(Y) + I_n$, so that the matrix $\bar{Y}$ is identical to Y except that all its diagonal entries are set to 1. Using this intermediate object, define the $(n \times n)$ matrix

$$\widetilde{Y} := (\bar{Y}(\bar{Y})^T)^{-1} - e_n e_n^T, \tag{2.36}$$

where e_n denotes the n^{th} standard basis vector. One can verify that this matrix has entries

$$[\widetilde{Y}]_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \text{ or } i = j = n \\ 2 & \text{if } 1 < i = j < n \\ -1 & \text{if } i = j + 1 \text{ or } i = j - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, it is equal to the graph Laplacian[6] of an undirected chain graph on n nodes. Consequently, from standard results in spectral graph theory [26], the eigenvalues of $\widetilde{Y}$ are given by $\{4\sin^2(\frac{\pi i}{n})\}_{i=0}^{n-1}$. Recall the elementary sandwich relationship $\frac{x}{2} \leq \sin x \leq x$, valid for every $x \in [0, \frac{\pi}{6}]$. Using this fact, we are guaranteed that

$$\frac{\pi^2 i^2}{n^2} \leq \lambda_{i+1}(\widetilde{Y}) \leq \frac{4\pi^2 i^2}{n^2} \quad \text{for all integers } i \in [1, \tfrac{n}{6}]. \tag{2.37}$$

We now use this intermediate result to establish the claimed bounds on the singular values of Y. Observe that the matrices $\widetilde{Y}$ and $(\bar{Y}(\bar{Y})^T)^{-1}$ differ only by the rank one matrix $e_n e_n^T$. Standard results in matrix perturbation theory [252] guarantee that a rank-one perturbation can shift the position (in the large-to-small ordering) of any eigenvalue by at most one. Consequently, the eigenvalues of the matrix $(\bar{Y}(\bar{Y})^T)^{-1}$ must be sandwiched as

$$\frac{\pi^2 (i-1)^2}{n^2} \leq \lambda_{i+1}((\bar{Y}(\bar{Y})^T)^{-1}) \leq \frac{4\pi^2 (i+1)^2}{n^2} \qquad \text{for all integers } i \in [1, \tfrac{n}{6} - 1].$$

[6] In particular, the Laplacian of a graph is given by $L = D - A$, where A is the graph adjacency matrix, and D is the diagonal degree matrix.

It follows that the singular values of $\bar{Y}$ are sandwiched as

$$\frac{n}{4\pi(i+1)} \leq \sigma_{n-i-1}(\bar{Y}) \leq \frac{n}{\pi(i-1)} \qquad \text{for all integers } i \in [1, \tfrac{n}{6}-1].$$

Observe that $\bar{Y} - Y$ is a $\{0, \frac{1}{2}\}$-valued diagonal matrix, and hence $|\!|\!|\bar{Y} - Y|\!|\!|_{\text{op}} \leq \frac{1}{2}$. Consequently, we have $\max_{i=1,\ldots,n} |\sigma_i(Y) - \sigma_i(\bar{Y})| \leq \frac{1}{2}$, from which it follows that

$$\frac{n}{4\pi(i+1)} - \frac{1}{2} \leq \sigma_{n-i-1}(Y) \leq \frac{n}{\pi(i-1)} + \frac{1}{2}$$

as claimed.

2.6.4 Proof of Theorem 3: CRL estimator

We now prove the upper bound for the CRL estimator, as stated in Theorem 3. In order to simplify the presentation, we assume without loss of generality that the true permutation of the n items is the identity permutation π_{id}. Let $\pi_{\text{CRL}} = (\pi_1, \ldots, \pi_n)$ denote the permutation obtained at the end of the second step of the CRL estimator. The following lemma proves a useful property of the outcomes of the first two steps.

Lemma 7. *With probability at least $1 - n^{-20}$, the permutation π_{CRL} obtained at the end of the second step of the estimator satisfies:*

$$\max_{i \in [n]} \sum_{\ell=1}^{n} |M^*_{i\ell} - M^*_{\pi_{CRL}(i)\ell}| \leq 2\sqrt{n}(\log n)^2.$$

See the end of this section for a proof of this lemma.

Given Lemma 7, let us complete the proof of the theorem. Let $\widehat{\Pi}$ denote the set of all permutations on n items which satisfy the condition in the statement of Lemma 7. Given that every entry of M^* lies in the interval $[0, 1]$, any permutation $\hat{\pi} \in \widehat{\Pi}$ satisfies

$$\begin{aligned} |\!|\!|M^* - \hat{\pi}(M^*)|\!|\!|_{\text{F}}^2 = \sum_{i \in [n]} \sum_{\ell \in [n]} (M^*_{i\ell} - M^*_{\hat{\pi}(i)\hat{\pi}(\ell)})^2 \;&\leq\; \sum_{i \in [n]} \sum_{\ell \in [n]} \mid M^*_{i\ell} - M^*_{\hat{\pi}(i)\hat{\pi}(\ell)} \mid \\ &\leq \sum_{i \in [n]} \sum_{\ell \in [n]} \mid M^*_{i\ell} - M^*_{\hat{\pi}(i)\ell} \mid + \sum_{i \in [n]} \sum_{\ell \in [n]} \mid M^*_{\hat{\pi}(i)\ell} - M^*_{\hat{\pi}(i)\hat{\pi}(\ell)} \mid, \end{aligned}$$

where the final expression is a result of the triangle inequality. Since M^* satisfies shifted skew-symmetry, we obtain

$$|\!|\!|M^* - \hat{\pi}(M^*)|\!|\!|_{\text{F}}^2 \leq\; 2 \sum_{i \in [n]} \sum_{\ell \in [n]} \mid M^*_{i\ell} - M^*_{\hat{\pi}(i)\ell} \mid . \tag{2.38}$$

From Lemma 7, each item contributes at most $2\sqrt{n}(\log n)^2$ to the error. As a consequence, we have the upper bound

$$|\!|\!| M^* - \hat{\pi}(M^*) |\!|\!|_{\mathrm{F}}^2 \leq 8n\sqrt{n}(\log n)^2. \tag{2.39}$$

Let us now analyze the third step of the CRL estimator. The problem of bivariate isotonic regression refers to estimation of the matrix $M^* \in \mathbb{C}_{\text{SST}}$ when the true underlying permutation of the items *is known* a priori. In our case, the permutation is known only approximately, so that we need also to track the associated approximation error.

Consider any (fixed) permutation $\hat{\pi} \in \widehat{\Pi}$. For clarity, we use $\widehat{M}_{\text{L}}(Y, \hat{\pi})$ to represent the least squares estimator under the permutation $\hat{\pi}$ for the observation matrix Y, that is,

$$\widehat{M}_{\text{L}}(Y, \hat{\pi}) := \underset{M \in \mathbb{C}_{\text{SST}}(\hat{\pi})}{\arg\min} |\!|\!| M - Y |\!|\!|_{\mathrm{F}}^2. \tag{2.40}$$

With this definition, we have the relation $\widehat{M}_{\text{CRL}} = \widehat{M}_{\text{L}}(Y, \pi_{\text{CRL}})$. We cannot bound the error of this estimate directly since the permutation π_{CRL} is not fixed, but dependent on the observed data Y. In order to derive the desired result, we first bound the error of the estimator $\widehat{M}_{\text{L}}(Y, \hat{\pi})$ when the permutation $\hat{\pi}$ is fixed.

Consider any matrix $M^* \in \mathbb{C}_{\text{SST}}(\pi_{\text{id}})$ under the identity permutation. We can then write

$$\begin{aligned} &|\!|\!| \widehat{M}_{\text{L}}(M^* + W, \hat{\pi}) - M^* |\!|\!|_{\mathrm{F}}^2 \\ &\quad = |\!|\!| \widehat{M}_{\text{L}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) + \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^* |\!|\!|_{\mathrm{F}}^2 \\ &\quad \leq 2 |\!|\!| \widehat{M}_{\text{L}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) |\!|\!|_{\mathrm{F}}^2 + 2 |\!|\!| \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^* |\!|\!|_{\mathrm{F}}^2. \end{aligned} \tag{2.41}$$

We separately bound the two terms on the right hand side of expression (2.41). First observe that the least squares step of the estimator $\widehat{M}_{\text{L}}$ (for a given permutation $\hat{\pi}$ in its second argument) is a projection onto the convex set $\mathbb{C}_{\text{SST}}(\hat{\pi})$, and hence we have the deterministic bound

$$|\!|\!| \widehat{M}_{\text{L}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) |\!|\!|_{\mathrm{F}}^2 \leq |\!|\!| M^* - \hat{\pi}(M^*) |\!|\!|_{\mathrm{F}}^2. \tag{2.42a}$$

In addition, we have

$$|\!|\!| \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^* |\!|\!|_{\mathrm{F}}^2 \leq 2 |\!|\!| \widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) - \hat{\pi}(M^*) |\!|\!|_{\mathrm{F}}^2 + 2 |\!|\!| \hat{\pi}(M^*) - M^* |\!|\!|_{\mathrm{F}}^2. \tag{2.42b}$$

At this point, recall the proof of Theorem 1. It follows as a corollary of Theorem 1 that

$$|\!|\!| \widehat{M}_{\text{L}}(M^* + W, \pi_{\text{id}}) - M^* |\!|\!|_{\mathrm{F}}^2 \leq cn(\log n)^3,$$

with probability at least $1 - e^{-3n(\log n)}$. There are three properties of the noise matrix W that are required for the proof of this bound in Theorem 1: (a) $\mathbb{E}[W] = 0$, (b) $\mid W_{ij} \mid \leq 1$ for every pair $i, j \in [n]$, and (c) the entries above the diagonal of W are independent (and those below are governed by skew-symmetry). For any fixed permutation $\hat{\pi}$, the matrix $\hat{\pi}^{-1}(W)$ also satisfies each of these properties. As a result, the same bound applies when the noise matrix is $\hat{\pi}^{-1}(W)$ instead of W:

$$\mathbb{P}\big(|\!|\!|\widehat{M}_{\text{L}}(M^* + \hat{\pi}^{-1}(W), \pi_{\text{id}}) - M^*|\!|\!|_{\text{F}}^2 \leq c_1 n(\log n)^3\big) \geq 1 - e^{-3n(\log n)}.$$

Applying permutation $\hat{\pi}$ to each of the matrices in the above inequality then yields the bound

$$\mathbb{P}\big(|\!|\!|\widehat{M}_{\text{L}}(\hat{\pi}(M^*) + W, \hat{\pi}) - \hat{\pi}(M^*)|\!|\!|_{\text{F}}^2 \leq c_1 n(\log n)^3\big) \geq 1 - e^{-3n\log n}. \tag{2.43}$$

In conjunction, the bounds (2.41), (2.42), and (2.43) imply that for any *fixed* $\hat{\pi} \in \widehat{\Pi}$,

$$\mathbb{P}\Big(|\!|\!|\widehat{M}_{\text{L}}(M^* + W, \hat{\pi}) - M^*|\!|\!|_{\text{F}}^2 \leq 9n\sqrt{n}(\log n)^2\Big) \geq 1 - e^{-3n\log n}. \tag{2.44}$$

Although we are guaranteed that $\pi_{\text{CRL}} \in \widehat{\Pi}$, we cannot apply the bound (2.44) directly to it, since π_{CRL} is a data-dependent quantity. In order to circumvent this issue, we need to obtain a uniform version of the bound (2.44). We do so by applying the union bound over all possible permutations in the set $\widehat{\Pi}$. Since the total number of permutations is at most $n!$, we obtain the bound

$$\mathbb{P}\Big[|\!|\!|\widehat{M}_{\text{CRL}} - M^*|\!|\!|_{\text{F}}^2 \leq 9n\sqrt{n}(\log n)^2 \mid \pi_{\text{CRL}} \in \widehat{\Pi}\Big] \geq 1 - e^{-\log n}.$$

Recalling that Lemma 7 ensures that $\mathbb{P}\big[\pi_{\text{CRL}} \in \widehat{\Pi}\big] \geq 1 - n^{-20}$, we have established the claim.

It remains to prove the auxiliary lemma stated above.

Proof of Lemma 7

We first prove that for any fixed item $i \in [n]$, the inequality holds with probability at least $1 - n^{-22}$. The claimed result then follows via a union bound over all items.

Consider any item $j > i$ such that

$$\sum_{\ell=1}^{n} M^*_{i\ell} - \sum_{\ell=1}^{n} M^*_{j\ell} > 2\sqrt{n}(\log n)^2. \tag{2.45}$$

An application of Bernstein's inequality then gives (see the proof of Theorem 7 in Chapter 3 for a detailed derivation) that

$$\mathbb{P}\big(\sum_{\ell=1}^{n} Y_{j\ell} \geq \sum_{\ell=1}^{n} Y_{i\ell} - \sqrt{n}\log n\big) \leq \frac{1}{n^{23}}. \tag{2.46}$$

Likewise, for any item $j < i$ such that $\sum_{\ell=1}^{n} M^*_{j\ell} - \sum_{\ell=1}^{n} M^*_{i\ell} > 2\sqrt{n}(\log n)^2$, we have $\mathbb{P}\big(\sum_{\ell=1}^{n} Y_{i\ell} \geq \sum_{\ell=1}^{n} Y_{j\ell} - \sqrt{n}\log n\big) \leq \frac{1}{n^{23}}$.

Now consider any $j \geq i$. In order for item i to be located in position j in the total order given by the count and randomize steps of the CRL estimator, there must be at least $(j-i)$ items in the set $\{i+1, \ldots, n\}$ whose row sums are at least $(\sum_{\ell=1}^{n} Y_{i\ell} - \sqrt{n}\log n)$. In particular, there must be at least one item in the set $\{j, \ldots, n\}$ such that its row sum is at least $(\sum_{\ell=1}^{n} Y_{i\ell} - \sqrt{n}\log n)$. It follows from our results above that under the condition (2.45), this event occurs with probability no more than $\frac{1}{n^{21}}$. Likewise when $j \leq i$, thereby proving the claim.

2.6.5 Proof of Theorem 4: High SNR Subclass

We now prove our results on the high SNR subclass of $\mathbb{C}_{\text{SST}}$, in particular establishing a lower bound and then analyzing the two-stage estimator described in Section 2.3.4 so as to obtain the upper bound.

Proof of lower bound

In order to prove the lower bound, we follow the proof of the lower bound of Theorem 1, with the only difference being that the vector $q \in \mathbb{R}^{n-1}$ is restricted to lie in the interval $[\frac{1}{2}+\gamma, 1]^{n-1}$.

Proof of upper bound

Without loss of generality, assume that the true matrix M^* is associated to the identity permutation. Recall that the second step of our procedure involves performing constrained regression over the set $\mathbb{C}_{\text{BISO}}(\widehat{\pi}_{\text{FAS}})$. The error in such an estimate is necessarily of two types: the usual estimation error induced by the noise in our samples, and in addition, some form of approximation error that is induced by the difference between $\widehat{\pi}_{\text{FAS}}$ and the correct identity permutation.

In order to formalize this notion, for any fixed permutation π, consider the constrained least-squares estimator

$$\widehat{M_\pi} \in \underset{M \in \mathbb{C}_{\text{BISO}}(\pi)}{\arg\min} \; |\!|\!| Y - M |\!|\!|_{\text{F}}^2. \tag{2.47}$$

Our first result provides an upper bound on the error matrix $\widehat{M_\pi} - M^*$ that involves both approximation and estimation error terms.

Lemma 8. *There is a universal constant $c_0 > 0$ such that error in the constrained LS estimate* (2.47) *satisfies the upper bound*

$$\frac{|\!|\!|\widehat{M_\pi} - M^*|\!|\!|_F^2}{c_0} \leq \underbrace{|\!|\!| M^* - \pi(M^*) |\!|\!|_F^2}_{\textit{Approx. error}} + \underbrace{n \log^2(n)}_{\textit{Estimation error}} \tag{2.48}$$

with probability at least $1 - c_1 e^{-c_2 n}$.

There are two remaining challenges in the proof. Since the second step of our estimator involves the FAS-minimizing permutation $\widehat{\pi}_{\text{FAS}}$, we cannot simply apply Lemma 8 to it directly. (The permutation $\widehat{\pi}_{\text{FAS}}$ is random, whereas this lemma applies to any fixed permutation). Consequently, we first need to extend the bound (2.48) to one that is uniform over a set that includes $\widehat{\pi}_{\text{FAS}}$ with high probability. Our second challenge is to upper bound the approximation error term $|\!|\!| M^* - \widehat{\pi}_{\text{FAS}}(M^*) |\!|\!|_{\text{F}}^2$ that is induced by using the permutation $\widehat{\pi}_{\text{FAS}}$ instead of the correct identity permutation.

In order to address these challenges, for any constant $c > 0$, define the set

$$\widehat{\Pi}(c) := \{\pi \mid \max_{i \in [n]} |i - \pi(i)| \leq c \log n\}.$$

This set corresponds to permutations that are relatively close to the identity permutation in the sup-norm sense. Our second lemma shows that any permutation in $\widehat{\Pi}(c)$ is "good enough" in the sense that the approximation error term in the upper bound (2.48) is well-controlled:

Lemma 9. *For any $M^* \in \mathbb{C}_{\text{BISO}}$ and any permutation $\pi \in \widehat{\Pi}(c)$, we have*

$$|\!|\!| M^* - \pi(M^*) |\!|\!|_F^2 \leq 2c'' n \log n, \tag{2.49}$$

where c'' is a positive constant that may depend only on c.

Taking these two lemmas as given, let us now complete the proof of Theorem 4. (We return to prove these lemmas at the end of this section.) Braverman and Mossel [23] showed that for the class $\mathbb{C}_{\text{HIGH}}(\gamma)$, there exists a positive constant c—depending on γ but independent of n—such that

$$\mathbb{P}\Big[\widehat{\pi}_{\text{FAS}} \in \widehat{\Pi}(c)\Big] \geq 1 - \frac{c_3}{n^2}. \tag{2.50}$$

From the definition of class $\widehat{\Pi}(c)$, there is a positive constant c' (whose value may depend only on c) such that its cardinality is upper bounded as

$$\text{card}(\widehat{\Pi}(c)) \leq n^{2c' \log n} \overset{(i)}{\leq} e^{.5 c_2 n},$$

where the inequality (i) is valid once the number of items n is larger than some universal constant. Consequently, by combining the union bound with Lemma 8 we conclude that, with probability at least $1 - c_1' e^{-c_2' n} - \frac{c_3}{n^2}$, the error matrix $\widehat{\Delta}_{\text{FAS}} := \widehat{M}_{\widehat{\pi}_{\text{FAS}}} - M^*$ satisfies the upper bound (2.48). Combined with the approximation-theoretic guarantee from Lemma 9, we find that

$$\begin{aligned}\frac{|\!|\!| \widehat{\Delta}_{\text{FAS}} |\!|\!|_{\text{F}}^2}{c_0} &\leq |\!|\!| M^* - \widehat{\pi}_{\text{FAS}}(M^*) |\!|\!|_{\text{F}}^2 + n \log^2(n) \\ &\leq c'' n \log n + + n \log^2(n),\end{aligned}$$

from which the claim follows.

It remains to prove the two auxiliary lemmas, and we do so in the following subsections.

Proof of Lemma 8: The proof of this lemma involves a slight generalization of the proof of the upper bound in Theorem 1 (see Section 2.6.2 for this proof). From the optimality of $\widehat{M}_\pi$ and feasibility of $\pi(M^*)$ for the constrained least-squares program (2.47), we are guaranteed that $\|Y - \widehat{M}_\pi\|_{\text{F}}^2 \leq \|Y - \pi(M^*)\|_{\text{F}}^2$. Introducing the error matrix $\widehat{\Delta}_\pi := \widehat{M}_\pi - M^*$, some algebraic manipulations yield the modified basic inequality

$$\|\widehat{\Delta}_\pi\|_{\text{F}}^2 \leq \|M^* - \pi(M^*)\|_{\text{F}}^2 + 2\langle\!\langle W,\ \widehat{M}_\pi - \pi(M^*)\rangle\!\rangle.$$

Let us define $\widehat{\Delta} := \widehat{M}_\pi - \pi(M^*)$. Further, for each choice of radius $t > 0$, recall the definitions of the random variable $Z(t)$ and event $\mathcal{A}_t$ from equations (2.26) and (2.29), respectively. With these definitions, we have the upper bound

$$\|\widehat{\Delta}_\pi\|_{\text{F}}^2 \leq \|M^* - \pi(M^*)\|_{\text{F}}^2 + 2Z\big(\|\widehat{\Delta}\|_{\text{F}}\big). \tag{2.51}$$

Lemma 3 proved earlier shows that the inequality $\mathbb{E}[Z(\delta_0)] \leq \frac{\delta_0^2}{2}$ is satisfied by $\delta_0 = c\sqrt{n}\log n$. In a manner identical to the proof in Section 2.6.2, one can show that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}\big] \ \leq\ 2e^{-c_1 t\delta_0} \quad \text{for all } t \geq \delta_0.$$

Given these results, we break the next step into two cases depending on the magnitude of $\widehat{\Delta}$. Case I: Suppose $\|\widehat{\Delta}\|_{\text{F}} \leq \sqrt{t\delta_0}$. In this case, we have

$$\begin{aligned}\|\widehat{\Delta}_\pi\|_{\text{F}}^2 &\leq 2\|M^* - \pi(M^*)\|_{\text{F}}^2 + 2\|\widehat{\Delta}\|_{\text{F}}^2\\ &\leq 2\|M^* - \pi(M^*)\|_{\text{F}}^2 + t\delta_0.\end{aligned}$$

Case II: Otherwise, we must have $\|\widehat{\Delta}\|_{\text{F}} > \sqrt{t\delta_0}$. Conditioning on the complement $\mathcal{A}_t^c$, our basic inequality (2.51) implies that

$$\begin{aligned}\|\widehat{\Delta}_\pi\|_{\text{F}}^2 &\leq \|M^* - \pi(M^*)\|_{\text{F}}^2 + 4\|\widehat{\Delta}\|_{\text{F}}\sqrt{t\delta_0}\\ &\leq \|M^* - \pi(M^*)\|_{\text{F}}^2 + \frac{\|\widehat{\Delta}\|_{\text{F}}^2}{8} + 32t\delta_0,\\ &\leq \|M^* - \pi(M^*)\|_{\text{F}}^2 + \frac{2\|\widehat{\Delta}_\pi\|_{\text{F}}^2 + 2\|M^* - \pi(M^*)\|_{\text{F}}^2}{8} + 32t\delta_0,\end{aligned}$$

with probability at least $1 - 2e^{-c_1 t\delta_0}$.

Finally, setting $t = \delta_0 = c\sqrt{n}\log(n)$ in either case and re-arranging yields the bound (2.48).

Proof of Lemma 9: For any matrix M and any value i, let M_i denote its i^{th} row. Also define the clipping function $b : \mathbb{Z} \to [n]$ via $b(x) = \min\{\max\{1, x\}, n\}$. Using this notation, we have

$$\begin{aligned}\|M^* - \pi(M^*)\|_{\text{F}}^2 &= \sum_{i=1}^{n} \|M_i^* - M_{\pi^{-1}(i)}^*\|_2^2\\ &\leq \sum_{i=1}^{n} \max_{0\leq j\leq c\log n}\{\|M_i^* - M_{b(i-j)}^*\|_2^2, \|M_i^* - M_{b(i+j)}^*\|_2^2\},\end{aligned}$$

where we have used the definition of the set $\widehat{\Pi}(c)$ to obtain the final inequality. Since M^* corresponds to the identity permutation, we have $M^*_1 \geq M^*_2 \geq \cdots \geq M^*_n$, where the inequalities are in the pointwise sense. Consequently, we have

$$\begin{aligned}
|\!|\!| M^* - \pi(M^*) |\!|\!|_{\mathrm{F}}^2 &\leq \sum_{i=1}^{n} \max\Big\{ \|M^*_i - M^*_{b(i-c\log n)}\|_2^2, \|M^*_i - M^*_{b(i+c\log n)}\|_2^2 \Big\} \\
&\leq 2 \sum_{i=1}^{n-c\log n} \|M^*_i - M^*_{i+c\log n}\|_2^2.
\end{aligned}$$

One can verify that the inequality $\sum_{i=1}^{k-1}(a_i - a_{i+1})^2 \leq (a_1 - a_k)^2$ holds for all ordered sequences of real numbers $a_1 \geq a_2 \geq \cdots \geq a_k$. As stated earlier, the rows of M^* dominate each other pointwise, and hence we conclude that

$$|\!|\!| M^* - \pi(M^*) |\!|\!|_{\mathrm{F}}^2 \leq 2c \log n \|M^*_1 - M^*_n\|_2^2 \;\leq\; 2cn\log n,$$

which establishes the claim (2.49).

2.6.6 Proof of Theorem 5: Parameter-based models

We now turn to our theorem giving upper and lower bounds on estimating pairwise probability matrices for parameter-based models. Let us begin with a proof of the claimed lower bound.

Lower bound

We prove our lower bound by constructing a set of matrices that are well-separated in Frobenius norm. Using this set, we then use an argument based on Fano's inequality (2.22) to lower bound the minimax risk. Underlying our construction of the matrix collection is a collection of Boolean vectors. For any two Boolean vectors $b, b' \in \{0,1\}^n$, let $D_{\mathrm{H}}(b,b') = \sum_{j=1}^{n} \mathbf{1}[b_j \neq b'_j]$ denote the Hamming distance between them.

Lemma 10. *For any fixed $\zeta \in (0, 1/4)$, there is a collection of Boolean vectors $\{b^1, \ldots, b^\eta\}$ such that*

$$\min\big\{ D_{\mathrm{H}}(b^j, b^k), D_{\mathrm{H}}(b^j, 0) \big\} \geq \lceil \zeta n \rceil \qquad \textit{for all distinct } j \neq k \in \{1,\ldots,\eta\}, \textit{ and} \tag{2.52a}$$

$$\eta \equiv \eta(\zeta) \geq \exp\Big\{ (n-1)\, D_{\mathrm{KL}}(2\alpha \| \frac{1}{2}) \Big\} - 1. \tag{2.52b}$$

See the end of this section for a proof of this lemma.

Given the collection $\{b^j, j \in [\eta(\zeta)]\}$ guaranteed by this lemma, we then define the collection of real vectors $\{w^j, j \in [\eta(\zeta)]\}$ via

$$w^j = \delta\Big(I - \frac{1}{n} 11^T\Big) b^j \qquad \text{for each } j \in [\eta(\zeta)],$$

where $\delta \in (0,1)$ is a parameter to be specified later in the proof. By construction, for each index $j \in [\eta(\zeta)]$, we have $\langle 1,\, w^j\rangle = 0$ and $\|w^j\|_\infty \leq \delta$. Based on these vectors, we then define the collection of matrices $\big\{M^j, j \in [\eta(\zeta)]\big\}$ via

$$[M^k]_{ij} := F([w^k]_i - [w^k]_j).$$

By construction, this collection of matrices is contained within our parameter-based family. We also claim that they are well-separated in Frobenius norm:

Lemma 11. *For any distinct pair $j, k \in [\eta(\zeta)]$, we have*

$$\frac{|\!|\!| M^j - M^k |\!|\!|_F^2}{n^2} \geq \frac{\zeta^2}{4}(F(\delta) - F(0))^2. \tag{2.53}$$

See the end of this section for a proof of this lemma.

In order to apply Fano's inequality, our second requirement is an upper bound on the mutual information $I(Y;J)$, where J is a random index uniformly distributed over the index set $[\eta] = \{1,\ldots,\eta\}$. By Jensen's inequality, we have $I(Y;J) \leq \frac{1}{\binom{\eta}{2}} \sum_{j\neq k} D_{\text{KL}}(\mathbb{P}^j \| \mathbb{P}^k)$, where $\mathbb{P}^j$ denotes the distribution of Y when the true underlying matrix is M^j. Let us upper bound these KL divergences.

For any pair of distinct indices $u, v \in [n]^2$, let x_{uv} be a differencing vector—that is, a vector whose components u and v are set as 1 and -1, respectively, with all remaining components equal to 0. We are then guaranteed that

$$\langle x_{uv},\, w^j\rangle = \delta\langle x_{uv},\, b^j\rangle, \quad \text{and} \quad F(\langle x_{uv},\, w^j\rangle) \in \big\{F(-\delta), F(0), F(\delta)\big\},$$

where $F(\delta) \geq F(0) \geq F(-\delta)$ by construction. Using these facts, we have

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}^j \| \mathbb{P}^k) &\overset{(i)}{\leq} 2 \sum_{u,v \in [n]} \frac{\big(F(\langle x_{uv},\, w^j\rangle) - F(\langle x_{uv},\, w^k\rangle)\big)^2}{\min\{F(\langle x_{uv},\, w^k\rangle), 1 - F(\langle x_{uv},\, w^k\rangle)\}} \\ &\leq 2n^2 \frac{(F(\delta) - F(-\delta))^2}{F(-\delta)} \\ &\leq 8n^2 \frac{(F(\delta) - F(0))^2}{F(-\delta)}, \end{aligned} \tag{2.54}$$

where the bound (i) follows from the elementary inequality $a \log \frac{a}{b} \leq (a-b)\frac{a}{b}$ for any two numbers $a, b \in (0,1)$.

This upper bound on the KL divergence (2.54) and lower bound on the Frobenius norm (2.53), when combined with Fano's inequality (2.22), imply that any estimator $\widehat{M}$ has its worst-case risk over our family lower bounded as

$$\sup_{j \in [\eta(\zeta)]} \frac{1}{n^2} \mathbb{E}\big[|\!|\!| \widehat{M} - M(w^j) |\!|\!|_{\text{F}}^2\big] \geq \frac{1}{8}\zeta^2 (F(\delta) - F(0))^2 \Big(1 - \frac{\frac{8}{F(-\delta)} n^2 (F(\delta) - F(0))^2 + \log 2}{n}\Big).$$

Choosing a value of $\delta > 0$ such that $(F(\delta) - F(0))^2 = \frac{F(-\delta)}{80n}$ gives the claimed result. (Such a value of δ is guaranteed to exist with $F(-\delta) \in [\frac{1}{4}, \frac{1}{2}]$ given our assumption that F is continuous and strictly increasing.)

The only remaining details are to prove Lemmas 10 and 11.

Proof of Lemma 10: The Gilbert-Varshamov bound [90, 258] guarantees the existence of a collection of vectors $\{b^0, \ldots, b^{\bar{T}-1}\}$ contained with the Boolean hypercube $\{0,1\}^n$ such that

$$\begin{aligned} \bar{T} &\geq 2^{n-1} \, \Big(\sum_{\ell=0}^{\lceil \zeta n \rceil - 1} \binom{n-1}{\ell} \Big)^{-1}, \qquad \text{and} \\ D_{\mathrm{H}}(b^j, b^k) &\geq \lceil \zeta n \rceil \qquad \text{for all } j \neq k, \; j,k \in [\bar{T}-1]. \end{aligned}$$

Moreover, their construction allows loss of generality that the all-zeros vector is a member of the set—say $b^0 = 0$. We are thus guaranteed that $D_{\mathrm{H}}(b^j, 0) \geq \lceil \zeta n \rceil$ for all $j \in \{1, \ldots, \bar{T}-1\}$.

Since $n \geq 2$ and $\alpha \in (0, \frac{1}{4})$, we have $\frac{\lceil \alpha n \rceil - 1}{n-1} \leq 2\alpha \leq \frac{1}{2}$. Applying standard bounds on the tail of the binomial distribution yields

$$\frac{1}{2^{n-1}} \sum_{\ell=0}^{\lceil \zeta n \rceil - 1} \binom{n-1}{\ell} \leq \exp\Big(-(n-1) D_{\mathrm{KL}}\big(\frac{\lceil \alpha n \rceil - 1}{n-1} \| \frac{1}{2}\big)\Big) \leq \exp\Big(-(n-1) D_{\mathrm{KL}}(2\alpha \| \frac{1}{2})\Big).$$

Consequently, the number of non-zero code words $\eta := \bar{T} - 1$ is at least

$$\eta(\zeta) := \exp\Big((n-1) D_{\mathrm{KL}}(2\alpha \| \frac{1}{2})\Big) - 1.$$

Thus, the collection $\{b^1, \ldots, b^\eta\}$ has the desired properties.

Proof of Lemma 11: By definition of the matrix ensemble, we have

$$|\!|\!| M(w^j) - M(w^k) |\!|\!|_{\mathrm{F}}^2 = \sum_{u,v \in [n]} (F(\langle x_{uv}, \, w^j \rangle) - F(\langle x_{uv}, \, w^k \rangle))^2. \tag{2.55}$$

By construction, the Hamming distances between the triplet of vectors $\{w^j, w^k, 0\}$ are lower bounded $D_{\mathrm{H}}(w^j, 0) \geq \zeta n$, $D_{\mathrm{H}}(w^k, 0) \geq \zeta n$ and $D_{\mathrm{H}}(w^j, w^k) \geq \zeta n$. We claim that this implies that

$$\operatorname{card}\Big\{ u \neq v \in [n]^2 \;\mid\; \langle x_{uv}, \, w^j \rangle \neq \langle x_{uv}, \, w^k \rangle \Big\} \geq \frac{\zeta^2}{4} \, n^2. \tag{2.56}$$

Taking this auxiliary claim as given for the moment, applying it to Equation (2.55) yields the lower bound $|\!|\!| M(w^1) - M(w^2) |\!|\!|_{\mathrm{F}}^2 \;\geq\; \frac{1}{4} \zeta^2 n^2 (F(\delta) - F(0))^2$, as claimed.

It remains to prove the auxiliary claim (2.56). We relabel $j = 1$ and $k = 2$ for simplicity in notation. For $(y, z) \in \{0, 1\} \times \{0, 1\}$, let set $\mathcal{I}_{yz} \subseteq [n]$ denote the set of indices on which w^1 takes value y and w^2 takes value z. We then split the proof into two cases:

Case 1: Suppose $\mid \mathcal{I}_{00} \cup \mathcal{I}_{11} \mid \geq \frac{\zeta n}{2}$. The minimum distance condition $D_{\mathrm{H}}(w^1, w^2) \geq \zeta n$ implies that $\mid \mathcal{I}_{01} \cup \mathcal{I}_{10} \mid \geq \zeta n$. For any $i \in \mathcal{I}_{00} \cup \mathcal{I}_{11}$ and any $j \in \mathcal{I}_{01} \cup \mathcal{I}_{10}$, it must be that $\langle x_{uv},\, w^1 \rangle \neq \langle x_{uv},\, w^2 \rangle$. Thus there are at least $\frac{\zeta^2}{2} n^2$ such pairs of indices.

Case 2: Otherwise, we may assume that $\mid \mathcal{I}_{00} \cup \mathcal{I}_{11} \mid < \frac{\zeta n}{2}$. This condition, along with the minimum Hamming weight conditions $D_{\mathrm{H}}(w^1, 0) \geq \zeta n$ and $D_{\mathrm{H}}(w^2, 0) \geq \zeta n$, gives $\mathcal{I}_{10} \geq \frac{\zeta n}{2}$ and $\mathcal{I}_{01} \geq \frac{\zeta n}{2}$. For any $i \in \mathcal{I}_{01}$ and any $j \in \mathcal{I}_{10}$, it must be that $\langle x_{uv},\, w^1 \rangle \neq \langle x_{uv},\, w^2 \rangle$. Thus there are at least $\frac{\zeta^2}{4} n^2$ such pairs of indices.

Upper bound

In our earlier work [220, Theorem 2b] we prove that when F is strongly log-concave and twice differentiable, then there is a universal constant c_1 such that the maximum likelihood estimator $\widehat{w}_{\mathrm{ML}}$ has mean squared error at most

$$\sup_{w^* \in [-1,1]^n, \langle w^*,\, 1 \rangle = 0} \mathbb{E}[\|\widehat{w}_{\mathrm{ML}} - w^*\|_2^2] \leq c_1. \tag{2.57}$$

Moreover, given the log-concavity assumption, the MLE is computable in polynomial-time. Let $M(\widehat{w}_{\mathrm{ML}})$ and $M(w^*)$ denote the pairwise comparison matrices induced, via Equation (2.4), by $\widehat{w}_{\mathrm{ML}}$ and w^*. It suffices to bound the Frobenius norm $|\!|\!| M(\widehat{w}_{\mathrm{ML}}) - M(w^*) |\!|\!|_{\mathrm{F}}$.

Consider any pair of vectors w^1 and w^2 that lie in the hypercube $[-1, 1]^n$. For any pair of indices $(i, j) \in [n]^2$, we have

$$((M(w^1))_{ij} - (M(w^2))_{ij})^2 = (F(w_i^1 - w_j^1) - F(w_i^2 - w_j^2))^2 \;\leq\; \zeta^2((w_i^1 - w_j^1) - (w_i^2 - w_j^2))^2,$$

where we have defined $\zeta := \max_{z \in [-1,1]} F'(z)$. Putting together the pieces yields

$$|\!|\!| M(w^1) - M(w^2) |\!|\!|_{\mathrm{F}}^2 \leq \zeta^2 (w^1 - w^2)^T (nI - 11^T)(w^1 - w^2) \;=\; n\zeta^2 \|w^1 - w^2\|_2^2. \tag{2.58}$$

Applying this bound with $w^1 = \widehat{w}_{\mathrm{ML}}$ and $w^2 = w^*$ and combining with the bound (2.57) yields the claim.

2.6.7 Proof of Theorem 6: Partial observations

We now turn to the proof of Theorem 6, which characterizes the behavior of different estimators for the partially observed case.

Proof of part (a)

In this section, we prove the lower and upper bounds stated in part (a).

Proof of lower bound: We begin by proving the lower bound in equation (2.20a). The Gilbert-Varshamov bound [90, 258] guarantees the existence of a set of vectors $\{b^1, \ldots, b^\eta\}$ in the Boolean cube $\{0,1\}^{\frac{n}{2}}$ with cardinality at least $\eta := 2^{cn}$ such that

$$D_{\rm H}(b^j, b^k) \geq \lceil 0.1n \rceil \qquad \text{for all distinct pairs } j, k \in [\eta] := \{1, \ldots, \eta\}.$$

Fixing some $\delta \in (0, \frac{1}{4})$ whose value is to be specified later, for each $k \in [\eta]$, we define a matrix $M^k \in \mathbb{C}_{\rm SST}$ with entries

$$[M^k]_{uv} = \begin{cases} \frac{1}{2} + \delta & \text{if } u \leq \frac{n}{2}, [b^k]_u = 1 \text{ and } v \geq \frac{n}{2} \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

for every pair of indices $u \leq v$. We complete the matrix by setting $[M^k]_{vu} = 1 - [M^k]_{uv}$ for all indices $u > v$.

By construction, for each distinct pair $j, k \in [\eta]$, we have the lower bound

$$|\!|\!| M^j - M^k |\!|\!|_{\rm F}^2 = n\delta^2 \|b^j - b^k\|_2^2 \; \geq \; c_0 n^2 \delta^2.$$

Let $\mathbb{P}^j$ and $\mathbb{P}^j_{uv}$ denote (respectively) the distributions of the matrix Y and entry Y_{uv} when the underlying matrix is M^j. Since the entries of Y are generated independently, we have $D_{\rm KL}(\mathbb{P}^j \| \mathbb{P}^k) = \sum\limits_{1 \leq u < v \leq n} D_{\rm KL}(\mathbb{P}^j_{uv} \| \mathbb{P}^k_{uv})$. The matrix entry Y_{uv} is generated according to the model

$$Y_{uv} = \begin{cases} 1 & \text{w.p. } p_{\rm obs} M^*_{uv} \\ 0 & \text{w.p. } p_{\rm obs}(1 - M^*_{uv}) \\ \text{not observed} & \text{w.p. } 1 - p_{\rm obs}. \end{cases}$$

Consequently, the KL divergence can be upper bounded as

$$\begin{aligned} & D_{\rm KL}(\mathbb{P}^j_{uv} \| \mathbb{P}^k_{uv}) \\ & = p_{\rm obs}\Big(M^j_{uv} \log \frac{M^j_{uv}}{M^k_{uv}} + (1 - M^j_{uv}) \log \frac{(1 - M^j_{uv})}{(1 - M^k_{uv})} \Big) \\ & \leq p_{\rm obs}\Big\{ M^j_{uv} \big(\frac{M^j_{uv} - M^k_{uv}}{M^k_{uv}} \big) + (1 - M^j_{uv}) \big(\frac{M^k_{uv} - M^j_{uv}}{1 - M^k_{uv}} \big) \Big\} & \text{(2.59a)} \\ & = p_{\rm obs} \; \frac{(M^j_{uv} - M^k_{uv})^2}{M^k_{uv}(1 - M^k_{uv})} & \text{(2.59b)} \\ & \leq 16 p_{\rm obs} \; (M^j_{uv} - M^k_{uv})^2, & \text{(2.59c)} \end{aligned}$$

where inequality (2.59a) follows from the fact that $\log(t) \leq t - 1$ for all $t > 0$; and inequality (2.59c) follows since the numbers $\{M^j_{uv}, M^k_{uv}\}$ both lie in the interval $[\frac{1}{4}, \frac{3}{4}]$. Putting together the pieces, we conclude that

$$D_{\rm KL}(\mathbb{P}^j \| \mathbb{P}^k) \leq c_1 p_{\rm obs} |\!|\!| M^j - M^k |\!|\!|_{\rm F}^2 \; \leq c_1' p_{\rm obs} n^2 \delta^2.$$

Thus, applying Fano's inequality (2.22) to the packing set $\{M^1, \ldots, M^\eta\}$ yields that any estimator $\widehat{M}$ has mean squared error lower bounded by

$$\sup_{k \in [\eta]} \frac{1}{n^2} \mathbb{E}[\|\!|\widehat{M} - M^k\|\!|_{\mathrm{F}}^2] \geq c_0 \delta^2 \Big(1 - \frac{c_1' p_{\mathrm{obs}} n^2 \delta^2 + \log 2}{cn}\Big).$$

Finally, choosing $\delta^2 = \frac{c_2}{2c_1 p_{\mathrm{obs}} n}$ yields the lower bound $\sup_{k \in [\eta]} \frac{1}{n^2} \mathbb{E}[\|\!|\widehat{M} - M^k\|\!|_{\mathrm{F}}^2] \geq c_3 \frac{1}{n p_{\mathrm{obs}}}$. Note that in order to satisfy the condition $\delta \leq \frac{1}{4}$, we must have $p_{\mathrm{obs}} \geq \frac{16 c_2}{2 c_1 n}$.

Proof of upper bound: For this proof, recall the linearized form of the observation model given in equations (2.19a), (2.21a), and (2.21b). We begin by introducing some additional notation. Letting Π denote the set of all permutations of n items. For each $\pi \in \Pi$, we define the set

$$\pi(\mathbb{C}_{\mathrm{BISO}}) := \big\{M \in [0,1]^{n \times n} \mid M_{k\ell} \geq M_{ij} \text{ whenever } \pi(k) \leq \pi(i) \text{ and } \pi(\ell) \geq \pi(j)\big\},$$

corresponding to the subset of permutation-based SST matrices that are faithful to the permutation π. We then define the estimator $M_\pi \in \underset{M \in \pi(\mathbb{C}_{\mathrm{BISO}})}{\arg\min} \|\!|Y' - M\|\!|_{\mathrm{F}}^2$, in terms of which the least squares estimator (2.19b) can be rewritten as

$$\widehat{M} \in \underset{\pi \in \Pi}{\arg\min} \|\!|Y' - M_\pi\|\!|_{\mathrm{F}}^2.$$

Define a set of permutations $\Pi' \subseteq \Pi$ as

$$\Pi' := \{\pi \in \Pi \mid \|\!|Y' - M_\pi\|\!|_{\mathrm{F}}^2 \leq \|\!|Y' - M^*\|\!|_{\mathrm{F}}^2\}.$$

Note that the set Π' is guaranteed to be non-empty since the permutation corresponding to $\widehat{M}$ always lies in Π'. We claim that for any $\pi \in \Pi'$, we have

$$\mathbb{P}\big(\|\!|M_\pi - M^*\|\!|_{\mathrm{F}}^2 \leq c_1 \frac{n}{p_{\mathrm{obs}}} \log^2 n\big) \geq 1 - e^{-3n \log n}, \tag{2.60}$$

for some positive universal constant c_1. Given this bound, since there are at most $e^{n \log n}$ permutations in the set Π', a union bound over all these permutations applied to (2.60) yields

$$\mathbb{P}\Big(\max_{\pi \in \Pi'} \|\!|M_\pi - M^*\|\!|_{\mathrm{F}}^2 > c_1 \frac{n}{p_{\mathrm{obs}}} \log^2 n\Big) \leq e^{-2n \log n}.$$

Since $\widehat{M}$ is equal to M_π for some $\pi \in \Pi'$, this tail bound yields the claimed result.

The remainder of our proof is devoted to proving the bound (2.60). By definition, any permutation $\pi \in \Pi'$ must satisfy the inequality

$$\|\!|Y - M_\pi\|\!|_{\mathrm{F}}^2 \leq \|\!|Y - M^*\|\!|_{\mathrm{F}}^2.$$

Letting $\widehat{\Delta}_\pi := M_\pi - M^*$ denote the error matrix, and using the linearized form (2.21a) of the observation model, some algebraic manipulations yield the basic inequality

$$\frac{1}{2}|||\widehat{\Delta}_\pi|||_{\mathrm{F}}^2 \leq \frac{1}{p_{\mathrm{obs}}}\langle\langle W',\ \widehat{\Delta}_\pi\rangle\rangle. \tag{2.61}$$

Now consider the set of matrices

$$\mathbb{C}_{\text{DIFF}}(\pi) := \Big\{\alpha(M - M^*) \mid M \in \pi(\mathbb{C}_{\text{BISO}}),\ \alpha \in [0,1]\Big\}, \tag{2.62}$$

and note that $\mathbb{C}_{\text{DIFF}}(\pi) \subseteq [-1,1]^{n\times n}$. (To be clear, the set $\mathbb{C}_{\text{DIFF}}(\pi)$ also depends on the value of M^*, but considering M^* as fixed, we omit this dependence from the notation for brevity.) For each choice of radius $t > 0$, define the random variable

$$Z_\pi(t) := \sup_{\substack{D\in\mathbb{C}_{\text{DIFF}}(\pi),\\ |||D|||_{\mathrm{F}}\leq t}} \frac{1}{p_{\mathrm{obs}}}\langle\langle D,\ W'\rangle\rangle. \tag{2.63}$$

Using the basic inequality (2.61), the Frobenius norm error $|||\widehat{\Delta}_\pi|||_{\mathrm{F}}$ then satisfies the bound

$$\frac{1}{2}|||\widehat{\Delta}_\pi|||_{\mathrm{F}}^2 \leq \frac{1}{p_{\mathrm{obs}}}\langle\langle W',\ \widehat{\Delta}_\pi\rangle\rangle\ \leq\ Z_\pi\big(|||\widehat{\Delta}_\pi|||_{\mathrm{F}}\big). \tag{2.64}$$

Thus, in order to obtain a high probability bound, we need to understand the behavior of the random quantity $Z_\pi(t)$.

One can verify that the set $\mathbb{C}_{\text{DIFF}}(\pi)$ is star-shaped, meaning that $\alpha D \in \mathbb{C}_{\text{DIFF}}(\pi)$ for every $\alpha \in [0,1]$ and every $D \in \mathbb{C}_{\text{DIFF}}(\pi)$. Using this star-shaped property, we are guaranteed that there is a non-empty set of scalars $\delta_0 > 0$ satisfying the critical inequality

$$\mathbb{E}[Z_\pi(\delta_0)] \leq \frac{\delta_0^2}{2}. \tag{2.65}$$

Our interest is in an upper bound to the smallest (strictly) positive solution δ_0 to the critical inequality (2.65), and moreover, our goal is to show that for every $t \geq \delta_0$, we have $|||\widehat{\Delta}|||_{\mathrm{F}} \leq c\sqrt{t\delta_0}$ with high probability.

For each $t > 0$, define the "bad" event

$$\mathcal{A}_t = \big\{\exists \Delta \in \mathbb{C}_{\text{DIFF}}(\pi) \mid |||\Delta|||_{\mathrm{F}} \geq \sqrt{t\delta_0} \quad \text{and} \quad \frac{1}{p_{\mathrm{obs}}}\langle\langle\Delta,\ W'\rangle\rangle \geq 2|||\Delta|||_{\mathrm{F}}\sqrt{t\delta_0}\big\}. \tag{2.66}$$

Using the star-shaped property of $\mathbb{C}_{\text{DIFF}}(\pi)$, it follows by a rescaling argument that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z_\pi(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}] \qquad \text{for all } t \geq \delta_0.$$

The following lemma helps control the behavior of the random variable $Z_\pi(\delta_0)$.

Lemma 12. *For any $\delta > 0$, the mean of $Z_\pi(\delta)$ is bounded as*

$$\mathbb{E}[Z_\pi(\delta)] \leq c_1 \frac{n}{p_{\text{obs}}} \log^2 n,$$

and for every $u > 0$, its tail probability is bounded as

$$\mathbb{P}\Big(Z_\pi(\delta) > \mathbb{E}[Z_\pi(\delta)] + u\Big) \leq \exp\Big(\frac{-c_0 u^2 p_{\text{obs}}}{\delta^2 + \mathbb{E}[Z_\pi(\delta)] + u}\Big),$$

where c_1 and c_0 are positive universal constants.

See the end of this section for a proof of this lemma.

From this lemma, we have the tail bound

$$\mathbb{P}\Big(Z_\pi(\delta_0) > \mathbb{E}[Z_\pi(\delta_0)] + \delta_0\sqrt{t\delta_0}\Big) \leq \exp\Big(\frac{-c_0(\delta_0\sqrt{t\delta_0})^2 p_{\text{obs}}}{\delta_0^2 + \mathbb{E}[Z_\pi(\delta_0)] + (\delta_0\sqrt{t\delta_0})}\Big), \quad \text{for all } t \geq \delta_0.$$

By the definition of δ_0 in equation (2.65), we have $\mathbb{E}[Z(\delta_0)] \leq \delta_0^2 \leq \delta_0\sqrt{t\delta_0}$ for any $t \geq \delta_0$, and consequently

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}] \;\leq\; \exp\Big(\frac{-c_0(\delta_0\sqrt{t\delta_0})^2 p_{\text{obs}}}{3\delta_0\sqrt{t\delta_0}}\Big), \quad \text{for all } t \geq \delta_0.$$

Consequently, either $|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}} \leq \sqrt{t\delta_0}$, or we have $|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}} > \sqrt{t\delta_0}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, our basic inequality implies that $\frac{1}{2}|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}}^2 \leq 2|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}}\sqrt{t\delta_0}$ and hence $|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_0}$. Putting together the pieces yields that

$$\mathbb{P}\big(|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_0}\big) \geq 1 - \exp\big(-c_0'\delta_0\sqrt{t\delta_0}p_{\text{obs}}\big), \quad \text{for all } t \geq \delta_0. \tag{2.67}$$

Finally, from the bound on the expected value of $Z_\pi(t)$ in Lemma 12, we see that the critical inequality (2.65) is satisfied for $\delta_0 = \sqrt{\frac{c_1 n}{p_{\text{obs}}}}\log n$. Setting $t = \delta_0 = \sqrt{\frac{c_1 n}{p_{\text{obs}}}}\log n$ in (2.67) yields

$$\mathbb{P}\Big(|\!|\!|\widehat{\Delta}_\pi|\!|\!|_{\text{F}} \leq 4\frac{c_1 n}{p_{\text{obs}}}\log^2 n\Big) \geq 1 - \exp\Big(-3n\log n\Big), \tag{2.68}$$

for some universal constant $c_1 > 0$, thus proving the bound (2.60).

It remains to prove Lemma 12.

Proof of Lemma 12 Bounding $\mathbb{E}[Z_\pi(\delta)]$: We establish an upper bound on $\mathbb{E}[Z_\pi(\delta)]$ by using Dudley's entropy integral, as well as some auxiliary results on metric entropy. We use the notation $\log N(\epsilon, \mathbb{C}, \rho)$ to denote the ϵ metric entropy of the class $\mathbb{C}$ in the metric ρ. Introducing the random variable $\widetilde{Z}_\pi := \sup_{D \in \mathbb{C}_{\text{DIFF}}(\pi)} \langle\!\langle D, \; W' \rangle\!\rangle$, note that we have $\mathbb{E}[Z_\pi(\delta)] \leq \frac{1}{p_{\text{obs}}} \mathbb{E}[\widetilde{Z}_\pi]$. The truncated form of Dudley's entropy integral inequality yields

$$\mathbb{E}[\widetilde{Z}_\pi] \leq c \Big\{ n^{-8} + \int_{\frac{1}{2}n^{-9}}^{2n} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(\pi), |\!|\!|.|\!|\!|_{\text{F}})} d\epsilon \Big\}, \tag{2.69}$$

where we have used the fact that the diameter of the set $\mathbb{C}_{\text{DIFF}}(\pi)$ is at most $2n$ in the Frobenius norm.

From our earlier bound (2.34), we are guaranteed that for each $\epsilon > 0$, the metric entropy is upper bounded as

$$\log N\Big(\epsilon, \{\alpha M \mid M \in \mathbb{C}_{\text{BISO}}, \alpha \in [0,1]\}, |\!|\!| \cdot |\!|\!|_{\text{F}}\Big) \leq 8 \frac{n^2}{\epsilon^2} \big(\log \frac{n}{\epsilon}\big)^2.$$

Consequently, we have

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(\pi), |\!|\!|.|\!|\!|_{\text{F}}) \leq 16 \frac{n^2}{\epsilon^2} \big(\log \frac{n}{\epsilon}\big)^2.$$

Substituting this bound on the metric entropy of $\mathbb{C}_{\text{DIFF}}(\pi)$ and the inequality $\epsilon \geq \frac{1}{2}n^{-9}$ into the Dudley bound (2.69) yields

$$\mathbb{E}[\widetilde{Z}_\pi] \leq cn(\log n)^2.$$

The inequality $\mathbb{E}[Z_\pi(\delta)] \leq \frac{1}{p_{\text{obs}}} \mathbb{E}[\widetilde{Z}_\pi]$ then yields the claimed result.

Bounding the tail probability of $Z_\pi(\delta)$: In order to establish the claimed tail bound, we use a Bernstein-type bound on the supremum of empirical processes due to Klein and Rio [129, Theorem 1.1c], which we state in a simplified form here.

Lemma 13. *Let $X := (X_1, \ldots, X_m)$ be any sequence of zero-mean, independent random variables, each taking values in $[-1, 1]$. Let $\mathcal{V} \subset [-1,1]^m$ be any measurable set of m-length vectors. Then for any $u > 0$, the supremum $X^\dagger = \sup_{v \in \mathcal{V}} \langle X, \; v \rangle$ satisfies the upper tail bound*

$$\mathbb{P}\big(X^\dagger > \mathbb{E}[X^\dagger] + u\big) \leq \exp\Big(\frac{-u^2}{2 \sup_{v \in \mathcal{V}} \mathbb{E}[\langle v, \; X \rangle^2] + 4\mathbb{E}[X^\dagger] + 3u}\Big).$$

We now invoke Lemma 13 with the choices $\mathcal{V} = \mathbb{C}_{\text{DIFF}}(\pi) \cap \mathbb{B}(\delta)$, $m = (n \times n)$, $X = W'$, and $X^\dagger = p_{\text{obs}} Z_\pi(\delta)$. The matrix W' has zero-mean entries belonging to the interval $[-1, +1]$, and are independent on and above the diagonal (with the entries below determined by the skew-symmetry condition). Then we have $\mathbb{E}[X^\dagger] \leq p_{\text{obs}}\mathbb{E}[Z_\pi(\delta)]$ and $\mathbb{E}[\langle\!\langle D,\ W' \rangle\!\rangle^2] \leq 4p_{\text{obs}}|\!|\!|D|\!|\!|_{\text{F}}^2 \leq 4p_{\text{obs}}\delta^2$ for every $D \in \mathcal{V}$. With these assignments, and some algebraic manipulations, we obtain that for every $u > 0$,

$$\mathbb{P}\Big[Z_\pi(\delta) > \mathbb{E}[Z_\pi(\delta)] + u\Big] \leq \exp\Big(\frac{-u^2 p_{\text{obs}}}{8\delta^2 + 4\mathbb{E}[Z_\pi(\delta)] + 3u}\Big),$$

as claimed.

Proof of part (b)

In order to prove the bound (2.20b), we analyze the SVT estimator $T_{\lambda_n}(Y')$ with the threshold $\lambda_n = 3\sqrt{\frac{n}{p_{\text{obs}}}}$. Naturally then, our analysis is similar to that of complete observations case from Section 2.6.3. Recall our formulation of the problem in terms of the observation matrix Y' along with the noise matrix W' from equations (2.19a), (2.21a) and (2.21b). The result of Lemma 4 continues to hold in this case of partial observations, translated to this setting. In particular, if $\lambda_n \geq \frac{1.01}{p_{\text{obs}}}|\!|\!|W'|\!|\!|_{\text{op}}$, then

$$|\!|\!|T_{\lambda_n}(Y') - M^*|\!|\!|_{\text{F}}^2 \leq c_1 \sum_{j=1}^{n} \min\big\{\lambda_n^2, \sigma_j^2(M^*)\big\},$$

where $c_1 > 0$ is a universal constant.

We now upper bound the operator norm of the noise matrix W'. Define a $(2n \times 2n)$ matrix

$$W'' = \frac{1}{\sqrt{p_{\text{obs}}}} \begin{bmatrix} 0 & W' \\ (W')^T & 0 \end{bmatrix}.$$

From equation (2.21b) and the construction above, we have that the matrix W'' is symmetric, with mutually independent entries above the diagonal that have a mean of zero, a variance upper bounded by 1, and entries bounded in absolute value by $\frac{\sqrt{n}}{\sqrt{c_4}\log^{3.5} n}$. Consequently, known results in random matrix theory (e.g., see [44, Theorem 3.4] or [250, Theorem 2.3.21]) yield the bound $|\!|\!|W''|\!|\!|_{\text{op}} \leq 2.01\sqrt{2n}$ with probability at least $1 - n^{-c_2}$, for some universal constant $c_2 > 1$. One can also verify that $|\!|\!|W''|\!|\!|_{\text{op}} = \frac{1}{\sqrt{p_{\text{obs}}}}|\!|\!|W'|\!|\!|_{\text{op}}$, thereby yielding the bound

$$\mathbb{P}\Big[|\!|\!|W'|\!|\!|_{\text{op}} > 2.01\sqrt{2np_{\text{obs}}}\Big] \leq n^{-c_2}.$$

With our choice $\lambda_n = 3\sqrt{\frac{n}{p_{\text{obs}}}}$, the event $\{\lambda_n \geq \frac{1.01}{p_{\text{obs}}}|\!|\!|W'|\!|\!|_{\text{op}}\}$ holds with probability at least $1 - n^{-c_2}$. Conditioned on this event, the approximation-theoretic result from Lemma 5 gives

$$\frac{1}{n^2}|\!|\!|T_{\lambda_n}(Y') - M^*|\!|\!|_{\text{F}}^2 \leq c\Big(\frac{s\lambda_n^2}{n^2} + \frac{1}{s}\Big)$$

with probability at least $1 - n^{-c_2}$. Substituting $\lambda_n = 3\sqrt{\frac{n}{p_{\text{obs}}}}$ in this bound and setting $s = \sqrt{p_{\text{obs}} n}$ yields the claimed result.

Proof of part (c)

As in our of proof of the fully observed case from Section 2.6.6, we consider the two-stage estimator based on first computing the MLE $\widehat{w}_{\mathrm{ML}}$ of w^* from the observed data, and then constructing the matrix estimate $M(\widehat{w}_{\mathrm{ML}})$ via Equation (2.4). Let us now upper bound the mean-squared error associated with this estimator.

Our observation model can be (re)described in the following way. Consider an Erdős-Rényi graph on n vertices with each edge drawn independently with a probability p_{obs}. For each edge in this graph, we obtain one observation of the pair of vertices at the end-points of that edge. Let L be the (random) Laplacian matrix of this graph, that is, $L = D - A$ where D is an $(n \times n)$ diagonal matrix with $[D]_{ii}$ being the degree of item i in the graph (equivalently, the number of pairwise comparison observations that involve item i) and A is the $(n \times n)$ adjacency matrix of the graph. Let $\lambda_2(L)$ denote the second largest eigenvalue of L. From Theorem 2(b) of our paper [220] on estimating parameter-based models,[7] for this graph, there is a universal constant c_1 such that the maximum likelihood estimator $\widehat{w}_{\mathrm{ML}}$ has mean squared error upper bounded as

$$\mathbb{E}[\|\widehat{w}_{\mathrm{ML}} - w^*\|_2^2 \mid L] \leq c_1 \frac{n}{\lambda_2(L)}.$$

The estimator $\widehat{w}_{\mathrm{ML}}$ is computable in a time polynomial in n.

Since $p_{\mathrm{obs}} \geq c_0 \frac{(\log n)^2}{n}$, known results on the eigenvalues of random graphs [48, 131, 184] imply that

$$\mathbb{P}\Big[\lambda_2(L) \geq c_2 n p_{\mathrm{obs}}\Big] \geq 1 - \frac{1}{n^4} \tag{2.70}$$

for a universal constant c_2 (that may depend on c_0). As shown earlier in Equation (2.58), for any valid score vectors w^1, w^2, we have $|\!|\!|M(w^1) - M(w^2)|\!|\!|_{\mathrm{F}}^2 \leq n\zeta^2 \|w^1 - w^2\|_2^2$ where $\zeta := \max_{z \in [-1,1]} F'(z)$ is a constant independent of n and p_{obs}. Putting these results together and performing some simple algebraic manipulations leads to the upper bound

$$\frac{1}{n^2}\mathbb{E}\Big[|\!|\!|M(\widehat{w}_{\mathrm{ML}}) - M^*|\!|\!|_{\mathrm{F}}^2\Big] \leq \frac{c_3 \zeta^2}{n p_{\mathrm{obs}}},$$

which establishes the claim.

2.6.8 Proof of Proposition 2: MST and WST models

We will derive an order one lower bound under the moderate stochastic transitivity condition. This result automatically implies the order one lower bound for weak stochastic transitivity.

[7] Note that the Laplacian matrix used in the statement of [220, Theorem 2(b)] is a scaled version of the matrix L introduced here, with each entry of L divided by the total number of observations.

The proof imposes a certain structure on a subset of the entries of M^* in a manner that $\Theta(n^2)$ remaining entries are free to take arbitrary values within the interval $[\frac{1}{2}, 1]$. This flexibility then establishes a minimax error of $\Theta(1)$ as claimed.

Let us suppose M^* corresponds to the identity permutation of the n items, and that this information is public knowledge. Set the entries of M^* above the diagonal in the following manner. For every $i \in [n]$ and every *odd* $j \in [n]$, set $M^*_{ij} = \frac{1}{2}$. For every $i \in [n]$ and every *even* $j \in [n]$, set $M^*_{ji} = \frac{1}{2}$. This information is also assumed to be public knowledge. Let $\mathcal{S} \subset [n]^2$ denote the set of all entries of M^* above the diagonal whose values were not assigned in the previous step. Let $|\mathcal{S}|$ denote the size of set $\mathcal{S}$. The entries below the diagonal are governed by the skew-symmetry constraints.

We first argue that every entry in $\mathcal{S}$ can take arbitrary values in the interval $[\frac{1}{2}, 1]$, and are not constrained by each other under the moderate stochastic transitivity condition. To this end, consider any entry $(i, k) \in \mathcal{S}$. Recall that the moderate stochastic transitivity condition imposes the following set of restrictions in M^*_{ik}: for every j, $M^*_{ik} \geq \min\{M^*_{ij}, M^*_{jk}\}$. From our earlier construction we have that for every odd value of j, $M^*_{ij} = \frac{1}{2}$ and hence the restriction simply reduces to $M^*_{ik} \geq \frac{1}{2}$. On the other hand, for every even value of j, our construction gives $M^*_{jk} = \frac{1}{2}$, and hence the restriction again reduces to $M^*_{ik} \geq \frac{1}{2}$. Given the absence of any additional restrictions, the error $\mathbb{E}[|||\widehat{M} - M^*|||_{\text{F}}^2] \geq c|\mathcal{S}|$. Finally, observe that every entry (i, k) where $i < k$, i is odd and k is even belongs to the set $\mathcal{S}$. It follows that $|\mathcal{S}| \geq \frac{n^2}{8}$, thus proving our claim.

2.6.9 Proof of Proposition 3: Other statistical models

The constructions governing the claimed relations are enumerated in Figure 2.2 and the details are provided below.

It is easy to see that since F is non-decreasing, the parameter-based class $\mathbb{C}_{\text{PAR}}$ is contained in the strong stochastic transitivity class $\mathbb{C}_{\text{SST}}$. We provide a formal proof of this statement for the sake of completeness. Suppose without loss of generality that $w_1 \geq \cdots \geq w_n$. Then we claim that the distribution of pairwise comparisons generated through this model result in a matrix, say M, that lies in the permutation-based SST model with the ordering following the identity permutation. This is because for any $i > j > k$,

$$\begin{aligned} w_i - w_k &\geq w_i - w_j \\ F(w_i - w_k) &\geq F(w_i - w_j) \\ M_{ik} &\geq M_{ij}. \end{aligned}$$

We now show the remaining relations with the four constructions indicated in Figure 2.2. While these constructions target some specific value of n, the results hold for any value n greater than that specific value. To see this, suppose we construct a matrix M for some $n = n_0$, and show that it lies inside (or outside) one of these classes. Consider any $n > n_0$, and define a $(n \times n)$ matrix M' as having M as the top-left $(n_0 \times n_0)$ block, $\frac{1}{2}$ on the remaining diagonal entries, 1 on the remaining entries above the diagonal and 0 on the remaining entries below the diagonal. This matrix M' will retain the properties of M in terms of lying inside (or outside, respectively) the claimed class.

In this proof, we use the notation $i \succ j$ to represent a greater preference for i as compared to j.

Construction 1

We construct a matrix M such that $M \in \mathbb{C}_{\text{FULL}}$ but $M \notin \mathbb{C}_{\text{SST}}$. Let $n = 3$. Consider the following distribution over permutations of 3 items $(1, 2, 3)$:

$$\begin{aligned} \mathbb{P}(1 \succ 2 \succ 3) &= \tfrac{2}{5}, \\ \mathbb{P}(3 \succ 1 \succ 2) &= \tfrac{1}{5}, \\ \mathbb{P}(2 \succ 3 \succ 1) &= \tfrac{2}{5}. \end{aligned}$$

This distribution induces the pairwise marginals

$$\begin{aligned} \mathbb{P}(1 \succ 2) &= \tfrac{3}{5}, \\ \mathbb{P}(2 \succ 3) &= \tfrac{4}{5}, \\ \mathbb{P}(3 \succ 1) &= \tfrac{3}{5}. \end{aligned}$$

Set $M_{ij} = \mathbb{P}(i \succ j)$ for every pair. By definition of the class $\mathbb{C}_{\text{FULL}}$, we have $M \in \mathbb{C}_{\text{FULL}}$.

A necessary condition for a matrix M to belong to the class $\mathbb{C}_{\text{SST}}$ is that there must exist at least one item, say item i, such that $M_{ij} \geq \frac{1}{2}$ for every item j. One can verify that the pairwise marginals enumerated above do not satisfy this condition, and hence $M \notin \mathbb{C}_{\text{SST}}$.

Construction 2

We construct a matrix M such that $M \in \mathbb{C}_{\text{SST}} \cap \mathbb{C}_{\text{FULL}}$ but $M \notin \mathbb{C}_{\text{PAR}}$. Let $n = 4$ and consider the following distribution over permutations of 4 items $(1, 2, 3, 4)$:

$$\begin{gathered} \mathbb{P}(3 \succ 1 \succ 2 \succ 4) = \frac{1}{8}, \qquad \mathbb{P}(1 \succ 2 \succ 4 \succ 3) = \frac{1}{8} \\ \mathbb{P}(2 \succ 1 \succ 4 \succ 3) = \frac{2}{8} \quad \text{and} \quad \mathbb{P}(1 \succ 2 \succ 3 \succ 4) = \frac{4}{8}. \end{gathered}$$

One can verify that this distribution leads to the following pairwise comparison matrix M (with the ordering of the rows and columns respecting the permutation $1 \succ 2 \succ 3 \succ 4$):

$$M := \frac{1}{8} \begin{bmatrix} 4 & 6 & 7 & 8 \\ 2 & 4 & 7 & 8 \\ 1 & 1 & 4 & 5 \\ 0 & 0 & 3 & 4 \end{bmatrix}.$$

It is easy to see that this matrix $M \in \mathbb{C}_{\text{SST}}$, and by construction $M \in \mathbb{C}_{\text{FULL}}$. Finally, the proof of Proposition 2 shows that $M \notin \mathbb{C}_{\text{PAR}}$, thereby completing the proof.

Construction 3

We construct a matrix M such that $M \in \mathbb{C}_{\text{PAR}}$ (and hence $M \in \mathbb{C}_{\text{SST}}$) but $M \notin \mathbb{C}_{\text{FULL}}$. First observe that any total ordering on n items can be represented as an $(n \times n)$ matrix in the permutation-based SST class such that all its off-diagonal entries take values in $\{0, 1\}$. The class $\mathbb{C}_{\text{FULL}}$ is precisely the convex hull of all such binary permutation-based SST matrices.

Let $B^1, \ldots, B^{n!}$ denote all $(n \times n)$ matrices in $\mathbb{C}_{\text{SST}}$ whose off-diagonal elements are restricted to take values in the set $\{0, 1\}$. The following lemma derives a property that any matrix in the convex hull of $B^1, \ldots, B^{n!}$ must satisfy.

Lemma 14. *Consider any $M \in \mathbb{C}_{SST}$, and consider three items $i, j, k \in [n]$ such that M respects the ordering $i \succ j \succ k$. Suppose $M_{ij} = M_{jk} = \frac{1}{2}$ and $M_{ik} = 1$. Further suppose that M can be written as*

$$M = \sum_{\ell \in [n!]} \alpha^\ell B^\ell, \tag{2.71}$$

where $\alpha^\ell \geq 0 \ \forall \ \ell$ and $\sum_{\ell=1}^{n!} \alpha^\ell = 1$. Then for any $\ell \in [n!]$ such that $\alpha^\ell > 0$, it must be that $B^\ell_{ij} \neq B^\ell_{jk}$.

The proof of the lemma is provided at the end of this section.

Now consider the following (7×7) matrix $M \in \mathbb{C}_{\text{SST}}$:

$$M := \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{2.72}$$

We will now show via proof by contradiction that M cannot be represented as a convex combination of the matrices $B^1, \ldots, B^{n!}$. We will then show that $M \in \mathbb{C}_{\text{PAR}}$.

Suppose one can represent M as a convex combination $M = \sum_{\ell \in [n!]} \alpha^\ell B^\ell$, where $\alpha^1, \ldots, \alpha^{n!}$ are non-negative scalars that sum to one. Consider any ℓ such that $\alpha^\ell \neq 0$. Let $B^\ell_{12} = b \in \{0, 1\}$. Let us derive some more constraints on B^ℓ. Successively applying Lemma 14 for the following values of i, j, k implies that B^ℓ must necessarily have the form (2.73) shown below. Here $\bar{b} := 1 - b$ and '$*$' denotes some arbitrary value that is irrelevant to the discussion at hand.

- $i = 1, j = 2, k = 3$ gives $B^\ell_{23} = \bar{b}$
- $i = 1, j = 2, k = 4$ gives $B^\ell_{24} = \bar{b}$

- $i = 2, j = 3, k = 5$ gives $B^\ell_{35} = b$
- $i = 2, j = 4, k = 6$ gives $B^\ell_{46} = b$
- $i = 3, j = 5, k = 6$ gives $B^\ell_{56} = \bar{b}$
- $i = 4, j = 6, k = 7$ gives $B^\ell_{67} = \bar{b}$.

Thus B^ℓ must be of the form

$$B^\ell = \begin{bmatrix} \frac{1}{2} & b & 1 & 1 & 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & \bar{b} & \bar{b} & 1 & 1 & 1 \\ 0 & b & \frac{1}{2} & * & b & 1 & 1 \\ 0 & b & * & \frac{1}{2} & * & b & 1 \\ 0 & 0 & \bar{b} & * & \frac{1}{2} & \bar{b} & 1 \\ 0 & 0 & 0 & \bar{b} & b & \frac{1}{2} & \bar{b} \\ 0 & 0 & 0 & 0 & 0 & b & \frac{1}{2} \end{bmatrix}. \tag{2.73}$$

Finally, applying Lemma 14 with $i = 5$, $j = 6$ and $k = 7$ implies that $B^\ell_{67} = b$, which contradicts the necessary condition in equation (2.73). We have thus shown that $M \notin \mathbb{C}_{\text{FULL}}$.

We now show that the matrix M constructed in equation (2.72) is contained in the class $\mathbb{C}_{\text{PAR}}$. Consider the following function $F : [-1, 1] \to [0, 1]$ in the definition of a parameter-based class:

$$F(x) = \begin{cases} 0 & \text{if } x < -0.25 \\ \frac{1}{2} & \text{if } -0.25 \leq x \leq 0.25 \\ 1 & \text{if } x > 0.25. \end{cases}$$

Let $n = 7$ with $w_1 = .9$, $w_2 = .7$, $w_3 = .6$, $w_4 = .5$, $w_5 = .4$, $w_6 = .3$ and $w_7 = .1$. One can verify that under this construction, the matrix of pairwise comparisons is identical to that in equation (2.72).

Proof of Lemma 14 In what follows, we show that $\sum_{\ell : B^\ell_{ij}=1, B^\ell_{jk}=1} \alpha^\ell = \sum_{\ell : B^\ell_{ij}=1, B^\ell_{jk}=1} \alpha^\ell = 0$. The result then follows immediately.

Consider some $\ell' \in [n!]$ such that $\alpha^{\ell'} > 0$ and $B^{\ell'}_{ij} = 0$. Since $M_{ik} = 1$, we must have $B^{\ell'}_{ik} = 1$. Given that $B^{\ell'}$ represents a total ordering of the n items, that is, $B^{\ell'}$ is a permutation-based SST matrix with boolean-valued its off-diagonal elements, $B^{\ell'}_{ij} = 0$ and $B^{\ell'}_{ik} = 1$ imply that $B^{\ell'}_{jk} = 1$. We have thus shown that $B^{\ell'}_{jk} = 1$ whenever $B^{\ell'}_{ij} = 0$. This result has two consequences. The first consequence is that $\sum_{\ell : B^\ell_{ij}=0, B^\ell_{jk}=0} \alpha^\ell = 0$. The second consequence employs the additional fact that $M_{ij} = \frac{1}{2}$ and hence $\sum_{\ell : B^\ell_{ij}=0} \alpha^\ell = \frac{1}{2}$, and then gives $\sum_{\ell:B^\ell_{ij}=0,B^\ell_{jk}=1}\alpha^\ell = \frac{1}{2}$. Building on, we have

$$\frac{1}{2} = M_{jk} = \sum_{\ell:B^\ell_{ij}=0,B^\ell_{jk}=1}\alpha^\ell + \sum_{\ell:B^\ell_{ij}=1,B^\ell_{jk}=1}\alpha^\ell,$$

and hence we have $\sum_{\ell:B^\ell_{ij}=1,B^\ell_{jk}=1}\alpha^\ell = 0$, thus completing the proof.

Construction 4

We construct a matrix M such that $M \in \mathbb{C}_{\text{SST}}$ but $M \notin \mathbb{C}_{\text{FULL}}$ and $M \notin \mathbb{C}_{\text{PAR}}$. Consider $n = 11$. Let M_2 denote the (4×4) matrix of Construction 2 and let M_3 denote the (7×7) matrix of construction 3. Consider the (11×11) matrix M of the form

$$M := \begin{bmatrix} M_2 & 1 \\ 0 & M_3 \end{bmatrix}.$$

Since $M_2 \in \mathbb{C}_{\text{SST}}$ and $M_3 \in \mathbb{C}_{\text{SST}}$, it is easy to see that $M \in \mathbb{C}_{\text{SST}}$. Since $M_2 \notin \mathbb{C}_{\text{PAR}}$ and $M \notin \mathbb{C}_{\text{FULL}}$, it follows that $M \notin \mathbb{C}_{\text{PAR}}$ and $M \notin \mathbb{C}_{\text{FULL}}$. This construction completes the proof of Proposition 3.

2.A Appendix: Relation to other error metrics

In this section, we show how estimation of the pairwise-comparison-probability matrix M^* under the squared Frobenius norm implies estimates and bounds under other error metrics. In particular, we investigate relations between estimation of the true underlying ordering under the Spearman's footrule and the Kemeny metrics, and estimation of the matrix M^* under the Kullback-Leibler divergence metric.

2.A.1 Recovering the true ordering

Recall that the permutation-based SST class assumes the existence of some true ordering of the n items. The pairwise-comparison probabilities are then assumed to be faithful to this ordering. In this section, we investigate the problem of estimating this underlying ordering.

In order to simplify notation, we assume without loss of generality that this true underlying ordering is the identity permutation of the n items, and denote the identity permutation as π_{id}. Recall the set $\mathbb{C}_{\text{BISO}}$ of *bivariate isotonic matrices*, that is, permutation-based SST matrices that are faithful to the identity permutation:

$$\mathbb{C}_{\text{BISO}} = \{M \in [0,1]^{n\times n} \mid M_{ij} = 1 - M_{ji} \text{ for all } (i,j) \in [n]^2, \text{ and } M_{i\ell} \geq M_{j\ell} \text{ whenever } i < j.\}$$

Then we have that $M^* \in \mathbb{C}_{\text{BISO}}$. Let π be any permutation of the n items. For any matrix $M \in \mathbb{R}^{n\times n}$ and any integer $i \in [n]$ we let M_i denote the i^{th} row of M.

Two of the most popular metrics of measuring the error between two such orderings are the Spearman's footrule and the Kemeny (or Kendall tau) distance, defined as follows. Spearman's footrule measures the total displacement of all items in π as compared to π_{id}, namely

$$\text{Spearman's footrule}(\pi, \pi_{\text{id}}) := \sum_{i=1}^{n} \mid \pi(i) - i \mid .$$

On the other hand, the Kemeny distance equals the total number of pairs whose relative positions are different in the two orderings, namely,

$$\text{Kemeny}(\pi, \pi_{\text{id}}) := \sum_{1 \leq i < j \leq n} \mathbf{1}\{\text{sign}(\pi(i) - \pi(j)) \neq \text{sign}(i - j)\},$$

where "sign" denotes the sign function, that is, $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$ and $\text{sign}(x) = 0$ if $x = 0$. The Kemeny distance is also known as the Kendall tau metric.

Before investigating the two aforementioned metrics, we remark on one important aspect of the problem of estimating the order of the items. Observe that if the rows of M^* corresponding to some pair of items (i, j) are very close to each other (say, in a pointwise sense), then it is hard to estimate the relative position of item i with respect to item j. On the other hand, if the two rows are far apart then differentiating between the two items is easier. Consequently, it is reasonable to consider a metric that penalizes errors in the inferred permutation based on the relative values of the rows of M^*. To this end, we define a reweighted version of Spearman's footrule as

$$\text{Matrix-reweighted Spearman's footrule}_{M^*}(\pi, \pi_{\text{id}}) := |||\pi(M^*) - M^*|||_{\text{F}}^2 = \sum_{i=1}^{n} \|M^*_{\pi(i)} - M^*_i\|_2^2.$$

Given these definitions, the following proposition now relates the squared Frobenius norm metric to the other aforementioned metrics.

Proposition 2.A. *Any two matrices $M^* \in \mathbb{C}_{\text{BISO}}$, and $M \in \mathbb{C}_{\text{SST}}$ with π as its underlying permutation, must satisfy the following bound on the matrix-reweighted Spearman's footrule:*

$$|||M^* - \pi(M^*)|||_F^2 \leq 4|||M^* - M|||_F^2.$$

Proposition 2.B. *Consider any matrix $M^* \in \mathbb{C}_{\text{BISO}}$ that satisfies $\|M^*_i - M^*_{i+1}\|_2^2 \geq \gamma^2$ for some constant $\gamma > 0$ and for every $i \in [n-1]$. Then for any permutation π, the Spearman's footrule distance from the identity permutation is upper bounded as*

$$\sum_{i=1}^{n} \mid i - \pi(i) \mid \leq \frac{1}{\gamma^2} |||M^* - \pi(M^*)|||_F^2.$$

Conversely, there exists a matrix $M^* \in \mathbb{C}_{BISO}$ that satisfies $\|M^*_i - M^*_{i+1}\|_2^2 = \gamma^2$ for every $i \in [n-1]$ such that for every permutation π, the Spearman's footrule distance from the identity permutation is lower bounded as

$$\sum_{i=1}^{n} \mid i - \pi(i) \mid \geq \frac{1}{4\gamma^2} \|M^* - \pi(M^*)\|_F^2.$$

Proposition 2.C ([64])**.** *The Kemeny distance of any permutation π from the identity permutation π_{id} is sandwiched as*

$$\frac{1}{2}\sum_{i=1}^{n} \mid i - \pi(i) \mid \leq \sum_{1 \leq i < j \leq n} \mathbf{1}\{sign(\pi(i) - \pi(j)) \neq sign(i-j)\} \leq \sum_{i=1}^{n} \mid i - \pi(i) \mid .$$

As a consequence of this proposition, an upper bound on the error in estimation of M^* under the squared Frobenius norm yields identical upper bounds (with some constant factors) under the other three metrics.
A few remarks are in order:

(a) Treating M^* as the true pairwise comparison probability matrix and M as its estimate, Proposition 2.A assumes that M also lies in the matrix class $\mathbb{C}_{\text{SST}}$. This set-up is known as proper learning in some of the machine learning literature.

(b) The γ-separation condition of Proposition 2.B is satisfied in the models assumed in several earlier works [23, 263].

The remainder of this subsection is devoted to the proof of these claims.

Proof of Proposition 2.A

For any matrix M and any permutation π of n items, let $\pi(M)$ denote the matrix resulting from permuting the rows of M by π. With this notation, we have

$$\|\pi(M^*) - M^*\|_{\text{F}}^2 \leq 2\|\pi(M^*) - M\|_{\text{F}}^2 + 2\|M - M^*\|_{\text{F}}^2 = 2\|M^* - \pi^{-1}(M)\|_{\text{F}}^2 + 2\|M - M^*\|_{\text{F}}^2.$$

We now show that

$$\|M^* - \pi^{-1}(M)\|_{\text{F}}^2 \leq \|M^* - M\|_{\text{F}}^2, \tag{2.74}$$

which would then imply the claimed result. As shown below, the inequality (2.74) is a consequence of the fact that M^* and $\pi^{-1}(\widehat{M})$ both lie in the permutation-based SST class and have the same underlying ordering of the rows. More generally, we claim that for any two matrices $M \in \mathbb{C}_{\text{BISO}}$ and $M' \in \mathbb{C}_{\text{BISO}}$,

$$\pi_{\text{id}} \in \arg\min_{\widetilde{\pi}} \|M - \widetilde{\pi}(M')\|_{\text{F}}^2, \tag{2.75}$$

where the minimization is carried out over all permutations of n items. To see this, consider any two matrices M and M' in $\mathbb{C}_{\text{BISO}}$ and let π' be a minimizer of $|\!|\!| M - \pi(M') |\!|\!|_{\text{F}}^2$. If $\pi' \neq \pi_{\text{id}}$, then there must exist some item $i \in [n-1]$ such that item $(i+1)$ is ranked higher than item i in π'. Consequently,

$$\begin{aligned}\|M_i - M'_{i+1}\|_2^2 + \|M_{i+1} - M'_i\|_2^2 - \|M_i - M'_i\|_2^2 - \|M_{i+1} - M'_{i+1}\|_2^2 \\ = 2\langle\!\langle M_i - M_{i+1},\ M'_i - M'_{i+1} \rangle\!\rangle \geq 0,\end{aligned}$$

where the final inequality follows from the fact that $M \in \mathbb{C}_{\text{BISO}}$ and $M' \in \mathbb{C}_{\text{BISO}}$. It follows that the new permutation obtained by swapping the positions of items i and $(i+1)$ in π' (which now ranks item i higher than item $(i+1)$) is also a minimizer of $|\!|\!| M - \pi(M') |\!|\!|_{\text{F}}^2$. A recursive application of this argument yields that π_{id} is also a minimizer of $|\!|\!| M - \pi(M') |\!|\!|_{\text{F}}^2$.

Proof of Proposition 2.B

We first prove the upper bound on the Spearman's footrule metric. Due to the monotonicity of the rows and the columns of M^*, we have the lower bound

$$|\!|\!| M^* - \pi(M^*) |\!|\!|_{\text{F}}^2 \geq \sum_{\ell=1}^{n} \|M^*_\ell - M^*_{\pi(\ell)}\|_2^2.$$

Now consider any $\ell \in [n]$ such that $\pi(\ell) > \ell$. Then we have

$$\|M^*_\ell - M^*_{\pi(\ell)}\|_2^2 = \| \sum_{i=\ell}^{\pi(\ell)-1} (M^*_i - M^*_{i+1}) \|_2^2 \overset{(i)}{\geq} \sum_{i=\ell}^{\pi(\ell)-1} \|M^*_i - M^*_{i+1}\|_2^2 \overset{(ii)}{\geq} \gamma^2 |\pi(i) - i|,$$

where the inequality (i) is a consequence of the fact that for every $i \in [n-1]$, every entry of the vector $(M^*_i - M^*_{i+1})$ is non-negative, and the inequality (ii) results from the assumed γ-separation condition on the rows of M^*. An identical argument holds when $\pi(\ell) < \ell$. This argument completes the proof of the upper bound.

We now move on to the lower bound on Spearman's footrule. To this end, consider the matrix $M^* \in \mathbb{C}_{\text{BISO}}$ with its entries given as:

$$[M^*]_{ij} = \begin{cases} \frac{1}{2} + \frac{\gamma}{\sqrt{2}} & \text{if } i < j \\ \frac{1}{2} & \text{if } i = j \\ \frac{1}{2} - \frac{\gamma}{\sqrt{2}} & \text{if } i > j. \end{cases}$$

One can verify that this matrix M^* satisfies the required condition $\|M^*_i - M^*_{i+1}\|_2^2 = \gamma^2$ for every $i \in [n-1]$. One can also compute that this matrix also satisfies the condition $|\!|\!| M^* - \pi(M^*) |\!|\!|_{\text{F}} = 4\gamma^2 \sum_{\ell=1}^{n} |\ell - \pi(\ell)|$, thereby yielding the claim.

Proof of Proposition 2.C

It is well known [64] that the Kemeny distance and Spearman's footrule distance between two permutation lie within a factor of 2 of each other.

2.A.2 Estimating comparison probabilities under Kullback-Leibler divergence

Let $\mathbb{P}_M$ denote the probability distribution of the observation matrix $Y \sim \{0,1\}^{n \times n}$ obtained by independently sampling entry Y_{ij} from a Bernoulli distribution with parameter M_{ij}. The Kullback-Leibler (KL) divergence between $\mathbb{P}_M$ and $\mathbb{P}_{M'}$ is given by

$$D_{\text{KL}}(\mathbb{P}_M \| \mathbb{P}_{M'}) = M_{ij} \log \frac{M_{ij}}{M'_{ij}} + (1 - M_{ij}) \log \frac{1 - M_{ij}}{1 - M'_{ij}}.$$

Before we establish the connection with the squared Frobenius norm, we make one assumption on the pairwise comparison probabilities that is standard in the literature on estimation from pairwise comparisons [46, 98, 177, 220]. We assume that every entry of M^* is bounded away from $\{0, 1\}$. In other words, we assume the existence of some known constant-valued parameter $\epsilon \in (0, \frac{1}{2}]$ whose value is independent of n, such that $M^*_{ij} \in (\epsilon, 1-\epsilon)$ for every pair (i, j). Given this assumption, for any estimator M of M^*, we clip each of its entries and force them to lie in the interval $(\epsilon, 1-\epsilon)$.[8] The following proposition then relates the Kullback-Leibler divergence metric to estimation under the squared Frobenius norm.

Proposition 3. *The probability distributions induced by any two probability matrices M^* and M must satisfy the sandwich inequalities:*

$$|\!|\!| M - M^* |\!|\!|_F^2 \leq D_{\text{KL}}(\mathbb{P}_M \| \mathbb{P}_{M^*}) \leq \frac{1}{\epsilon(1-\epsilon)} |\!|\!| M - M^* |\!|\!|_F^2,$$

where for the upper bound we have assumed that every entry of the matrices lies in $(\epsilon, 1-\epsilon)$.

The proof of the proposition follows from standard upper and lower bounds on the natural logarithm (2.59b). As a consequence of this result, any upper or lower bound on $|\!|\!| M - M^* |\!|\!|_{\text{F}}^2$ therefore automatically implies an identical upper or lower bound on $D_{\text{KL}}(\mathbb{P}_M \| \mathbb{P}_{M^*})$ up to constant factors.

[8] This clipping step does not increase the estimation error.

Chapter 3

Ranking and Top-k Recovery

"Find the five premier citizenry for the job. Make them work together for an outcome par excellence."
– Nikola Tesla

3.1 Introduction

Ranking problems involve a collection of n items, and some unknown underlying total ordering of these items. In many applications, one may observe (noisy) comparisons between various pairs of items. Examples include matches between football teams in tournament play; consumer's preference ratings in marketing; and certain types of voting systems in politics. Given a set of such noisy comparisons between items, it is often of interest to find the true underlying ordering of all n items, or alternatively, given some given positive integer $k < n$, to find the subset of k most highly rated items. These two problems are the focus of this chapter.

There is a substantial body of literature on the problem of finding approximate rankings based on noisy pairwise comparisons. A number of papers (e.g., [23, 71, 121]) consider models in which the probability of a pairwise comparison agreeing with the underlying order is identical across all pairs. These results break down when for one or more pairs, the probability of agreeing with the underlying ranking is either comes close to or is exactly equal to $\frac{1}{2}$. Another set of papers [98, 105, 177, 220, 244] work using parameter-based models of pairwise comparisons, and address the problem of recovering the parameters associated to every individual item. The works [3, 65, 107, 172] consider mixture models, in which every pairwise comparison is associated to a certain individual making the comparison, and it is assumed that the preferences across individuals can be described by a low-dimensional model.

Most related to our work are the papers [46, 192, 193, 263], which we discuss in more detail here. Wauthier et al. [263] analyze a weighted counting algorithm to recover approximate rankings; their analysis applies to a specific model in which the pairwise comparison between any pair of items remains faithful to their relative positions in the true ranking with a probability common across all pairs. They consider recovery of an approximate ranking (under Kendall's tau and maximum displacement metrics), but do not provide results on exact recovery. As the analysis of this chapter shows, their bounds are quite loose: their results are tight only when there are a total of at least $\Theta(n^2)$ comparisons. The pair of papers [192, 193] by Rajkumar et al. consider ranking under several models and several metrics. In the part that is common with our setting, they show that the counting algorithm is consistent in terms of recovering the full ranking, which automatically implies consistency in exactly recovering the top k items. They obtain upper bounds on the sample complexity in terms of a separation threshold that is identical to a parameter Δ_k defined subsequently in this chapter (see Section 3.3). However, as our analysis shows, their bounds are loose by at least an order of magnitude. They also assume a certain high-SNR condition on the probabilities, an assumption that is not imposed in our analysis.

Finally, in very recent work on this problem, Chen and Suh [46] proposed an algorithm called the Spectral MLE for exact recovery of the top k items. They showed that, if the pairwise observations are assumed to drawn according to the parameter-based Bradley-Terry-Luce (BTL) model [20, 154], the Spectral MLE algorithm recovers the k items correctly with high probability under certain regularity conditions. In addition, they also show, via matching lower bounds, that their regularity conditions are tight up to constant factors. While these guarantees are attractive, it is natural to ask how such an algorithm behaves when the data is not drawn from the BTL model. In real-world instances of pairwise ranking data, it is often found that parameter-based models, such as the BTL model and its variants, fail to provide accurate fits (for instance, see the papers [11, 59, 163, 255] and references therein).

With this context, the main contribution of this chapter is to analyze a classical counting-based method for ranking, often called the Copeland method [53], and to show that it is simple, optimal and robust. Our analysis does not require that the data-generating mechanism follow either the BTL or other parameter-based assumptions, nor other regularity conditions such as stochastic transitivity. We show that the Copeland counting algorithm has the following properties:

- Simplicity: The algorithm is simple, as it just orders the items by the number of pairwise comparisons won. As we will subsequently see, the execution time of this counting algorithm is several orders of magnitude lower as compared to prior work.

- Optimality: We derive conditions under which the counting algorithm achieves the stated goals, and by means of matching information-theoretic lower bounds, show that these conditions are tight.

- Robustness: The guarantees that we prove do not require any assumptions on the pairwise-comparison probabilities, and the counting algorithm performs well for various classes of data sets. In contrast, we find that the spectral MLE algorithm performs poorly when the data is not drawn from the BTL model.

In doing so, we consider three different instantiations of the problem of set-based recovery: (i) Recovering the top k items perfectly; (ii) Recovering the top k items allowing for a certain Hamming error tolerance; and (iii) a more general recovery problem for set families that satisfy a natural "set-monotonicity" condition. In order to tackle this third problem, we introduce a general framework that allows us to treat a variety of problems in the literature in an unified manner.

The remainder of this chapter is organized as follows. We begin in Section 3.2 with a more precise formulation of the problem. Section 3.3 presents our main theoretical results. Section 3.4 provides the results of experiments on both simulated and real-world data sets. We present a concluding discussion in Section 3.5. Finally, we provide all proofs in Section 3.6.

3.2 Problem setting

In this section, we provide a more formal statement of the problem along with background on various types of ranking models.

3.2.1 Problem statement

Given an integer $n \geq 2$, we consider a collection of n items, indexed by the set $[n] := \{1, \ldots, n\}$. For each pair $i \neq j$, we let M_{ij} denote the probability that item i wins the comparison with item j. We assume that that each comparison necessarily results in one winner, meaning that

$$M_{ij} + M_{ji} = 1, \qquad \text{and} \quad M_{ii} = \frac{1}{2},$$

where we set the diagonal for concreteness.

For any item $i \in [n]$, we define an associated score τ_i as

$$\tau_i := \frac{1}{n} \sum_{j=1}^{n} M_{ij}. \tag{3.1}$$

In words, the score τ_i of any item $i \in [n]$ corresponds to the probability that item i beats an item chosen uniformly at random from all n items.

Given a set of noisy pairwise comparisons, our goals are (a) to recover the k items with the maximum values of their scores; and (b) to recover the full ordering of all the items as defined by the score vector. The notion of ranking items via their scores (3.1) generalizes the explicit rankings under popular models in the literature. Indeed, as we discuss shortly, most models of pairwise comparisons considered in the literature either implicitly or explicitly assume that the items are ranked according to their scores. Note that neither the scores $\{\tau_i\}_{i \in [n]}$ nor the matrix $M := \{M_{ij}\}_{i,j \in [n]}$ of probabilities is assumed to be known.

More concretely, we consider a random-design observation model defined as follows. Each pair is associated with a random number of noisy comparisons, following a binomial distribution with parameters (r, p_{obs}), where $r \geq 1$ is the number of trials and $p_{\text{obs}} \in (0, 1]$ is the probability of making a comparison on any given trial. Thus, each pair (i, j) is associated with a binomial random variable with parameters (r, p_{obs}) that governs the number of comparisons between the pair of items. We assume that the observation sequences for different pairs are independent. Note that in the special case $p_{\text{obs}} = 1$, this random binomial model reduces to the case in which we observe exactly r observations of each pair; in the special case $r = 1$, the set of pairs compared form an (n, p_{obs}) Erdős-Rényi random graph.

In this chapter, we begin in Section 3.3.2 by analyzing the problem of exact recovery. More precisely, for a given matrix M of pairwise probabilities, suppose that we let $\mathcal{S}_k^*$ denote the (unknown) set of k items with the largest values of their respective scores, assumed to be unique for concreteness.

Given noisy observations specified by the pairwise probabilities M, our goal is to establish conditions under which there exists some algorithm $\widehat{\mathcal{S}}_k$ that identifies k items based on the outcomes of various comparisons such that the probability $\mathbb{P}_M(\widehat{\mathcal{S}}_k = \mathcal{S}_k^*)$ is very close to one. In the case of recovering the full ranking, our goal is to identify conditions that ensure that the probability $\mathbb{P}_M\big(\bigcap_{k \in [n]} (\widehat{\mathcal{S}}_k = \mathcal{S}_k^*)\big)$ is close to one.

In Section 3.3.3, we consider the problem of recovering a set of k items that approximates $\mathcal{S}_k^*$ with a minimal Hamming error For any two subsets of $[n]$, we define their Hamming distance D_{H}, also referred to as their Hamming error, to be the number of items that belong to exactly one of the two sets—that is

$$D_{\text{H}}(A, B) = \text{card}\Big(\{A \cup B\} \backslash \{A \cap B\}\Big). \tag{3.2}$$

For a given user-defined tolerance parameter $h \geq 0$, we derive conditions that ensure that $D_{\text{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2h$ with high probability.

Finally, we generalize our results to the problem of satisfying any a general class of requirements on set families. These requirement are specified in terms of which k-sized subsets of the items are allowed, and is required to satisfy only one natural condition, that of set-monotonicity, meaning that replacing an item in an allowed set with a higher rank item should also be allowed. See Section 3.3.4 for more details on this general framework.

3.2.2 A range of pairwise comparison models

To be clear, our work makes no assumptions on the form of the pairwise comparison probabilities. However, so as to put our work in context of the literature, let us briefly review some standard models uesd for pairwise comparison data.

Parameter-based models: A broad class of parameter-based models, including the Bradley-Terry-Luce (BTL) model as a special case [20, 154], are based on assuming the existence of "quality" parameter $w_i \in \mathbb{R}$ for each item i, and requiring that the probability of an item beating another is a specific function of the difference between their values. In the BTL model, the probability M_{ij} that i beats j is given by the logistic model

$$M_{ij} = \frac{1}{1 + e^{-(w_i - w_j)}}. \tag{3.3a}$$

More generally, parameter-based models assume that the pairwise comparison probabilities take the form

$$M_{ij} = F(w_i - w_j), \tag{3.3b}$$

where $F : \mathbb{R} \rightarrow [0, 1]$ is some strictly increasing cumulative distribution function.

By construction, any parameter-based model has the following property: if $w_i > w_j$ for some pair of items (i, j), then we are also guaranteed that $M_{i\ell} > M_{j\ell}$ for every item ℓ. As a consequence, we are guaranteed that $\tau_i > \tau_j$, which implies that ordering of the items in terms of their quality vector $w \in \mathbb{R}^n$ is identical to their ordering in terms of the score vector $\tau \in \mathbb{R}^n$. Consequently, if the data is actually drawn from a parameter-based model, then recovering the top k items according to their scores is the same as recovering the top k items according their respective quality parameters.

Strong Stochastic Transitivity (SST) class: The class of strong stochastic transitivity (SST) models, studied in more detail in Chapter 2, is a superset of parameter-based models. It does not assume the existence of a quality vector, nor does it assume any specific form of the probabilities as in equation (3.3a). Instead, the SST class is defined by assuming the existence of a total ordering of the n items, and imposing the inequality constraints $M_{i\ell} \geq M_{j\ell}$ for every pair of items (i, j) where i is ranked above j in the ordering, and every item ℓ. One can verify that an ordering by the scores $\{\tau_i\}_{i \in [n]}$ of the items lead to an ordering of the items that is consistent with that defined by the SST class.

A very general permutation-based model: From the discussion above, we see that in a broad class of models for pairwise ranking, the total ordering defined by the scores (3.1) coincides with the underlying ordering used to define the models. In this chapter, we analyze the performance of a counting algorithm with a focus on a very general permutation-based model that imposes very little conditions on the family of pairwise probabilities. Our permutation-based model assumes that there exists an unknown permutation of the items which is the ranking to be recovered, and that the scores (3.1) associated to the n items are ordered according to this permutation. Formally, denoting the unknown underlying permutation as π^*, our permutation-based model is governed by the assumption that $\tau_i \geq \tau_j$ whenever $\pi^*(i) < \pi^*(j)$.

The next three sections establish theoretical guarantees on the recovery of the top k items under various requirements.

3.3 Main results

In this section, we present our main theoretical results on top-k recovery under the three settings described earlier. Note that the three settings are ordered in terms of increasing generality, with the advantage that the least general setting leads to the simplest form of theoretical claim. We begin with a formal description of the counting algorithm that we analyze in this chapter.

3.3.1 Copeland counting algorithm

The analysis of this chapter focuses on a simple counting-based algorithm, often called the Copeland method [53]. It can be also be viewed as a special case of the Borda count method [18], which applies more generally to observations that consist of rankings of two or more items. Here we describe how this method applies to the random-design observation model introduced earlier.

More precisely, for each distinct $i, j \in [n]$ and every integer $\ell \in [r]$, let $Y_{ij}^{\ell} \in \{-1, 0, +1\}$ represent the outcome of the ℓ^{th} comparison between the pair i and j, defined as

$$Y_{ij}^{\ell} = \begin{cases} 0 & \text{no comparison between } (i.j) \text{ in trial } \ell \\ +1 & \text{if comparison is made and item } i \text{ beats } j \\ -1 & \text{if comparison is made and item } j \text{ beats } i. \end{cases} \tag{3.4}$$

Note that this definition ensures that $Y_{ij}^{\ell} = -Y_{ji}^{\ell}$. For $i \in [n]$, the quantity

$$N_i := \sum_{j \in [n]} \sum_{\ell \in [r]} \mathbf{1}\{Y_{ij}^{\ell} = 1\} \tag{3.5}$$

corresponds to the number of pairwise comparisons won by item i. Here we use $\mathbf{1}\{\cdot\}$ to denote the indicator function that takes the value 1 if its argument is true, and the value 0 otherwise. For each integer k, the vector $\{N_i\}_{i=1}^{n}$ of number of pairwise wins defines a k-sized subset

$$\widetilde{\mathcal{S}}_k = \Big\{ i \in [n] \mid N_i \text{ is among the } k \text{ highest number of pairwise wins} \Big\}, \tag{3.6}$$

corresponding to the set of k items with the largest values of N_i. Otherwise stated, the set $\widetilde{\mathcal{S}}_k$ corresponds to the rank statistics of the top k-items in the pairwise win ordering. (If there are any ties, we resolve them by choosing the indices with the smallest value of i.)

3.3.2 Thresholds for exact recovery of the top k items

We begin with the goal of exactly recovering the k top-ranked items. As one might expect, the difficulty of this problem turns out to depend on the degree of separation between the top k items and the remaining items. More precisely, let us use (k) and $(k+1)$ to denote the indices of the items that are ranked k^{th} and $(k+1)^{th}$ respectively. With this notation, the *k-separation threshold* Δ_k is given by

$$\Delta_k := \tau_{(k)} - \tau_{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} M_{(k)i} - \frac{1}{n}\sum_{i=1}^{n} M_{(k+1)i}. \tag{3.7}$$

In words, the quantity Δ_k is the difference in the probability of item (k) beating another item chosen uniformly at random, versus the same probability for item $(k+1)$.

As shown by the following theorem, success or failure in recovering the top k entries is determined by the size of Δ_k relative to the number of items n, observation probability p_{obs} and number of repetitions r. In particular, consider the family of matrices

$$\mathcal{F}_k(\alpha; n, p_{\text{obs}}, r) := \left\{ M \in [0,1]^{n\times n} \mid M + M^T = 11^T, \text{ and } \Delta_k \geq \alpha\sqrt{\frac{\log n}{n p_{\text{obs}} r}} \right\}. \tag{3.8}$$

To simplify notation, we often adopt $\mathcal{F}_k(\alpha)$ as a convenient shorthand for this set, where its dependence on (n, p_{obs}, r) should be understood implicitly.

With this notation, the achievable result in part (a) of the following theorem is based on the estimator that returns the set $\widetilde{\mathcal{S}}_k$ of the the k items defined by the number of pairwise comparisons won, as defined in equation (3.6). On the other hand, the information theoretic lower bound established in part (b) applies to *any estimator*, meaning any measurable function of the observations.

Theorem 7. *(a) For any $\alpha \geq 8$, the maximum pairwise win estimator $\widetilde{\mathcal{S}}_k$ from equation* (3.6) *satisfies*

$$\sup_{M\in\mathcal{F}_k(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \leq \frac{1}{n^{14}}. \tag{3.9a}$$

(b) Conversely, suppose that $n \geq 7$ and $p_{\text{obs}} \geq \frac{\log n}{2nr}$. Then for any $\alpha \leq \frac{1}{7}$, the error probability of any *estimator $\widehat{\mathcal{S}}_k$ is lower bounded as*

$$\sup_{M\in\mathcal{F}_k(\alpha)} \mathbb{P}_M\big[\widehat{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \geq \frac{1}{7}. \tag{3.9b}$$

Remarks: First, it is important to note that the negative result in part (b) holds even if the supremum is further restricted to a particular parameter-based sub-class of $\mathcal{F}_k(\alpha)$, such as the pairwise comparison matrices generated by the BTL model, or by the SST model. Our proof of the lower bound for exact recovery is based on a generalization of a construction introduced by Chen and Suh [46], one adapted to the general definition (3.7) of the separation threshold Δ_k.

Second, we note that in the regime $p_{\text{obs}} < \frac{\log n}{2nr}$, standard results from random graph theory [70] can be used to show that there are at least $\sqrt{n}$ items (in expectation) that are never compared to any other item. Of course, estimating the rank is impossible in this pathological case, so we omit it from consideration.

Third, the two parts of the theorem in conjunction show that the counting algorithm is essentially optimal. The only room for improvement is in the difference between the value 8 of α in the achievable result, and the value $\frac{1}{7}$ in the lower bound.

Theorem 7 can also be used to derive guarantees for recovery of other functions of the underlying ranking. Here we consider the problem of identifying the ranking of all n items, say denoted by the permutation π^*. In this case, we require that each of the separations $\{\Delta_j\}_{j=1}^{n-1}$ are suitably lower bounded: more precisely, we study models M that belong to the intersection $\cap_{j=1}^{n-1} \mathcal{F}_j(\gamma)$.

Corollary 1. *Let $\widetilde{\pi}$ be the permutation of the items specified by the number of pairwise comparisons won. Then for any $\alpha \geq 8$, we have*

$$\sup_{M \in \cap_{j=1}^{n-1} \mathcal{F}_j(\alpha)} \mathbb{P}_M\big[\widetilde{\pi} \neq \pi^*\big] \leq \frac{1}{n^{13}}.$$

Moreover, the separation condition on $\{\Delta_j\}_{j=1}^{n-1}$ that defines the set $\cap_{j=1}^{n-1} \mathcal{F}_j(\alpha)$ is unimprovable beyond constant factors.

This corollary follows from the equivalence between correct recovery of the ranking and recovering the top k items for every value of $k \in [n]$.

Detailed comparison to related work: In the remainder of this subsection, we make a detailed comparison to the related works [46, 192, 193, 263] that we briefly discussed earlier in Section 3.1.

Wauthier et al. [263] analyze a weighted counting algorithm for approximate recovery of rankings; they work under a model in which $M_{ij} = \frac{1}{2} + \gamma$ whenever item i is ranked above item j in an assumed underlying ordering. Here the parameter $\gamma \in (0, \frac{1}{2}]$ is independent of (i, j), and as a consequence, the best ranked item is assumed to be as likely to meet the worst item as it is to beat the second ranked item, for instance. They analyze approximate ranking under Kendall tau and maximum displacement metrics. In order to have a displacement upper bounded by by some $\delta > 0$, their bounds require the order of $\frac{n^5}{\delta^2 \gamma^2}$ pairwise comparisons. In comparison, our model is more general in that we do not impose the γ-condition on the pairwise probabiltiies. When specialized to the γ-model, the quantities $\{\Delta_j\}_{j=1}^{n}$ in our analysis takes the form $\Delta_j = \frac{2\gamma}{n}$, and Corollary 1 shows that $\frac{n \log n}{\min_{j\in[n]} \Delta_j^2} = \frac{n^3 \log n}{\gamma^2}$ observations are sufficient to recover the exact total ordering. Thus, for any constant δ, Corollary 1 guarantees recover with a multiplicative factor of order $\frac{n^2}{\log n}$ smaller than that established by Wauthier et al. [263].

The pair of papers [192, 193] by Rajkumar et al. consider ranking under several models and several metrics. For the subset of their models common with our setting—namely, Bradley-Terry-Luce and the so-called low noise models—they show that the counting algorithm is consistent in terms of recovering the full ranking or the top subset of items. The guarantees are obtained under a low-noise assumpotion: namely, that the probability of any item i beating j is at least $\frac{1}{2} + \gamma$ whenever item i is ranked higher than item j in an assumed underlying ordering. Their guarantees are based on a sample size of at least $\frac{\log n}{\gamma^2 \mu^2}$, where μ is a parameter lower bounded as $\mu \geq \frac{1}{n^2}$. Once again, our setting allows for the parameter γ to be arbitrarily close to zero, and furthermore as one can see from the discussion above, our bounds are much stronger. Moreover, while Rajkumar et al. focus on upper bounds alone, we also prove matching lower bounds on sample complexity showing that our results are unimprovable beyond constant factors. It should be noted that Rajkumar et al. also provide results for other types of ranking problems that lie outside the class of models treated in the current chapter.

Most recently, Chen and Suh [46] show that if the pairwise observations are assumed to drawn according to the parameter-based Bradley-Terry-Luce (BTL) model (3.3a), then their proposed Spectral MLE algorithm recovers the k items correctly with high probability when a certain separation condition on the parameters $\{w_i\}_{i=1}^n$ of the BTL model is satisfied. In addition, they also show, via matching lower bounds, that this separation condition are tight up to constant factors. In real-world instances of pairwise ranking data, it is often found that parameter-based models, such as the BTL model and its variants, fail to provide accurate fits [11, 59, 163, 255]. Our results make no such assumptions on the noise, and furthermore, our notion of the ordering of the items in terms of their scores (3.1) strictly generalizes the notion of the ordering with respect to the BTL parameters. In empirical evaluations presented subsequently, we see that the counting algorithm is significantly more robust to various kinds of noise, and takes several orders of magnitude lesser time to compute.

Finally, in addition to the notion of exact recovery considered so far, in the next two subsections we also derive tight guarantees for the Hamming error metric and more general metrics inspired by the requirements of many relevant applications [8, 73, 106, 126, 166, 168].

3.3.3 Approximate recovery under Hamming error

In the previous section, we analyzed performance in terms of exactly recovering the k-ranked subset. Although exact recovery is suitable for some applications (e.g., a setting with high stakes, in which any single error has a large price), there are other settings in which it may be acceptable to return a subset that is "close" to the correct k-ranked subset. In this section, we analyze this problem of approximate recovery when closeness is measured under the Hamming error. More precisely, for a given threshold $h \in [0, k)$, suppose that our goal is to output a set k-sized set $\widehat{\mathcal{S}}_k$ such that its Hamming distance to the set $\mathcal{S}^*_k$ of the true top k items, as defined in equation (3.2), is bounded as

$$D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}^*_k) \leq 2h. \tag{3.10}$$

Our goal is to establish conditions under which it is possible (or impossible) to return an estimate $\widehat{\mathcal{S}}_k$ satisfying the bound (3.10) with high probability.[1]

As before, we use $(1), \ldots, (n)$ to denote the permutation of the n items in decreasing order of their scores. With this notation, the following quantity plays a central role in our analysis:

$$\Delta_{k,\,h} := \tau_{(k-h)} - \tau_{(k+h+1)}. \tag{3.11a}$$

Observe that $\Delta_{k,\,h}$ is a generalization of the quantity Δ_k defined previously in equation (3.7); more precisely, the quantity Δ_k corresponds to $\Delta_{k,\,h}$ with $h = 0$. We then define a generalization of the family $\mathcal{F}_k(\alpha; n, p_{\mathrm{obs}}, r)$, namely

$$\mathcal{F}_{k,h}(\alpha; n, p_{\mathrm{obs}}, r) := \left\{ M \in [0,1]^{n \times n} \;\middle|\; M + M^T = 11^T, \text{ and } \Delta_{k,\,h} \geq \alpha \sqrt{\frac{\log n}{n p_{\mathrm{obs}} r}} \right\}. \tag{3.11b}$$

As before, we frequently adopt the shorthand $\mathcal{F}_{k,h}(\alpha)$, with the dependence on (n, p_{obs}, r) being understood implicitly.

Theorem 8. *(a) For any $\alpha \geq 8$, the maximum pairwise win set $\widetilde{\mathcal{S}}_k$ satisfies*

$$\sup_{M \in \mathcal{F}_{k,h}(\alpha)} \mathbb{P}_M\big[D_{\mathrm{H}}(\widetilde{\mathcal{S}}_k, \mathcal{S}^*_k) > 2h\big] \leq \frac{1}{n^{14}}. \tag{3.12a}$$

(b) Conversely, in the regime $p_{\mathrm{obs}} \geq \frac{\log n}{2nr}$ and for given constants $\nu_1, \nu_2 \in (0,1)$, suppose that $2h \leq \frac{1}{1+\nu_2} \min\{n^{1-\nu_1}, k, n-k\}$. Then for any $\alpha \leq \frac{\sqrt{\nu_1 \nu_2}}{14}$, any estimator $\widehat{\mathcal{S}}_k$ has error at least

$$\sup_{M \in \mathcal{F}_{k,h}(\alpha)} \mathbb{P}_M\big[D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}^*_k) > 2h\big] \geq \frac{1}{7}, \tag{3.12b}$$

for all n larger than a constant $c(\nu_1, \nu_2)$.

[1] The requirement $h < k$ is sensible because if $h \geq k$, the problem is trivial: any two k-sized sets $\widehat{\mathcal{S}}_k$ and $\mathcal{S}^*_k$ satisfy the bound $D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}^*_k) \leq 2k \leq 2h$.

This result is similar to that of Theorem 7, except that the relaxation of the exact recovery condition allows for a less constrained definition of the separation threshold $\Delta_{k,h}$. As with Theorem 7, the lower bound in part (b) applies even if probability matrix M is restricted to lie in a parameter-based model (such as the BTL model), or the more general SST class. The counting algorithm is thus optimal for estimation under the relaxed Hamming metric as well.

Finally, it is worth making a few comments about the constants appearing in these claims. We can weaken the lower bound on Δ_k required in Theorem 8(a) at the expense of a lower probability of success; for instance, if we instead require that $\alpha \geq 4$, then the probability of error is guaranteed to be at most n^{-2}. Subsequently in the chapter, we provide the results of simulations with $n = 500$ items and $\alpha = 4$. On the other hand, in Theorem 8(b), if we impose the stronger upper bound $\alpha = \mathcal{O}(1/\sqrt{h \log n})$, then we can remove the condition $h \leq n^{1-\nu_1}$.

3.3.4 An abstract form of k-set recovery

In earlier sections, we investigated recovery of the top k items either exactly or under a Hamming error. Exact recovery may be quite strict for certain applications, whereas the property of Hamming error allowing for a few of the top k items to be replaced by *arbitrary* items may be undesirable. Indeed, many applications have requirements that go beyond these metrics; for instance, see the papers [8, 73, 106, 126, 166, 168] and references therein for some examples. In this section, we generalize the notion of exact or Hamming-error recovery in order to accommodate a fairly general class of requirements.

Both the exact and approximate Hamming recovery settings require the estimator to output a set of k items that are either exactly or approximately equal the true set of top k items. When is the estimate deemed successful? One way to think about the problem is as follows. The specified requirement of exact or approximate Hamming recovery is associated to a set of k-sized subsets of the n possible ranks. The estimator is deemed successful if the true ranks of the chosen k items equals one of these subsets. In our notion of generalized recovery, we refer to such sets as *allowed sets*. For example, in the case $k = 3$, we might say that the set $\{1, 4, 10\}$ is allowed, meaning that an output consisting of the "first", "fourth" and "tenth" ranked items is considered correct.

In more generality, let $\mathfrak{S}$ denote a family of k-sized subsets of $[n]$, which we refer to as *family of allowed sets*. Notice that any allowed set is defined by the *positions* of the items in the true ordering and not the items themselves.[2] Once some true underlying ordering of the n items is fixed, each element of the family $\mathfrak{S}$ then specifies a set of the items themselves. We use these two interpretations depending on the context — the definition in terms of positions to specify the requirements, and the definition in terms of the items to evaluate an estimator for a given underlying probability matrix M.

[2] In case of two or more items with identical scores, the choice of any of these items is considered valid.

We let $\mathcal{S}_k^\dagger$ denote a k-set estimate, meaning a function that given a set of observations as input, returns a k-sized subset of $[n]$ as output.

Definition 6 ($\mathfrak{S}$-respecting estimators)**.** *For any family $\mathfrak{S}$ of allowed sets, a k-set estimate $\mathcal{S}_k^\dagger$* respects *its structure if the set of k positions of the items in $\mathcal{S}_k^\dagger$ belongs to the set family $\mathfrak{S}$.*

Our goal is to determine conditions on the set family $\mathfrak{S}$ under which there exist estimators $\mathcal{S}_k^\dagger$ that respect its structure. In order to illustrate this definition, let us return to the examples treated thus far:

Example 1 (Exact and approximate Hamming recovery)**.** *The requirement of exact recovery of the top k items has $\mathfrak{S}$ consisting of exactly one set, the set of the top k positions $\mathfrak{S} = \{[k]\}$. In the case of recovery with a Hamming error at most $2h$, the set $\mathfrak{S}$ of all allowed sets consists all k-sized subsets of $[n]$ that contain at least $(k-h)$ positions in the top k positions. For instance, in the case $h = 1$, $k = 2$ and $n = 4$, we have*

$$\mathfrak{S} = \Big\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\}\Big\}.$$

Apart from these two requirements, there are several other requirements for top-k recovery popular in the literature [8, 39, 73, 106, 126, 166, 168]. Let us illustrate them with another example:

Example 2. *Let $\pi^* : [n] \rightarrow [n]$ denote the true underlying ordering of the n items. The following are four popular requirements on the set $\mathcal{S}_k^\dagger$ for top-k identification, with respect to the true permutation π^*, for a pre-specified parameter $\epsilon \geq 0$.*

(i) All items in the set $\mathcal{S}_k^\dagger$ must be contained contained within the top $(1+\epsilon)k$ entries:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon)k. \tag{3.13a}$$

(ii) The rank of any item in the set $\mathcal{S}_k^\dagger$ must lie within a multiplicative factor $(1+\epsilon)$ of the rank of any item not in the set $\mathcal{S}_k^\dagger$:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon) \min_{j \in [n]\backslash \mathcal{S}_k^\dagger} \pi(j). \tag{3.13b}$$

(iii) The rank of any item in the set $\mathcal{S}_k^\dagger$ must lie within an additive factor ϵ of the rank of any item not in the set $\mathcal{S}_k^\dagger$:

$$\max_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq \min_{j \in [n]\backslash \mathcal{S}_k^\dagger} \pi^*(j) + \epsilon. \tag{3.13c}$$

(iv) The sum of the ranks of the items in the set $\mathcal{S}_k^\dagger$ must be contained within a factor $(1+\epsilon)$ of the sums of ranks of the top k entries:

$$\sum_{i \in \mathcal{S}_k^\dagger} \pi^*(i) \leq (1+\epsilon)\frac{1}{2}k(k+1). \tag{3.13d}$$

Note that each of these requirements reduces to the exact recovery requirement when $\epsilon = 0$. Moreover, each of these requirements can be rephrased in terms of families of allowed sets. For instance, if we focus on requirement (i), then any k-sized subset of the top $(1+\epsilon)k$ positions is an allowable set.

In this chapter, we derive conditions that govern k-set recovery for allowable set systems that satisfy a natural "monotonicity" condition. Informally, the monotonicity condition requires that the set of k items resulting from replacing an item in an allowed set with a higher ranked item must also be an allowed set. More precisely, for any set $\{t_1, \ldots, t_k\} \subseteq [n]$, let $\Lambda(\{t_1, \ldots, t_k\}) \subseteq 2^{[n]}$ be the set defined by all of its monotone transformations—that is

$$\Lambda(\{t_1, \ldots, t_k\}) := \Big\{\{t'_1, \ldots, t'_k\} \subseteq [n] \mid t'_j \leq t_j \text{ for every } j \in [k]\Big\}.$$

Using this notation, we have the following:

Definition 7 (Monotonic set systems). *The set $\mathfrak{S}$ of allowed sets is a* monotonic set system *if*

$$\Lambda(T) \subseteq \mathfrak{S} \quad \textit{for every } T \in \mathfrak{S}. \tag{3.14}$$

One can verify that condition (3.14) is satisfied by the settings of exact and Hamming-error recovery, as discussed in Example 1. The condition is also satisfied by all four requirements discussed in Example 2.

The following theorem establishes conditions under which one can (or cannot) produce an estimator that respects an allowable set requirement. In order to state it, recall the score $\tau_i := \frac{1}{n}\sum_{j=1}^n M_{ij}$, as previously defined in equation (3.1) for each $i \in [n]$. For notational convenience, we also define $\tau_i := -\infty$ for every $i > n$. Consider any monotonic family of allowed sets $\mathfrak{S}$, and for some integer $\beta \geq 1$, let $T^1, \ldots, T^\beta \in \mathfrak{S}$ such that $\mathfrak{S} = \underset{b \in [\beta]}{\cup} \Lambda(T^b)$. For every $b \in [\beta]$, let $t_1^b < \cdots < t_k^b$ denote the entries of T^b. We then define the critical threshold based on the scores:

$$\Delta_{\mathfrak{S}} := \max_{b \in [\beta]} \min_{j \in [k]} (\tau_{(j)} - \tau_{(k + t_j^b - j + 1)}). \tag{3.15}$$

The term $\Delta_{\mathfrak{S}}$ is a further generalization of the quantities Δ_k and $\Delta_{k,h}$ defined in earlier sections. We also define a generalization $\mathcal{F}_{\mathfrak{S}}(\cdot)$ of the families $\mathcal{F}_k(\cdot)$ and $\mathcal{F}_k(\cdot)$ as

$$\mathcal{F}_{\mathfrak{S}}(\alpha; n, p_{\text{obs}}, r) := \left\{ M \in [0,1]^{n \times n} \mid M + M^T = 11^T \text{ and } \Delta_{\mathfrak{S}} \geq \alpha\sqrt{\frac{\log n}{n p_{\text{obs}} r}} \right\}. \tag{3.16}$$

As before, we use the shorthand $\mathcal{F}_{\mathfrak{S}}(\alpha)$, with the dependence on (n, p_{obs}, r) being understood implicitly.

Theorem 9. *Consider any allowable set requirement specified by a monotonic set class $\mathfrak{S}$.*

(a) For any $\alpha \geq 8$, the maximum pairwise win set $\widetilde{\mathcal{S}}_k$ satisfies

$$\sup_{M \in \mathcal{F}_{\mathfrak{S}}(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \notin \mathfrak{S}\big] \leq \frac{1}{n^{13}}.$$

(b) Conversely, in the regime $p_{\mathrm{obs}} \geq \frac{\log n}{2nr}$, and for given constants $\mu_1 \in (0,1), \mu_2 \in (\frac{3}{4}, 1]$, suppose that $\max_{b \in [\beta]} t^b_{\lceil \mu_2 k \rceil} \leq \frac{n}{2}$ and $8(1-\mu_2)k \leq n^{1-\mu_1}$. Then for any α smaller than a constant $c_1(\mu_1, \mu_2) > 0$, any estimator $\widehat{\mathcal{S}}_k$ has error at least

$$\sup_{M \in \mathcal{F}_{\mathfrak{S}}(\alpha)} \mathbb{P}_M\big[\widetilde{\mathcal{S}}_k \notin \mathfrak{S}\big] \geq \frac{1}{15}, \tag{3.17}$$

for all n larger than a constant $c_4(\mu_1, \mu_2)$.

A few remarks on the lower bound are in order. First, the lower bound continues to hold even if the probability matrix M is restricted to follow a parameter-based model such as BTL or restricted to lie in the SST class. Second, in terms of the threshold for α, the lower bound holds with $c_1(\mu_1, \mu_2) = \frac{1}{15}\sqrt{\mu_1 \min\big\{\frac{1}{4(1-\mu_2)-1}, \frac{1}{2}\big\}}$. Third, it is worth noting that one must necessarily impose some conditions for the lower bound, along the lines of those required in Theorem 9(b) for the allowable sets to be "interesting" enough.

As a concrete illustration, consider the requirement defined by the parameters $b = 1$, $k = 1$ and $\mathfrak{S} = \Lambda(\{n - \sqrt{n}\})$. For $\mu_1 = \mu_2 = \frac{9}{10}$, this requirement satisfies the condition $8(1-\mu_2)k \leq n^{1-\mu_1}$ but violates the condition $t_{\lceil \mu_2 k \rceil} \leq \frac{n}{2}$. Now, a selection of $k = 1$ item made uniformly at random (independent of the data) satisfies this allowable set requirement with probability $1 - \frac{1}{\sqrt{n}}$. Given the success of such a random selection algorithm in this parameter regime, we see that the lower bounds therefore cannot be universal, but must require some conditions on the allowable sets.

3.4 Simulations and experiments

In this section, we empirically evaluate the performance of the counting algorithm and compare it with the Spectral MLE algorithm via simulations on synthetic data, as well as on datasets from the Amazon Mechanical Turk crowdsourcing platform.

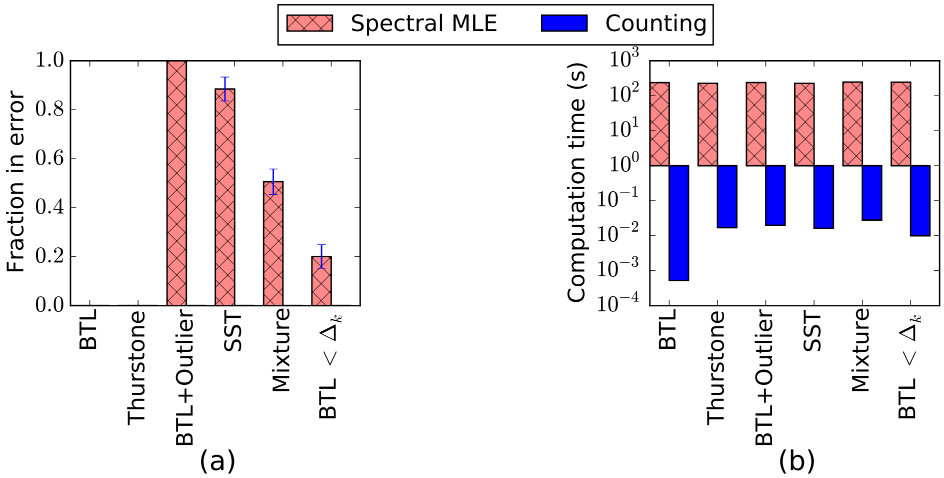


Figure 3.1: Simulation results comparing Spectral MLE and the counting algorithm in terms of error rates for exact recovery of the top k items, and computation time. (a) Histogram of fraction of instances where the algorithm failed to recover the k items correctly, with each bar being the average value across 50 trials. The counting algorithm has 0% error across all problems, while the spectral MLE is accurate for parameter-based models (BTL, Thurstone), but increasingly inaccurate for other models. (b) Histogram plots of the maximum computation time taken by the counting algorithm and the minimum computation time taken by Spectral MLE across all trials. Even though this maximum-to-minimum comparison is unfair to the counting algorithm, it involves five or more orders of magnitude less computation.

3.4.1 Simulations

We begin with simulations using synthetically generated data with $n = 500$ items and observation probability $p_{\text{obs}} = 1$, and with pairwise comparison models ranging over six possible types. Panel (a) in Figure 3.1 provides a histogram plot of the associated error rates (with a bar for each one of these six models) in recovering the $k = n/4 = 125$ items for the counting algorithm versus the Spectral MLE algorithm. Each bar corresponds to the average over 50 trials. Panel (b) compares the CPU times of the two algorithms. The value of α (and in turn, the value of r) in the first five models is as derived in Section 3.3.2. In more detail, the six model types are given by:

(I) *Bradley-Terry-Luce (BTL) model:* Recall that the theoretical guarantees for the Spectral MLE algorithm [46] are applicable to data that is generated from the BTL model (3.3a), and as guaranteed, the Spectral MLE algorithm gives a 100% accuracy under this model. The counting algorithm also obtains a 100% accuracy, but importantly, the counting algorithm requires a computational time that is five orders of magnitude lower than that of Spectral MLE.

(II) *Thurstone model:* The Thurstone model [253] is another parameter-based model, with the function F in equation (3.3b) set as the cumulative distribution function of the standard Gaussian distribution. Both Spectral MLE and the counting algorithm gave 100% accuracy under this model.

(III) *BTL parameter-based model with one (non-transitive) outlier:* This model is identical to BTL, with one modification. Comparisons among $(n-1)$ of the items follow the BTL parameter-based model as before, but the remaining item always beats the first $\frac{n}{4}$ items and always loses to each of the other items. We see that the counting algorithm continues to achieve an accuracy of 100% as guaranteed by Theorem 7. The departure from the BTL model however prevents the Spectral MLE algorithm from identifying the top k items.

(IV) *Strong stochastic transitivity (SST) model:* We simulate the "independent bands" construction of Section 2.4 (from Chapter 2) in the SST class. Spectral MLE is often unsuccessful in recovering the top k items, while the counting algorithm always succeeds.

(V) *Mixture of BTL models:* Consider two sets of people with opposing preferences. The first set of people have a certain ordering of the items in their mind and their preferences follow a BTL parameter-based model under this ordering. The second set of people have the opposite ordering, and their preferences also follow the BTL model under this opposite ordering. The overall preference probabilities is a mixture between these two sets of people. In the simulations, we observe that the counting algorithm is always successful while the Spectral MLE method often fails.

(VI) *BTL with violation of separation condition:* We simulate the BTL parameter-based model, but with a choice of parameter r small enough that the value of α is about one-tenth of its recommended value in Section 3.3.2. We observe that the counting algorithm incurs lower errors than the Spectral MLE algorithm, thereby demonstrating its robustness.

To summarize, the performance of the two algorithms can be contrasted in the following way. When our stated lower bounds on α are satisfied, then consistent with our theoretical claims, the Copeland counting algorithm succeeds irrespective of the form of the pairwise probability distributions. The Spectral MLE algorithm performs well when the pairwise comparison probabilities are faithful to parameter-based models, but is often unsuccessful otherwise. Even when the condition on α is violated, the performance of the counting algorithm remains superior to that of the Spectral MLE.[3] In terms of computational complexity, for every instance we simulated, the counting algorithm took several orders of magnitude less time as compared to Spectral MLE.

[3] Note that part (b) of Theorem 7 is a minimax converse meaning that it appeals to the worst case scenario.

3.4.2 Experiments

In this section, we describe experiments on real world datasets from the Amazon Mechanical Turk commercial crowdsourcing platform.

The first experiment we consider is using data from a cell-counting experiment where we use the ground truth from [40] and crowdsourced data from [234]. In this experiment, there are $n = 23$ images containing a number of cells, and workers are asked to compare randomly chosen pairs of images and indicate the image they think has a fewer number of cells. There are a total of 704 pairwise comparisons in this dataset, and the fraction of incorrect pairwise comparisons as compared to the ground truth is 10%. For three other experiments, we used the "cardinal versus ordinal" dataset from [220]; three of the experiments performed in that paper are suitable for the evaluations here—namely, ones in which each question has a ground truth, and the pairs of items are chosen uniformly at random. The three experiments tested the workers' general knowledge, audio, and visual understanding, and the respective tasks involved: (i) identifying the pair of cities with a greater geographical distance, (ii) identifying the higher frequency key of a piano, and (iii) identifying spelling mistakes in a paragraph of text. The number of items n in the three experiments were 16, 10 and 8 respectively. The total number of pairwise comparisons were 408, 265 and 184 respectively. The fraction of pairwise comparisons whose outcomes were incorrect (as compared to the ground truth) in the raw data are 17%, 20% and 40% respectively.

We compared the performance of the counting algorithm with that of the Spectral MLE algorithm for these four experiments under various metrics. Figure 3.2 shows the results of the experiments – the error bars are produced via 50 iterations of subsampling the data and executing the estimators on this subsampled data. We see that the counting algorithm almost always outperforms Spectral MLE. Moreover, the Spectral MLE algorithm required about 5 orders of magnitude more computation time as compared to counting. The counting algorithm thus performs well on simulated as well as real data. It outperforms Spectral MLE not only when the number of items is large (as in the simulations) but also when the problem sizes are small as seen in these experiments.

3.5 Discussion

In this chapter, we analyzed the problem of recovering the k most highly ranked items based on observing noisy comparisons. We proved that an algorithm that simply selects the items that win the maximum number of comparisons is, up to constant factors, an information-theoretically optimal procedure. Our results also extend to recovering the entire ranking of the items as a corollary. In empirical evaluations, this algorithm takes several orders of magnitude lower computation time while providing higher accuracy as compared to prior work. The results of this chapter thus underscore the philosophy of Occam's razor that the simplest answer is often correct.

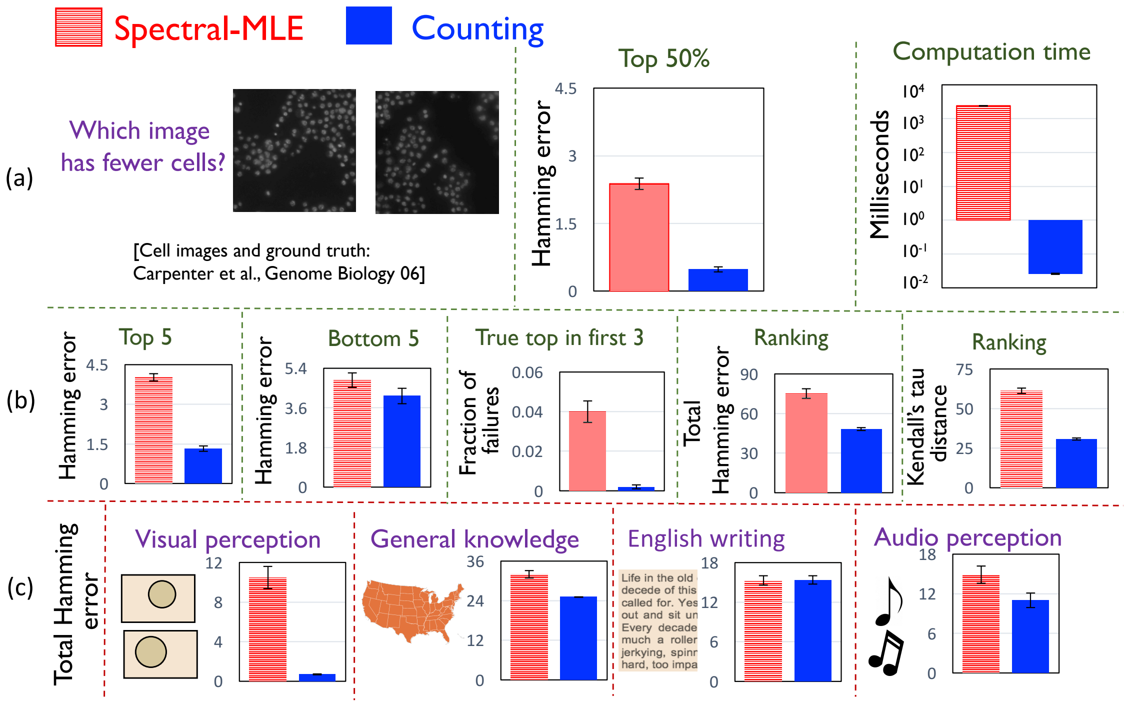


Figure 3.2: Evaluation of Spectral MLE and the counting algorithm on four datasets from Amazon Mechanical Turk. (a) Cell-counting experiment example, Hamming error for top 50%-recovery, and computation time; (b) Other requirements and metrics for the cell-counting experiment; (c) Four other experiments.

We conclude this chapter by discussing an interesting and practically useful open problem that arises from our work. The notion of allowable sets introduced in this chapter apply to recovery of k-sized subsets of the items; such a formulation and associated results may apply to recovery of partial or total orderings of the items. A parallel line of literature (e.g., [30, 100, 110, 118]) studies settings in which the pairs to be compared can be chosen sequentially in a data-dependent manner, but to the best of our knowledge, this line of literature considers only the metric of exact recovery of the top k items. It is of interest to investigate the Hamming and allowable set recovery problems in such an active setting.

3.6 Proofs

We now turn to the proofs of our main results. We continue to use the notation $[i]$ to denote the set $\{1, \ldots, i\}$ for any integer $i \geq 1$. We ignore floor and ceiling conditions unless critical to the proof.

3.6.1 Proof of Theorem 7: Exact recovery of top k items

We begin with the proof of Theorem 7, dividing our argument into two parts.

Proof of part (a)

For any pair of items (i, j), let us encode the outcomes of the r trials by an i.i.d. sequence $V_{ij}^{(\ell)} = [X_{ij}^{(\ell)} \quad X_{ji}^{(\ell)}]^T$ of random vectors, indexed by $\ell \in [r]$. Each random vector follows the distribution

$$\mathbb{P}\big[x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}\big] = \begin{cases} 1 - p_{\text{obs}} & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (0,0) \\ p_{\text{obs}} M_{ij} & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (1,0) \\ p_{\text{obs}}(1 - M_{ij}) & \text{if } (x_{ij}^{(\ell)},\, x_{ji}^{(\ell)}) = (0,1) \\ 0 & \text{otherwise.} \end{cases}$$

With this encoding, the variable $W_a := \sum_{\ell \in [r]} \sum_{z \in [n]\backslash\{a\}} X_{aj}^{(r)}$ encodes the number of wins for item a.

Consider any item $a \in \mathcal{S}_k^*$ which ranks among the top k in the true underlying ordering, and any item $b \in [n]\backslash\mathcal{S}_k^*$ which ranks outside the top k. We claim that with high probability, item a will win more pairwise comparisons than item b. More precisely, let $\mathcal{E}_{ba}$ denote the event that item b wins at least as many pairwise comparisons than a. We claim that

$$\mathbb{P}(\mathcal{E}_{ba}) \overset{(i)}{\leq} \exp\left(-\frac{\frac{1}{2}(rp_{\text{obs}} n\Delta_k)^2}{rp_{\text{obs}} n(2 - \Delta_k) + \frac{2}{3} rp_{\text{obs}} n\Delta_k}\right) \overset{(ii)}{\leq} \frac{1}{n^{16}}. \tag{3.18}$$

Given this bound, the probability that the counting algorithm will rank item b above a is no more than n^{-16}. Applying the union bound over all pairs of items $a \in \mathcal{S}_k^*$ and $b \in [n]\backslash\mathcal{S}_k^*$ yields $\mathbb{P}\big[\widetilde{\mathcal{S}}_k \neq \mathcal{S}_k^*\big] \leq n^{-14}$ as claimed.

We note that inequality (ii) in equation (3.18) follows from inequality (i) combined with the condition on Δ_k that arises by setting $\alpha \geq 8$ as assumed in the hypothesis of the theorem. Thus, it remains to prove inequality (i) in equation (3.18). By definition of $\mathcal{E}_{ba}$, we have

$$\mathbb{P}(\mathcal{E}_{ba}) = \mathbb{P}\Big(\underbrace{\sum_{\ell \in [r]} \sum_{z \in [n]\backslash\{b\}} X_{bz}^{(\ell)}}_{W_b} - \underbrace{\sum_{\ell \in [r]} \sum_{z \in [n]\backslash\{a\}} X_{az}^{(\ell)}}_{W_a} \geq 0\Big). \tag{3.19}$$

It is convenient to recenter the random variables. For every $\ell \in [r]$ and $z \in [n]\backslash\{a, b\}$, define the zero-mean random variables

$$\overline{X}_{az}^{(\ell)} = X_{az}^{(\ell)} - \mathbb{E}[X_{az}^{(\ell)}] = X_{az}^{(\ell)} - p_{\text{obs}} M_{az} \quad \text{and} \quad \overline{X}_{bz}^{(\ell)} = X_{bz}^{(\ell)} - \mathbb{E}[X_{bz}^{(\ell)}] = X_{bz}^{(\ell)} - p_{\text{obs}} M_{bz}.$$

Also, let

$$\overline{X}_{ab}^{(\ell)} = (X_{ab}^{(\ell)} - X_{ba}^{(\ell)}) - \mathbb{E}[X_{ab}^{(\ell)} - X_{ba}^{(\ell)}] \;=\; (X_{ab}^{(\ell)} - X_{ba}^{(\ell)}) - (p_{\text{obs}} M_{ab} - p_{\text{obs}} M_{ba}).$$

We then have

$$\mathbb{P}(\mathcal{E}_{ba}) = \mathbb{P}\Bigg(\sum_{\ell\in[r]}\Big(\sum_{z\in[n]\backslash\{a,b\}} \overline{X}_{bz}^{(\ell)} - \sum_{z\in[n]\backslash\{a,b\}} \overline{X}_{az}^{(\ell)} - \overline{X}_{ab}^{(\ell)}\Big) \geq rp_{\text{obs}} \sum_{z\in[n]} \Big(M_{az} - M_{bz}\Big)\Bigg).$$

Since $a \in \mathcal{S}_k^*$ and $b \in [n]\backslash\mathcal{S}_k^*$, from the definition of Δ_k, we have $n\Delta_k \leq \sum\limits_{z\in[n]} (M_{az} - M_{bz})$, and consequently

$$\mathbb{P}\left(\mathcal{E}_{ba}\right) \leq \mathbb{P}\left(\sum_{\ell\in[r]}\Big(\sum_{z\in[n]\backslash\{a,b\}} \overline{X}_{bz}^{(\ell)} - \sum_{z\in[n]\backslash\{a,b\}} \overline{X}_{az}^{(\ell)} - \overline{X}_{ab}^{(\ell)}\Big) \geq rp_{\text{obs}} n\Delta_k\right). \tag{3.20}$$

By construction, all the random variables in the above inequality are zero-mean, mutually independent, and bounded in absolute value by 2. These properties alone would allow us to obtain a tail bound by Hoeffding's inequality; however, in order to obtain the stated result (3.18), we need the more refined result afforded by Bernstein's inequality (e.g., [19]). In order to derive a bound of Bernstein type, the only remaining step is to bound the second moments of the random variables at hand. Some straightforward calculations yield

$$\mathbb{E}[(-\overline{X}_{az}^{(\ell)})^2] \leq p_{\text{obs}} M_{az}, \qquad \mathbb{E}[(\overline{X}_{bz}^{(\ell)})^2] \leq p_{\text{obs}} M_{bz}, \quad \text{and} \quad \mathbb{E}[(\overline{X}_{ab}^{(\ell)})^2] \leq p_{\text{obs}} M_{ab} + p_{\text{obs}} M_{ba}.$$

It follows that

$$\begin{aligned}\sum_{z\in[n]\backslash\{a,b\}} \mathbb{E}[(-\overline{X}_{az}^{(\ell)})^2] + &\sum_{z\in[n]\backslash\{a,b\}} \mathbb{E}[(\overline{X}_{bz}^{(\ell)})^2] + \mathbb{E}[(\overline{X}_{ab}^{(\ell)})^2] \\ &\leq p_{\text{obs}}\left(\sum_{z\in[n]\backslash\{a,b\}} (M_{az} + M_{bz}) + M_{ab} + M_{ba}\right) \\ &\overset{(iii)}{\leq} p_{\text{obs}}\left(2\sum_{z\in[n]} M_{az} - n\Delta_k\right) \\ &\overset{(iv)}{<} p_{\text{obs}} n(2 - \Delta_k),\end{aligned}$$

where the inequality (iii) follows from the definition of Δ_k, and step (iv) follows because $M_{az} \leq 1$ for every z and $M_{aa} = \frac{1}{2}$. Applying the Bernstein inequality now yields the stated bound (3.18)(i).

Proof of part (b)

The symmetry of the problem allows us to assume, without loss of generality, that $k \leq \frac{n}{2}$. We prove a lower bound by first constructing a ensemble of $n - k + 1$ different problems, and considering the problem of distinguishing between them. For each $a \in \{k-1, k, \ldots, n\}$, let us define the k-sized subset $\mathcal{S}^*[a] := \{1, \ldots, k-1\} \cup \{a\}$, and the associated matrix of pairwise probabilities

$$M^a_{ij} := \begin{cases} \frac{1}{2} & \text{if } i, j \in \mathcal{S}^*[a], \text{ or } i, j \notin \mathcal{S}^*[a] \\ \frac{1}{2} + \delta & \text{if } i \in \mathcal{S}^*[a] \text{ and } j \notin \mathcal{S}^*[a] \\ \frac{1}{2} - \delta & \text{if } i \notin \mathcal{S}^*[a] \text{ and } j \in \mathcal{S}^*[a], \end{cases}$$

where $\delta \in (0, \frac{1}{2})$ is a parameter to be chosen. We use $\mathbb{P}^a$ to denote probabilities taken under pairwise comparisons drawn according to the model M^a.

One can verify that the construction above falls in the intersection of parameter-based models and the SST model. In the case of parameter-based models, this construction amounts to having the parameters associated to every item in $\mathcal{S}^*_k$ to have the same value, and those associated to every item in $[n]\backslash\mathcal{S}^*_k$ to have the same value. Also observe that for every such distribution $\mathbb{P}^a$, the associated k-separation threshold $\Delta_k = \delta$.

Any given set of observations can be described by the collection of random variables $Y = \{Y^{(\ell)}_{ij}, j > i \in [n], \ell \in [r]\}$. When the true underlying model is $\mathbb{P}^a$, the random variable $Y^{(\ell)}_{ij}$ follows the distribution

$$Y^{(\ell)}_{ij} = \begin{cases} 0 & \text{with probability } 1 - p_{\text{obs}} \\ i & \text{with probability } p_{\text{obs}} M^a_{ij} \\ j & \text{with probability } p_{\text{obs}}(1 - M^a_{ij}). \end{cases}$$

The random variables $\{Y^{(\ell)}_{ij}\}_{i,j \in [n], i<j, \ell \in [r]}$ are mutually independent, and the distribution $\mathbb{P}^a$ is a product distribution across pairs $\{i > j\}$ and repetitions $\ell \in [r]$.

Let $A \in \{k, \ldots, n\}$ follow a uniform distribution over the index set, and suppose that given $A = a$, our observations Y has components drawn according to the model $\mathbb{P}^a$. Consequently, the marginal distribution of Y is the mixture distribution $\frac{1}{L}\sum_{a=1}^{L} \mathbb{P}^a$ over all $L = n - k + 1$ models. Based on observing Y, our goal is to recover the correct index $A = a$ of the underlying model, which is equivalent to recovering the planted subset $\mathcal{S}^*[a]$. We use the Fano bound (2.22) to lower bound the error bound associated with any test for this problem. In order to apply Fano's inequality, the following result provides control over the Kullback-Leibler divergence between any pair of probabilities involved.

Lemma 15. *For any distinct pair $a, b \in \{k, \ldots, n\}$, we have*

$$D_{\text{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq \frac{2np_{\text{obs}}r}{\frac{1}{4\delta^2} - 1}. \tag{3.21}$$

See the end of this section for the proof of this claim.

Given this bound on the Kullback-Leibler divergence, Fano's inequality (2.22) implies that any estimator ϕ of A has error probability lower bounded as

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{\frac{2np_{\text{obs}}r}{\frac{1}{4\delta^2}-1} + \log 2}{\log(n-k+1)} \;\geq\; \frac{1}{7}.$$

Here the final inequality holds whenever $\delta \leq \frac{1}{7}\sqrt{\frac{\log n}{np_{\text{obs}}r}}$, $p_{\text{obs}} \geq \frac{\log n}{2nr}$, $n \geq 7$ and $k \leq \frac{n}{2}$. The condition $p_{\text{obs}} \geq \frac{\log n}{2nr}$ also ensures that $\delta < \frac{1}{2}$ thereby ensuring that our construction is valid. It only remains to prove Lemma 15.

Proof of Lemma 15

Since the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$ are formed by components that are independent across edges $i > j$ and repetitions $\ell \in [r]$, we have

$$D_{\text{KL}}(\mathbb{P}^a \| \mathbb{P}^b) = \sum_{\ell \in [r]} \sum_{1 \leq i < j \leq n} D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(\ell)}) \| \mathbb{P}^b(X_{ij}^{(\ell)})) \;=\; r \sum_{1 \leq i < j \leq n} D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})),$$

where the second equality follows since the r trials are all independent and identically distributed.

We now evaluate each individual term in right hand side of the above equation. Consider any $i, j \in [n]$. We divide our analysis into three disjoint cases:
Case I: Suppose that $i, j \in [n] \backslash \{a, b\}$. The distribution of $X_{ij}^{(1)}$ is identical across the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$. As a result, we find that

$$D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) = 0.$$

Case II: Suppose that $i = a,\ j \in [n] \backslash \{a, b\}$ or $i = b,\ j \in [n] \backslash \{a, b\}$. We then have

$$D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq p_{\text{obs}} \frac{\delta^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Case III: Suppose that $i = a,\ j = b$. We then have

$$D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq p_{\text{obs}} \frac{(2\delta)^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Combining the bounds from all three cases, we find that the KL divergence is upper bounded as

$$\frac{1}{r} D_{\text{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 2(n-2) p_{\text{obs}} \frac{\delta^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)} + p_{\text{obs}} \frac{(2\delta)^2}{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)}.$$

Some simple algebraic manipulations yield the claimed result.

3.6.2 Proof of Corollary 1: Ranking

We now turn to the proof of Corollary 1. Beginning with the claim of sufficiency, it is easy to see that the ranking is correctly recovered whenever the top k items are correctly recovered for every value of $k \in [n]$. Consequently, one can apply the union bound to (3.9a) over all values of $k \in [n]$ and this gives the desired upper bound.

Now turning to the claim of necessity, we first introduce some notation to aid in subsequent discussion. Defining the parameter $\Delta_0 := \min_{j\in[n-1]}(\tau_{(j)} - \tau_{(j+1)})$, we have shown that the lower bound

$$\Delta_0 \geq 8\sqrt{\frac{\log n}{np_{\text{obs}}r}}$$

is sufficient to guarantee exact recovery of the full ranking. Further, one must also have

$$\Delta_0 \leq \frac{1}{n-1}\sum_{j=1}^{n-1}(\tau_{(j)} - \tau_{(j+1)}) = \frac{1}{n-1}(\tau_{(1)} - \tau_{(n)}) \leq \frac{1}{n-1}.$$

Here we show that these two requirements are tight up to constant factors, meaning that for any value of Δ_0 satisfying $\Delta_0 \leq \frac{1}{9}\sqrt{\frac{\log n}{np_{\text{obs}}r}}$ and $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$, there are instances where recovery of the underlying ranking fails with probability at least $\frac{1}{70}$ for any estimator.

Consider the following ensemble of $(n-1)$ different problems, indexed by $a \in [n-1]$. For every value of $a \in [n-1]$, define a permutation π^a of the n items as

$$\pi^a(i) = \begin{cases} i+1 & \text{if } i = a \\ i-1 & \text{if } i = a+1 \\ i & \text{otherwise.} \end{cases}$$

In words, the permutation π^a equals the identity permutation except for the swapping of items a and $(a+1)$. Define an associated matrix of pairwise-comparison probabilities M^a as

$$M^a_{ij} = \frac{1}{2} - (\pi^a(i) - \pi^a(j))\Delta_0,$$

and $M^a_{ji} = 1 - M^a_{ij}$. Let $\mathbb{P}^a$ denote the probabilities taken under pairwise comparisons drawn according to the model M^a. The condition $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$ ensures that this construction is a valid probability distribution. One can then compute that under distribution $\mathbb{P}^a$, the score τ^a_i of any item i equals

$$\tau^a_i = \frac{1}{2} - \big(\pi^a(i) - \frac{n+1}{2}\big)\Delta_0.$$

One can also verify that for any $a \in [n-1]$, and any $i \in [n-1]$, we have

$$\tau^a_{\pi^a(i)} - \tau^a_{\pi^a(i+1)} = \Delta_0,$$

where we have used the fact that $\pi^a(\pi^a(i)) = i$. The requirement imposed by the hypothesis is thus satisfied.

We now use Fano's inequality (2.22) obtain the claimed lower bound. In order to apply this result, we first obtain an upper bound on the Kullback-Leibler divergence between the probability distributions of the observed data under any pair of problems constructed above.

Lemma 16. *For any distinct pair $a, b \in [n-1]$, we have*

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 50 n p_{\mathrm{obs}} r \Delta_0^2.$$

See the end of this section for the proof of this claim.

Given this bound on the Kullback-Leibler divergence, the Fano bound (2.22) implies that any method ϕ for identifying the true ranking has error probability

$$\mathbb{P}[\phi(Y) \neq A] \geq 1 - \frac{50 n p_{\mathrm{obs}} r \Delta_0^2 + \log 2}{\log(n-1)} \geq \frac{1}{70},$$

where the final inequality holds whenever $\Delta_0 \leq \frac{1}{9}\sqrt{\frac{\log n}{n p_{\mathrm{obs}} r}}$ and $n \geq 9$.

The only remaining detail is the proof of Lemma 16.

Proof of Lemma 16

Since the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$ are formed by components that are independent across edges $i > j$ and repetitions $\ell \in [r]$, we have

$$D_{\mathrm{KL}}(\mathbb{P}^a \| \mathbb{P}^b) = \sum_{\ell \in [r]} \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(\ell)}) \| \mathbb{P}^b(X_{ij}^{(\ell)})) \; = \; r \sum_{1 \leq i < j \leq n} D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})),$$

where the second equality follows since the r trials are all independent and identically distributed.

We now evaluate each individual term in right hand side of the above equation. Consider any $i, j \in [n]$. We divide our analysis into three disjoint cases:
<u>Case I</u>: Suppose that $i, j \in [n] \backslash \{a, a+1, b, b+1\}$. The distribution of $X_{ij}^{(1)}$ is identical across the distributions $\mathbb{P}^a$ and $\mathbb{P}^b$. As a result, we find that

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) = 0.$$

<u>Case II</u>: Alternatively, suppose $i \in \{a, a+1, b, b+1\}$ and $j \in [n] \backslash \{a, a+1, b, b+1\}$ or if $j \in \{a, a+1, b, b+1\}$ and $i \in [n] \backslash \{a, a+1, b, b+1\}$. Then we have

$$D_{\mathrm{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq 5 p_{\mathrm{obs}} \Delta_0^2,$$

where we have used the fact that $\mathbb{P}^a(X_{ij}^{(1)})$ and $\mathbb{P}^b(X_{ij}^{(1)})$ both take values in $[\frac{7}{18}, \frac{11}{18}]$ since $\Delta_0 \leq \frac{1}{9}\frac{1}{n-1}$.
Case III: Otherwise, suppose that both $i, j \in \{a, a+1, b, b+1\}$. Then we have

$$D_{\text{KL}}(\mathbb{P}^a(X_{ij}^{(1)}) \| \mathbb{P}^b(X_{ij}^{(1)})) \leq 20 p_{\text{obs}} \Delta_0^2.$$

Combining the bounds from the three cases, we find that the KL divergence is upper bounded as

$$\frac{1}{r} D_{\text{KL}}(\mathbb{P}^a \| \mathbb{P}^b) \leq 40(n-4) p_{\text{obs}} \Delta_0^2 + 240 p_{\text{obs}} \Delta_0^2 \leq 50 n p_{\text{obs}} \Delta_0^2,$$

where we have used the assumption $n \geq 9$ to obtain the final inequality.

3.6.3 Proof of Theorem 8: Hamming error

We now turn to the proof of Theorem 8, beginning with part (a).

Proof of part (a)

Without loss of generality, we can assume that the true underlying ranking is the identity ranking, that is, item i is ranked at position i for every $i \in [n]$. Given the the lower bound $\alpha \geq 8$ is satisfied, Theorem 7 ensures that with probability at least $1 - n^{-16}$, the counting estimator $\widetilde{\mathcal{S}}_k$ ranks every item in $\{1, \ldots, k-h\}$ higher than every item in the set $\{k+h+1, \ldots, n\}$. Thus, we are guaranteed that either $\widetilde{\mathcal{S}}_k \subseteq [k+h]$ and/or $[k-h] \subseteq \widetilde{\mathcal{S}}_k$. One can verify either case leads to $|\widetilde{\mathcal{S}}_k \cap [k]| \geq k-h$, thereby proving the claimed result.

Proof of part (b)

We assume without loss of generality that $k \leq \frac{n}{2}$. (Otherwise, one can equivalently study the problem of recovering the last k items.) Since the case $h = 0$ is already covered by Theorem 7(b), we may also assume that $h \geq 1$.

The proof involves construction of $\eta \geq 1$ sets of probability matrices $\{M^a\}_{a \in [\eta]}$ of the pairwise comparisons with the following two properties:

(i) For every $a \in [\eta]$, let $S_k^a \subseteq [n]$ denote the set of the top k items under the a^{th} set of distributions. Then for every k-sized set $S \in [n]$,

$$\sum_{a=1}^{\eta} \mathbf{1}\{D_{\text{H}}(S, S_k^a) \leq 2h\} \leq 1.$$

(ii) If the underlying distribution a is chosen uniformly at random from this set of η distributions, then any estimator that attempts to identify the underlying distribution $a \in [\eta]$ errs with probability at least $\frac{1}{7}$.

Now consider any estimator $\widehat{\mathcal{S}}_k$ for identifying the top k items $\mathcal{S}_k^*$. Given property (i), whenever the estimator is successful under the Hamming error requirement $D_{\mathrm{H}}(\widehat{\mathcal{S}}_k, \mathcal{S}_k^*) \leq 2h$, it must be able to uniquely identify the index $a \in [\eta]$ of the underlying distribution of pairwise comparison probabilities. However, property (ii) mandates that any estimator for identifying the underlying distribution errs with a probability at least $\frac{1}{7}$. Assuming that such sets of probability distributions satisfying these two properties exist, putting these results together yields the claimed result.

We now proceed to construct probability distributions satisfying the two aforementioned properties. Consider any positive number Δ_0 satisfying the upper bound

$$\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1 \nu_2 \log n}{n p_{\mathrm{obs}} r}}. \tag{3.22}$$

The η matrices $\{M^a\}_{a \in [\eta]}$ of probability distributions we construct differ only in a permutation of their rows and columns, and modulo this permutation, have identical values. In other words, these η distributions differ only in the identities of the n items and the values of the pairwise-comparison probabilities $M^a_{(i)(j)}$ among the ordered sequence of the n items are identical across all distributions $a \in [\eta]$.

For any ordering $(1), \ldots, (n)$ of the n items, for every $a \in [\eta]$, set

$$M^a_{(i)(j)} = \begin{cases} \frac{1}{2} + \Delta_0 & \text{if } i \in [k] \text{ and } j \notin [k] \\ \frac{1}{2} - \Delta_0 & \text{if } i \notin [k] \text{ and } j \in [k] \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{3.23}$$

Note that the upper bound (3.22) on Δ_0, coupled with the assumption $p_{\mathrm{obs}} \geq \sqrt{\frac{\log n}{2nr}}$, ensures that $\Delta_0 < \frac{1}{3}$ and hence that our definition (3.23) leads to a valid set of probabilities. Given this construction, the scores of the n items are $\tau_{(1)} = \cdots = \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n)} + \Delta_0$. The bound (3.22) ensures that the condition $\alpha \leq \frac{\sqrt{\nu_1 \nu_2}}{14}$ required by the hypothesis of the theorem is satisfied.

It remains to specify the ordering of the n items in each set of probability distributions. This specification relies on the following lemma, that in turn uses a coding-theoretic result due to Levenshtein [149]. It applies in the regime $2h \leq \frac{1}{1+\nu_2} \min\{n^{1-\nu_1}, k, n-k\}$ for some constants $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, and when n is larger than a (ν_1, ν_2)-dependent constant.

Lemma 17. *Under the previously given conditions, there exists a subset* $\{b^1, \ldots, b^L\} \subseteq \{0,1\}^{n/2}$ *with cardinality* $L \geq e^{\frac{9}{10}\nu_1 \nu_2 h \log n}$, *and such that*

$$D_{\mathrm{H}}(b^j, \mathbf{0}) = 2(1+\nu_2)h, \quad \textit{and} \quad D_{\mathrm{H}}(b^j, b^k) > 4h \quad \textit{for all } j \neq k \in [L].$$

We prove this lemma at the end of this section. Given this lemma, we now complete the proof of the theorem. Map the $\frac{n}{2}$ items $\{\frac{n}{2}+1, \ldots, n\}$ to the $\frac{n}{2}$ bits in each of the strings given by Lemma 17. For each $\ell \in [e^{\frac{9}{10}\nu_1\nu_2 h \log n}]$, let B_ℓ denote the $2(1+\nu_2)h$-sized subset of $\{\frac{n}{2}+1,\ldots,n\}$ corresponding to the $2(1+\nu_2)h$ positions equalling 1 in the ℓ^{th} string. Also define sets $A_\ell = \{1,\ldots,k-2(1+\nu_2)h\}$ and $C_\ell = [n]\backslash(A_\ell \cup B_\ell)$. We note that this construction is valid since $2h \leq \frac{1}{1+\nu_2}k$.

We now construct $\eta = e^{\frac{9}{10}\nu_1\nu_2 h \log n}$ sets of pairwise comparison probability distributions $M^1,\ldots,M^\eta$ and show that these sets satisfy the two required properties. As mentioned earlier, each matrix of comparison-probabilities M^ℓ takes values as given in (3.23), but differs in the underlying ordering of the n items. In particular, associate the set $\ell \in [\eta]$ of distributions to any ordering of the n items that ranks every item in A_ℓ higher than every item in B_ℓ, and every item in B_ℓ in turn higher than every item in C_ℓ. Then for any ℓ, the set of top k items is given by $A_\ell \cup B_\ell$. From the guarantees provided by Lemma 17, for any distinct $\ell, m \in [\eta]$, we have $D_{\mathrm{H}}(A_\ell \cup B_\ell, A_m \cup B_m) \geq 4h+1$. This construction consequently satisfies the first required property.

We now show that the construction also satisfies the second property: namely, it is difficult to identify the true index. We do so using Fano's inequality (2.22), for which we denote the probability distribution of the observations due to any matrix M^ℓ, $\ell \in [\eta]$, as $\mathbb{P}^\ell$.

We first derive an upper bound on the Kullback-Leibler divergence between any two distributions $\mathbb{P}^\ell$ and $\mathbb{P}^m$ of the observations. Observe that $\mathbb{P}^\ell(i \succ j) \neq \mathbb{P}^m(i \succ j)$ only if $i \in B_\ell \cup B_m$ or $j \in B_\ell \cup B_m$. In this case, we have $D_{\mathrm{KL}}(\mathbb{P}^\ell(i \succ j) \| \mathbb{P}^m(i \succ j)) \leq \frac{4\Delta_0^2}{\frac{1}{4}-\Delta_0^2}$. Since both sets B_ℓ and B_m have a cardinality of $2(1+\nu_2)h$, aggregating over all possible observations across all pairs, we obtain that

$$D_{\mathrm{KL}}(\mathbb{P}^\ell \| \mathbb{P}^m) \leq 4(1+\nu_2)hnp_{\mathrm{obs}}r\frac{4\Delta_0^2}{\frac{1}{4}-\Delta_0^2}. \tag{3.24}$$

In the regime $p_{\mathrm{obs}} \geq \frac{\log n}{2nr}$ and $\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1\nu_2 \log n}{np_{\mathrm{obs}}r}}$, we have $\Delta_0 \leq \frac{1}{14\sqrt{2}}$. Substituting the inequality $\Delta_0 \leq \frac{1}{14}\sqrt{\frac{\nu_1 \log n}{np_{\mathrm{obs}}r}}$ in the numerator and $\frac{1}{4}-\Delta_0^2 \geq \frac{1}{4} - \left(\frac{1}{14\sqrt{2}}\right)^2$ in the denominator of the right hand side of the bound (3.24), we find that

$$D_{\mathrm{KL}}(\mathbb{P}^\ell \| \mathbb{P}^m) \leq \frac{3}{4}\nu_1\nu_2 h \log n.$$

Now suppose that we drawn Y from some distribution chosen uniformly at random from $\{\mathbb{P}^1,\ldots,\mathbb{P}^\eta\}$. Applying Fano's inequality (2.22) ensures that any test ϕ for estimating the index A of the chosen distribution must have error probability lower bounded as

$$\mathbb{P}\big[\phi(Y) \neq A\big] \geq \left(1 - \frac{\frac{3}{4}\nu_1\nu_2 h \log n + \log 2}{\frac{9}{10}\,\nu_1\nu_2 h \log n}\right) \geq \frac{1}{7}.$$

Here the final inequality holds as long as n is larger than some universal constant.

Proof of Lemma 17

We divide the proof into two cases depending on the value of h.

Case I: $h \geq \frac{1}{2\nu_1\nu_2}$: Let L denote the number of binary strings of length m_0 such that each has a Hamming weight w_0 and each pair has a Hamming distance at least d_0. It is known [111, 149] that L can be lower bounded as:

$$\eta \geq \frac{\binom{m_0}{w_0}}{\sum_{i=0}^{\lfloor \frac{d_0-1}{2} \rfloor} \binom{w_0}{j}\binom{m_0-w_0}{j}} \geq \frac{\left(\frac{m_0}{w_0}\right)^{w_0}}{\frac{d_0+1}{2}\left(\frac{ew_0}{\min\{d_0,w_0\}/2}\right)^{\min\{d_0,w_0\}/2}\left(\frac{em_0}{\min\{d_0,m_0\}/2}\right)^{\min\{d_0,m_0\}/2}}.$$

Note that for the setting at hand, we have $m_0 = \frac{n}{2}$, $w_0 = 2(1+\nu_2)h$ and $d_0 = 4h+1$. Since $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, we have the chain of inequalities

$$w_0 < d_0 \leq 4n^{1-\nu_1} \overset{(i)}{<} \frac{n}{2} = m_0,$$

where the inequality (i) holds when n is large enough. These relations allow for the simplification:

$$\begin{aligned}\log\eta &\geq \log\left\{\frac{\left(\frac{m_0}{w_0}\right)^{w_0}}{\frac{d_0+1}{2}\left(\frac{ew_0}{w_0/2}\right)^{w_0/2}\left(\frac{em_0}{d_0/2}\right)^{d_0/2}}\right\} \\ &= (w_0 - d_0/2)\log m_0 - w_0\log w_0 + \frac{d_0}{2}\log d_0 - \frac{d_0+w_0}{2}\log(2e) - \log((d_0+1)/2).\end{aligned}$$

Substituting the values of w_0, d_0 and m_0 and then simplifying yields

$$\begin{aligned}\log\eta &\geq (2\nu_2 h - \frac{1}{2})\log\frac{n}{2} - 2(1+\nu_2)h\log(2(1+\nu_2)h) + (2h+\frac{1}{2})\log(4h+1) \\ &\quad - (((3+\nu_2)h) + \frac{1}{2})\log(2e) - \log(2h+1) \\ &\geq (2\nu_2 h - \frac{1}{2})\log\frac{n}{2} - 2\nu_2 h\log(2(1+\nu_2)h) - c_1'h,\end{aligned}$$

where c_1' is a constant whose value depends only on (ν_1, ν_2). In the regime $\frac{1}{\nu_1\nu_2} \leq 2h \leq \frac{n^{1-\nu_1}}{1+\nu_2}$, some algebraic manipulations then yield

$$\log\eta \geq (2\nu_1\nu_2 h - \frac{1}{2})\log\frac{n}{2} - c_1'h \geq \nu_1\nu_2 h(\log n - \log 2 - c_1') \geq \frac{9}{10}\nu_1\nu_2 h\log n,$$

where the final inequality holds when n is large enough.

Case II: $h < \frac{1}{2\nu_1\nu_2}$ Consider a partition of the $\frac{n}{2}$ bits into $\frac{n}{4(1+\nu_2)h}$ sets of size $2(1+\nu_2)h$ each. Define an associated set of $\frac{n}{4(1+\nu_2)h}$ sets of binary strings, each of length $\frac{n}{2}$, with the

i^{th} string having ones in the positions corresponding to the i^{th} set in the partition and zeros elsewhere. Then each of these strings have a Hamming weight of $2(1+\nu_2)h$, and every pair has a Hamming distance at least $4(1+\nu_2)h > 4h$. The total number of such strings equals

$$\exp\Big(\log\frac{n}{4(1+\nu_2)h}\Big) \overset{(i)}{\geq} \exp\Big(\log n - \log(\frac{2(1+\nu_2)}{\nu_1\nu_2})\Big) \overset{(ii)}{\geq} \exp\Big(\frac{9}{10}\log n\Big) \overset{(iii)}{>} \exp\Big(1.8\nu_1\nu_2 h\log n\Big),$$

where the inequalities (i) and (iii) are a result of operating in the regime $h < \frac{1}{2\nu_1\nu_2}$ and the inequality (ii) assumes that n is greater than a (ν_1,ν_2)-dependent constant.

3.6.4 Proof of Theorem 9: General error

We now turn to the proof of Theorem 9.

Proof of part (a)

For every $i \in [n]$, let (i) denote the item ranked i according to their latent scores, as defined in equation (3.1). Recall from the proof of Theorem 7 that for any $u < v \in [n]$, the condition

$$\tau_{(u)} - \tau_{(v)} \geq 8\sqrt{\frac{\log n}{np_{\text{obs}}r}}$$

ensures that with probability at least $1-n^{-14}$, every item in the set $\{(1),\ldots,(u)\}$ wins more comparisons than every item in the set $\{(v),\ldots,(n)\}$. Consequently, if the set $\widetilde{\mathcal{S}}_k$ contains any item in $\{(v),\ldots,(n)\}$, then it must contain the entire set $\{(1),\ldots,(u)\}$. In other words, at least one of the following must be true: either $\{(1),\ldots,(u)\} \subseteq \widetilde{\mathcal{S}}_k$ or $\widetilde{\mathcal{S}}_k \subseteq \{(1),\ldots,(v-1)\}$. Consequently, in the regime $v = k+t-u+1$ for any $1 \leq u \leq k$ and $u \leq t \leq n$, we have that

$$|\widetilde{\mathcal{S}}_k \cap \{(1),\ldots,(t)\}| \geq u. \tag{3.25}$$

Now consider any $b \in [\beta]$ that satisfies the condition

$$\min_{j\in[k]}(\tau_{(j)} - \tau_{(k+t_j^b-j+1)}) \geq 8\sqrt{\frac{\log n}{np_{\text{obs}}r}}.$$

For any $j \in [k]$, setting $u = j$ and $v = (k+t_j^b-j+1)$ in (3.25), and applying the union bound over all values of $j \in [k]$ yields that

$$|\widetilde{\mathcal{S}}_k \cap \{(1),\ldots,(t_j^b)\}| \geq j \quad \text{for every } j \in [k],$$

with probability at least $1-n^{-13}$. Consequently, we have that

$$\mathbb{P}\big(\widetilde{\mathcal{S}}_k \in \Lambda(T_b)\big) \geq 1-n^{-13},$$

completing the proof of the claim.

Proof of part (b)

In the regime $t^b_{\mu_2 k} \leq \frac{n}{2}$ for every $b \in [\beta]$, it suffices to show that any estimator $\widehat{\mathcal{S}}_k$ will incur an error lower bounded as

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (n/2)\}| < \mu_2 k\big) \geq \frac{1}{15},$$

where (i) denotes the item ranked i according to their latent scores according to equation (3.1).

Our proof relies on the result and proof of the Hamming error case analyzed in Theorem 8(b). To this end, let us set the parameter h of Theorem 8(b) as $h = 2(1-\mu_2)k$. We claim that this value of h lies in the regime $h \leq \frac{1}{2(1+\nu_2)} \min\{k, n-k, n^{1-\nu_1}\}$ for some values $\nu_1 \in (0,1)$ and $\nu_2 \in (0,1)$, as required by Theorem 8(b). This claim follows from the fact that

$$h = 2(1-\mu_2)k \leq \frac{1}{2(1+\nu_2)} k,$$

for $\nu_2 = \min\{\frac{1}{4(1-\mu_2)} - 1, \frac{1}{2}\} \in (0,1)$. Furthermore,

$$h = 2(1-\mu_2)k \overset{(i)}{\leq} \frac{n^{1-\mu_1}}{4} \overset{(ii)}{\leq} \frac{1}{2(1+\nu_2)} n^{1-\nu_1}$$

for $\nu_1 = \frac{9}{10}\mu_1 \in (0,1)$, where (i) is a result of our assumption $8(1-\mu_2)k \leq n^{1-\mu_1}$ and (ii) holds when n is large enough. This assumption also implies that $k \leq n-k$ for a large enough value of n. We have now verified operation in the regime required by Theorem 8(b).

The construction in the proof of Theorem 8 is based on setting

$$\tau_{(1)} = \cdots \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n)} + \Delta_0,$$

for any real number Δ_0 in the interval $\Big(0,\ \frac{1}{14}\sqrt{\frac{\nu_1 \nu_2 \log n}{n p_{\mathrm{obs}} r}}\ \Big]$. This condition is also satisfied in our construction due to the assumed upper bound $\alpha \leq \frac{1}{15}\sqrt{\mu_1 \min\big\{\frac{1}{4(1-\mu_2)-1}, \frac{1}{2}\big\}}$. Consequently, the result of Theorem 8(b) implies that in this setting, any estimator $\widehat{\mathcal{S}}_k$ will incur a Hamming error greater than $h = 2(1-\nu_2)k$ with probability at least $\frac{1}{7}$, or equivalently,

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (k)\}| < (2\mu_2 - 1)k\big) \geq \frac{1}{7}.$$

Under this event, the estimator $\widehat{\mathcal{S}}_k$ contains at most $(2\mu_2 - 1)k - 1$ items from the set of top k items. In order to ensure it gets at least $\mu_2 k$ items from $\{(1), \ldots, (n/2)\}$, the remaining $2(1-\mu_2)k + 1$ chosen items must have at least $(1-\mu_2)k + 1$ items from $\{(k+1), \ldots, (n/2)\}$. However, in the construction, items $(k+1), \ldots, (n)$ are indistinguishable from each other, and hence by symmetry these $2(1-\mu_2)k + 1$ chosen items must contain at least $(1-\mu_2)k + 1$

items from the set $\{(n/2+1), \ldots, (n)\}$ with probability at least $\frac{1}{2}$. Putting these arguments together, we obtain that under this construction, any estimator $\widehat{\mathcal{S}}_k$ has error probability lower bounded as

$$\mathbb{P}\big(|\widehat{\mathcal{S}}_k \cap \{(1), \ldots, (n/2)\}| < \mu_2 k\big) \geq \frac{1}{14}. \tag{3.26}$$

It remains to deal with a subtle technicality. The construction above involves items $(k+1), \ldots, (n)$ with identical scores. Recall that in the definition of the user-defined requirement, in case of multiple items with identical scores, we considered the choice of either of such items as valid. The following lemma helps overcome this issue. In order to state the lemma, we define $\|\!|M\|\!|_\infty := \max_{(i,j)\in[n]^2} |M_{ij}|$ for a matrix $M \in \mathbb{R}^{n\times n}$.

Lemma 18. *Consider any two $(n \times n)$ matrices M^a and M^b of pairwise probabilities such that*

$$\|\!|M^a - M^b\|\!|_\infty \leq \epsilon, \quad \|\!|M^a\|\!|_\infty \geq \epsilon, \text{ and } \|\!|M^b\|\!|_\infty \geq \epsilon \tag{3.27a}$$

for some $\epsilon \in [0,1]$. Then for any k-sized sets of items $T_1, \ldots, T_\beta \subseteq [n]$, *and any estimator $\widehat{\mathcal{S}}_k$, we have*

$$|\ \mathbb{P}_{M^a}(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\}) - \mathbb{P}_{M^b}(\widehat{\mathcal{S}}_k \in \{T_1, \ldots, T_\beta\})\ | \leq 6^{n^2 r}\epsilon. \tag{3.27b}$$

See Section 3.6.4 for the proof of this claim.

Now consider an $(n \times n)$ pairwise probability matrix M' whose entries takes values

$$M'_{(i)(j)} = \begin{cases} \frac{1}{2} + \Delta_0 + \epsilon & \text{if } i \in [k] \text{ and } j \in [n]\backslash[n/2] \\ \frac{1}{2} + \Delta_0 & \text{if } i \in [k] \text{ and } j \in [n/2]\backslash[k] \\ \frac{1}{2} + \epsilon & \text{if } i \in [n/2]\backslash[k] \text{ and } j \in [n]\backslash[n/2] \\ \frac{1}{2} & \text{otherwise,} \end{cases}$$

and $M'_{ji} = 1 - M'_{ij}$, whenever $i \leq j$. Set $\epsilon = 7^{-n^2 r}$.

One can verify that under the probability matrix M', the scores of the n items satisfy the relations

$$\tau_{(1)} = \cdots = \tau_{(k)} = \tau_{(k+1)} + \Delta_0 = \cdots = \tau_{(n/2)} + \Delta_0 = \tau_{(n/2+1)} + \Delta_0 + \epsilon = \cdots = \tau_{(n)} + \Delta_0 + \epsilon.$$

The set of items $\{(1), \ldots, (n/2)\}$ are thus explicitly distinguished from the items $\{(n/2+1), \ldots, (n)\}$. We now call upon Lemma 18 with $M^a = M'$, and M^b as the matrix of probabilities constructed in the proof of Theorem 8, where both sets have the same ordering of the items. This assignment is valid given that $\Delta_0 < \frac{1}{3}$ and $\epsilon = 7^{-n^2 r}$. Lemma 18 then implies that any estimator that is $\mathfrak{S}$-respecting with probability at least $1 - \frac{1}{15}$ under M^b must also be $\mathfrak{S}$-respectiin with probability at least $1 - \frac{1}{14.5}$ under M^a. But by equation (3.26), the latter condition is impossible, which implies our claimed lower bound.

Proof of Lemma 18

Let $\mathbb{P}^a$ and $\mathbb{P}^b$ denote the probabilities induced by the matrices M^a and M^b respectively. Consider any fixed observation $Y_1 \subseteq \{0,1,\phi\}^{r(n\times n)}$, where ϕ denotes the absence of an observation. Given the bounds (3.27a), some algebra leads to

$$\mid \mathbb{P}^a(Y=Y_1) - \mathbb{P}^b(Y=Y_1) \mid \leq 2^{n^2 r}\epsilon, \tag{3.28}$$

where $\mathbb{P}^a(Y=Y_1)$ and $\mathbb{P}^b(Y=Y_1)$ denote the probabilities of observing Y_1 under $\mathbb{P}^a$ and $\mathbb{P}^b$, respectively.

Now consider any estimator $\widehat{\mathcal{S}}_k$, which is permitted to be randomized. Let $\eta \leq 3^{n^2 r}$ denote the total number of possible values of the observation Y, and let $\{Y_1,\ldots,Y_\eta\} = \{0,1,\phi\}^{r(n\times n)}$ denote the set of all possible valid values of the observation. For each $i \in [\eta]$, let $q_i \in [0,1]$ denote the probability that the estimator $\widehat{\mathcal{S}}_k$ succeeds in satisfying the given requirement when the data observed equals Y_i. (Recall that the given requirement is in terms of the actual items and not their positions.) Then we have

$$\begin{aligned}
\big|\mathbb{P}^1(\widehat{\mathcal{S}}_k \in \{T_1,\ldots,T_\beta\}) - \mathbb{P}^2(\widehat{\mathcal{S}}_k \in \{T_1,\ldots,T_\beta\})\big| &= \big|\sum_{i=1}^{\eta}\mathbb{P}^1(Y=Y_i)q_i - \sum_{i=1}^{\eta}\mathbb{P}^2(Y=Y_i)q_i\big| \\
&\leq \sum_{i=1}^{\eta} \mid \mathbb{P}^1(Y=Y_i) - \mathbb{P}^2(Y=Y_i) \mid q_i \\
&\overset{(i)}{\leq} \sum_{i=1}^{\eta} 2^{n^2 r}\epsilon q_i \overset{(ii)}{\leq} 6^{n^2 r}\epsilon,
\end{aligned}$$

as claimed, where step (i) follows from our earlier bound (3.28) and step (ii) uses the bound $\eta \leq 3^{n^2 r}$.

Chapter 4

Labeling and Classification

"Often are the amalgamated judgements of many better than the assessment of one."

– Marie Curie

4.1 Introduction

Recent years have witnessed a surge of interest in the use of crowdsourcing for labeling massive datasets. Expert labels are often difficult or expensive to obtain at scale, and crowdsourcing platforms allow for the collection of labels from a large number of low-cost workers. This paradigm, while enabling several new applications of machine learning, also introduces some key challenges: first, low-cost workers are often non-experts and the labels they produce can be quite noisy, and second, data collected in this fashion has a high amount of heterogeneity with significant differences in the quality of labels across workers and tasks. Thus, it is important to develop realistic models and scalable algorithms for aggregating and drawing meaningful inferences from the noisy labels obtained via crowdsourcing.

This chapter focuses on objective labeling tasks involving binary choices, meaning that each question or task is associated with a single correct binary answer or label.[1] There is a vast literature on the problem of estimation from noisy crowdsourced labels [56, 84, 85, 88, 116, 117, 151, 274]. This past work is based primarily on the classical Dawid-Skene model [60]. The Dawid-Skene model is a paramter-based model in which each worker i is associated with a single scalar parameter $q_i^{\mathrm{DS}} \in [0,1]$, and it is assumed that the probability that worker i answers any question j correctly is given by the same scalar q_i^{DS}. Thus, the

[1] In this chapter, we use the terms {question, task}, and {answer, label} in an interchangeable manner.

Dawid-Skene model imposes a homogeneity condition on the questions, one which is often not satisfied in practical applications where some questions may be more difficult than others.[2]

Accordingly, in this chapter, we propose and analyze a more general permutation-based model that allows the noise in the answer to depend on the particular question-worker pair. Within the context of such models, we propose and analyze a variety of estimation algorithms. One possible metric for analysis is the Hamming error, and there is a large body of past work [56, 84, 85, 88, 116, 117, 274] that provides sufficient conditions that guarantee zero Hamming error—meaning that every question is answered correctly—with high probability. Although the Hamming error can be suitable for the analysis of Dawid-Skene style models, we argue in the sequel that it is less appropriate for the heterogenous settings studied in this chapter. Instead, when tasks have heterogenous difficulties, it is more natural to use a weighted metric that also accounts for the underlying difficulty of the tasks. Concretely, an estimator should be penalized less for making an error on a question that is intrinsically more difficult. In this chapter, we introduce and provide analysis under such a difficulty-weighted error metric.

From a high-level perspective, the contributions of this chapter can be summarized as follows:

- We introduce a new "permutation-based" model for crowd-labeled data, and a new difficulty-weighted metric that extends the popular Hamming metric.
- We provide upper and lower bounds on the minimax error, sharp up to logarithmic factors, for estimation under the permutation-based model. Our bounds lead to the useful implication that the generality afforded by the proposed permutation-based model as compared to the popular parameter-based Dawid-Skene model enables more robust estimation, and surprisingly, there is only a small statistical price to be paid for this flexibility.
- We provide a computationally-efficient estimator that achieves the minimax limits over the permutation-based model when an approximate ordering of the workers in terms of their abilities is known.
- We provide a computationally-efficient estimator, termed the OBI-WAN estimator, that is consistent over the permutation-based model class. Moreover, it is optimal over an intermediate setting between the parameter-based Dawid-Skene and the permutation-based models, which allows for task heterogeneity but in a restricted, parameter-based manner. As a special case, our sharp upper bounds on the estimation error of OBI-WAN also apply uniformly over the parameter-based Dawid-Skene model, while prior known guarantees fall short of establishing such uniform bounds.

The remainder of this chapter is organized as follows. In Section 4.2, we provide some background, setup the problems we address in this chapter, and provide an overview of related literature. Section 4.3 is devoted to our main results. We present numerical simulations

[2] To be clear, the original model by Dawid and Skene [60] also allows for asymmetric errors across different classes. In this chapter, we focus on the setting with symmetric error probabilities, that has popularly come to be known as the "one-coin Dawid-Skene model", and is considered in many past theoretical works [56, 88, 116, 117].

in Section 4.4. We conclude the chapter with a discussion of future research directions in Section 4.5. We present proofs of our main results in Section 4.6.

4.2 Problem setting

We begin with some background on existing crowd labeling models, followed by an introduction to our proposed models; we conclude with a discussion of related work.

4.2.1 Observation model

Consider a crowdsourcing system that consists of n workers and d questions. We assume every question has two possible answers, denoted by $\{-1, +1\}$, of which exactly one is correct. We let $x^* \in \{-1, 1\}^d$ denote the collection of correct answers to all d questions. We model the question-answering via an unknown matrix $Q^* \in [0, 1]^{n \times d}$ whose $(i, j)^{th}$ entry, Q^*_{ij}, represents the probability that worker i answers question j correctly. Otherwise, with probability $1 - Q^*_{ij}$, worker i gives the incorrect answer to question j. For future reference, note that the parameter-based Dawid-Skene model involves a special case of such a matrix, namely one of the form $Q^* = q^{\text{DS}} 1^T$, where the vector $q^{\text{DS}} \in [0, 1]^n$ corresponds to the vector of correctness probabilities, with a single scalar associated with each worker.

We denote the response of worker i to question j by a variable $Y_{ij} \in \{-1, 0, 1\}$, where we set $Y_{ij} = 0$ if worker i is not asked question j, and set Y_{ij} to the answer (-1 or 1) provided by the worker otherwise. We also assume that worker i is asked question j with probability $p_{\text{obs}} \in [0, 1]$, independently for every pair $(i, j) \in [n] \times [d]$, and that a worker is never asked the same question twice. We also make the standard assumption that given the values of x^* and Q^*, the entries of Y are all mutually independent. In summary, we observe a matrix Y which has independent entries distributed as

$$Y_{ij} = \begin{cases} x^*_j & \text{with probability } p_{\text{obs}}\, Q^*_{ij} \\ -x^*_j & \text{with probability } p_{\text{obs}}\, (1 - Q^*_{ij}) \\ 0 & \text{with probability } (1 - p_{\text{obs}}). \end{cases}$$

Given this random matrix Y, our goal is to estimate the binary vector $x^* \in \{-1, 1\}^d$ of true labels.

Obtaining non-trivial guarantees for this problem requires that some structure be imposed on the probability matrix Q^*. The parameter-based Dawid-Skene model is one form of such structure: it requires that the probability matrix Q^* be rank one, with identical columns all equal to $q^{\text{DS}} \in \mathbb{R}^n$. As noted previously, this structural assumption on Q^* is very strong. It assumes that each worker has a fixed probability of answering a question correctly, and is likely to be violated in settings where some questions are more difficult than others.

Accordingly, in this chapter, we study a more general permutation-based model of the following form. We assume that there are two underlying orderings, both of which are unknown to us: first, a permutation $\pi^* : [n] \to [n]$ that orders the n workers in terms of their (latent) abilities, and second, a permutation $\sigma^* : [d] \to [d]$ that orders the d questions with respect to their (latent) difficulties. In terms of these permutations, we assume that the probability matrix Q^* obeys the following conditions:

- **Worker monotonicity:** For every pair of workers i and i' such that $\pi^*(i) < \pi^*(i')$ and every question j, we have $Q^*_{ij} \geq Q^*_{i'j}$.
- **Question monotonicity:** For every pair of questions j and j' such that $\sigma^*(j) < \sigma^*(j')$ and every worker i, we have $Q^*_{ij} \geq Q^*_{ij'}$.

In other words, the permutation-based model assumes the existence of a permutation of the rows and columns such that each row and each column of the permuted matrix Q^* has non-increasing entries. The rank of the resulting matrix is allowed to be as large as $\min\{n, d\}$. It is straightforward to verify that the parameter-based Dawid-Skene model corresponds to a particular type of such probability matrices, restricted to have identical columns.

It should be noted that none of these models are identifiable without further constraints. For instance, changing x^* to $-x^*$ and Q^* to $(11^T - Q^*)$ does not change the distribution of the observation matrix Y. In the context of the parameter-based Dawid-Skene model, several papers [85, 116, 117, 274] have resolved this issue by requiring that $\frac{1}{n}\sum_{i=1}^{n} q_i^{\mathrm{DS}} \geq \frac{1}{2} + \mu$ for some constant value $\mu > 0$. Although this condition resolves the lack of identifiability, the underlying assumption—namely that every question is answerable by a subset of the workers—can be violated in practice. In particular, one frequently encounters questions that are too difficult to answer by any of the hired workers, and for which the worker's answers are near uniformly random (e.g., see the papers [69, 235]). On the other hand, empirical observations also show that workers in crowdsourcing platforms, as opposed to being adversarial in nature, at worst provide random answers to labeling tasks [69, 82, 83, 270]. On this basis, it is reasonable to assume that for every worker i and question j we have that $Q^*_{ij} \geq \frac{1}{2}$. We make this assumption throughout this chapter.

In summary, we let $\mathbb{C}_{\mathrm{Perm}}$ denote the set of all possible values of matrix Q^* under the proposed permutation-based model, that is,

$$\mathbb{C}_{\mathrm{Perm}} := \big\{Q \in [0.5, 1]^{n \times d} \; | \text{there exist permutations } (\pi, \sigma) \text{ such that} \\ \text{question and worker monotonocity hold}\big\}.$$

For future reference, we use

$$\mathbb{C}_{\mathrm{DS}} := \big\{Q \in \mathbb{C}_{\mathrm{Perm}} \mid Q = q^{\mathrm{DS}} 1^T \text{ for some } q^{\mathrm{DS}} \in [0.5, 1]^n\big\},$$

to denote the subset of such matrices that are realizable under the parameter-based Dawid-Skene assumption.

4.2.2 Evaluating estimators

In this section, we introduce the criteria used to evaluate estimators in this chapter. In formal terms, an estimator $\widehat{x}$ is a measurable function that maps any observation matrix

Y to a vector in the Boolean hypercube $\{-1,1\}^d$. The most popular way of assessing the performance of such an estimator is in terms of its (normalized) *Hamming error*

$$D_{\mathrm{H}}(\widehat{x}, x^*) := \frac{1}{d}\sum_{j=1}^{d} \mathbf{1}\{\widehat{x}_j \neq x^*_j\}, \tag{4.1}$$

where $\mathbf{1}\{\widehat{x}_j \neq x^*_j\}$ denotes a binary indicator which takes the value 1 if $\widehat{x}_j \neq x^*_j$, and 0 otherwise. A potential deficiency of the Hamming error is that it places a uniform weight on each question. As mentioned earlier, there are applications of crowdsourcing in which some subset of the questions are very difficult, and no hired worker can answer reliably. In such settings, any estimator will have an inflated Hamming error, not due to any particular deficiencies of the estimator, but rather due to the intrinsic hardness of the assigned collection of questions. This error inflation will obscure possible differences between estimators.

With this issue in mind, we propose an alternative error measure that weights the Hamming error with the difficulty of each task. A more general class of error measures takes the form

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) = \frac{1}{d}\sum_{j=1}^{d} \mathbf{1}\{\widehat{x}_j \neq x^*_j\}\Psi(Q^*_{1j}, \ldots, Q^*_{nj}), \tag{4.2}$$

for some function $\Psi : [0,1]^n \to \mathbb{R}_+$ which captures the difficulty of estimating the answer to a question.

The Q^*-loss: In order to choose a suitable function Ψ, we note that past work on the parameter-based Dawid-Skene model [56, 85, 88, 116, 117] has shown that the quantity

$$\frac{1}{n}\sum_{i=1}^{n}(2q_i^{\mathrm{DS}} - 1)^2, \tag{4.3}$$

popularly known as the *collective intelligence* of the crowd, is central to characterizing the overall difficulty of the crowd-sourcing problem under the parameter-based Dawid-Skene assumption. A natural generalization, then, is to consider the weights

$$\Psi(Q^*_{1j}, \ldots, Q^*_{nj}) = \frac{1}{n}\sum_{i=1}^{n}\left(2Q^*_{ij} - 1\right)^2 \qquad \text{for each task } j \in [d], \tag{4.4a}$$

which characterizes the difficulty of task j for a given collection of workers. This choice gives rise to the *Q^*-loss function*

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) := \frac{1}{d}\sum_{j=1}^{d}\Big(\mathbf{1}\{\widehat{x}_j \neq x^*_j\}\ \frac{1}{n}\sum_{i=1}^{n}(2Q^*_{ij} - 1)^2\Big) \tag{4.4b}$$

$$= \frac{1}{dn}\|\!|(Q^* - \frac{1}{2}11^T)\ \mathrm{diag}(\widehat{x} - x^*)|\!\|_{\mathrm{F}}^2, \tag{4.4c}$$

where $\mathrm{diag}(\widehat{x}-x^*)$ denotes the matrix in $\mathbb{R}^{d\times d}$ whose diagonal entries are given by the vector $\widehat{x}-x^*$. Note that under the parameter-based Dawid-Skene model (in which $Q^* = q^{\mathrm{DS}}1^T$), this loss function reduces to

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) = \Big(\frac{1}{n}\sum_{i=1}^{n}(2q_i^{\mathrm{DS}}-1)^2\Big)\ \underbrace{\Big(\frac{1}{d}\sum_{j=1}^{d}\mathbf{1}\{\widehat{x}_j \neq x_j^*\}\Big)}_{D_{\mathrm{H}}(\widehat{x},x^*)},$$

corresponding to the normalized Hamming error rescaled by the collective intelligence.

For future reference, let us summarize some properties of the function $\mathcal{L}_{Q^*}$: (a) it is symmetric in its arguments $(x^*, \widehat{x})$, and satisfies the triangle inequality; (b) it takes values in the interval $[0,1]$; and (c) if for every question $j \in [d]$, there exists a worker $\ell \in [n]$ such that $Q^*_{\ell j} > \frac{1}{2}$, then $\mathcal{L}_{Q^*}$ defines a metric; if not, it defines a pseudo-metric.

Minimax risk: Given the loss function $\mathcal{L}_{Q^*}$, we evaluate the performance of estimators in terms of their uniform risk properties over a particular class $\mathbb{C}$ of probability matrices. More formally, for an estimator $\widehat{x}$ and class $\mathbb{C} \subseteq [0,1]^{n\times d}$ of possible values of Q^*, the uniform risk of $\widehat{x}$ over class $\mathbb{C}$ is

$$\sup_{x^*\in\{-1,1\}^d}\ \sup_{Q^*\in\mathbb{C}}\ \mathbb{E}[\mathcal{L}_{Q^*}(\widehat{x}, x^*)], \tag{4.5}$$

where the expectation is taken over the randomness in the observations Y for the given values of x^* and Q^*. The smallest value of the expression (4.5) across all estimators is the minimax risk.

Regime of interest: In this chapter, we focus on understanding the minimax risk as well as the risk of various computationally efficient estimators. We work in a non-asymptotic framework where we are interested in evaluating the risk in terms of the triplet (n, d, p_{obs}). We assume that $p_{\mathrm{obs}} \geq \frac{1}{n}$, which ensures that on average, at least one worker answers any question. We also operate in the regime $d \geq n$, which is commonplace in practical applications. Indeed, as also noted in earlier works [274], typical medium or large-scale crowdsourcing tasks employ tens to hundreds of workers, while the number of questions is on the order of hundreds to many thousands. We assume that the value of p_{obs} is known. This is a mild assumption since it is straightforward to estimate p_{obs} very accurately using its empirical expectation.

4.2.3 Related work

Having set up our model and notation, let us now relate it to past work in the area. For the problem of crowd labeling, the parameter-based Dawid-Skene model [60] is the dominant model, and has been widely studied [56, 85, 88, 116, 117, 151, 274]. Some papers have studied models beyond the parameter-based Dawid-Skene model. In a recent work, Khetan and Oh [125] analyze an extension of th parameter-based Dawid-Skene model where a vector $\widetilde{q} \in \mathbb{R}^n$, capturing the abilities of the workers, is supplemented with a second vector $h^* \in [0,1]^d$, and the likelihood of worker i correctly answering question j is set as $\widetilde{q}_i(1 - h^*_j + (1 - \widetilde{q}_i)h^*_j$. Although this model now has $(n + d)$ parameters instead of just n as in the parameter-based Dawid-Skene model, it retains parametric-type assumptions. Each worker and each question is described by a single parameter, and in this model the probability of correctness takes a specific form governed by these parameters. In contrast, in the permutation-based model each worker-question pair is described by a single parameter. Our permutation-based model forms a strict superset of this class. Our permutation-based model forms a strict superset of this class. Zhou et al. [275, 276] propose a model based on a certain minimax entropy principle, and Whitehill et al. [265] propose a parameter-based model, that also incorporate question difficulties; however, these algorithms have yet to be rigorously analyzed. While the present chapter addresses the setting of binary labels with symmetric error probabilities, several of these prior works also address settings with more than two classes, and where the probability of error of a worker may be asymmetric across the classes. We defer a further detailed comparison of our main results with those in earlier works to Section 4.3.4.

A related problem in the context of crowdsourcing is to estimate pairwise outcome probabilities from pairwise comparison data (Chapter 2). The permutation-based SST model considered in Chapter 2 is closely related to the permutation-based model for the workers assumed in the present chapter. However, the current chapter involves an unknown set of labels, as well as a significantly different observation model: in particular, the observed data couples the unknown matrix Q^* with the unknown labels. Moreover, rather than estimating the unknown probabilities Q^*, our primary goal in this chapter is to estimate these underlying labels, for which significantly different algorithmic ideas and proof techniques are required.

Finally, the problem of aggregating labels of crowdsourcing workers is conceptually similar to that of aggregating the outputs of multiple weak classifiers, each solving multiple classification problems [186]. Our results for this crowdsourcing problem may also shed light on fundamental theoretical properties and algorithm design guidelines for the classifier-aggregation problem.

4.3 Main results

We now turn to the statement of our main results. As noted earlier, our results are focused on the practically relevant regime where we have that:

$$p_{\text{obs}} \geq \frac{1}{n} \quad \text{and} \quad d \geq n. \tag{R}$$

We use c, c_1, c_2, c_4, c_0 to denote positive universal constants that are independent of all other problem parameters. Recall that the Q^*-loss takes values in the interval $[0,1]$.

4.3.1 Minimax risk for estimation under the permutation-based model

We begin by proving sharp upper and lower bounds on the minimax risk for the permutation-based model $\mathbb{C}_{\text{Perm}}$. The upper bound is obtained via an analysis of the following least squares estimator

$$(\widetilde{x}_{\text{LS}}, \widetilde{Q}_{\text{LS}}) \in \underset{x \in \{-1,1\}^d,\ Q \in \mathbb{C}_{\text{Perm}}}{\arg\min} \quad |\!|\!| p_{\text{obs}}^{-1}\, Y - (2Q - 11^T)\, \text{diag}(x) |\!|\!|_{\text{F}}^2. \tag{4.6}$$

The intuition behind this estimator is as follows. One can show (see the proof of Theorem 10(a) for details) that the unknowns x^* and Q^* are related to the mean of the observed matrix Y as $\mathbb{E}[Y] = p_{\text{obs}}(2Q^* - 11^T)\, \text{diag}(x^*)$. Consequently, the estimate $(\widetilde{x}_{\text{LS}}, \widetilde{Q}_{\text{LS}})$ computed via the program (4.6) equals the true solution (x^*, Q^*) when Y is replaced by its population version.

We do not know of a computationally efficient way to compute this estimate. Nonetheless, our statistical analysis provides a benchmark for comparing other computationally-efficient estimators, to be discussed in subsequent sections. The following result holds in the regime (R):

Theorem 10. *(a) For any $x^* \in \{-1,1\}^d$ and any $Q^* \in \mathbb{C}_{Perm}$, the least squares estimator $\widetilde{x}_{LS}$ has error at most*

$$\mathcal{L}_{Q^*}(\widetilde{x}_{LS}, x^*) \leq c_1 \frac{1}{np_{\text{obs}}} \log^2 d, \tag{4.7a}$$

with probability at least $1 - e^{-c_0 d \log(dn)}$.
(b) Conversely, any estimator $\widehat{x}$ has error at least

$$\sup_{Q^* \in \mathbb{C}_{DS}} \ \sup_{x^* \in \{-1,1\}^d} \mathbb{E}[\mathcal{L}_{Q^*}(\widehat{x}, x^*)] \geq c_2 \frac{1}{np_{\text{obs}}}. \tag{4.7b}$$

The lower bound holds even if the true matrix Q^* is known to the estimator.

The result of Theorem 10 has a number of important consequences. Since the permutation-based class $\mathbb{C}_{\text{Perm}}$ is significantly richer than the parameter-based Dawid-Skene class $\mathbb{C}_{\text{DS}}$, one might expect that estimation over $\mathbb{C}_{\text{Perm}}$ might require a significantly larger sample size to achieve the same accuracy. However, Theorem 10 shows that this is *not* the case: the lower bound (4.7b) holds even when the supremum over matrices Q^* is restricted to the parameter-based Dawid-Skene model $\mathbb{C}_{\text{DS}} \subset \mathbb{C}_{\text{Perm}}$. Consequently, we see that estimation over the more general permutation-based model leads to (at worst) a logarithmic penalty in the required sample size. Thus, making the restrictive assumption that the data is drawn from the parameter-based Dawid-Skene model yields little statistical advantage as compared to making the more relaxed assumption of the permutation-based model.

We note that the least squares estimator analyzed in part (a) also yields an accurate estimate of the probability matrix Q^* in the Frobenius norm, useful in settings where the calibration of workers or questions might be of interest. Again, this result holds in the regime (R):

Corollary 2. *(a) For any $x^* \in \{-1,1\}^d$ and any $Q^* \in \mathbb{C}_{Perm},$, the least squares estimate $\widetilde{Q}_{LS}$ has error at most*

$$\frac{1}{dn}|\!|\!|\widetilde{Q}_{LS} - Q^*|\!|\!|_F^2 \leq c_1 \frac{1}{np_{\text{obs}}} \log^2 d, \tag{4.8a}$$

with probability at least $1 - e^{-c_0 d \log(dn)}$.
(b) Conversely, for any answer vector $x^* \in \{-1,1\}^d$, any estimator $\widehat{Q}$ has error at least

$$\sup_{Q^* \in \mathbb{C}_{Perm}} \mathbb{E}[\frac{1}{dn}|\!|\!|\widehat{Q} - Q^*|\!|\!|_F^2] \geq c_2 \frac{1}{np_{\text{obs}}}. \tag{4.8b}$$

This lower bound holds even if the true answer vector x^* is known to the estimator.

We do not know if there exist computationally-efficient estimators that can achieve the upper bound on the sample complexity established in Theorem 10(a) uniformly over the entire permutation-based model class. In the following sections, we design and analyze polynomial-time estimators that address interesting subclasses of the permutation-based model.

4.3.2 The WAN estimator: When workers' ordering is (approximately) known

Several organizations employ crowdsourcing workers only after a thorough testing and calibration process. This section is devoted to a setting in which the workers are calibrated, in the sense that it is known how they are ordered in terms of their respective abilities. More formally, recall from Section 4.2.1 that any matrix $Q^* \in \mathbb{C}_{\text{Perm}}$ is associated with two permutations: a permutation of the workers in terms of their abilities, and a permutation of the questions in terms of their difficulty. In this section, we assume that the permutation of the workers is (approximately) known to the estimation algorithm. Note that the estimator does *not* know the permutation of the questions, nor does it know the values of the entries of Q^*.

Given a permutation π of the workers, our estimator consists of two steps, which we refer to as Windowing and Aggregating Naïvely, respectively, and accordingly term the procedure as the WAN estimator:

- Step 1 (Windowing): Compute the integer

$$k_{\text{WAN}} \in \underset{k \in \{p_{\text{obs}}^{-1} \log^{1.5}(dn), \ldots, n\}}{\arg\max} \sum_{j \in [d]} \mathbf{1}\Big\{\big|\sum_{i \in [k]} Y_{\pi^{-1}(i)j}\big| \geq \sqrt{k p_{\text{obs}} \log^{1.5}(dn)}\Big\}. \tag{4.9a}$$

- Step 2 (Aggregating Naïvely): Set $\widehat{x}_{\text{WAN}}(\pi)$ as a majority vote of the best k_{WAN} workers—that is

$$[\widehat{x}_{\text{WAN}}(\pi)]_j \in \arg\max_{b \in \{-1,1\}} \quad \sum_{i=1}^{k_{\text{WAN}}} \mathbf{1}\{Y_{\pi^{-1}(i)j} = b\} \qquad \text{for every } j \in [d]. \tag{4.9b}$$

The windowing step finds a value k_{WAN} such that the answers of the best k_{WAN} workers to most questions are significantly biased towards one of the options, thereby indicating that these workers are knowledgeable (or at least, are in agreement with each other). The second step then simply takes a majority vote of this set of the best k_{WAN} workers. We remark that it is important to choose a reasonably good value of k_{WAN} (as done in Step 1) since a much larger value could include many random workers thereby increasing the noise in the input to the second step, whereas too small a value could eliminate too much of the "signal". Both steps can be carried out in time $\mathcal{O}(nd)$.

For the case when π is an approximate ordering, we establish an oracle bound on the error. For every $j \in [d]$, let Q^*_j denote the j^{th} column of Q^*; for any ordering π of the workers, let Q^π_j denote the vector obtained by permuting the entries of Q^*_j in the order given by π, that is, with the first entry of Q^π_j corresponding to the best worker according to π, and so on. Also recall the notation π^* representing the true permutation of the workers in terms of their actual abilities. As with all of our theoretical results, the following claim holds in the regime (R):

Theorem 11. *For any $Q^* \in \mathbb{C}_{Perm}$ and any $x^* \in \{-1,1\}^d$, suppose the WAN estimator is provided with the permutation π of workers. Then for every question $j \in [d]$ such that*

$$\|Q^*_j - \frac{1}{2}\|_2^2 \geq \frac{5\log^{2.5}(dn)}{p_{\text{obs}}}, \quad \textit{and} \quad \|Q^\pi_j - Q^{\pi^*}_j\|_2 \leq \frac{\|Q^*_j - \frac{1}{2}\|_2}{\sqrt{9\log(dn)}}, \tag{4.10a}$$

we have

$$\mathbb{P}([\widehat{x}_{\mathit{WAN}}(\pi)]_j = x^*_j) \geq 1 - e^{-c_0 \log^{1.5}(dn)}. \tag{4.10b}$$

Consequently, if π is the correct permutation of the workers, then

$$\mathcal{L}_{Q^*}(\widehat{x}_{\mathit{WAN}}(\pi), x^*) \leq c_1 \frac{1}{np_{\text{obs}}} \log^{2.5} d, \tag{4.10c}$$

with probability at least $1 - e^{-c_0' \log^{1.5}(dn)}$.

At this point, we recall the lower bound of Theorem 10(b) on the estimation error in the Q^*-loss allows for any estimator. Moreover, it applies to estimators that know not only the ordering of the workers, but also the entire matrix Q^*. This lower bound matches the upper bound (4.10c) of Theorem 11, and the two results in conjunction imply that the bound (4.10c) is sharp up to logarithmic factors.

We also note that the conditions (4.10a) required for the result of Theorem 11 are sharp up to logarithmic factors. The required approximation guarantee $\|Q_j^{\pi} - Q_j^{\pi^*}\|_2 \leq \frac{\|Q_j^* - \frac{1}{2}\|_2}{\sqrt{9\log(dn)}}$, if weakened to $\|Q_j^{\pi} - Q_j^{\pi^*}\|_2 \leq 2\|Q_j^* - \frac{1}{2}\|_2$, would allow for any arbitrary permutation π. This is because every permutation π satisfies $\|Q_j^{\pi} - Q_j^{\pi^*}\|_2 \leq \|Q_j^{\pi} - \frac{1}{2}\|_2 + \|Q_j^{\pi^*} - \frac{1}{2}\|_2 = 2\|Q_j^* - \frac{1}{2}\|_2$. Secondly, there exist constants $c_0 > 0$ and $c_2 > 0$ such that if one were guaranteed a lower bound of only $\frac{c_0}{p_{\text{obs}}}$ on $\|Q_j^* - \frac{1}{2}\|_2^2$ instead of the stated condition of $\frac{5\log^{2.5}(dn)}{p_{\text{obs}}}$, then there exists a $Q^* \in \mathbb{C}_{\text{DS}}$ satisfying this weaker condition such that any estimator $\widehat{x}$ incurs an error at least $\mathbb{P}(\widehat{x}_j \neq x_j^*) \geq c_2$. Furthermore, this lower bound holds not only when the ordering of workers is exactly known, but even when the entire matrix Q^* is known. The proof for this claim follows from the construction in the proof of Theorem 10(b).

The result of Theorem 11 for the WAN algorithm has the following useful implication for the setting when the ordering of workers is *unknown* (under either of the models $\mathbb{C}_{\text{DS}}$ or $\mathbb{C}_{\text{Perm}}$). For any $Q^* \in \mathbb{C}_{\text{Perm}}$, there exists a set of workers $S_{Q^*} \subseteq [n]$ such that an estimator $\widehat{x}_S$ that takes a majority vote of the answers of the workers in S_{Q^*}, has risk at most

$$\mathcal{L}_{Q^*}(\widehat{x}_S, x^*) \leq c_1 \frac{1}{np_{\text{obs}}} \log^{2.5} d,$$

with high probability. Consequently, it suffices to design an estimator that only identifies a set of good workers and computes a majority vote of their answers. The estimator need not attempt to infer the values of the entries of Q^*, as is otherwise required, for instance, to compute maximum likelihood estimates. The estimator we propose in the next section is based on this observation.

4.3.3 The OBI-WAN estimator

In this section, we return to the setting where the ordering of the workers is *unknown*. In addition to the parameter-based Dawid-Skene and the permutation-based models introduced earlier, we also study the estimation problem in an intermediate model that lies between these two models. This intermediate model introduces a parameter $h_j^* \in [0, 1]$ that captures the difficulty of each question $j \in [d]$, along with parameters $\widetilde{q} \in \mathbb{R}^n$ associated to the workers as in the parameter-based Dawid-Skene model. Under this intermediate xmodel, the probability that worker $i \in [n]$ correctly answers question $j \in [d]$ (when the worker is asked the question) is given by

$$\mathbb{P}(Y_{ij} = x_j^*) = \widetilde{q}_i(1 - h_j^*) + \frac{1}{2}\, h_j^*, \quad \forall\ (i, j) \text{ such that } Y_{ij} \neq 0. \tag{4.11}$$

Intuitively, the parameter h_j^* corresponds to the difficulty of question j. When $h_j^* = 1$, the worker is purely stochastic and provides a random guess, while for smaller values of h_j^* the worker is more likely to provide a correct answer

.This modeling assumption leads to the parameter-based class

$$\mathbb{C}_{\text{Int}} := \left\{ Q = \widetilde{q}(1 - h)^T + \frac{1}{2}\, 1h^T \mid \text{for some } \widetilde{q} \in [\frac{1}{2},\ 1]^n,\ h \in [0, 1]^d \right\}.$$

Note that we have the nested relations $\mathbb{C}_{\text{DS}} \subset \mathbb{C}_{\text{Int}} \subset \mathbb{C}_{\text{Perm}}$; the parameter-based Dawid-Skene model is a special case of $\mathbb{C}_{\text{Int}}$ corresponding to $h = 0$.

Up to a bijective transformation of the parameters, the model (4.11) is identical to a recent model proposed independently by Khetan and Oh [125], where the probability of a correct answer is assumed to be $\widetilde{q}_i(1 - h_j^*) + (1 - \widetilde{q}_i)h_j^*$. The two models however arise from different conceptual motivations: Khetan and Oh consider the probability of correctness as a convex combination of the worker's behavior $\widetilde{q}_i$ and the opposite behavior $(1 - \widetilde{q}_i)$, whereas our consideration of rarity of adversarial behavior leads to the probability of correctness set as a convex combination of the worker's behavior $\widetilde{q}_i$ and random responses $\frac{1}{2}$.

We now describe a computationally efficient estimator, and establish sharp guarantees on its statistical risk for the intermediate parameter-based model $\mathbb{C}_{\text{Int}}$, as well as guarantees on its consistency under the permutation-based model. Our analysis of this estimator also makes contributions in the specific context of the parameter-based Dawid-Skene model. In particular, the guarantees established for computationally efficient estimators in prior works (e.g., [56, 84, 85, 88, 116, 117, 125, 274]) fall short of translating to uniform guarantees over the parameter-based Dawid-Skene model $\mathbb{C}_{\text{DS}}$ in the Q^*-loss; see Section 4.3.4 for further details. Our result in this section fills this gap by establishing sharp uniform bounds on the statistical risk over the entire parameter-based Dawid-Skene class $\mathbb{C}_{\text{DS}}$, and more generally over the entire class $\mathbb{C}_{\text{Int}}$.

Our proposed estimator operates in two steps. The first step performs an Ordering Based on Inner-products (OBI), that is, computes an ordering of the workers based on an inner product with the data. The second step calls upon the WAN estimator from Section 4.3.2 with this ordering. We thus term our proposed estimator as the OBI-WAN estimator, $\widehat{x}_{\text{OBI-WAN}}$. In order to make its description precise, we augment the notation of the WAN estimator $\widehat{x}_{\text{WAN}}(\pi)$ to let $\widehat{x}_{\text{WAN}}(\pi, Y)$ to denote the estimate given by $\widehat{x}_{\text{WAN}}(\pi)$ operating on Y when given the permutation π of workers.

An important technical issue is that re-using the observed data Y to both determine an appropriate ordering of workers as well as to estimate the desired answers, results in a violation of important independence assumptions. We resolve this difficulty by partitioning the set of questions into two sets, and using the ordering estimated from one set to estimate the desired answers for the other set and vice versa. We provide a careful error analysis for this partitioning-based estimator in the sequel. Formally, the OBI-WAN estimator $\widehat{x}_{\text{OBI-WAN}}$ is defined by the following steps:

- Step 0 (preliminary): Split the set of d questions into two sets, T_0 and T_1, with every question assigned to one of the two sets uniformly at random. Let Y_0 and Y_1 denote the corresponding submatrices of Y, containing the columns of Y associated to questions in T_0 and T_1 respectively.

- Step 1 (OBI): For $\ell \in \{0, 1\}$, let

$$u_\ell \in \arg\max_{\|u\|_2 = 1} \|Y_\ell^T u\|_2$$

denote the top eigenvector of $Y_\ell Y_\ell^T$; in order to resolve the global sign ambiguity of eigenvectors, we choose the global sign so that $\sum_{i\in[n]}[u_\ell]_i^2\mathbf{1}\{[u_\ell]_i > 0\} \geq \sum_{i\in[n]}[u_\ell]_i^2\mathbf{1}\{[u_\ell]_i < 0\}$. Let π_ℓ be the permutation of the n workers in order of the respective entries of u_ℓ (with ties broken arbitrarily).

- Step 2 (WAN): Compute the quantities

$$\widehat{x}_{\text{OBI-WAN}}(T_0) := \widehat{x}_{\text{WAN}}(Y_0, \pi_1), \quad \text{and} \quad \widehat{x}_{\text{OBI-WAN}}(T_1) := \widehat{x}_{\text{WAN}}(Y_1, \pi_0),$$

corresponding to estimates of the answers for questions in the sets T_0 and T_1, respectively.

The following theorem provides guarantees on this estimator, again in the regime (R).

Theorem 12.*(a)* *Uniformly optimal over $\mathbb{C}_{Int}$: For any $Q^* \in \mathbb{C}_{Int}$ and any $x^* \in \{-1,1\}^d$, the error incurred by the estimate $\widehat{x}_{\textit{OBI-WAN}}$ is upper bounded as*

$$\mathcal{L}_{Q^*}(\widehat{x}_{\textit{OBI-WAN}}, x^*) \leq c_1 \frac{1}{np_{\text{obs}}} \log^{2.5} d, \tag{4.12a}$$

with probability at least $1 - e^{-c_0 \log^{1.5}(dn)}$.

(b) Uniformly consistent over $\mathbb{C}_{Perm}$: For any $Q^* \in \mathbb{C}_{Perm}$ and any $x^* \in \{-1,1\}^d$, the estimate $\widehat{x}_{\textit{OBI-WAN}}$ has error at most

$$\mathcal{L}_{Q^*}(\widehat{x}_{\textit{OBI-WAN}}, x^*) \leq c_1 \frac{1}{\sqrt{np_{\text{obs}}}} \log d, \tag{4.12b}$$

with probability at least $1 - e^{-c_0 \log^{1.5}(dn)}$.

Recall that the statistical lower bound established earlier in Theorem 10(b) is also applicable to the classes $\mathbb{C}_{\text{DS}}$ and $\mathbb{C}_{\text{Int}}$. Consequently, the upper bound of Theorem 12 is sharp over these two classes.

Guarantees for OBI-WAN under the Dawid-Skene model for the Hamming error

In this section, we present results relating the performance of the OBI-WAN estimator to the settings considered in most prior works on this topic. Most of this chapter focuses on the permutation-based model, the Q^*-loss and does not account for adversarial workers. In the following theorem, we present optimality guarantees of the OBI-WAN estimator, in terms of the popular Hamming error, when data is actually faithful to the parameter-based Dawid-Skene model, and in a setting where the workers may also be adversarial (that is, where $q_i^{\text{DS}} < \frac{1}{2}$ for some workers $i \in [n]$). In particular, we show that the OBI-WAN estimator incurs a zero Hamming error under the parameter-based Dawid-Skene model when the collective intelligence (see Equation (4.3)) is sufficiently high. Our results show that OBI-WAN is optimal up to logarithmic factors, and that it also has appealing adaptivity properties.

We introduce some notation in order to describe the result involving adversarial workers. For the vector $q^{\mathrm{DS}} \in [0,1]^n$, we define two associated vectors $q^{\mathrm{DS}+}, q^{\mathrm{DS}-} \in [0,1]^n$ as $q_i^{\mathrm{DS}+} = \max\{q_i^{\mathrm{DS}}, \frac{1}{2}\}$ and $q_i^{\mathrm{DS}-} = \min\{q_i^{\mathrm{DS}}, \frac{1}{2}\}$ for every $i \in [n]$. Then we have $(q^{\mathrm{DS}} - \frac{1}{2}) = (q^{\mathrm{DS}+} - \frac{1}{2}) + (q^{\mathrm{DS}-} - \frac{1}{2})$, with $q^{\mathrm{DS}+}$ representing normal workers and $q^{\mathrm{DS}-}$ representing adversarial workers who are inclined to provide incorrect answers. As with all our theorems, the following result holds in the regime (R):

Theorem 13. *Consider any Dawid-Skene matrix of the form $Q^* = q^{DS}1^T$ for some $q^{DS} \in [0,1]^n$. Then:*

(a) If $\|q^{DS+} - \frac{1}{2}\|_2 \geq \|q^{DS-} - \frac{1}{2}\|_2 + \sqrt{\frac{4\log^{2.5}(dn)}{p_{\mathrm{obs}}}}$ and $(q^{DS} - \frac{1}{2})^T 1 \geq 0$, then for any $x^* \in \{-1,1\}^d$, the OBI-WAN *estimator satisfies*

$$\mathbb{P}(\widehat{x}_{\textit{OBI-WAN}} = x^*) \geq 1 - e^{-c_0 \log^{1.5}(dn)}. \tag{4.13a}$$

(b) Conversely, there exists a positive universal constant c such that for any $q^{DS} \in [\frac{1}{10}, \frac{9}{10}]^n$ with $\|q^{DS} - \frac{1}{2}\|_2 \leq \sqrt{\frac{c}{p_{\mathrm{obs}}}}$, any estimator $\widehat{x}$ has (normalized) Hamming error at least

$$\sup_{x^* \in \{-1,1\}^d} \mathbb{E}\big[\sum_{i=1}^d \frac{1}{d}\mathbf{1}\{\widehat{x}_i \neq x_i^*\}\big] \geq \frac{1}{10}. \tag{4.13b}$$

A couple of remarks are in order, and for the following discussion, consider the two mild conditions $\|q^{\mathrm{DS}+} - \frac{1}{2}\|_2 \geq 1.01\|q^{\mathrm{DS}-} - \frac{1}{2}\|_2$ and $(q^{\mathrm{DS}} - \frac{1}{2})^T 1 > 0$. First, we claim that under these mild conditions, the OBI-WAN estimator is *optimal* up to logarithmic factors. To see this, first observe that part (b) of Theorem 13 necessitates the condition $\|q^{\mathrm{DS}} - \frac{1}{2}\|_2 > \sqrt{\frac{c}{p_{\mathrm{obs}}}}$, for a positive universal constant c, for any non-trivial recovery guarantees. Now suppose that $\|q^{\mathrm{DS}} - \frac{1}{2}\|_2 > \sqrt{\frac{c' \log^{2.5}(dn)}{p_{\mathrm{obs}}}}$ for a large enough positive constant c'; observe that this condition is only a logarithmic factor away from the necessary condition. Then under the mild aforementioned conditions, we have $\|q^{\mathrm{DS}+} - \frac{1}{2}\|_2 \geq \|q^{\mathrm{DS}-} - \frac{1}{2}\|_2 + \sqrt{\frac{4\log^{2.5}(dn)}{p_{\mathrm{obs}}}}$. Part (a) of Theorem 13 then guarantees that the OBI-WAN estimator recovers the true answers x^* with high probability.

Secondly, Theorem 13 also shows that OBI-WAN has strong *adaptivity* guarantees under the parameter-based Dawid-Skene model. The guarantees match (up to logarithmic factors) those derived in past works such as [117]. Furthermore, our guarantees are applicable for all values of p_{obs}, as opposed to the restricted range required in past works. In more detail, for the parameter-based Dawid-Skene model, past works consider the regime $p_{\mathrm{obs}} \leq \frac{\log n}{n}$, and show that the algorithms (say, $\widehat{x}$) proposed therein incur an error upper bounded as

$$\begin{aligned}\frac{1}{d}\mathbb{E}[\sum_{i=1}^{d}\mathbf{1}\{\widehat{x}_i \neq x_i^*\}] &\leq e^{-(c_0 n p_{\text{obs}})(\frac{1}{n}\|q^{\text{DS}}-\frac{1}{2}\|_2^2)} \qquad (4.14)\\ &\leq e^{-c_0 \log n} \qquad \text{in the regime } p_{\text{obs}} \leq \tfrac{\log n}{n}.\end{aligned}$$

Observe that this bound is no better than the bound of $\frac{1}{2}$ achieved by random guessing unless

$$\|q^{\text{DS}} - \frac{1}{2}\|_2^2 \geq \frac{c_2}{p_{\text{obs}}},$$

where $c_2 > 0$ is a universal constant. On the other hand, part (a) of Theorem 13 above shows that when $\|q^{\text{DS}} - \frac{1}{2}\|_2^2 \geq \frac{4\log^{2.5}(dn)}{p_{\text{obs}}}$, the error incurred by OBI-WAN is upper bounded as

$$\frac{1}{d}\mathbb{E}[\sum_{i=1}^{d}\mathbf{1}\{\widehat{x}_i \neq x_i^*\}] \leq e^{-c_0 \log^{1.5}(dn)}.$$

Furthermore, the OBI-WAN algorithm and the associated guarantees are not restricted to the regime $p_{\text{obs}} \leq \frac{\log n}{n}$.

The OBI-WAN estimator is thus not only consistent under the permutation-based model, but also has strong adaptivity guarantees under the Dawid-Skene model.

One application of Theorem 13 is to the setting that has been the focus of this chapter, where we have no adversarial workers. In this case, $q^{\text{DS}-} = 0$, and $q^{\text{DS}+} = q^{\text{DS}}$, and the upper and lower bounds match upto a logarithmic factor. The upper bound shows that when $\|q^{\text{DS}} - \frac{1}{2}\|_2 \geq \sqrt{\frac{4\log^{2.5}(dn)}{p_{\text{obs}}}}$ the Hamming error is vanishingly small while the lower bound shows that there is a universal constant c such that the Hamming error is essentially as large as possible when $\|q^{\text{DS}} - \frac{1}{2}\|_2 \leq \sqrt{\frac{c}{p_{\text{obs}}}}$.

The results of Theorem 12 and Theorem 13 in conjunction show that the OBI-WAN estimator not only has optimal guarantees (up to logarithmic factors) in terms of the models and metrics popular in past literature, but is also efficient in terms of the more general models and metric introduced here.

4.3.4 Past work and the Q^*-loss

Several past works have introduced computationally-efficient estimation algorithms, and provided theoretical guarantees for these algorithms under the parameter-based Dawid-Skene model. These guarantees apply to the Hamming metric, and usually quantify the sample complexity required for exact recovery of all the questions with high probability. In this section, we consider the implications for such guarantees for the goal of this chapter—namely, that of establishing uniform guarantees under the Q^*-loss. We find that guarantees from earlier works—for the purposes of establishing uniform guarantees over the parameter-based Dawid-Skene model in the Q^*-loss—are either inapplicable, or lead to sub-optimal guarantees.

To be fair, some of this past work applies to settings more general than our chapter, including problems with more than two classes, and problems where the probability of error of a worker may be asymmetric across the classes. The present chapter, on the other hand, considers the setting of binary labels with symmetric error probabilities, and accordingly, all comparison made in this section pertain to this setting. We note that the various prior works make different assumptions regarding the choice of questions assigned to each worker, and in order to bring these works under the same umbrella, we assume that each of the n workers answers each of the d questions (that is, $p_{\text{obs}} = 1$) unless specified otherwise. As indicated earlier, in this section we restrict attention to the parameter-based Dawid-Skene model $\mathbb{C}_{\text{DS}}$.

Note that when the guarantee claimed in a past work requires certain additional conditions that are not satisfied, one can always appeal to the naïve bound

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) \leq \frac{1}{n}\|q^{\text{DS}} - \frac{1}{2}\|_2^2. \tag{4.15}$$

Thus, in all of comparisons with past work, we take the minimum of this bound, and the bound provided by their work. We show below that in each of the prior works, this augmented guarantee has weaker scaling than the bound strictly weaker scaling than the scaling of

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) \leq \frac{1}{n}\log^{2.5} d, \tag{4.16}$$

achieved by the OBI-WAN estimator for the parameter-based Dawid-Skene model (see Theorem 12(a)) when $p_{\text{obs}} = 1$.

Ghosh et al. [88] The guarantees for recovery provided in the paper [88] require the lower bound

$$\|q^{\text{DS}} - \frac{1}{2}\|_2^2 \geq c_0\sqrt{n \log n} \tag{4.17}$$

to be satisfied, where c_0 is a positive universal constant. This requirement means that it is not possible to translate the bounds of [88] to a uniform bound over the entire parameter-based Dawid-Skene class in the Q^*-loss. For instance, for a DS matrix given by the vector

$$q_i^{\text{DS}} = \begin{cases} 1 & \text{if } i \leq \sqrt{n} \\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{4.18}$$

the guarantees of [88] are inapplicable, and the naïve bound of $\frac{1}{n}\|q^{\text{DS}} - \frac{1}{2}\|_2^2 = \frac{1}{\sqrt{n}}$ is sub-optimal.

Karger et al. [116, 117], Khetan and Oh [125] The guarantees from this set of works assume that $p_{\text{obs}} = \mathcal{O}(\frac{\log d}{d})$.[3] The assumption stems from the use of message passing algorithms, where the analysis requires a certain "locally tree-like" worker-question assignment graph which is guaranteed to hold in this regime. Moreover, the results of [116] apply to a particular subset of the parameter-based Dawid-Skene model, for which is it assumed that $q^{\text{DS}} \in \{\frac{1}{2}, 1\}^n$.

Let us evaluate these guarantees from the perspective of our requirements, namely to obtain uniform guarantees on the Q^*-loss under the parameter-based Dawid-Skene model across different values of the problem parameters. When $p_{\text{obs}} = \mathcal{O}(\frac{\log d}{d})$, then the trivial upper bound of 1 on the Q^*-loss is only a logarithmic factor away from the lower bound of $\frac{1}{np_{\text{obs}}}$ given by Theorem 10(b) in the present chapter. Consequently, any result will then be sandwiched between these two bounds, and can yield at most a logarithmic improvement over the trivial upper bound in this regime. On the other hand, the guarantees derived in [116, 117, 125] are loose when p_{obs} takes larger values. For instance, when $p_{\text{obs}} \geq \frac{1}{\sqrt{n}}$, these bounds reduce to the trivial property that the number of answers decoded incorrectly is upper bounded by d. Consequently, in this regime, these analyses yield an upper bound of $\frac{1}{n}\|q^{\text{DS}} - \frac{1}{2}\|_2^2$; note that this bound could be as large as $\frac{1}{4}$.

Dalvi et al. [56] For the setting described in equation (4.18), the bound of Dalvi et al. only guarantees that the number of answers estimated incorrectly is upper bounded by cd, for some constant $c > 0$. This guarantee translates to a suboptimal bound of order $\frac{1}{\sqrt{n}}$ on the Q^*-loss.

Zhang et al. [274] Zhang et al. [274] assume the existence of three groups of workers such that the second largest singular value of a certain set of matrices capturing the correlations between the probabilities of correctness of workers in the groups are all lower bounded by a parameter, denoted as σ_L. Their results require, among other conditions, that $d \geq (\sigma_L)^{-13}$. It turns out that for a large number of settings of interest, this condition is quite prohibitive. Here is a simple example to illustrate this issue. Suppose that

$$q_i^{\text{DS}} = \begin{cases} 1 & \text{if } i \leq \sqrt{n}\log d \\ \frac{1}{2} & \text{otherwise} \end{cases}. \tag{4.19}$$

In order to apply the bounds of [274] to this setting, we must have $d \geq n^{14}$. One can see that this condition is prohibitive, even when the number of workers n is as small as 10. The naïve bound of $\frac{1}{n}\|q^{\text{DS}} - \frac{1}{2}\|_2^2 = \frac{\log d}{4\sqrt{n}}$ is also suboptimal. We note that on the other hand, the

[3] The setting analyzed in these papers is slightly different from ours when $p_{\text{obs}} < 1$. Specifically, the paper [125] assumes that the sets of questions assigned to the workers are chosen based on a certain regular random bipartite graph, with each worker answering dp_{obs} questions and each question being answered by np_{obs} workers. We think that the assumptions on the worker-question assignment in [125] and those made in the present chapter may have similar guarantees. In the spirit of allowing for a comparison between the two works, we consider their guarantees as applicable for our setting as well.

problem (4.19) is not actually hard: a simple analysis of the majority voting algorithm leads to a guarantee that all the questions will be decoded correctly with a high probability.

Gao et al. [84] Gao et al. [84] present an algorithm and associated guarantees to estimate the true labels under the parameter-based Dawid-Skene model when the worker abilities q^{DS} are (approximately) known. In order to estimate the value of q^{DS}, they employ one of the two following methods: (a) The algorithm of Zhang et al. [274], which results in the same limitations as those for the guarantees of [274] discussed earlier; and (b) An estimator based on the work of Gao and Zhou [85] that prohibits settings where most labels in may have the same true value, thereby yielding only the naïve bound of 1 on the minimax risk of estimation under the Q^*-loss.

Majority voting Finally, let us comment on a relatively simple estimator—namely, the majority voting estimator. It computes the sign vector $\widetilde{x}_{\mathrm{MV}} \in \{-1, +1\}^d$ with entries

$$[\widetilde{x}_{\mathrm{MV}}]_j \in \arg\max_{b \in \{-1,1\}} \sum_{i=1}^{n} \mathbf{1}\{Y_{ij} = b\} \qquad \text{for all } j \in [d].$$

Here we use $\mathbf{1}\{\cdot\}$ to denote the indicator function. In Appendix 4.A, we show that the majority voting estimator also incurs an expected Q^*-loss lower bounded as order $\frac{1}{\sqrt{n}}$ under the parameter-based Dawid-Skene model.

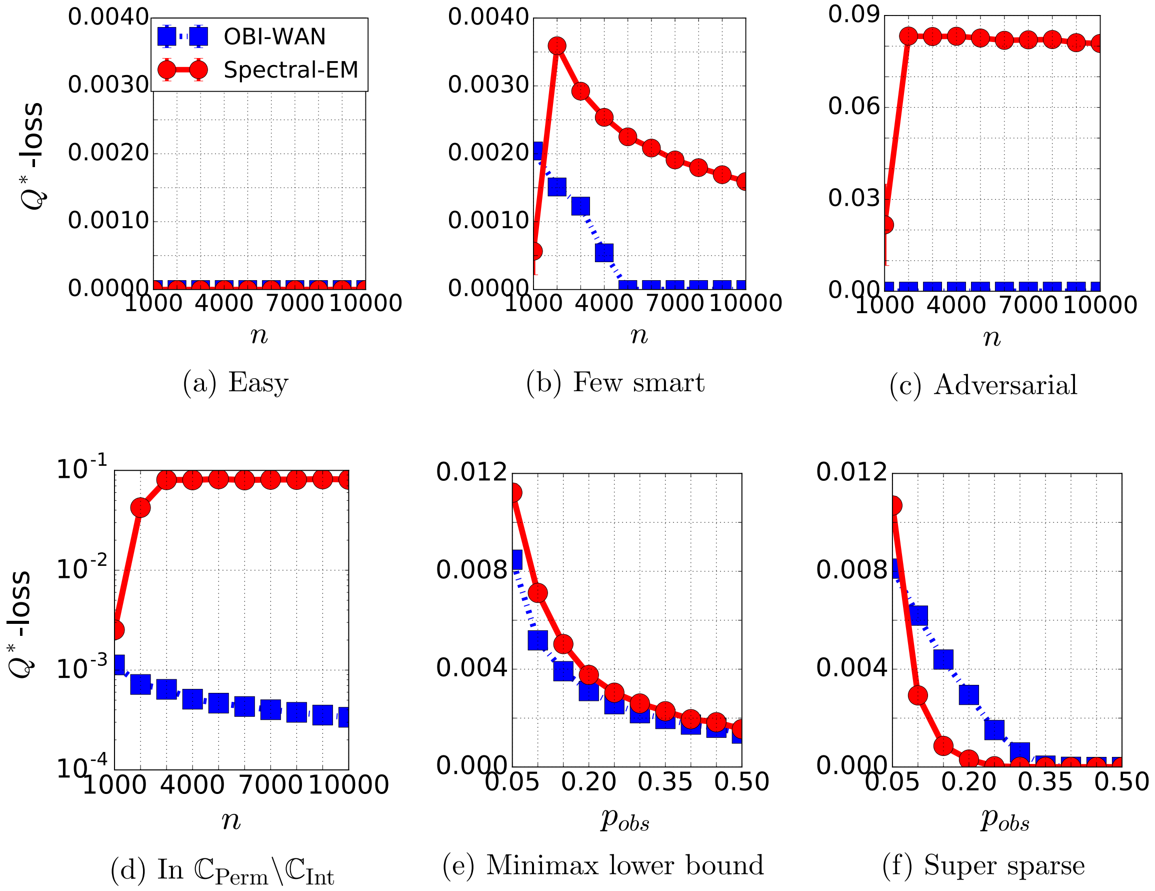
4.4 Simulations

In this section, we present numerical simulations comparing our proposed OBI-WAN estimator (introduced in Section 4.3.3) to the Spectral-EM estimator due to Zhang et al. [274], which to the best of our knowledge, has the strongest established guarantees in the literature. For the Spectral-EM estimator, we used an implementation provided by the authors of the paper [274]. The code for the OBI-WAN estimator as well as the constituent WAN estimator is available on the author's website.

The results from our simulations are plotted in Figure 4.1. The plots in the six panels (a) through (f) of the figure are discussed below.

(a) Easy: $Q^* = q^{\mathrm{DS}} 1^T \in \mathbb{C}_{\mathrm{DS}}$ where $q_i^{\mathrm{DS}} = \frac{9}{10}$ if $i < \frac{n}{2}$, and $q_i^{\mathrm{DS}} = \frac{1}{2}$ otherwise. The parameter n is varied, and the regime of operation is $(d = n,\ p_{\mathrm{obs}} = 1)$. In this setting, both estimators correctly recover x^*.

(b) Few smart: $Q^* = q^{\mathrm{DS}} 1^T \in \mathbb{C}_{\mathrm{DS}}$ where $q_i^{\mathrm{DS}} = \frac{9}{10}$ if $i < \sqrt{n}$, and $q_i^{\mathrm{DS}} = \frac{1}{2}$ otherwise. The parameter n is varied, and the regime of operation $(d = n,\ p_{\mathrm{obs}} = 1)$. Even though the data is drawn from the parameter-based Dawid-Skene model, the error of Spectral-EM is much higher than that of the OBI-WAN estimator. Recall that the OBI-WAN estimator has



(a) Easy

(b) Few smart

(c) Adversarial

(d) In $\mathbb{C}_{\text{Perm}} \backslash \mathbb{C}_{\text{Int}}$

(e) Minimax lower bound

(f) Super sparse

Figure 4.1: Results from numerical simulations comparing the OBI-WAN and Spectral-EM estimators. The plots in panels (a)-(d) measure the Q^*-loss as a function of n, and the plots in panels (e)-(f) measure the Q^*-loss as a function of p_{obs}. Each point is an average of over 20 trials. Recall that when Q^* follows the parameter-based Dawid-Skene model, as in panels (a)-(c), (e)-(f), the Hamming error is proportional to the Q^*-loss. Also note that the Y-axis of panel (d) is plotted on a logarithmic scale.

uniform guarantees of recovery over the entire parameter-based Dawid-Skene class, unlike the estimators in prior literature.

(c) Adversarial: $Q^* = q^{\text{DS}} 1^T \in \mathbb{C}_{\text{DS}}$ where $q_i^{\text{DS}} = \frac{9}{10}$ if $i < \frac{n}{4} + \sqrt{n}$, $q_i^{\text{DS}} = \frac{1}{10}$ if $i > \frac{3n}{4}$, and $q_i^{\text{DS}} = \frac{1}{2}$ otherwise. The parameter n is varied, and the regime of operation is ($d = n$, $p_{\text{obs}} = 1$). This set of simulations moves beyond the assumption that the entries of Q^* are lower bounded by $\frac{1}{2}$, and allows for adversarial workers. The OBI-WAN estimator is successful in such a setting as well.

(d) In $\mathbb{C}_{\text{Perm}}$ but outside $\mathbb{C}_{\text{Int}}$: $Q^*_{ij} = \frac{9}{10}$ if ($i < \sqrt{n}$ or $j < \frac{d}{2}$), and $Q^*_{ij} = \frac{1}{2}$ otherwise. The parameter n is varied, and the regime of operation is ($d = n,\ p_{\text{obs}} = 1$). Here we have $Q^* \in \mathbb{C}_{\text{Perm}} \backslash \mathbb{C}_{\text{Int}}$. The Q^*-loss incurred by the OBI-WAN estimator decays as $\frac{1}{\sqrt{n}}$, whereas the Q^*-loss of Spectral-EM grows remains a constant.

(e) Minimax lower bound: $Q^* = q^{\text{DS}} 1^T \in \mathbb{C}_{\text{DS}}$ where $q^{\text{DS}}_i = \frac{9}{10}$ if $i \leq \frac{5}{p_{\text{obs}}}$ and $q^{\text{DS}}_i = \frac{1}{2}$ otherwise. The parameter p_{obs} is varied, and the regime of operation is ($d = 1000,\ n = 1000$). This setting is the cause of the minimax lower bound of Theorem 10(b). The error of both estimators, in this case, behaves in an almost identical manner with a scaling of $\frac{1}{p_{\text{obs}}}$.

(f) Super sparse: $Q^* = q^{\text{DS}} 1^T \in \mathbb{C}_{\text{DS}}$ where $q^{\text{DS}}_i = \frac{9}{10}$ if $i \leq \frac{n}{10}$ and $q^{\text{DS}}_i = \frac{1}{2}$ otherwise. The parameter p_{obs} is varied, and the regime of operation is ($d = 1000,\ n = 1000$). We see that the OBI-WAN estimator incurs a relatively higher error when data is very sparse — more generally, we have observed a higher error when $p_{\text{obs}} = o(\frac{\log^2(dn)}{n})$, and this gap is also reflected in our upper bounds for the OBI-WAN estimator in Theorem 12(a) and Theorem 13(a) that are loose by precisely a polylogarithmic factor as compared to the associated lower bounds.

The relative benefits and disadvantages of of the proposed OBI-WAN estimator, as observed from the simulations, may be summarized as follows. In terms of limitations, the error of OBI-WAN is higher than prior works when p_{obs} is small (as observed in the super-sparse case) or when n and d are small (for instance, less than 200). On the positive side, the simulations reveal that the OBI-WAN estimator leads to accurate estimates in a variety of settings, providing uniform guarantees over the $\mathbb{C}_{\text{DS}}$ and $\mathbb{C}_{\text{Int}}$ classes, and demonstrating significant robustness in more general settings in comparison to the best known estimator in the literature.

4.5 Discussion

We proposed a flexible permutation-based model for the noise in crowdsourced labels, and by establishing fundamental theoretical guarantees on estimation, we showed that this model allows for robust and statistically efficient estimation of the true labels in comparison to the popular parameter-based Dawid-Skene model. We hope that this win-win feature of the permutation-based model will encourage researchers and practitioners to further build on the permutation-based core of this model. In addition, we proposed a new metric for theoretical evaluation of algorithms for this problem that eliminates drawbacks of the Hamming metric used in prior works. Using our approach towards estimation under such a general class, we proposed a robust estimator, OBI-WAN, that unlike the estimators in prior literature, has optimal uniform guarantees over the entire parameter-based Dawid-Skene model. In more general settings, the OBI-WAN estimator is uniformly optimal over the parameter-based class $\mathbb{C}_{\text{Int}}$ that is richer than the parameter-based Dawid-Skene model, and is uniformly consistent over the entire permutation-based model.

This work gives rise to several open problems that are theoretically challenging and potentially useful in practice. First, the problem of establishing optimal minimax risk under the permutation-based model for computationally-efficient estimators remains open. Second, the focus of Theorem 10 of the present chapter is on the global minimax error, and it is of interest to obtain sharp bounds on local adaptivity under the permutation-based model. Such adaptive bounds are obtained for Dawid-Skene and related parameter-based models in the papers [84, 117, 125, 274] as well as in Theorem 12 of the present chapter for the parameter-based Dawid-Skene model. Third, it will be useful to extend the proposed permutation-based model and associated algorithms to more general settings in crowdsourcing such as having multiple (more than two) choice questions. Finally, we considered a symmetric setting where the error probability is independent of the true answer, and extension to the asymmetric case remains open.

4.6 Proofs

In this section, we present the proofs of our theoretical results. In the proofs we ignore floors and ceilings unless critical to the proof. We assume that n and d are greater than some universal constants; the case of smaller values of these parameters are then directly implied by only changing the constant prefactors.

4.6.1 Proof of Theorem 10(a): Minimax upper bound

In this section, we prove the minimax upper bound stated in part (a) of Theorem 10. The proof is divided into two parts, where in the first part, we obtain an upper bound on the error term $|\!|\!|(2Q^* - 11^T)\mathrm{diag}(x^*) - (2\widetilde{Q}_{\mathrm{LS}} - 11^T)\mathrm{diag}(\widetilde{x}_{\mathrm{LS}})|\!|\!|_{\mathrm{F}}^2$, following which we convert this bound to one on $\mathcal{L}_{Q^*}(x^*, \widetilde{x}_{\mathrm{LS}})$.

We begin with the first part of the proof, where we bound the error in estimating the product term $(2Q^* - 11^T)\mathrm{diag}(x^*)$. Let us rewrite our observation model in a "linearized" fashion that is convenient for subsequent analysis. In particular, let us define a random matrix $W \in \mathbb{R}^{n \times d}$ with entries independently drawn from the distribution

$$W_{ij} = \begin{cases} 1 - p_{\mathrm{obs}}(2Q^*_{ij} - 1)x^*_j & \text{w.p.} \quad p_{\mathrm{obs}}\Big(Q^*_{ij}\big(\frac{1+x^*_j}{2}\big) + (1 - Q^*_{ij})\big(\frac{1-x^*_j}{2}\big)\Big) \\ -1 - p_{\mathrm{obs}}(2Q^*_{ij} - 1)x^*_j & \text{w.p.} \quad p_{\mathrm{obs}}\Big(Q^*_{ij}\big(\frac{1-x^*_j}{2}\big) + (1 - Q^*_{ij})\big(\frac{1+x^*_j}{2}\big)\Big) \\ -p_{\mathrm{obs}}(2Q^*_{ij} - 1)x^*_j & \text{w.p.} \quad 1 - p_{\mathrm{obs}}, \end{cases} \tag{4.20}$$

where "w.p." is a shorthand for "with probability". One can verify that $\mathbb{E}[W] = 0$, every entry of W is bounded by 2 in absolute value, and moreover that our observed matrix Y can be written in the form

$$\frac{1}{p_{\mathrm{obs}}} Y = (2Q^* - 11^T)\ \mathrm{diag}(x^*) + \frac{1}{p_{\mathrm{obs}}} W. \tag{4.21}$$

Let Π_n denote the set of all permutations of the n workers, and let Σ_d denote the set of all permutations of the d questions. For any pair of permutations $(\pi, \sigma) \in \Pi_n \times \Sigma_d$, define the set

$$\mathbb{C}_{\text{Perm}}(\pi, \sigma) := \Big\{ Q \in [0,1]^{n \times d} \; | Q_{ij} \geq Q_{i'j'} \text{ whenever } \pi(i) \leq \pi(i') \text{ and } \sigma(j) \leq \sigma(j') \Big\},$$

corresponding to the subset of $\mathbb{C}_{\text{Perm}}$ consisting of matrices that are faithful to the permutations π and σ. For any fixed $x \in \{-1,1\}^d$, $\pi \in \Pi_n$ and $\sigma \in \Sigma_d$, define the matrix

$$\widetilde{Q}(\pi, \sigma, x) \in \underset{Q \in \mathbb{C}_{\text{Perm}}(\pi,\sigma)}{\arg\min} \; \mathcal{C}(Q, x),$$
$$\text{where} \quad \mathcal{C}(Q, x) := |\!|\!| \frac{1}{p_{\text{obs}}} Y - (2Q - 11^T)\text{diag}(x) |\!|\!|_{\text{F}}^2.$$

Using this notation, we can rewrite the least squares estimator (4.6) in the compact form

$$(\widetilde{x}_{\text{LS}}, \widetilde{\pi}_{\text{LS}}, \widetilde{\sigma}_{\text{LS}}) \in \underset{\substack{(\pi,\sigma) \in \Pi_n \times \Sigma_d \\ x \in \{-1,1\}^d}}{\arg\min} \; \mathcal{C}(\widetilde{Q}(\pi,\sigma,x), x), \quad \text{and} \quad \widetilde{Q}_{\text{LS}} = \widetilde{Q}(\widetilde{\pi}_{\text{LS}}, \widetilde{\sigma}_{\text{LS}}, \widetilde{x}_{\text{LS}}).$$

For the purposes of analysis, let us define the set

$$\mathcal{P} := \Big\{ (\pi, \sigma, x) \in \Pi_n \times \Sigma_d \times \{-1,1\}^d \mid \mathcal{C}(\widetilde{Q}(\pi,\sigma,x), x) \leq \mathcal{C}(Q^*, x^*) \Big\}. \tag{4.22}$$

With this set-up, we claim that it is sufficient to show the following: fix a triplet $(\pi, \sigma, x) \in \mathcal{P}$, for this fixed triplet there is a universal constant c_1 such that

$$\mathbb{P}\big(|\!|\!| (2\widetilde{Q}(\pi,\sigma,x) - 11^T)\text{diag}(x - x^*) |\!|\!|_{\text{F}}^2 \leq c_1 \frac{d}{p_{\text{obs}}} \log^2 d \big) \geq 1 - e^{-4d\log(dn)}. \tag{4.23}$$

Given this bound, since the cardinality of the set $\mathcal{P}$ is upper bounded by $e^{3d\log d}$ (since $d \geq n$), a union bound over all these permutations applied to (4.23) yields

$$\mathbb{P}\big(\max_{(\pi,\sigma,x) \in \mathcal{P}} |\!|\!| (2\widetilde{Q}(\pi,\sigma,x) - 11^T)\text{diag}(x - x^*) |\!|\!|_{\text{F}}^2 \leq c_1 \frac{d \log^2 d}{p_{\text{obs}}} \big) \geq 1 - e^{-d\log(dn)}.$$

The set $\mathcal{P}$ is guaranteed to be non-empty since the true permutations π^* and σ^* corresponding to Q^* and the true answer x^* always lie in $\mathcal{P}$, and consequently, the above tail bound yields the claimed result.

The remainder of our analysis is devoted to proving the bound (4.23). Given any triplet $(\pi, \sigma, x) \in \mathcal{P}$, we define the matrices

$$V^* := (2Q^* - 11^T)\, \text{diag}(x^*), \quad \text{and} \quad \widetilde{V}(\pi,\sigma,x) := (2\widetilde{Q}(\pi,\sigma,x) - 11^T)\, \text{diag}(x).$$

Henceforth, for brevity, we refer to the matrix $\widetilde{V}(\pi,\sigma,x)$ simply as $\widetilde{V}$ and the matrix $\widetilde{Q}(\pi,\sigma,x)$ simply as $\widetilde{Q}$, since the values of the associated quantities (π,σ,x) are fixed and clear from context.

Applying the linearized form (4.21) of our observation model to the inequality that defines the set (4.22), some simple algebraic manipulations yield

$$\frac{1}{2}\|V^* - \widetilde{V}\|_{\mathrm{F}}^2 \leq \frac{1}{p_{\mathrm{obs}}}\langle\langle V^* - \widetilde{V},\ W\rangle\rangle. \tag{4.24}$$

The following lemma uses this inequality to obtain an upper bound on the quantity $\frac{1}{2}\|V^* - \widetilde{V}\|_{\mathrm{F}}^2$.

Lemma 19. *There exists a universal constant $c_1 > 0$ such that*

$$\mathbb{P}\Big(\|V^* - \widetilde{V}\|_F^2 \leq c_1 \frac{d \log^2 d}{p_{\mathrm{obs}}}\Big) \geq 1 - e^{-4d\log(dn)}. \tag{4.25}$$

See the end of this section for the proof of this lemma. This completes the first part of the proof.

In the second part of the proof, we now convert our bound (4.25) on the Frobenius norm $\|V^* - \widetilde{V}\|_{\mathrm{F}}$ into one on the error in estimating x^* under the Q^*-loss. The following lemma is useful for this conversion:

Lemma 20. *For any pair of matrices $A_1, A_2 \in \mathbb{R}_+^{n\times d}$ and any pair of vectors $v_1, v_2 \in \{-1, 1\}^d$, we have*

$$\|A_1\ diag(v_1 - v_2)\|_F^2 \leq 4\|A_1\ diag(v_1) - A_2\ diag(v_2)\|_F^2. \tag{4.26}$$

See the end of this section for the proof of this claim.

Recall our assumption that every entry of the matrices Q^* and $\widetilde{Q}$ is at least $\frac{1}{2}$. Consequently, we can apply Lemma 20 with $A_1 = (Q^* - \frac{1}{2}11^T)$, $A_2 = (\widetilde{Q} - \frac{1}{2}11^T)$, $v_1 = x^*$ and $v_2 = x$ to obtain the inequality

$$\begin{aligned}\|(Q^* - \frac{1}{2}11^T)\mathrm{diag}(x^* - x)\|_{\mathrm{F}}^2 &\leq 4\|(Q^* - \frac{1}{2}11^T)\mathrm{diag}(x^*) - (\widetilde{Q} - \frac{1}{2}11^T)\mathrm{diag}(\widehat{x})\|_{\mathrm{F}}^2 \\ &= 4\|V^* - \widetilde{V}\|_{\mathrm{F}}^2. \end{aligned} \tag{4.27}$$

Coupled with Lemma 19, this bound yields the desired result (4.23).

Proof of Lemma 19

Our proof of this lemma closely follows along the lines of the proof of Theorem 1 in Chapter 2.

Denote the error in the estimate as $\widehat{\Delta} := \widetilde{V} - V^*$. Then from the inequality (4.24), have

$$\frac{1}{2}\|\widehat{\Delta}\|_{\mathrm{F}}^2 \leq \frac{1}{p_{\mathrm{obs}}}\langle\langle W,\ \widehat{\Delta}\rangle\rangle. \tag{4.28}$$

For the quadruplet (π, σ, x, V^*) under consideration, define the set

$$\mathbb{V}_{\text{DIFF}}(\pi, \sigma, x, V^*) := \Big\{\alpha(V - V^*) \; | V = (2Q - 11^T)\text{diag}(x), \\ Q \in \mathbb{C}_{\text{Perm}}(\pi, \sigma), \; \alpha \in [0, 1]\Big\}.$$

Since the terms π, σ, x and V^* are fixed for the purposes of this proof, we will use the abbreviated notation $\mathbb{V}_{\text{DIFF}}$ for $\mathbb{V}_{\text{DIFF}}(\pi, \sigma, x, V^*)$.

For each choice of radius $t > 0$, define the random variable

$$Z(t) := \sup_{\substack{D \in \mathbb{V}_{\text{DIFF}}, \\ \|D\|_{\text{F}} \leq t}} \frac{1}{p_{\text{obs}}} \langle\!\langle D, \; W \rangle\!\rangle. \tag{4.29a}$$

Using the basic inequality (4.28), the Frobenius norm error $\|\widehat{\Delta}\|_{\text{F}}$ then satisfies the bound

$$\frac{1}{2}\|\widehat{\Delta}\|_{\text{F}}^2 \leq \frac{1}{p_{\text{obs}}} \langle\!\langle W, \; \widehat{\Delta} \rangle\!\rangle \; \leq \; Z\big(\|\widehat{\Delta}\|_{\text{F}}\big). \tag{4.29b}$$

Thus, in order to obtain a high probability bound, we need to understand the behavior of the random quantity $Z(t)$.

One can verify that the set $\mathbb{V}_{\text{DIFF}}$ is star-shaped, meaning that $\alpha D \in \mathbb{V}_{\text{DIFF}}$ for every $\alpha \in [0, 1]$ and every $D \in \mathbb{V}_{\text{DIFF}}$. Using this star-shaped property, we are guaranteed that there is a non-empty set of scalars $\delta_{n,d} > 0$ satisfying the critical inequality

$$\mathbb{E}[Z(\delta_{n,d})] \leq \frac{\delta_{n,d}^2}{2}. \tag{4.29c}$$

Our interest is in an upper bound to the smallest (strictly) positive solution $\delta_{n,d}$ to the critical inequality (4.29c), and moreover, our goal is to show that for every $t \geq \delta_{n,d}$, we have $\|\widehat{\Delta}\|_{\text{F}} \leq c\sqrt{t\delta_{n,d}}$ with high probability.

Define a "bad" event

$$\mathcal{A}_t := \Big\{\exists \Delta \in \mathbb{V}_{\text{DIFF}} \mid \|\Delta\|_{\text{F}} \geq \sqrt{t\delta_{n,d}} \quad \text{and} \quad \frac{1}{p_{\text{obs}}} \langle\!\langle \Delta, \; W \rangle\!\rangle \geq 2\|\Delta\|_{\text{F}}\sqrt{t\delta_{n,d}}\Big\}. \tag{4.30}$$

Using the star-shaped property of $\mathbb{V}_{\text{DIFF}}$, it follows by a rescaling argument that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_{n,d}) \geq 2\delta_{n,d}\sqrt{t\delta_{n,d}}] \qquad \text{for all } t \geq \delta_{n,d}.$$

The following lemma helps control the behavior of the random variable $Z(\delta_{n,d})$.

Lemma 21. *For any $\delta > 0$, the mean of $Z(\delta)$ is upper bounded as*

$$\mathbb{E}[Z(\delta)] \leq c_1 \frac{n + d}{p_{\text{obs}}} \log^2(nd), \tag{4.31a}$$

and for every $u > 0$, its tail probability is bounded as

$$\mathbb{P}\Big(Z(\delta) > \mathbb{E}[Z(\delta)] + u\Big) \leq \exp\Big(\frac{-c_2 u^2 p_{\text{obs}}}{\delta^2 + \mathbb{E}[Z(\delta)] + u}\Big), \tag{4.31b}$$

where c_1 and c_2 are positive universal constants.

See the end of this section for the proof of this lemma.

Setting $u = \delta_{n,d}\sqrt{t\delta_{n,d}}$ in the tail bound (4.31b), we find that

$$\mathbb{P}\big(Z(\delta_{n,d}) > \mathbb{E}[Z(\delta_{n,d})] + \delta_{n,d}\sqrt{t\delta_{n,d}}\big) \leq \exp\Big(\frac{-c_2(\delta_{n,d}\sqrt{t\delta_{n,d}})^2 p_{\text{obs}}}{\delta_{n,d}^2 + \mathbb{E}[Z(\delta_{n,d})] + \delta_{n,d}\sqrt{t\delta_{n,d}}}\Big), \text{ for all } t > 0.$$

By the definition of $\delta_{n,d}$ in (4.29c), we have $\mathbb{E}[Z(\delta_{n,d})] \leq \delta_{n,d}^2 \leq \delta_{n,d}\sqrt{t\delta_{n,d}}$ for any $t \geq \delta_{n,d}$, and with these relations we obtain the bound

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_{n,d}) \geq 2\delta_{n,d}\sqrt{t\delta_{n,d}}] \ \leq \ \exp\big(-\frac{c_2}{3}\delta_{n,d}\sqrt{t\delta_{n,d}}p_{\text{obs}}\big), \quad \text{for all } t \geq \delta_{n,d}.$$

Consequently, either $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq \sqrt{t\delta_{n,d}}$, or we have $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} > \sqrt{t\delta_{n,d}}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, our basic inequality implies that $\frac{1}{2}|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2 \leq 2|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}\sqrt{t\delta_{n,d}}$ and hence $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_{n,d}}$. Putting together the pieces yields that

$$\mathbb{P}\big(|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_{n,d}}\big) \geq 1 - \exp\big(-\frac{c_2}{3}\delta_{n,d}\sqrt{t\delta_{n,d}}p_{\text{obs}}\big), \qquad \text{valid for all } t \geq \delta_{n,d}. \tag{4.32}$$

Finally, from the bound on the expected value of $Z(t)$ in Lemma 21, we see that the critical inequality (4.29c) is satisfied for $\delta_{n,d} = \sqrt{\frac{2c_1(n+d)}{p_{\text{obs}}}}\log(nd)$. Setting $t = \delta_{n,d} = \sqrt{\frac{2c_1(n+d)}{p_{\text{obs}}}}\log(nd)$ in equation (4.32) yields the claimed result.

Proof of Lemma 20

Consider any four scalars $a_1 \geq 0, a_2 \geq 0, b_1 \in \{-1, 1\}$ and $b_2 \in \{-1, 1\}$. If $b_1 = b_2$ then

$$(a_1 b_1 - a_1 b_2)^2 = 0 \leq (a_1 b_1 - a_2 b_2)^2.$$

Otherwise we have $b_1 = -b_2$. In this case, since a_1 and a_2 have the same sign,

$$(a_1 b_1 - a_2 b_2)^2 \geq (a_1 b_1)^2 = \frac{1}{4}(a_1 b_1 - a_1 b_2)^2.$$

The two results above in conjunction yield the inequality $(a_1(b_1 - b_2))^2 \leq 4(a_1 b_1 - a_2 b_2)^2$. Applying the above argument to each entry of the matrices $A_1\text{diag}(v_1 - v_2)$ and $(A_1\text{diag}(v_1) - A_2\text{diag}(v_2))$ yields the claim.

Proof of Lemma 21

We need to prove the upper bound (4.31a) on the mean, as well as the tail bound (4.31b).

Upper bounding the mean: We upper bound the mean by using Dudley's entropy integral, as well as some auxiliary results on metric entropy. Given a set $\mathbb{C}$ equipped with a metric ρ and a tolerance parameter $\epsilon \geq 0$, we let $\log N(\epsilon, \mathbb{C}, \rho)$ denote the ϵ-metric entropy of the class $\mathbb{C}$ in the metric ρ.

With this notation, the truncated form of Dudley's entropy integral inequality[4] yields

$$\mathbb{E}[Z(\delta)] \leq \frac{c}{p_{\text{obs}}} \Big\{ d^{-8} + \int_{\frac{1}{2}d^{-9}}^{2\sqrt{nd}} \sqrt{\log N(\epsilon, \mathbb{V}_{\text{DIFF}}, |\!|\!|.|\!|\!|_{\text{F}})} (\Delta\epsilon) \Big\}. \tag{4.33}$$

The upper limit of $2\sqrt{nd}$ in the integration is due to the fact $|\!|\!|D|\!|\!|_{\text{F}} \leq 2\sqrt{nd}$ for every $D \in \mathbb{V}_{\text{DIFF}}$.

It is known [221] that the metric entropy of the set $\mathbb{V}_{\text{DIFF}}$ is upper bounded as

$$\log N(\epsilon, \mathbb{V}_{\text{DIFF}}, |\!|\!|.|\!|\!|_{\text{F}}) \leq 8 \frac{\max\{n,d\}^2}{\epsilon^2} \big(\log \frac{\max\{n,d\}}{\epsilon} \big)^2 \qquad \text{for each } \epsilon > 0.$$

Combining this upper bound with the Dudley entropy integral (4.33), and observing that the integration has $\epsilon \geq \frac{1}{2}d^{-9}$, the claimed upper bound (4.31a) follows.

Bounding the tail probability of $Z(\delta)$: In order to establish the claimed tail bound (4.31b), we use a Bernstein-type bound on the supremum of empirical processes due to Klein and Rio [129, Theorem 1.1c]. In particular, this result applies to a random variable of the form $X^\dagger = \sup_{v \in \mathcal{V}} \langle X, \, v \rangle$, where $X = (X_1, \ldots, X_m)$ is a vector of independent random variables taking values in $[-1, 1]$, and $\mathcal{V}$ is some subset of $[-1, 1]^m$. Their theorem guarantees that for any $t > 0$,

$$\mathbb{P}\big(X^\dagger > \mathbb{E}[X^\dagger] + t \big) \leq \exp\left(\frac{-t^2}{2 \sup\limits_{v \in \mathcal{V}} \mathbb{E}[\langle v, \, X \rangle^2] + 4\mathbb{E}[X^\dagger] + 3t} \right). \tag{4.34}$$

In our setting, we apply this tail bound with the choices

$$X = \frac{1}{2} W, \quad \text{and} \quad X^\dagger = \frac{1}{2} \sup_{\substack{D \in \mathbb{V}_{\text{DIFF}}, \\ |\!|\!|D|\!|\!|_{\text{F}} \leq \delta}} \langle\!\langle D, \, W \rangle\!\rangle = \frac{1}{2} p_{\text{obs}} Z(\delta).$$

The entries of the matrix W are independently distributed with a mean of zero and a variance of at most $4p_{\text{obs}}$, and are bounded in absolute value by 2. As a result, we have

[4] Here we use $(\Delta\epsilon)$ to denote the differential of ϵ, so as to avoid confusion with the number of questions d.

$\mathbb{E}[\langle\!\langle D,\ W\rangle\!\rangle^2] \leq 4p_{\text{obs}}|\!|\!|D|\!|\!|_{\text{F}}^2 \leq 4p_{\text{obs}}\delta^2$ for every $D \in \mathbb{V}_{\text{DIFF}}$. With these assignments, inequality (4.34) guarantees that

$$\mathbb{P}\big(p_{\text{obs}}Z(\delta) > p_{\text{obs}}\mathbb{E}[Z(\delta)] + p_{\text{obs}}u\big) \leq \exp\Big(\frac{-(up_{\text{obs}})^2}{2p_{\text{obs}}\delta^2 + 2p_{\text{obs}}\mathbb{E}[Z(\delta)] + 3up_{\text{obs}}}\Big),$$

for all $u > 0$, and some algebraic simplifications yield the claimed result.

4.6.2 Proof of Theorem 10(b): Minimax lower bound

We now turn to the proof of the minimax lower bound. For a numerical constant $\delta \in (0, \frac{1}{4})$ whose precise value is determined later, define the probability matrix $Q^* \in [0,1]^{n\times d}$ with entries

$$Q^*_{ij} = \begin{cases} \frac{1}{2} + \delta & \text{if } i \leq \frac{1}{p_{\text{obs}}} \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{4.35}$$

One may assume that the matrix Q^* is known to any estimator under consideration.

The Gilbert-Varshamov bound [90, 258] guarantees that for a universal constant $c > 0$, there is a collection $\eta = \exp(cd)$ binary vectors—that is, a collection of vectors $\{x^1, \ldots, x^\eta\}$ all belonging to the Boolean hypercube $\{-1,1\}^d$—such that the normalized Hamming distance (4.1) between any pair of vectors in this set is lower bounded as

$$D_{\text{H}}(x^\ell, x^{\ell'}) \geq \frac{1}{10}, \qquad \text{for every } \ell, \ell' \in [\eta].$$

For each $\ell \in [\eta]$, let $\mathbb{P}^\ell$ denote the probability distribution of Y induced by setting $x^* = x^\ell$. For the choice of Q^* specified in (4.35), following some algebra, we obtain a upper bound on the Kullback-Leibler divergence between any pair of distributions from this collection as

$$D_{\text{KL}}(\mathbb{P}^\ell \| \mathbb{P}^{\ell'}) \leq c'd\delta^2 \qquad \text{for every } \ell \neq \ell' \in [\eta],$$

for another constant $c' > 0$. Combining the above observations with Fano's inequality [54] yields that any estimator $\widehat{x}$ has expected normalized Hamming error lower bounded as

$$\mathbb{E}[D_{\text{H}}(\widehat{x}, x^*)] \geq \frac{1}{20}\Big(1 - \frac{c'd\delta^2 + \log 2}{\log \eta}\Big).$$

Consequently, for the choice of Q^* given by (4.35), the Q^*-loss is lower bounded as

$$\mathbb{E}[\mathcal{L}_{Q^*}(\widehat{x}, x^*)] = \frac{4\delta^2}{p_{\text{obs}}}\frac{\mathbb{E}[D_{\text{H}}(\widehat{x}, x^*)]}{n} \geq \frac{4\delta^2}{20np_{\text{obs}}}\Big(1 - \frac{c'd\delta^2 + \log 2}{cd}\Big) \overset{(i)}{\geq} \frac{c''}{np_{\text{obs}}},$$

for some constant $c'' > 0$ as claimed. Here inequality (i) follows by setting δ to be a sufficiently small positive constant (depending on the values of c' and c'').

4.6.3 Proof of Corollary 2(a): Upper bound for estimating Q^*

In the proof of Theorem 10(a), we showed that there is a constant $c_1 > 0$ such that

$$|\!|\!|(2Q^* - 11^T)x^* - (2\widetilde{Q}_{\rm LS} - 11^T)\widetilde{x}_{\rm LS}|\!|\!|_{\rm F}^2 \leq c_1 \frac{d}{p_{\rm obs}} \log^2 d,$$

with probability at least $1 - e^{-d\log(dn)}$. Since all entries of the matrices $2Q^* - 11^T$ and $2\widetilde{Q}_{\rm LS} - 11^T$ are non-negative, and since every entry of the vectors x^* and $\widetilde{x}_{\rm LS}$ lies in $\{-1, 1\}$, some algebra yields the bound

$$\big((2Q^*_{ij} - 1) - (2[\widetilde{Q}_{\rm LS}]_{ij} - 1)\big)^2 \leq \big((2Q^*_{ij} - 1)x^*_j - (2[\widetilde{Q}_{\rm LS}]_{ij} - 1)[\widetilde{x}_{\rm LS}]_j\big)^2,$$

for every $i \in [n],\ j \in [d]$. Combining these inequalities yields the claimed bound.

4.6.4 Proof of Corollary 2(b): Lower bound for estimating Q^*

We begin by constructing a set, of cardinality η, of possible matrices Q^*, for some integer $\eta > 1$, and subsequently we show that it is hard to identify the true matrix if drawn from this set. We begin by defining a η-sized collection of vectors $\{h^1, \ldots, h^\eta\}$, all contained in the set $[\frac{1}{2}, 1]^d$, as follows. The Gilbert-Varshamov bound [90, 258] guarantees a constant $c \in (0, 1)$ such that there exists set of $\eta = \exp(cd)$ vectors, $v^1, \ldots, v^\eta \in \{-1, 1\}^d$ with the property that the normalized Hamming distance (4.1) between any pair of these vectors is lower bounded as

$$D_{\rm H}(v^\ell, v^{\ell'}) \geq \frac{1}{10}, \qquad \text{for every } \ell, \ell' \in [\eta].$$

Fixing some $\delta \in (0, \frac{1}{4})$, let us define, for each $\ell \in [\eta]$, the vector $h^\ell \in \mathbb{R}^d$ with entries

$$[h^\ell]_j := \begin{cases} \frac{1}{2} + \delta & \text{if } [v^\ell]_j = 1 \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

For each $\ell \in [\eta]$, define the matrix $Q^\ell = 1(h^\ell)^T$, and let $\mathbb{P}^\ell$ denote the probability distribution of the observed data Y induced by setting $Q^* = Q^\ell$ and $x^* = 1$. Since the entries of Y are all independent, some algebra leads to the following upper bound on the Kullback-Leibler divergence between any pair of distributions from this collection:

$$D_{\rm KL}(\mathbb{P}^\ell \| \mathbb{P}^{\ell'}) \leq 4 p_{\rm obs} n d \delta^2 \qquad \text{for every } \ell \neq \ell' \in [\eta].$$

Moreover, some simple calculation shows that the squared Frobenius norm distance between any two matrices in this collection is lower bounded as

$$|\!|\!|Q^\ell - Q^{\ell'}|\!|\!|_{\rm F}^2 \geq \frac{1}{10} d n \delta^2 \qquad \text{for every } \ell \neq \ell' \in [\eta].$$

Combining the above observations with Fano's inequality [54] yields that any estimator $\widehat{Q}$ for Q^* has mean squared error lower bounded as

$$\mathbb{E}[\|\!|Q^* - \widehat{Q}\|\!|_{\mathrm{F}}^2] \geq \frac{1}{20} dn\delta^2 \Big(1 - \frac{4p_{\mathrm{obs}} dn\delta^2 + \log 2}{\log \eta}\Big) \geq c' \frac{d}{p_{\mathrm{obs}}},$$

where we have set $\delta^2 = \frac{c''}{p_{\mathrm{obs}} n}$ for a small enough positive constant c'', where c' is another positive constant whose value may depend only on c and c''.

4.6.5 Proof of Theorem 11: When ordering of workers is approximately known

We begin by stating a key auxiliary lemma, which is somewhat more general than what is required for the current proof. For any matrix $Q^* \in \mathbb{C}_{\mathrm{Perm}}$ and worker permutation π, we define the set

$$J := \Big\{ j \in [d] \mid \exists k_j \geq \frac{\log^{1.5}(dn)}{p_{\mathrm{obs}}} \;\; \text{s.t.} \;\; \sum_{i=1}^{k_j} (Q^*_{\pi^{-1}(i)j} - \frac{1}{2}) \geq \frac{3}{4} \sqrt{\frac{k_j}{p_{\mathrm{obs}}} \log^{1.5}(dn)} \Big\}. \tag{4.36}$$

Note that this set corresponds to a subset of questions that are relatively "easy", in a certain sense specified by Q^*.

Lemma 22. *For the set J, the WAN estimator satisfies the bound*

$$\mathbb{P}\Big([\widehat{x}_{\textit{WAN}}(\pi)]_{j_0} = x^*_{j_0} \quad \textit{for all } j_0 \in J \Big) \geq 1 - e^{-c \log^{1.5}(dn)}.$$

See the end of this section for the proof of this claim.

Lemma 22 guarantees that the WAN estimator correctly answers all questions that are relatively easy. Note that the set (4.36) is defined in terms of the ℓ_1-norm of subvectors of columns of $Q^* - \frac{1}{2}$, whereas the conditions

$$\|Q^*_j - \frac{1}{2}\|_2^2 \geq \frac{5 \log^{2.5}(dn)}{p_{\mathrm{obs}}} \quad \text{and} \quad \|Q^\pi_j - Q^{\pi^*}_j\|_2 \leq \frac{\|Q^*_j - \frac{1}{2}\|_2}{\sqrt{9 \log(dn)}} \tag{4.37}$$

in the theorem claim are in terms of the ℓ_2-norm of the columns of Q^*. The following lemma allows us to connect the ℓ_1 and ℓ_2-norm constraints for any vector in a general class.

Lemma 23. *For any vector $v \in [0,1]^n$ such that $v_1 \geq \ldots \geq v_n$, there must be some $\alpha \geq \lceil \frac{1}{2} \|v\|_2^2 \rceil$ such that*

$$\sum_{i=1}^{\alpha} v_i \geq \sqrt{\frac{\alpha \|v\|_2^2}{2 \log n}}. \tag{4.38}$$

See the end of this section for the proof of this claim.

Using these two lemmas, we can now complete the proof of the theorem. We may assume without loss of generality that the rows of Q^* are ordered to be non-decreasing downwards along any column, that is, that π^* is the identity permutation. Consider any question $j \in [d]$ for which the permutation π satisfies the bounds (4.37). For any $\ell \in [n]$, let $g_\ell \in \mathbb{R}^n$ denote a vector with ones in its first ℓ positions and zeros elsewhere. The Cauchy-Schwarz inequality implies that $(Q^\pi_j - \frac{1}{2})^T g_\ell \geq (Q^*_j - \frac{1}{2})^T g_\ell - \sqrt{\ell}\|Q^\pi_j - Q^*_j\|_2$. By applying Lemma 23 to the vector $Q^*_j - \frac{1}{2}$, we are guaranteed the existence of some value $k \geq \frac{5\log^{2.5}(dn)}{2p_{\text{obs}}}$ such that $(Q^*_j - \frac{1}{2})^T g_k \geq \|Q^*_j - \frac{1}{2}\|_2 \sqrt{\frac{k}{2\log n}}$. Consequently, we have the lower bound

$$\begin{aligned}(Q^\pi_j - \frac{1}{2})^T g_k \;&\geq\; \|Q^*_j - \frac{1}{2}\|_2 \sqrt{\frac{k}{2\log n}} - \sqrt{k}\|Q^\pi_j - Q^*_j\|_2 \\ &\overset{(i)}{\geq}\; .37\|Q^*_j - \frac{1}{2}\|_2 \sqrt{\frac{k}{\log(dn)}} \;\overset{(ii)}{\geq}\; \frac{3}{4}\sqrt{\frac{k}{p_{\text{obs}}}\log^{1.5}(dn)},\end{aligned}$$

where inequalities (i) and (ii) follow from conditions (4.37). Consequently, we can apply Lemma 22 for every such question j, thereby yielding the claimed result.

Proof of Lemma 22

Observe that the windowing step of the WAN estimator identifies a group of k_{WAN} workers such that their aggregate responses towards questions are biased (towards either answer $\{-1, 1\}$) by at least $\sqrt{k_{\text{WAN}} p_{\text{obs}} \log^{1.5}(dn)}$. We first derive three properties associated with having such a bias. These properties involve function $\gamma_\pi : [n] \times [d] \times \{-1, 1\} \to \mathbb{R}$, where $\gamma_\pi(k, j, x)$ represents the amount of bias in the responses of the top $k \in [n]$ workers for question $j \in [d]$ towards the answer $x \in \{-1, 1\}$:

$$\gamma_\pi(k, j, x) := \sum_{i=1}^{k} (\mathbf{1}\{Y_{\pi^{-1}(i)j} = x\} - \mathbf{1}\{Y_{\pi^{-1}(i)j} = -x\}) = x \sum_{i=1}^{k} Y_{\pi^{-1}(i)j}.$$

A straightforward application of the Bernstein inequality [15], using the fact that the entries of the observed matrix Y are all independent, with moments bounded as

$$\mathbb{E}[Y_{ij}] = 2p_{\text{obs}}(Q^*_{ij} - \frac{1}{2})x^*_j, \quad \text{and} \quad \mathbb{E}[Y_{ij}^2] = p_{\text{obs}},$$

ensures that all three properties stated below are satisfied with probability at least $1 - e^{c\log^{1.5}(dn)}$ for every question $j \in [d]$ and every $k \in \{p_{\text{obs}}^{-1}\log^{1.5}(dn), \ldots, n\}$. For the remainder of the proof we work conditioned on the event where the following properties hold:

(P1) Sufficient condition for bias towards correct answer: If $\sum_{i=1}^{k}(Q^*_{\pi^{-1}(i)j} - \frac{1}{2}) \geq \frac{3}{4}\sqrt{\frac{k\log^{1.5}(dn)}{p_{\text{obs}}}}$, then $\gamma_\pi(k, j, x^*_j) \geq \sqrt{k p_{\text{obs}} \log^{1.5}(dn)}$.

(P2) Necessary condition for bias towards any answer $x \in \{-1, 1\}$: $\gamma_\pi(k, j, x) \geq \sqrt{k p_{\text{obs}} \log^{1.5}(dn)}$ only if $x = x^*_j$ and $\sum_{i=1}^{k}(Q^*_{\pi^{-1}(i)j} - \frac{1}{2}) \geq \frac{1}{4}\sqrt{\frac{k \log^{1.5}(dn)}{p_{\text{obs}}}}$.

(P3) Sufficient condition for aggregate to be correct: If $\sum_{i=1}^{k}(Q^*_{\pi^{-1}(i)j} - \frac{1}{2}) \geq \frac{1}{4}\sqrt{\frac{k \log^{1.5}(dn)}{p_{\text{obs}}}}$, then $\gamma_\pi(k, j, x^*_j) > 0$.

We now show that when these three properties hold, for any question $j_0 \in J$, we must have that $[\widehat{x}_{\text{WAN}}(\pi)]_{j_0} = x^*_{j_0}$. In particular, we do so by exihibiting a question that is at least as hard as j_0 on which the WAN estimator is definitely correct, and use the above properties to conclude that it therefore must also be correct on the question j_0.

Recall that by the definition (4.36) of J, for any question $j_0 \in J$, it must be the case that there exists a $k_{j_0} \geq p_{\text{obs}}^{-1} \log^{1.5}(dn)$ such that

$$\sum_{i=1}^{k_{j_0}}(Q^*_{\pi^{-1}(i)j} - \frac{1}{2}) \geq \frac{3}{4}\sqrt{\frac{k_{j_0}}{p_{\text{obs}}} \log^{1.5}(dn)}. \tag{4.39}$$

We define an associated set J_0 as the set of questions that are at least as easy as question j_0 according to the underlying permutation σ^*, that is,

$$J_0 := \{j \in [d] \mid \sigma^*(j) \leq \sigma^*(j_0)\}.$$

By the monotonicity of the columns of Q^*, every question in J_0 also satisfies condition (4.39). For each positive integer k, define the set

$$J(k) := \Big\{j \in [d] \Big| \gamma_\pi(k, j, x) \geq \sqrt{k p_{\text{obs}} \log^{1.5}(dn)} \quad \text{for some} \quad x \in \{-1, 1\}\Big\}.$$

Property (P1) ensures that every question in the set J_0 is also in the set $J(k_{j_0})$. We then have

$$|J(k_{\text{WAN}})| \overset{(i)}{\geq} |J(k_{j_0})| \geq |J_0|,$$

where step (i) uses the optimality of k_{WAN} for the optimization problem in equation (4.9a). Given this, there are two possibilities: either (1) we have the equality $J(k_{\text{WAN}}) = J_0$, or (2) the set $J(k_{\text{WAN}})$ contains some question not in the set J_0. We address each of these possibilities in turn.

Case 1: It suffices to observe by Properties (P1)–(P3), that the aggregate of the top k_{WAN} workers is correct on every question in the set $J(k_{\text{WAN}})$ and this implies that it must be the case that $[\widehat{x}_{\text{WAN}}(\pi)]_{j_0} = x^*_{j_0}$ as desired.

Case 2: In this case, there is some question $j' \notin J_0$ such that $\gamma_\pi(k_{\text{WAN}}, j, x) \geq \sqrt{k_{\text{WAN}} p_{\text{obs}} \log^{1.5}(dn)}$ for some $x \in \{-1, 1\}$. Property (P2) guarantees that $\sum_{i=1}^{k_{\text{WAN}}}(Q^*_{\pi^{-1}(i)j'} - \frac{1}{2}) \geq \frac{1}{4}\sqrt{\frac{k_{\text{WAN}} \log^{1.5}(dn)}{p_{\text{obs}}}}$

and that $x = x^*_j$. Now, since every question easier than j_0 is in the set J_0, question j' must be more difficult than j_0, which implies that

$$\sum_{i=1}^{k_{\text{WAN}}} (Q^*_{\pi^{-1}(i)j_0} - \frac{1}{2}) \geq \frac{1}{4}\sqrt{\frac{k_{\text{WAN}} \log^{1.5}(dn)}{p_{\text{obs}}}}.$$

Applying Property (P3), we can then conclude that $[\widehat{x}_{\text{WAN}}(\pi)]_{j_0} = x^*_{j_0}$ as desired.

Proof of Lemma 23

We partition the proof into two cases depending on the value of $\|v\|_2^2$.
Case 1: First, suppose that $\frac{1}{2}\|v\|_2^2 \geq e$. In this case, we proceed via proof by contradiction. If the claim were false, then we would have

$$\sqrt{\frac{\alpha\|v\|_2^2}{2\log n}} > \sum_{i=1}^{\alpha} v_i \geq \alpha v_\alpha \qquad \text{for every } \alpha \geq \lceil \tfrac{1}{2}\|v\|_2^2 \rceil.$$

It would then follow that

$$\begin{aligned}\sum_{i=1}^n v_i^2 \;=\; \sum_{i=1}^{\lceil \frac{1}{2}\|v\|_2^2\rceil - 1} v_i^2 + \sum_{i=\lceil \frac{1}{2}\|v\|_2^2\rceil}^{n} v_i^2 &\overset{\text{(i)}}{\leq} \lceil \frac{1}{2}\|v\|_2^2 \rceil - 1 + \sum_{i=\lceil \frac{1}{2}\|v\|_2^2\rceil}^{n} v_i^2 \\ &< \frac{1}{2}\|v\|_2^2 \; + \sum_{i=\lceil \frac{1}{2}\|v\|_2^2\rceil}^{n} \frac{\|v\|_2^2}{2i\log n},\end{aligned}$$

where step (i) uses the fact that $v_i \in [0,1]$. Using the standard bound $\sum_{i=a}^b \frac{1}{i} \leq \log(\frac{eb}{a})$ and the assumption $\lceil \frac{1}{2}\|v\|_2^2 \rceil \geq e$, we find that

$$\frac{1}{2}\|v\|_2^2 \; + \sum_{i=\lceil \frac{1}{2}\|v\|_2^2\rceil}^{n} \frac{\|v\|_2^2}{2i\log n} \leq \|v\|_2^2.$$

The resulting chain of inequalities contradicts the definition of $\|v\|_2^2$.

Case 2: Otherwise, we may assume that $\frac{1}{2}\|v\|_2^2 < e$. Observe that the case $v = 0$ trivially satisfies the claim with $\alpha = 1$, and hence we restrict attention to non-zero vectors. Define a vector $v' \in [0,1]^n$ as

$$v' = \frac{1}{v_1} v.$$

We first prove the claim of the lemma for the vector v', that is, we prove that there exists some value $\alpha \geq \lceil \frac{1}{2} \|v'\|_2^2 \rceil$ such that

$$\sum_{i=1}^{\alpha} v'_i \geq \sqrt{\frac{\alpha \|v'\|_2^2}{2 \log n}}. \tag{4.40}$$

Observe that $1 = v'_1 \geq \cdots \geq v'_n \geq 0$. If $\frac{1}{2}\|v'\|_2^2 \geq e$, then our claim (4.40) is proved via the analysis of Case 1 above. Otherwise, we have that $\frac{1}{2}\|v'\|_2^2 \leq e$ and $v'_1 = 1$. Setting $\alpha = 1$, we obtain the inequalities

$$\sum_{i=1}^{\alpha} v'_i = 1 \quad \text{and} \quad \sqrt{\frac{\alpha \|v'\|_2^2}{2 \log n}} \leq 1,$$

where we have used the assumption that n is large enough (concretely, $n \geq 16$). We have thus proved the bound (4.40), and it remains to translate this bound on v' to an analogous bound on the vector v. Observe that since $v_1 \leq 1$, we have the relation $\|v'\|_2 \geq \|v\|_2$. Using the same value of α as that derived for vector v', we then obtain from (4.40) that this value $\alpha \geq \lceil \frac{1}{2}\|v'\|_2^2 \rceil \geq \lceil \frac{1}{2}\|v\|_2^2 \rceil$ satisfies

$$v_1 \sum_{i=1}^{\alpha} v'_i \geq v_1 \sqrt{\frac{\alpha \|v'\|_2^2}{2 \log n}},$$

which establishes the claim.

4.6.6 Proof of Theorem 12 (a): OBI-WAN under intermediate class

Define the vector $r^* := \widetilde{q} - \frac{1}{2}$. We split the proof into two cases, depending on whether or not the condition

$$\|r^*\|_2 \|1 - h^*\|_2 \geq \sqrt{\frac{C d \log^{2.5}(dn)}{p_{\text{obs}}}} \tag{4.41}$$

is satisfied. Here $C > 20$ is a constant, whose value is specified later in the proof. (In particular, see equation (4.48) in Lemma 24.)

Case 1

First, suppose that condition (4.41) is *violated*. For each $\widehat{x} \in \{-1, 1\}^d$, we then have

$$\mathcal{L}_{Q^*}(\widehat{x}, x^*) \leq \frac{1}{dn} \|r^*\|_2^2 \|1 - h^*\|_2^2 \leq \frac{6C \log^{2.5} d}{n p_{\text{obs}}},$$

as claimed, where we have made use of the fact that $d \geq n$.

Case 2

In this second case, we may assume that condition (4.41) holds, and we do so throughout the remainder of this section. Our proof of this case is divided into three parts, each corresponding to one of the three steps in the OBI-WAN algorithm. The first step is to derive certain properties of the split of the questions. The second step is to derive approximation-guarantees on the outcome of the OBI step. The third and final step is to show that this approximation guarantee ensures that the output of the WAN estimator meets the claimed error guarantee.

Step 1: Analyzing the split Our first step is to exhibit a useful property of the split of the questions—namely, that with high probability, the questions in the two sets T_0 and T_1 have a similar total difficulty.

The random sets (T_0, T_1) chosen in the first step can be obtained as follows: first generate an i.i.d. sequence $\{\epsilon_j\}_{j=1}^d$ of equiprobable $\{0,1\}$ variables, and then set $T_\ell := \{j \in [d] \mid \epsilon_j = \ell\}$ for $\ell \in \{0,1\}$. Note that we have $\mathbb{E}[\sum_{j\in[d]}(1-h_j^*)^2\epsilon_j] = \frac{1}{2}\|1-h^*\|_2^2$, and $\mathbb{E}[\sum_{j\in[d]}((1-h_j^*)^2\epsilon_j)^2] = \frac{1}{2}\sum_{j\in[d]}(1-h_j^*)^4 \leq \frac{1}{2}\|1-h^*\|_2^2$. Applying Bernstein's inequality then guarantees that

$$\mathbb{P}\big(\sum_{j\in T_\ell}(1-h_j^*)^2 > \frac{2}{3}\|1-h^*\|_2^2\big) \leq \exp\big(-c\|1-h^*\|_2^2\big) \qquad \text{for each } \ell \in \{0,1\},$$

where c is a positive universal constant. We are thus guaranteed that

$$\frac{1}{3}\|1-h^*\|_2^2 \leq \sum_{j\in T_\ell}(1-h_j^*)^2 \leq \frac{2}{3}\|1-h^*\|_2^2 \qquad \text{for both } \ell \in \{1,2\}, \tag{4.42}$$

with probability at least $1 - e^{-cC\frac{\log^{2.5} d}{p_{\text{obs}}}}$, where we have used the fact that $\|1-h^*\|_2^2 \geq \frac{Cd\log^{2.5} d}{p_{\text{obs}}\|r^*\|_2^2} \geq \frac{C\log^{2.5} d}{p_{\text{obs}}}$. Now define the error event

$$\mathcal{E} := \Big\{\mathcal{L}_{Q^*}(\widehat{x}_{\text{OBI-WAN}}, x^*) > \frac{6C\log^{2.5} d}{np_{\text{obs}}}\Big\}.$$

Combining the sandwich relation (4.42) with the union bound, we find that

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \sum_{\text{partitions } \widetilde{T}_0,\widetilde{T}_1} \mathbb{P}(\mathcal{E} \mid T_0 = \widetilde{T}_0, T_1 = \widetilde{T}_1)\mathbb{P}(T_0 = \widetilde{T}_0, T_1 = \widetilde{T}_1)\\ &\leq \sum_{\substack{\text{partitions } \widetilde{T}_0,\widetilde{T}_1\\ \text{satisfying (4.42)}}} \mathbb{P}(\mathcal{E} \mid T_0 = \widetilde{T}_0, T_1 = \widetilde{T}_1)\mathbb{P}(T_0 = \widetilde{T}_0, T_1 = \widetilde{T}_1) + e^{-cC\frac{\log^{2.5} d}{p_{\text{obs}}}}.\end{aligned}$$

Consequently, in the rest of the proof we consider any partition $(\widetilde{T}_0, \widetilde{T}_1)$ that satisfies the sandwich bound (4.42) and derive an upper bound on the error conditioned on this partition. In other words, it suffices to prove the following bound for any partition $(\widetilde{T}_0, \widetilde{T}_1)$ satisfying (4.42):

$$\mathbb{P}(\mathcal{E} \mid T_0 = \widetilde{T}_0, T_1 = \widetilde{T}_1) \leq e^{-c' \log^{1.5}(dn)}, \tag{4.43}$$

for some positive universal constant c' whose value may depend only on C. We note that conditioned on the partition $(\widetilde{T}_0, \widetilde{T}_1)$, and for any fixed values of Q^* and x^*, the responses of the workers to the questions in one set are statistically independent of the responses in the other set. Consequently, we describe the proof for any one of the two partitions, and the overall result is implied by a union bound of the error guarantees for the two partitions. We use the notation ℓ to denote either one of the two partitions in the sequel, that is, $\ell \in \{0, 1\}$.

Step 2: Guarantees for the OBI step Assume without loss of generality that the rows of the matrix Q^* are ordered according to the abilities of the corresponding workers, that is, the entries of $\widetilde{q}$ are arranged in a non-increasing order. Recall that π_ℓ denotes the permutation of the workers in order of their respective values in u_ℓ. Let $\widetilde{r}_\ell \in \mathbb{R}^n$ denote the vector obtained by permuting the entries of r^* in the order given by π_ℓ. Thus the entries of $\widetilde{r}_\ell$ are identical to those of r^* up to a permutation; the ordering of the entries of $\widetilde{r}_\ell$ is identical to the ordering of the entries of u_ℓ. The following lemma—central for the proof of this theorem—establishes a deterministic relation between these vectors. The proof of this lemma combines matrix perturbation theory with some careful algebraic arguments.

Lemma 24. *Suppose that condition* (4.41) *holds for a sufficiently large constant* $C > 0$*. Then for any split* (T_0, T_1) *satisfying the relation* (4.42)*, we have*

$$\mathbb{P}\Big(\|\widetilde{r}_\ell - r^*\|_2^2 > \frac{\|r^*\|_2^2}{9\log(dn)}\Big) \leq e^{-c \log^{1.5} d}. \tag{4.44}$$

See the end of this section for the proof of this claim.

At this point, we are now ready to apply the bound for the WAN estimator from Theorem 11.

Step 3: Guarantees for the WAN step Recall that for any choice of index $\ell \in \{0, 1\}$, the OBI step operates on the set T_ℓ of questions, and the WAN step operates on the alternate set $T_{1-\ell}$. Consequently, conditioned on the partition $(\widetilde{T}_0, \widetilde{T}_1)$, the outcomes $Y_{1-\ell}$ of the comparisons in set $(1-\ell)$ are statistically independent of the permutation π_ℓ obtained from set ℓ in the OBI step.

Consider any question $j \in T_{1-\ell}$ that satisfies the inequality $\|(1-h_j^*)r^*\|_2^2 \geq \frac{5\log^{2.5}(dn)}{p_{\text{obs}}}$. We now claim that this question j satisfies the pair of conditions (4.10a) required by the statement of Theorem 11. First observe that $(1-h_j^*)r^*$ is simply the j^{th} column of the matrix $(Q^* - \frac{1}{2})$, we have $\|Q_j^* - \frac{1}{2}\|_2^2 \geq \frac{5\log^{2.5}(dn)}{p_{\text{obs}}}$. The first condition in (4.10a) is thus satisfied.

In order to establish the second condition, observe that a rescaling of the inequality (4.44) by the non-negative scalar $(1-h^*_j)$ yields the bound

$$\|(1-h^*_j)\widetilde{r}_\ell - (1-h^*_j)r^*\|_2^2 \leq \frac{\|(1-h^*_j)r^*\|_2^2}{9\log(dn)} \qquad \text{for every } j \in T_{1-\ell}. \tag{4.45}$$

Recall our notational assumption that the entries of $\widetilde{q}$ (and hence the rows of Q^*) are arranged in order of the workers' abilities, and that Q^π is a matrix obtained by permuting the rows of Q^* according to a given permutation π. Also observe that the vector $(1-h^*_j)\widetilde{r}_\ell$ equals the j^{th} column of $(Q^{\pi_\ell} - \frac{1}{2})$, where π_ℓ is the permutation of the workers obtained from the OBI step. Consequently, the approximation guarantee (4.45) implies that $\|Q^{\pi_\ell}_j - Q^*_j\|_2 \leq \frac{\|Q^*_j\|_2}{\sqrt{9\log(dn)}}$. Thus the second condition in equation (4.10a) is also satisfied for the question j under consideration.

Applying the result of Theorem 11 for the WAN step, we obtain that this question j is decoded correctly with a probability at least $1 - e^{-c\log^{1.5}(dn)}$, for some positive constant c. Since this argument holds for every question j satisfying $\|(1-h^*_j)r^*\|_2^2 \geq \frac{5\log^{2.5}(dn)}{p_{\text{obs}}}$, the total contribution from the remaining questions to the Q^*-loss is at most $\frac{5\log^{2.5}(dn)}{p_{\text{obs}}n}$. A union bound over all questions and both values of $\ell \in \{0,1\}$ then yields the claim that the aggregate Q^*-loss is at most $\frac{5\log^{2.5}(dn)}{p_{\text{obs}}n}$ with probability at least $1 - e^{-c'\log^{1.5}(dn)}$, for some positive constant c', as claimed in (4.43).

Proof of Lemma 24

The proof of this lemma consists of three main steps:

(i) First, we show that u_ℓ is a good approximation for the vector of worker abilities r^* up to a global sign.

(ii) We then show that the global sign is correctly identified with high probability.

(iii) The final step in the proof is to convert this guarantee to one on the permutation induced by u_ℓ.

Step 1 We first show that the vector u_ℓ approximates r^* up to a global sign. Assume without loss of generality that $x^*_j = 1$ for every question $j \in [d]$. As in the proof of Theorem 10(a), we begin by rewriting the model in a "linearized" fashion which is convenient for our analysis. Let Q^*_0 and Q^*_1 denote the submatrices of Q^* obtained by splitting its columns according to the sets T_0 and T_1. Then we have for $\ell \in \{0,1\}$,

$$\frac{1}{p_{\text{obs}}}Y_\ell = (2Q^*_\ell - 11^T)\,\text{diag}(x^*) + \frac{1}{p_{\text{obs}}}W_\ell, \tag{4.46}$$

where conditioned on T_0 and T_1, the noise matrices $W_0, W_1 \in \mathbb{R}^{n\times d}$ have entries independently drawn from the distribution (4.20). One can verify that the entries of W_0 and W_1 have a mean of zero, second moment upper bounded by $4p_{\text{obs}}$, and their absolute values are upper bounded by 2.

We now require a standard result on the perturbation of eigenvectors of symmetric matrices [246]. Consider a symmetric and positive semidefinite matrix $M \in \mathbb{R}^{d\times d}$, a second symmetric matrix $\Delta M \in \mathbb{R}^{d\times d}$, and let $\widetilde{M} = M + \Delta M$. Let $v \in \mathbb{R}^d$ be an eigenvector associated to the largest eigenvalue of M. Likewise define $\widetilde{v} \in \mathbb{R}^d$ as an eigenvector associated to the largest eigenvalue of $\widetilde{M}$. Then we are guaranteed [246] that

$$\min\{\|\widetilde{v} - v\|_2, \|\widetilde{v} + v\|_2\} \leq \frac{2\|\!|\Delta M|\!\|_{\text{op}}}{\max\{\lambda_1(M) - \lambda_2(M) - 2\|\!|\Delta M|\!\|_{\text{op}}, 0\}}, \tag{4.47}$$

where $\lambda_1(M)$ and $\lambda_2(M)$ denote the largest and second largest eigenvalues of M, respectively.

In order to apply the bound (4.47), we define the matrix $R^*_\ell := Q^*_\ell - \frac{1}{2}11^T$, as well as the matrices

$$\widetilde{M} := \frac{1}{p_{\text{obs}}^2} Y_\ell Y_\ell^T, \quad M = 4R^*_\ell (R^*_\ell)^T, \quad \text{and}$$
$$\Delta M := \frac{2}{p_{\text{obs}}} W_\ell (R^*_\ell)^T + \frac{2}{p_{\text{obs}}} R^*_\ell W_\ell^T + \frac{1}{p_{\text{obs}}^2} W_\ell W_\ell^T.$$

Using our linearized observation model (4.46), it is straightforward to verify that these choices satisfy the condition $\widetilde{M} = M + \Delta M$, so that the bound (4.47) can be applied.

Recall that for any matrix $Q^* \in \mathbb{C}_{\text{Int}}$, we have $Q^* = \widetilde{q}(1-h^*)^T + \frac{1}{2}(h^*)^T$ for some vectors $\widetilde{q} \in [\frac{1}{2}, 1]^n$ and $h^* \in [0,1]^d$. Also recall our definition of the associated quantity $r^* \in [0, \frac{1}{2}]^n$ as $r^* = \widetilde{q} - \frac{1}{2}$. We denote the magnitude of the vector r^* as $\rho := \|r^*\|_2$.

With the notation introduced above, we are ready to apply the bound (4.47). First observe that the matrix R^*_ℓ has a rank of one, and consequently $\|\!|R^*_\ell|\!\|_{\text{op}} = \rho\sqrt{\sum_{j\in T_\ell}(1-h^*_j)^2}$. Conditioned on the bound (4.42), we obtain

$$\sqrt{\frac{1}{3}}\rho\|1-h^*\|_2 \leq \|\!|R^*_\ell|\!\|_{\text{op}} \leq \sqrt{\frac{2}{3}}\rho\|1-h^*\|_2.$$

Moreover, the entries of the matrix W_ℓ are independent, zero-mean, and have a second moment upper bounded by $4p_{\text{obs}}$. Consequently, known results on random matrices [12, Remark 3.13] guarantee that

$$\|\!|W_\ell|\!\|_{\text{op}} \leq c\sqrt{\max\{d,n\}p_{\text{obs}}\log^{1.5} d} \leq c\sqrt{dp_{\text{obs}}\log^{1.5} d},$$

with probability at least $1 - e^{-c\log^{1.5} d}$, where we have used the fact that $d \geq n$ and $p_{\text{obs}} \geq \frac{1}{n}$. These inequalities, in turn, imply that the top eigenvalue of M is lower bounded as $\lambda_1(M) =$

$|||R^*|||_{\text{op}}^2 \geq \frac{1}{3}\rho^2\|1 - h^*\|_2^2$, the second eigenvalue vanishes (that is, $\lambda_2(M) = 0$), and moreover that

$$\begin{aligned} |||\Delta M|||_{\text{op}} &\leq \frac{2}{p_{\text{obs}}}|||R^*|||_{\text{op}}|||W|||_{\text{op}} + \frac{1}{p_{\text{obs}}^2}|||W|||_{\text{op}}^2 \\ &\leq c' \frac{\sqrt{d \log^{1.5} d}}{p_{\text{obs}}}(\rho\|1 - h^*\|_2\sqrt{p_{\text{obs}}} + \sqrt{d \log^{1.5} d}). \end{aligned}$$

Recall the lower bound $\rho\|1 - h^*\|_2 \geq \sqrt{\frac{Cd \log^{2.5} d}{p_{\text{obs}}}}$, assumed in the statement of the lemma. Using these facts and doing some algebra, we find that with probability at least $1 - e^{-c \log^{1.5} d}$, for any pair of sets T_0 and T_1 satisfying (4.42), we have the bound

$$\min\{\|u_\ell - \frac{1}{\rho}r^*\|_2^2, \|u_\ell + \frac{1}{\rho}r^*\|_2^2\} \leq \frac{1}{36}\frac{1}{\rho^2\|1 - h_j^*\|_2^2}\frac{d \log^{1.5} d}{p_{\text{obs}}}, \tag{4.48}$$

where the prefactor $\frac{1}{36}$ is obtained by setting the constant $C > 20$ to a large enough value.

Step 2 We now verify that the global sign is correctly identified. Recall our selection

$$\sum_{j=1}^n [u_\ell]_j^2 \mathbf{1}\{[u_\ell]_j > 0\} \geq \sum_{j=1}^n [u_\ell]_j^2 \mathbf{1}\{[u_\ell]_j < 0\}.$$

Since every entry of the vector r^* is non-negative, we have the inequality

$$\|u_\ell + \frac{1}{\rho}r^*\|_2^2 \geq \sum_{j=1}^n [u_\ell]_j^2 \mathbf{1}\{[u_\ell]_j > 0\} \geq \sum_{j=1}^n [u_\ell]_j^2 \mathbf{1}\{[u_\ell]_j < 0\},$$

and consequently,

$$\|u_\ell + \frac{1}{\rho}r^*\|_2^2 \geq \frac{1}{2}\|u_\ell\|_2^2. \tag{4.49a}$$

On the other hand, a version of the triangle inequality yields

$$2\|u_\ell\|_2^2 + 2\|u_\ell + \frac{1}{\rho}r^*\|_2^2 \geq \|\frac{1}{\rho}r^*\|_2^2 = 1 \tag{4.49b}$$

Now suppose that $\|u_\ell - \frac{1}{\rho}r^*\|_2^2 \geq \|u_\ell + \frac{1}{\rho}r^*\|_2^2$. Then from our earlier result (4.48), we have the bound

$$\|u_\ell + \frac{1}{\rho}r^*\|_2^2 \leq \frac{d \log^{1.5} d}{36\rho^2\|1 - h^*\|_2^2 p_{\text{obs}}}, \tag{4.49c}$$

with probability at least $1 - e^{-c\log^{1.5}(dn)}$. Putting together the inequalities (4.49a), (4.49b) and (4.49c) and rearranging some terms yields the inequality

$$\rho^2 \|1 - h^*\|_2^2 \leq \frac{d \log^{1.5} d}{9 p_{\text{obs}}}.$$

This requirement contradicts our initial assumption $\rho^2\|1 - h^*\|_2^2 \geq \frac{Cd\log^{2.5} d}{p_{\text{obs}}}$, with $C > 20$, thereby proving that $\|u_\ell - \frac{1}{\rho} r^*\|_2^2 < \|u_\ell + \frac{1}{\rho} r^*\|_2^2$. Substituting this inequality into equation (4.48) yields the bound

$$\|u_\ell - \frac{1}{\rho} r^*\|_2^2 \leq \frac{1}{36\rho^2 \|1 - h_j^*\|_2^2} \frac{d \log^{1.5} d}{p_{\text{obs}}}. \tag{4.50}$$

Step 3 The final step of this proof is to convert the approximation guarantee (4.50) on u_ℓ to an approximation guarantee on the vector $\widetilde{r}_\ell$ (which, recall, is a permutation of r^* according to the permutation induced by u_ℓ). An additional lemma is useful for this step:

Lemma 25. *For any $\ell \in \{0, 1\}$, we have $\|\widetilde{r}_\ell - r^*\|_2 \leq 2\|\rho u_\ell - r^*\|_2$.*

See the end of this section for the proof of this claim.

Combining Lemma 25 with the inequality (4.50) yields that for any choice of the set T_0 and T_1 satisfying the condition (4.42), with probability at least $1 - e^{-c\log^{1.5} d}$, we have

$$\|\widetilde{r}_\ell - r^*\|_2^2 \; \leq \; \frac{1}{18\|1 - h^*\|_2^2} \frac{d \log^{1.5} d}{p_{\text{obs}}} \; \overset{(i)}{\leq} \; \frac{\|r^*\|_2^2}{18 \log(dn)}.$$

Here, inequality (i) follows from our earlier assumption that $\|r^*\|_2 \|1 - h^*\|_2 \geq \sqrt{\frac{Cd\log^{2.5} d}{p_{\text{obs}}}}$ with $C > 20$.

Proof of Lemma 25

Recall that the two vectors $\widetilde{r}_\ell$ and r^* are identical up to a permutation. Now suppose $\widetilde{r}_\ell \neq r^*$. Then there must exist some position $i \in [n-1]$ such that $[r^*]_i < [r^*]_{i+1}$ and $[\widetilde{r}_\ell]_i \geq [\widetilde{r}_\ell]_{i+1}$. Define the vector $\widetilde{r}'$ obtained by interchanging the entries in positions i and $(i+1)$ in r^*. The difference $\Delta := \|\widetilde{r}' - \rho u_\ell\|_2^2 - \|r^* - \rho u_\ell\|_2^2$ then can be bounded as

$$\begin{aligned}
\Delta &= ([\widetilde{r}']_i - \rho[u_\ell]_i)^2 + ([\widetilde{r}']_{i+1} - \rho[u_\ell]_{i+1})^2 - ([r^*]_i - \rho[u_\ell]_i)^2 - ([r^*]_{i+1} - \rho[u_\ell]_{i+1})^2 \\
&= ([r^*]_{i+1} - \rho[u_\ell]_i)^2 + ([r^*]_i - \rho[u_\ell]_{i+1})^2 - ([r^*]_i - \rho[u_\ell]_i)^2 - ([r^*]_{i+1} - \rho[u_\ell]_{i+1})^2 \\
&= 2\rho([r^*]_{i+1} - [r^*]_i)([u_\ell]_{i+1} - [u_\ell]_i) \\
&\leq 0,
\end{aligned}$$

where the final inequality uses the fact that the ordering of the entries in the two vectors $\widetilde{r}_\ell$ and u_ℓ are identical, which in turn implies that $[u_\ell]_i \geq [u_\ell]_{i+1}$. We have thus shown an interchange of the entries i and $(i+1)$ in r^*, which brings it closer to the permutation of $\widetilde{r}_\ell$, cannot increase the distance to the vector ρu_ℓ. A recursive application of this argument leads to the inequality $\|\widetilde{r}_\ell - \rho u_\ell\|_2 \leq \|r^* - \rho u_\ell\|_2$. Applying the triangle inequality then yields

$$\|\widetilde{r}_\ell - r^*\|_2 \leq \|\widetilde{r}_\ell - \rho u_\ell\|_2 + \|\rho u_\ell - r^*\|_2 \leq 2\|\rho u_\ell - r^*\|_2,$$

as claimed.

4.6.7 Proof of Theorem 12(b): OBI-WAN under permutation-based model

First, suppose that $p_{\text{obs}} < \frac{\log^{1.5}(dn)}{n}$. In this case, we have

$$\mathcal{L}_{Q^*}(\widehat{x}_{\text{OBI-WAN}}, x^*) \leq 1 \leq \frac{1}{\sqrt{np_{\text{obs}}}} \log(dn),$$

and the claim follows immediately.

Otherwise, we may assume that $p_{\text{obs}} \geq \frac{\log^{1.5}(dn)}{n}$. For any index $\ell \in \{0,1\}$, consider an arbitrary permutation π_ℓ. Observe that conditioned on the split (T_0, T_1), the data $Y_{1-\ell}$ is independent of the choice of the permutation π_ℓ. Now consider any question $j \in T_{1-\ell}$ that satisfies

$$\sum_{i=1}^{n} (Q^*_{ij} - \frac{1}{2})^2 \geq \frac{3}{2}\sqrt{\frac{n}{p_{\text{obs}}}} \log(dn). \tag{4.51a}$$

Lemma 22 (with the associated parameter $k_j = n$) then guarantees that

$$\mathbb{P}([\widehat{x}_{\text{WAN}}(\pi)]_j \neq x^*_j) \leq e^{-c\log^{1.5}(dn)}. \tag{4.51b}$$

All remaining questions can contribute a total of at most $\frac{3}{2}\frac{1}{\sqrt{np_{\text{obs}}}} \log(dn)$ to the Q^*-loss. Consequently, a union bound over the probabilities (4.51b) for all questions (in T_0 and T_1) that satisfy the bound (4.51a) yields the claimed result.

4.6.8 Proof of Theorem 13(a): OBI-WAN under Dawid-Skene

Throughout the proof, we make use the notation previously introduced in the proof of Theorem 12(a). As in this same proof, we condition on some choice of T_0 and T_1 that satisfies (4.42). The proof of this theorem follows the same structure as the proof of Theorem 12(a) and the lemmas within it. However, we must make additional arguments in order

to account for adversarial workers. In the remainder of the proof, we consider any $\ell \in \{0, 1\}$, and then apply the union bound across both values of ℓ.
Our proof consists of the three steps:

(1) We first show that the vector u_ℓ is a good approximation to $(q^{\mathrm{DS}} - \frac{1}{2})$ up to a global sign.

(2) Second, we show that the global sign of r^* is indeed recovered correctly.

(3) Third, we establish guarantees on the performance of the WAN estimator for our setting.

We work through each of these steps in turn.

Step 1 We first show that the vector u_ℓ is a good approximation to $q^{\mathrm{DS}} - \frac{1}{2}$ up to a global sign. When $Q^* = q^{\mathrm{DS}} 1^T$, we can set the vector $h^* = 0$ in the proof of Theorem 12(a). We also have $r^* = q^{\mathrm{DS}} - \frac{1}{2}$. With these assignments, the the arguments up to equation (4.48) in Lemma 24 continue to apply even for the present setting where $q^{\mathrm{DS}} \in [0, 1]^n$. From these arguments, we obtain the following approximation guarantee (4.48) on recovering r^* up to a global sign:

$$\min\{\|u_\ell - \frac{1}{\rho} r^*\|_2^2, \|u_\ell + \frac{1}{\rho} r^*\|_2^2\} \leq \frac{1}{36} \frac{1}{\rho^2} \frac{\log^{1.5} d}{p_{\mathrm{obs}}}, \tag{4.52}$$

with probability at least $1 - e^{-c \log^{1.5} d}$.

Step 2 The next step of the proof is to show that the global sign of r^* is indeed recovered correctly. Define two pairs of vectors $\{u_\ell^+,\ u_\ell^-\}$ and $\{r_\ell^{*+},\ r_\ell^{*-}\}$, all lying in the unit cube $[0, 1]^n$, with entries

$$\begin{aligned} [u_\ell^+]_i &:= \max\{[u_\ell^+]_i, 0\} \quad \text{and} \quad [u_\ell^-]_i := \min\{[u_\ell^-]_i, 0\} \quad \text{for every } i \in [n]; \\ [r_\ell^{*+}]_i &:= \max\{[r_\ell^{*+}]_i, 0\}, \quad \text{and} \quad [r_\ell^{*-}]_i := \min\{[r_\ell^{*-}]_i, 0\} \quad \text{for every } i \in [n]. \end{aligned}$$

From the conditions assumed in the statement of the theorem, we have $\|r^{*+}\|_2 \geq \|r^{*-}\|_2 + \sqrt{\frac{4 \log^{2.5}(dn)}{p_{\mathrm{obs}}}}$, whereas from the choice of u in the OBI-WAN estimator, we have $\|u^+\|_2 \geq \|u^-\|_2$. One can also verify that

$$\|u_\ell + \frac{1}{\rho} r^*\|_2^2 \geq \|u_\ell^+ + \frac{1}{\rho} r^{*-}\|_2^2 + \|u_\ell^- + \frac{1}{\rho} r^{*+}\|_2^2. \tag{4.53a}$$

Now suppose that $\|\frac{1}{\rho} r^{*+}\|_2 \geq \|u_\ell^-\|_2 + \sqrt{\frac{\log^{2.5}(dn)}{\rho^2 p_{\mathrm{obs}}}}$. Then from the triangle inequality, we obtain the bound

$$\|u_\ell^- + \frac{1}{\rho} r^{*+}\|_2 \ \geq\ \|\frac{1}{\rho} r^{*+}\|_2 - \|u_\ell^-\|_2 \ \geq\ \sqrt{\frac{\log^{2.5}(dn)}{\rho^2 p_{\mathrm{obs}}}}. \tag{4.53b}$$

Otherwise we have that $\|\frac{1}{\rho}r^{*+}\|_2 < \|u_\ell^-\|_2 + \sqrt{\frac{\log^{2.5}(dn)}{\rho^2 p_{\text{obs}}}}$. In this case, we have

$$\begin{aligned}\|u_\ell^+ + \frac{1}{\rho}r^{*-}\|_2 \geq \|u_\ell^+\|_2 - \|\frac{1}{\rho}r^{*-}\|_2 &\geq \|u_\ell^-\|_2 - \|\frac{1}{\rho}r^{*+}\|_2 + 2\sqrt{\frac{\log^{2.5}(dn)}{\rho^2 p_{\text{obs}}}} \\ &\geq \sqrt{\frac{\log^{2.5}(dn)}{\rho^2 p_{\text{obs}}}}. \end{aligned} \tag{4.53c}$$

Putting together the conditions (4.53a), (4.53b) and (4.53c), we obtain the bound $\|u_\ell + \frac{1}{\rho}r^*\|_2^2 \geq \frac{\log^{2.5}(dn)}{\rho^2 p_{\text{obs}}}$. In conjunction with the result of equation (4.52), this bound guarantees the correct detection of the global sign, that is, $\|u_\ell - \frac{1}{\rho}r^*\|_2^2 \leq \frac{1}{36}\frac{1}{\rho^2}\frac{\log^{1.5} d}{p_{\text{obs}}}$. The deterministic inequality afforded by Lemma 25 then guarantees that

$$\|\widetilde{r}_\ell - r^*\|_2^2 \leq \frac{1}{18}\frac{\log^{1.5} d}{p_{\text{obs}}}, \tag{4.54}$$

and this completes the analysis of the OBI part of the estimator.

Step 3 In the third step, we establish guarantees on the performance of the WAN estimator for our setting. Recall that since the WAN estimator uses the permutation given by $\widetilde{r}_\ell$ and with this permutation, acts on the observation $Y_{1-\ell}$ of the other set of questions, the noise $W_{1-\ell}$ is statistically independent of the choice of $\widetilde{r}_\ell$, when conditioned on the split (T_0, T_1). Assume without loss of generality that $x^* = 1$ and that the rows of Q^* are arranged according to the worker abilities, meaning that $q_i^{\text{DS}} \geq q_{i'}^{\text{DS}}$ for every $i < i'$, or in other words, $r_i^* \geq r_{i'}^*$ for every $i < i'$. Recall our earlier notation of $g_k \in \{0,1\}^n$ denoting a vector with ones in its first k positions and zeros elsewhere.

Now from the proof of Lemma 22 the following two properties ensure that the WAN estimator decodes every question correctly with probability at least $1 - e^{-c\log^{1.5}(dn)}$: (i) There exists some value $k \geq p_{\text{obs}}^{-1}\log^{1.5}(dn)$ such that $\langle \widetilde{r}_\ell,\, g_k\rangle \geq \frac{3}{4}\sqrt{\frac{k\log^{1.5}(dn)}{p_{\text{obs}}}}$, and (ii) for every $k \in [n]$, it must be that $\langle \widetilde{r}_\ell,\, g_k\rangle > -\frac{1}{4}\sqrt{\frac{k\log^{1.5}(dn)}{p_{\text{obs}}}}$. Let us first address property (i). Lemma 23 guarantees the existence of some value $k \geq \lceil \frac{1}{2}\|r^*\|_2^2 \rceil$ such that

$$\langle r^{*+},\, g_k\rangle \geq \frac{\sqrt{k}\|r^{*+}\|_2}{\sqrt{\log(dn)}}.$$

If there exist multiple such values of k, then choose the smallest such value. Since the vector r^* has its entries arranged in order, and since $\|r^{*+}\|_2 \geq \|r^{*-}\|_2$, we obtain the following relations for this chosen value of k:

$$\langle r^*,\, g_k\rangle = (r^{*+})^T g_k \geq \frac{\sqrt{k}\|r^{*+}\|_2}{\sqrt{\log(dn)}} \geq \frac{\|r^*\|_2}{2}\sqrt{\frac{k}{\log(dn)}} \geq \sqrt{\frac{\log^{2.5}(dn)}{p_{\text{obs}}}\frac{k}{\log(dn)}}.$$

The Cauchy-Schwarz inequality then implies

$$\langle \widetilde{r}_\ell,\, g_k\rangle \geq \langle r^*,\, g_k\rangle - \sqrt{k}\|\widetilde{r}_\ell - r^*\|_2 \overset{(i)}{\geq} \frac{3}{4}\sqrt{\frac{k\log^{1.5}(dn)}{p_{\text{obs}}}},$$

where the inequality (i) also uses our earlier bound (4.54), thereby proving the first property. Now towards the second property, we use the condition $\langle r^*,\, 1\rangle \geq 0$. Since the entries of r^* are arranged in order, we have $\langle r^*,\, g_k\rangle \geq 0$ for every $k \in [n]$. Applying the Cauchy-Schwarz inequality yields

$$\langle \widetilde{r}_\ell,\, g_k\rangle \geq \langle r^*,\, g_k\rangle - \sqrt{k}\|\widetilde{r}_\ell - r^*\|_2 \overset{(ii)}{>} -\frac{1}{4}\sqrt{\frac{k\log^{1.5}(dn)}{p_{\text{obs}}}},$$

where the inequality (ii) also uses our earlier bound (4.54), thereby proving the second property. This argument completes the proof of part (a).

4.6.9 Proof of Theorem 13(b): Lower bound for Dawid-Skene

The Gilbert-Varshamov bound [90, 258] guarantees existence of a set of η vectors, $x^1, \ldots, x^\eta \in \{-1, 1\}^d$ such that the normalized Hamming distance (4.1) between any pair of vectors in this set is lower bounded as $D_{\text{H}}(x^\ell, x^{\ell'}) \geq 0.25, \qquad$ for every $\ell, \ell' \in [\eta]$, where $\eta = \exp(c_1 d)$ for some constant $c_1 > 0$. For each $\ell \in [\eta]$, let $\mathbb{P}^\ell$ denote the probability distribution of Y induced by setting $x^* = x^\ell$. When $Q^* = q^{\text{DS}}1^T$ for some $q^{\text{DS}} \in [\frac{1}{10}, \frac{9}{10}]^n$, we have an upper bound on the Kullback-Leibler divergence between any pair of distributions $\ell \neq \ell' \in [\eta]$ as $D_{\text{KL}}(\mathbb{P}^\ell \| \mathbb{P}^{\ell'}) \;\leq\; 25\; p_{\text{obs}} d\; \|q^{\text{DS}} - \frac{1}{2}\|_2^2 \;\leq\; 25cd$, where we have used the assumption $\|q^{\text{DS}} - \frac{1}{2}\|_2^2 \leq \frac{c}{p_{\text{obs}}}$. Putting the above observations together into Fano's inequality [54] yields a lower bound on the expected value of the normalized Hamming error (4.1) for any estimator $\widehat{x}$ as:

$$\mathbb{E}[D_{\text{H}}(\widehat{x}, x^*)] \geq \frac{1}{8}\Big(1 - \frac{25cd + \log 2}{c_1 d}\Big) \overset{(i)}{\geq} \frac{1}{10},$$

as claimed, where inequality (i) results from setting the value of c as a small enough positive constant.

4.A Appendix: The majority voting estimator

In this section, we analyze the majority voting estimator, given by

$$[\widetilde{x}_{\text{MV}}]_j \in \arg\max_{b\in\{-1,1\}} \sum_{i=1}^n \mathbf{1}\{Y_{ij} = b\} \qquad \text{for every } j \in [d].$$

Here we use $\mathbf{1}\{\cdot\}$ to denote the indicator function. The following theorem provides bounds on the risk of majority voting under the Q^*-semimetric loss in the regime of interest (R).

Proposition 4. *For the majority vote estimator, the uniform risk over the Dawid-Skene class is lower bounded as*

$$\sup_{x^* \in \{-1,1\}^d} \sup_{Q^* \in \mathbb{C}_{DS}} \mathbb{E}[\mathcal{L}_{Q^*}(\widetilde{x}_{MV}, x^*)] \geq c_2 \frac{1}{\sqrt{np_{\text{obs}}}}, \tag{4.55}$$

for some positive constant c_2.

A comparison of the bound (4.55) with the results of Theorem 10, Theorem 12(a) and Theorem 13 shows that the majority voting estimator is suboptimal in terms of the sample complexity. Since this suboptimality holds for the (smaller) Dawid-Skene model class, it also holds for the (larger) intermediate model class, as well as the permutation-based model class.
The remainder of this section is devoted to the proof of this claim.

Proof of Proposition 4

We begin with a lower bound due to Feller [77] (see also [161, Theorem 7.3.1]) on the tail probability of a sum of independent random variables.

Lemma 26 (Feller)**.** *There exist positive universal constants c_1 and c_2 such that for any set of independent random variables $X_1, \ldots, X_n$ satisfying $\mathbb{E}[X_i] = 0$ and $|X_i| \leq M$ for every $i \in [n]$, if $\sum_{i=1}^n \mathbb{E}[(X_i)^2] \geq c_1$ then*

$$\mathbb{P}\big(\sum_{i=1}^n X_i > t\big) \geq c_2 \exp\Big(\frac{-t^2}{12 \sum_{i=1}^n \mathbb{E}[(X_i)^2]}\Big),$$

for every $t \in [0, \frac{\sum_{i=1}^n \mathbb{E}[(X_i)^2]}{M^2 \sqrt{c_1}}]$.

In what follows, we use Lemma 26 to derive the claimed lower bound on the error incurred by the majority voting algorithm. To this end, let $S \subset [n]$ denote the set of some $|S| = \sqrt{\frac{n}{2p_{\text{obs}}}}$ workers. Consider the following value of matrix Q^*:

$$Q^*_{ij} = \begin{cases} 1 & \text{if } i \in S \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

Then for any question $j \in [d]$, we have $\sum_{i=1}^n (2Q^*_{ij} - 1)^2 = \sqrt{\frac{n}{2p_{\text{obs}}}}$.

Now suppose that $x^*_j = -1$ for every question $j \in [d]$. Then for every $i \in S$, the observations are distributed as

$$Y_{ij} = \begin{cases} 0 & \text{with probability } 1 - p_{\text{obs}} \\ -1 & \text{with probability } p_{\text{obs}}, \end{cases}$$

and for every $i \notin S$, as

$$Y_{ij} = \begin{cases} 0 & \text{with probability } 1 - p_{\text{obs}} \\ -1 & \text{with probability } 0.5 p_{\text{obs}} \\ 1 & \text{with probability } 0.5 p_{\text{obs}}. \end{cases}$$

Consider any question $j \in [d]$. Then in this setting, the majority voting estimator incorrectly estimates the value of x^*_j when $\sum_{i=1}^n Y_{ij} > 0$. We now use Lemma 26 to obtain a lower bound on the probability of the occurrence of this event. Some simple algebra yields

$$\sum_{i=1}^n \mathbb{E}[Y_{ij}] = -|S| p_{\text{obs}} \qquad \text{and} \qquad \sum_{i=1}^n \mathbb{E}[(Y_{ij})^2] = n p_{\text{obs}}.$$

In order to satisfy the conditions required by the lemma, we assume that $np_{\text{obs}} > c_1$. Note that this condition makes the problem strictly easier than the condition $np_{\text{obs}} \geq 1$ assumed otherwise, and affects the lower bounds by at most a constant factor c_1. An application of Lemma 26 with $t = -\sum_{i=1}^n \mathbb{E}[Y_{ij}] = |S| p_{\text{obs}}$ now yields

$$\mathbb{P}(\sum_{i=1}^n Y_{ij} > 0) \geq c_2 \exp\Big(\frac{-|S|^2 p_{\text{obs}}^2}{12 n p_{\text{obs}}} \Big) \overset{(i)}{\geq} c',$$

for some constant $c' > 0$ that may depend only on c_1 and c_2, where inequality (i) is a consequence of the choice $|S| = \sqrt{\frac{n}{2p_{\text{obs}}}}$.

Now that we have established a constant-valued lower bound on the probability of error in the estimation of x^*_j for every $j \in [d]$, for the value of Q^* under consideration, we have

$$\mathbb{P}([\widetilde{x}_{\text{MV}}]_j \neq x^*_j) \sum_{i=1}^n (Q^*_{ij} - \frac{1}{2})^2 \geq \sqrt{\frac{n}{2p_{\text{obs}}}} c',$$

and consequently $\mathbb{E}[\mathcal{L}_{Q^*}(\widetilde{x}_{\text{MV}}, x^*)] \geq \frac{c'}{\sqrt{2np_{\text{obs}}}}$, as claimed.

Acknowledgements

We thank the authors of the paper [274] for sharing their implementation of their Spectral-EM algorithm.

Chapter 5

Matrix Completion and Recommendations

"Being complete is a state of utmost fulfillment."
– C. V. Raman

5.1 Introduction

We consider the problem of noisy matrix completion wherein a structured matrix must be reconstructed from partial and noisy observations. The archetypal example of this setting is in the context of recommender systems [133], and other applications include understanding images [145], credit risk monitoring [257], fluorescence spectroscopy [95], and modeling signal-adaptive audio effects [215]. We refer the reader to the surveys [58, 92] for an overview of the vast literature on this topic.

We use a particular variant of a recommender system application as a running example throughout the chapter. In this context, suppose there are $n \geq 2$ users and $d \geq 2$ items. There is an unknown matrix $M^* \in [0,1]^{n \times d}$ that captures the users' preferences for the items. Specifically, the $(i,j)^{th}$ entry of M^*, M^*_{ij}, represents the probability that user i likes item j. The problem is to estimate this preference matrix $M^* \in [0,1]^{n \times d}$ from observing users' likes or dislikes for some subset of the items.

Following many of the seminal works [34, 35, 44, 122, 206, 245] in this area, we consider a random design observation model in this chapter. The standard random design setup we consider is associated to a parameter $p_{\text{obs}} \in (0,1]$, such that for any user-item pair (i,j), we observe user i's rating for item j with probability p_{obs}. When an entry (i,j) is observed, we get access to only a binary rating ($\{\text{like, dislike}\}$ or $\{0,1\}$), which arises as a Bernoulli realization of the true preference M^*_{ij}.[1] More formally, we observe a matrix $Y \in \{0, \frac{1}{2}, 1\}^{n \times d}$,

[1] Our results and proofs readily extend to any rating scheme with bounded values, such as five-star ratings. We focus on the binary case for purposes of brevity.

where

$$Y_{ij} = \begin{cases} 1 & \text{with probability } p_{\text{obs}} M^*_{ij} \qquad \text{(user } i \text{ likes item } j) \\ 0 & \text{with probability } p_{\text{obs}}(1 - M^*_{ij}) \qquad \text{(user } i \text{ dislikes item } j) \\ \frac{1}{2} & \text{with probability } 1 - p_{\text{obs}} \qquad \text{(no data available)}, \end{cases} \tag{5.1}$$

for every $(i, j) \in [n] \times [d]$. The goal is to estimate the underlying matrix M^* from the observed matrix Y.

It is not hard to see that one cannot hope to do achieve the aforementioned estimation goal in any non-trivial way without any modeling assumptions on the matrix M^*. In what follows, we discuss some modeling assumptions—we begin with the typical low non-negative rank assumption, followed by our proposed low "permutation-rank" model.

Non-negative rank: The problem of non-negative low-rank matrix completion assumes that the matrix M^* takes the form

$$M^* = UV^T,$$

for some matrices $U \in \mathbb{R}_+^{n\times r}$ and $V \in \mathbb{R}_+^{d\times r}$. Here, $r \in \{0, \ldots, \min\{d, n\}\}$ is a parameter termed the "non-negative" rank of the matrix. It is often assumed that the value r of the non-negative rank is known and that r is much smaller than $\min\{d, n\}$, but in this chapter we make no such assumptions. For any value of $r \in \{0, \ldots, \min\{d, n\}\}$, we let $\mathbb{C}_{\text{NR}}(r)$ denote the set of all matrices with a non-negative factorization of rank at most r,

$$\mathbb{C}_{\text{NR}}(r) := \{M \in [0, 1]^{n\times d} \mid M = UV^T,\ U \in \mathbb{R}_+^{n\times r},\ V \in \mathbb{R}_+^{d\times r}\}.$$

For any matrix M, the smallest value of r such that $M \in \mathbb{C}_{\text{NR}}(r)$ is termed its non-negative rank. We denote the non-negative rank of any matrix M as $\overline{r}(M)$.

Observe that any matrix $M \in \mathbb{C}_{\text{NR}}(r)$ can alternatively be written as

$$M = \sum_{\ell=1}^{r} u^{(\ell)}(v^{(\ell)})^T,$$

for some vectors $u^{(\ell)} \in \mathbb{R}_+^n,\ v^{(\ell)} \in \mathbb{R}_+^d$ such that $u^{(\ell)}(v^{(\ell)})^T \in [0, 1]^{n\times d}$ for every $\ell \in [r]$. Such a decomposition is interpreted as the existence of some r "features," where for each feature ℓ, the d entries of vector $v^{(\ell)}$ represent the contribution of feature ℓ to the d respective items, and the n entries of vector $u^{(\ell)}$ represent the amounts by which the n respective users are influenced by feature ℓ. The popular overview article by Koren, Bell, and Volinsky [133] provides an explanation for this assumption:

> "Latent factor models are an alternative approach that tries to explain the ratings by characterizing both items and users on, say, 20 to 100 factors inferred from the ratings patterns. In a sense, such factors comprise a computerized alternative to

> the aforementioned human-created song genes. For movies, the discovered factors might measure obvious dimensions such as comedy versus drama, amount of action, or orientation to children; less well-defined dimensions such as depth of character development or quirkiness; or completely uninterpretable dimensions. For users, each factor measures how much the user likes movies that score high on the corresponding movie factor."

Let us now delve deeper into this model[2], and continue the context of movie recommendations for concreteness. Suppose there are r features that govern the movie watching experience, where examples of such features are the amount of comedy content or the depth of character development. For any user $i \in [n]$ and any feature $\ell \in [r]$, we let $u_i^{(\ell)} \in \mathbb{R}_+$ denote the "affinity" of user i towards feature ℓ, and for any movie $j \in [d]$, we let $v_j^{(\ell)} \in \mathbb{R}_+$ denote the amount of content associated to feature ℓ in movie j. The conventional low non-negative rank model then assumes that the affinity of user i towards movie j conditioned on feature ℓ equals

$$u_i^{(\ell)} v_j^{(\ell)}.$$

Observe that this model is of a parameter-based form in that for any given feature ℓ, the entire behavior of any user or any movie is governed by a single parameter each ($u_i^{(\ell)}$ and $v_j^{(\ell)}$ for user i and item j respectively). Moreover, such an assumption has some unnatural implications. For instance, consider any two movies, say A and B, and any two users, say X and Y. Then conditioned on any feature ℓ, we have the implication

$$\frac{\text{Preference of user } X \text{ for movie } A}{\text{Preference of user } X \text{ for movie } B} = \frac{\text{Preference of user } Y \text{ for movie } A}{\text{Preference of user } Y \text{ for movie } B}.$$

In words, the low non-negative rank model inherently leads to the unrealistic assumption that for any given feature, the ratio of preferences for any pair of movies is *identical for all users*. Likewise, for any given feature, the ratio of preferences of any pair of users is *identical for all movies*. In this chapter, we assume the following more general model that overcomes these issues.

Permutation-rank: The permutation-rank model is also associated to an integer parameter $\rho \in \{0, \ldots, \min\{n, d\}\}$ that we will term as the permutation rank. In order to define this model, we first define some primitives. Let $\mathbb{C}_{\text{PR}}(0)$ denote the set of all matrices that have a permutation-rank of zero – this is a singleton set containing the all-zero matrix as its only element. Now let $\mathbb{C}_{\text{PR}}(1)$ denote the set of matrices that have a permutation rank of at most 1, that is,

$$\begin{aligned}\mathbb{C}_{\text{PR}}(1) := \{&M \in [0,1]^{n \times d} \mid \exists \text{ permutations } \pi_1 : [n] \to [n] \text{ and } \pi_2 : [d] \to [d] \text{ such that}\\ &M_{ij} \geq M_{i'j'} \text{ for every quadruple } (i, j, i', j') \text{ such that } \pi_1(i) \geq \pi_1(i') \text{ and } \pi_2(j) \geq \pi_2(j') \ \}.\end{aligned}$$

[2] A slightly different, alternative interpretation is discussed in Appendix 5.A.

In words, a non-zero matrix is said to have a permutation rank of 1 if there exists a permutation of its rows and columns such that the entries of the resulting matrix are non-decreasing down any column and to the right along any row. Observe that any matrix with the conventional (non-negative) rank equal to 1 also belongs to the set $\mathbb{C}_{\text{PR}}(1)$. Moreover, a matrix in $\mathbb{C}_{\text{PR}}(1)$ can have any non-negative rank—the set of matrices with a permutation-rank of 1 also includes matrices with a full non-negative rank.

With these primitives, we are now ready to define the more general class of matrices with permutation rank ρ. To this end, for any value $\rho \in \{0, \ldots, \min\{n, d\}\}$, we define the set

$$\mathbb{C}_{\text{PR}}(\rho) := \Big\{ M \in [0,1]^{n \times d} \;\Big|\; M = \textstyle\sum_{\ell=1}^{\rho} Q^{\ell} \text{ for some matrices } Q^1, \ldots, Q^{\rho} \in \mathbb{C}_{\text{PR}}(1) \Big\},$$

of matrices having a permutation-rank at most ρ. Note that the permutations in $\mathbb{C}_{\text{PR}}(1)$ are allowed to be different for each of the constituent matrices $Q^1, \ldots, Q^{\rho}$. For any matrix M, the smallest value of ρ such that $M \in \mathbb{C}_{\text{PR}}(\rho)$ is termed its permutation-rank, and is denoted as $\overline{\rho}(M)$.

Revisiting the example of movie recommendations, the interpretation of this more general permutation-rank model is that conditioned on any feature $\ell \in [r]$, the preference ordering across movies continues to be consistent for different users, but the values of these preferences *need not* be identical scalings of each other.

Observe that the conventional non-negative matrix-completion setting $\mathbb{C}_{\text{NR}}(r)$ is a special case of the permutation-rank matrix-completion setting where each matrix Q^{ℓ} is restricted to be of rank one. Whenever $r < \min\{d, n\}$, we have the strict inclusion $\mathbb{C}_{\text{NR}}(r) \subset \mathbb{C}_{\text{PR}}(r)$.

Outline and main contributions: Having discussed the limitations of the non-negative rank model, and having defined the new permutation-based model that overcomes these issues, we now outline our contributions in the remainder of the chapter. In Section 5.2 we present our main results on the problem of estimating the matrix M^* (in the Frobenius norm) from partial and noisy observations. Specifically, we present a certain regularized least squares estimator which we prove is minimax-optimal for estimation over the permutation-rank model. We also show that surprisingly, even if one considers the more restrictive non-negative rank model, and even if the rank is known, no estimator can yield a lower statistical error (up to logarithmic factors). We also analyze the computationally efficient Singular Value Thresholding (SVT) algorithm and show that this algorithm yields consistent estimates over the permutation-rank model, in addition to yielding the optimal estimate under the non-negative rank model. In Section 5.3, we establish some interesting properties of the permutation-rank model, and also derive certain relationships of this model with the non-negative rank model. We present a concluding discussion in Section 5.4. In Section 5.5 we present the proofs of our theoretical results.

5.2 Main results on estimating M^*

In this section, we present our main theoretical results on estimating the underlying matrix M^* from the observations Y.

5.2.1 Statistically optimal estimation

In what follows, we establish sharp guarantees on the estimation of matrices in $\mathbb{C}_{\rm NR}$ and $\mathbb{C}_{\rm PR}$ from observations of the form (5.1). Our upper bounds employ the regularized least squares estimator $\widehat{M}_{\rm LSReg}$, that operates on the observed data Y as follows. The estimator first computes

$$Y' := \frac{1}{p_{\rm obs}} Y - \frac{1-p_{\rm obs}}{2p_{\rm obs}} 11^T. \tag{5.2a}$$

Now letting $\overline{\rho}(M)$ denote the permutation-rank of any matrix M, the estimator then computes

$$\widehat{M}_{\rm LSReg} \in \mathop{\arg\min}_{M\in[0,1]^{n\times d}} \left(|\!|\!| Y' - M |\!|\!|_{\rm F}^2 + \frac{\overline{\rho}(M)\max\{n,d\}\log^{2.01} d}{p_{\rm obs}} \right) \tag{5.2b}$$

as the final estimate.

Observe that importantly, the estimator $\widehat{M}_{\rm LSReg}$ does *not* need to know the value of the true permutation-rank of the underlying matrix. Note that while the estimator as stated assumes to know $p_{\rm obs}$, this is not a critical issue since if unknown, this parameter can be estimated accurately from the data.

We now present sharp oracle inequalities on the statistical risk of low permutation-rank matrix estimation, also showing that the estimator $\widehat{M}_{\rm LSReg}$ is statistically optimal up to logarithmic factors. In order to state our results cleanly, we introduce the notation $B^{\rm P}(\rho,\epsilon)$ to denote the set of all matrices that are at most ϵ away from some matrix with permutation-rank ρ,

$$B^{\rm P}(\rho,\epsilon) := \big\{ M \in [0,1]^{n\times d} \mid \exists M' \in [0,1]^{n\times d} \text{ s.t. } \overline{\rho}(M') \leq \rho \text{ and } |\!|\!| M - M' |\!|\!|_{\rm F} \leq \epsilon \big\}.$$

We also use $B^{\rm N}(r,\epsilon)$ to denote the set of all matrices that are at most ϵ away from some matrix with non-negative-rank r,

$$B^{\rm N}(r,\epsilon) := \big\{ M \in [0,1]^{n\times d} \mid \exists M' \in [0,1]^{n\times d} \text{ s.t. } \overline{r}(M') \leq r \text{ and } |\!|\!| M - M' |\!|\!|_{\rm F} \leq \epsilon \big\}.$$

In stating the following theorem, as well as throughout the remainder of the chapter, we use c, c', c_1 etc. to denote positive universal constants. The values of these constants may differ for different results.

Theorem 14. *(a) For any matrix $M^* \in [0,1]^{n \times d}$, the error incurred by the regularized least squares estimator $\widehat{M}_{LSReg}$ is upper bounded as*

$$\frac{1}{nd} |||\widehat{M}_{LSReg} - M^*|||_F^2 \leq c_1 \min \Big\{ \frac{\epsilon^2}{nd} + \frac{\rho \ \log^{2.01}(nd)}{\min\{n,d\} p_{\text{obs}}}, 1 \Big\}, \tag{5.3a}$$

with probability at least $1 - e^{-c_0 \max\{n,d\} \log(\max\{nd\})}$, for any (ρ, ϵ) such that $M^* \in B^P(\rho, \epsilon)$.
(b) Conversely, for any integer $r \in [\max\{n,d\}]$, any value $\epsilon \geq 0$, and any estimator $\widehat{M}$, there exists a matrix $M^* \in B^N(r, \epsilon)$ such that the estimator $\widehat{M}$ incurs an error at least

$$\mathbb{E}\big[\frac{1}{nd} |||\widehat{M} - M^*|||_F^2\big] \geq c_2 \min \Big\{ \frac{\epsilon^2}{nd} + \frac{r}{\min\{n,d\} p_{\text{obs}}}, 1 \Big\}. \tag{5.3b}$$

The oracle inequalities in Theorem 14 can now be used to obtain sharp bounds on the minimax risk for the problem of matrix completion over the sets $\mathbb{C}_{\text{NR}}$ and $\mathbb{C}_{\text{PR}}$.

Remark 1 (Minimax risk). *Part (a) of Theorem 14 implies that for any value of $\rho \in [\min\{n,d\}]$, the error incurred by the regularized least squares estimator $\widehat{M}_{LSReg}$ is upper bounded as*

$$\sup_{M^* \in \mathbb{C}_{PR}(\rho)} \frac{1}{dn} |||\widehat{M}_{LSReg} - M^*|||_F^2 \leq c_1 \min \Big\{ \frac{\rho \ \log^{2.01}(nd)}{\min\{n,d\} p_{\text{obs}}}, 1 \Big\}, \tag{5.4a}$$

with probability at least $1 - e^{-c_0 (n+d) \log(nd)}$. The deterministic $\frac{1}{nd} |||\widehat{M}_{LSReg} - M^*|||_F^2 \leq 1$ further implies a similar bound on the risk incurred by the estimator $\widehat{M}_{LSReg}$,

$$\sup_{M^* \in \mathbb{C}_{PR}(\rho)} \frac{1}{dn} \mathbb{E}[|||\widehat{M}_{LSReg} - M^*|||_F^2] \leq c_1' \min \Big\{ \frac{\rho \ \log^{2.01}(nd)}{\min\{n,d\} p_{\text{obs}}}, 1 \Big\}. \tag{5.4b}$$

Conversely, part (b) of Theorem 14 implies that for any $r \in [\max\{n,d\}]$, any estimator $\widehat{M}$ incurs an error lower bounded as

$$\sup_{M^* \in \mathbb{C}_{NR}(r)} \frac{1}{dn} \mathbb{E}[|||\widehat{M} - M^*|||_F^2] \geq c_2 \min \Big\{ \frac{r}{\min\{n,d\} p_{\text{obs}}}, 1 \Big\}. \tag{5.4c}$$

We have thus established a sharp characterization of the minimax risk, up to logarithmic factors. An important consequence of our oracle and minimax results is the multi-fold benefit of moving from the restrictive non-negative-rank assumptions to the strictly and significantly more general permutation-rank assumptions. Fitting a permutation-rank k model when the true matrix actually has a non-negative rank of k incurs very little additional (overfitting) error. On the other hand, as we show in the next section, fitting a non-negative rank k model when the true matrix actually has a permutation-rank of k can incur a very high (model mismatch) error.

A special case of our present problem is equivalent to the setting considered earlier in Chapter 2, corresponding to the case when the value of ρ is known and equal to 1, the matrix M^* is square with $n = d$, and all entries of M^* satisfy the shifted-skew-symmetry condition $M^*_{ij} + M^*_{ji} = 1$. The proof of the upper bound of Theorem 14(a) employs the proof framework established in Chapter 2. Our result of Theorem 14(a) in the present chapter, in turn, augments the minimax results of Theorem 1 in Chapter 2 by providing oracle inequalities for the problem considered therein. Our result also provides sharp guarantees on the estimation of "mixtures" of different permutations in the setting of Chapter 2.

5.2.2 Computationally efficient estimator

At this point, we do not know how to compute the regularized least squares estimator (5.2b) in an efficient manner. Consequently, in this section we turn to another estimator – the singular value thresholding (SVT) estimator. Singular value thresholding has been used either directly or as a subroutine in several past papers on the conventional low-rank matrix completion problem (see, for example, [31, 44, 67]). An algorithm for fast computation of the SVT is provided in [32].

In what follows, we show that for the setting considered in this chapter, the SVT estimator is consistent for estimation under the permutation-rank model (with an error suboptimal by a factor of $\sqrt{\min\{n,d\}p_{\text{obs}}}$) and is also simultaneously optimal for estimation under the non-negative-rank model. Moreover, the estimator does not need to now the value of ρ or r nor does it need to know whether the underlying matrix is drawn from a permutation-rank model or the conventional non-negative-rank model.

Let us first describe the SVT estimator.[3] From the observation matrix $Y \in \{0, \frac{1}{2}, 1\}^{n \times d}$, we first obtain the transformed observation matrix Y' as in (5.2a). Let the singular value decomposition of this matrix be $Y' = UDV^T$, where the $(n \times d)$ matrix D is diagonal and the $(n \times d)$ matrices U and V are orthonormal. For a threshold $\lambda > 0$ to be specified, define another diagonal matrix T_λ with entries

$$[T_\lambda]_{jj} = \begin{cases} 0 & \text{if } D_{jj} < \lambda \\ D_{jj} - \lambda & \text{if } D_{jj} \geq \lambda, \end{cases} \tag{5.5}$$

for every $j \in [\max\{n, d\}]$. Finally, the SVT estimator is given by

$$\widehat{M}_{\text{SVT}} = UT_\lambda V^T.$$

The following theorem now establishes guarantees for the singular value thresholding estimator, showing that it provides a consistent (but slightly suboptimal) estimate under the permutation-rank model. Using the same proof framework, for the sake of completeness, we also derive the previously known guarantees [44, 132] on for optimal estimation under the non-negative rank model.

[3] This estimator is identical to the SVT estimator studied in Section 2.3.2 in Chapter 2, but we nevertheless describe it here for the sake of completeness.

Theorem 15. *Suppose that* $p_{\text{obs}} \geq \frac{1}{\min\{n,d\}} \log^7(nd)$.
(a) For any $\rho \in [\min\{n,d\}]$ *and any matrix* $M^* \in \mathbb{C}_{PR}(\rho)$*, the soft-SVT estimator* $\widehat{M}_{SVT}$ *with threshold* $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ *incurs an error upper bounded as*

$$\frac{1}{nd}|||\widehat{M}_{SVT} - M^*|||_F^2 \leq c_1 \frac{\rho}{\sqrt{\min\{n,d\}p_{\text{obs}}}}, \tag{5.6a}$$

with probability at least $1 - e^{-c_0 \max\{n,d\}}$.
(b) For any $r \in [\min\{n,d\}]$ *and any matrix* $M^* \in \mathbb{C}_{NR}(r)$*, the soft-SVT estimator* $\widehat{M}_{SVT}$ *with threshold* $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ *incurs an error upper bounded as*

$$\frac{1}{nd}|||\widehat{M}_{SVT} - M^*|||_F^2 \leq c_1 \frac{r}{\min\{n,d\}p_{\text{obs}}}, \tag{5.6b}$$

with probability at least $1 - e^{-c_0 \max\{n,d\}}$.

Observe that the bound (5.6a) on the risk of the SVT estimator for the permutation-based model has a $\sqrt{\min\{n,d\}}$ term in the denominator, as opposed to a $\min\{n,d\}$ term in the oracle risk established in Theorem 14 and the minimax risk established in Remark 1. On the other hand, a comparison with the results of Section 5.2.1 yields that the bound (5.6b) for the SVT estimator is optimal with respect to the non-negative rank model. Importantly, the estimator automatically adapts to the container set $\mathbb{C}_{\text{PR}}(\cdot)$ or $\mathbb{C}_{\text{NR}}(\cdot)$ as well as to the true value of r or ρ.

5.3 Properties of permutation-rank model

In this section, we derive some more insights on the proposed permutation-rank model.

5.3.1 Comparison between permutation-rank model and non-negative-rank model

We begin by comparing the permutation-rank model with the conventional non-negative rank model. To this end, first observe that the definitions of the two models immediately imply that the permutation-rank of any matrix is always upper bounded by its non-negative rank, that is, for any matrix M:

$$\overline{\rho}(M) \quad \leq \quad \overline{r}(M),$$

A natural question that now arises is whether, in addition to this simple relation, there is any additional general condition that constrains the two notions of the matrix rank. The following proposition shows that there is no other guaranteed relation between the two notions of matrix rank.

In order to state the proposition, we introduce two additional pieces of notation: For any integer $k \geq 0$, we let J_k denote an upper triangular matrix of size $(k \times k)$ with all entries on and above the diagonal set as 1, and let I_k denote the identity matrix of size $(k \times k)$.

Proposition 5. *For any values $0 < \rho \leq r \leq \min\{n, d\}$, there exists matrices whose permutation-rank is ρ and non-negative rank is r. For instance, the following block matrix M of size $(n \times d)$,*

$$M := \begin{bmatrix} J_{r-\rho+1} & 0 & 0 \\ 0 & I_{\rho-1} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

has $\overline{r}(M) = r$ and $\overline{\rho}(M) = \rho$.

We now investigate a second relation between the two models, towards which we begin with the simple observation that for every positive integer $k < \min\{d, n\}$, there is a strict inclusion

$$\mathbb{C}_{\mathrm{NR}}(k) \subset \mathbb{C}_{\mathrm{PR}}(k).$$

Furthermore, recall from our discussion earlier that the assumptions of the permutation-rank model are much less restrictive than the assumptions of the non-negative rank model. With this context, a natural question that arises is how badly would an estimator that fits a non-negative rank of k be biased when the true underlying matrix may actually have a permutation rank of k. The following proposition answers this question by quantifying the distance between non-negative-rank and permutation-rank models.

Proposition 6. *Consider any positive integer $k \leq \frac{1}{2}\min\{d, n\}$, and any estimator $\widetilde{M_k}$ that outputs a matrix in $\mathbb{C}_{NR}(k)$. The error incurred by this estimator when the true matrix lies in the set $\mathbb{C}_{PR}(k)$ is lower bounded as*

$$\sup_{M^* \in \mathbb{C}_{PR}(k)} \frac{1}{dn} |\!|\!| M - \widetilde{M_k} |\!|\!|_F^2 \geq c_3 \frac{1}{k}, \tag{5.7}$$

with probability 1.

This result is a consequence of the following bound on the Hausdorff distance between the two sets, which is proved as a part of the proof of Proposition 6:

$$\sup_{M_1 \in \mathbb{C}_{\mathrm{PR}}(k)} \inf_{M_2 \in \mathbb{C}_{\mathrm{NR}}(k)} \frac{1}{dn} |\!|\!| M_1 - M_2 |\!|\!|_{\mathrm{F}}^2 \geq c_3 \frac{1}{k}, \tag{5.8}$$

whenever $k \leq \frac{1}{2}\min\{d, n\}$.

Observe that when k is a constant (but n and d are allowed to grow), the right hand sides of the bounds (5.7) and (5.8) also equal a constant, which is the largest possible error and the largest possible order-wise gap between any pair of matrices in $[0, 1]^{n \times d}$.

5.3.2 No "good" convex approximation

In this section, we investigate a question about an important property of the permutation-based set, and in particular, its primitive $\mathbb{C}_{\text{PR}}(1)$. With the often-pursued goal of optimization over the set $\mathbb{C}_{\text{PR}}(1)$ in mind, a natural question that arises is: Is the set $\mathbb{C}_{\text{PR}}(1)$ is convex? If not, then does it at least have a "good" convex approximation? The following proposition answers these questions in the negative.

Proposition 7. *There is a constant $c > 0$ such that every convex set $\mathbb{C} \subseteq \mathbb{R}^{n \times n}$ must necessarily satisfy*

$$\frac{1}{nd} \max\Big\{ \sup_{M_1 \in \mathbb{C}_{PR}(1)} \inf_{M_2 \in \mathbb{C}} |\!|\!| M_1 - M_2 |\!|\!|_F^2 \ , \ \sup_{M_2 \in \mathbb{C}} \inf_{M_1 \in \mathbb{C}_{PR}(1)} |\!|\!| M_1 - M_2 |\!|\!|_F^2 \Big\} \geq c.$$

A specific example of a convex set $\mathbb{C}$ is the convex hull of $\mathbb{C}_{\text{PR}}(1)$. Then by definition we have the relation $\sup\limits_{M_1 \in \mathbb{C}_{\text{PR}}(1)} \inf\limits_{M_2 \in \mathbb{C}} |\!|\!| M_1 - M_2 |\!|\!|_{\text{F}}^2 = 0$. Consequently, Proposition 7 implies that $\sup\limits_{M_2 \in \mathbb{C}} \inf\limits_{M_1 \in \mathbb{C}_{\text{PR}}(1)} |\!|\!| M_1 - M_2 |\!|\!|_{\text{F}}^2 = \Theta(nd)$, thereby suggesting that the convex hull of $\mathbb{C}_{\text{PR}}(1)$ is a much larger set than $\mathbb{C}_{\text{PR}}(1)$ itself.

The proof of Proposition 7 relies on a more general result that we derive, relating a certain notion of inherent (lack of) convexity of a set to the Hausdorff distance between that set and any convex approximation.

Note that this result does not preclude the possibility that an optimization procedure over a convex approximation to $\mathbb{C}_{\text{PR}}(1)$ converges close enough to some element of $\mathbb{C}_{\text{PR}}(1)$ itself. We leave the investigation of this possibility to future work.

5.3.3 On the uniqueness of decomposition

In this section, we investigate conditions for the uniqueness of the decomposition of any matrix into its constituent components that have a permutation-rank of one.

In the conventional setting of low non-negative rank matrix completion, there has been considerable interest in the conditions required for uniqueness of the decomposition of matrices into their constituent non-negative rank-one matrices [7, 68, 91, 141, 251]. In this section, we consider an analogous question under the permutation-rank setting. In more detail, consider any matrix $M \in [0,1]^{n \times d}$ with a permutation-rank decomposition

$$M = \sum_{\ell=1}^{\overline{\rho}(M)} M^{(\ell)}, \tag{5.9}$$

where $M^{(\ell)} \in \mathbb{C}_{\text{PR}}(1)$ for every $\ell \in [\overline{\rho}(M)]$. Then under what conditions on the matrix M is the set $\{M^{(1)}, \ldots, M^{(\overline{\rho}(M))}\}$ of constituent matrices unique?

Proposition 8. *A necessary condition for the uniqueness of a permutation-rank decomposition* (5.9) *is that for every coordinate $(i, j) \in [n] \times [d]$, there is at most one $\ell \in [\rho]$ such that $M_{ij}^{(\ell)}$ is non-zero and distinct from all other entries of $M^{(\ell)}$.*

Note that the necessary condition continues to hold even if we restrict attention to only symmetric matrices.

The result of Proposition 8 indicates that any sufficient condition(s) for uniqueness of the decomposition will be extremely strong. Moreover, we believe that the conditions for sufficiency may be significantly stronger than those necessitated by Proposition 8. The reason for such drastic requirements for uniqueness is the high-degree of flexibility offered by the permutation-rank model.

We illustrate the condition necessitated by Proposition 8 by means of a simple example. Consider the following matrix M with $n = d = 2$ and $\overline{\rho}(M) = 2$ and decomposition into $M^{(1)}, M^{(2)} \in \mathbb{C}_{\text{PR}}(1)$:

$$M := \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix} = \begin{bmatrix} 0 & .3 \\ .3 & .9 \end{bmatrix} + \begin{bmatrix} 1 & .3 \\ .3 & .1 \end{bmatrix}$$

As described in the statement of Proposition 8, the condition specified therein is required to hold for every coordinate of the matrix. Let us first evaluate this condition for coordinate $(1, 1)$. Since $M_{11}^{(1)} = 0$, there is at most one $\ell \in \{1, 2\}$ such that $M_{11}^{(\ell)}$ is non-zero. The coordinate $(1, 1)$ therefore passes the necessary condition for uniqueness. Moving on to coordinate $(1, 2)$, we have $M_{12}^{(1)} = M_{21}^{(1)}$ and hence there is at most one $\ell \in \{1, 2\}$ such that $M_{12}^{(\ell)}$ is distinct from all other entries of $M^{(\ell)}$. The coordinate $(1, 2)$ also passes the condition necessary for uniqueness. The argument for coordinate $(1, 2)$ also applies to coordinate $(2, 1)$ since the matrices involved are symmetric. We finally test coordinate $(2, 2)$. Observe that $M_{22}^{(1)} \notin \{0, M_{11}^{(1)}, M_{12}^{(1)}, M_{21}^{(1)}\}$ and $M_{22}^{(2)} \notin \{0, M_{11}^{(2)}, M_{12}^{(2)}, M_{21}^{(2)}\}$. As a consequence, for both $\ell = 1$ and $\ell = 2$, we have that $M_{22}^{(\ell)}$ is non-zero and distinct from all other entries of $M^{(\ell)}$. The condition necessary for uniqueness is thus violated. Indeed, as guaranteed by Proposition 8, there exits other decomposition of M, for instance,

$$M = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix} = \begin{bmatrix} 0 & .4 \\ .4 & .9 \end{bmatrix} + \begin{bmatrix} 1 & .2 \\ .2 & .1 \end{bmatrix}.$$

5.4 Discussion

We see that the conventional low-rank models for matrix completion and denoising are equivalent to parameter-based assumptions with undesirable implications. We propose a new permutation-rank approach and argue, by means of a philosophical discussion as well as theoretical guarantees, that this approach offers significant benefits at little additional cost.

We established benefits of the permutation-based approach for the matrix completion problem under the random design observation setting. In the literature, the classical low (non-negative) rank matrix completion problem has recently been studied under other observation models such as weighted random sampling [178], fixed design [109, 130], streaming/active learning [9, 113, 272], or biased observation models [104], which are also of interest in the context of permutation-rank matrix completion.

5.5 Proofs

In this section, we present the proofs of the claimed results. We assume without loss of generality that the values of n and d are large enough (that is, greater than certain constants) – otherwise the results continue to hold with different constant prefactors.

5.5.1 Proof of Theorem 14(a): Oracle upper bound

The proof of this theorem builds on the framework established in the proof of Theorem 1 in Chapter 2. In particular, the problem setting of Theorem 1 is a special case of the present problem, restricted to the case of $n = d$, $\rho = 1$, and establishing control over the minimax risk. In the present proof, we employ several additional techniques in order to generalize to the setting under consideration in the present chapter.

Let us assume without loss of generality that $n \leq d$.

One can verify that the matrix Y' can equivalently be written in a linearized form as

$$Y' = M^* + \frac{1}{p_{\text{obs}}} W', \tag{5.10a}$$

where W' has entries that are independent, and are distributed as

$$[W']_{ij} = \begin{cases} p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) + \frac{1}{2} & \text{with probability } p_{\text{obs}}[M^*]_{ij} \\ p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) - \frac{1}{2} & \text{with probability } p_{\text{obs}}(1 - [M^*]_{ij}) \\ p_{\text{obs}}(\frac{1}{2} - [M^*]_{ij}) & \text{with probability } 1 - p_{\text{obs}}. \end{cases} \tag{5.10b}$$

We begin by introducing some additional notation in order to accommodate the arbitrary permutation-rank of M^* and the fact that each constituent component in $\mathbb{C}_{\text{PR}}(1)$ can have any arbitrary permutation. For any pair of permutations $\pi : [n] \rightarrow [n]$ and $\sigma : [d] \rightarrow [d]$, we first define the set

$$\begin{aligned} \mathbb{C}_{\text{PR}}(1; \pi, \sigma) := \{M \in \mathbb{C}_{\text{PR}}(1) \mid & \text{ rows and columns of } M \text{ are ordered} \\ & \text{according to } \pi \text{ and } \sigma \text{ respectively}\}. \end{aligned}$$

Now let Π denote the set of all possible permutations of d items, and let Σ denote the set of all possible permutations of the n users. Consider any value $k \in [n]$, any sequence $\widetilde{\Pi}^{(k)} := (\pi_1, \ldots, \pi_k) \in \Pi^k$ and any sequence $\widetilde{\Sigma}^{(k)} := (\sigma_1, \ldots, \sigma_k) \in \Sigma^k$. Define the associated sets

Define an associated set

$$\mathbb{C}_{\mathrm{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) := \Big\{ M = \sum_{\ell=1}^{k} M^{(\ell)} \Big| \; M^{(\ell)} \in \mathbb{C}_{\mathrm{PR}}(1; \pi_\ell, \sigma_\ell) \text{ for every } \ell \in [k] \Big\}.$$

We then define the estimator

$$M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} \in \underset{M \in \mathbb{C}_{\mathrm{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})}{\arg\min} \; \|\!| Y' - M |\!\|_{\mathrm{F}}^2,$$

using which the least squares estimator (5.2b) can equivalently be rewritten as

$$\widehat{M}_{\mathrm{LSReg}} \in \underset{k \in [n]}{\arg\min} \; \underset{\substack{\widetilde{\Pi}^{(k)} \subseteq \Pi^k, \\ \widetilde{\Sigma}^{(k)} \subseteq \Sigma^k}}{\arg\min} \; \|\!| Y' - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!\|_{\mathrm{F}}^2 + \frac{kd \log^{2.01} d}{p_{\mathrm{obs}}}.$$

Define $M_0 \in [0,1]^{n \times d}$ as the matrix

$$M_0 \in \underset{M \in [0,1]^{n \times d}}{\arg\min} \Big(\|\!| M - M^* |\!\|_{\mathrm{F}}^2 + \frac{\overline{\rho}(M) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big),$$

and an associated set $\widetilde{\Gamma}$ as

$$\begin{aligned}\widetilde{\Gamma} := \Big\{ (k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in [n] \times \Pi^k \times \Sigma^k \Big| \|\!| Y' - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!\|_{\mathrm{F}}^2 + \frac{kd \log^{2.01} d}{p_{\mathrm{obs}}} \\ \leq \|\!| Y' - M_0 |\!\|_{\mathrm{F}}^2 + \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big\}.\end{aligned}$$

Note that the set $\widetilde{\Gamma}$ is guaranteed to be non-empty since the parameter and permutations corresponding to M_0 always lie in $\widetilde{\Gamma}$. We will subsequently show the following bound for any $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$:

$$\mathbb{P}\Big(\|\!| M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0 |\!\|_{\mathrm{F}}^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big) \geq 1 - e^{-4kd \log d}, \tag{5.11}$$

for some positive universal constant c_1. Under our assumption of $d \geq n$, for any value of k the cardinality of the set $\widetilde{\Gamma}$ restricted to any k is at most $e^{2kd \log d}$. Hence a union bound over all $k \in [n]$ and all permutations, applied to (5.11) yields

$$\mathbb{P}\Big(\max_{(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}} \|\!| M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0 |\!\|_{\mathrm{F}}^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big) \geq 1 - e^{-d \log d}.$$

Since $\widehat{M}_{\mathrm{LSReg}}$ is equal to $M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}$ for some $k \in [n]$ and some $(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$, this tail bound yields

$$\mathbb{P}\Big(\|\!| \widehat{M}_{\mathrm{LSReg}} - M_0 |\!\|_{\mathrm{F}}^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big) \geq 1 - e^{-d \log d}.$$

Finally, applying the triangle inequality yields the claimed result

$$\mathbb{P}\Big(|\!|\!| \widehat{M}_{\mathrm{LSReg}} - M^* |\!|\!|_{\mathrm{F}}^2 \leq 2 |\!|\!| M^* - M_0 |\!|\!|_{\mathrm{F}}^2 + 2c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big) \geq 1 - e^{-d \log d}.$$

The remainder of our proof is devoted to proving the claim (5.11). By definition, any $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$ must satisfy the inequality

$$|\!|\!| Y - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!|\!|_{\mathrm{F}}^2 + \frac{k d \log^{2.01} d}{p_{\mathrm{obs}}} \leq |\!|\!| Y - M_0 |\!|\!|_{\mathrm{F}}^2 + \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}}.$$

Denoting the error in the estimate as $\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} := M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0$, and using the linearized form (5.10a), some algebraic manipulations yield the basic inequality

$$\frac{1}{2} |\!|\!| \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!|\!|_{\mathrm{F}}^2 \leq \frac{1}{p_{\mathrm{obs}}} \langle\!\langle W', \ \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} \rangle\!\rangle + \frac{1}{2} \frac{(\overline{\rho}(M_0) - k) d \log^{2.01} d}{p_{\mathrm{obs}}}. \tag{5.12}$$

Now consider the set of matrices

$$\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) := \Big\{ \alpha (M - M_0) \mid M \in \mathbb{C}_{\mathrm{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}), \ \alpha \in [0,1] \Big\}, \tag{5.13}$$

and note that $\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) \subseteq [-1,1]^{n \times d}$. For each choice of radius $t > 0$, define the random variable

$$Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(t) := \sup_{\substack{M_{\mathrm{DIFF}} \in \mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0), \\ |\!|\!| M_{\mathrm{DIFF}} |\!|\!|_{\mathrm{F}} \leq t}} \frac{1}{p_{\mathrm{obs}}} \langle\!\langle M_{\mathrm{DIFF}}, \ W' \rangle\!\rangle. \tag{5.14}$$

Using the basic inequality (5.12), the Frobenius norm error $|\!|\!| \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!|\!|_{\mathrm{F}}$ then satisfies the bound

$$\frac{1}{2} |\!|\!| \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!|\!|_{\mathrm{F}}^2 \leq \ Z_{\widetilde{\Pi}, \widetilde{\Sigma}}\big(|\!|\!| \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} |\!|\!|_{\mathrm{F}} \big) + \frac{1}{2} \frac{(\overline{\rho}(M_0) - k) d \log^{2.01} d}{p_{\mathrm{obs}}}. \tag{5.15}$$

Thus, in order to obtain our desired bound, we need to understand the behavior of the random quantity $Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(t)$.

Dy definition, the set $\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$ is "star-shaped", meaning that $\alpha M_{\mathrm{DIFF}} \in \mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ for every $\alpha \in [0,1]$ and every $M_{\mathrm{DIFF}} \in \mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$. Using this star-shaped property, we are guaranteed that $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)]$ grows at most linearly with δ. We are then in turn guaranteed the existence of some scalar $\delta_0 > 0$ satisfying the critical inequality

$$\mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0)] \leq \frac{\delta_0^2}{2}. \tag{5.16}$$

Our interest is in an upper bound to the smallest (strictly) positive solution δ_0 to the critical inequality (5.16), and moreover, our goal is to show that for every $t \geq \delta_0$, we have $|||\widehat{\Delta}|||_{\mathrm{F}} \leq c\sqrt{t\delta_0}$ with high probability. To this end, define a "bad" event $\mathcal{A}_t$ as

$$\mathcal{A}_t = \left\{\exists \Delta \in \mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) \mid |||\Delta|||_{\mathrm{F}} \geq \sqrt{t\delta_0} \quad \text{and} \quad \frac{1}{p_{\mathrm{obs}}}\langle\langle \Delta,\ W' \rangle\rangle \geq 2|||\Delta|||_{\mathrm{F}}\sqrt{t\delta_0}\right\}. \tag{5.17}$$

Using the star-shaped property of $\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$, it follows by a rescaling argument that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}] \qquad \text{for all } t \geq \delta_0.$$

The following lemma helps control the behavior of the random variable $Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0)$.

Lemma 27. *For any $\delta > 0$, the mean of $Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)$ is bounded as*

$$\mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)] \leq c_1 \frac{\max\{k, \overline{\rho}(M_0)\} d}{p_{\mathrm{obs}}} \log^2 d,$$

and for every $u > 0$, its tail probability is bounded as

$$\mathbb{P}\Big(Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)] + u\Big) \leq \exp\Big(\frac{-c_2 u^2 p_{\mathrm{obs}}}{\delta^2 + \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)] + u}\Big),$$

where c_1 and c_2 are positive universal constants.

From this lemma, we have the tail bound

$$\mathbb{P}\Big(Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0)] + \delta_0\sqrt{t\delta_0}\Big) \leq \exp\Big(\frac{-c_2(\delta_0\sqrt{t\delta_0})^2 p_{\mathrm{obs}}}{\delta_0^2 + \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0)] + (\delta_0\sqrt{t\delta_0})}\Big),$$

for all $t > 0$. By the definition of δ_0 in (5.16), we have $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0)] \leq \delta_0^2 \leq \delta_0\sqrt{t\delta_0}$ for all $t \geq \delta_0$, and consequently

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta_0) \geq 2\delta_0\sqrt{t\delta_0}] \ \leq\ \exp\Big(\frac{-c_2(\delta_0\sqrt{t\delta_0})^2 p_{\mathrm{obs}}}{3\delta_0\sqrt{t\delta_0}}\Big),$$

for all $t \geq \delta_0$. Now we must have either $|||\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}|||_{\mathrm{F}} \leq \sqrt{t\delta_0}$, or we have $|||\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}|||_{\mathrm{F}} > \sqrt{t\delta_0}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, our basic inequality implies that

$$\frac{1}{2}|||\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}|||_{\mathrm{F}}^2 \leq 2|||\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}|||_{\mathrm{F}}\sqrt{t\delta_0} + \frac{1}{2}\frac{(\overline{\rho}(M_0) - k)d\log^{2.01} d}{p_{\mathrm{obs}}},$$

and hence

$$|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}} \leq 4\sqrt{t\delta_0} + \sqrt{\frac{(\overline{\rho}(M_0)-k)d\log^{2.01} d}{p_{\mathrm{obs}}}}.$$

Putting together the pieces yields the bound

$$\mathbb{P}\Big(|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}}^2 \leq 32t\delta_0 + 2\frac{(\overline{\rho}(M_0)-k)d\log^{2.01} d}{p_{\mathrm{obs}}}\Big) \geq 1 - \exp\big(-c_2\delta_0\sqrt{t\delta_0}p_{\mathrm{obs}}\big), \tag{5.18}$$

for all $t \geq \delta_0$. Finally, from the bound on the expected value of $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(t)$ in Lemma 27, we see that the critical inequality (5.16) is satisfied for

$$\delta_0 = \sqrt{\frac{c_1\max\{\overline{\rho}(M_0),k\}d}{p_{\mathrm{obs}}}}\log d.$$

Setting $t = c'\delta_0$ in (5.18) for a large enough constant c' yields

$$\mathbb{P}\Big(|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}} \leq \frac{c_1'\overline{\rho}(M_0)d}{p_{\mathrm{obs}}}\log^2 d\Big) \geq 1 - \exp\Big(-4\max\{\overline{\rho}(M_0),k\}d\log d\Big), \tag{5.19}$$

for some constant $c_1' > 0$, thus proving the bound (5.11).
It remains to prove Lemma 27.

Proof of Lemma 27 Bounding $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)]$: We establish an upper bound on $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)]$ by using Dudley's entropy integral, as well as some auxiliary results on metric entropy. We use the notation $\log N(\epsilon, \mathbb{C}, |\!|\!|\cdot|\!|\!|_{\mathrm{F}})$ to denote the ϵ metric entropy of class $\mathbb{C} \subset \mathbb{R}^{n\times d}$ in the Frobenius norm metric $|\!|\!|\cdot|\!|\!|_{\mathrm{F}}$.

For convenience of analysis, we introduce a new random variable

$$\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}} := \sup_{M_{\mathrm{DIFF}}\in\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)};M_0)} \langle\!\langle M_{\mathrm{DIFF}},\, W'\rangle\!\rangle.$$

Then by definition, we have $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] \leq \frac{1}{p_{\mathrm{obs}}}\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}]$ for every $\delta > 0$. In addition, since $M_0 \in \mathbb{C}_{\mathrm{PR}}(k)$, it can be decomposed as $M_0 = \sum_{\ell=1}^{k} M_0^{(\ell)}$, for some matrices $M_0^{(1)},\ldots,M_0^{(k)} \in \mathbb{C}_{\mathrm{PR}}(1)$.

We introduce some additional notation for ease of exposition. If $\overline{\rho}(M_0) < k$, then let $M_0^{(\overline{\rho}(M_0)+1)},\ldots,M_0^{k}$ denote all-zero matrices. Hence we can equivalently write $M_0 = \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} M_0^{(\ell)}$. On the other hand, if $\overline{\rho}(M_0) > k$ then let $\pi_{k+1},\ldots,\pi_{\overline{\rho}(M_0)}$ be arbitrary (but fixed) permutations of n items and $\sigma_{k+1},\ldots,\sigma_{\overline{\rho}(M_0)}$ be arbitrary (but fixed) permutations of d items. With this notation in place, we have the following deterministic upper bound on the value of the random variable $\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}$:

$$\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}} \leq \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} \sup_{[M_{\mathrm{DIFF}}]_\ell\in\mathbb{C}_{\mathrm{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})} \langle\!\langle [M_{\mathrm{DIFF}}]_\ell,\, W'\rangle\!\rangle.$$

We also recall our assumption that $d \geq n$ without loss of generality. Now the truncated form of Dudley's entropy integral inequality yields[4]

$$\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}] \leq \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} c\left\{d^{-8} + \int_{\frac{1}{2}d^{-9}}^{2d} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)}), |\!|\!|.|\!|\!|_{\text{F}})}(\Delta\epsilon)\right\}, \tag{5.20}$$

where we have used the fact that the diameter of the set $\mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})$ is at most $2d$ in the Frobenius norm.

In Lemma 3 derived earlier in Chapter 2, we derived a bound on the metric entropy of the set $\mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})$ as:

$$\log N\Big(\epsilon, \mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)}), |\!|\!| \cdot |\!|\!|_{\text{F}}\Big) \leq 16\frac{d^2}{\epsilon^2}\big(\log\frac{d}{\epsilon}\big)^2,$$

for any $\epsilon > 0$ and $\ell \in [k]$. Substituting this bound on the metric entropy into the Dudley bound (5.20) yields

$$\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}] \leq c' \max\{\overline{\rho}(M_0),k\} d \log^2 d.$$

The inequality $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] \leq \frac{1}{p_{\text{obs}}}\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}]$ then yields the claimed result.

Bounding the tail probability of $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)$: In order to establish the claimed tail bound, we use a Bernstein-type bound on the supremum of empirical processes due to Klein and Rio [129, Theorem 1.1c], which we state in a simplified form here.

Lemma 28. *Let $X := (X_1, \ldots, X_m)$ be any sequence of zero-mean, independent random variables, each taking values in $[-1,1]$. Let $\mathcal{V} \subset [-1,1]^m$ be any measurable set of m-length vectors. Then for any $u > 0$, the supremum $X^\dagger = \sup_{v\in\mathcal{V}}\langle X,\, v\rangle$ satisfies the upper tail bound*

$$\mathbb{P}\big(X^\dagger > \mathbb{E}[X^\dagger] + u\big) \leq \exp\Big(\frac{-u^2}{2\sup_{v\in\mathcal{V}}\mathbb{E}[\langle v,\, X\rangle^2] + 4\mathbb{E}[X^\dagger] + 3u}\Big).$$

We now call upon Lemma 28 setting $\mathcal{V} = \{M_{\text{DIFF}} \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)};M_0) \mid |\!|\!|M_{\text{DIFF}}|\!|\!|_{\text{F}} \leq \delta\}$, $X = W'$, and $X^\dagger = p_{\text{obs}} Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)$. The entries of the matrix W' are mutually independent, have a mean of zero, and are bounded by 1 in absolute value. Then we have $\mathbb{E}[X^\dagger] = p_{\text{obs}}\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)]$ and $\mathbb{E}[\langle\!\langle M_{\text{DIFF}},\; W'\rangle\!\rangle^2] \leq 4p_{\text{obs}}|\!|\!|M_{\text{DIFF}}|\!|\!|_{\text{F}}^2 \leq 4p_{\text{obs}}\delta^2$ for every $M_{\text{DIFF}} \in \mathcal{V}$. With these assignments, and some algebraic manipulations, we obtain that for every $u > 0$,

$$\mathbb{P}\big(Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] + u\big) \leq \exp\Big(\frac{-u^2 p_{\text{obs}}}{8\delta^2 + 4\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] + 3u}\Big),$$

as claimed.

[4] Here we use $(\Delta\epsilon)$ to denote the differential of ϵ, so as to avoid confusion with the number of columns d.

5.5.2 Proof of Theorem 14(b): Oracle lower bound

Assume without loss of generality that $d \geq n$. Throughout the proof, we ignore floor and ceiling conditions as these are not critical to the proof and affect the lower bound by only a constant factor.

The Gilbert-Varshamov bound [90, 258] from coding theory guarantees existence of

$$\eta := \exp\Big(c(dr + p_{\text{obs}}\epsilon^2)\Big)$$

binary vectors $g^1, \ldots, g^\eta$, each of length $(dr + p_{\text{obs}}\epsilon^2)$, such that the Hamming distance between any pair of vectors in this set is lower bounded as

$$D_{\text{H}}(g^\ell, g^{\ell'}) \geq \frac{dr + p_{\text{obs}}\epsilon^2}{10}.$$

For some $\delta \in (0, \frac{1}{4})$ whose value is specified later, define a related set of vectors $\widetilde{g}^1, \ldots, \widetilde{g}^\eta$ as

$$\widetilde{g}^\ell_j = \begin{cases} \frac{1}{2} + \delta & \text{if } g^\ell_j = 1 \\ \frac{1}{2} - \delta & \text{if } g^\ell_j = 0, \end{cases}$$

for every $\ell \in [\eta]$ and $j \in [dr + p_{\text{obs}}\epsilon^2]$. Next define a set of "low rank" matrices $G^1, \ldots, G^\eta \in [0,1]^{n \times d}$ where the matrix G^ℓ is obtained as follows. For each $\ell \in [\eta]$, arrange the first rd entries of vector $\widetilde{g}^\ell$ as the entries of an $(r \times d)$ matrix—this arrangement may be done in an arbitrary manner as long as it is consistent across every $\ell \in [\eta]$. Now append a $(\frac{p_{\text{obs}}\epsilon^2}{d} \times d)$ matrix at the bottom, whose entries comprise the last $p_{\text{obs}}\epsilon^2$ entries of the vector $\widetilde{g}^\ell$—again, this arrangement may be done in an arbitrary manner as long as it is consistent across every $\ell \in [\eta]$. Now stack $\frac{1}{p_{\text{obs}}}$ copies of the resulting $\big((r + \frac{p_{\text{obs}}\epsilon^2}{d}) \times d\big)$ matrix on top of each other to form a $\big((\frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}) \times d\big)$ matrix. Note that our assumption $\epsilon^2 + \frac{r \max\{n,d\}}{p_{\text{obs}}} \leq nd$, along with the assumption $d \geq n$, implies that $n \geq \frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}$. Append $(n - (\frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}))$ rows of all zeros at the bottom of this matrix, and denote the resultant $(n \times d)$ matrix as G^ℓ.

We now show that $G^\ell \in B^{\text{N}}(r, \epsilon)$ for every $\ell \in [\eta]$, that is, we show that the matrix $G^\ell \in [0,1]^{n \times d}$ can be decomposed into a sum of a low-rank matrix (of non-negative rank at most r) and a sparse matrix (number of non-zero entries at most ϵ^2). First we set to zero the entries in G^ℓ which correspond to the last $p_{\text{obs}}\epsilon^2$ entries of the vector $\widetilde{g}^\ell$. Let us denote the resulting matrix as $\widetilde{G}^\ell$. Each row of the matrix $\widetilde{G}^\ell$ is either all zero or is identical to one among the first r rows of G^ℓ. Consequently we have $\overline{r}(\widetilde{G}^\ell) \leq r$. Also observe that in the matrix $(G^\ell - \widetilde{G}^\ell)$, the number of non-zero entries is at most $\frac{1}{p_{\text{obs}}} \times p_{\text{obs}}\epsilon^2 = \epsilon^2$, and furthermore, each of these entries lie in the interval $[0,1]$. Hence we have $|\!|\!|G^\ell - \widetilde{G}^\ell|\!|\!|_{\text{F}}^2 \leq \epsilon^2$. The matrix G^ℓ thus satisfies all the requirements for membership in the set $B^{\text{N}}(r, \epsilon)$.

For every $\ell \in [\eta]$, let $\mathbb{P}^\ell$ denote the probability distribution of the matrix Y obtained by setting $M^* = G^\ell$. One can verify that the set of matrices $G^1, \ldots, G^\eta$ constructed above has the following two properties, for every pair $\ell \neq \ell' \in [\eta]$:

$$D_{\mathrm{KL}}(\mathbb{P}^{\ell} \| \mathbb{P}^{\ell'}) \leq c'\delta^2 p_{\mathrm{obs}} \big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \big),$$

and

$$|\!|\!| G^{\ell} - G^{\ell'} |\!|\!|_{\mathrm{F}}^2 \geq \frac{\delta^2}{10} \big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \big).$$

Substituting these relations in Fano's inequality [54] yields that when M^* is drawn uniformly at random from the set $\{G^1, \ldots, G^{\eta}\}$, any estimate $\widehat{M}$ for M^* incurs an error lower bounded as

$$\mathbb{E}[|\!|\!| \widehat{M} - M^* |\!|\!|_{\mathrm{F}}^2] \geq \frac{\delta^2}{20} \Big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \Big) \Big(1 - \frac{c'\delta^2 p_{\mathrm{obs}} \big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \big) + \log 2}{c(dr + p_{\mathrm{obs}}\epsilon^2)} \Big) \overset{(i)}{\geq} c'' \big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \big),$$

where inequality (i) is obtained by choosing δ^2 as a small enough constant (that depends only on c and c'). Recalling our assumption $d \geq n$, and consequently replacing d by $\max\{n, d\}$ in the bound yields the claimed result.

5.5.3 Proof of Theorem 15: SVT Estimator

This proof builds on the framework of the proof of a result in our earlier work [221, Theorem 2] corresponding to the case of $n = d$ and $\rho = 1$. We introduce certain additional tricks in order to generalize the proof for general values of ρ and to obtain a sharp dependence on ρ.

Assume without loss of generality that $n \leq d$.

Recall from earlier (5.10a) that we can write our observation model as $Y' = M^* + \frac{1}{p_{\mathrm{obs}}} W'$, where $W' \in [-1, 1]^{n \times d}$ is a zero-mean matrix with mutually independent entries. The distribution (5.10b) of the entries is reproduced here for convenience:

$$[W']_{ij} = \begin{cases} p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) + \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}[M^*]_{ij} \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) - \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}(1 - [M^*]_{ij}) \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) & \text{with probability } 1 - p_{\mathrm{obs}}. \end{cases} \tag{5.21}$$

For any matrix X, let $\sigma_1(X), \sigma_2(X), \ldots$ denote its singular values in descending order.

Our proof of the upper bound hinges upon the following three lemmas. The first lemma is Lemma 4 from Chpater 2 which states that if $\lambda \geq \frac{1.01 |\!|\!| W' |\!|\!|_{\mathrm{op}}}{p_{\mathrm{obs}}}$, then

$$|\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\mathrm{F}}^2 \leq c \sum_{j=1}^{n} \min \big\{ \lambda^2, \sigma_j^2(M^*) \big\}$$

with probability at least $1 - c_1 e^{-c'n}$, where c, c_1 and c' are positive universal constants.

Our second lemma is an approximation-theoretic result that bounds the tail of the singular values of any matrix with a given permutation-rank or non-negative rank. The proof of this lemma builds on a construction due to Chatterjee [44] with several additional techniques in order to obtain a result that is sharp enough for our purposes.

Lemma 29. *(a) For any matrix $M \in \mathbb{C}_{PR}(\rho)$ and any $s \in \{1, 2, \ldots, n-1\}$, we have*

$$\sum_{j=s+1}^{n} \sigma_j^2(M) \leq \frac{nd\rho^2}{s}.$$

(b) For any matrix $M \in \mathbb{C}_{NR}(r)$ and any $s \in \{1, 2, \ldots, n-1\}$, we have

$$\sum_{j=s+1}^{n} \sigma_j^2(M) \leq nd \max\big\{\frac{r-s}{r}, 0\big\}.$$

Our third lemma controls the noise term W'.

Lemma 30. *The operator norm of the noise matrix W' distributed as* (5.21) *is upper bounded as*

$$\mathbb{P}\big(|\!|\!| W' |\!|\!|_{op} > 2.01\sqrt{p_{\text{obs}}(n+d)}\big) \leq e^{-c' \max\{n,d\}}.$$

Based on these three lemmas, we now complete the proof of the theorem. From Lemma 30 we see that the choice $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ guarantees that $\lambda \geq \frac{1.01|\!|\!| W'|\!|\!|_{\text{op}}}{p_{\text{obs}}}$ with probability at least $1 - e^{-c' \max\{n,d\}}$. Consequently, the condition required for an application of Lemma 4 is satisfied, and applying this lemma then yields the upper bound

$$|\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\text{F}}^2 \leq c \sum_{j=1}^{n} \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M^*)\big\}$$

with probability at least $1 - e^{-c' \max\{n,d\}}$. Applying Lemma 29 yields that with probability at least $1 - e^{-c' \max\{n,d\}}$, it must be that

$$|\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\text{F}}^2 \leq c \min_{s \in [n]} \big(\frac{sd}{p_{\text{obs}}} + \frac{\rho^2 nd}{s}\big), \qquad \text{if } M^* \in \mathbb{C}_{\text{PR}}(\rho),$$

and

$$|\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\text{F}}^2 \leq c \min_{s \in [n]} \big(\frac{sd}{p_{\text{obs}}} + nd \max\big\{1 - \frac{s}{r}, 0\big\}\big), \qquad \text{if } M^* \in \mathbb{C}_{\text{NR}}(r).$$

For the case when $M^* \in \mathbb{C}_{\text{PR}}(\rho)$, setting $s = \lceil \rho\sqrt{p_{\text{obs}} n} \rceil$ and performing some algebra shows that

$$\mathbb{P}\Big[\frac{1}{nd} |\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\text{F}}^2 > \frac{c_1 \rho}{\sqrt{p_{\text{obs}} d}}\Big] \leq e^{-c' \max\{n,d\}}.$$

When $M^* \in \mathbb{C}_{\mathrm{NR}}(r)$, setting $s = r$ shows that

$$\mathbb{P}\Big[\frac{1}{nd} |\!|\!| \widehat{M}_{\text{SVT}} - M^* |\!|\!|_{\mathrm{F}}^2 > \frac{c_1 r}{p_{\mathrm{obs}} d}\Big] \leq e^{-c' \max\{n,d\}}.$$

Recalling our assumption that $d \geq n$ and substituting $n = \min\{n, d\}$ and $d = \max\{n, d\}$ yields the claimed result.

Proof of Lemma 5 Part (a): Without loss of generality, assume that $d \geq n$.

We begin with an upper bound on the tail of the singular values of any matrix in $\mathbb{C}_{\mathrm{PR}}(1)$, that is, that has a permutation-rank of 1. The proof of this bound uses a construction due to Chatterjee [44] for a rank $\widetilde{s}$ approximation of any matrix in $\mathbb{C}_{\mathrm{PR}}(1)$, for any value $\widetilde{s} \in [n]$. We first reproduce Chatterjee's construction.

For a given matrix $M \in \mathbb{C}_{\mathrm{PR}}(1)$, define the vector $\tau \in \mathbb{R}^d$ of column sums—namely, with entries $\tau_j = \sum_{i=1}^n [M]_{ij}$ for $j \in [d]$. Using this vector, define a rank $\widetilde{s}$ approximation $\widetilde{M}$ to M by grouping the columns according to the vector τ according to the following procedure:

- Observing that each $\tau_j \in [0, n]$, divide the full interval $[0, n]$ into $\widetilde{s}$ groups—say of the form $[0, n/\widetilde{s}), [n/\widetilde{s}, 2n/\widetilde{s}), \ldots [(\widetilde{s}-1)n/\widetilde{s}, n]$. If τ_j falls into the interval α for some $\alpha \in [\widetilde{s}]$, then map column j to the group G_α of indices.

- For each $\alpha \in [\widetilde{s}]$ such that group G_α is non-empty, choose a particular column index $j' =\in G_\alpha$ in an *arbitrary* fashion. For every other column index $j \in G_\alpha$, set $\widetilde{M}_{ij} = M_{ij'}$ for all $i \in [n]$.

By construction, the matrix $\widetilde{M}$ has at most $\widetilde{s}$ distinct rows, and hence rank at most $\widetilde{s}$. Now consider any column $j \in [d]$ and suppose that $j \in G_\alpha$. Let j' denote the column chosen for the group G_α in the second step of the construction. Since $M \in \mathbb{C}_{\mathrm{PR}}(1)$, we must either have $M_{ij} \geq M_{ij'} = \widetilde{M}_{ij}$ for every $i \in [n]$, or $M_{ij} \leq M_{ij'} = \widetilde{M}_{ij}$ for every $i \in [n]$. Then we are guaranteed that

$$\sum_{i=1}^n |\widetilde{M}_{ij} - M_{ij}| =| \sum_{i=1}^n (\widetilde{M}_{ij} - M_{ij}) |= |\tau_{j'} - \tau_j| \leq \frac{n}{\widetilde{s}}, \tag{5.22}$$

where we have used the fact the pair $(\tau_j, \tau_{j'})$ must lie in an interval of length at most $n/\widetilde{s}$. This completes the description of Chatterjee's construction.

In what follows, we use Chatterjee's result in order to obtain our claimed bound on the tail of the spectrum of any matrix $M \in \mathbb{C}_{\mathrm{PR}}(\rho)$. We modify the result in a specific critical manner that allows us to obtain the desired dependence on the parameter ρ. Recall that any matrix $M \in \mathbb{C}_{\mathrm{PR}}(\rho)$ can be decomposed as

$$M = \sum_{\ell=1}^{\rho} M^{(\ell)},$$

for some matrices $M^{(1)}, \ldots, M^{(\rho)} \in \mathbb{C}_{\mathrm{PR}}(1)$. For every $\ell \in [r]$, let $\widetilde{M}^{(\ell)}$ be a rank $\widetilde{s} = \frac{s}{\rho}$ approximation of $M^{(\ell)}$ obtained from Chatterjee's construction above, but with the following additional detail. Observe that in Chatterjee's construction, the choice of column j' from group G_α is arbitrary. For our construction, we will make a specific choice of this column: we choose the column whose entries have the smallest values among all columns in the group G_α. With this choice, we have the property

$$\widetilde{M}^{(\ell)}_{ij} \leq M^{(\ell)}_{ij} \qquad \text{for every } \ell \in [\rho],\ i \in [n], j \in [d]. \tag{5.23}$$

Now let $\widetilde{M} := \sum_{\ell=1}^{\rho} \widetilde{M}^{(\ell)}$. Since every entry of every matrix $\widetilde{M}^{(\ell)}$ is non-negative, we have that every entry of $\widetilde{M}$ is also non-negative. We also claim that

$$\widetilde{M}_{ij} = \sum_{\ell=1}^{\rho} \widetilde{M}^{(\ell)}_{ij} \overset{(i)}{\leq} \sum_{\ell=1}^{\rho} M^{(\ell)}_{ij} = M_{ij} \leq 1,$$

where the inequality (i) is a consequence of the set of inequalities (5.23). Thus we have that $\widetilde{M} \in [0,1]^{n\times d}$, and that the rank of $\widetilde{M}$ is at most $\rho\widetilde{s}$. This result then yields the bound

$$\sum_{j=\rho\widetilde{s}+1}^{n} \sigma_j^2(M) \leq |\!|\!|M - \widetilde{M}|\!|\!|_{\mathrm{F}}^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{d} |M_{ij} - \widetilde{M}_{ij}|.$$

Applying the triangle inequality, we further bound this quantity as

$$\sum_{j=\rho\widetilde{s}+1}^{n} \sigma_j^2(M) \leq \sum_{i=1}^{n}\sum_{j=1}^{d} |\sum_{\ell=1}^{\rho}(M^{(\ell)}_{ij} - \widetilde{M}^{(\ell)}_{ij})| \leq \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{\ell=1}^{\rho} |M^{(\ell)}_{ij} - \widetilde{M}^{(\ell)}_{ij}| \overset{(i)}{\leq} \frac{\rho n d}{\widetilde{s}} = \frac{\rho^2 n d}{s},$$

where the bound (i) follows from (5.22), and equation (ii) is a result of our choice $\widetilde{s} = \frac{s}{\rho}$.

Part (b): This result follows directly from the facts that the rank of M is at most r, and the square of its Frobenius norm is at most nd.

Proof of Lemma 30 Define an $((n+d)\times(n+d))$ matrix W'' as

$$W'' = \frac{1}{\sqrt{p_{\mathrm{obs}}}} \begin{bmatrix} 0 & W' \\ (W')^T & 0 \end{bmatrix}.$$

From (5.21) and the construction above, we have that the matrix W'' is symmetric, with mutually independent entries above the diagonal that have a mean of zero and a variance upper bounded by 1. Consequently, known results in random matrix theory (e.g., see [44, Theorem 3.4] or [250, Theorem 2.3.21]) yield the bound $|\!|\!|W''|\!|\!|_{\mathrm{op}} \leq 2.01\sqrt{n+d}$ with probability at least $1 - e^{-c\max\{n,d\}}$. One can also verify that $|\!|\!|W''|\!|\!|_{\mathrm{op}} = \frac{1}{\sqrt{p_{\mathrm{obs}}}}|\!|\!|W'|\!|\!|_{\mathrm{op}}$, yielding the claimed result.

5.5.4 Proof of Proposition 5: Every rank is possible

We recall that for any integer $k \geq 0$, the notation J_k denotes an upper triangular matrix of size $(k \times k)$ with all entries on and above the diagonal set as 1, and I_k denotes the identity matrix of size $(k \times k)$. We also recall the following block matrix M, of size $(n \times d)$, defined in the statement of the proposition:

$$M = \begin{bmatrix} J_{r-\rho+1} & 0 & 0 \\ 0 & I_{\rho-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In the remainder of the proof, we show that $\overline{r}(M) = r$ and $\overline{\rho}(M) = \rho$. Using the ideas in the construction of M and the associated proof to follow, one can construct many other matrices that have the non-negative rank and the permutation-rank equaling r and ρ respectively, for any given value $1 \leq \rho \leq r \leq \min\{n, d\}$.

We partition the proof into four parts.

Proof of $\overline{r}(M) \leq r$: One can write M as a sum of r matrices, each having a non-negative rank of one: for each non-zero row, consider a component matrix comprising that row and zeros elsewhere. Consequently, we also have $\overline{r}(M) \leq r$. We have thus established that the non-negative rank of this matrix equals exactly r.

Proof of $\overline{r}(M) \geq r$: Towards the claim regarding the non-negative rank, observe that the (normal) rank of M^* equals r. Since the rank of any matrix is a lower bound on its non-negative rank, we have that $\overline{r}(M) \geq r$.

Proof of $\overline{\rho}(M) \leq \rho$: Observe that the $(n \times d)$ matrix with $J_{r-\rho+1}$ as its top-left submatrix and 0 elsewhere has a permutation-rank of 1. Moreover, any $(n \times d)$ matrix with exactly one entry as 1 and the remaining entries 0 also has a permutation-rank of 1, and hence a $(n \times d)$ matrix with $I_{\rho-1}$ as its submatrix and zeros elsewhere has a permutation-rank of at most $(\rho - 1)$. Putting these arguments together, we obtain the bound $\overline{\rho}(M) \leq \rho$.

Proof that $\overline{\rho}(M) \geq \rho$: Suppose that $M = \sum_{\ell=1}^{\overline{\rho}(M)} M^{(\ell)}$ for some $M^{(1)}, \ldots, M^{(\rho)} \in \mathbb{C}_{\text{PR}}(1)$.

First observe that for every $\ell \in [M^{(\rho)}]$, we have $M^{(\ell)} \in \mathbb{C}_{\text{PR}}(1) \subset [0,1]^{n \times d}$, and therefore $M_{ij}^{(\ell)} \geq 0$. This observation then implies that for any $(i,j) \in [n] \times [d]$, we must have the relation

$$M_{ij} = 0 \quad \Rightarrow \quad M_{ij}^{(\ell)} = 0 \quad \text{for every } \ell \in [\overline{\rho}(M)].$$

Secondly, observe that the matrix

$$I_{2\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

does not belong to $\mathbb{C}_{\text{PR}}(1)$. It

It follows that any matrix containing $I_{2\times 2}$ as a submatrix cannot belong to the set $\mathbb{C}_{\text{PR}}(1)$. It further follows that for any positive integer k, the matrix $I_{k\times k}$ must have a permutation rank of at least k. Finally, observe that the matrix M defined earlier contains $I_{\rho\times\rho}$ as its submatrix, given by the intersection of rows $\{r-\rho,\ldots,r\}$ with the columns $\{r-\rho,\ldots,r\}$. It follows that M must have a permutation rank of at least ρ, thereby proving the claim.

5.5.5 Proof of Proposition 6: Bias of non-negative-rank fitting estimator

We assume for ease of exposition that r divides n and d. Otherwise, since $r \leq \frac{1}{2}\min\{n,d\}$, one may take floors or ceilings and this will change the result only by a constant factors.

First consider the block matrix $M \in [0,1]^{\frac{n}{r}\times\frac{d}{r}}$:

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \tag{5.24}$$

where each of the four blocks is of size $(\frac{n}{2r} \times \frac{d}{2r})$. The following lemma shows that the best rank-1 approximation to M has a large approximation error:

Lemma 31. *For the matrix M defined in* (5.24)*, for any vectors $u^{()} \in \mathbb{R}^n$ and $v^{()} \in \mathbb{R}^d$, it must be that*

$$|\!|\!| M - u^{()}v^{()^T} |\!|\!|_F^2 \geq c\frac{nd}{r^2},$$

where $c > 0$ is a universal constant.

We use the matrix M defined in (5.24) to build the following block matrix $M' \in \mathbb{C}_{\text{PR}}(r)$:

$$M' := \begin{bmatrix} M & 0 & \cdots & 0 \\ 0 & M & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M \end{bmatrix}.$$

In words, the matrix M' is a block-diagonal matrix where the diagonal has r copies of M.

Due to the block diagonal structure of M', the singular values of M' are simply r copies of the singular values of its constituent matrix M. Consequently, we have that for any matrix $\widetilde{M} \in \mathbb{C}_{\text{NR}}(r)$:

$$|\!|\!| \widetilde{M} - M' |\!|\!|_{\text{F}}^2 \geq r(|\!|\!| M |\!|\!|_{\text{F}}^2 - |\!|\!| M |\!|\!|_{\text{op}}^2) \overset{(i)}{\geq} c\frac{nd}{r},$$

as claimed, where the inequality (i) is a consequence of Lemma 31.

Proof of Lemma 31 Consider any value $i \in [\frac{n}{2r}]$ and $j \in [\frac{d}{2r}]$. Then we claim that

$$(M_{i,j} - [u^{()}v^{()^T}]_{i,j})^2 + (M_{i+\frac{n}{2r},j} - [u^{()}v^{()^T}]_{i+\frac{n}{2r},j})^2 + (M_{i,j+\frac{d}{2r}} - [u^{()}v^{()^T}]_{i,j+\frac{d}{2r}})^2 \\ + (M_{i+\frac{n}{2r},j+\frac{d}{2r}} - [u^{()}v^{()^T}]_{i+\frac{n}{2r},j+\frac{d}{2r}})^2 \geq 0.01. \tag{5.25}$$

If not, then for the choice of M in (5.24), we must have $[u^{()}v^{()^T}]_{i,j} \in (0.9, 1.1)$, $[u^{()}v^{()^T}]_{i+\frac{n}{2r},j} \in (0.9, 1.1)$, $[u^{()}v^{()^T}]_{i,j+\frac{d}{2r}} \in (0.9, 1.1)$ and $[u^{()}v^{()^T}]_{i+\frac{n}{2r},j+\frac{d}{2r}} < 0.1$. However, since $[u^{()}v^{()^T}]_{i',j'} = u^{()}_{i'}v^{()}_{j'}$ for every coordinate (i', j'), we also have

$$[u^{()}v^{()^T}]_{i,j} \times [u^{()}v^{()^T}]_{i+\frac{n}{2r},j+\frac{d}{2r}} = [u^{()}v^{()^T}]_{i+\frac{n}{2r},j} \times [u^{()}v^{()^T}]_{i,j+\frac{d}{2r}},$$

which contradicts the required ranges of the individual coordinates.

Summing the bound (5.25) over all values of $i \in [\frac{n}{2r}]$ and $j \in [\frac{d}{2r}]$ yields the claimed result.

5.5.6 Proof of Proposition 7: No "good" convex approximation

Consider any set $\mathbb{S}$ and any convex set $\mathbb{C}$. We begin with a key lemma that establishes a relation between the Hausdorff distance of $\mathbb{S}$ from $\mathbb{C}$ and a proposed notion of the inherent convexity of $\mathbb{S}$.

Lemma 32. *For any set $\mathbb{S} \subseteq [0,1]^{n \times d}$ and any convex set $\mathbb{C} \subseteq [0,1]^{n \times d}$, it must be that*

$$\max\Big\{ \sup_{M \in \mathbb{S}} \inf_{M' \in \mathbb{C}} |\!|\!| M - M' |\!|\!|_F^2 \ , \ \sup_{M' \in \mathbb{C}} \inf_{M \in \mathbb{S}} |\!|\!| M - M' |\!|\!|_F^2 \Big\} \\ \geq \frac{2}{9} \sup_{M_1 \in \mathbb{S},\ M_2 \in \mathbb{S}} \inf_{M_0 \in \mathbb{S}} |\!|\!| \frac{1}{2}(M_1 + M_2) - M_0 |\!|\!|_F^2. \tag{5.26}$$

See the end of this section for a proof of this claim.

The left hand side of inequality (5.26) is the Hausdorff distance between the sets $\mathbb{S}$ and $\mathbb{C}$ in terms of the squared Frobenius norm The right hand side of the inequality represents a notion of the inherent convexity of the set $\mathbb{S}$.

With this lemma in place, we now complete the remainder of the proof. To this end, we set $\mathbb{S} = \mathbb{C}_{\text{PR}}(1)$, and let $\mathbb{C}$ be any convex set of $[0,1]$-valued $(n \times d)$ matrices.

We now construct a pair of matrices $M_1 \in \mathbb{C}_{\text{PR}}(1)$ and $M_2 \in \mathbb{C}_{\text{PR}}(1)$ that we use to lower bound the right hand side of (5.26). Define $M_1 \in \mathbb{C}_{\text{PR}}(1)$ as

$$[M_1]_{ij} = \begin{cases} 1 & \text{if } i \leq \frac{n}{2},\ j \leq \frac{d}{2} \\ 0 & \text{otherwise,} \end{cases}$$

and matrix $M_2 \in \mathbb{C}_{\text{PR}}(1)$ as

$$[M_2]_{ij} = \begin{cases} 1 & \text{if } i > \frac{n}{2},\ j > \frac{d}{2} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the entries of the matrix $\frac{1}{2}(M_1 + M_2)$ are given by:

$$[\frac{1}{2}(M_1 + M_2)]_{ij} = \begin{cases} \frac{1}{2} & \text{if } (i \leq \frac{n}{2},\ j \leq \frac{d}{2}) \text{ or } (i > \frac{n}{2},\ j > \frac{d}{2}) \\ 0 & \text{otherwise.} \end{cases}$$

Now consider any pair of integers $(i, j) \in [\lfloor n/2 \rfloor] \times [\lfloor d/2 \rfloor]$. Then the (2×2) submatrix of $\frac{1}{2}(M_1 + M_2)$ formed by its entries (i, j), $(i + \lceil n/2 \rceil, j)$, $(i, j + \lceil d/2 \rceil)$ and $(i + \lceil n/2 \rceil, j + \lceil d/2 \rceil)$ equals

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

It is easy to verify that there is a constant $c > 0$ such that the squared Frobenius norm distance between this rescaled identity matrix and any (2×2) matrix in $\mathbb{C}_{\text{PR}}(1)$ is at least c. Since this argument holds for any choice of $(i, j) \in [\lfloor n/2 \rfloor] \times [\lfloor d/2 \rfloor]$, summing up the errors across each of these sets of entries yields

$$|\!|\!| \frac{1}{2}(M_1 + M_2) - M |\!|\!|_{\text{F}}^2 \geq c' nd, \qquad \text{for every matrix } M \in \mathbb{C}_{\text{PR}}(1),$$

where $c' > 0$ is a universal constant. Finally, substituting this bound in Lemma 32 yields the claimed result.

It remains to prove Lemma 32.

Proof of Lemma 32. For the given sets $\mathbb{S}$ and $\mathbb{C}$, let d_{Haus} denote the Hausdorff distance between the two sets in the squared Frobenius norm, that is,

$$d_{\text{Haus}}(\mathbb{S}, \mathbb{C}) := \max \Big\{ \sup_{M \in \mathbb{S}} \inf_{M' \in \mathbb{C}} |\!|\!| M - M' |\!|\!|_{\text{F}}^2 \ ,\ \sup_{M' \in \mathbb{C}} \inf_{M \in \mathbb{S}} |\!|\!| M - M' |\!|\!|_{\text{F}}^2 \Big\}$$

Consider any matrices $M_1 \in \mathbb{S}$ and $M_2 \in \mathbb{S}$. From the definition of d_{Haus}, we know that there exist matrices $\widetilde{M_1} \in \mathbb{C}$ and $\widetilde{M_2} \in \mathbb{C}$ such that

$$|\!|\!| M_i - \widetilde{M_i} |\!|\!|_{\text{F}}^2 \leq d_{\text{Haus}}(\mathbb{S}, \mathbb{C}), \quad \text{for } i \in \{1, 2\}. \tag{5.27}$$

Since $\mathbb{C}$ is a convex set, we also have $\frac{1}{2}(\widetilde{M_1} + \widetilde{M_2}) \in \mathbb{C}$. Then from the definition of d_{Haus}, we also know that there exists a matrix $M_0 \in \mathbb{S}$ such that

$$|\!|\!| \frac{1}{2}(\widetilde{M_1} + \widetilde{M_2}) - M_0 |\!|\!|_{\text{F}}^2 \leq d_{\text{Haus}}(\mathbb{S}, \mathbb{C}). \tag{5.28}$$

Finally, applying the triangle inequality to the bounds (5.27) and (5.28) yields

$$\begin{aligned} |\!|\!| \frac{1}{2}(M_1 + M_2) - M_0 |\!|\!|_{\text{F}}^2 &\leq 3 |\!|\!| \frac{1}{2}(\widetilde{M_1} + \widetilde{M_2}) - M_0 |\!|\!|_{\text{F}}^2 + \frac{3}{4} |\!|\!| M_1 - \widetilde{M_1} |\!|\!|_{\text{F}}^2 + \frac{3}{4} |\!|\!| M_2 - \widetilde{M_2} |\!|\!|_{\text{F}}^2 \\ &\leq \frac{9}{2} d_{\text{Haus}}(\mathbb{S}, \mathbb{C}). \end{aligned}$$

5.5.7 Proof of Proposition 8: Necessary condition for unique decomposition

Suppose there exists a coordinate pair (i, j) such the stated condition is violated. Then there must exist two distinct values $\ell_1 \in [\rho]$ and $\ell_2 \in [\rho]$ that satisfy the following three conditions:
(a) $M_{ij}^{(\ell_1)} > 0$ and $M_{ij}^{(\ell_2)} > 0$,
(b) $M_{ij}^{(\ell_1)}$ is distinct from all other entries in $M^{(\ell_1)}$, and
(c) $M_{ij}^{(\ell_2)}$ is distinct from all other entries in $M^{(\ell_2)}$.

In addition, the fact that $M_{ij}^{(\ell_1)} + M_{ij}^{(\ell_2)} \in (0, 1)$ for every coordinate (i, j), along with condition (a) above, imply a fourth condition:
(d) $M_{ij}^{(\ell_1)} \in (0, 1)$ and $M_{ij}^{(\ell_2)} \neq 1$.

Now, conditions (b)–(d) in tandem imply the existence of some value $\epsilon > 0$ such that *all* of the following properties hold:
(i) $(M_{ij}^{(\ell_1)} + \epsilon) \in [0, 1]$,
(ii) $(M_{ij}^{(\ell_2)} - \epsilon) \in [0, 1]$, and
(iii) setting the $(i, j)^{th}$ entries of the matrices $M^{(\ell_1)}$ and $M_{ij}^{(\ell_2)}$ as $(M_{ij}^{(\ell_1)} + \epsilon)$ and $(M_{ij}^{(\ell_2)} - \epsilon)$ respectively does not change the ordering of the entries within matrices $M^{(\ell_1)}$ and $M^{(\ell_2)}$.

As a result of properties (i)–(iii), the decomposition where $M^{(\ell_1)}$ and $M^{(\ell_2)}$ are replaced by these new matrices with replaced $(i, j)^{th}$ entry is a different, valid permutation-rank decomposition of M^*.

5.A Appendix: Alternative interpretation of the non-negative rank model

In the non-negative rank model described in the introduction, one may wonder why the affinity of a user to a movie conditioned on a feature must be modeled as the product $u_i^{(\ell)} v_j^{(\ell)}$, where $u_i^{(\ell)}$ is user i's affinity for feature ℓ and $v_j^{(\ell)}$ is movie j's connection to feature ℓ. Secondly, one may also wonder why the net affinity of a user to a movie is the sum of the affinities across the features $\sum_{\ell=1}^{r} u_i^{(\ell)} v_j^{(\ell)}$. These two modeling assumptions may sometimes be confusing, and hence in what follows, we present an alternative interpretation of the low non-negative rank model for the recommender systems application.

Consider any feature $\ell \in [r]$. The affinities of users towards movies conditioned on this feature is a matrix $X^{(\ell)} \in [0, 1]^{n \times d}$, whose $(i, j)^{\text{th}}$ entry $X_{ij}^{(\ell)}$ is the probability that user i likes movie j when asked to judge only based on feature ℓ. The matrix $X^{(\ell)}$ is assumed to have a (non-negative) rank of 1.

Now, every user is assumed to have his/her own way of weighing features to decide which movies he/she likes. Specifically, any user $i \in [n]$ is associated to values $\alpha_i^{(1)}, \ldots, \alpha_i^{(r)}$ such that $\alpha_i^{(\ell)} \geq 0$ for every $\ell \in [r]$ and $\sum_{\ell=1}^{r} \alpha_i^{(\ell)} = 1$, where for any i and ℓ, the value $\alpha_i^{(\ell)}$

represents the weight that user i puts on feature ℓ. Then the probability that user i likes any movie j is assumed to be the convex combination

$$\sum_{\ell=1}^{r} \alpha_i^{(\ell)} X_{ij}^{(\ell)}.$$

This completes the description of the model. Let us now verify that the resulting user-movie matrix has a non-negative rank of r. Since $X^{(\ell)}$ has a non-negative rank of 1, we can write $X^{(\ell)} = u^{(\ell)}(v^{(\ell)})^T$ for some vectors $u^{(\ell)}$ and $v^{(\ell)}$. Then the i^{th} row of the net user-movie matrix equals $\sum_{\ell=1}^{r} \alpha_i^{(\ell)} u_i^{(\ell)} (v^{(\ell)})^T$, and hence the net user-movie matrix equals

$$\sum_{\ell=1}^{r} \widetilde{u}^{(\ell)}(v^{(\ell)})^T, \qquad \text{where} \quad \widetilde{u}^{(\ell)} = \begin{bmatrix} \alpha_1^{(\ell)} u_1^{(\ell)} \\ \vdots \\ \alpha_n^{(\ell)} u_n^{(\ell)} \end{bmatrix}.$$

This completes the alternative description of the non-negative rank model.

One can observe that the restriction $\sum_{\ell=1}^{r} \alpha_i^{(\ell)} = 1$ makes this model slightly more restrictive than the non-negative rank model described earlier in the main text. However, our lower bounds (Theorem 14(b)) on the risk for the non-negative rank model continue to apply to this model as well.

Part II

Incentives: Unique Multiplicative Mechanisms

Chapter 6

Double or Nothing

> *"Life may appear like a gamble. Although it is not very much so."*
>
> – Charles Darwin

6.1 Introduction

Complex machine learning tools such as deep learning are gaining increasing popularity and are being applied to a wide variety of problems. These tools require large amounts of labeled data [38, 62, 101, 205]. These large labeling tasks are being performed by coordinating crowds of semi-skilled workers through the Internet. This is known as crowdsourcing. Generating large labeled data sets through crowdsourcing is inexpensive and fast as compared to employing experts. Furthermore, given the current platforms for crowdsourcing such as Amazon Mechanical Turk and many others, the initial overhead of setting up a crowdsourcing task is minimal. Crowdsourcing as a means of collecting labeled training data has now become indispensable to the engineering of intelligent systems. The crowdsourcing of labels is also often used to supplement automated algorithms, to perform the tasks that are too difficult to accomplish by machines alone [14, 80, 124, 139, 259].

Most workers in crowdsourcing are not experts. As a consequence, labels obtained from crowdsourcing typically have a significant amount of error [119, 260, 261]. It is not surprising that there is significant emphasis on having higher quality labeled data for machine learning algorithms, since a higher amount of noise implies requirement of more labels for obtaining the same accuracy in practice. Moreover, several algorithms and settings are not very tolerant of data that is noisy [10, 99, 153, 157]; for instance, the paper [153] concludes that "a range of different types of boosting algorithms that optimize a convex potential function satisfying mild conditions cannot tolerate random classification noise." Recent efforts have focused on developing statistical techniques to post-process the noisy labels in order to improve its quality (see Chapter 4 for a detailed discussion). However, when the inputs to these

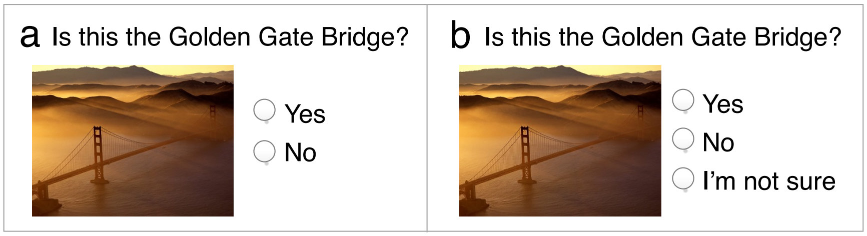

a Is this the Golden Gate Bridge?
○ Yes
○ No
b Is this the Golden Gate Bridge?
○ Yes
○ No
○ I'm not sure

Figure 6.1: Different interfaces for a task that requires the worker to answer the question "Is this the Golden Gate Bridge?": (a) the conventional interface; (b) with an option to skip.

algorithms are very erroneous, it is difficult to guarantee that the processed labels will be reliable enough for subsequent use by machine learning or other applications. In order to avoid "garbage in, garbage out", we take a complementary approach to this problem: cleaning the data at the time of collection.

We consider crowdsourcing settings where the workers are paid for their services, such as in the popular crowdsourcing platforms of Amazon Mechanical Turk (`mturk.com`), Crowdflower (`crowdflower.com`) and other commercial platforms, as well as internal crowdsourcing platforms of companies such as Google, Facebook and Microsoft. These commercial platforms have gained substantial popularity due to their support for a diverse range of tasks for machine learning labeling, varying from image annotation and text recognition to speech captioning and machine translation. We consider problems that are objective in nature, that is, have a definite answer. Figure 6.1a depicts an example of such a question where the worker is shown a set of images, and for each image, the worker is required to identify if the image depicts the Golden Gate Bridge.

Our approach builds on the simple insight that in typical crowdsourcing setups, workers are simply paid in proportion to the amount of tasks they complete. As a result, workers attempt to answer questions that they are not sure of, thereby increasing the error rate of the labels. For the questions that a worker is not sure of, her answers could be very unreliable [108, 119, 260, 261]. To ensure acquisition of only high-quality labels, we wish to encourage the worker to skip the questions about which she is unsure, for instance, by providing an explicit "I'm not sure" option for every question (see Figure 6.1b); we will thus often refer to this setting as the skip-based setting. Given this additional option, one must also ensure that the worker is indeed incentivized to skip the questions that she is not confident about, and to attempt the questions for which she is sure enough. The goal is to design payment mechanisms that incentivize the worker to respond in this manner. As we will see later, this significantly improves the aggregate quality of the labels that are input to the machine learning algorithms. We term any payment mechanism that incentivizes the worker to do so as an "incentive compatible" mechanism.

In addition to incentive compatibility, preventing spammers is another desirable requirement from incentive mechanisms in crowdsourcing. Spammers are workers who answer randomly without regard to the question being asked, in the hope of earning some free money, and are known to exist in large numbers on crowdsourcing platforms [17, 119, 260, 261]. The presence of spammers can significantly affect the performance of any machine learning algorithm that is trained on this data. It is thus of interest to deter spammers by paying them as low as possible. An intuitive objective, to this end, is to ensure a minimum possible payment to spammers who answer randomly. For instance, in a task with binary-choice questions, a spammer is expected to have half of the attempted answers incorrect; one may thus wish to set the payment to its minimum possible value if half or more of the attempted answers are wrong. In this chapter, however, we impose *strictly and significantly weaker requirement*, and then show that there is one and only one incentive-compatible mechanism that can satisfy this weak requirement. Our requirement is referred to as the "no-free-lunch" axiom. It says that if *all* the questions attempted by the worker are answered incorrectly, then the payment must be the minimum possible. We term this condition the "no-free-lunch" axiom.

We propose a payment mechanism for the aforementioned setting ("incentive compatibility" plus "no-free-lunch"), and show that surprisingly, this is the *only* possible mechanism. We also show that additionally, our mechanism makes the smallest possible payment to spammers among all possible incentive compatible mechanisms that may or may not satisfy the no-free-lunch axiom. Interestingly, our payment mechanism takes a multiplicative form: the evaluation of the worker's response to each question is a certain score, and the final payment is a *product* of these scores. This mechanism has additional appealing features in that it is simple to compute, and is also simple to explain to the workers. Our mechanism is applicable to any type of objective questions, including multiple choice annotation questions, transcription tasks, etc. We also demonstrate via empirical evaluations that our theoretical guarantees do translate to practice.

Related literature. The framework of "strictly proper scoring rules" [24, 94, 137, 216] provides a general theory for eliciting information for settings where this information can subsequently be verified by the mechanism designer, for example, by observing the true value some time in the future. In our work, this verification is performed via the presence of some "gold standard" questions in the task. Consequently, our mechanisms can also be called "strictly proper scoring rules". It is important to note that the framework of strictly proper scoring rules, however, provides a large collection of possible mechanisms and does not guide the choice of a specific mechanism from this collection [94]. In this work, we show that for the crowdsourcing setups considered, under a very mild condition we term the "no-free-lunch" axiom, the mechanism proposed in this chapter is the one and only strictly proper scoring rule.

Interestingly, proper scoring rules have another interesting connection with machine learning techniques: to quote the paper [28], "proper scoring rules comprise most loss functions currently in use: log-loss, squared error loss, boosting loss, and as limiting cases cost-weighted misclassification losses." The present chapter does not investigate this aspect of proper scoring rules, and we refer the reader to the papers [27, 28, 164] for more details.

In this chapter, we assume the existence of some 'gold standard' questions whose answers are known apriori to the system designer. As a result, the payment to a worker is determined solely by her own work. There is a parallel line of literature that explores the design of mechanisms that operate in the absence of any gold standard questions. The idea in the aforementioned line of literature is to reward the agents based on certain criteria that compares certain elicited data from the agents with each other, and typically involves asking agents to predict other agents' responses. The mechanisms designed often provide weaker guarantees (such as that of truth-telling being a Nash equilibrium) due to the absence of a gold standard answer to compare with. This line of literature includes work on peer-prediction [57, 171], the Bayesian truth serum [188] and prediction markets [50, 267], and related subsequent works [115].

The design of statistical inference algorithms for denoising the data obtained from workers is an active topic of research; see Chapter 4 for more details. In addition, several machine learning algorithms accommodating errors in the data have also been designed [5, 36, 47, 146]. These algorithms are typically oblivious to the elicitation procedure. Our work nicely complements this line of research in that these inference algorithms may now additionally employ the higher quality data and the specific structure of the elicited data for an improved denoising efficiency.

Another relevant problem in crowdsourcing is that of choosing which workers to hire or efficiently matching workers to tasks, and such problems are studied in the papers [4, 102, 271, 277] under different contexts. Our work assumes that a worker is already matched, and focuses on incentivizing that worker to respond in a certain manner. A recent line of work has focussed on elicitation of data from multiple agents in order to perform certain specific estimation tasks [33, 61, 75, 156]. In contrast, our goal is to ensure that workers censor their own low-quality (raw) data, without restricting our attention to any specific downstream algorithm or task.

Organization. The organization of this chapter is as follows. We present the formal problem setting in Section 6.2. In Section 6.3 we consider the skip-based setting: We present our proposed mechanism and show that it is the only mechanism which satisfies the requirements discussed above. In Section 6.4 we evaluate the proposed schemes via synthetic simulations. We discuss various modeling choices made in our work as well as direction for future research in Section 6.5. In Section 6.6 we present proofs of our theoretical results.

This chapter also contains an appendix (Section 6.A) in which we prove that imposing a requirement that is only slightly stronger than our proposed no-free-lunch axiom leads to impossibility results. Subsequently in Chapter 7 (Section 7.4), we also present experiments using data from Amazon Mechanical Turk that reveal significant improvements in the quality of data using our skip-based setting and multiplicative mechanisms.

6.2 Problem setting

In the crowdsourcing setting that we consider, one or more workers perform a *task*, where a task consists of multiple *questions*. The questions are objective, by which we mean, each question has precisely one correct answer. Examples of objective questions include multiple-choice classification questions such as Figure 6.1, questions on transcribing text from audio or images, etc.

For any possible answer to any question, we define the worker's *confidence about an answer* as the probability, according to her belief, of this answer being correct. In other words, one can assume that the worker has (in her mind) a probability distribution over all possible answers to a question, and the confidence for an answer is the probability of that answer being correct. As a shorthand, we also define the *confidence about a question* as the confidence for the answer that the worker is most confident about for that question. We assume that the worker's confidences for different questions are independent. In the interface we consider, for each question, the worker can either choose to 'skip' the question or provide an answer (Figure 6.1b). Our goal is that for every question, the worker should be incentivized to skip if her confidence for that question is below a certain pre-defined threshold, otherwise select the answer that she is most confident about.

Let N denote the total number of questions in the task. Among these questions, we assume the existence of some "gold standard" questions, that is, a set of questions whose answers are known to the requester. Let G $(1 \leq G \leq N)$ denote the number of gold standard questions. The G gold standard questions are assumed to be distributed uniformly at random in the pool of N questions (of course, the worker does not know which G of the N questions form the gold standard). The payment to a worker for a task is computed after receiving her responses to all the questions in the task. The payment is based on the worker's performance on the gold standard questions. Since the payment is based on known answers, the payments to different workers do not depend on each other, thereby allowing us to consider the presence of only one worker without any loss in generality.

Let $x_1, \ldots, x_G$ denote the evaluations of the answers that the worker gives to the G gold standard questions, and let f denote the scoring rule, i.e., a function that determines the payment to the worker based on these evaluations $x_1, \ldots, x_G$. In the skip-based setting we consider in this chapter, we have $x_i \in \{-1, 0, +1\}$ for all $i \in [G]$. Here, "0" denotes that the worker skipped the question, "-1" denotes that the worker attempted to answer the question and that answer was incorrect, and "$+1$" denotes that the worker attempted to answer the question and that answer was correct. The payment function is $f : \{-1, 0, +1\}^G \rightarrow \mathbb{R}$.

The payment is further associated to two parameters, $\alpha_{\max}$ and $\alpha_{\min}$. The parameter $\alpha_{\max}$ denotes the *budget*, i.e., the maximum amount that is paid to any individual worker for this task:

$$\max_{x_1,\ldots,x_G} f(x_1, \ldots, x_G) = \alpha_{\max}.$$

The amount $\alpha_{\max}$ is thus the amount of compensation paid to a perfect worker for her work. Further, one may often also have the requirement of paying a certain minimum amount to any worker. The parameter $\alpha_{\min}$ ($\leq \alpha_{\max}$) denotes this minimum payment: the payment function must also satisfy

$$\min_{x_1,\ldots,x_G} f(x_1, \ldots, x_G) \geq \alpha_{\min}.$$

For instance, crowdsourcing platforms today allow payments to workers, but do not allow imposing penalties: this condition gives $\alpha_{\min} = 0$.

We assume that the worker attempts to maximize her overall expected payment. In what follows, the expression 'the worker's expected payment' will refer to the expected payment from the worker's point of view, and the expectation will be taken with respect to the worker's confidences about her answers and the uniformly random choice of the G gold standard questions among the N questions in the task. A payment function f is called *incentive compatible* if the expected payment of the worker under this payment function is *strictly* maximized when the worker answers in the manner desired.[1]

We now explain what we mean by the phrase "manner desired" in the context of our skip-based setting. Let $T \in (0, 1)$ be some predefined "threshold" value. The goal is to design payment mechanisms that incentivize the worker to skip the questions for which her confidence is lower than T, and answer those for which her confidence is higher than T. [2] Moreover, for the questions that she attempts to answer, she must be incentivized to select the answer that she believes is most likely to be correct. The value of T is chosen apriori based on factors such as budget constraints, the targeted quality of labels, and/or the choice of the algorithm used to subsequently aggregate the responses of multiple workers. In this chapter, we assume that the value of the threshold T is specified to us.

In the remainder of this section, we formally define the concepts of the worker's expected payment and incentive compatibility; the reader interested in understanding the chapter at a higher level may skip directly to the next section without loss in continuity.

Let Ω denote the set of options for each question. We assume that Ω is a finite set, for instance, the set $\{\text{Yes}, \text{No}\}$ for a task with binary-choice questions, or the set of all strings of at most a certain length for a task with textual responses. Let $Q \in [0, 1]^{|\Omega| \times N}$ denote the beliefs of a worker for the N questions asked. Specifically, for any question $i \in [N]$ and any option $\omega \in \Omega$, let $Q_{\omega,i}$ represent the probability, according to the worker's belief, that option ω is the correct answer to question i. Then from the law of total probability, any valid Q must have $\sum_{\omega\in\Omega} Q_{\omega,i} = 1$ for every $i \in [N]$. The value of Q is unknown to the mechanism.

Let us first define the notion of the expected payment (from the worker's point of view) for any given response of the worker to the questions. For any question $i \in [N]$, define a variable $\xi_i \in \{0, 1\}$ that is set as 0 if the worker skips question i and is set as 1 if the worker

[1] Such a notion of incentive compatibility is often called "strict incentive compatibility"; we drop the prefix term "strict" for brevity.

[2] In the event that the confidence about a question is exactly equal to T, the worker may choose to answer or skip.

attempts question i. Further, for every question $i \in [N]$ such that $\xi_i \neq 0$, let $\omega_i \in \Omega$ denote the option selected by the worker; whenever $\xi_i = 0$, indicating a skip, we let ω_i take any arbitrary value in Ω. Furthermore, let $p_i = Q_{\omega_i,i}$ denote the probability, according to the worker's belief, that the chosen option ω_i is the correct answer to question i. For notational purposes, we also define a vector $E = (\epsilon_1, \ldots, \epsilon_G) \in \{-1, 1\}^G$. Then for the given responses, for the worker beliefs Q, and under payment mechanism f, the worker' expected payment $\Gamma_{Q,f} : (\{0, 1\} \times \Omega)^N \to \mathbb{R}$ is given by the expression:

$$\begin{aligned}&\Gamma_{Q,f}(\xi_1,\ \omega_1,\ \ldots,\ \xi_N,\ \omega_N)\\&\qquad = \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq\{1,\ldots,N\}}} \sum_{E\in\{-1,1\}^G} \left(f(\epsilon_1\xi_{j_1}, \ldots, \epsilon_G\xi_{j_G}) \prod_{i=1}^{G} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1 - p_{j_i})^{\frac{1-\epsilon_i}{2}} \right). \quad (6.1)\end{aligned}$$

In the expression (6.1), the outermost summation corresponds to the expectation with respect to the randomness arising from the unknown positions of the gold standard questions. The inner summation corresponds to the expectation with respect to the worker's beliefs about the correctness of her responses. Note that the right hand side of (6.1) implicitly depends on $(\omega_1, \ldots, \omega_N)$ through the values $(p_1, \ldots, p_N)$. Also note that for every question i such that $\xi_i = 0$, the right hand side of (6.1) does not depend on the values of ω_i and p_i; this is because the choice $\xi_i = 0$ of skipping question i implies that the worker did not select any particular option.

We will now use the the definition of the expected payment of the worker to define the notion of incentive compatibility. To this end, for any valid probabilities Q, let $\mathcal{A}(Q) \subseteq (\{0, 1\} \times \Omega)^N$ denote an associated set of "desired" responses. By this we mean that every $a \in (\{0, 1\} \times \Omega)^N$ represents a possible response to the set of N questions, and the goal is to incentivize the worker to provide any one response in the set $\mathcal{A}(Q)$. Then a mechanism f is termed incentive compatible if

$$\Gamma_{Q,f}(a) > \Gamma_{Q,f}(a') \quad \text{for every } a \in \mathcal{A}(Q), \text{ every } a' \notin \mathcal{A}(Q), \text{ and every valid } Q.$$

The goal is to design mechanisms that are incentive compatible, that is, incentivize the workers to respond in the desired manner that is discussed above.

6.3 Main results

In this section we present our main results for the skip-based setting that is considered in this chapter.

6.3.1 The no-free-lunch axiom

Recall from the previous that the goal is to design an incentive compatible mechanism for the skip-based setting. In order for practical deployment, we must somehow decide on any one particular mechanism. However, the space of all possible mechanisms for this problem may be rather wide. Thus in order to narrow down our search, we begin by imposing the following additional simple and natural requirement:

Axiom 1 (No-free-lunch Axiom). *If all the answers attempted by the worker in the gold standard are wrong, then the payment is the minimum possible. More formally,* $f(x_1, \ldots, x_G) = \alpha_{\min}$ *for every evaluation* $(x_1, \ldots, x_G)$ *such that* $0 < \sum_{i=1}^{G} \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^{G} \mathbf{1}\{x_i = -1\}$.

One may expect a payment mechanism to impose the restriction of minimum payment to spammers who answer randomly. For instance, in a task with binary-choice questions, a spammer is expected to have 50% of the attempted answers incorrect; one may thus wish to set a the minimum possible payment if 50% or more of the attempted answers were incorrect. The no-free-lunch axiom which we impose is however a *significantly weaker condition*, mandating minimum payment if *all* attempted answers are incorrect.

6.3.2 Payment mechanism

We now present our proposed payment mechanism in Algorithm 1.

Algorithm 1 Incentive mechanism for skip-based setting

```
• Inputs:
    ▸ Threshold $T$
    ▸ Budget parameters $\alpha_{\max}$ and $\alpha_{\min}$
    ▸ Evaluations $(x_1, \ldots, x_G) \in \{-1, 0, +1\}^G$ of the worker's answers to the $G$ gold
      standard questions
• Set $\alpha_{-1} = 0,\ \alpha_0 = 1,\ \alpha_{+1} = \frac{1}{T}$
• The payment is
      $f(x_1, \ldots, x_G) = \kappa \prod_{i=1}^{G} \alpha_{x_i} + \alpha_{\min},$
  where $\kappa = (\alpha_{\max} - \alpha_{\min}) T^G$.
```

The proposed mechanism has a *multiplicative* form: each answer in the gold standard is given a score based on whether it was correct (score = $\frac{1}{T}$), incorrect (score = 0) or skipped (score = 1), and the final payment is simply a product of these scores (scaled and shifted by constants). The mechanism is easy to describe to workers: For instance, if $T = \frac{1}{2}$, $G = 3$, $\alpha_{\max} = 80$ cents and $\alpha_{\min} = 0$ cents, then the description reads:

> *"The reward starts at 10 cents. For every correct answer in the 3 gold standard questions, the reward will double. However, if any of these questions are answered*

> *incorrectly, then the reward will become zero. So please use the 'I'm not sure' option wisely."*

Observe how this payment rule is similar to the popular 'double or nothing' paradigm [182].

The algorithm makes a minimum payment if *one or more* attempted answers in the gold standard are wrong. Note that this property is significantly stronger than the no-free-lunch axiom which we originally required, where we wanted a minimum payment only when *all* attempted answers were wrong. Surprisingly, as we prove shortly, Algorithm 1 is the only incentive-compatible mechanism that satisfies no-free-lunch.

The following theorem shows that our mechanism is indeed guaranteed to satisfy the stated requirements.

Theorem 16. *The mechanism of Algorithm 1 is incentive-compatible and satisfies the no-free-lunch axiom.*

We see that this mechanism thus incentivizes a worker to skip the questions for which her confidence is below T, while answering those for which her confidence is greater than T. In the latter case, the worker is incentivized to select the answer which she thinks is most likely to be correct.

6.3.3 Uniqueness of our mechanism

While we started out with a very weak condition of no-free-lunch of that requires a minimum payment when *all* attempted answers are wrong, the mechanism proposed in Algorithm 1 is significantly more strict and pays the minimum amount when *any* of the attempted answers is wrong. A natural question that arises is: can we design an alternative mechanism satisfying incentive compatibility and no-free-lunch that operates somewhere in between? The following theorem answers this question in the negative.

Theorem 17. *The mechanism of Algorithm 1 is the only incentive-compatible mechanism that satisfies the no-free-lunch axiom.*

Theorem 17 gives a strong result despite imposing very weak requirements. To see this, recall our earlier discussion on deterring spammers, that is, making a low payment to workers who answer randomly. For instance, when the task comprises binary-choice questions, one may wish to design mechanisms which make the minimum possible payment when the responses to 50% or more of the questions in the gold standard are incorrect. The no-free-lunch axiom is a much weaker requirement, and the only mechanism that can satisfy this requirement is the mechanism of Algorithm 1.

The proof of Theorem 17 is based on the following key lemma, establishing a condition that any incentive-compatible mechanism must necessarily satisfy. Note that this lemma does *not* require the no-free-lunch axiom.

Lemma 33. *Any incentive-compatible mechanism f must satisfy, for every gold standard question $i \in \{1, \ldots, G\}$ and every $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \in \{-1, 0, 1\}^{G-1}$,*

$$Tf(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_G) + (1-T)f(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_G) \\ = f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_G) \ .$$

The proof of Lemma 33 is provided in Section 6.6.3 along with the proofs of all other theoretical results.

6.3.4 Optimality against spamming behavior

As discussed earlier, crowdsourcing tasks, especially those with multiple choice questions, often encounter spammers who answer randomly without heed to the question being asked. For instance, under a binary-choice setup, a spammer will choose one of the two options uniformly at random for every question. A highly desirable objective in crowdsourcing settings is to deter spammers. To this end, one may wish to impose a condition of making the minimum possible payment when the responses to 50% or more of the attempted questions in the gold standard are incorrect. A second desirable metric could be to minimize the expenditure on a worker who simply skips all questions. While the aforementioned requirements were deterministic functions of the worker's responses, one may alternatively wish to impose requirements that depend on the distribution of the worker's answering process. For instance, a third desirable feature would be to minimize the expected payment to a worker who answers all questions uniformly at random. We now show that interestingly, our unique multiplicative payment mechanism *simultaneously* satisfies all these requirements. The result is stated assuming a multiple-choice setup, but extends trivially to non-multiple-choice settings.

Theorem 17.R (Distributional). *Consider any value $A \in \{0, \ldots, G\}$. Among all incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), Algorithm 1 pays strictly the smallest amount to a worker who skips some A of the questions in the the gold standard, and chooses answers to the remaining $(G - A)$ questions uniformly at random.*

Theorem 17.S (Deterministic). *Consider any value $B \in (0, 1]$. Among all incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), Algorithm 1 pays strictly the smallest amount to a worker who gives incorrect answers to a fraction B or more of the questions attempted in the gold standard.*

We see from this result that the multiplicative payment mechanism of Algorithm 1 thus possesses very useful properties geared to deter spammers, while ensuring that a good worker will be paid a high enough amount. To illustrate this point, let us compare the mechanism of Algorithm 1 with the popular additive class of payment mechanisms.

Example 3. *Consider the popular class of "additive" mechanisms, where the payments to a worker are added across the gold standard questions. This additive payment mechanism offers a reward of $\frac{\alpha_{\max}}{G}$ for every correct answer in the gold standard, $\frac{\alpha_{\max}T}{G}$ for every question skipped, and* 0 *for every incorrect answer. Importantly, the final payment to the worker is the* sum *of the rewards across the G gold standard questions. One can verify that this additive mechanism is incentive compatible. One can also see that that as guaranteed by our theory, this additive payment mechanism does not satisfy the no-free-lunch axiom.*

Suppose each question involves choosing from two options. Let us compute the payment that these two mechanisms make under a spamming behavior of choosing the answer randomly to each question. Given the 50% *likelihood of each question being correct, on can compute that the additive mechanism makes a payment of $\frac{\alpha_{\max}}{2}$ in expectation. On the other hand, our mechanism pays an expected amount of only $\alpha_{\max}2^{-G}$. The payment to spammers thus reduces exponentially with the number of gold standard questions under our mechanism, whereas it does not reduce at all in the additive mechanism.*

Now, consider a different means of exploiting the mechanism(s) where the worker simply skips all questions. To this end, observe that if a worker skips all the questions then the additive payment mechanism will make a payment of $\alpha_{\max}T$. On the other hand, the proposed payment mechanism of Algorithm 1 pays an exponentially smaller amount of $\alpha_{\max}T^G$ (recall that $T < 1$).

6.4 Simulations

In this section, we present synthetic simulations the effects of our setting and our mechanism on the final label quality. Experiments using real-world data from crowdsourcing are described in Section 7.4 of Chapter 7, where the skip-based setting is compared with the standard baseline as well as a confidence-based setting that forms the focus of Chapter 7.

In this section, we employ synthetic simulations to understand the effects of various distributions of the confidences and labeling errors. We consider binary-choice questions in this set of simulations. Whenever a worker answers a question, her confidence for the correct answer is drawn from a distribution $\mathcal{P}$ independent of all else. We investigate the effects of the following five choices of the distribution $\mathcal{P}$:

- The uniform distribution on the support $[0.5, 1]$.
- A triangular distribution with lower end-point 0.2, upper end-point 1 and a mode of 0.6.
- A beta distribution with parameter values $\alpha = 5$ and $\beta = 1$.
- The hammer-spammer distribution [117]: uniform on the discrete set $\{0.5, 1\}$.
- A truncated Gaussian distribution: a truncation of $\mathcal{N}(0.75, 0.5)$ to the interval $[0, 1]$.

We compare (a) the setting where workers attempt every question, with (b) the setting where workers skip questions for which their confidence is below a certain threshold T. In this set of simulations, we set $T = 0.75$. In either setting, we aggregate the labels obtained

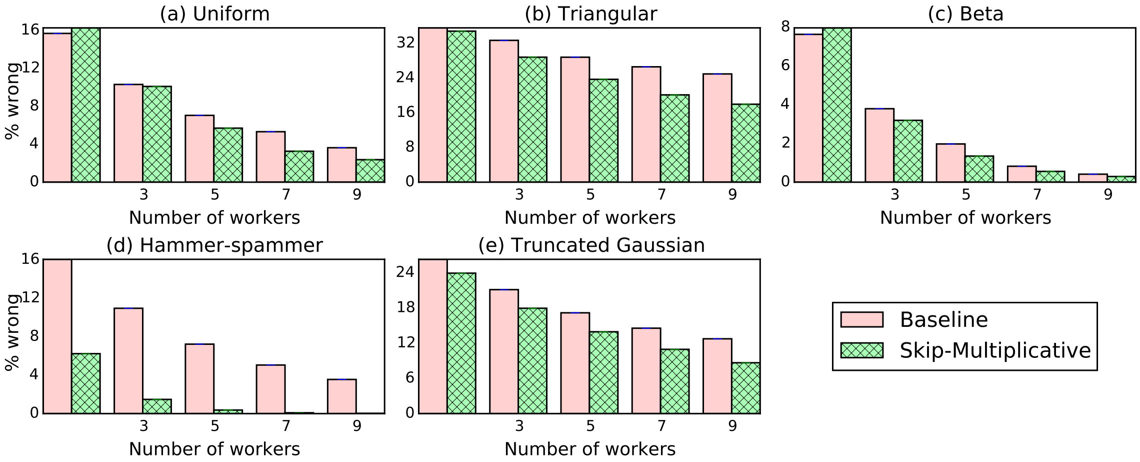


Figure 6.2: Error under different interfaces for synthetic simulations of five distributions of the workers' error probabilities.

from the workers for each question via a majority vote on the two classes. Ties are broken by choosing one of the two options uniformly at random.

Figure 6.2 depicts the results from these simulations. Each bar represents the fraction of questions that are labeled incorrectly, and is an average across 50,000 trials. (The standard error of the mean is too small to be visible.) We see that the skip-based setting consistently outperforms the conventional setting, and the gains obtained are moderate to high depending on the underlying distribution of the workers' errors. In particular, the gains are quite striking under the hammer-spammer model: this result is not surprising since the mechanism (ideally) screens the spammers out and leaves only the hammers who answer perfectly.

The setup of the simulations described above assumes that the workers confidences equal the true error probabilities. In practice, however, the workers may have incorrect beliefs. The setup also assumes that ties are broken randomly; however in practice, ties may be broken in a more systematic manner by eliciting additional labels for only these hard questions. We now present a second set of simulations that mitigates these biases. In particular, when a worker has a confidence of p, the actual probability of error is assumed to be drawn from a Gaussian distribution with mean p and standard deviation 0.1, truncated to $[0, 1]$. In addition, when evaluating the performance of the majority voting procedure, we consider a tie as having an error of 0.4. Figure 6.3 depicts the results of these simulations. We observe that the results from these simulations are very similar to those obtained in the earlier simulation setup of Figure 6.2.

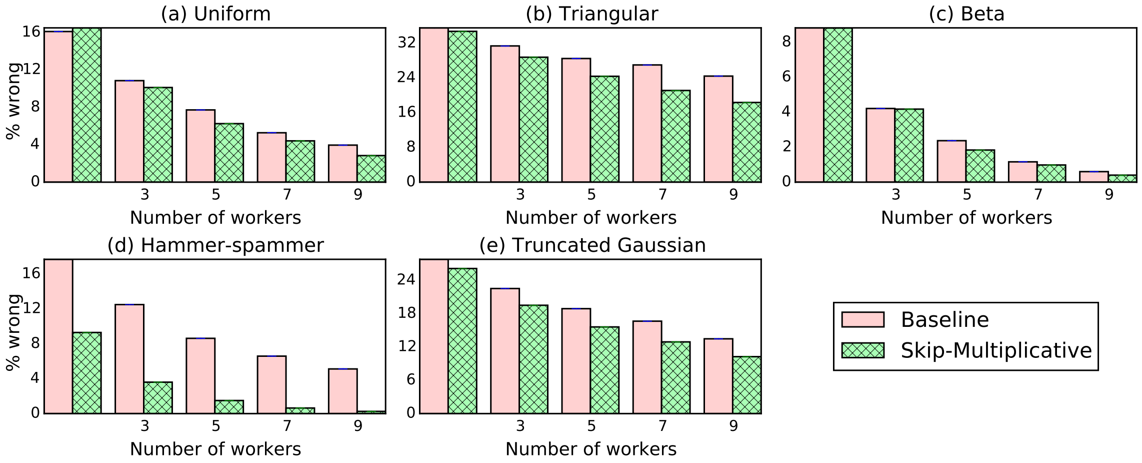


Figure 6.3: Errors under a model that is a perturbation of the first experiment, where the worker's confidence is a noisy version of the true error probability and where ties are considered different from random decisions.

6.5 Discussion

Given the uniqueness in theory, simplicity, and good empirical performance, we envisage our 'multiplicative' mechanisms to be of interest to machine learning researchers and practitioners who use crowdsourcing to collect labeled data. In this concluding section, we first discuss the modeling assumptions that we made in this chapter, followed by a discussion on open problems.

Modelling assumptions

When forming the model for our problem, as in any other field of theoretical research, we had to make certain assumptions and choices. In what follows, we discuss the reasons for the modeling choices we made.

- *Use of gold standard questions.* We assume the existence of gold standard questions in the task, i.e., a subset of questions to which the answers are known to the system designer. The existence of gold standard is commonplace in crowdsourcing platforms [45, 143].

- *Error rate vs. sample size under skips* Consider the skip-based setting. As discussed earlier, allowing the workers to skip questions they are unsure of reduces the error rates in the data obtained. However, the higher data quality trades off with the sample size, i.e., the amount of (non-skipped) data obtained is lower. In our simulations in this chapter as well as in the experiments discussed in the next chapter, we see that the tradeoff is very favorable towards our skip-based setting in that we see a significant drop in the aggregate error with the use of our skip-based interface and mechanism.

- *Workers aiming to maximize their expected payments:* We assume that the workers aim to maximize their expected payments. In many other problems in game theory, one often makes the assumption that people are "risk-averse", and aim to maximize the expected value of some "utility function" of their payments. However in the context of the crowdsourcing settings we consider here, we believe that the assumption of workers maximizing their expected payments is a perfectly reasonable assumption. The reason is that each such task in crowdsourcing lasts for a handful of minutes and is worth a few tens of cents. Workers typically perform tens to hundreds of tasks per day, and consequently their empirical hourly wages very quickly converge to their expectation.

- *Workers knowing their confidences:* We understand that in practice the workers will have noisy or granular estimates of their own beliefs. The mathematical assumption of workers knowing their precise confidences is an idealization intended for mathematical tractability. This is one of the reasons why we only elicit a quantized value of the workers' beliefs (in terms of skipping or choosing one of a finite number of confidence levels), and not try to ask for a precise value.

- *Eliciting a quantized version of the beliefs:* We do not directly attempt to elicit the values of the beliefs of the workers, but instead ask them to indicate only a quantization by means of either attempting or skipping the question. In the next chapter, we extend our results to a finer quantization where we also ask for the confidence level of the worker in terms of a finite number of choices such as {"I'm not sure", "moderately sure", "absolutely sure"}. We prefer this quantization to direct assessment to real-valued probability, motivated by the extensive literature in psychology on the coarseness of human perception and processing (e.g., [114, 169, 220, 238]) establishing that humans are more comfortable at providing quantized responses. This notion is verified by experiments on Amazon Mechanical Turk in the paper [220] where it is observed that people are more consistent when giving ordinal answers (comparing pairs of items) as opposed to when they are asked for numeric evaluations.

Open problems

We discuss two sets of open problems, one from the practical perspective and another on the theoretical front.

First, in the chapter, we assumed that the number of total questions N in a task, the number of gold standard questions G, and the threshold T for skipping (or the number and thresholds of the different confidence levels) were provided to the mechanism. While these parameters may be chosen by hand by a system designer based on her own experience, a more principled design of these parameters is an important question. The choices for these parameters may have to be made based on certain tradeoffs. For instance, a higher value of G reduces the variance in the payments but uses more resources in terms of gold standard questions. Or for instance, more number of threshold levels L would increase the amount of information obtained about the workers' beliefs, but also increase the noise in the workers' estimates of her own beliefs.

A second open problem is the design of inference algorithms that can exploit the specific structure of the skip-based setting. There are several algorithms and theoretical analyses in the literature for aggregating data from multiple workers in the baseline setting (see Chapter 4). A useful direction of research in the future is to develop algorithms and theoretical guarantees that incorporate information about the workers' confidences. For instance, for the skip-based setting considered in this chapter, the missing labels are not missing "at random" but are correlated with the difficulty of the task. Designing algorithms that can exploit this information judiciously (e.g., via confidence-weighed worker/item constraints in the minimax entropy method of [276]) is a useful direction of future research.

6.6 Proofs

In this section, we prove the claimed theoretical results whose proofs are not included in the main text of the chapter.

The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of these proofs, we can assume without loss of generality that $\alpha_{\min} = 0$.

6.6.1 Proof of Theorem 16: Our mechanism works

The proposed payment mechanism satisfies no-free-lunch since the payment is $\alpha_{\min}$ when there are one or more wrong answers in the gold standard. It remains to show that the mechanism is incentive compatible. To this end, observe that the property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\alpha_{\min} = 0$.

We will first assume that, for every question that the worker does not skip, she selects the answer which she believes is most likely to be correct. Under this assumption we will show that the worker is incentivized to skip the questions for which her confidence is smaller than T and attempt if it is greater than T. Finally, we will show that the mechanism indeed incentivizes the worker to select the answer which she believes is most likely to be correct for the questions that she doesn't skip. In what follows, we will employ the notation $\kappa = \alpha_{\max} T^G$.

Let us first consider the case when $G = N$. Let $p_1, \ldots, p_N$ be the confidences of the worker for questions $1, \ldots, N$ respectively. Further, let $p_{(1)} \geq \cdots \geq p_{(m)} > T > p_{(m+1)} \geq \cdots \geq p_{(N)}$ be the ordered permutation of these confidences (for some number m). Let $\{(1), \ldots, (N)\}$ denote the corresponding permutation of the N questions. If the mechanism is incentive compatible, then the expected payment received by this worker should be maximized when the worker answers questions $(1), \ldots, (m)$ and skips the rest. Under the mechanism proposed in Algorithm 1, this action fetches the worker an expected payment of

$$\kappa \frac{p_{(1)}}{T} \cdots \frac{p_{(m)}}{T}.$$

Alternatively, if the worker answers the questions $\{i_1, \ldots, i_\beta\}$, with $p_{i_1} > \cdots > p_{i_\nu} > T > p_{i_{\nu+1}} > \cdots > p_{i_\beta}$ for some value ν, then the expected payment is

$$\begin{aligned} p_{i_1} \cdots p_{i_\beta} \frac{\kappa}{T^\beta} &= \kappa \frac{p_{i_1}}{T} \cdots \frac{p_{i_\beta}}{T} && (6.2) \\ &\leq \kappa \frac{p_{i_1}}{T} \cdots \frac{p_{i_\nu}}{T} && (6.3) \\ &\leq \kappa \frac{p_{(1)}}{T} \cdots \frac{p_{(m)}}{T} && (6.4) \end{aligned}$$

where inequality (6.3) holds because $\frac{p_{i_j}}{T} \leq 1 \;\; \forall \; j > \nu$ and holds with equality only when $\beta = \nu$. Inequality (6.4) is a result of $\frac{p_{(j)}}{T} \geq 1 \;\; \forall \; j \leq m$ and holds with equality only when $\nu = m$. It follows that the expected payment is (strictly) maximized when $i_1 = (1), \ldots, i_\beta = (m)$ as required.

The case of $G < N$ is a direct consequence of the result for $G = N$, as follows. When $G < N$, from a worker's point of view, the set of G questions is distributed uniformly at random in the superset of N questions. However, for every set of G questions, the relations (6.2), (6.3), (6.4) and their associated equality/strict-inequality conditions hold. The expected payment is thus (strictly) maximized when the worker answers the questions for which her confidence is greater than T and skips those for which her confidence is smaller than T.

One can see that for every question that the worker chooses to answer, the expected payment increases with an increase in her confidence. Thus, the worker is incentivized to select the answer that she thinks is most probably correct.

Finally, since $\kappa = \alpha_{\max} T^G > 0$ and $T \in (0, 1)$, the payment is always non-negative and satisfies the $\alpha_{\max}$-budget constraint.

6.6.2 Proof of Theorem 17: Uniqueness of our mechanism

The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\alpha_{\min} = 0$.

We will first prove that any incentive-compatible mechanism satisfying the no-free-lunch axiom must make a zero payment if one or more answers in the gold standard are incorrect. The proof proceeds by induction on the number of skipped questions S in the gold standard. Let us assume for now that in the G questions in the gold standard, the first question is answered incorrectly, the next $(G - 1 - S)$ questions are answered by the worker and have arbitrary evaluations, and the remaining S questions are skipped. The proof proceeds by an induction on S. Suppose $S = G - 1$. In this case, the only attempted question is the first question and the answer provided by the worker to this question is incorrect. The no-free-lunch axiom necessitates a zero payment in this case, thus satisfying the base case of our induction hypothesis. Now we prove the hypothesis for some S under the assumption of it being true when the number of questions skipped in the gold standard is $(S+1)$ or more. From Lemma 33 (with $i = G - S - 1$) we have

$$\begin{aligned}Tf(-1, y_2, \ldots, y_{G-S-2}, 1, 0, \ldots, 0) + (1-T)f(-1, y_2, \ldots, y_{G-S-2}, -1, 0, \ldots, 0)& \\ = f(-1, y_2, \ldots, y_{G-S-2}, 0, 0, \ldots, 0)& \\ = 0,&\end{aligned}$$

where the final equation is a consequence of our induction hypothesis: The induction hypothesis is applicable since $f(-1, y_2, \ldots, y_{G-S-2}, 0, 0, \ldots, 0)$ corresponds to the case when the last $(S+1)$ questions are skipped and the first question is answered incorrectly. Now, since the payment f must be non-negative and since $T \in (0,1)$, it must be that

$$f(-1, y_2, \ldots, y_{G-S-2}, 1, 0, \ldots, 0) = 0,$$

and

$$f(-1, y_2, \ldots, y_{G-S-2}, -1, 0, \ldots, 0) = 0.$$

This completes the proof of our induction hypothesis. Furthermore, each of the arguments above hold for any permutation of the G questions, thus proving the necessity of zero payment when any one or more answers are incorrect.

We will now prove that when no answers in the gold standard are incorrect, the payment must be of the form described in Algorithm 1. Let κ be the payment when all G questions in the gold standard are skipped. Let C be the number of questions answered correctly in the gold standard. Since there are no incorrect answers, it follows that the remaining $(G - C)$ questions are skipped. Let us assume for now that the first C questions are answered correctly and the remaining $(G - C)$ questions are skipped. We repeatedly apply Lemma 33, and the fact that the payment must be zero when one or more answers are wrong, to get

$$\begin{aligned}f(\underbrace{1,\ldots,1}_{C-1}, 1, \underbrace{0,\ldots,0}_{G-C}) &= \frac{1}{T}f(\underbrace{1,\ldots,1}_{C-1}, 0, \underbrace{0,\ldots,0}_{G-C}) - \frac{1-T}{T}f(\underbrace{1,\ldots,1}_{C-1}, -1, \underbrace{0,\ldots,0}_{G-C}) \\ &= \frac{1}{T}f(\underbrace{1,\ldots,1}_{C-1}, 0, \underbrace{0,\ldots,0}_{G-C}) \\ &\vdots \\ &= \frac{1}{T^C}f(\underbrace{0,\ldots,0}_{G}) \\ &= \frac{1}{T^C}\kappa\ .\end{aligned}$$

In order to abide by the budget, we must have the maximum payment as $\alpha_{\max} = \kappa \frac{1}{T^G}$. It follows that $\kappa = \alpha_{\max} T^G$. Finally, the arguments above hold for any permutation of the G questions, thus proving the uniqueness of the mechanism of Algorithm 1.

6.6.3 Proof of Lemma 33: The workhorse lemma

First we consider the case of $G = N$. In the set $\{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G\}$, for some $(\eta, \gamma) \in \{0, \ldots, G-1\}^2$ such that $\eta + \gamma + 1 \leq G$, suppose there are η elements with a value 1, γ elements with a value -1, and $(G - 1 - \eta - \gamma)$ elements with a value 0. Let us assume for now that $i = \eta + \gamma + 1$, $y_1 = 1, \ldots, y_\eta = 1, y_{\eta+1} = -1, \ldots, y_{\eta+\gamma} = -1, y_{\eta+\gamma+2} = 0, \ldots, y_G = 0$.

Suppose the worker has confidences $(p_1, \ldots, p_{\eta+\gamma}) \in (T, 1]^{\eta+\gamma}$ for the first $(\eta + \gamma)$ questions, a confidence of $q \in (0, 1]$ for the next question, and confidences smaller than T for the remaining $(G - \eta - \gamma - 1)$ questions. The mechanism must incentivize the worker to answer the first $(\eta + \gamma)$ questions and skip the last $(G - \eta - \gamma - 1)$ questions; for question $(\eta + \gamma + 1)$, it must incentivize the worker to answer if $q > T$ and skip if $q < T$. Supposing the worker indeed attempts the first $(\eta + \gamma)$ questions and skips the last $(G - \eta - \gamma - 1)$ questions, let $\boldsymbol{x} = \{x_1, \ldots, x_{\eta+\gamma}\} \in \{-1, 1\}^{\eta+\gamma}$ denote the the evaluation of the worker's answers to the first $(\eta + \gamma)$ questions. Define quantities $\{r_j\}_{j \in [\eta+\gamma]}$ as $r_j = 1 - p_j$ for $j \in \{1, \ldots, \eta\}$, and $r_j = p_j$ for $j \in \{\eta + 1, \eta + \gamma\}$. The requirement of incentive compatibility necessitates

$$
\begin{aligned}
q & \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma}, 1, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1 - r_j)^{\frac{1+x_j}{2}} \right) \\
& \quad + (1-q) \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma}, -1, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1 - r_j)^{\frac{1+x_j}{2}} \right) \\
\overset{q<T}{\underset{q>T}{\lessgtr}} & \qquad \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma}, 0, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1 - r_j)^{\frac{1+x_j}{2}} \right) .
\end{aligned}
$$

The left hand side of this expression is the expected payment if the worker chooses to answer question $(\eta + \gamma + 1)$, while the right hand side is the expected payment if she chooses to skip it. For any real-valued variable q, and for any real-valued constants a, b and c,

$$
aq \overset{q<c}{\underset{q>c}{\lessgtr}} b \quad \Rightarrow \quad ac = b \, .
$$

As a result,

$$T \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma}, 1, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right)$$
$$+ (1-T) \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma}, -1, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right)$$
$$- \sum_{\boldsymbol{x} \in \{-1,1\}^{\eta+\gamma}} \left(f(x_1, \ldots, x_\eta, -x_{\eta+1}, \ldots, -x_{\eta+\gamma-1}, 0, 0, \ldots, 0) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) = 0. \tag{6.5}$$

The left hand side of (6.5) represents a polynomial in $(\eta+\gamma)$ variables $\{r_j\}_{j=1}^{\eta+\gamma}$ which evaluates to zero for all values of the variables within a $(\eta+\gamma)$-dimensional solid Euclidean ball. Thus, the coefficients of the monomials in this polynomial must be zero. In particular, the constant term must be zero. The constant term appears when $x_j = 1 \; \forall \; j$ in the summations in (6.5). Setting the constant term to zero gives

$$\begin{aligned} &Tf(x_1 = 1, \ldots, x_\eta = 1, -x_{\eta+1} = -1, \ldots, -x_{\eta+\gamma} = -1, 1, 0, \ldots, 0) \\ &\quad + (1-T) f(x_1 = 1, \ldots, x_\eta = 1, -x_{\eta+1} = -1, \ldots, -x_{\eta+\gamma} = -1, -1, 0, \ldots, 0) \\ &\qquad - f(x_1 = 1, \ldots, x_\eta = 1, -x_{\eta+1} = -1, \ldots, -x_{\eta+\gamma} = -1, 0, 0, \ldots, 0) = 0 \end{aligned}$$

as desired. Since the arguments above hold for any permutation of the G questions, this completes the proof for the case of $G = N$.

Now consider the case $G < N$. Let $g : \{-1, 0, 1\}^N \rightarrow \mathbb{R}_+$ represent the expected payment given an evaluation of all the N answers, when the identities of the gold standard questions are unknown. Here, the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1, \ldots, x_N) \in \{-1, 0, 1\}^N$ are the evaluations of the worker's answers to the N questions then the expected payment is

$$g(x_1, \ldots, x_N) = \frac{1}{\binom{N}{G}} \sum_{(i_1, \ldots, i_G) \subseteq \{1, \ldots, N\}} f(x_{i_1}, \ldots, x_{i_G}) \; . \tag{6.6}$$

Notice that when $G = N$, the functions f and g are identical.

In the set $\{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G\}$, for some $(\eta, \gamma) \in \{0, \ldots, G-1\}^2$ with $\eta + \gamma < G$, suppose there are η elements with a value 1, γ elements with a value -1, and $(G - 1 - \eta - \gamma)$ elements with a value 0. Let us assume for now that $i = \eta + \gamma + 1$, $y_1 = 1, \ldots, y_\eta = 1, y_{\eta+1} = -1, \ldots, y_{\eta+\gamma} = -1, y_{\eta+\gamma+2} = 0, \ldots, y_G = 0$.

Suppose the worker has confidences $\{p_1, \ldots, p_{\eta+\gamma}\} \in (T, 1]^{\eta+\gamma}$ for the first $(\eta + \gamma)$ of the N questions, a confidence of $q \in (0, 1]$ for the next question, and confidences smaller than T for the remaining $(N - \eta - \gamma - 1)$ questions. The mechanism must incentivize the worker to answer the first $(\eta+\gamma)$ questions and skip the last $(N-\eta-\gamma-1)$ questions; for the $(\eta+\gamma+1)^{\text{th}}$ question, the mechanism must incentivize the worker to answer if $q > T$ and skip if $q < T$. Supposing the worker indeed attempts the first $(\eta+\gamma)$ questions and skips the last $(N-\eta-\gamma-1)$ questions, let $\boldsymbol{x} = \{x_1, \ldots, x_{\eta+\gamma}\} \in \{-1,1\}^{\eta+\gamma}$ denote the the evaluation of the worker's answers to the first $(\eta+\gamma)$ questions. Define quantities $\{r_j\}_{j\in[\eta+\gamma]}$ as $r_j = 1-p_j$ for $j \in \{1,\ldots,\eta\}$, and $r_j = p_j$ for $j \in \{\eta+1, \eta+\gamma\}$. The requirement of incentive compatibility necessitates

$$\begin{aligned}
q & \sum_{\boldsymbol{x}\in\{-1,1\}^{\eta+\gamma}} \left(g(x_1,\ldots,x_\eta,-x_{\eta+1},\ldots,-x_{\eta+\gamma},1,0,\ldots,0) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
& + (1-q) \sum_{\boldsymbol{x}\in\{-1,1\}^{\eta+\gamma}} \left(g(x_1,\ldots,x_\eta,-x_{\eta+1},\ldots,-x_{\eta+\gamma},-1,0,\ldots,0) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) \\
\underset{q>T}{\overset{q<T}{\lesseqgtr}} & \sum_{\boldsymbol{x}\in\{-1,1\}^{\eta+\gamma}} \left(g(x_1,\ldots,x_\eta,-x_{\eta+1},\ldots,-x_{\eta+\gamma},0,0,\ldots,0) \prod_{j\in[\eta+\gamma]} r_j^{\frac{1-x_j}{2}} (1-r_j)^{\frac{1+x_j}{2}} \right) .
\end{aligned} \tag{6.7}$$

Again, applying the fact that for any real-valued variable q and for any real-valued constants a, b and c, $aq \underset{q>c}{\overset{q<c}{\lesseqgtr}} b \quad \Rightarrow \; ac = b$, we get that

$$\begin{aligned}
& Tg(x_1=1,\ldots,x_\eta=1,-x_{\eta+1}=-1,\ldots,-x_{\eta+\gamma}=-1,1,0,\ldots,0) \\
& \quad + (1-T)g(x_1=1,\ldots,x_\eta=1,-x_{\eta+1}=-1,\ldots,-x_{\eta+\gamma}=-1,-1,0,\ldots,0) \\
& \qquad - g(x_1=1,\ldots,x_\eta=1,-x_{\eta+1}=-1,\ldots,-x_{\eta+\gamma}=-1,0,0,\ldots,0) = 0 .
\end{aligned} \tag{6.8}$$

The proof now proceeds via induction on the quantity $(G-\eta-\gamma-1)$, i.e., on the number of skipped questions in $\{y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_G\}$. We begin with the case of $(G-\eta-\gamma-1) = G-1$ which implies $\eta=\gamma=0$. In this case (6.8) simplifies to

$$Tg(1,0,\ldots,0) + (1-T)g(-1,0,\ldots,0) = g(0,0,\ldots,0) .$$

Applying the expansion of function g in terms of function f from (6.6) gives

$$\begin{aligned}
T\left(c_1 f(1,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right) + (1-T)\left(c_1 f(-1,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right) \\
= \left(c_1 f(0,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right)
\end{aligned}$$

for constants $c_1 > 0$ and $c_2 > 0$ that respectively denote the probabilities that the first question is picked and not picked in the set of G gold standard questions. Cancelling out the common terms on both sides of the equation, we get the desired result

$$Tf(1,0,\ldots,0) + (1-T)f(-1,0,\ldots,0) = f(0,0,\ldots,0) .$$

Next, we consider the case when $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard, and assume that the result is true when more than $(G - \eta - \gamma - 1)$ questions are skipped in the gold standard. In (6.8), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first where all of the first $(\eta + \gamma + 1)$ questions are included in the gold standard, and the second where one or more of the first $(\eta + \gamma + 1)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than $(G - \eta - \gamma - 1)$ questions skipped in the gold standard and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of (6.8). The remainder comprises only evaluations of function f for arguments in which the first $(\eta + \gamma + 1)$ questions are included in the gold standard: since the last $(N - \eta - \gamma - 1)$ questions are skipped by the worker, the remainder evaluates to

$$\begin{aligned} &Tc_3 f(y_1, \ldots, y_{\eta+\gamma}, 1, 0, \ldots, 0) + (1-T)c_3 f(y_1, \ldots, y_{\eta+\gamma}, -1, 0, \ldots, 0) \\ &\qquad\qquad\qquad\qquad\qquad\qquad = c_3 f(y_1, \ldots, y_{\eta+\gamma}, 0, 0, \ldots, 0) \end{aligned}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

Proof of Theorem 6.3.4: Minimum payment to spammers

The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of this proof, we can assume without loss of generality that $\alpha_{\min} = 0$.

Part A (Distributional). Let m denote the number of options in each question. One can verify that under the mechanism of Algorithm 1, a worker who skips A questions and answers the rest uniformly at random will get a payment of $\frac{\alpha_{\max} T^A}{m^{G-A}}$ in expectation. This expression arises due to the fact that Algorithm 1 makes a zero payment if any of the attempted answers are incorrect, and a payment of $\alpha_{\max} T^A$ if the worker skips A questions and answers the rest correctly. Under uniformly random answers, the probability of the latter event is $\frac{1}{m^{G-A}}$.

Now consider any other mechanism, and denote it as f'. Let us suppose without loss of generality that the worker attempts the first $(G - A)$ questions. Since the payment must be non-negative, a repeated application of Lemma 33 gives

$$\begin{aligned} f'(\underbrace{1, \ldots, 1}_{G-A}, 0, \ldots, 0) &\geq T f'(\underbrace{1, \ldots, 1}_{G-A+1}, 0, \ldots, 0) \qquad (6.9) \\ &\;\;\vdots \\ &\geq T^A f'(1, \ldots, 1) \\ &= T^A \alpha_{\max}, \qquad (6.10) \end{aligned}$$

where equation (6.10) is a result of the $\alpha_{\max}$-budget constraint. Since there is a $\frac{1}{m^{G-A}}$ chance of the $(G-A)$ attempted answers being correct, the expected payment under any other mechanism f' must be at least $\frac{\alpha_{\max} T^A}{m^{G-A}}$.

We will now show that if any mechanism f' that makes an expected payment of $\frac{\alpha_{\max} T^A}{m^{G-A}}$ to such a spammer, then the mechanism must be identical to Algorithm 1. We split the proof of this part into two cases, depending on the value of the parameter A.

Case I ($A < G$): In order to make an expected payment of $\frac{\alpha_{\max} T^A}{m^{G-A}}$, the mechanism must achieve the bound (6.10) with equality, and furthermore, the mechanism must have zero payment if any of the $(G-A)$ attempted questions are answered incorrectly. In other words, the mechanism f' under consideration must satisfy

$$f'(y_1, \ldots, y_{G-A}, 0, \ldots, 0) = 0 \qquad \forall (y_1, \ldots, y_{G-A}) \in \{-1, 1\}^{G-A} \backslash \{1\}^{G-A}.$$

A repeated application of Lemma 33 then implies

$$f'(0, 0, \ldots, -1) = 0. \tag{6.11}$$

Note that so far we considered the case when the worker attempts the first $(G-A)$ questions. The arguments above hold for any choice of the $(G-A)$ attempted questions, and consequently the results shown so far in this proof hold for all permutations of the arguments to f'. In particular, the mechanism f' must make a zero payment when any $(G-1)$ questions in the gold standard are skipped and the remaining question is answered incorrectly. Another repeated application of Lemma 33 to this result gives

$$f'(y_1, \ldots, y_G) = 0 \qquad \forall (y_1, \ldots, y_G) \in \{0, -1\}^G \backslash \{0\}^G.$$

This condition is precisely the no-free-lunch axiom, and in Theorem 17 we had shown that Algorithm 1 is the only incentive-compatible mechanism that satisfies this axiom. It follows that our mechanism, Algorithm 1 strictly minimizes the expected payment in the setting under consideration.

Case II ($A = G$): In order to achieve the bound (6.10) with equality, the mechanism f' must also achieve the bound (6.9) with equality. Noting that we have $A = G$ in this case, it follows that the mechanism f' must satisfy

$$f'(-1, 0, \ldots, 0) = 0.$$

This condition is identical to (6.11) established for Case I earlier, and the rest of the argument now proceeds in a manner identical to the subsequent arguments in Case I.

Part B (Deterministic). Given our result of Theorem 17, the proof for the deterministic part is straightforward. Algorithm 1 makes a payment of zero when one or more of the answers to questions in the gold standard are incorrect. Consequently, for every value of parameter $B \in (0, 1]$, Algorithm 1 makes a zero payment when a fraction B or more of the attempted answers are incorrect. Any other mechanism doing so must satisfy the no-free-lunch axiom. In Theorem 17 we had shown that Algorithm 1 is the only incentive-compatible mechanism that satisfies this axiom. It follows that our mechanism, Algorithm 1, strictly minimizes the payment in the event under consideration.

6.A Appendix: A stronger no-free-lunch

Recall that the no-free-lunch axiom under the skip-based mechanism of Section 6.3 requires the payment to be the minimum possible if all attempted answers in the gold standard are incorrect. However, a worker who skips all the questions may still receive a payment. One may thus wish to impose a stronger requirement instead, where the minimum payment is made to workers who make no useful contribution. This is the primary focus of this section.

In order to accommodate this requirement, let us define the following axiom which is slightly stronger than the no-free-lunch axiom defined earlier.

Strong-no-free-lunch: If none of the answers in the gold standard are correct, then the payment is $\alpha_{\min}$. More formally, strong-no-free-lunch imposes the condition $f(x_1, \ldots, x_G) = \alpha_{\min}$ for every evaluation $(x_1, \ldots, x_G)$ that satisfies $\sum_{i=1}^{G} \mathbf{1}\{x_i > 0\} = 0$.

The strong-no-free-lunch axiom is only slightly stronger than the no-free-lunch axiom proposed in Section 6.3. The strong-no-free-lunch axiom can equivalently be written as imposing requiring the payment to be the minimum possible for every evaluation that satisfies $\sum_{i=1}^{G} \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^{G} \mathbf{1}\{x_i = -1\}$. From this interpretation, one can see that to the set of events necessitating the minimum payment under the no-free-lunch axiom, the strong-no-free-lunch axiom adds only one extra event – the event of the worker skipping all questions. Unfortunately, it turns out that this minimal strengthening of the requirements is associated to impossibility results.

In this section we show that no mechanism satisfying the strong-no-free-lunch axiom can be incentive compatible in general. The only exception is the case when (a) all questions are in the gold standard ($G = N$), and (b) it is guaranteed that the worker has a confidence greater than T for at least one of the N questions. These conditions are, however, impractical for the crowdsourcing setup under consideration in this chapter. We will first prove the impossibility results under the strong-no-free-lunch axiom. For the sake of completeness (and also to satisfy mathematical curiosity), we will then provide a (unique) mechanism that is incentive-compatible and satisfies the strong-no-free-lunch axiom for the skip-based setting under the two conditions listed above. The proofs of each of the claims made in this section are provided at the end of this section.

In this section, we will call any worker whose confidences for all of the N questions is lower than T as an *unknowledgeable worker*, and call the worker a *knowledgeable worker* otherwise.

Proposition 9. *No payment mechanism satisfying the strong-no-free-lunch axiom can incentivize an unknowledgeable worker to skip all questions. As a result, no mechanism satisfying the strong-no-free-lunch axiom can be incentive-compatible.*

The proof of this proposition, and that of all other theoretical claims made in this section, are presented at the end of this section.

The impossibility result of Proposition 9 relies on trying to incentivize an unknowledgeable worker to act as desired. Since no mechanism can be incentive compatible for unknowledgeable workers, we will now consider only workers who are knowledgeable. The following proposition shows that the strong-no-free-lunch axiom is too strong even for this relaxed setting.

Proposition 10. *When $G < N$, there exists no mechanism that is incentive-compatible for knowledgeable workers and satisfies the strong-no-free-lunch axiom.*

Given this impossibility result for $G < N$, we are left with $G = N$ which means that the true answers to all the questions are known a priori. This condition is clearly not applicable to a crowdsourcing setup; nevertheless, it is mathematically interesting and may be applicable to other scenarios such as testing and elicitation of beliefs about future events.

Proposition 11 below presents a mechanism for this case and proves its uniqueness. We previously saw that an unknowledgeable worker cannot be incentivized to skip all the questions (even when $G = N$). Thus, in our payment mechanism, we do the next best thing: Incentivize the unknowledgeable worker to answer only one question, that which she is most confident about, while incentivizing the knowledgeable worker to answer questions for which her confidence is greater than T and skip those for which her confidence is smaller than T.

Proposition 11. *Let C be the number of correct answers and W be the number of wrong answers (in the gold standard). Let the payment be $\alpha_{\min}$ if $W > 0$ or $C = 0$, and be $(\alpha_{\max} - \alpha_{\min})T^{G-C} + \alpha_{\min}$ otherwise. Under this mechanism, when $G = N$, an unknowledgeable worker is incentivized to answer only one question, that for which her confidence is the maximum, and a knowledgeable worker is incentivized to answer the questions for which her confidence is greater than T and skip those for which her confidence is smaller than T. Furthermore, when $G = N$, this mechanism is the one and only mechanism that obeys the strong-no-free-lunch axiom and is incentive-compatible for knowledgeable workers.*

Proofs

In the remainder of this section, we prove the various claims regarding the strong no-free-lunch axiom studied in this section.

Proof of Proposition 9

If the worker skips all questions, then the expected payment is zero under the strong-no-free-lunch axiom. On the other hand, in order to incentivize knowledgeable workers to select answers whenever their confidences are greater than T, there must exist some situation in which the payment is strictly larger than zero. Suppose the payment is strictly positive when questions $\{1, \ldots, z\}$ are answered correctly, questions $\{z+1, \ldots, z'\}$ are answered incorrectly, and the remaining questions are skipped. If the confidence of the unknowledgeable worker is in the interval $(0, T)$ for every question, then attempting to answer questions $\{1, \ldots, z'\}$ and skipping the rest fetches her a payment that is strictly positive in expectation. Thus, this unknowledgeable worker is incentivized to answer at least one question.

Proof of Proposition 10

Consider a (knowledgeable) worker who has a confidence of $p \in (T, 1]$ for the first question, $q \in (0, 1)$ for the second question, and confidences in the interval $(0, T)$ for the remaining questions. Suppose the worker attempts to answer the first question (and selects the answer the believes is most likely to be correct) and skips the last $(N - 2)$ questions as desired. Now, in order to incentivize her to answer the second question if $q > T$ and skip the second question if $q < T$, the payment mechanism must satisfy

$$\begin{aligned} &pqg(1, 1, 0, \ldots, 0) + (1 - p)qg(-1, 1, 0, \ldots, 0) + p(1 - q)g(1, -1, 0, \ldots, 0) \\ &\quad + (1 - p)(1 - q)g(-1, -1, 0, \ldots, 0) \underset{q>T}{\overset{q<T}{\lessgtr}} pg(1, 0, 0, \ldots, 0) + (1 - p)g(-1, 0, 0, \ldots, 0) \ . \end{aligned}$$

For any real-valued variable q, and for any real-valued constants a, b and c,

$$aq \underset{q>c}{\overset{q<c}{\lessgtr}} b \quad \Rightarrow \quad ac = b \ .$$

As a result we have

$$\begin{aligned} &pTg(1, 1, 0, \ldots, 0) + (1 - p)Tg(-1, 1, 0, \ldots, 0) + p(1 - T)g(1, -1, 0, \ldots, 0) \\ &\quad + (1 - p)(1 - T)g(-1, -1, 0, \ldots, 0) - pg(1, 0, 0, \ldots, 0) - (1 - p)g(-1, 0, 0, \ldots, 0) = 0 \ . \end{aligned}$$

The left hand side of this equation is a polynomial in variable p and takes a value of zero for all values of p in a one-dimensional box $(T, 1]$. It follows that the monomials of this polynomial must be zero, and in particular the constant term must be zero:

$$Tg(-1, 1, 0, \ldots, 0) + (1 - T)g(-1, -1, 0, \ldots, 0) - g(-1, 0, 0, \ldots, 0) = 0 \ .$$

The strong-no-free-lunch axiom implies $f(-1, -1, 0, \ldots, 0) = f(-1, 0, \ldots, 0) = f(0, \ldots, 0) = 0$, and hence $g(-1, -1, 0, \ldots, 0) = g(-1, 0, 0, \ldots, 0) = 0$. Since $T \in (0, 1)$, we have

$$\begin{aligned} 0 &= g(-1, 1, 0, \ldots, 0) \\ &= c_1 f(-1, 1, 0, \ldots, 0) + c_2 f(-1, 0, \ldots, 0) + c_2 f(1, 0, \ldots, 0) \ , \end{aligned} \tag{6.12}$$

for some constants $c_1 > 0$ and $c_2 > 0$ that represent the probability that the first two questions are included in the gold standard, and the probability that the first (or, second) but not the second (or, first) questions are included in the gold standard. Since f is a non-negative function, it must be that

$$f(1, 0, \ldots, 0) = 0 \ .$$

Now suppose a (knowledgeable) worker has a confidence of $p \in (T, 1]$ for the first question and confidences lower than T for the remaining $(N - 1)$ questions. Suppose the worker chooses to skip the last $(N-1)$ questions as desired. In order to incentivize the worker to answer the first question, the mechanism must satisfy for all $p \in (T,1]$,

$$\begin{aligned}0 &< pg(1,0,\ldots,0) + (1-p)g(-1,0,\ldots,0) - g(0,0,\ldots,0)\\ &= pc_3 f(1,0,\ldots,0) + pc_4 f(0,0,\ldots,0) + (1-p)c_3 f(-1,0,\ldots,0)\\ &\qquad + (1-p)c_4 f(0,0,\ldots,0) - f(0,0,\ldots,0)\\ &= 0,\end{aligned}$$

where $c_3 > 0$ and $c_4 > 0$ are some constants. The final equation is a result of the strong-no-free-lunch axiom and the fact that $f(1,0,\ldots,0) = 0$ as proved above. This yields a contradiction, and hence no incentive-compatible mechanism f can satisfy the strong-no-free-lunch axiom when $G < N$ even when allowed to address only knowledgeable workers.

Finally, as a sanity check, note that if $G = N$ then $c_2 = 0$ in (6.12). The proof above thus doesn't hold when $G = N$.

Proof of Proposition 11

We will first show that the mechanism works as desired.

First consider the case when the worker is unknowledgeable and her confidences are of the form $T > p_{(1)} \geq p_{(2)} \geq p_{(3)} \geq \cdots \geq p_{(G)}$. If she answers only the first question, then her expected payment is

$$\kappa \frac{p_{(1)}}{T} \ .$$

Let us now see her expected payment if she doesn't follow this answer pattern. The strong-no-free-lunch axiom implies that if the worker doesn't answer any question then her expected payment is zero. Suppose the worker chooses to answer questions $\{i_1, \ldots, i_z\}$. In that case, her expected payment is

$$\begin{aligned}\kappa \frac{p_{i_1} \cdots p_{i_z}}{T^z} &= \kappa \frac{p_{i_1}}{T} \cdots \frac{p_{i_z}}{T}\\ &\leq \kappa \left(\frac{p_{(1)}}{T}\right)^z \qquad (6.13)\\ &\leq \kappa \frac{p_{(1)}}{T} \ , \qquad (6.14)\end{aligned}$$

where (6.14) uses the fact that $p_{(1)} < T$. The inequality in (6.14) becomes an equality only when $z = 1$. Now when $z = 1$, the inequality in (6.13) becomes an equality only when $i_1 = (1)$. Thus the unknowledgeable worker is incentivized to answer only one question – the one that she has the highest confidence in.

Now consider a knowledgeable worker and suppose her confidences are of the form $p_{(1)} \geq \cdots \geq p_{(m)} > T > p_{(m+1)} \geq \cdots \geq p_{(G)}$ for some $m \geq 1$. If the worker answers questions $(1), \ldots, (m)$ as desired, her expected payment is

$$\kappa \frac{p_{(1)}}{T} \cdots \frac{p_{(m)}}{T} \ .$$

Now let us see what happens if the worker does not follow this answer pattern. The strong-no-free-lunch axiom implies that if the worker doesn't answer any question then her expected payment is zero. Now, if she answers some other set of questions, say questions $\{i_1, \ldots, i_z\}$ with $p_{(1)} \leq p_{i_1} < \cdots < p_{i_y} \leq p_{(m)} < p_{i_{y+1}} < \cdots p_{i_z} \leq p_{(G)}$. The expected payment in that case is

$$\begin{aligned} \kappa \frac{p_{i_1} \cdots p_{i_z}}{T^z} &= \kappa \frac{p_{i_1}}{T} \cdots \frac{p_{i_z}}{T} \\ &\leq \kappa \frac{p_{i_1}}{T} \cdots \frac{p_{i_y}}{T} \qquad (6.15) \\ &\leq \kappa \frac{p_{(1)}}{T} \cdots \frac{p_{(m)}}{T} \qquad (6.16) \end{aligned}$$

where inequality (6.15) is a result of $\frac{p_{i_j}}{T} \leq 1 \ \ \forall\, j > y$ and holds with equality only when $y = z$. Inequality (6.16) is a result of $\frac{p_{(j)}}{T} \geq 1 \ \ \forall\, j \leq m$ and holds with equality only when $y = m$. Thus the expected payment is maximized when $i_1 = (1), \ldots, i_z = (m)$ as desired. Finally, the payment strictly increases with an increase in the confidences, and hence the worker is incentivized to always consider the answer that she believes is most likely to be correct.

We will now show that this mechanism is unique.

The necessary conditions derived in Lemma 33, when restricted to the case of $G = N$ and $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \neq \{0\}^{N-1}$, is also applicable to the present setting. This is because the strong-no-free-lunch axiom assumed here is a stronger condition than the no-free-lunch axiom considered in Lemma 33, and moreover, $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \neq \{0\}^{N-1}$ avoids the use of unknowledgeable workers in the proof of Lemma 33. It follows that for every question $i \in \{1, \ldots, G\}$ and every $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \in \{-1, 0, 1\}^{G-1} \backslash \{0\}^{G-1}$, it must be that

$$\begin{aligned} &Tf(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_G) + (1-T) f(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_G) \\ &\qquad = f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_G) \ . \qquad (6.17) \end{aligned}$$

We claim that the payment must be zero whenever the number of incorrect answers $W > 0$. The proof proceeds by induction on the number of correct answers C. First suppose $C = 0$ (and $W > 0$). Then all questions are either wrong or skipped, and hence by the strong-no-free-lunch axiom, the payment must be zero. Now suppose the payment is necessarily zero whenever $W > 0$ and the total number of correct answers is $(C-1)$ or lower, for some $C \in [G-1]$. Consider any evaluation $(y_1, \ldots, y_G) \in \{-1, 0, 1\}^G$ in which the number of incorrect answers is more than zero and the number of correct answers is C. Suppose $y_i = 1$ for some $i \in [G]$, and $y_j = -1$ for some $j \in [G] \backslash \{i\}$. Then from the induction hypothesis, we have $f(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_G) = f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_G) = 0$. Applying (6.17) and noting that $T \in (0, 1)$, we get that $f(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_G) = 0$ as claimed. This result also allows us to simplify (6.17) to: For every question $i \in \{1, \ldots, G\}$ and every $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \in \{-1, 0, 1\}^{G-1} \backslash \{0\}^{G-1}$,

$$f(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_G) = \frac{1}{T} f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_G) \ . \qquad (6.18)$$

We now show that when $C > 0$ and $W = 0$, the payment must necessarily be of the form described in the statement of Proposition 11. The proof again proceeds via an induction on the number of correct answers C (≥ 1). Define a quantity $\kappa > 0$ as

$$\kappa = Tf(1, 0, \ldots, 0)\ . \tag{6.19}$$

Now consider the payment $f(1, y_2, \ldots, y_G)$ for some $(y_2, \ldots, y_G) \in \{0,1\}^{G-1} \backslash \{0\}^{G-1}$ with C correct answers. Applying (6.18) repeatedly (once for every i such that $y_i = 1$), we get

$$f(1, y_2, \ldots, y_G) = \frac{\kappa}{T^C}\ .$$

Unlike other results in this chapter, at this point we cannot claim the result to hold for all permutations of the questions. This is because we have defined the quantity κ in an asymmetric manner (6.19), in terms of the payment function when the *first* question is correct and the rest are skipped. In what follows, we will prove that the result claimed in the statement of Proposition 11 indeed holds for all permutations of the questions.

From equation (6.18) we have

$$\begin{aligned} f(0, 1, 0, \ldots, 0) &= Tf(1, 1, 0, \ldots, 0) \\ &= f(1, 0, 0, \ldots, 0) \\ &= \kappa\ . \end{aligned}$$

Thus the payment must be κ even if the second answer in the gold standard is correct and the rest are skipped. In fact, the argument holds when any one answer in the gold standard is correct and the rest are skipped. Thus the definition of κ is not restricted to the first question alone as originally defined in (6.19), but holds for all permutations of the questions. This allows the other arguments above to be applicable to any permutation of the questions. Finally, the budget constraint of $\alpha_{\max}$ fixes the value of κ to that claimed, thereby completing the proof.

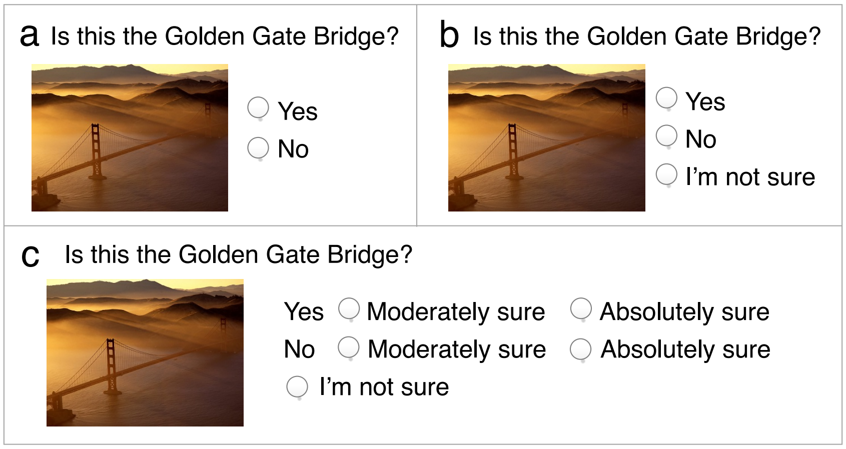
Chapter 7

Eliciting Confidences

> *"Be confident about your abilities. Be aware of your limitations."*
>
> – Vikram Sarabhai

7.1 Introduction

In this chapter, we consider crowdsourcing settings where the worker must perform a "task" comprising multiple questions. We assume that each question is objective, meaning that it has exactly one correct answer. In addition to eliciting the answers that the worker thinks



a Is this the Golden Gate Bridge?

○ Yes
○ No

b Is this the Golden Gate Bridge?

○ Yes
○ No
○ I'm not sure

c Is this the Golden Gate Bridge?

Yes ○ Moderately sure ○ Absolutely sure
No ○ Moderately sure ○ Absolutely sure
○ I'm not sure

Figure 7.1: An interface to elicit the answer and the worker's (quantized) confidence level: (a) standard interface used in crowdsourcing, (b) skip-based interface of Chapter 6 which forms a special case of the setting of this chapter, and (c) the confidence-based interface considered in this chapter.

are correct, we also wish to elicit the worker's confidence for each question; see Figure 7.1 for an illustration. For every question, the worker is asked for a quantized confidence level associated to the answer that he/she provides – for instance, by asking the worker to indicate whether he/she has a mild, moderate, or high confidence for his/her answer as illustrated in Figure 7.1(c). We term this setting as a "confidence-based" setting.

The goal is to design payment mechanisms that are "incentive compatible", that is, incentivize the worker to report the answer they think is most likely to be correct and their own confidence level honestly. In general, there may be many mechanisms which may be incentive compatible, we wish to choose a mechanism for deployment in a principled manner, and to that end, we propose a mild and natural "no-free-lunch" requirement on any payment mechanism. We then design a mechanism that takes a "multiplicative" form and show that our mechanism is the only mechanism that is possible.

For the reader who read Chapter 6, let us also provide a brief comparison with the setting of the skip-based setting considered therein. In Chapter 6 we considered an interface where the worker may answer any question if his/her confidence is high enough, or skip it otherwise. Given that we are asking the worker to make decisions based on his/her confidence, in this chapter we aim to directly elicit the confidence of the worker. The no-free-lunch axiom of this chapter is even weaker than that of Chapter 6: the no-free-lunch axiom of this chapter requires the minimum possible payment only if the worker indicates the *highest confidence level* for *all* the questions she attempts *and* if *all* these responses are incorrect.

Organization. The organization of this chapter is as follows. We present the formal problem setting in Section 7.2. In Section 7.3, we construct a mechanism for the confidence-based setting, which takes a multiplicative form, and prove its uniqueness. In Section 7.4 we present experiments using data from Amazon Mechanical Turk to evaluate the potential of our setting and algorithm to work in practice. We present a concluding discussion in Section 7.5. Finally, in Section 7.6 we provide proofs of our theoretical results.

7.2 Problem setting

We retain all of the notation and terminology from Chapter 6 which we reproduce here for completeness; we extend the formulation of Chapter 6 from the skip-based to a confidence-based setting.

General setting and terminology

In the crowdsourcing setting that we consider, one or more workers perform a *task*, where a task consists of multiple *questions*. The questions are objective, by which we mean, each question has precisely one correct answer. Examples of objective questions include multiple-choice classification questions such as Figure 7.1, questions on transcribing text from audio or images, etc.

For any possible answer to any question, we define the worker's *confidence about an answer* as the probability, according to her belief, of this answer being correct. In other words, one can assume that the worker has (in her mind) a probability distribution over all possible answers to a question, and the confidence for an answer is the probability of that answer being correct. As a shorthand, we also define the *confidence about a question* as the confidence for the answer that the worker is most confident about for that question. We assume that the worker's confidences for different questions are independent.

Let N denote the total number of questions in the task. Among these questions, we assume the existence of some "gold standard" questions, that is, a set of questions whose answers are known to the requester. Let G $(1 \leq G \leq N)$ denote the number of gold standard questions. The G gold standard questions are assumed to be distributed uniformly at random in the pool of N questions (of course, the worker does not know which G of the N questions form the gold standard). The payment to a worker for a task is computed after receiving her responses to all the questions in the task. The payment is based on the worker's performance on the gold standard questions. Since the payment is based on known answers, the payments to different workers do not depend on each other, thereby allowing us to consider the presence of only one worker without any loss in generality.

Confidence-based setting

In the confidence-based setting that we consider in this chapter, for each question, the worker can either 'skip' the question or provide an answer, and in the latter case, indicate her confidence for this answer as a number in $\{1, \ldots, L\}$. We term this indicated confidence as the 'confidence-level'. Here, L represents the highest confidence-level, and 'skip' is considered to be a confidence-level of 0. For instance, the interface of Figure 7.1c has $L = 3$. [1] Note that we do not solicit an answer if the worker indicates a confidence-level of 0 (skip), but the worker must provide an answer if she indicates a confidence-level of 1 or higher.

The reader who has read Chapter 6 will see from the aforementioned definition that the confidence-based setting is a generalization of the skip-based setting. The skip-based setting corresponds to $L = 1$.

The goal is to ensure that for a given set of intervals that partition $[0, 1]$, for every question the worker is incentivized to indicate 'skip' or choose the appropriate confidence-level when her confidence for that question falls in the corresponding interval. We specify these intervals by means of a set of "threshold" parameters $\{S_l, T_l\}_{l=1}^{L}$ that determine the confidence-levels that the workers should indicate. We assume that these thresholds are specified to us, and will use them to design the payment mechanism in a principled manner. In particular, we will require specification of two reference points for each confidence level, and this specification generalizes the skip-based setting.

[1] When the task is presented to the workers, the word 'skip' or the numbers $\{1, \ldots, L\}$ are replaced by more comprehensible phrases such as "I don't know", "moderately sure", "absolutely sure", etc.

- The first set of thresholds specifies a comparison of any confidence level with the skipping option as a fixed reference. To this end, recall that in the skip-based setting, the threshold T specified when the worker should skip a question and when she should attempt to answer. This is generalized to the confidence-based setting where for every level $l \in [L]$, a fixed threshold S_l specifies the 'strength' of confidence-level l: If restricted to only the two options of skipping or selecting confidence-level l for any question, the worker should be incentivized to select confidence-level l if her confidence is higher than S_l and skip if her confidence is lower than S_l.

- The second set of thresholds specifies a comparison of any confidence level with its neighbors. If a worker decides to not skip a question, she must choose one of multiple confidence-levels. A set $\{T_l\}_{l=1}^{L}$ of thresholds specify the boundaries between different confidence-levels. In particular, when the confidence of the worker for a question lies in (T_{l-1}, T_{l+1}), then the worker must be incentivized to indicate confidence-level $(l-1)$ if her confidence is lower than T_l and to indicate confidence-level l if her confidence is higher than T_l. This includes selecting level L if her confidence is higher than T_L and selecting level 0 if her confidence is lower than T_1.

We will call a payment mechanism as incentive-compatible if it satisfies the two requirements listed above, and also incentivizes the worker to select the answer that she believes is most likely to be correct for every question for which her confidence is higher than T_1.

The problem setting inherently necessitates certain restrictions in the choice of the thresholds. Since we require the worker to choose a higher level when her confidence is higher, the thresholds must necessarily be monotonic and satisfy $0 < S_1 < S_2 < \cdots < S_L < 1$ and $0 < T_1 < T_2 < \cdots < T_L < 1$. Also observe that the definitions of S_1 and T_1 coincide, and hence $S_1 = T_1$. Additionally, we can show (see Appendix 7.A) that for incentive-compatible mechanisms to exist, it must be that $T_l > S_l \ \forall \ l \in \{2, \ldots, L\}$. As a result, the thresholds must also satisfy $T_1 = S_1, \ T_2 > S_2, \ldots, T_L > S_L$. These thresholds may be chosen based on various factors of the problem at hand, for example, on the post-processing algorithms, any statistics on the distribution of worker abilities, budget constraints, etc. In this chapter, we will assume that these values are given to us.

Payment function

Let $x_1, \ldots, x_G$ denote the evaluations of the answers that the worker gives to the G gold standard questions, and let f denote the scoring rule, i.e., a function that determines the payment to the worker based on these evaluations $x_1, \ldots, x_G$.

We let any answer $i \in [G]$ take values in the set $x_i \in \{-L, \ldots, +L\}$. Here, we set $x_i = 0$ if the worker skipped the question, and for $l \in \{1, \ldots, L\}$, we set $x_i = l$ if the question was answered correctly with confidence l and $x_i = -l$ if the question was answered incorrectly with confidence l. The function $f : \{-L, \ldots, +L\}^G \to \mathbb{R}$ specifies the payment to be made to the worker.

As before, the payment is further associated to two parameters, $\alpha_{\max}$ and $\alpha_{\min}$. The parameter $\alpha_{\max}$ denotes the *budget*, i.e., the maximum amount that is paid to any individual worker for this task:

$$\max_{x_1,\ldots,x_G} f(x_1,\ldots,x_G) = \alpha_{\max}.$$

The amount $\alpha_{\max}$ is thus the amount of compensation paid to a perfect worker for her work. Further, one may often also have the requirement of paying a certain minimum amount to any worker. The parameter $\alpha_{\min}$ ($\leq \alpha_{\max}$) denotes this minimum payment: the payment function must also satisfy

$$\min_{x_1,\ldots,x_G} f(x_1,\ldots,x_G) \geq \alpha_{\min}.$$

For instance, crowdsourcing platforms today allow payments to workers, but do not allow imposing penalties: this condition gives $\alpha_{\min} = 0$.

We assume that the worker attempts to maximize her overall expected payment. In what follows, the expression 'the worker's expected payment' will refer to the expected payment from the worker's point of view, and the expectation will be taken with respect to the worker's confidences about her answers and the uniformly random choice of the G gold standard questions among the N questions in the task. A payment function f is called *incentive compatible* if the expected payment of the worker under this payment function is *strictly* maximized when the worker answers in the manner desired.[2]

Incentive compatibility

In the remainder of this section, we formally define the concepts of the worker's expected payment and incentive compatibility; the reader interested in understanding the chapter at a higher level may skip directly to the next section without loss in continuity.

Let Ω denote the set of options for each question. We assume that Ω is a finite set, for instance, the set $\{\text{Yes}, \text{No}\}$ for a task with binary-choice questions, or the set of all strings of at most a certain length for a task with textual responses. Let $Q \in [0,1]^{|\Omega| \times N}$ denote the beliefs of a worker for the N questions asked. Specifically, for any question $i \in [N]$ and any option $\omega \in \Omega$, let $Q_{\omega,i}$ represent the probability, according to the worker's belief, that option ω is the correct answer to question i. Then from the law of total probability, any valid Q must have $\sum_{\omega \in \Omega} Q_{\omega,i} = 1$ for every $i \in [N]$. The value of Q is unknown to the mechanism.

Let us first define the notion of the expected payment (from the worker's point of view) for any given response of the worker to the questions. For any question $i \in [N]$, suppose the worker indicates the confidence-level $\xi_i \in \{0,\ldots,L\}$. For every question $i \in [N]$ such that $\xi_i \neq 0$, let $\omega_i \in \Omega$ denote the option selected by the worker; whenever $\xi_i = 0$, indicating a skip, we let ω_i take any arbitrary value in Ω. Furthermore, let $p_i = Q_{\omega_i,i}$ denote the probability, according to the worker's belief, that the chosen option ω_i is the correct answer

[2] Such a notion of incentive compatibility is also called a strictly proper scoring rule.

to question i. For notational purposes, we also define a vector $E = (\epsilon_1, \ldots, \epsilon_G) \in \{-1, 1\}^G$. Then for the given responses, for the worker beliefs Q, and under payment mechanism f, the worker' expected payment $\Gamma_{Q,f} : (\{0, \ldots, L\} \times \Omega)^N \rightarrow \mathbb{R}$ is given by the expression:

$$\Gamma_{Q,f}(\xi_1,\ \omega_1,\ \ldots,\ \xi_N,\ \omega_N) \\ = \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G) \\ \subseteq\{1,\ldots,N\}}} \sum_{E\in\{-1,1\}^G} \left(f(\epsilon_1\xi_{j_1}, \ldots, \epsilon_G\xi_{j_G}) \prod_{i=1}^{G} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}} \right). \quad (7.1)$$

In the expression (7.1), the outermost summation corresponds to the expectation with respect to the randomness arising from the unknown positions of the gold standard questions. The inner summation corresponds to the expectation with respect to the worker's beliefs about the correctness of her responses. Note that the right hand side of (7.1) implicitly depends on $(\omega_1, \ldots, \omega_N)$ through the values $(p_1, \ldots, p_N)$. Also note that for every question i such that $\xi_i = 0$, the right hand side of (7.1) does not depend on the values of ω_i and p_i; this is because the choice $\xi_i = 0$ of skipping question i implies that the worker did not select any particular option.

We will now use the the definition of the expected payment of the worker to define the notion of incentive compatibility. To this end, for any valid probabilities Q, let $\mathcal{A}(Q) \subseteq (\{0, \ldots, L\} \times \Omega)^N$ denote an associated set of "desired" responses. By this we mean that every $a \in (\{0, \ldots, L\} \times \Omega)^N$ represents a possible response to the set of N questions, and the goal is to incentivize the worker to provide any one response in the set $\mathcal{A}(Q)$. Then a mechanism f is termed incentive compatible if

$$\Gamma_{Q,f}(a) > \Gamma_{Q,f}(a') \quad \text{for every } a \in \mathcal{A}(Q), \text{ every } a' \notin \mathcal{A}(Q), \text{ and every valid } Q.$$

The goal is to design mechanisms that are incentive compatible, that is, incentivize the workers to respond in a certain manner. The specific choice of "desired responses" for the skip-based and the confidence-based settings are discussed subsequently in their respective sections. We begin with the skip-based setting.

7.3 Main results

In this section we present the main theoretical results of this chapter. This makes the case of having only a 'skip' as considered in Chapter 6 a special case of this setting, and corresponds to $L = 1$.

7.3.1 No-free-lunch axiom

We now present a simple and natural requirement that we impose on any payment mechanism in order to enable us to narrow down on a good mechanism.

Axiom 2 (**Generalized-no-free-lunch axiom).** *If* all *the answers attempted by the worker in the gold standard are selected as the highest confidence-level (level L), and* all *of them turn out to be wrong, then the payment is $\alpha_{\min}$. More formally, we require the mechanism f to satisfy $f(x_1, \ldots, x_G) = \alpha_{\min}$ for every evaluation $(x_1, \ldots, x_G)$ that satisfies $0 < \sum_{i=1}^{G} \mathbf{1}\{x_i \neq 0\} = \sum_{i=1}^{G} \mathbf{1}\{x_i = -L\}$.*

Observe that the no-free-lunch axiom for the confidence-based setting introduced here is even weaker than that considered for the skip-based setting in Chapter 6. It reduces to the no-free-lunch axiom of the skip-based setting when $L = 1$, but more generally, is applicable only when the worker has selected the highest confidence level for every attempted question.

7.3.2 Payment mechanism

The proposed payment mechanism is described in Algorithm 2.

Algorithm 2 Incentive mechanism for the confidence-based setting

- Inputs:
  - ▸ Thresholds $S_1, \ldots, S_L$ and $T_1, \ldots, T_L$
  - ▸ Budget parameters $\alpha_{\max}$ and $\alpha_{\min}$
  - ▸ Evaluations $(x_1, \ldots, x_G) \in \{-L, \ldots, +L\}^G$ of the worker's answers to the G gold standard questions
- Set $\alpha_{-L}, \ldots, \alpha_L$ as
  - ▸ $\alpha_L = \frac{1}{S_L}$, $\alpha_{-L} = 0$
  - ▸ For $l \in \{L-1, \ldots, 1\}$,

$$\alpha_l = \frac{(1-S_l)T_{l+1}\alpha_{l+1} + (1-S_l)(1-T_{l+1})\alpha_{-(l+1)} - (1-T_{l+1})}{T_{l+1} - S_l} \quad \text{and} \quad \alpha_{-l} = \frac{1 - S_l\alpha_l}{1 - S_l}$$

  - ▸ $\alpha_0 = 1$
- The payment is

$$f(x_1, \ldots, x_G) = \kappa \prod_{i=1}^{G} \alpha_{x_i} + \alpha_{\min}$$

  where $\kappa = (\alpha_{\max} - \alpha_{\min}) \left(\frac{1}{\alpha_L}\right)^G$.

The following theorem shows that this mechanism indeed incentivizes a worker to select answers and confidence-levels as desired.

Theorem 18. *The mechanism of Algorithm 2 is incentive-compatible and satisfies the generalized-no-free-lunch axiom.*

Remark 2. *The mechanism of Algorithm 2 also ensures a condition stronger than the 'boundary-based' definition of the thresholds $\{T_l\}_{l\in[L]}$ given earlier. Under this mechanism, for every $l \in [L-1]$ the worker is incentivized to select confidence-level l (over all else) whenever her confidence lies in the interval (T_l, T_{l+1}), select confidence-level 0 (over all else) whenever her confidence is lower than T_1 and select confidence-level L (over all else) whenever her confidence is higher than T_L.*

7.3.3 Uniqueness of this mechanism

We prove that the mechanism of Algorithm 2 is unique, that is, no other incentive-compatible mechanism can satisfy the generalized-no-free-lunch axiom.

Theorem 19. *The payment mechanism of Algorithm 2 is the only incentive-compatible mechanism that satisfies the generalized-no-free-lunch axiom.*

The proof of Theorem 19 is conceptually similar to that of the skip-based setting considered in the previous chapter, but involves resolving several additional complexities that arise due to elicitation from multiple confidence levels.

7.4 Experiments

In this section, we review experiments conducted in [235] on the Amazon Mechanical Turk crowdsourcing platform to evaluate the performance of the new skip-based (Chapter 6) and confidence-based (this chapter) interfaces and mechanisms in real-world scenarios. The data collection procedure is described in detail in Appendix B of paper [235].

There are nine experiments (tasks) ranging from image annotation to text and speech recognition. All experiments involve objective questions, and the responses elicited were multiple choice in five of the experiments and freeform text in the rest. For each experiment, three settings were tested: (i) the baseline conventional setting (Figure 7.1a) with a mechanism of paying a fixed amount per correct answer, (ii) our skip-based setting (Figure 7.1b) with our multiplicative mechanism, and (iii) our confidence-based setting (Figure 7.1c) with our confidence-based mechanism. For each mechanism in each experiment, 35 workers independently performed the task, thus amounting to a total of 945 worker-tasks.

Results: Raw data Figure 7.2 plots, for the baseline, skip-based and confidence-based mechanisms for all nine experiments, the (i) fraction of questions that were answered incorrectly, (ii) fraction of questions that were answered incorrectly among those that were attempted, (iii) the average payment to a worker (in cents), and (iv) break up of the answers in terms of the fraction of answers in each confidence level. The payment for various tasks

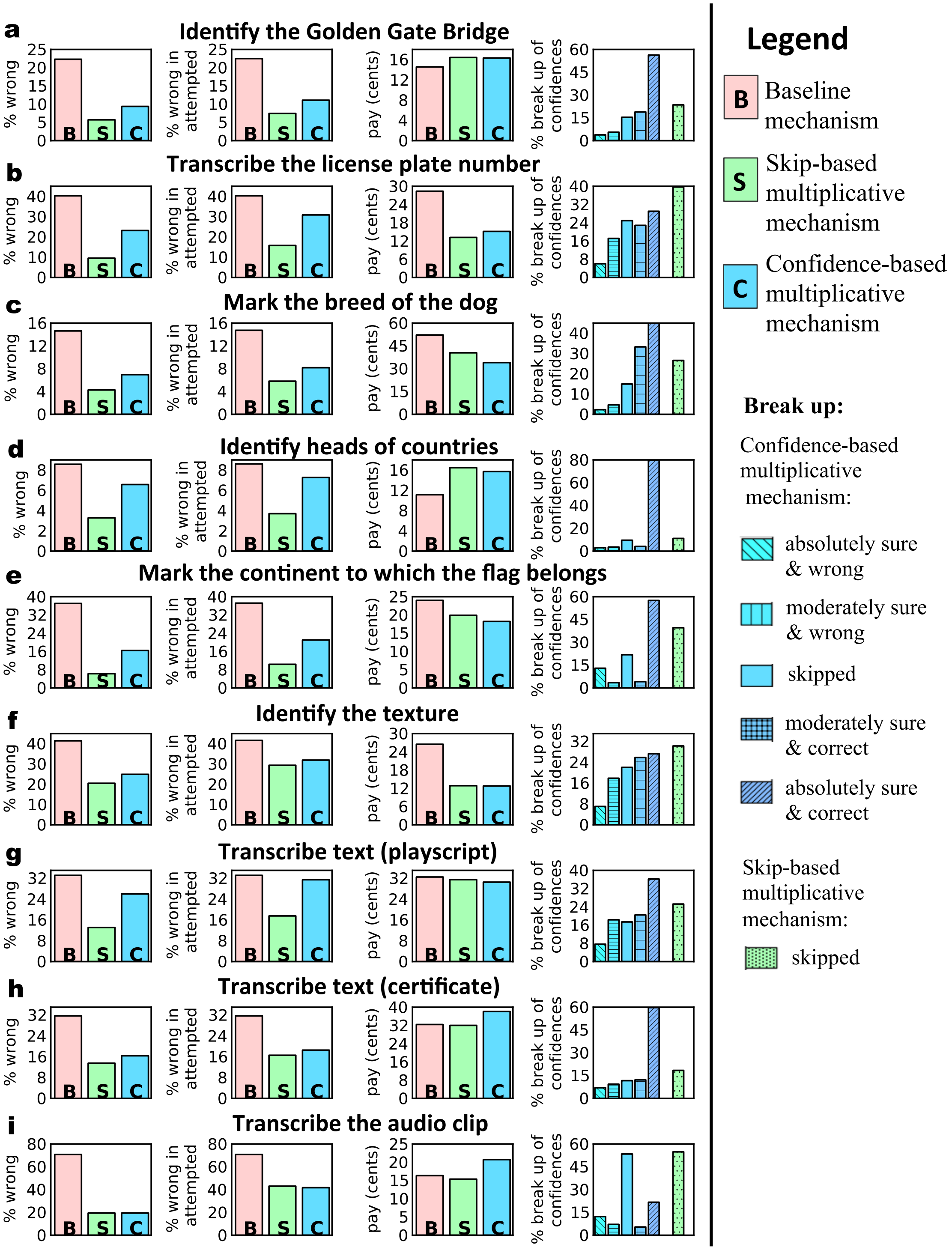


Figure 7.2: The error-rates in the raw data and payments in the nine experiments indexed as (a)–(i). Each individual bar in the plots corresponds to one mechanism in one experiment and is generated from 35 distinct workers (this totals to 945 worker-tasks).

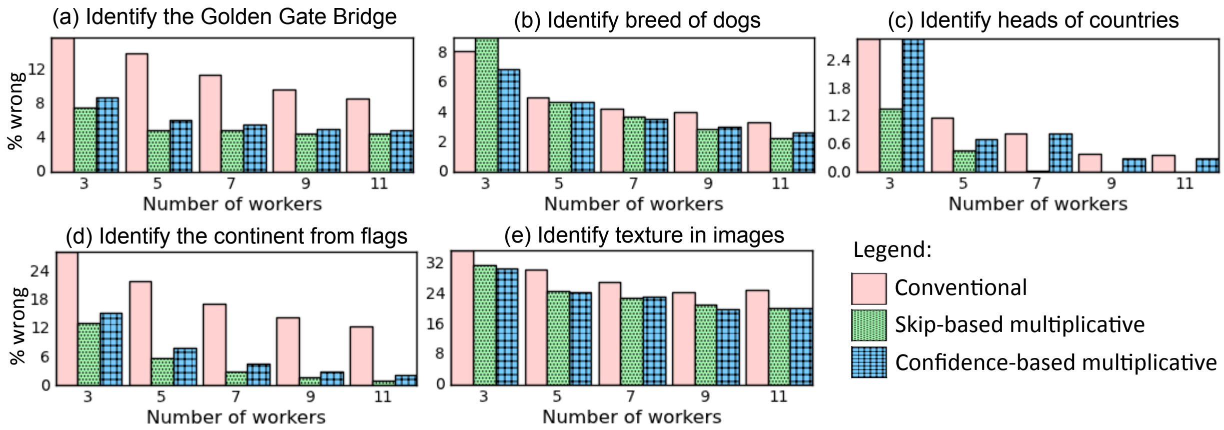


Figure 7.3: Error-rates in the aggregated data in the five experiments involving multiple-choice tasks.

plotted in Figure 7.2 is computed as the average of the payments across 100 (random) selections of the gold standard questions, in order to prevent any distortion of the results due to the randomness in the choice of the gold standard questions.

The figure shows that the amount of errors among the attempted questions is much lower in the skip and the confidence-based settings than the baseline setting. Also observe that in the confidence-based setting, as expected, the answers selected with higher confidence-levels are more correct. The total money spent under each of these settings is similar, with the skip and the confidence-based settings faring better in most cases. We also elicited feedback from the workers, in which we received several positive comments (and no negative comments). Examples of comments that were received are: “I was wondering if it would possible to increase the maximum number of HITs I may complete for you. As I said before, they were fun to complete. I think I did a good job completing them, and it would be great to complete some more for you.”; “I am eagerly waiting for your bonus.”; “Enjoyable. Thanks.”

Results: Aggregated data We saw in the raw data that under the skip-based setting, the amount of error among the attempted questions was significantly lower than the amount of error in the baseline setting. However, the skip-based setting was also associated, by design, to lesser amount of data by virtue of questions being skipped by the workers. A natural question that arises is how the baseline and the skip-based mechanisms will compare in terms of the final data quality, i.e., the amount of error once data from multiple workers is aggregated.

To this end, we considered the five experiments that consisted of multiple-choice questions. We let a parameter `num_workers` take values in {3, 5, 7, 9, 11}. For each of the five experiments and for each of the five values of `num_workers`, we perform the following actions 1,000 times: for each question, we choose `num_workers` workers and perform a majority vote on their responses. If the correct answer for that question does not lie in the set of options given by the majority, we consider it as an accuracy of zero. Otherwise, if there are m options tied in the majority vote, and the correct answer is one of these m, then we consider it as an accuracy of $\frac{100}{m}\%$ (hence, 100% if the correct answer is the only answer picked by the majority vote). We average the accuracy across all questions and across all iterations.

We choose majority voting as the means of aggregation since (a) it is the simplest and still most popular aggregation method, and (b) to enable an apples-to-apples comparison design since while more advanced aggregation algorithms have been developed for the baseline setting without the skip (see, for instance, Chapter 4), design of analogous algorithms for the new skip-based setting hasn't been explored yet.

The results are presented in Figure 7.3. We see that in most cases, our skip-based mechanism induces a lower labelling error at the aggregate level than the baseline. Furthermore, in many of the instances, the reduction is two-fold or higher.

All in all, in the experiments, we observe a substantial reduction in the error-rates while expending the same or lower amounts and receiving no negative comments from the workers, suggesting that these mechanisms can work; the fundamental theory underlying the mechanisms ensures that the system cannot be gamed in the long run. Our proposed settings and mechanisms thus have the potential to provide much higher quality labeled data as input to machine learning algorithms.

7.5 Discussion

We conclude this chapter with a discussion on an important open problem – the choice of the number of confidence levels L. In the chapter we assume that the number of confidence levels L is specified to us, and we provide mechanisms for any given choice of L, but assume that the value of L is provided to us. There are at least two effects that one must account for when making this choice. An increase in the value of L is good because we may obtain additional nuanced information about the workers' beliefs. However, on the other hand, workers would now require a greater time and effort in order to provide select the confidence level, and moreover, the multitude of options may make their answers more noisy. In other words, both the "signal" and the "noise" in the data increase with an increase in the value of L, and lead to an interesting trade-off. It is also of interest to design means of choosing the thresholds $\{S_l, T_l\}_{l=1}^{L}$ once the value of L is set.

7.6 Proofs

In this section, we prove the claimed theoretical results whose proofs are not included in the main text of the chapter.

The property of incentive-compatibility does not change upon any shift of the mechanism by a constant value or any scaling by a positive constant value. As a result, for the purposes of these proofs, we can assume without loss of generality that $\alpha_{\min} = 0$.

7.6.1 Proof of Theorem 18: Working of Algorithm 2

We first state three properties that the constants $\{\alpha_l\}_{l=-L}^{L}$ defined in Algorithm 2 must satisfy.

Lemma 34. *For every $l \in \{0, \ldots, L-1\}$*

$$T_{l+1}\alpha_{l+1} + (1 - T_{l+1})\alpha_{-(l+1)} = T_{l+1}\alpha_l + (1 - T_{l+1})\alpha_{-l} \ , \tag{7.2}$$

and

$$S_{l+1}\alpha_{l+1} + (1 - S_{l+1})\alpha_{-(l+1)} = \alpha_0 = 1 \ . \tag{7.3}$$

Lemma 35. *$\alpha_L > \alpha_{L-1} > \cdots > \alpha_{-L} = 0$.*

Lemma 36. *For any $m \in \{1, \ldots, L\}$, any $p > T_m$ and any $z < m$,*

$$p\alpha_m + (1-p)\alpha_{-m} > p\alpha_z + (1-p)\alpha_{-z} \ , \tag{7.4}$$

and for any $m \in \{0, \ldots, L-1\}$, any $p < T_{m+1}$ and any $z > m$,

$$p\alpha_m + (1-p)\alpha_{-m} > p\alpha_z + (1-p)\alpha_{-z} \ . \tag{7.5}$$

The proof of these results are available at the end of this section. Given these lemmas, we now complete the proof of Theorem 18.

The choice of $\alpha_{-L} = 0$ made in Algorithm 2 ensures that the payment is zero whenever any answer in the gold standard evaluates to $-L$. This choice ensures that the no-free-lunch axiom is satisfied. One can easily verify that the payment lies in the interval $[0, \alpha_{\max}]$. It remains to prove that the proposed mechanism is incentive-compatible.

Define $E = (\epsilon_1, \ldots, \epsilon_G) \in \{-1, 1\}^G$ and $E_{\backslash 1} = (\epsilon_2, \ldots, \epsilon_G)$. Suppose the worker has confidences $p_1, \ldots, p_N$ for her N answers. For some $(s(1), \ldots, s(N)) \in \{0, \ldots, L\}^N$ suppose $p_i \in (T_{s(i)},\ T_{s(i)+1})\ \forall\ i \in \{1, \ldots, N\}$, i.e., $s(1), \ldots, s(N)$ are the correct confidence-levels for her answers. Consider any other set of confidence-levels $s'(1), \ldots, s'(N)$. When the mechanism of Algorithm 2 is employed, the expected payment (from the point of view of the worker) on selecting confidence-levels $s(1), \ldots, s(N)$ is

$$\mathbb{E}[\text{Pay}] = \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq\{1,\ldots,N\}}} \sum_{E\in\{-1,1\}^G} \prod_{i=1}^{G} \alpha_{\epsilon_i s(j_i)} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}} \tag{7.6}$$

$$= \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq\{1,\ldots,N\}}} \sum_{E_{\setminus 1}\in\{-1,1\}^{G-1}} \left(p_{j_1}\alpha_{s(j_1)} + (1-p_{j_1})\alpha_{-s(j_1)}\right) \prod_{i=2}^{G} \alpha_{\epsilon_i s(j_i)} (p_{j_i})^{\frac{1+\epsilon_i}{2}} (1-p_{j_i})^{\frac{1-\epsilon_i}{2}}$$

$$\vdots$$

$$= \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq\{1,\ldots,N\}}} \prod_{i=1}^{G} \left(p_{j_i}\alpha_{s(j_i)} + (1-p_{j_i})\alpha_{-s(j_i)}\right) \tag{7.7}$$

$$> \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq\{1,\ldots,N\}}} \prod_{i=1}^{G} \left(p_{j_i}\alpha_{s'(j_i)} + (1-p_{j_i})\alpha_{-s'(j_i)}\right)$$

which is the expected payment under any other set of confidence-levels $s'(1), \ldots, s'(N)$. The last inequality is a consequence of Lemma 36.

An argument similar to the above also proves that for any $m \in \{1, \ldots, L\}$, if allowed to choose between only skipping and confidence-level m, the worker is incentivized to choose confidence-level m over skip if her confidence is greater S_m, and choose skip over level m if if her confidence is smaller than S_m. Finally, from Lemma 35 we have $\alpha_L > \cdots > \alpha_{-L} = 0$. It follows that the expected payment (7.7) is strictly increasing in each of the values $p_1, \ldots, p_N$. Thus the worker is incentivized to report the answer that she thinks is most likely to be correct.

Proof of Lemma 34

Algorithm 2 states that $\alpha_{-l} = \frac{1-\alpha_l S_l}{1-S_l}$ for all $l \in [L]$. A simple rearrangement of the terms in this expression gives (7.3).

Towards the goal of proving (7.2), we will first prove an intermediate result:

$$\alpha_l > 1 > \alpha_{-l} \;\forall\; l \in \{L, \ldots, 1\}\ . \tag{7.8}$$

The proof proceeds via an induction on $l \in \{L, \ldots, 2\}$. The case of $l = 1$ will be proved separately. The induction hypothesis involves two claims: $\alpha_l > 1 > \alpha_{-l}$ and $T_l\alpha_l + (1 - T_l)\alpha_{-l} > 1$. The base case is $l = L$ for which we know that $\alpha_L = \frac{1}{S_L} > 1 > 0 = \alpha_{-L}$ and $T_l\alpha_l + (1 - T_l)\alpha_{-l} = \frac{T_l}{S_l} > 1$. Now suppose that the induction hypothesis is true for $(l + 1)$. Rearranging the terms in the expression defining α_l in Algorithm 2 and noting that

$1 > T_{l+1} > S_l$, we get

$$\begin{aligned}\alpha_l &= \frac{(1-S_l)\left(T_{l+1}\alpha_{l+1} + (1-T_{l+1})\alpha_{-(l+1)}\right) - (1-T_{l+1})}{(1-S_l)-(1-T_{l+1})} \\ &> \frac{(1-S_l)-(1-T_{l+1})}{(1-S_l)-(1-T_{l+1})} \\ &= 1\ .\end{aligned} \tag{7.9}$$

From (7.3) we see that the value 1 is a convex combination of α_l and α_{-l}. Since $\alpha_l > 1$ and $S_l \in (0,1)$, it must be that $\alpha_{-l} < 1$. Furthermore, since $T_l > S_l$ we get

$$\begin{aligned}T_l\alpha_l + (1-T_l)\alpha_{-l} &> S_l\alpha_l + (1-S_l)\alpha_{-l} \\ &= 1\ .\end{aligned}$$

This proves the induction hypothesis. Let us now consider $l = 1$. If $L = 1$ then we have $\alpha_L = \frac{1}{S_L} > 1 > 0 = \alpha_{-L}$ and we are done. If $L > 1$ then we have already proved that $\alpha_2 > 1 > \alpha_{-2}$ and $T_2\alpha_2 + (1-T_2)\alpha_{-2} > 1$. An argument identical to (7.9) onwards proves that $\alpha_1 > 1 > \alpha_{-1}$.

Now that we have proved $\alpha_l > \alpha_{-l} \forall\ l \in [L]$, we can rewrite the expression defining α_{-l} in Algorithm 2 as

$$S_l = \frac{1-\alpha_{-l}}{\alpha_l - \alpha_{-l}}\ .$$

Substituting this expression for S_l in the definition of α_l in Algorithm 2 and making some simple rearrangements gives the desired result (7.2).

Proof of Lemma 35

We have already shown (7.8) in the proof of Lemma 34 above that $\alpha_l > 1 > \alpha_{-l}\ \forall\ l \in [L]$.

Next we will show that $\alpha_{l+1} > \alpha_l$ and $\alpha_{-(l+1)} < \alpha_{-l}\ \forall\ l \geq 0$. First consider $l = 0$, for which Algorithm 2 sets $\alpha_0 = 1$, and we have already proved that $\alpha_1 > 1 > \alpha_{-1}$.

Now consider some $l > 0$. Observe that since $S_l\alpha_l + (1-S_l)\alpha_{-l} = 1$ (Lemma 34), $S_{l+1} > S_l$ and $\alpha_l > \alpha_{-l}$, it must be that

$$S_{l+1}\alpha_l + (1-S_{l+1})\alpha_{-l} > 1\ . \tag{7.10}$$

From Lemma 34, we also have

$$S_{l+1}\alpha_{l+1} + (1-S_{l+1})\alpha_{-(l+1)} = 1\ . \tag{7.11}$$

Subtracting (7.10) from (7.11) we get

$$S_{l+1}(\alpha_{l+1} - \alpha_l) + (1-S_{l+1})(\alpha_{-(l+1)} - \alpha_{-l}) < 0\ . \tag{7.12}$$

From Lemma 34 we also have

$$T_{l+1}\alpha_{l+1} + (1 - T_{l+1})\alpha_{-(l+1)} = T_{l+1}\alpha_l + (1 - T_{l+1})\alpha_{-l} \tag{7.13}$$

$$\Rightarrow \quad T_{l+1}(\alpha_{l+1} - \alpha_l) + (1 - T_{l+1})(\alpha_{-(l+1)} - \alpha_{-l}) = 0 \ . \tag{7.14}$$

Subtracting (7.12) from (7.14) gives

$$(T_{l+1} - S_{l+1})[(\alpha_{l+1} - \alpha_l) + (\alpha_{-l} - \alpha_{-(l+1)})] > 0 \ . \tag{7.15}$$

Since $T_{l+1} > S_{l+1}$ by definition, it must be that

$$\alpha_{l+1} - \alpha_l > \alpha_{-(l+1)} - \alpha_{-l} \ . \tag{7.16}$$

Now, rearranging the terms in (7.13) gives

$$(\alpha_{l+1} - \alpha_l)T_{l+1} = -(\alpha_{-(l+1)} - \alpha_{-l})(1 - T_{l+1}) \ . \tag{7.17}$$

Since $T_{l+1} \in (0, 1)$, it follows that the terms $(\alpha_{l+1} - \alpha_l)$ and $(\alpha_{-(l+1)} - \alpha_{-l})$ have opposite signs. Using (7.16) we conclude that $\alpha_{l+1} - \alpha_l > 0$ and $\alpha_{-(l+1)} - \alpha_{-l} < 0$.

Proof of Lemma 36

Let us first prove (7.4). First consider the case $z = m - 1$. From Lemma 34 we know that

$$T_m\alpha_{m-1} + (1 - T_m)\alpha_{-(m-1)} \quad = \quad T_m\alpha_m + (1 - T_m)\alpha_{-m},$$

and with some rearrangement,

$$\begin{aligned} 0 &= T_m(\alpha_m - \alpha_{m-1}) + T_m(\alpha_{-(m-1)} - \alpha_{-m}) - (\alpha_{-(m-1)} - \alpha_{-m}) \\ &< p(\alpha_m - \alpha_{m-1}) + p(\alpha_{-(m-1)} - \alpha_{-m}) - (\alpha_{-(m-1)} - \alpha_{-m}) \ , \end{aligned} \tag{7.18}$$

where (7.18) is a consequence of $p > T_m$ and Lemma 35. A simple rearrangement of the terms in (7.18) gives (7.4). Now, for any $z < m$, recursively apply this result to get

$$\begin{aligned} p\alpha_m + (1 - p)\alpha_{-m} &> p\alpha_{m-1} + (1 - p)\alpha_{-(m-1)} \\ &> p\alpha_{m-2} + (1 - p)\alpha_{-(m-2)} \\ &\vdots \\ &> p\alpha_z + (1 - p)\alpha_{-z} \ . \end{aligned}$$

Let us now prove (7.5). We first consider the case $z = m + 1$. From Lemma 34 we know that

$$\begin{aligned} T_{m+1}\alpha_m + (1 - T_{m+1})\alpha_{-m} &= T_{m+1}\alpha_{m+1} + (1 - T_{m+1})\alpha_{-(m+1)} \\ \Rightarrow \qquad 0 &= T_{m+1}(\alpha_{m+1} - \alpha_m) + T_{m+1}(\alpha_{-m} - \alpha_{-(m+1)}) - (\alpha_{-m} - \alpha_{-(m+1)}) \\ &> p(\alpha_{m+1} - \alpha_m) + p(\alpha_{-m} - \alpha_{-(m+1)}) - (\alpha_{-m} - \alpha_{-(m+1)}) \ , \end{aligned} \tag{7.19}$$

where (7.19) is a consequence of $p < T_{m+1}$ and Lemma 35. A simple rearrangement of the terms in (7.19) gives (7.5). For any $z > m$, applying this result recursively gives

$$\begin{aligned} p\alpha_m + (1-p)\alpha_{-m} &> p\alpha_{m+1} + (1-p)\alpha_{-(m+1)} \\ &> p\alpha_{m+2} + (1-p)\alpha_{-(m+2)} \\ &\vdots \\ &> p\alpha_z + (1-p)\alpha_{-z} \ . \end{aligned}$$

7.6.2 Proof of Theorem 19: Uniqueness of Algorithm 2

We will first define one additional piece of notation. Let $g : \{-L, \ldots, L\}^N \rightarrow \mathbb{R}_+$ denote the expected payment given an evaluation of the N answers, where the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1, \ldots, x_N) \in \{-L, \ldots, L\}^N$ are the evaluations of the worker's answers to the N questions then the expected payment is

$$g(x_1, \ldots, x_N) = \frac{1}{\binom{N}{G}} \sum_{(i_1, \ldots, i_G) \subseteq \{1, \ldots, N\}} f(x_{i_1}, \ldots, x_{i_G}) \ . \tag{7.20}$$

Notice that when $G = N$, the functions f and g are identical.

The proof of uniqueness is based on a certain condition necessitated by incentive-compatibility stated in the form of Lemma 37 below. Note that this lemma does *not* require the generalized-no-free-lunch axiom, and may be of independent interest.

Lemma 37. *Any incentive-compatible mechanism must satisfy, for every question $i \in \{1, \ldots, G\}$, every $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_G) \in \{-L, \ldots, L\}^{G-1}$, and every $m \in \{1, \ldots, L\}$,*

$$\begin{aligned} &T_m f(y_1, \ldots, y_{i-1}, m, y_{i+1}, \ldots, y_G) + (1 - T_m) f(y_1, \ldots, y_{i-1}, -m, y_{i+1}, \ldots, y_G) \\ &\quad = T_m f(y_1, \ldots, y_{i-1}, m-1, y_{i+1}, \ldots, y_G) + (1 - T_m) f(y_1, \ldots, y_{i-1}, -(m-1), y_{i+1}, \ldots, y_G) \end{aligned} \tag{7.21a}$$

and

$$\begin{aligned} &S_m f(y_1, \ldots, y_{i-1}, m, y_{i+1}, \ldots, y_G) + (1 - S_m) f(y_1, \ldots, y_{i-1}, -m, y_{i+1}, \ldots, y_G) \\ &\quad = f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_G) \ . \end{aligned} \tag{7.21b}$$

Note that (7.21a) and (7.21b) coincide when $m = 1$, since $T_1 = S_1$ by definition. See the end of this section for a proof of this lemma.

We first prove that any incentive compatible mechanism that satisfies the no-free-lunch axiom must give a zero payment when one or more questions are selected with a confidence

L and turn out to be incorrect. Let us assume for now that in the G questions in the gold standard, the first question is answered incorrectly with a confidence of L, the next $(G-1-S)$ questions are answered by the worker and have arbitrary evaluations, and the remaining S questions are skipped. The proof proceeds by an induction on S. If $S = G-1$, the only attempted question is the first question and this is incorrect with confidence L. The no-free-lunch axiom necessitates a zero payment in this case, thus satisfying the base case of our induction hypothesis. Now we prove the hypothesis for some S under the assumption that the hypothesis is true for every $S' > S$. From Lemma 33 with $m = 1$, we have

$$\begin{aligned} &T_1 f(-L, y_2, \ldots, y_{G-S-1}, 1, 0, \ldots, 0) + (1-T_1) f(-L, y_2, \ldots, y_{G-S-1}, -1, 0, \ldots, 0) \\ &\qquad = T_1 f(-L, y_2, \ldots, y_{G-S-1}, 0, 0, \ldots, 0) \ + (1-T_1) f(-L, y_2, \ldots, y_{G-S-1}, 0, 0, \ldots, 0) \\ &\qquad = f(-L, y_2, \ldots, y_{G-S-1}, 0, 0, \ldots, 0) \\ &\qquad = 0 \ , \end{aligned} \tag{7.22}$$

where the final equation (7.22) is a consequence of our induction hypothesis given the fact that $f(-L, y_2, \ldots, y_{G-S-1}, 0, 0, \ldots, 0)$ corresponds to the case when the last $(S+1)$ questions are skipped and the first question is answered incorrectly with confidence L. Now, since the payment f must be non-negative and since $T \in (0,1)$, it must be that

$$f(-L, y_2, \ldots, y_{G-S-1}, 1, 0, \ldots, 0) = 0$$

and

$$f(-L, y_2, \ldots, y_{G-S-1}, -1, 0, \ldots, 0) = 0 \ .$$

Repeatedly applying the same argument to $m = 2, \ldots, L$ gives that for every value of m, it must be that $f(-L, y_2, \ldots, y_{G-S-1}, m, 0, \ldots, 0) = f(-L, y_2, \ldots, y_{G-S-1}, -m, 0, \ldots, 0) = 0$. This completes the proof of our induction hypothesis. Observe that each of the aforementioned arguments hold for any permutation of the G questions, thus proving the necessity of zero payment when any one or more answers are incorrect.

We now prove that when no answers in the gold standard are incorrect with confidence L, the payment must be of the form described in Algorithm 1. Let κ denote the payment when all G questions in the gold standard are skipped, i.e.,

$$\kappa = f(0, \ldots, 0) \ .$$

Now consider any $S \in \{0, \ldots, G-1\}$ and any $(y_1, \ldots, y_{G-S-1}, m) \in \{-L, \ldots, L\}^{G-S}$. The payments $\{f(y_1, \ldots, y_{G-S-1}, m, 0, \ldots, 0)\}_{m=-L}^{L}$ must satisfy the $(2L-1)$ linear constraints arising out of Lemma 37 and must also satisfy $f(y_1, \ldots, y_{G-S-1}, -L, 0, \ldots, 0) = 0$. This comprises a total of $2L$ linearly independent constraints on the $(2L+1)$ values $\{f(y_1, \ldots, y_{G-S-1}, m, 0, \ldots, 0)\}_{m=-L}^{L}$. The only set of solutions that meet these constraints are

$$f(y_1, \ldots, y_{G-S-1}, m, 0, \ldots, 0) = \alpha_m f(y_1, \ldots, y_{G-S-1}, 0, 0, \ldots, 0),$$

where the constants $\{\alpha_m\}_{m=-L}^{L}$ are as specified in Algorithm 2. Applying this argument G times, starting from $S = 0$ to $S = G - 1$, gives

$$f(y_1, \ldots, y_G) = \kappa \prod_{j=1}^{G} \alpha_{y_j} \ .$$

Finally, the budget requirement necessitates $\alpha_{\max} = \kappa \left(\alpha_L\right)^G$, which mandates the value of κ to be $\alpha_{\max}\left(\frac{1}{\alpha_L}\right)^G$. This is precisely the mechanism described in Algorithm 2.

Proof of Lemma 37: Necessary condition for any incentive-compatible mechanism

First consider the case of $G = N$. For every $j \in \{1, \ldots, i-1, i+1, \ldots, G\}$, define

$$r_j = \begin{cases} 1 - p_j & \text{if} \quad y_j \geq 0 \\ p_j & \text{if} \quad y_j < 0 \ . \end{cases}$$

Define $E_{\backslash i} = \{\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_G\}$. For any $l \in \{-L, \ldots, L\}$ let $\Lambda_l \in \mathbb{R}_+$ denote the expected payment (from the worker's point of view) when her answer to the i^{th} question evaluates to l:

$$\Lambda_l = \sum_{E_{\backslash i} \in \{-1,1\}^{G-1}} \left(f(y_1\epsilon_1, \ldots, y_{i-1}\epsilon_{i-1}, l, y_{i+1}\epsilon_{i+1}, \ldots, y_G\epsilon_G) \prod_{j \in [G]\backslash\{i\}} r_j^{\frac{1-\epsilon_j}{2}} (1 - r_j)^{\frac{1+\epsilon_j}{2}} \right) \quad (7.23)$$

Consider a worker who has confidences $\{p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_G\} \in (0,1)^{G-1}$ for questions $\{1, \ldots, i-1, i+1, \ldots, G\}$ respectively, and for question i suppose she has a confidence of $q \in (T_{m-1}, T_{m+1})$. For question i, we must incentivize the worker to select confidence-level m if $q > T_m$, and to select $(m-1)$ if $q < T_m$. This necessitates

$$q\Lambda_m + (1-q)\Lambda_{-m} \underset{q>T_m}{\overset{q<T_m}{\lessgtr}} q\Lambda_{m-1} + (1-q)\Lambda_{-(m-1)} \ . \quad (7.24)$$

Also, for question i, the requirement of level m having a higher incentive as compared to skipping when $q > S_m$ and vice versa when $q < S_m$ necessitates

$$q\Lambda_m + (1-q)\Lambda_{-m} \underset{q>S_m}{\overset{q<S_m}{\lessgtr}} \Lambda_0 \ . \quad (7.25)$$

Now, note that for any real-valued variable q, and for any real-valued constants a, b and c,

$$aq \underset{q>c}{\overset{q<c}{\lessgtr}} b \quad \Rightarrow \quad ac = b \ .$$

Applying this fact to (7.24) and (7.25) gives

$$(T_m\Lambda_m + (1-T_m)\Lambda_{-m}) - (T_m\Lambda_{m-1} + (1-T_m)\Lambda_{-(m-1)}) = 0\ , \tag{7.26}$$

$$(S_m\Lambda_m + (1-S_m)\Lambda_{-m}) - \Lambda_0 = 0\ . \tag{7.27}$$

From the definition of Λ_l in (7.23), we see that the left hand sides of (7.26) and (7.27) are both polynomials in $(G-1)$ variables $\{r_j\}_{j\in[G]\backslash\{i\}}$ and take a value of zero for all values of the variables in a $(G-1)$-dimensionall solid ball. Thus, each of the coefficients (of the monomials) in both polynomials must be zero, and in particular, the constant terms must also be zero. Observe that in both these polynomials, the constant term arises only when $\epsilon_j = 1\ \forall\ j \in [G]\backslash\{i\}$ (which makes the exponent of r_j to be 0 and that of $(1-r_j)$ to be 1). Thus, setting the constant term to zero in the two polynomials results in

$$\begin{aligned}&T_m f(y_1,\ldots,y_{i-1},m,y_{i+1},\ldots,y_G) + (1-T_m)f(y_1,\ldots,y_{i-1},-m,y_{i+1},\ldots,y_G)\\ &\quad = T_m f(y_1,\ldots,y_{i-1},m-1,y_{i+1},\ldots,y_G) + (1-T_m)f(y_1,\ldots,y_{i-1},-(m-1),y_{i+1},\ldots,y_G)\end{aligned} \tag{7.28}$$

and

$$\begin{aligned}&S_m f(y_1,\ldots,y_{i-1},m,y_{i+1},\ldots,y_G) + (1-S_m)f(y_1,\ldots,y_{i-1},-m,y_{i+1},\ldots,y_G)\\ &\quad = f(y_1,\ldots,y_{i-1},0,y_{i+1},\ldots,y_G)\end{aligned} \tag{7.29}$$

thus proving the claim for the case of $G = N$.

Now consider the case when $G < N$. In order to simplify notation, let us assume $i = 1$ without loss of generality (since the arguments presented hold for any permutation of the questions). Suppose a worker's answers to questions $\{2,\ldots,G\}$ evaluate to $(y_2,\ldots,y_G) \in \{-L,\ldots,L\}^{G-1}$, and further suppose that the worker skips the remaining $(N-G)$ questions. By going through arguments identical to those for $G = N$, but with f replaced by g, we get the necessity of

$$\begin{aligned}&T_m g(m,y_2,\ldots,y_G,0,\ldots,0) + (1-T_m)g(-m,y_2,\ldots,y_G,0,\ldots,0)\\ &\quad = T_m g(m-1,y_2,\ldots,y_G,0,\ldots,0) + (1-T_m)g(-(m-1),y_2,\ldots,y_G,0,\ldots,0)\end{aligned} \tag{7.30}$$

and

$$S_m g(m,y_2,\ldots,y_G,0,\ldots,0) + (1-S_m)g(-m,y_2,\ldots,y_G,0,\ldots,0) = g(0,y_2,\ldots,y_G,0,\ldots,0)\ . \tag{7.31}$$

We will now use this result in terms of function g to get an equivalent result in terms of function f. For some $S \in \{0,\ldots,G-1\}$, suppose $y_{G-S+1} = 0,\ldots,y_G = 0$. The remaining proof proceeds via an induction on S. We begin with $S = G-1$. In this case, (7.30) and (7.31) simplify to

$$\begin{aligned}&T_m g(m,0,\ldots,0) + (1-T_m)g(-m,0,0,\ldots,0)\\ &\quad = T_m g(m-1,0,\ldots,0) + (1-T_m)g(-(m-1),0,\ldots,0)\end{aligned}$$

and

$$S_m g(m,0,\ldots,0) + (1-S_m) g(-m,0,\ldots,0) = g(0,0,\ldots,0) \ .$$

Applying the definition of function g from (7.20) leads to

$$\begin{aligned}T_m\left(c_1 f(m,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right) + (1-T_m)\left(c_1 f(-m,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right)\\ = T_m\left(c_1 f(m-1,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right)\\ + (1-T_m)\left(c_1 f(-(m-1),0,\ldots,0) + c_2 f(0,0,\ldots,0)\right),\end{aligned}$$

and

$$\begin{aligned}S_m\left(c_1 f(m,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right) + (1-S_m)\left(c_1 f(-m,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right)\\ = \left(c_1 f(0,0,\ldots,0) + c_2 f(0,0,\ldots,0)\right)\end{aligned}$$

for constants $c_1 > 0$ and $c_2 > 0$ that respectively denote the probabilities that the first question is picked and not picked in the set of G gold standard questions. Canceling out the common terms on both sides of the equation, we get the desired results

$$\begin{aligned}T_m f(m,0,\ldots,0) + (1-T_m) f(-m,0,\ldots,0)\\ = T_m f(m-1,0,\ldots,0) + (1-T_m) f(-(m-1),0,\ldots,0)\end{aligned}$$

and

$$S_m f(m,0,\ldots,0) + (1-S_m) f(-m,0,\ldots,0) = f(0,0,\ldots,0) \ .$$

Next, we consider the case of a general $S \in \{0,\ldots,G-2\}$ and assume that the result is true when $y_{G-S} = 0,\ldots,y_G = 0$. In (7.30) and (7.31), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first where all of the first $(G-S)$ questions are included in the gold standard, and the second where one or more of the first $(G-S)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than S questions skipped in the gold standard, i.e., when $y_{G-S} = 0,\ldots,y_G = 0$, and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of (7.30) and (7.31). The remainder comprises only evaluations of function f for arguments in which the first $(G-S)$ questions are included in the gold standard: since the last $(N-G+S)$ questions are skipped by the worker, the remainder evaluates to

$$\begin{aligned}&T_m c_3 f(y_1,\ldots,y_{i-1},m,y_{i+1},\ldots,y_G) + (1-T_m) c_3 f(y_1,\ldots,y_{i-1},-m,y_{i+1},\ldots,y_G)\\ &\quad = T_m c_3 f(y_1,\ldots,y_{i-1},m-1,y_{i+1},\ldots,y_G) + (1-T_m) c_3 f(y_1,\ldots,y_{i-1},-(m-1),y_{i+1},\ldots,y_G) \ ,\\ &S_m c_3 f(y_1,\ldots,y_{i-1},m,y_{i+1},\ldots,y_G) + (1-S_m) c_3 f(y_1,\ldots,y_{i-1},-m,y_{i+1},\ldots,y_G)\\ &\quad = c_3 f(y_1,\ldots,y_{i-1},0,y_{i+1},\ldots,y_G) \ ,\end{aligned}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

7.A Appendix: Necessity of $T_l > S_l$ for the problem to be well defined

We now show that the restriction $T_l > S_l$ was necessary when defining the thresholds in Section 7.3.

Proposition 12. *Incentive-compatiblity necessitates $T_l > S_l \ \forall \ l \in \{2, \ldots, L\}$, even in the absence of the generalized-no-free-lunch axiom.*

First observe that the proof of Lemma 37 did not employ the generalized-no-free-lunch axiom, neither did it assume $T_l > S_l$. We will thus use the result of Lemma 37 to prove our claim.

Suppose the confidence of the worker for all but the first question is lower than T_1 and that the worker decides to skip all these questions. Suppose the worker attempts the first question. In order to ensure that the worker selects the answer that she believes is most likely to be true, it must be that

$$f(l, 0, \ldots, 0) > f(-l, 0, \ldots, 0) \quad \forall l \in [L] \ .$$

We now call upon Lemma 37 where we set $i = 1$, $m = l$, $y_2 = \ldots, y_G = 0$. Using the fact that $T_l > T_{l-1} \ \forall l \in \{2, \ldots, L\}$, we get

$$\begin{aligned}
T_l f(l, 0, \ldots, 0) &+ (1 - T_l) f(-l, 0, \ldots, 0) \\
&= T_l f(l-1, 0, \ldots, 0) + (1 - T_l) f(-(l-1), 0, \ldots, 0) \\
&> T_{l-1} f(l-1, 0, \ldots, 0) + (1 - T_{l-1}) f(-(l-1), 0, \ldots, 0) \\
&= T_{l-1} f(l-2, 0, \ldots, 0) + (1 - T_{l-1}) f(-(l-2), 0, \ldots, 0) \\
&> T_{l-2} f(l-2, 0, \ldots, 0) + (1 - T_{l-2}) f(-(l-2), 0, \ldots, 0) \\
&\vdots \\
&> T_1 f(1, 0, \ldots, 0) + (1 - T_1) f(-1, 0, \ldots, 0) \\
&= f(0, \ldots, 0) \\
&= S_l f(l, 0, \ldots, 0) + (1 - S_l) f(-l, 0, \ldots, 0).
\end{aligned}$$

Since $f(l, 0, \ldots, 0) > f(-l, 0, \ldots, 0)$, we have our desired result.

Chapter 8

Approval Voting

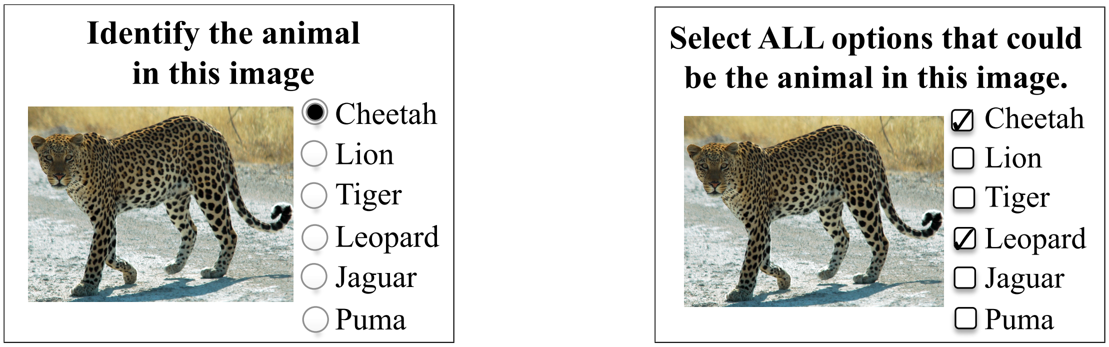
> *"Give them many options, give them many paths, give them the freedom that they may ask."*
>
> – Ada Lovelace

8.1 Introduction

In the big data era, with the ever increasing complexity of machine learning models such as deep learning, the demand for large amounts of labeled data is growing at an unprecedented scale. These labeling tasks used to be done by domain experts. It is not hard to imagine that the limited pool of experts would limit the size of the datasets. In the modern day, these massive labeling tasks are performed through commercial web services such as Amazon Mechanical Turk, where millions of crowdsourcing workers or annotators perform tasks in exchange for monetary payments [211]. Unfortunately, the data obtained via crowdsourcing is typically highly erroneous [119, 260, 261] due to the lack of expertise of workers, lack of appropriate incentives, and often the lack of an appropriate interface for the workers to express their knowledge. The typical crowdsourcing labeling task consists of a set of items such as images to be labeled, and each item is associated with a set of exclusive categories. Each worker is required to select the category that she believes is most likely to be correct. More formally, it involves eliciting the mode of the worker's belief. Such a "single-selection" crowdsourcing setting has been studied extensively, both empirically and theoretically.

In this chapter, we consider an alternative "approval-voting" means of eliciting labels from the workers. Approval voting [21, 120, 185, 264] is a form of voting in which each voter can "approve of" (that is, select) multiple candidates. No further preferences among these candidates is specified by the voter. In our context of crowdsourcing, the approval voting interface allows workers to select multiple options for every question.[1] See Figure 8.1

[1] The literature on psychology often refers to approval voting as "subset selection".



Identify the animal in this image

● Cheetah
○ Lion
○ Tiger
○ Leopard
○ Jaguar
○ Puma

Select ALL options that could be the animal in this image.

☑ Cheetah
□ Lion
□ Tiger
☑ Leopard
□ Jaguar
□ Puma

(a) (b)

Figure 8.1: Illustration of a task with (a) the standard single selection interface, and (b) an approval-voting interface.

for an example. Approval voting is known to have many advantages over single-selection systems in psychology and social choice theory [21, 49, 51, 52, 89, 103, 140]. First, an approval voting interface is easy to understand [140]. Approval voting provides workers more flexibility to express their beliefs, for instance, allowing the worker to express any confusion between multiple options, instead of being forced to select one of the options. Approval voting also utilizes the expertise of workers with partial knowledge more effectively. For instance, Coombs [51] posits that "It seems to be a common experience of individuals taking objective tests to feel confident about *eliminating some of the wrong alternatives and then guess from among the remaining ones.*" Further, Coombs argues that "Individuals taking the test should be instructed to *cross out all the alternatives which they consider wrong.*" This is precisely what we intend to do in this chapter.

Under this approval-voting interface, we require a worker to select the options that she feels are "quite likely" to be correct (we formalize this in subsequent sections). In the setting of crowdsourcing, as compared to single-selection, selecting multiple options would allow for obtaining more information about the partial knowledge of these non-expert workers. This additional information is particularly valuable for difficult labeling questions, allowing for the identification of the sources of difficulty.

Let us illustrate the utility of this setting in crowdsourcing by means of an example illustrated in Figure 8.1. The question requires the worker to identify the animal in the image — a leopard in this case. Suppose there are two workers. The first worker believes the true label to be either "Cheetah" or "Leopard", but certainly not any other option; the second worker is confused about some other aspect of the image, and believes the true label to be either "Jaguar" or "Leopard", but certainly none of the others. If each worker is allowed to select only a single answer (Figure 8.1(a)), and if the workers choose one of the two options they are confused about uniformly at random, there is a 25% chance that the first worker selects "Cheetah" and the second worker selects "Jaguar". Moreover, there is a 50% chance that one worker chooses "Leopard" and the other chooses some other option.

In each of these cases, their responses will thus not provide any definitive answer about the true label. In contrast, under the approval voting interface (Figure 8.1(b)), we allow the worker to select both the options that they are confused about, that is, ("Cheetah", "Leopard") from the first worker and ("Jaguar", "Leopard") from the other worker, then "Leopard" becomes a clear winner. Indeed, "Leopard" is the language in Figure 8.1. For our second example, we continue to consider the question in Figure 8.1. Now suppose that one worker knows the correct answer to be "Leopard" for sure, while the second worker is completely confused between "Cheetah", "Jaguar" and "Leopard." In a single-selection setting, the second worker may select one of the other three options at random, and in the case "Cheetah" is not selected, it provides an inconclusive set of answers from the two workers comprising two different options. On the other hand, in the approval voting setting, the second worker is allowed to communicate her confusion by selecting all three options under consideration, that allows for inference of "Leopard" as the correct answer.

Despite the flexibility it offers in eliciting partial knowledge, approval voting alone may not suffice for high quality crowdsourcing. A worker may have no incentive to truthfully disclose her knowledge on the crowdsourcing question. For instance, an immediate concern with the approval voting setup is that a (spammer) worker may simply choose all provided options for each question such that the correct answer is then guaranteed to lie in the set of selected options.[2] In order to incentivize the workers to report their answers truthfully, we need to couple approval voting with an appropriate payment mechanism that is a strictly proper scoring rule, such that a worker receives her maximum expected payment if and only if she truthfully discloses her partial knowledge (that is, the support of her belief) on the crowdsourcing question. Moreover, we want the mechanism to be "frugal", that is, *to pay as little as possible for poor-quality work.*

In crowdsourcing tasks comprising objective questions, it is a standard practice [83] to include "gold standard questions," that is, questions to which the system designer (or, the principal in game-theoretic terms) already knows the answers. The gold standard questions are mixed at random within the actual questions, and the worker is unaware of which questions are the gold standard. These gold standard questions are employed to verify the answers provided by the workers, and form the basis of the payments made to the workers. The gold standard questions are typically generated by experts (who are often much more expensive than crowdsourcing workers), or are obtained as an aggregate of the answers of a large number of crowdsourcing workers. In this chapter, we will not concern ourselves with the source of these gold standard questions, but only assume that we have access to a set of gold standard questions to which we know the correct answers.

The framework of scoring rules [24, 94, 137, 216] considers the design of payment mechanisms (that is, scoring rules) to elicit predictions about an event whose actual outcome will be observed in the future. The payment is a function of the agent's response and the outcome of the event. The payment is called a "strictly proper scoring rule" if its expecta-

[2] In fact, we received questions regarding precisely this action in several of our initial presentations on this work.

tion, with respect to the belief of the agent about the event, is strictly maximized when the agent reports her true belief.[3] Within this context, the first goal of the chapter, in informal terms, is to design strictly proper scoring rules that incentivize the worker to select only those options which she thinks are relatively more likely to be correct. While proper scoring rules have previously been studied under quite generic settings, this general theory provides a very broad class of scoring rules, and does not specify any particular scoring rule for use. In the application of crowdsourcing, we have a budget issue in reality. This leads to the second goal of the chapter — choose the strictly proper scoring rule which is optimal in terms of frugality.

Summary of results and organization

We consider two settings in the context of incentivizing the worker, for each question, to select options which she thinks are "quite likely" to be correct. The two settings differ in the precise meaning of the term "quite likely": The second setting is a generalization of the first, but this added generality allows for only weaker guarantees. We now describe these settings and summarize our results at a higher level.

The first setting we consider involves eliciting every option which she believes could possibly be correct. Mathematically, we formulate this problem as eliciting the support of the beliefs of workers for each question under a certain "coarse beliefs" assumption. As mentioned above, this setup coincides with the existing psychology studies on multiple-choice tests. We propose a scoring rule that we show is strictly proper for eliciting the support of the beliefs. Along the lines of our second goal, we then prove that our proposed scoring rule has several useful properties: (1) it makes the minimum payment under a spamming or low-quality work, as compared to any possible strictly proper scoring rule; (2) it is the only strictly proper scoring rule that can satisfy a simple and intuitive condition which we term "no-free-lunch"; (3) it is robust in the sense that, even when the coarse beliefs assumption is violated, the scoring rule does not break down, but continues to incentivize workers to act in a certain desirable manner.

In the second part of the chapter, we consider a general setting that is associated with a given parameter σ, and for each question, involves eliciting the set of options whose likelihood of correctness is more than σ according to the worker's belief. This setting is more general than the first setting since: (1) the first setting turns out to be a special case corresponding to $\sigma = 0$; and (2) we do not make the coarse beliefs assumption. We design a scoring rule for this setting and show that it is strictly proper. In addition, we also prove that under certain restricted conditions, our proposed rule is the only possible strictly proper scoring rule.

The rest of the chapter is organized as follows. We begin with a description of the formal problem setting and the goals of the chapter in Section 8.2. We present our main theoretical

[3] In this chapter, we use the term "strictly proper scoring rule" instead of using "(strictly) incentive compatible" as in Chapter 6 and Chapter 7 in order to remain consistent with our paper associated to the results of this chapter. In the context of our work, the two terms mean the same and can be used interchangably.

results in Section 8.3. Here, in Section 8.3.1 we present theoretical results on the problem of eliciting the support of the worker, and in Section 8.3.2 we present theory on the problem of eliciting options whose belief-probabilities exceed a given threshold. We conclude the main text of the chapter with a discussion on empirical evaluations and future work in Section 8.4.

8.2 Problem setting

Consider $N \geq 1$ questions, each of which has B options ($2 \leq B < \infty$) to choose from. For each option, exactly one of the B options is correct. We assume that these N questions contain G ($1 \leq G \leq N$) "gold standard" questions, that is, questions to which the mechanism designer knows the answers a priori. These gold standard questions are assumed to be mixed uniformly at random among the N questions, and the worker is evaluated based on her performance on these G questions. For every individual question, we assume that the worker has, in her mind, a distribution over the B options representing her beliefs of the probabilities of the respective options being correct. We assume that these belief-distributions of a worker are independent across questions [89].

8.2.1 Payment function (scoring rule)

As mentioned earlier, the worker's performance is evaluated based on her responses to the gold standard questions. For any question in the gold standard, we denote the evaluation of the worker's performance on this question by a value in the set $\{-(B-1), \ldots, B\}$: the magnitude of this value represents the number of options she had selected, and the sign is positive if the correct answer was in the set of selected options and negative otherwise. For instance, if the worker selected four options for a certain gold standard question but none of them was correct, then the evaluation of this response is denoted as "-4"; if the worker selects two options for a gold standard question and one of them turns out to be the correct option then the evaluation of this response is denoted as "$+2$".

We will assume that the payments are bounded, that is, any payment must lie in the interval $[\alpha_{\min}, \alpha_{\max}]$, for some values $\alpha_{\min}$ and $\alpha_{\max} > \alpha_{\min}$. The choice of the two parameters $\alpha_{\min}$ and $\alpha_{\max}$ may be made keeping various factors in mind, such as guidelines of the crowdsourcing platform used, the budget constraints, and the minimum wage. We will assume that the values of the two parameters are given to us.

Let

$$f : \{-(B-1), \ldots, B\}^G \rightarrow [\alpha_{\min}, \alpha_{\max}]$$

denote the payment function. We will use the terms "payment function" and "scoring rule" interchangeably throughout the chapter. It is this function f which must be designed in order to incentivize the worker. In order to bring all possible scoring rules on an equal footing, we fix $\alpha_{\max}$ as the payment for the best possible outcome, which is when the worker selects exactly the correct option (and nothing else) for each question in the gold standard:

$$f(1,\ldots,1)=\alpha_{\max}. \tag{8.1}$$

Throughout the chapter, we assume all scoring rules to satisfy (8.1), along with the requirement to lie in the interval $[\alpha_{\min},\alpha_{\max}]$.

In the sequel, we use the notation f^* and $f^\#$ to denote the two scoring rules proposed in this chapter, and f to denote any general scoring rule. We will say that the worker has "not attempted" a certain question if the worker selects all the B options for that question.

8.2.2 Expected payment

A quantity central to our analysis is the *expected payment*, where the expectation is from the point of view of the worker, and is taken over the randomness in the choice of the G gold standard questions among the N questions, and over the N probability distributions representing her beliefs for the N questions. Let us formalize this notion. Suppose that for question $i\in[N]$, the worker has selected some $y_i\in[B]$ of the B options. Further, let $s_i\in[0,1]$ denote the probability, under the worker's beliefs, that the correct answer to question i lies in this set of y_i selected options. In other words, s_i denotes the sum of the beliefs for the y_i options selected by the worker (consequently, the sum of the beliefs for the options not selected is $(1-s_i)$). Then from the worker's point of view, her expected payment for this selection is

$$\frac{1}{\binom{N}{G}}\sum_{(j_1,\ldots,j_G)\subseteq[N]}\ \sum_{(\epsilon_1,\ldots,\epsilon_G)\in\{-1,1\}^G}\left(f(\epsilon_1 y_{j_1},\ldots,\epsilon_G y_{j_G})\prod_{i=1}^{G}(1-s_{j_i})^{\mathbf{1}\{\epsilon_i=-1\}}s_{j_i}^{\mathbf{1}\{\epsilon_i=1\}}\right). \tag{8.2}$$

The outer summation in (8.2) corresponds to the expectation with respect to the random distribution of the G gold standard questions in the N total questions. The inner summation in (8.2) corresponds to the expectation with respect to the worker's beliefs of her choices being correct. In more detail, the values $\epsilon_1,\ldots,\epsilon_G$ iterate over all possible events regarding whether or not the correct answer lies in the selected set of options for each of the G gold standard questions, and the term $\prod_{i=1}^{G}(1-s_{j_i})^{\mathbf{1}\{\epsilon_i=-1\}}s_{j_i}^{\mathbf{1}\{\epsilon_i=1\}}$ represents the probability (according to the worker's beliefs) of the occurrence of each such event. For instance, if the worker selects all options ($y_i=B$) for every question $i\in[N]$, then the correct answer must necessarily lie in the set of selected options (that is, $s_i=1$ for every $i\in[G]$), and then the payment evaluates to exactly $f(B,\ldots,B)$.

Given the presence of gold standard questions, the performance of any worker can be verified based on only her own answers (without depending on the answers of other workers). The payments made to different workers thus do not depend on each other, and hence we consider only one worker without loss of generality.

In this chapter, we assume that the worker aims to maximize her expected reward, where the expectation is taken over the randomness in the choice of the G standard questions, and in terms of the beliefs of the worker regarding the correctness of various options for each question.

8.2.3 Goal

At a higher level, our goal is to incentivize the worker, for each question, to select all options that she believes are quite likely to be correct. The chapter is split into two parts, depending on the specifics of this goal, as detailed below.

First setting: Eliciting support of beliefs

The first part of the chapter (Section 8.3.1) addresses the goal of eliciting, for every question, the *support* of the worker's distribution over the B options. In other words, we wish to incentivize the worker such that for each question, the worker should select the *smallest subset of the set of options* such that the correct answer according to her belief lies in the selected subset. Such a requirement is motivated by studies such as [51] discussed earlier, and also due to its virtue of being quite simple to describe to the workers.

Formally, suppose that for any question $i \in [N]$, the worker believes that the probability of option $b \in [B]$ being correct is p_{ib}, for some non-negative values $p_{i1}, \ldots, p_{iB}$ that sum to one. Then the goal is to incentivize the worker, for each question $i \in [N]$, to select precisely the set of options

$$\{b \in [B] \mid p_{ib} \neq 0\}. \tag{8.3}$$

The worker is incentivized to do so by means of a payment mechanism that forms a strictly proper scoring rule.

Definition 8 (Strictly proper scoring rule for support elicitation)**.** *A payment function is a strictly proper scoring rule for the problem of eliciting the support of beliefs if the expected payment* (8.2) *from the worker's point of view is strictly maximized when she selects all options* (8.3) *for which her belief is non-zero.*

As we show in Appendix 8.A.1, there is no strictly proper scoring rule for this goal in the absence of any additional assumptions. To this end, we make a certain assumption of "coarse beliefs" that is motivated by various findings in psychology that are similar in spirit. There is an extensive literature in psychology establishing the coarseness of processing and perception in humans. For instance, Miller's celebrated paper [169] establishes the information and storage capacity of humans, that an average human being can typically distinguish at most about seven states. This granualrity of human computation is verified in many subsequent experiments [212, 238]. The paper [114] establishes the ineffectiveness of finer-granularity response elicitation. Mullainathan et al. [174] hypothesize that humans often group things into categories; this hypothesis is experimentally verified in the paper [240] in a specific setting. In the same spirit of coarseness of human processing, we make the following assumption.

Consider some (fixed and known) value $\rho > 0$, and assume that the probability of any option for any question, according to the worker's belief, is either zero or greater than

ρ.[4] Since one must necessarily take into account situations when a worker is totally clueless about a question, that is, when her belief is distributed uniformly over all options, we restrict $\rho < \frac{1}{B}$. To summarize, we make the following "coarse belief" assumption.

Definition 9 (Coarse belief assumption)**.** *The worker's belief for any option for any question lies in the set* $\{0\} \cup (\rho, 1]$ *for some (fixed and known) value* $\rho \in \left(0, \frac{1}{B}\right)$*.*

Our first set of results (Section 8.3.1) address the goal of designing strictly proper scoring rules for support elicitation under the coarse beliefs assumption. In general, there may be many such strictly proper scoring rules possible, and hence importantly, we also show that our designed mechanism is strictly "optimal" in the sense of a certain notion of frugality and "unique" in that it is the only mechanism that satisfies a simple and intuitive additional requirement. We present this mechanism, along with the precise definitions and statements of uniqueness and optimality in Section 8.3.1.

Second setting: Eliciting thresholded beliefs

Our second set of results (Section 8.3.2) consider a more general setting that also does not make the coarse beliefs assumption. We assume to be given the value of a parameter $\sigma \in (0, 1)$. For any question $i \in [N]$, let p_{ib} denote the probability of option $b \in [B]$ being correct according to the worker's belief. Here, $p_{i1}, \ldots, p_{iB}$ are non-negative values that sum to one. Then for each question $i \in [N]$, we want the worker to select precisely the set of options

$$\{b \in [B] \mid p_{ib} > \sigma\}, \tag{8.4a}$$

while not selecting the options

$$\{b \in [B] \mid p_{ib} < \sigma\}. \tag{8.4b}$$

Then the goal is to design payment mechanisms that are strictly proper scoring rules that incentivize this behavior.

Definition 10 (Strictly proper scoring rule to elicit thresholded beliefs)**.** *A payment function is a strictly proper scoring rule if the expected payment* (8.2) *from the worker's point of view is strictly maximized when she selects all options* (8.4a) *for which her belief is more than* σ *and does not select all options* (8.4b) *for which her belief is lower than* σ*.*

We note that the worker is allowed to act either way for options for which her belief is exactly σ. We do not impose a requirement from the scoring rule when the worker's belief equals σ for any option since this is a boundary case that is impossible to incentivize; see Appendix 8.A.2 for a proof of this claim.

[4] The impossibility shown in Appendix 8.A.1, discussed earlier in Section 8.2.3, pertains to the case $\rho = 0$.

In Section 8.3.2, we provide a payment mechanism that forms a strictly proper scoring rule for eliciting thresholded beliefs, along with additional results on its "uniqueness". We remark that the absence of additional assumptions (such as that of coarse beliefs), the results on uniqueness/optimality are weaker than that established towards elicitation of the support in Section 8.3.1.

8.3 Main results

We present our main results in this section. In Section 8.3.1, we consider the requirement of eliciting the full support of the beliefs of the worker. Under a "coarse beliefs" assumption regarding on the beliefs of people, we design a strictly proper scoring rule and show strong guarantees associated to it. Subsequently in Section 8.3.2, we forgo this coarse beliefs assumption and design strictly proper scoring rules that elicit options with beliefs above a chosen (strictly positive) threshold.

8.3.1 Eliciting support under a coarse beliefs assumption

In this section, we address the goal of eliciting the full support of the workers' beliefs, assuming a coarseness of belief that assigns a value of zero to very low probability categories, as detailed in Section 8.2.3. Since the support of the belief distribution must necessarily contain at least one item, we will mandate the worker to select at least one option for each question. Consequently, the scoring rules considered in this section are functions mapping the set $\{-(B-1),\ldots,-1,1,\ldots,B\}$ to the interval $[\alpha_{\min},\alpha_{\max}]$.

In what follows, we first present our proposed strictly proper scoring rule for this problem, and subsequently derive certain motivating properties for our scoring rule.

Proposed strictly proper scoring rule

We begin by presenting our proposed scoring rule, denoted by f^*, as Scoring rule 3. Here, the function $\mathbf{1} : \{True, False\} \rightarrow \{0,1\}$ is the indicator function, defined as $\mathbf{1}\{x\} = 1$ if x is true, and $\mathbf{1}\{x\} = 0$ otherwise.

The payment is based only on the evaluation of the worker's responses to the gold standard questions. It is easy to describe the mechanism in words: The payment is $\alpha_{\min}$ plus

(i) 0 if the correct answer is not selected for one or more questions, otherwise

(ii) κ^* reduced by $(100\rho)\%$ for each option selected.

Part (i) of the payment rule is a result of the indicator function $\mathbf{1}\{x_i \geq 1\}$ in the description of Scoring rule 3. Part (ii) arises due to the term $(1-\rho)^{|x_i|}$. The term κ^* is only used to ensure that the $(\alpha_{\max},\alpha_{\min})$-conditions are satisfied.

Observe that our proposed scoring rule takes a very interesting "multiplicative" structure.

Scoring rule 3 Strictly proper scoring rule to elicit support of beliefs in approval voting

- **Input:** Evaluations of the worker's answers to the G gold standard questions $(x_1, \ldots, x_G)$
- **Output:** The worker's payment

$$f^*(x_1, \ldots, x_G) = \kappa^* \prod_{i=1}^{G} \Big((1-\rho)^{|x_i|} \ \mathbf{1}\{x_i \geq 1\} \Big) + \alpha_{\min},$$

where $\kappa^* = \frac{\alpha_{\max} - \alpha_{\min}}{(1-\rho)^G}$

Theorem 20. *Under the coarse-beliefs assumption, Scoring rule 3 is strictly proper.*

Our multiplicative scoring rule is thus theoretically guaranteed to work.

Frugality

As discussed earlier, popular crowdsourcing platforms suffer from the problem of low-quality data, often due to the presence of spammers and the like. To this end, it is imperative to have scoring rules that make a low payment for such work, while retaining a high-enough payment for good workers. Such a scoring rule will ensure that, first, there is minimal expenditure on such spamming behavior, and second, that the low payment will deincentivize spammers from taking up the task. With this motivation, we study such frugality properties of our proposed scoring rules, and contrast them with all other strictly proper scoring rules.[5]

Consider the problem of eliciting the support of the worker's beliefs for each question under the coarse beliefs assumption. Recall our proposed payment function f^* described in Scoring rule 3. Also, recall the condition $f(1, \ldots, 1) = \alpha_{\max}$ and $f(x) \in [\alpha_{\min}, \alpha_{\max}]$ for every x imposed on any scoring rule considered here. Then we have the following guarantees on our scoring rule f^*, as compared to any other strictly proper scoring rule.

Theorem 21. *For any $N \geq G$, among all strictly proper scoring rules, our scoring rule f^* spends the strictly minimum possible amount to a worker who does not attempt any question, that is,*

$$f(B, \ldots, B) > f^*(B, \ldots, B),$$

for every strictly proper scoring rule f.

[5] In this chapter, the term "frugality" is used only in a colloquial sense, although the connotation here is similar in spirit to other formal notions of frugality [6, 249] that address the amount of over-payment by a mechanism.

Theorem 21 addresses the most concerning spamming behavior — that of skipping all questions — and shows that our proposed strictly proper scoring rule pays a *strictly* smaller amount under such a behavior as compared to any other strictly proper scoring rule. In accordance with our earlier discussion, this result is important in practice since a system using our strictly proper scoring rule will waste he minimum amount of money on such spamming behavior, and moreover, the lower payment may serve to deincentivize spammers from taking up tasks that employ our strictly proper scoring rule.

The proof of Theorem 21 first shows that any strictly proper scoring rule f must satisfy $f(B,\ldots,B) \geq f^*(B,\ldots,B)$. The proof then goes on to show that any strictly proper scoring rule f with $f(B,\ldots,B) = f^*(B,\ldots,B)$ must necessarily satisfy $f(x) = f^*(x)$ for every argument x, that is, must be identical to Scoring rule 3. This claim is proved via (thee levels of) induction on x.

The following theorem now proves more properties of our scoring rule.

Theorem 22. *(a) When $N = G$, for any $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$, we have*

$$f(x_1,\ldots,x_G) \geq f^*(x_1,\ldots,x_G).$$

(b) For any $N \geq G$, consider any value $\gamma \in [G]$. Consider any strictly proper scoring rule f such that

$$f(x_1,\ldots,x_G) = f^*(x_1,\ldots,x_G)$$

for every $(x_1,\ldots,x_G) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$ that satisfies $\sum_{i=1}^G \mathbf{1}\{x_i = 1\} \geq \gamma$. Then it must be that

$$f(x_1{'},\ldots,x_G{'}) \geq f^*(x_1{'},\ldots,x_G{'})$$

for every $(x_1{'},\ldots,x_G{'}) \in \{-(B-1),\ldots,-1,1,\ldots,B\}^G$ that satisfies $\sum_{i=1}^G \mathbf{1}\{x_i{'} = 1\} \geq \gamma - 1$.

Let us parse the two parts of the theorem. Part (a) of Theorem 22 demonstrates that our scoring rule is pointwise the most frugal among all strictly proper scoring rules under the setting $N = G$. While the setting of $N = G$ is not directly useful for our crowdsourcing setting in practice, it is important since it forms the basis for all the theoretical guarantees. Part (b) then goes on to show that any other strictly proper scoring rule that pays the same amount as Scoring rule 3 for a certain quality of work, must pay at least as much as our scoring rule for a worse quality. The notion of quality in the statement of this part is determined by the number of questions to which the worker chose only the correct answer $\sum_{i=1}^G \mathbf{1}\{x_i = 1\}$.

Each of the three parts follow from one or more applications application of the following lemma.

Lemma 38. *Consider some $y, y' \in [B]^N$ and some $\mathcal{I} \subseteq [N]$ such that $y_i = y'_i + 1$ for all $i \in \mathcal{I}$, and $y_i = y'_i$ for all $i \notin \mathcal{I}$. Then any strictly proper scoring rule f must necessarily satisfy*

$$\frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) \geq \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}). \tag{8.5}$$

Furthermore, a necessary condition for the above inequality to be satisfied with equality is

$$f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}) = \alpha_{\min} \tag{8.6}$$

for all $(j_1,\ldots,j_G) \subseteq [N]$, and all $\{(\epsilon_1,\ldots,\epsilon_G) \in \{-1,1\}^G \backslash \{1\}^G \mid \epsilon_i = 1$ whenever $j_i \notin \mathcal{I}\}$.

The lemma derives lower bounds on the payment made by any strictly proper scoring rule under any evaluation y as compared to another evaluation y' that differs from y only in the questions in some set $\mathcal{I}$. In particular, the left hand side of (8.5) is simply the expected payment under a strictly proper scoring rule f under an evaluation y. The right hand side is a rescaling of the expected payment under the evaluation y', where the payment is rescaled by a factor $(1-\rho)$ for every additional option selected in y' as compared to y.

The second part (8.6) of the lemma then shows that any strictly proper scoring rule that achieves (8.5) with equality must make a minimum payment whenever any of the questions in the set $\mathcal{I}$ does not have the correct answer selected.

Robustness to the coarse beliefs assumption: Incentives with finer beliefs

In the earlier subsections, we made the "coarse belief" assumption that the worker's belief for any option, when non-zero, is at least ρ. We then designed a strictly proper scoring rule, Scoring rule 3, with respect to eliciting the supports of the beliefs of the worker. A natural question then arises is: How does this scoring rule perform if the coarse beliefs assumption is violated? We address this issue in the present section, showing that our scoring rule does not break down, but rather continues to incentivize workers to act in a certain desirable way. In more detail, if Scoring rule 3 (for a certain value of ρ) is encountered by a worker who may have arbitrary beliefs, the scoring rule incentivizes the worker to select all options for which the relative belief of the worker is high enough.

Theorem 23. *Under Scoring rule 3, for any question, a worker with beliefs $1 \geq p_1 \geq \ldots \geq p_B \geq 0$ for the B options is incentivized to select options $\{1,\ldots,m\}$ for that question, where*

$$m = \arg\max_{z\in[B]} \left(\frac{p_z}{\sum_{i=1}^{z} p_i} > \rho \right).$$

It is not hard to interpret this incentivized action. The worker selects options one by one in decreasing order of her beliefs as long as the selected option contributes a fraction more than ρ to the total belief of the selected options.

In order to understand the result of Theorem 23 slightly better, let us perform a sanity check and verify that the earlier result of Theorem 20 for "coarse beliefs" is indeed a special case of Theorem 23. To this end, suppose the beliefs of the worker for any particular question are $p_1 \geq \cdots p_m > \rho > p_{m+1} = \cdots = p_B = 0$ for some $m \in [B]$. Then we have

$$\frac{p_z}{\sum_{i=1}^{z} p_i} = \frac{0}{\sum_{i=1}^{z} p_i} = 0 < \rho \qquad \text{for all } z \geq m+1,$$

and

$$\frac{p_z}{\sum_{i=1}^{z} p_i} \geq \frac{p_z}{1} > \rho \qquad \text{for all } z \leq m.$$

It follows that under the result of Theorem 23, a worker with "coarse beliefs" will be incentivized to select precisely the support of her beliefs.

The proof of Theorem 23 proof first computes the expected payment under the response described in Theorem 23, and then by means of some careful algebraic arguments, shows that any other answer will lead to a strictly lower payment.

An axiomatic derivation

We conclude this section with an alternative axiomatic derivation of our mechanism when accommodating workers with arbitrary beliefs. The derivation involves a "no-free-lunch axiom" similar to that considered in Chapter 6 and Chapter 7, which when adapted to our approval-voting based setting is defined as follows. Recall from Section 8.2 that we say that a worker has "not attempted" a certain question if the worker selects all the B options for that question; otherwise we say that the worker attempted that question. We will also say that the response of a worker to a question is "wrong" if the correct option does not lie in the set of options that the worker selected for that question.

Definition 11 (No-free-lunch axiom). *If the response to every attempted question in the gold standard turns out to be wrong, then the worker gets the minimum payment, namely,*

$$f(x_1, \ldots, x_G) = \alpha_{\min} \qquad \forall \;\; (x_1, \ldots, x_G) \in \{-(B-1), \ldots, -1, B\}^G \backslash \{B\}^G.$$

First, observe that the no-free-lunch axiom is a very mild condition. For instance, even if a worker selects $\frac{B}{2}$ options uniformly at random for each question, there is only a $\frac{1}{2^G}$ chance that the no-free-lunch axiom will come into play. Second, the axiom is quite intuitive: the only condition where the axiom applies is when for every question, the worker either does not attempt or provides the incorrect answer, in which case the worker is providing no useful information.

The no-free-lunch axiom is quantitatively different from the notions of frugality we studied earlier in Section 8.3.1. However, both these notions have the same qualitative goal, namely to minimize the expenditure when no useful data is obtained, while providing higher payments to workers providing better data. Interestingly, as we show below, both these notions lead to the same (unique) strictly proper scoring rule under our setting of approval voting.

Theorem 24. *Consider no assumptions on the minimum value of the belief, and suppose the workers must be incentivized to select options* $\{1, \ldots, m\}$ *where* $m = \arg\max_z \left(\frac{p_z}{\sum_{i=1}^{z} p_i} > \rho \right)$. *Then, Scoring rule 3 is the one and only strictly proper scoring rule that satisfies the no-free-lunch axiom.*

The proof of Theorem 24 relies on the following lemma, which provides another necessary condition (in addition to Lemma 38) that must necessarily be satisfied by any strictly proper scoring rule (that may or may not satisfy no-free-lunch).

Lemma 39. *Any strictly proper scoring rule* f *must satisfy*

$$\begin{aligned} &f(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_G) \\ &\qquad = (1-\rho) f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) + \rho f(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_G), \end{aligned}$$

for every $i \in [G]$ *and* $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in \{-(B-1), \ldots, -1, 1, \ldots, B\}^{G-1}$, $x_i \in [B-1]$.

Note that the lemma does *not* use the no-free-lunch condition. In words, the lemma considers any arbitrary evaluations of the answers to the gold standard questions $\{1, \ldots, i-1, i+1, G\}$. Fixing the evaluations of these questions, it says that the payment under any strictly proper scoring rule for the evaluation $(x_{i+1} + 1)$ of question i must be a convex combination of the payments under the evaluations x_i and $-x_i$ for question i. Moreover, the coefficient in that convex combination must precisely equal the parameter ρ.

The proof of this lemma is provided in Appendix 8.5.6. A repeated application of this lemma, along with the no-free-lunch axiom, leads to the result of Theorem 24.

8.3.2 Eliciting options with beliefs above a threshold

In this section, we consider the setting of incentivizing the worker to select all options for which her belief is strictly greater σ, for some fixed parameter $\sigma \in (0, 1)$, as detailed in Section 8.2.3. We do *not* assume the restriction of coarseness of the beliefs.

Before proceeding ahead, we must specify certain pedantic details of the problem setting. Let us define two integers $s_{\min}$ and $s_{\max}$ as $s_{\min} = \mathbf{1}\{\sigma < \frac{1}{B}\}$ and $s_{\max} = \min\{\lceil \frac{1}{\sigma} \rceil - 1, B\}$. Consider any question. Observe that if if $\sigma < \frac{1}{B}$ then it is meaningless to let the worker select zero options since the belief for at least one option must be $\frac{1}{B}$ or higher. Also observe that for any value of $\sigma \in (0, 1)$, it is meaningless to allow the worker to select $\lceil \frac{1}{\sigma} \rceil$ or more options, since it is mathematically impossible for those many options to have probabilities more than σ. As a result, we will mandate the worker to select at least $s_{\min}$ and at most $s_{\max}$

options for any question. The goal remains to design the payment function $f(x_1, \ldots, x_G)$ when $|x_i| \in \{s_{\min}, \ldots, s_{\max}\}$ for every $i \in [G]$.

Finally, we note some special cases which we exclude from the subsequent analysis. The case of $\sigma = 0$ degenerates to the impossibility result of eliciting the support in the absence of the course beliefs assumption (Appendix 8.A.1) discussed earlier in Section 8.2.3. If $B = 2$ or if $\sigma \geq \frac{1}{2}$, the setting degenerates to the "skip-based" single-selection setting studied in Chapter 6. Hence we focus on the case of $B \geq 3$ and $\sigma \in (0, \frac{1}{2})$ in the rest of this section.

Proposed scoring rule

Our proposed scoring rule for the setting of this section is provided as Scoring rule 4. For convenience of notation, we denote this scoring rule as $f^{\#}$.

Scoring rule 4 Incentive mechanism for the alternative problem formulation

- **Input:** Evaluations of the worker's answers to the G gold standard questions $(x_1, \ldots, x_G)$
- **Output:** The worker's payment

$$f^{\#}(x_1, \ldots, x_G) = \kappa^{\#} \prod_{i=1}^{G} \Big((B - |x_i| - 1)\sigma + \mathbf{1}\{x_i \geq 1\} \Big) + \alpha_{\min},$$

where $\kappa^{\#} = \frac{\alpha_{\max} - \alpha_{\min}}{((B - s_{\min} - 1)\sigma + \mathbf{1}\{s_{\min} \geq 1\})^G}$

For any question $i \in [G]$, the component $\big((B - |x_i| - 1)\sigma + \mathbf{1}\{x_i \geq 1\}\big)$ of the scoring rule $f^{\#}$ penalizes the selection of an incorrect option by σ and rewards the selection of the correct option by 1. The overall payment is then a product of these components over all gold standard questions. The constant $\kappa^{\#}$ simply serves to scale the payment in order to accommodate the $(\alpha_{\min}, \alpha_{\max})$-requirements.

The following theorem now proves guarantees associated to our scoring rule.

Theorem 25. *Consider any $\sigma \in (0, \frac{1}{2})$, $N \geq G \geq 1$ and $B \geq 3$. Then Scoring rule 4 is a strictly proper scoring rule.*

The proof of this result first computes the expected payment in case of honest responses, and then via some algebraic arguments shows that every other response must lead to a strictly smaller payment.

Uniqueness

In this section, we address our second goal of choosing a strictly proper scoring rule among many possible options. In particular, we show that the core structure of Scoring rule 4 must necessarily be a part of any strictly proper scoring rule.

Theorem 26. *Consider any $\sigma \in (0, \frac{1}{2})$ and any $B \geq 3$. When $G = 1$, our proposed scoring rule, Scoring rule 4, is the one and only possible strictly proper scoring rule upto a constant shift and positive scaling.*

The proof of this result shows that any strictly proper scoring rule f must necessarily satisfy the following four sets of equations (when $G = 1$):

$$\begin{aligned} f(m+1) &= (1-\sigma)f(m) + \sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 1\}, \\ f(m+2) &= (1-2\sigma)f(m) + 2\sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 2\}, \\ f(-s_{\max}) &= f(s_{\max}) - f(s_{\max} - 1) + f(-(s_{\max} - 1)), \qquad \text{and} \\ f(0) &= \sigma f(1) + (1-\sigma)f(-1). \end{aligned}$$

These four sets of conditions together leave only two degrees of freedom for the choice of the payment function f, and hence uniquely characterize the scoring rule upto a constant shift and scale.

While we do not have a complete answer as to what the “best” or “unique” mechanism is for general values of N and G, but going by the results in Section 8.3.1, we conjecture that Scoring rule 4 may also possess some more attractive properties along the lines of Scoring rule 3.

8.4 Discussion

Our goal is to deliver high quality labels for machine learning applications, at low costs, by means of incentive mechanisms or aggregation algorithms or both. In this chapter, we pursue the former approach. We take an approval-voting based means of gathering labeled data from crowdsourcing. We design an incentive mechanism via a principled theoretical approach, and prove appealing properties of optimality and uniqueness of our proposed mechanism. Preliminary experiments conducted on Amazon Mechanical Turk corroborate the usefulness of this mechanism for practical scenarios. Our mechanism may also draw more experts to the crowdsourcing platform since their compensation will be significantly higher than that of mediocre workers, unlike most compensation mechanisms in current use.

Experiments performed on Amazon Mechanical Turk in [237] reveal that (a) workers to make judicious use of the flexibility offered by the approval voting interface, (b) the presence of a strictly proper scoring rule does make a statistically significant difference as compared to a scoring rule that is not strictly proper, and (c) the workers did not have any real objections to the approval voting interface or the multiplicative mechanism.

We conclude this chapter with a discussion on closely related topics that merit investigation in the future.

Aggregation of labels. For the traditional single-selection setting, there is a long line of work on statistical methods to aggregate redundant noisy data from multiple workers; see Chapter 4 for more details. An open problem is the design of aggregation algorithms for approval-voting-based data: algorithms that can exploit the specific structure of the responses that arise as a result of the approval voting interface and the proposed mechanism. There is indeed work on aggregation algorithms [22, 37, 160, 189] and probabilistic models [66, 74, 159, 207] for approval-voting in the context of social choice theory; their objective, however, is primarily of fairness and strategy-proofing of the voting procedure, as opposed to our goal of denoising data obtained from multiple heterogeneous workers as required for labeling tasks in crowdsourcing.

Choosing the right interface. There are tradeoffs between various interfaces for crowdsourcing. For instance, the approval voting interface elicits the support of the belief whereas the single selection interface elicits the mode. Choosing among these two interfaces would depend on the application under consideration, and moreover, one may adaptively switch between the two depending on the data obtained. A natural question that one may further ask is, why not elicit the entire belief distribution itself? While the entire belief distribution seems to supercede the support and the mode, stating the distribution will also require much more time and effort from the workers, and often also suffer from a higher noise. These tradeoffs must be taken into account when choosing the interface for the application at hand.

The coarse beliefs parameter. One may wish to evaluate the value of ρ by explicitly asking workers on the crowdsourcing platform for this value. However, it is noted in the literature (e.g., see the paper [220] for experiments on Amazon Mechanical Turk) that the cardinal representations that humans provide are not always consistent with their respective mental beliefs, and are far noisier. This phenomenon suggests the requirement of developing alternative methods of evaluating this parameter. Indeed, measurement is considered one of the most difficult parts of behavioral research.

8.5 Proofs

In this section, we present proofs of the various theoretical results presented in the chapter.

8.5.1 Proof of Lemma 38: The workhorse lemma

Consider any arbitrary strictly proper scoring rule f satisfying the $(\alpha_{\min}, \alpha_{\max})$ conditions. Without loss of generality let $\alpha_{\min} = 0$.

Consider a real number $\rho_0 \in (\rho, \frac{1}{B})$, whose precise value will be specified later. Consider a worker such that for every question $i \in \mathcal{I}$, her belief is ρ_0 for the first option and $\frac{1-\rho_0}{y_i-1}$ for each of the last $(y_i - 1)$ options. For every question $i \notin \mathcal{I}$, her belief is uniformly distributed among the first y_i options. Note that this satisfies the coarse beliefs assumption since

$$\frac{1-\rho_0}{y_i-1} \overset{(i)}{\geq} \frac{1-\rho_0}{B-1} \overset{(ii)}{>} \frac{1-\frac{1}{B}}{B-1} \geq \frac{1}{B} \overset{(iii)}{>} \rho,$$

where the inequality (i) follows from the fact that $y_i \leq B$, inequality (ii) follows from the assumption $\rho_0 < \frac{1}{B}$, and inequality (iii) is an assumption on ρ (see Definition 9).

If this worker selects precisely the support of her beliefs for every question then her expected payment Ψ_1 is

$$\Psi_1 = \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}). \tag{8.7}$$

We will compare the aforementioned action to another action, where for each question $i \in \mathcal{I}$, the worker selects only the last $y'_i = (y_i - 1)$ options but not the first option; for each question $i \notin \mathcal{I}$, the worker selects the support of her belief. Under this action, the expected payment Ψ_2 equals

$$\Psi_2 = \frac{1}{\binom{N}{G}} \sum_{\substack{(j_1,\ldots,j_G)\\ \subseteq[N]}} \sum_{\substack{(\epsilon_1,\ldots,\epsilon_G)\\ \in\{-1,1\}^G}} \mathbf{1}\{\{j_i \mid \epsilon_i = -1\} \subseteq \mathcal{I}\}(1-\rho_0)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho_0^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G}). \tag{8.8}$$

In the expression (8.8), the outer summation represents the expectation over the random choice of the G gold standard questions among the N questions. The inner summation represents the expectation with respect to the correctness or incorrectness of the answers to the G gold standard questions: for any question i, $\epsilon_i = 1$ captures the event where the i^{th} question in the gold standard is answered correctly and $\epsilon_i = -1$ represents the event of this question being answered incorrectly. The term $\mathbf{1}\{\{j_i \mid \epsilon_i = -1\} \subseteq \mathcal{I}\}$ ensures that only the questions in $\mathcal{I}$ can be wrong, since it is only these questions for which the worker has selected a subset of her belief's support.

Since $f(x) \geq 0$ for every valid argument x, we can lower bound Ψ_2 as

$$\Psi_2 \geq \frac{1}{\binom{N}{G}} \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}). \tag{8.9}$$

A strictly proper scoring rule must incentivize the worker to perform the first action (over the second), i.e, must have $\Psi_1 > \Psi_2$. Thus from (8.7) and (8.9), we get

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) > \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}). \tag{8.10}$$

Note that (8.10) must hold for all $\rho_0 > \rho$. The left hand side of (8.10) does not involve ρ_0 whereas the right hand side is continuous in ρ_0. It follows that

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1},\ldots,y_{j_G}) \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1},\ldots,y'_{j_G}). \tag{8.11}$$

This proves the first part (8.5) of the lemma.

We now move on to prove the second part (8.6) of the lemma. We prove the claimed result by means of a contradiction argument. Suppose that $f(\epsilon_1 y'_{j_1},\ldots,\epsilon_G y'_{j_G})$ is strictly positive for some $(j_1, \ldots, j_G) \subseteq [N]$, $\{(\epsilon_1, \ldots, \epsilon_G) \in \{-1,1\}^G \backslash \{1\}^G \mid \epsilon_i = 1$ whenever $j_i \notin \mathcal{I}\}$. Then using the fact that $f(x) \geq 0$ for all x, from (8.8), we obtain the inequality

$$\Psi_2 \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho_0)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}) + (1-\rho_0)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho_0^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1}, \ldots, \epsilon_G y'_{j_G}).$$

Note that this inequality is the equivalent of (8.9) in the first part of the lemma, but also accounts for the additional strictly positive term. Following arguments identical to those in the first part, we have the following tighter version of (8.11):

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1}, \ldots, y_{j_G}) \geq \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}) + (1-\rho)^{|\mathcal{I}\cap\{j_i|\epsilon_i=1\}|}\rho^{|\mathcal{I}\cap\{j_i|\epsilon_i=-1\}|} f(\epsilon_1 y'_{j_1}, \ldots, \epsilon_G y'_{j_G}).$$

Since $f(\epsilon_1 y'_{j_1}, \ldots, \epsilon_G y'_{j_G}) > 0$, we then have

$$\sum_{(j_1,\ldots,j_G)\subseteq[N]} f(y_{j_1}, \ldots, y_{j_G}) > \sum_{(j_1,\ldots,j_G)\subseteq[N]} (1-\rho)^{|\mathcal{I}\cap\{j_1,\ldots,j_G\}|} f(y'_{j_1}, \ldots, y'_{j_G}),$$

thereby contradicting the hypothesis of the equality in (8.5) assumed in the second part of the lemma, and hence proving the claimed result.

8.5.2 Proof of Theorem 20: Working of our scoring rule

Without loss of generality, we may assume that $\alpha_{\min} = 0$ since in our setting, the property of being strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. We also remind the reader that the "expected payment" always refers to the expectation with respect to the worker's beliefs regarding the correctness of various options, and the randomness in the choice of the G gold standard questions in the N questions.

First consider the case $N = G = 1$. In this case, Scoring rule 3 reduces to

$$f^*(x) = \alpha(1-\rho)^{(x_1-1)}\mathbf{1}\{x_1 \geq 1\}.$$

Suppose without loss of generality that the worker's beliefs for the B options are $p_1 \geq \cdots \geq p_m > \rho > p_{m+1} = \cdots = p_B = 0$ for some $m \in [B]$. A strictly proper scoring rule must strictly maximize the worker's expected payment when she selects the support of her belief, that is, when she selects exactly the options $\{1, \ldots, m\}$. The expected payment, Ψ_0, under this selection is

$$\begin{aligned}\Psi_0 &= \alpha \sum_{i=1}^{m} p_i (1-\rho)^{m-1} \\ &= \alpha(1-\rho)^{m-1}.\end{aligned}$$

Instead, now suppose the worker selects some other set of options $\{o_1, \ldots, o_\ell\} \subseteq [B]$, $\{o_1, \ldots, o_\ell\} \neq [m]$. Then her expected payment, Ψ_1, under the proposed mechanism for this selection is

$$\begin{aligned}\Psi_1 &= \alpha \sum_{i=1}^{\ell} p_{o_i} (1-\rho)^{\ell-1} \\ &\leq \alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}, \end{aligned} \tag{8.12}$$

where the final inequality is a result of the ordering $p_1 \geq \cdots \geq p_B$. If $\ell = m$ then the inequality in (8.12) is strict since $p_j < p_i$ for all $(j > m,\, i \leq m)$. Thus the expected payment under the choice $\ell = m$ but with a selection different from the support is strictly smaller than Ψ_0. Also observe that the expected payment when $\ell > m$ is upper bounded by $\alpha(1-\rho)^{\ell-1}$, which is strictly smaller than Ψ_0. Let us now consider the remaining case of $\ell < m$. Since $p_i > \rho$ for all $i \in [m]$, we have

$$\begin{aligned}\Psi_1 &< \alpha \left(\sum_{i=1}^{m} p_i - (m-\ell)\rho \right) (1-\rho)^{\ell-1} \\ &= \alpha \left(1 - (m-\ell)\rho\right) (1-\rho)^{\ell-1} \\ &\overset{(i)}{\leq} \alpha (1-\rho)^{m-\ell} (1-\rho)^{\ell-1} \\ &= \Psi_0, \end{aligned}$$

where inequality (i) follows from the algebraic fact that $(1-aw) \leq (1-w)^a$ for every $a \geq 1$ and every $w \in [0,1]$. This completes the proof for the case $N = G = 1$.

Let us now consider the case of $N = G \geq 1$. By our assumption of the independence of the beliefs of the worker across the questions, the expected payment equals

$$\prod_{i=1}^{G} \mathbf{E} \left[\alpha (1-\rho)^{(x_i - 1)} \mathbf{1}\{x_i \geq 1\} \right],$$

where the expectation pertains to the randomness in the signs of $x_1, \ldots, x_G$. Since the payments are non-negative, if each individual component in the product is maximized then the product is also necessarily maximized. Each individual component simply corresponds to the setting of $N = G = 1$ discussed earlier. Thus calling upon our earlier result, we get that the expected payment for the case $N = G > 1$ is strictly maximized when the worker acts as desired for every question.

Let us finally consider the case of $N > G \geq 1$. Recall from (8.2) that the expected payment for the general case is a cascade of two expectations: the outer expectation is with respect to the uniformly random distribution of the G gold standard questions among the N total questions, while the inner expectation is taken over the worker's beliefs of the different questions conditioned on the choice of the gold standard questions. The arguments above for the case $N = G$ prove that every individual term in the inner expectation is maximized when the worker acts as desired. The expected payment is thus maximized when the worker acts as desired.

8.5.3 Proof of Theorem 21: Minimum pay if all questions skipped

Without loss of generality, assume that $\alpha_{\min} = 0$ since in our setting, the property of incentive compatibility is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. Consider any strictly proper scoring rule f such that $f(1, \ldots, 1) = \alpha$. The proof uses Lemma 38, which was stated earlier at the end of Section 8.3.1 and is proved in Section 8.5.1.

First, a non-strict inequality: Consider any $x_0 \in [B-1]$. Applying Lemma 38 with $y = (x_0 + 1, \ldots, x_0 + 1)$, $y' = (x_0, \ldots, x_0)$ and $\mathcal{I} = [G]$ gives

$$f(x_0 + 1, \ldots, x_0 + 1) \geq (1-\rho)^G f(x_0, \ldots, x_0).$$

A repeated application of this inequality for every $x_0 \in [B-1]$ gives

$$\begin{aligned} f(B, \ldots, B) &\geq (1-\rho)^G f(B-1, \ldots, B-1) \geq \cdots \geq (1-\rho)^{(B-1)G} f(1, \ldots, 1) \\ &= (1-\rho)^{(B-1)G} \alpha. \end{aligned}$$

Scoring rule 3 achieves this lower bound on $f(B, \ldots, B)$ with equality, thereby completing the proof.

Strict inequality: In the remainder of the proof, we show that any strictly proper scoring rule that achieves

$$f(B, \ldots, B) = (1-\rho)^{G(B-1)} \alpha$$

must be identical to our Scoring rule 3. In other words, we show that such a scoring rule f must satisfy $f(x) = f^*(x)$ for every evaluation x. We partition the rest of the proof into two cases, depending on whether $x_G > 0$ or $x_G < 0$.

In what follows, we only consider the set of evaluations x whose elements are non-decreasing, that is, $x_1 \geq x_2 \geq \cdots \geq x_G$. The proof for any other ordering of the elements follows in an identical manner.

Case of $x_G > 0$: We first consider any x such that $x_G > 0$. Then due to the monotonicity of the entries of x, we must have that $x_1 \geq \cdots \geq x_G > 0$. We define the following notation that we will subsequently use for our induction arguments.

- Let $\gamma(x)$ denote the number of distinct entries in x:

$$\gamma(x) := 1 + \sum_{i=1}^{G-1} \mathbf{1}\{x_i \neq x_{i+1}\}$$

- Let $\sigma(x)$ denote the size of the last jump in x:

$$\sigma(x) := x_j - x_{j+1} \qquad \text{where } j = \arg\max_{i\in[G-1]} x_i \neq x_{i+1}$$

- Let $\beta(x)$ denote the numeric value of x in a B-ary number system:

$$\beta(x) := \sum_{i=1}^{G} B^{G-i}(x_i - 1).$$

For example, if $B = 5$, $G = 5$ and $x = (5,5,4,1,1)$ then $\gamma(x) = |\{5,4,1\}| = 3$, $\sigma(x) = 4 - 1 = 3$ (where $j = 3$), and $\beta(x) = 4\cdot 5^4 + 4\cdot 5^3 + 3\cdot 5^2 + 0\cdot 5^1 + 0\cdot 5^0 = 3075$. The proof involves three nested levels of induction: on γ, on σ and then on β.

Base case for induction on γ: We first induct on γ. The base case is the set $\{x|\gamma(x) = 1\}$, that is, the set of vectors which have the same value for all its components. We now show that $f(x) = f^*(x)$ for every x such that $\gamma(x) = 1$. To this end, observe that from the definition of γ, the only values of x that have $\gamma(x) = 1$ are those that have all elements identical. Consider any $x_0 \in [B-1]$. Applying Lemma 38 with $y = (x_0+1,\ldots,x_0+1)$ and $y' = (x_0,\ldots,x_0)$ gives

$$f(x_0+1,\ldots,x_0+1) \geq (1-\rho)^G f(x_0,\ldots,x_0).$$

Since this inequality is true for every $x_0 \in [B-1]$, we have the sandwich inequalities

$$f(B,\ldots,B) \geq (1-\rho)^{(B-x_0)G} f(x_0,\ldots,x_0) \geq (1-\rho)^{(B-1)G} f(1,\ldots,1).$$

Setting $f(B,\ldots,B) = (1-\rho)^{(B-1)G}\alpha$ and $f(1,\ldots,1) = \alpha$ implies that each of the above inequalities is in fact an equality, thereby proving the base case

$$f(x_0,\ldots,x_0) = (1-\rho)^{x_0 G} f(1,\ldots,1) = f^*(x_0,\ldots,x_0).$$

Induction step for induction on γ: Now suppose our hypothesis of $f(x) = f^*(x)$ is true for all $\{x|\gamma(x) \leq \gamma_0 - 1\}$ for some $\gamma_0 \in \{2,\ldots,B\}$. We will now prove that the hypothesis $f(x) = f^*(x)$ is also true for all $\{x|\gamma(x) = \gamma_0\}$. Towards this goal, we now induct on σ, that is, we prove the hypothesis separately for every value of σ.

Base case for induction on σ: Observe the following set-relation $\{x|\gamma(x) = \gamma_0 - 1\} = \{x|\gamma(x) = \gamma_0, \sigma = 0\}$. Due to the assumed induction hypothesis $f(x) = f^*(x)$ for every $\{x|\gamma(x) = \gamma_0 - 1\}$, we have $f(x) = f^*(x)$ for every $\{x|\gamma(x) = \gamma_0, \sigma = 0\}$, thereby proving the base case of $\sigma = 0$.

Induction step for induction on σ: Now suppose that the hypothesis is true for all $\{x|\gamma(x) = \gamma_0,\ \sigma(x) \leq \sigma_0 - 1\}$ for some $\sigma_0 \in [B-1]$. We will prove that the hypothesis remains true for all $\{x|\gamma(x) = \gamma_0,\ \sigma(x) = \sigma_0\}$. We prove this statement is true for all values of β via an induction on β.

Base case for induction on β: Recall that we have restricted our attention to those x which have their elements in a descending order. Observe that the element with the minimum value of β in the set $\{x|\gamma(x) = \gamma_0,\ \sigma(x) = \sigma_0\}$ is $(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 1, \ldots, 1)$. We will prove the hypothesis for this element as the base case for our induction on β. Applying Lemma 38 with $y = (\gamma_0+\sigma_0-1, \ldots, \sigma_0+2, \sigma_0+1, 1, \ldots, 1)$ and $y' = (\gamma_0+\sigma_0-1, \ldots, \sigma_0+2, \sigma_0, 1, \ldots, 1)$ gives the inequality

$$\begin{aligned}
&c_1 f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1) + c_1' f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, 1, 1, \ldots, 1)\\
&\qquad + \sum_{s \subsetneq \{\gamma_0+\sigma_0-1,\ldots,\sigma_0+2\}} (c_s f(s, 1, 1, \ldots, 1) + c_s' f(s, \sigma_0 + 1, 1, \ldots, 1))\\
&\geq c_1(1-\rho) f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) + c_1' f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, 1, 1, \ldots, 1)\\
&\qquad + \sum_{s \subsetneq \{\gamma_0+\sigma_0-1,\ldots,\sigma_0+2\}} (c_s f(s, 1, 1, \ldots, 1) + c_s'(1-\rho) f(s, \sigma_0, 1, \ldots, 1)), \qquad (8.13)
\end{aligned}$$

for some positive constants c_1, c_1', c_s, c_s' (which represent the probabilities of the respective set of G questions being chosen as the G gold standard questions). Now, for any $s \subsetneq \{\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2\}$, observe that $\gamma(s, \sigma_0 + 1, 1, \ldots, 1) \leq \gamma_0 - 1$ and $\gamma(s, \sigma_0, 1, \ldots, 1) \leq \sigma_0 - 1$. Thus we have

$$\begin{aligned}
f(s, \sigma_0 + 1, 1, \ldots, 1) &\overset{(i)}{=} f^*(s, \sigma_0 + 1, 1, \ldots, 1)\\
&= (1-\rho) f^*(s, \sigma_0, 1, \ldots, 1)\\
&\overset{(ii)}{=} (1-\rho) f(s, \sigma_0, 1, \ldots, 1), \qquad (8.14)
\end{aligned}$$

where equations (i) and (ii) follow from our induction hypothesis. Also, $\gamma(\gamma_0+\sigma_0-1, \ldots, \sigma_0+2, \sigma_0, 1, \ldots, 1) = \gamma_0$ and $\sigma(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) = \sigma_0 - 1$. Consequently, our induction hypothesis yields

$$\begin{aligned}
f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1) &= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0, 1, \ldots, 1)\\
&= (1-\rho)^{\gamma_0+\sigma_0-2+\cdots+\sigma_0+1+\sigma_0-1}\alpha. \qquad (8.15)
\end{aligned}$$

Substituting (8.14) and (8.15) in (8.13) and canceling out common terms yields the inequality

$$\begin{aligned}
f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1) &\geq (1-\rho)^{\gamma_0+\sigma_0-2+\cdots+\sigma_0}\alpha\\
&= f^*(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1). \qquad (8.16)
\end{aligned}$$

We now derive a matching upper bound on $f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 2, \sigma_0 + 1, 1, \ldots, 1)$. Applying Lemma 38 with $y = (\gamma_0+\sigma_0-1, \ldots, \sigma_0+1, 2, \ldots, 2)$ and $y' = (\gamma_0+\sigma_0-1, \ldots, \sigma_0+1, 1, \ldots, 1)$ gives

$$\begin{aligned}
&c_1 f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 2, \ldots, 2) + \sum_{s \subsetneq \{\gamma_0+\sigma_0-1,\ldots,\sigma_0+1\}} c_s f(s, 2, \ldots, 2)\\
&\geq c_1(1-\rho)^{G-\gamma+1} f(\gamma_0 + \sigma_0 - 1, \ldots, \sigma_0 + 1, 1, \ldots, 1) + \sum_{s \subsetneq \{\gamma_0+\sigma_0-1,\ldots,\sigma_0+1\}} c_s(1-\rho)^{G-|s|} f(s, 1, \ldots, 1), \qquad (8.17)
\end{aligned}$$

for some positive constants c_1, c_s. Now, for any $s \subsetneq \{\gamma_0+\sigma_0-1,\ldots,\sigma_0+2\}$, observe that $\gamma(s,2,\ldots,2) \leq \gamma_0 - 1$ and $\gamma(s,1,\ldots,1) \leq \sigma_0 - 1$. Thus we have

$$f(s,2,\ldots,2) \overset{(i)}{=} f^*(s,2,\ldots,2) = (1-\rho)^{G-|s|} f(s,1,\ldots,1) \overset{(ii)}{=} (1-\rho)^{G-|s|} f(s,1,\ldots,1), \tag{8.18}$$

where equations (i) and (ii) follow from our induction hypothesis. Also note that $\gamma(\gamma_0+\sigma_0-1,\ldots,\sigma_0+1,2,\ldots,2) \leq \gamma_0$ and $\sigma(\gamma_0+\sigma_0-1,\ldots,\sigma_0+1,2,\ldots,2) = \sigma_0-1$, which allows us to apply our induction hypothesis to get

$$\begin{aligned} f(\gamma_0+\sigma_0-1,\ldots,\sigma_0+1,2,\ldots,2) &= f^*(\gamma_0+\sigma_0-1,\ldots,\sigma_0+1,2,\ldots,2) \\ &= (1-\rho)^{\gamma_0+\sigma_0-2+\ldots+\sigma_0+G-\gamma+1}\alpha. \end{aligned} \tag{8.19}$$

Substituting (8.22) and (8.19) in (8.17) and canceling out common terms yields the upper bound

$$\begin{aligned} f(\gamma_0+\sigma_0-1,\ldots,\sigma_0+2,\sigma_0+1,1,\ldots,1) &\leq (1-\rho)^{\gamma_0+\sigma_0-2+\cdots+\sigma_0}\alpha \\ &= f^*(\gamma_0+\sigma_0-1,\ldots,\sigma_0+2,\sigma_0+1,1,\ldots,1) \end{aligned} \tag{8.20}$$

The bounds (8.16) and (8.20) in conjunction imply that the hypothesis is true for $x = (\gamma_0+\sigma_0-1,\ldots,\sigma_0+2,\sigma_0+1,1,\ldots,1)$, which is the base case for our induction on β.
Induction step for induction on β: Now consider some x^* such that $\gamma(x^*) = \gamma_0$, $\sigma(x^*) = \sigma_0$ and $\beta(x^*) = \beta_0$, for some β_0. One can verify that any such x^* must necessarily take the form

$$x^* = ({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*,\ldots,\sigma_0+{x_G}^*}_{m_1},{x_G}^*,\ldots,{x_G}^*), \tag{8.21}$$

with ${x_1}^* \geq {x_2}^* \geq \cdots \geq {x_m}^* > \sigma_0 + {x_G}^*$ for some $m \geq 0$, $m_1 \geq 1$, $m+m_1 < G$.

Suppose the hypothesis $f(x) = f^*(x)$ is true for every $\{x|\gamma(x)=\gamma_0,\ \sigma(x)=\sigma_0,\ \beta(x) \leq \beta_0 - 1\}$. In what follows, we show that we must have $f(x) = f^*(x)$ for every $\{x|\gamma(x) = \gamma_0,\ \sigma(x)=\sigma_0,\ \beta(x)=\beta_0\}$.

Applying Lemma 38 to f with the choices

$$\begin{aligned} y &= ({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*,\ldots,\sigma_0+{x_G}^*}_{m_1},{x_G}^*,\ldots,{x_G}^*), \quad \text{and} \\ y' &= ({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*) \end{aligned}$$

gives the inequality

$$
\begin{aligned}
&c_1 f({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*,\ldots,\sigma_0+{x_G}^*}_{m_1},{x_G}^*,\ldots,{x_G}^*)\\
&\qquad+\sum_{s\subsetneq\{{x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*,\ldots,\sigma_0+{x_G}^*}_{m_1}\}} c_s f(s,{x_G}^*,\ldots,{x_G}^*)\\
&\qquad\geq c_1(1-\rho)^{m_1} f({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*)\\
&\qquad+\sum_{s\subsetneq\{{x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1}\}} c_s(1-\rho)^{\sum_i \mathbf{1}\{s_i=\sigma_0+{x_G}^*-1\}} f(s,{x_G}^*,\ldots,{x_G}^*), \qquad (8.22)
\end{aligned}
$$

for some positive constants c_1, c_s. Recall that x^* takes the form (8.21) and has $\gamma(x^*)=\gamma_0$, $\sigma(x^*)=\sigma_0$ and $\beta(x^*)=\beta_0$. Thus we have

$$
\gamma({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*)=\begin{cases}\gamma_0-1 & \text{if } \sigma_0=1\\ \gamma_0 & \text{otherwise,}\end{cases}
$$

and hence the induction hypothesis is satisfied in the first case of $\sigma_0=1$. In the second case of $\sigma_0\neq 1$, we have

$$
\sigma({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*)=\sigma_0-1,
$$

and hence the induction hypothesis is satisfied in the second case as well. Thus we have

$$
\begin{aligned}
&f({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*)\\
&\qquad\qquad=f^*({x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1},{x_G}^*,\ldots,{x_G}^*)\\
&\qquad\qquad=(1-\rho)^{\sum_{i=1}^m({x_i}^*-1)+m_1(\sigma_0+{x_G}^*-2)+(G-m_1-m)({x_G}^*-1)}\alpha. \qquad (8.23)
\end{aligned}
$$

Consider any $s\subsetneq\{{x_1}^*,\ldots,{x_m}^*,\underbrace{\sigma_0+{x_G}^*-1,\ldots,\sigma_0+{x_G}^*-1}_{m_1}\}$. We claim that the induction hypothesis is satisfied for $(s,{x_G}^*,\ldots,{x_G}^*)$. To this end, define the quantity $\mathfrak{m}_1(s)$ as

$$
\mathfrak{m}_1(s):=\sum_i \mathbf{1}\{s_i=\sigma_0+{x_G}^*-1\}.
$$

Observe that if $\mathfrak{m}_1(s)>0$ then it must either be that $\gamma(s,{x_G}^*,\ldots,{x_G}^*)\leq\gamma_0-1$ or it must be that $\sigma(s,{x_G}^*,\ldots,{x_G}^*)\leq\sigma_0-1$; if $\mathfrak{m}_1(s)=0$ then $\gamma(s,{x_G}^*,\ldots,{x_G}^*)\leq\gamma_0-1$. Now for any value $s \subsetneq \{x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}\}$, also define the quantity $\widetilde{\mathfrak{m}}_1(s)$ as

$$\widetilde{\mathfrak{m}}_1(s) := \sum_i \mathbf{1}\{s_i = \sigma_0 + x_G{}^*\}.$$

Observe that if $\widetilde{\mathfrak{m}}_1(s) > 0$ then either $\gamma(s, x_G{}^*, \dots, x_G{}^*) \leq \gamma_0 - 1$ or $\beta(s, x_G{}^*, \dots, x_G{}^*) \leq \beta_0 - 1$; if $\widetilde{\mathfrak{m}}_1(s) = 0$ then $\gamma(s, x_G{}^*, \dots, x_G{}^*) \leq \gamma_0 - 1$. Consequently from our induction hypothesis we have the series of equations

$$\begin{aligned}
&\sum_{s \subsetneq \{x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}\}} c_s f(s, x_G{}^*, \dots, x_G{}^*) \\
&= \sum_{s \subsetneq \{x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}\}} c_s f^*(s, x_G{}^*, \dots, x_G{}^*) \\
&= \sum_{s \subsetneq \{x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^* - 1, \dots, \sigma_0 + x_G{}^* - 1}_{m_1}\}} c_s (1-\rho)^{\sum_i \mathbf{1}\{s_i = \sigma_0 + x_G{}^* - 1\}} f^*(s, x_G{}^*, \dots, x_G{}^*) \\
&= \sum_{s \subsetneq \{x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^* - 1, \dots, \sigma_0 + x_G{}^* - 1}_{m_1}\}} c_s (1-\rho)^{\sum_i \mathbf{1}\{s_i = \sigma_0 + x_G{}^* - 1\}} f(s, x_G{}^*, \dots, x_G{}^*).
\end{aligned} \tag{8.24}$$

Substituting (8.23) and (8.24) in (8.22) and canceling out common terms gives

$$\begin{aligned}
f(x_1{}^*, \dots, x_m{}^*, &\underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^*, \dots, x_G{}^*) \\
&\geq (1-\rho)^{m_1} f(x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^* - 1, \dots, \sigma_0 + x_G{}^* - 1}_{m_1}, x_G{}^*, \dots, x_G{}^*) \\
&= (1-\rho)^{\sum_{i=1}^m (x_i{}^* - 1) + m_1(\sigma_0 + x_G{}^* - 1) + (G - m_1 - m)(x_G{}^* - 1)} \alpha.
\end{aligned} \tag{8.25}$$

We now employ Lemma 38 to derive a matching upper bound to (8.25). Setting

$$\begin{aligned}
y &= (x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \dots, x_G{}^* + 1), \mathit{and} \\
y' &= (x_1{}^*, \dots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \dots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^*, \dots, x_G{}^*)
\end{aligned}$$

in Lemma 38 yields the inequality

$$\begin{aligned}
&c_1 f(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \ldots, x_G{}^* + 1) \\
&\qquad + \sum_{s \subsetneq \{x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}\}} c_s f(s, x_G{}^* + 1, \ldots, x_G{}^* + 1) \\
&\geq c_1 (1-\rho)^{m_1} f(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^*, \ldots, x_G{}^*) \\
&\qquad + \sum_{s \subsetneq \{x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}\}} c_s (1-\rho)^{G-|s|} f(s, x_G{}^*, \ldots, x_G{}^*), \qquad (8.26)
\end{aligned}$$

for some positive constants c_1, c_s. Observe that

$$\gamma(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \ldots, x_G{}^* + 1) = \begin{cases} \gamma_0 - 1 & \text{if } \sigma_0 = 1 \\ \gamma_0 & \text{otherwise,} \end{cases}$$

and that the induction hypothesis is satisfied in the first case of $\sigma = 1$. In the second case of $\sigma \neq 1$, we have

$$\sigma(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \ldots, x_G{}^* + 1) = \sigma_0 - 1,$$

and hence the induction hypothesis is satisfied in this case as well. Thus from our induction hypothesis, we have

$$\begin{aligned}
&f(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \ldots, x_G{}^* + 1) \\
&\qquad = f^*(x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}, x_G{}^* + 1, \ldots, x_G{}^* + 1) \\
&\qquad = (1-\rho)^{\sum_{i=1}^m (x_i{}^* - 1) + m_1(\sigma_0 + x_G{}^* - 1) + (G - m_1 - m)(x_G{}^* - 2)} \alpha. \qquad (8.27)
\end{aligned}$$

Now consider any $s \subsetneq \{x_1{}^*, \ldots, x_m{}^*, \underbrace{\sigma_0 + x_G{}^*, \ldots, \sigma_0 + x_G{}^*}_{m_1}\}$. We now show that our induction hypothesis is satisfied for $(s, x_G{}^*, \ldots, x_G{}^*)$ as well as $(s, x_G{}^* + 1, \ldots, x_G{}^* + 1)$. To this end, recall our notation of $\widetilde{\mathfrak{m}}_1(s) := \sum_i \mathbf{1}\{s_i = \sigma_0 + x_G{}^*\}$. If $\sigma_0 = 1$ or if $\widetilde{\mathfrak{m}}_1(s) = 0$ then $\gamma(s, x_G{}^* + 1, \ldots, x_G{}^* + 1) \leq \gamma_0 - 1$; if $\sigma > 1$ and $\widetilde{\mathfrak{m}}_1(s) > 0$ then $\gamma(s, x_G{}^* + 1, \ldots, x_G{}^* + 1) \leq \gamma_0$ and $\sigma(s, x_G{}^* + 1, \ldots, x_G{}^* + 1) \leq \sigma_0 - 1$. If $\widetilde{\mathfrak{m}}_1(s) = 0$ then $\gamma(s, x_G{}^*, \ldots, x_G{}^*) \leq \gamma_0 - 1$, otherwise $\gamma(s, x_G{}^*, \ldots, x_G{}^*) \leq \gamma_0$, $\sigma(s, x_G{}^*, \ldots, x_G{}^*) = \sigma_0$ and $\beta(s, x_G{}^*, \ldots, x_G{}^*) \leq \beta_0 - 1$.

These terms thus satisfy our induction hypothesis and hence we have

$$\begin{aligned} f(s, {x_G}^* + 1, \ldots, {x_G}^* + 1) &= f^*(s, {x_G}^* + 1, \ldots, {x_G}^* + 1) \\ &= (1-\rho)^{G-|s|} f^*(s, {x_G}^*, \ldots, {x_G}^*) \\ &= (1-\rho)^{G-|s|} f(s, {x_G}^*, \ldots, {x_G}^*). \end{aligned} \tag{8.28}$$

Substituting (8.27) and (8.28) in (8.26) gives us an upper bound

$$\begin{aligned} f({x_1}^*, \ldots, {x_m}^*, \underbrace{\sigma_0 + {x_G}^*, \ldots, \sigma_0 + {x_G}^*}_{m_1}, {x_G}^*, \ldots, {x_G}^*) \\ &\leq (1-\rho)^{\sum_{i=1}^{m}({x_i}^*-1) + m_1(\sigma_0 + {x_G}^* - 1) + (G - m_1 - m)({x_G}^* - 1)} \alpha \\ &= f^*({x_1}^*, \ldots, {x_m}^*, \underbrace{\sigma_0 + {x_G}^*, \ldots, \sigma_0 + {x_G}^*}_{m_1}, {x_G}^*, \ldots, {x_G}^*). \end{aligned} \tag{8.29}$$

The matching bounds (8.25) and (8.29) together complete the proof of induction on β. Moving back up in our nesting of inductions, this also completes the proof for $\{x | x_i \geq 0 \ \forall \ i \in [G]\}$.
Case of $x_G < 0$: We now address the remaining case of $\{x \mid \min_{i \in [G]} x_i < 0\}$, and show that $f(x) = f^*(x) = 0$ for all such x. The arguments above for the case $\{x \mid \min_{i \in [G]} x_i > 0\}$ imply that for any strictly proper scoring rule f, the inequality (8.5) in the statement of Lemma 38 must be satisfied with an equality. This allows us to employ the second part of Lemma 38, and we do so in the following manner. For every $i \in [G]$, let $y_i = y'_i = x_i$ if $x_i > 0$, and $y_i - 1 = y'_i = |x_i|$ otherwise; set $y_i = y'_i = B$ for all $i \in \{G+1, \ldots, N\}$. Then (8.6) in the statement of Lemma 38 yields $f(x_1, \ldots, x_G) = 0$, thus completing the proof.

8.5.4 Proof of Theorem 22: Minimal expenditure under f^*

In this section, we prove that our Scoring rule 4 makes a minimal payment as compared to any other strictly proper scoring rule. The proof uses Lemma 38, which was stated earlier in Section 8.3.1 and is proved in Section 8.5.1.

Proof of part (a)

We assume without loss of generality that $\alpha_{\min} = 0$.

First consider the case of any x such that $x_i < 0$ for some $i \in [G]$. For this case we have

$$f(x) \geq 0 = f^*(x),$$

thereby proving our claim.

Now consider the remaining case where $x_i > 0$ for every $i \in [G]$. For the setting $N = G$ under consideration, the inequality (8.5) simplifies to

$$f(y_1, \ldots, y_G) \geq (1-\rho)^{|\mathcal{I}|} f(y'_1, \ldots, y'_G),$$

for any set $\mathcal{I} \in [G]$, and any $y, y' \in [B]^G$ such that $y_i = y'_i + 1$ for every $i \in \mathcal{I}$ and $y_i = y'_i$ otherwise. A repeated application of this inequality yields the bound

$$f(y_1, \ldots, y_G) \geq (1-\rho)^{\sum_{i=1}^G (y_i - 1)} f(1, \ldots, 1) = (1-\rho)^{\sum_{i=1}^G (y_i - 1)} \alpha_{\max} = f^*(y_1, \ldots, y_G),$$

for any $y \in [B]^G$, thereby proving the claimed result.

Proof of part (b)

We assume without loss of generality that $\alpha_{\min} = 0$. Also assume that $N > G$, since the case of $N = G$ follows directly from Theorem 21(a).

First consider the case of any x such that $x_i < 0$ for some $i \in [G]$. For this case we have

$$f(x) \geq 0 = f^*(x),$$

thereby proving our claim.

Now consider the remaining case where $x_i > 0$ for every $i \in [G]$. Define a function $\nu : [B]^G \to \{0, \ldots, G\}$ that measures the number of entries equaling 1 in the input, that is,

$$\nu(x) = \sum_{i=1}^G \mathbf{1}\{x_i = 1\},$$

for every $x \in [B]^G$.

First suppose that $\nu(x) = G$. In this case, we must have $x = (1, \ldots, 1)$. Consequently, we have $f(x) = \alpha_{\max} = f^*(x)$.

Now for some value $\nu_0 \in [G]$, suppose that $f(x) = f^*(x)$ for every $\{x \mid \nu(x) \geq \nu_0\}$, as in the statement of the theorem. In what follows, we show that any x satisfying $\nu(x) = \nu_0 - 1$ must also satisfy $f(x') > f^*(x')$.

Consider any $x \in [B]^G$ such that $\nu(x) = \nu_0 - 1$. We assume, without loss of generality, that $x_1 \geq \cdots \geq x_G$ (≥ 0). Then we have $x_{G-\nu_0+2} = \cdots = x_G = 1$ and $x_{G-\nu_0+1} > 1$. Define a vector $y = (x, 1, \ldots, 1) = (x_1, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1)$. We now apply Equation (8.5) from Lemma 38 with $y' = (x_1 - 1, x_2, \ldots, x_{G-\nu_0+1}, 1, \ldots, 1)$ and $\mathcal{I} = \{1\}$ to obtain the bound

$$\begin{aligned}
c_1 &\sum_{(j_2,\ldots,j_G)\subseteq\{2,\ldots,N\}} f(x_1, y_{j_2}, \ldots, y_{j_G}) + c_2 \sum_{(j_1,\ldots,j_G)\subseteq\{2,\ldots,N\}} f(y_{j_1}, \ldots, y_{j_G}) \\
&\geq c_1 \sum_{(j_2,\ldots,j_G)\subseteq\{2,\ldots,N\}} (1-\rho) f(x_1 - 1, y_{j_2}, \ldots, y_{j_G}) + c_2 \sum_{(j_1,\ldots,j_G)\subseteq\{2,\ldots,N\}} f(y_{j_1}, \ldots, y_{j_G}),
\end{aligned}$$

for some constants $c_1 > 0$ and $c_2 > 0$ whose values depend only on N and G. Canceling out common terms, we are left with

$$\sum_{(j_2,\ldots,j_G)\subseteq\{2,\ldots,N\}} f(x_1, y_{j_2}, \ldots, y_{j_G}) \geq \sum_{(j_2,\ldots,j_G)\subseteq\{2,\ldots,N\}} (1-\rho) f(x_1 - 1, y_{j_2}, \ldots, y_{j_G}). \qquad (8.30)$$

Both the left and right hand sides of this inequality involve linear combinations of the function f evaluated at various points. For our chosen values of y and y', observe that whenever $\{2,\ldots,G-\nu_0+1\} \not\subseteq \{j_2,\ldots,j_G\}$, we must have $\nu(x_1,y_{j_2},\ldots,y_{j_G}) \geq \nu_0$ and $\nu(x_1-1,y_{j_2},\ldots,y_{j_G}) \geq \nu_0$. Consequently, for any such value of $(j_2,\ldots,j_G)$, we have

$$\begin{aligned} f(x_1,&y_{j_2},\ldots,y_{j_G}) = \\ &f^*(x_1,y_{j_2},\ldots,y_{j_G}) = (1-\rho)f^*(x_1-1,y_{j_2},\ldots,y_{j_G}) = (1-\rho)f(x_1-1,y_{j_2},\ldots,y_{j_G}). \end{aligned}$$

For any remaining value of $(j_2,\ldots,j_G)$ (such that $\{2,\ldots,G-\nu_0+1\} \subseteq \{j_2,\ldots,j_G\}$), we have $(y_{j_2},\ldots,y_{j_G}) = (x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1)$. Substituting these relations in (8.30) and canceling out common terms leaves us with the bound

$$f(x_1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1) \geq (1-\rho)f(x_1-1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1).$$

A repeated application of this inequality for different values of x_1 then yields

$$\begin{aligned} f(x_1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1) &\geq (1-\rho)^{x_1-1}f(1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1) \\ &\overset{(i)}{=} (1-\rho)^{x_1-1}f^*(1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1) \\ &= f^*(x_1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1), \end{aligned}$$

as claimed, where equation (i) is a because $\nu(1,x_2,\ldots,x_{G-\nu_0+1},1,\ldots,1) \geq \nu_0$.

8.5.5 Proof of Theorem 23: Performance in absence of coarse belief assumption

Without loss of generality, assume that $\alpha_{\min} = 0$ since in our setting, the property of being strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$.

First consider the case of $N = G = 1$. For this case, Scoring rule 3 reduces to $f^*(x) = \alpha(1-\rho)^{(x_1-1)}\mathbf{1}\{x_1 \geq 0\}$. Suppose without loss of generality that the worker's beliefs for the B options are $p_1 \geq \cdots \geq p_B$. Define integer m as $m = \arg\max_{z\in[B]}\left(\frac{p_{(z)}}{\sum_{i=1}^{z} p_{(i)}} > \rho\right)$.

Suppose a worker decides to select some ℓ of the B options, say options $\{o_1,\ldots,o_\ell\} \subseteq [B]$. Then it is easy to see that her expected payment,

$$\alpha\sum_{i=1}^{\ell} p_{o_i}(1-\rho)^{\ell-1},$$

is maximized when she selects options $\{1,\ldots,\ell\}$, i.e., the ℓ options that are most likely to be correct. Under the monotonicity $p_1 \geq \cdots \geq p_B$, it is easy to see that for any fixed choice of $\ell \in [B]$, the expected payment is maximized when the worker selects options $\{1,\ldots,\ell\}$.

It remains to show that among all choices of $\ell \in [B]$, the expected payment is maximized when the worker selects $\ell = m$. Let Ψ_ℓ denote the expected payment when the worker selects the first ℓ options:

$$\Psi_\ell = \alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}.$$

Hence for any $\ell \in \{2, \ldots, B\}$, we have

$$\frac{\Psi_{\ell-1}}{\Psi_\ell} = \frac{\alpha \sum_{i=1}^{\ell-1} p_i (1-\rho)^{\ell-2}}{\alpha \sum_{i=1}^{\ell} p_i (1-\rho)^{\ell-1}} = \frac{1}{1-\rho}\left(1 - \frac{p_\ell}{\sum_{i=1}^{\ell} p_i}\right).$$

We know that $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} < \rho$ whenever $\ell > m$, and $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} > \rho$ when $\ell = m$. Furthermore, since p_ℓ decreases with ℓ and $\sum_{i=1}^{\ell} p_i$ increases with ℓ, it must also be that $\frac{p_\ell}{\sum_{i=1}^{\ell} p_i} > \rho$ for all $\ell < m$. Thus we have $\frac{\Psi_\ell}{\Psi_{\ell-1}} > 1$ for all $\ell \leq m$ and $\frac{\Psi_\ell}{\Psi_{\ell-1}} < 1$ for all $\ell > m$, or in other words, we have

$$\cdots < \Psi_{m-2} < \Psi_{m-1} < \Psi_m > \Psi_{m+1} > \Psi_{m+2} > \cdots.$$

It follows that the worker is incentivized to choose $\ell = m$.

Let us now consider the case of $N = G \geq 1$. By our assumption of the independence of the beliefs of the worker across the questions, the expected payment equals

$$\prod_{i=1}^{G} \mathbf{E}\left[\alpha(1-\rho)^{(x_i-1)}\mathbf{1}\{x_i \geq 0\}\right].$$

Since the payments are non-negative, if each individual component in the product is maximized then the product is also necessarily maximized. Each individual component simply corresponds to the setting of $N = G = 1$ discussed earlier. Thus calling upon our earlier result, we get that the expected payment for the case $N = G \geq 1$ is maximized when the worker acts as desired for every question.

Let us finally consider the general case of $N \geq G \geq 1$. Recall from (8.2) that the expected payment for the general case is a cascade of two expectations: the outer expectation is with respect to the uniformly random distribution of the G gold standard questions among the N total questions, while the inner expectation is taken over the worker's beliefs of the different questions conditioned on the choice of the gold standard questions and restricts attention to only these G questions. The arguments above for the case $N = G$ prove that every individual term in the inner expectation is maximized when the worker acts as desired. The outer expectation does not affect this argument. The expected payment is thus maximized when the worker acts as desired.

8.5.6 Proof of Lemma 39: A necessary condition for any strictly proper scoring rule

First consider the case of $G = N$. Consider some $\eta, \gamma \in \{0, \ldots, G-1\}$ with $\eta + \gamma < G$. Suppose $i = \eta+\gamma+1$, $x_1, \ldots, x_\eta \in [B-1]$, $x_{\eta+1}, \ldots, x_{\eta+\gamma} \in -[B-1]$ and $x_{\eta+\gamma+2}, \ldots, x_N = B$.

For every question $j \in [\eta + \gamma]$, suppose the worker's belief is $\delta_j \in (0, \rho)$ for the last option and $\frac{1-\delta_j}{|x_j|}$ each for the first $|x_j|$ options. One can verify that since $\delta_j < \rho < \frac{1}{B}$ and $|x_j| \leq B - 1$, it must be that $\frac{1-\delta_j}{|x_j|} > \delta_j$, and that the requirement of being a strictly proper scoring rule requires incentivizing the worker to select the first $|x_j|$ options. Suppose the worker does so. Now for every question $j' \in \{\eta + \gamma + 2, \ldots, N\}$, suppose the belief of the worker is uniform across all B options. The worker should be incentivized to select all B options in this case; suppose the worker does so. Finally, for question i, suppose the worker's belief is $\delta \in (\frac{\rho}{2}, \frac{3\rho}{2})$ for the last option and $\frac{1-\delta}{|x_i|}$ each for the first $|x_i|$ options. Then the worker must be incentivized to select the first $|x_i|$ options alone if $\delta < \rho$, and select the last option along with the first $|x_i|$ options if $\delta > \rho$.

Define $\{r_j\}_{j \in [\eta+\gamma]}$ as $r_j = \delta_j$ for $j \in [\eta]$, and $r_j = 1 - \delta_j$ for $j \in \{\eta + 1, \eta + \gamma\}$. Let $\boldsymbol{\epsilon} := \{\epsilon_1, \ldots, \epsilon_{\eta+\gamma}\} \in \{-1, 1\}^{\eta+\gamma}$. The requirement of incentivizing for question i necessitates

$$
\begin{aligned}
&(1-\delta) \sum_{\boldsymbol{\epsilon} \in \{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i, B, \ldots, B) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&\quad + \delta \sum_{\boldsymbol{\epsilon} \in \{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, -x_i, B, \ldots, B) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) \\
&\overset{\delta>\rho}{\underset{\delta<\rho}{\lessgtr}} \sum_{\boldsymbol{\epsilon} \in \{-1,1\}^{\eta+\gamma}} \left(f(\epsilon_1 x_1, \ldots, \epsilon_\eta x_\eta, \epsilon_{\eta+1} x_{\eta+1}, \ldots, \epsilon_{\eta+\gamma} x_{\eta+\gamma}, x_i + 1, B, \ldots, B) \prod_{j \in [\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}} (1-r_j)^{\frac{1+\epsilon_j}{2}} \right) .
\end{aligned}
$$

The left hand side of this expression is the expected payment if the worker chooses the first $|x_i|$ options for question $(\eta+\gamma+1)$, while the right hand side is the expected payment if she chooses the first $|x_i|$ options as well as the last option. For any real-valued variable q, and for any real-valued constants a, b and c,

$$
aq \overset{q<c}{\underset{q>c}{\lessgtr}} b \quad \Rightarrow \quad ac = b \ .
$$

With $q = 1 - \delta$ in this argument, we get

$$(1-\rho)\sum_{\epsilon\in\{-1,1\}^{\eta+\gamma}}\left(f(\epsilon_1 x_1,\ldots,\epsilon_\eta x_\eta,\epsilon_{\eta+1}x_{\eta+1},\ldots,\epsilon_{\eta+\gamma}x_{\eta+\gamma},x_i,B,\ldots,B)\prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}}(1-r_j)^{\frac{1+\epsilon_j}{2}}\right)$$
$$+\rho\sum_{\epsilon\in\{-1,1\}^{\eta+\gamma}}\left(f(\epsilon_1 x_1,\ldots,\epsilon_\eta x_\eta,\epsilon_{\eta+1}x_{\eta+1},\ldots,\epsilon_{\eta+\gamma}x_{\eta+\gamma},-x_i,B,\ldots,B)\prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}}(1-r_j)^{\frac{1+\epsilon_j}{2}}\right)$$
$$-\sum_{\epsilon\in\{-1,1\}^{\eta+\gamma}}\left(f(\epsilon_1 x_1,\ldots,\epsilon_\eta x_\eta,\epsilon_{\eta+1}x_{\eta+1},\ldots,\epsilon_{\eta+\gamma}x_{\eta+\gamma},x_i+1,B,\ldots,B)\prod_{j\in[\eta+\gamma]} r_j^{\frac{1-\epsilon_j}{2}}(1-r_j)^{\frac{1+\epsilon_j}{2}}\right)=0. \tag{8.31}$$

The left hand side of (8.31) represents a polynomial in $(\eta+\gamma)$ variables $\{r_j\}_{j=1}^{\eta+\gamma}$ which evaluates to zero for all values of the variables within an $(\eta+\gamma)$-dimensional solid ball. Thus, the coefficients of the monomials in this polynomial must be zero. In particular, the constant term must be zero. The constant term appears when $\epsilon_j = 1 \ \forall\ j$ in the summations in (8.31). Setting the constant term to zero gives

$$(1-\rho)f(x_1,\ldots,x_{\eta+\gamma},x_{\eta+\gamma+1},B,\ldots,B)+\rho f(x_1,\ldots,x_{\eta+\gamma},-x_{\eta+\gamma+1},B,\ldots,B)$$
$$-f(x_1,\ldots,x_{\eta+\gamma},x_{\eta+\gamma+1}+1,B,\ldots,B)=0$$

as desired. Since the arguments above hold for any permutation of the N questions, this completes the proof for the case of $G = N$.

Now consider the case $G < N$. Let $g : \{-(B-1),\ldots,-1,1,\cdots,B\}^N \rightarrow \mathbb{R}_+$ represent the expected payment given an evaluation of all the N answers, when the identities of the gold standard questions are unknown. Here, the expectation is with respect to the (uniformly random) choice of the G gold standard questions. If $(x_1,\ldots,x_N) \in \{-(B-1),\ldots,-1,1,\cdots,B\}^N$ are the evaluations of the worker's answers to the N questions then the expected payment is

$$g(x_1,\ldots,x_N) = \frac{1}{\binom{N}{G}}\sum_{(i_1,\ldots,i_G)\subseteq\{1,\ldots,N\}} f(x_{i_1},\ldots,x_{i_G}). \tag{8.32}$$

Applying the same arguments to g as done to f above, gives

$$(1-\rho)g(x_1,\ldots,x_{\eta+\gamma},x_{\eta+\gamma+1},B,\ldots,B)+\rho g(x_1,\ldots,x_{\eta+\gamma},-x_{\eta+\gamma+1},B,\ldots,B)$$
$$-g(x_1,\ldots,x_{\eta+\gamma},x_{\eta+\gamma+1}+1,B,\ldots,B)=0. \tag{8.33}$$

The proof now proceeds via an induction on the quantity $(G-\eta-\gamma-1)$. We begin with the case of $(G-\eta-\gamma-1) = G-1$ which implies $\eta=\gamma=0$. In this case (8.31) simplifies to

$$(1-\rho)g(x_1,B,\ldots,B)+\rho g(-x_1,B,\ldots,B) = g(x_1+1,B,\ldots,B).$$

Applying the expansion of function g in terms of function f from (8.32) for some $x_1 \in [B-1]$ gives

$$(1-\rho)\left(c_1 f(x_1, B, \ldots, B) + c_2 f(B, B, \ldots, B)\right) + \rho\left(c_1 f(-x_1, B, \ldots, B) + c_2 f(B, B, \ldots, B)\right) \\ = c_1 f(x_1+1, B, \ldots, B) + c_2 f(B, B, \ldots, B)$$

for constants $c_1 > 0$ and $c_2 > 0$ that respectively represent the probabilities that the first question is picked and not picked in the set of G gold standard questions. Cancelling out the common terms on both sides of the equation, we get the desired result

$$(1-\rho) f(x_1, B, \ldots, B) + \rho f(-x_1, B, \ldots, B) = f(x_1+1, B, \ldots, B).$$

Next, we consider the case when $(G-\eta-\gamma-1)$ questions are skipped in the gold standard, and assume that the result is true when more than $(G-\eta-\gamma-1)$ questions are skipped in the gold standard. In (8.33), the functions g decompose into a sum of the constituent f functions. These constituent functions f are of two types: the first where all of the first $(\eta+\gamma+1)$ questions are included in the gold standard, and the second where one or more of the first $(\eta+\gamma+1)$ questions are not included in the gold standard. The second case corresponds to situations where there are more than $(G-\eta-\gamma-1)$ questions skipped in the gold standard and hence satisfies our induction hypothesis. The terms corresponding to these functions thus cancel out in the expansion of (8.33). The remainder comprises only evaluations of function f for arguments in which the first $(\eta+\gamma+1)$ questions are included in the gold standard. Since the last $(N-\eta-\gamma-1)$ questions are skipped by the worker, the remainder evaluates to

$$(1-\rho) c_3 f(x_1, \ldots, x_{\eta+\gamma}, x_i, B, \ldots, B) + \rho c_3 f(x_1, \ldots, x_{\eta+\gamma}, -x_i, B, \ldots, B) \\ = c_3 f(x_1, \ldots, x_{\eta+\gamma}, x_i+1, B, \ldots, B) \tag{8.34}$$

for some constant $c_3 > 0$. Dividing throughout by c_3 gives the desired result.

Finally, the arguments above hold for any permutation of the first G questions, thus completing the proof.

8.5.7 Proof of Theorem 24: Uniqueness under no-free-lunch

Without loss of generality, assume that $\alpha_{\min} = 0$ since the property of a scoring rule being a strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$.

Consider any strictly proper scoring rule f that satisfies the no-free-lunch condition. We first show that the mechanism must necessarily make a zero payment when one more more questions in the gold standard are attempted incorrectly. To this end, observe that since $f \geq 0$ and $\rho \in (0,1)$, the statement of Lemma 39 necessitates that for every $i \in [G]$ and

$(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in \{-(B-1), \ldots, B\}^{G-1}$, $x_i \in [B-1]$:

$$\begin{aligned}&\text{If } f(x_1, \ldots, x_{i-1}, x_i+1, x_{i+1}, \ldots, x_G) = 0\\ &\qquad \text{then } f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) = f(x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_G) = 0.\end{aligned}$$

A repeated application of this argument implies:

$$\text{If } f(x_1, \ldots, x_{i-1}, B, x_{i+1}, \ldots, x_G) = 0 \quad \text{then } f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G) = 0,$$

for all $x_i \in \{-(B-1), \ldots, -1, 1, \ldots, B-1\}$.

Now consider any evaluation $(x_1, \ldots, x_G)$ which has at least one incorrect answer. Suppose without loss of generality that the first question is the one answered incorrectly, i.e., $x_1 \leq -1$. The no-free-lunch condition then makes $f(x_1, B, \ldots, B) = 0$. Applying our arguments from above we get that $f(x_1, x_2, \ldots, x_G) = 0$ for every value of $(x_2, \ldots, x_G) \in \{-(B-1), \ldots, -1, 1, \ldots, B\}$.

Substituting this necessary condition in Lemma 39, we get that for every question $i \in \{1, \ldots, G\}$ and every $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_G) \in [B]^{G-1}$, $x_i \in [B-1]$, it must be that

$$f(x_1, \ldots, x_{i-1}, x_i+1, x_{i+1}, \ldots, x_G) = (1-\rho) f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_G).$$

Substituting $f(1, \ldots, 1) = \alpha$, we obtain the claimed result $f = f^*$.

8.5.8 Proof of Theorem 25: Strictly proper scoring rule to elicit thresholded beliefs

Without loss of generality, assume that $\alpha_{\min} = 0$ since the property of a scoring rule being a strictly proper is invariant to any constant shift and positive scale of the payment. We adopt the succinct notation of $\alpha := \alpha_{\max} - \alpha_{\min}$. Also recall that the term "expected payment" always refers to the expectation with respect to the worker's beliefs regarding the correctness of various options, and the randomness in the choice of the G gold standard questions in the N questions.

First consider the case of $N = G = 1$. Suppose that the worker's beliefs for the B options are $p_1, \ldots, p_B$. It is easy to verify that the expected payment, Ψ_0, when the worker selects the options $\{o_1, \ldots, o_m\}$, for some m, equals

$$B\sigma + \sum_{i=1}^{B} (p_{o_i} - \sigma).$$

Consequently, the selection of any option o_i such that $p_{o_i} < \sigma$ contributes a term $p_{o_i} - \sigma < 0$ to the expected payment, whereas the selection of any option o_j such that $p_{o_j} > \sigma$ contributes a positive amount $p_{o_i} - \sigma > 0$. It follows that the payment is strictly maximized when the worker selects all options whose beliefs are greater than σ, and does not select any option whose belief is lower than σ.

The arguments above complete the proof for the case $N = G = 1$. The extension to $N \geq G \geq 1$ follow in a manner identical to the analogous extension in the proof of Theorem 20.

8.5.9 Proof of Theorem 26: Uniqueness of Scoring rule 4

Let f denote any strictly proper scoring rule. Consider any $m \in \{1, \ldots, s_{\max} - 1\}$. Consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \cdots = p_{m+1} = \frac{1-\sigma-\delta}{m}$ and $p_{m+2} = \cdots = p_B = 0$, for some value of δ in the neighborhood of 0. For the values of m under consideration, one can verify that $\sigma < \frac{1-\sigma}{m} < 1$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$ we have $0 \le \sigma + \delta \le 1$ and $\sigma < \frac{1-\sigma-\delta}{m} \le 1$. In order to achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $(m+1)$ if $\delta > 0$, and select options 2 through $(m+1)$ if $\delta < 0$. The scoring rule f therefore must satisfy the pair of inequalities

$$f(m+1) \underset{\delta>0}{\overset{\delta<0}{\lesseqgtr}} (1-\sigma-\delta)f(m) + (\sigma+\delta)f(-m).$$

Since the right hand side of the expression above is linear in δ but the left hand side is a constant, we must have

$$f(m+1) = (1-\sigma)f(m) + \sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 1\}. \tag{8.35}$$

We will return to this set of equations later.

Next consider any $m \in \{1, \ldots, s_{\max} - 2\}$. Consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \sigma + \delta$, $p_3 = \cdots = p_{m+2} = \frac{1-2\sigma-2\delta}{m}$ and $p_{m+3} = \cdots = p_B = 0$, for some value of δ in the neighborhood of 0. For the values of m under consideration, one can verify that $\sigma < \frac{1-2\sigma}{m} < 1$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$ we have $0 \le \sigma + \delta \le 1$ and $\sigma < \frac{1-2\sigma-2\delta}{m} \le 1$. In order to achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $(m+2)$ if $\delta > 0$, and select options 3 through $(m+2)$ if $\delta < 0$. The mechanism f thus must satisfy

$$f(m+2) \underset{\delta>0}{\overset{\delta<0}{\lesseqgtr}} (1-2\sigma-2\delta)f(m) + (2\sigma+2\delta)f(-m).$$

Since the right hand side of the expression above is linear in δ but the left hand side is a constant, we must have

$$f(m+2) = (1-2\sigma)f(m) + 2\sigma f(-m) \qquad \text{for all } m \in \{1, \ldots, s_{\max} - 2\}. \tag{8.36}$$

It follows from (8.35) and (8.36) that the values of $f(m)$ for every $m \in \{-(s_{\max} - 1), \ldots, -1, 1, \ldots, s_{\max} - 2\}$ can be expressed in terms of a linear combination of $f(s_{\max})$ and $f(s_{\max} - 1)$. We will now prove that the same holds true for $f(-s_{\max})$ and $f(0)$ as well, whenever these quantities are defined.

The quantity $f(-s_{\max})$ is defined only when $s_{\max} < B$. The reason is that when $s_{\max} = B$, $f(-s_{\max}) = f(-B)$ corresponds to a scenario where all the options are selected and the correct option is not, which is impossible. Now consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \cdots = p_{s_{\max}} = \frac{1-\sigma-\delta-\epsilon}{s_{\max}-1}$, $p_{s_{\max}+1} = \epsilon$, and $p_{s_{\max}+2} = \cdots = p_B = 0$, for some values of

$\epsilon \geq 0$ and δ in the neighborhood of 0. From the definition of $s_{\max}$, one can easily verify that $\sigma < \frac{1-\sigma-\epsilon}{s_{\max}-1} < 1$ whenever $s_{\max} > 1$. Consequently, there exist some values $\delta_{\max} > 0$ and $\epsilon_{\max} \in (0, \sigma)$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$ and for every $\epsilon \in [0, \epsilon_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and when $s_{\max} > 1$, we also have $\sigma < \frac{1-\sigma-\delta-\epsilon}{s_{\max}-1} \leq 1$. In order to achieve the stated goal, we would thus require to incentivize the worker to select options 1 through $s_{\max}$ if $\delta > 0$, and select options 2 through $s_{\max}$ if $\delta < 0$. The mechanism f therefore must satisfy

$$(1-\epsilon)f(s_{\max}) + \epsilon f(-s_{\max}) \overset{\delta<0}{\underset{\delta>0}{\lesseqgtr}} (1-\sigma-\delta-\epsilon)f(s_{\max}-1) + (\sigma+\delta+\epsilon)f(-(s_{\max}-1)).$$

Since the right hand side of the expression above is linear in δ but the left hand side does not depend onδ, we must have

$$(1-\epsilon)f(s_{\max}) + \epsilon f(-s_{\max}) = (1-\sigma-\epsilon)f(s_{\max}-1) + (\sigma+\epsilon)f(-(s_{\max}-1)).$$

Since this equation must be true for every $\epsilon \in [0, \epsilon_{\max}]$, we must have

$$-f(s_{\max}) + f(-s_{\max}) = -f(s_{\max}-1) + f(-(s_{\max}-1)).$$

Thus the term $f(-s_{\max})$, whenever applicable, can also be written as a linear combination of $f(s_{\max})$ and $f(s_{\max}-1)$.

The quantity $f(0)$ is defined only when $\sigma > \frac{1}{B}$. The reason is that when $\sigma \leq \frac{1}{B}$, it is mathematically impossible for the beliefs for all the B options to be less than or equal to σ (recall our assumption that no belief equals exactly σ). Now consider the set of beliefs $p_1 = \sigma + \delta$, $p_2 = \cdots = p_B = \frac{1-\sigma-\delta}{B-1}$, for some value of δ in the neighborhood of 0. One can verify that in this case of $\sigma > \frac{1}{B}$, it must be that $0 < \frac{1-\sigma}{B-1} < \sigma$. Consequently, there exists some value $\delta_{\max} > 0$ such that for every $\delta \in [-\delta_{\max}, \delta_{\max}]$, we have $0 \leq \sigma + \delta \leq 1$ and $0 \leq \frac{1-\sigma-\delta}{B-1} < \sigma$. In order to achieve the stated goal, we would thus require to incentivize the worker to select option 1 if $\delta > 0$, and select no options if $\delta < 0$. The mechanism f therefore must satisfy

$$(\sigma+\delta)f(1) + (1-\sigma-\delta)f(-1) \overset{\delta<0}{\underset{\delta>0}{\lesseqgtr}} f(0).$$

Since the left hand side of the expression above is linear in δ but the right hand side is a constant, we must have

$$\sigma f(1) + (1-\sigma)f(-1) = f(0).$$

Thus the term $f(0)$, whenever applicable, can also be written as a linear combination of $f(s_{\max})$ and $f(s_{\max}-1)$.

From the arguments above, we get that the design of f has only two degrees of freedom. Given that our claim is only up to some shift and scale, the claim is proved.

8.A Appendix: Auxiliary negative results

In this section we present a pair of auxiliary results that were referred to in the main text of the chapter.

8.A.1 Impossibility of support elicitation without the coarse belief assumption

In the setting of eliciting support of beliefs (Section 8.2.3 and Section 8.3.1), we made a coarse-beliefs assumption that the probability of correctness of any option, according to the worker's belief, must either be zero or exceed a certain threshold ρ. The following proposition shows that there exists no strictly proper scoring rule in the absence of this assumption.

Proposition 13. *For any N, G and $B \geq 2$, there is no strictly proper scoring rule towards incentivizing the worker to select precisely the support of her distribution for each question.*

To put this negative result in perspective with the positive results of Section 8.3.1, observe that $\rho = 0$ reduces Scoring rule 3 to $f^*(x_1, \ldots, x_G) = \kappa^* \prod_{i=1}^{G} \mathbf{1}\{x_i \geq 1\} + \alpha_{\min}$. One can see that it no longer remains a strictly proper scoring rule: the worker is incentivized to simply select all options for every question. The impossibility result of Theorem 13 proves that every possible mechanism must necessarily suffer this fate.

Proof of Proposition 13

We assume that there indeed exists some strictly proper scoring rule f, and prove a contradiction.

Let us first consider the special case of $N = G = 1$ and $B = 2$. Since $N = G = 1$, there is only one question. Let $p_1 > 0.5$ be the probability, according to the belief of the worker, that option 1 is correct; the worker then believes that option 2 is correct with probability $(1 - p_1)$.

When $p_1 = 1$, we need the worker to select option 1 alone. Thus we need

$$f(1) > f(2).$$

When $p_1 \in (0.5, 1)$, we require the worker to select options 1 and 2, as opposed to selecting option 1 alone. For this we need

$$p_1 f(1) + (1 - p_1) f(-1) < f(2)$$

It follows that we need

$$(1 - p_1)(f(1) - f(-1)) > f(1) - f(2). \tag{8.37}$$

However, the inequality (8.37) is satisfied only when $f(1) > f(-1)$ and $(1-p_1) > \frac{f(1)-f(2)}{f(1)-f(-1)}$. Thus for any given payment function f, a worker with belief $(1-p_1) \in (0, \frac{f(1)-f(2)}{f(1)-f(-1)})$ will not be incentivized to select the support of her belief. This yields a contradiction.

We now move on to the general case of $N \geq G \geq 1$ and $B \geq 2$. Consider a worker who is clueless about questions 2 through N (i.e., her belief is uniform across all options for these questions). Suppose this worker selects all B options for these questions as desired. For the first question, suppose that the worker is sure that options $3, \ldots, B$ are incorrect. We are now left with the first question and the first two options for this question. Letting X denote a random variable representing the evaluation of the worker's response to the first question, the expected payment then is

$$\frac{G}{N}\mathbb{E}[f(X, B, \ldots, B)] + (1 - \frac{G}{N})f(B, \ldots, B).$$

The expectation in the first term is taken with respect to the randomness in X. Defining

$$\tilde{f}(X) := \frac{G}{N}f(X, B, \ldots, B) + (1 - \frac{G}{N})f(B, \ldots, B),$$

and applying the same arguments to $\tilde{f}$ as those for f for the case of $N = G = 1$, $B = 2$ above gives the desired contradiction. This thus completes the proof of impossibility.

8.A.2 Impossibility of thresholded-belief elicitation when a belief exactly equals the threshold

Recall that when defining a strictly proper scoring rule for the setting of eliciting options with beliefs above a certain threshold σ (Section 8.2.3), we did not restrict the scoring rule to any specific choice when the probability of the correctness of an option equaled exactly σ. This is because, as one would intuitively expect, incentivizing a certain action at the boundary value of σ may not be possible. The following proposition provides a formal proof for this claim.

Proposition 14. *For any $N \geq G \geq 1$, there is no strictly proper scoring rule in the absence of this assumption.*

The remainder of this section is devoted to the proof of this claim.

Proof of Proposition 14

Let us first prove the result for the case of $N = G = 1$. The result of Theorem 26 implies that if there does exist a strictly proper scoring rule for this setting, then it must be Scoring rule 4 up to a constant shift and positive scale. Consider a worker with the belief $p_1 = 1-\sigma$, $p_2 = \sigma$ and $p_3 = \cdots p_B = 0$. Since $\sigma < \frac{1}{2}$, under a strictly proper scoring rule, the expected payment must be strictly larger if the worker selects only option 1 as compared to the expected payment when the worker selects options 1 and 2. However, one can compute that under Scoring rule 4, the expected payment in the two cases is identical. It follows that under any possible strictly proper scoring rule, the expected payment must be identical in the two following two actions of the worker (a) selecting only option 1, and (b) selecting options 1 and 2. It follows that there is no strictly proper scoring rule.

We now move on to the general case of $N \geq G \geq 1$. Let f denote any strictly proper scoring rule for the setting at hand. Consider a worker who knows the answers to questions 2 through N with a belief of 1 in each case. Suppose that for each of these $(N-1)$ questions, this worker selects the respective options that she thinks are correct. We are now left with the first question. Letting X denote a random variable representing the evaluation of the worker's response to the first question, the expected payment from the worker's point of view is

$$\frac{G}{N}\mathbb{E}[f(X, 1, \ldots, 1)] + (1 - \frac{G}{N})f(1, \ldots, 1).$$

The expectation in the first term is taken with respect to the randomness in X. Defining

$$\tilde{f}(X) := \frac{G}{N}f(X, 1, \ldots, 1) + (1 - \frac{G}{N})f(1, \ldots, 1),$$

and applying the same arguments to $\tilde{f}$ as those for f for the case of $N = G = 1$ above gives the desired contradiction. This completes the proof.

Part III

Conclusions

Chapter 9

Conclusions

"Please stop attributing your made up quotes to my fellow scientists."

– Issac Newton

Learning from people is the next frontier for data science. There are two primary challenges associated to learning from people. One challenge is to design estimation algorithms that are robust to modeling assumptions. To address this challenge, in this thesis we propose permutation-based models and estimators which we prove provide strong guarantees under very little assumptions. We show that this permutation-based approach has a multitude of benefits as compared to the classical approach involving restrictive parameter-based assumptions. The second challenge is to elicit high quality data from people, and in this thesis we design multiplicative incentives that we prove are the one and only mechanisms that can provably guarantee that the natural requirements of crowdsourcing platforms are met. All in all, this thesis contributes to the fundamental understanding of the problems in this area, designs algorithms that yield notable improvements in practice, and has also had real-world immediate impact.

The general principles of permutation-based models and estimators and unique multiplicative mechanisms have implications well beyond our present motivation of crowdsourcing. In many applications in machine learning and statistics, one often assumes models that are of the parameter-based form. Such models are popular partly because they are quite intuitive to write down, and partly because they are often analytically more tractable. However, instead if one were to consider rich enough models like permutation-based models then one can obtain a broader perspective and richer insights into the problem that can lead to obtain superior results. As an illustration, Appendix 9.A presents a permutation-based generalization of the classical normal means model. Moving on to game theory and decision theory, it is often the case that a mechanism is employed because it is incentive compatible or because truth telling is an equilibrium. However, it is usually the case that the mechanism employed may be just one of many possible mechanisms that has this property; it is seldom established that there is no other mechanism that can dominate the one under consideration.

Our simple no-free-lunch requirement and associated proof techniques are tools that can be valuable towards this goal in many applications. For instance, in Appendix 9.B we show how our proof techniques and ideas can be used to design unique grading schemes for objective examinations.

There are several open problems that emanate from this thesis, and we enumerate a few of them here. In Part I, we primarily considered a "random design" setting where the choice of the pairs compared or questions asked or entries observed is made uniformly at random. It remains to evaluate the performance of permutation-based models for other observation models such as weighted random sampling, fixed design, streaming or active learning, or biased observation models. In our paper [100], we have made progress in this front for the problem of ranking in an active setting where the pairs to be compared can be chosen based on outcomes of past comparisons. We show that even in this active setting, the same story holds: the restrictive assumptions of parameter-based models offer very little help as compared to the permutation-based models.

A second open problem is to close the gap between the error guarantees of the statistically optimal estimators and what can be achieved by computationally efficient estimators in the settings of Chapters 2, 4, and 5. We do not know at this point whether this gap is really fundamental or whether there exist computationally efficient estimators that can achieve the statistically optimal rates. There are however two key exceptions on such a gap. For the problem of minimizing the sample complexity of ranking, we know from Chapter 3 and our paper [100] that there is no gap – the statistically optimal algorithm (up to logarithmic factors) is also computationally efficient. On the other hand, for estimation of pairwise comparison probabilities in a manner that can adapt to underlying smoothness in the true probabilities, we know from our companion paper [223] that there is indeed a provable gap (conditioned on the planted clique hardness conjecture).

In Part II of this thesis, we designed incentive mechanisms that operate using some gold standard questions in the tasks. However, it is often the case that there are no gold standard questions available or that gold standard questions are too expensive to create. An interesting problem is to construct mechanisms that operate in the absence of gold standard questions, for instance, by comparing the answers of a worker to those of others in some fashion. There is a long line of literature on this topic [57, 171, 188, 190, 191, 239]. However, unlike the multiplicative mechanisms of this thesis, to the best of our knowledge none of the incentive-compatible mechanisms for settings without gold standard questions are simple enough to be understood by workers on crowdsourcing platforms like Amazon Mechanical Turk. In some of our own recent work [115], we have obtained mechanisms that are simper than the rest, but we still think there is a long way to go.

A second open problem that concerns practical deployment of incentive mechanisms is that of choosing the hyperparameters in the various mechanisms. In this thesis, we assumed that the threshold for skip, number and thresholds of confidence levels, the maximum pay for a task, and the number of gold standard questions in a task are all given to us. It is not known how to choose these values in a principled manner, possibly in a way that can adapt across tasks.

The final open problem we discuss is that of designing estimation and aggregation schemes that operate mindful of the incentive mechanisms used to collect the data as well as any additional information obtained from the interface used. For instance, the questions skipped in the skip-based mechanism of Chapter 6 are not skipped randomly but depend on the underlying question itself. Likewise, the self-reported confidences in the mechanism of Chapter 7 and the set of selected options in the mechanism of Chapter 8 are all driven by the underlying mechanism and interface. In contrast, the estimation algorithms designed in Chapter 4 for labeling tasks as well as in other chapters of Part I operate agnostic of this information. Establishing the fundamental limits of the use of this information in the estimation process and designing practical algorithms for this purpose remain worthwhile open problems.

We conclude with two longer-term directions of future research. The first direction is on robust statistics and machine learning. In various problems in machine learning and statistics, one often makes many assumptions that may not be fundamental to the problem. These include assumptions of existence of specific parameters or utility functions or specific priors etc. In order to remove these restrictive assumptions, it is of interest to identify the fundamental requirements of any such problem, and design models and estimation algorithms that provide maximal accuracy under minimal assumptions. In other words, it will be useful to design algorithms that are provably pareto optimal in the tradeoff between the modeling assumptions and the error under the assumed model. Secondly, in this thesis we considered settings with structured data of the form of ranking and classification. An important direction of future research is learning from unstructured human data. Human-centered systems with such unstructured data have massive potential, but also face many tall challenges. Such data arises in citizen science, healthcare, search and rescue, language translation, as well as in artificial intelligence in terms of imitation learning. It is of interest to approach these problems from a foundational mathematical perspective, and use the insights obtained therein to design practical algorithms for these practically useful settings.

9.A Appendix: Permutation-based generalization of the normal means model

In this section we consider the classical setup of the normal means model. In the normal means model, there are d latent "means" $m := [m_1, \ldots, m_d]^T \in \mathbb{R}^d$ that are required to be estimated. The setting assumes to make n independent observations of each mean value, that is, observe $Y \in \mathbb{R}^{n \times d}$ where the entries of Y are mutually independent and $Y_{ij} \sim \mathcal{N}(m_i, 1)$.

In accordance with the motivation of this section to connect with Part I, we assume that the vector of unknown means m has entries bounded as $\|m\|_\infty \leq 1$. We comment on removing this assumption later. Then we define the normal means model $\mathbb{C}_{\text{NORM}}$ as

$$\mathbb{C}_{\text{NORM}} := \{M \in [0,1]^{n \times d} \mid M = 1m^T \text{ for some } m \in [-1,1]^d\}.$$

The normal means model is useful for a variety of settings. For instance, consider the problem of modeling and estimating the average profit due to various items in a store chain (such as Walmart). Here d represents the number of items, m_j represents the true latent mean of the profit due to item $j \in [d]$, and the n associated observations $Y_{1j}, \ldots, Y_{nj}$ represent the actual profits associated to item j in n different branches of the store.

Observe that in the example application of the store chain, the standard normal means setting implies the assumption that for each item, every branch of the store chain has the same latent expected profit. This assumption may often be (severely) violated, for instance, if one branch is in a highly populated area while another is in an area with a lower population density. In this case, it may be reasonable to assume that the relative profits of any two items are consistent across the branches, and that the relative profits in any two branches are consistent across the different products.

The observations above leads us to the following "permutation-based" relaxation of the normal means model. Consider a matrix $M^* \in \mathbb{R}^{n \times d}$ of the underlying means of profits across branches and products, where $[M^*]_{ij}$ represents the latent mean profit due to product $i \in [n]$ in branch $j \in [n]$ of the store. Formalizing the above discussion, we assume that $M^* \in \mathbb{C}_{\text{PERM}}$, where $\mathbb{C}_{\text{PERM}}$ is defined as

$$\begin{aligned}\mathbb{C}_{\text{PERM}} := \{M \in [-1,1]^{n \times d} \; | M_{ij} \geq M_{i'j'} \text{ whenever } \pi(i) < \pi(i')\text{and } \sigma(j) < \sigma(j'), \\ \text{for some permutations } \pi \text{ and } \sigma\}.\end{aligned}$$

As before, the observation consists of a matrix $Y \in \mathbb{R}^{n \times d}$ with $Y_{ij} \sim \mathcal{N}([M^*]_{ij}, 1)$ for every pair (i, j), independent of all else.

It is easy to see that the permutation-based model is more general than the classical model. The following theorem now presents bounds relating to estimation of the vector $m^* \in \mathbb{C}_{\text{NORM}}$ and more generally the matrix $M^* \in \mathbb{C}_{\text{PERM}}$.

Proposition 15. *(a) Suppose that m^* is known to follow the normal means model $\mathbb{C}_{NORM}$, and consider any estimator $\widehat{m}$ for m^*. The estimator must incur an average per-entry error of at least*

$$\sup_{m^* \in \mathbb{C}_{NORM}} \mathbb{E}\Big[\frac{1}{d}\sum_{j=1}^{d}(\widehat{m}_j - m_j^*)^2\Big] \geq c_2 \frac{1}{n}.$$

(b) Suppose that $d \geq n$. Then over the permutation-based normal means model, the average per-entry error of the least squares estimator $\widehat{M}_{LS} := \arg\min_{M \in \mathbb{C}_{PERM}} \lvert\!\lvert\!\lvert Y - M \rvert\!\rvert\!\rvert_F$ is at most

$$\sup_{M^* \in \mathbb{C}_{PERM}} \mathbb{E}\Big[\frac{1}{n}\sum_{j=1}^{n}\frac{1}{d}\sum_{i=1}^{d}([\widehat{M}_{LS}]_{ij} - M_{ij}^*)^2\Big] \leq c_1 \frac{1}{n}\log^2 d.$$

Returning to our example applications of the store-chain, the number of items d in a typical store will generally outnumber the number of branches n. The bounds of parts (a) and (b) are tight in this regime, showing that the average per-entry error in either case is of the order of $\frac{1}{n}$. A consequence of this result is that if one is fine with a possible logarithmic factor increase in the error, then from a statistical viewpoint, the permutation-based normal means model may be preferable due to its added generality and comparable error.

The proof of Proposition 15(a) is standard folklore, and can be established through the framework of Fano's inequality that is used in most proofs of lower bounds in Part I this thesis. The proof of Proposition 15(b) is similar to the analysis of the least squares estimator in Theorem 1(a) with one exception. The bound (2.30) in the proof of Theorem 1(a) relies on the fact that the entries of the noise are bounded in the interval $[-1, 1]$ thereby allowing for the use of the relevant concentration bounds from [144, Theorem 5.9], [213]; in order to accommodate the present setting with (unbounded) normally distributed noise, we must instead use the tail bound [136, Theorem 2.2]. Finally, one may remove the assumption that M^* is bounded by considering a an estimator that finds the least squares estimate within a infinity norm ball of $\text{polylog}(n, d)$ around the observed matrix Y, along with the monotonicity constraints. Using the fact that the maximum of nd standard normal variables is at most $\text{polylog}(n, d)$ with high probability, the bound of Proposition 15(b) may be worsened by at most logarithmic factors.

All in all, we see that even though the setting of the classical normal means model that is not naturally connected to the theme of learning from people explored in this thesis, the high-level uses and benefits of permutation-based models and estimators carry over naturally.

9.B Appendix: Unique grading schemes for objective examinations

Consider the grading of a homework or an examination. Consider a given grading scheme that evaluates every question as either "correct" or "incorrect", and computes the score of a student based on the evaluation of his/her answers to the set of questions asked. This grading scheme may have been designed and revised over the years to ensure that it possesses many appealing properties. Now suppose you wish to extend this setting to one where the student gets an option to skip any question that she is not confident about. (The meanings of confidence and incentivization are as defined earlier in Part II.) Is there a rigorous way to extend the original grading scheme to this new setting, while preserving the appealing properties of the original grading scheme, and incentivizing the students to skip questions (only) when their confidence is below a certain threshold? In this section, we answer this question in the affirmative, and show that there is one and only one way to do so.

Consider an exam or a homework with N questions. An evaluation of an answer as correct is denoted by $+1$ and as incorrect is denoted by -1. Let $g : \{-1, 1\}^N \rightarrow \mathbb{R}$ denote the grading scheme that takes as input the evaluation of the student's responses to the N questions and outputs the final score. The only requirement on g that we impose is that if a student knows the answer to a question and is 100% confident about it, then the scheme g must incentivize the student to provide the answer she thinks is correct. Note that under this setting, a question that is skipped by a student is considered as answered incorrectly.

Now suppose you wish to extend this setting to one where the student gets an option to skip any question she is not sure of. In particular, we consider some threshold $T \in (0, 1)$ such that we want the student to skip a question if her confidence about the answer is smaller than T and answer if it is greater than T. We take an axiomatic approach towards the design of the grading scheme for this setting, and impose two simple and natural conditions on the grading scheme. The goal is to design a grading scheme $f : \{-1, 0, 1\}^N \to \mathbb{R}$ that satisfies the two following requirements:

- **Backward compatibility:** When no questions are skipped, the grade should be identical to what the grading scheme g would have given, that is, $f(x_1, \ldots, x_N) = g(x_1, \ldots, x_N)$ for every $(x_1, \ldots, x_N) \in \{-1, 1\}^N$. This requirement ensures that all the features of the earlier grading scheme are retained.
- **Skipping criterion:** For a fixed threshold $T \in (0, 1)$, for any question, if the student's confidence in any answer is more than T then the student should be incentivized to give that answer, otherwise the student should be incentivized to skip the question.

In what follows, we present an algorithm to obtain a grading scheme that incorporates skipping of questions from a grading scheme without it, and prove that this is the only scheme that is backward compatible and satisfies the skipping criterion.

Unique grading scheme

The proposed grading scheme f is constructed follows. Consider any set of evaluations $(y_1, \ldots, y_N) \in \{-1, 0, 1\}^N$. For every $i \in \{1, \ldots, N\}$, let $A_i = \{-1, 1\}$ if $y_i = 0$ and $A_i = \{y_i\}$ otherwise. Finally, set

$$f(y_1, \ldots, y_N) = \sum_{(x_1,\ldots,x_N) \in A_1 \times \cdots \times A_N} g(x_1, \ldots, x_N) T^{\sum_{i=1}^N \mathbf{1}\{x_i = 1,\ y_i = 0\}} (1 - T)^{\sum_{i=1}^N \mathbf{1}\{x_i = -1,\ y_i = 0\}}.$$

The expression for f in the proposed scheme may appear somewhat complicated at first, but it has a quite simple interpretation. For every question that is skipped by a student, take the convex combination of the scores for the case where the student answers that question correctly, with a weight T, and the case where the student answers that question incorrectly, with a weight $(1 - T)$. In other words, the scheme may equivalently be written in the following recursive form:

$$\begin{aligned} &f(y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_N) \\ &\quad = T f(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_N) + (1 - T) f(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_N). \end{aligned}$$

The following theorem proves that our proposed grading scheme is the one and only solution.

Proposition 16. *The proposed grading scheme is the one and only grading scheme that is backward compatible and satisfies the skipping criterion.*

The proof of this result follows directly from the results established earlier in this chapter. The proof that it satisfies the skipping criterion follows by simply evaluating the expected score under various actions as done in the proof of Theorem 16. The uniqueness follows from a recursive application of Lemma 33.

Finally, we note that the proposed unique grading scheme f is as simple or complex as the original scheme g. For instance, the range of scores provided under the new scheme is identical to the range under the original scheme. The following pair of important subclasses illustrate additional interesting properties.

Homogeneous grading scheme: Let us consider a popular subclass where the original grading scheme g puts an equal weight on every question, that is, $g(x_1, \ldots, x_N) = \hat{g}(r, w)$ where $r = \sum_{i=1}^N \mathbf{1}\{x_i = 1\}$ is the number of right answers and $w = \sum_{i=1}^N \mathbf{1}\{x_i = -1\}$ is the number of wrong answers.

Proposition 16 then implies that the one and only grading scheme f that meets the skipping criterion and is backward compatible is as follows:

$$\hat{f}(r, w, s) = \sum_{k=0}^{s} \binom{s}{k} T^k (1-T)^{s-k} g(r+k, w+s-k),$$

where $s = \sum_{i=1}^N \mathbf{1}\{x_i = 0\}$ is the number of skipped questions. Observe that this scheme is also homogeneous, that is, the final score depends only on the number of right, wrong, and skipped answers.

Let us first interpret this resulting scheme. Consider some values of r, w and s. If the student had attempted and answered all the s skipped questions incorrectly then her score would have been $g(r, w+s)$. On the other hand, if the student had attempted answered all the s skipped questions correctly then her score would have been $g(r+s, w)$. The score in the case when the student skips these s questions is simply a convex combination of these two terms and all the terms in between. In particular, our provably unique scheme when applied to r correct and s skipped answers, simply equals the expected value of the score under the original scheme g if each skipped question was actually attempted and each provided answer independently had a probability T of being correct.

Additive grading scheme: Next we consider a subclass in which each question is allowed to be evaluated in a different manner, but the final score is a sum of the scores in each individual question, that is, $g(x_1, \ldots, x_N) = \tilde{g}_0 + \sum_{i=1}^N \tilde{g}_i(x_i)$, for some constant $\tilde{g}_0 \in \mathbb{R}$ and functions $\tilde{g}_1, \ldots, \tilde{g}_N : \{-1, 1\} \rightarrow \mathbb{R}$.

Proposition 16 then implies that the one and only grading scheme f that meets the skipping criterion and is backward compatible is:

$$f(y_1, \ldots, y_N) = \tilde{f}_0 + \sum_{i=1}^{N} f_i(y_i),$$

where $\tilde{f}_0 = \tilde{g}_0$, and for every $i \in \{1, \ldots, N\}$,

$$f_i(y_i) = \begin{cases} \tilde{g}_i(y_i) & \text{if } y_i \in \{-1, 1\} \\ T\tilde{g}_i(1) + (1 - T)\tilde{g}_i(-1) & \text{otherwise.} \end{cases}$$

The additive nature of the original scheme is thus retained.

Bibliography

[1] N. Ailon, M. Charikar, and A. Newman. "Aggregating inconsistent information: ranking and clustering". In: *Journal of the ACM (JACM)* 55.5 (2008), p. 23.

[2] N. Alon. "Ranking tournaments". In: *SIAM Journal on Discrete Mathematics* 20.1 (2006), pp. 137–142.

[3] A. Ammar and D. Shah. "Efficient rank aggregation using partial data". In: *ACM SIGMETRICS Performance Evaluation Review.* 2012.

[4] N. Anari, G. Goel, and A. Nikzad. "Mechanism Design for Crowdsourcing: An Optimal 1-1/e Competitive Budget-Feasible Mechanism for Large Markets". In: *Foundations of Computer Science (FOCS).* 2014, pp. 266–275.

[5] D. Angluin and P. Laird. "Learning from noisy examples". In: *Machine Learning* 2.4 (1988), pp. 343–370.

[6] A. Archer and É. Tardos. "Truthful mechanisms for one-parameter agents". In: *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on.* IEEE. 2001, pp. 482–491.

[7] S. Arora, R. Ge, R. Kannan, and A. Moitra. "Computing a nonnegative matrix factorization–provably". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing.* ACM. 2012, pp. 145–162.

[8] B. Babcock and C. Olston. "Distributed top-k monitoring". In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data.* 2003, pp. 28–39.

[9] M.-F. Balcan and H. Zhang. "Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling". In: *Advances In Neural Information Processing Systems.* 2016.

[10] J. Baldridge and A. Palmer. "How well does active learning actually work?: Time-based evaluation of cost-reduction strategies for language documentation". In: *Conference on Empirical Methods in Natural Language Processing.* 2009, pp. 296–305.

[11] T. P. Ballinger and N. T. Wilcox. "Decisions, error and heterogeneity". In: *The Economic Journal* 107.443 (1997), pp. 1090–1105.

[12] A. S. Bandeira and R. van Handel. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries". In: *arXiv preprint arXiv:1408.6185* (2014).

[13] W. Barnett. “The modern theory of consumer behavior: Ordinal or cardinal?” In: *The Quarterly Journal of Austrian Economics* 6.1 (2003), pp. 41–65.

[14] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, et al. “Soylent: a word processor with a crowd inside”. In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology.* ACM. 2010, pp. 313–322.

[15] S. Bernstein. “On a modification of Chebyshev’s inequality and of the error formula of Laplace”. In: *Ann. Sci. Inst. Sav. Ukraine, Sect. Math* 1.4 (1924), pp. 38–49.

[16] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, et al. “VizWiz: nearly real-time answers to visual questions”. In: *ACM symposium on User interface software and technology.* 2010, pp. 333–342.

[17] J. Bohannon. “Social Science for Pennies”. In: *Science* 334.6054 (2011), pp. 307–307.

[18] J. C. de Borda. “Mémoire sur les élections au scrutin”. In: (1781).

[19] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

[20] R. A. Bradley and M. E. Terry. “Rank analysis of incomplete block designs: I. The method of paired comparisons”. In: *Biometrika* (1952), pp. 324–345.

[21] S. J. Brams and P. C. Fishburn. “Approval voting”. In: *American Political Science Review* 72.03 (1978), pp. 831–847.

[22] S. J. Brams and D. M. Kilgour. “Satisfaction approval voting”. In: *Voting Power and Procedures.* Springer, 2014, pp. 323–346.

[23] M. Braverman and E. Mossel. “Noisy sorting without resampling”. In: *Proc. ACM-SIAM symposium on Discrete algorithms.* 2008, pp. 268–276.

[24] G. W. Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[25] G. Bril, R. Dykstra, C. Pillers, and T. Robertson. “Isotonic regression in two independent variables”. In: *Journal of the Royal Statistical Society* (1984).

[26] A. E. Brouwer and W. H. Haemers. *Spectra of graphs.* Springer, 2011.

[27] P. Bühlmann and T. Hothorn. “Boosting algorithms: Regularization, prediction and model fitting”. In: *Statistical Science* (2007), pp. 477–505.

[28] A. Buja, W. Stuetzle, and Y. Shen. “Loss functions for binary class probability estimation and classification: Structure and applications”. In: *Working draft, November* (2005).

[29] J. D. Burger, E. Doughty, S. Bayer, D. Tresner-Kirsch, B. Wellner, J. Aberdeen, et al. “Validating candidate gene-mutation relations in MEDLINE abstracts via crowdsourcing”. In: *Data Integration in the Life Sciences* (2012), pp. 83–91.

[30] R. Busa-Fekete, B. Szorenyi, W. Cheng, P. Weng, and E. Hüllermeier. “Top-k selection based on adaptive sampling of noisy preferences”. In: *International Conference on Machine Learning.* 2013.

[31] J.-F. Cai, E. J. Candès, and Z. Shen. “A singular value thresholding algorithm for matrix completion”. In: *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982.

[32] J.-F. Cai and S. Osher. “Fast singular value thresholding without singular value decomposition”. In: *UCLA CAM Report* 5 (2010).

[33] Y. Cai, C. Daskalakis, and C. H. Papadimitriou. “Optimum Statistical Estimation with Strategic Data Sources”. In: *Conference on Learning Theory.* 2015.

[34] E. J. Candes and B. Recht. “Exact Matrix Completion via Convex Optimization”. In: *Found. Comput. Math.* 9.6 (Dec. 2009), pp. 717–772.

[35] E. J. Candès and T. Tao. “The Power of Convex Relaxation: Near-optimal Matrix Completion”. In: *IEEE Trans. Inf. Theor.* 56.5 (May 2010), pp. 2053–2080.

[36] I. J. Cano, Y. A. Dimitriadis, S. E. Gómez, and C. J. López. “Learning from noisy information in FasArt and FasBack neuro-fuzzy systems.” In: *Neural networks: the official journal of the International Neural Network Society* 14.4-5 (2001), pp. 407–425.

[37] I. Caragiannis, D. Kalaitzis, and E. Markakis. “Approximation Algorithms and Mechanism Design for Minimax Approval Voting.” In: *AAAI.* 2010.

[38] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell. “Coupled semi-supervised learning for information extraction”. In: *Proceedings of the third ACM international conference on Web search and data mining.* ACM. 2010, pp. 101–110.

[39] D. Carmel, D. Cohen, R. Fagin, E. Farchi, M. Herscovici, Y. S. Maarek, et al. “Static index pruning for information retrieval systems”. In: *ACM SIGIR conference on Research and development in information retrieval.* 2001.

[40] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, et al. “CellProfiler: image analysis software for identifying and quantifying cell phenotypes”. In: *Genome biology* 7.10 (2006), R100.

[41] M. Cattelan. “Models for paired comparison data: A review with emphasis on dependent data”. In: *Statistical Science* 27.3 (2012), pp. 412–433.

[42] S. Chatterjee, A. Guntuboyina, and B. Sen. “On matrix estimation under monotonicity constraints”. In: *arXiv:1506.03430* (2015).

[43] S. Chatterjee and S. Mukherjee. “On Estimation in Tournaments and Graphs under Monotonicity Constraints”. In: *arXiv preprint arXiv:1603.04556* (2016).

[44] S. Chatterjee. “Matrix estimation by universal singular value thresholding”. In: *The Annals of Statistics* 43.1 (2014), pp. 177–214.

[45] J. J. Chen, N. J. Menezes, A. D. Bradley, and T. North. "Opportunities for crowdsourcing research on amazon mechanical turk". In: *Interfaces* 5.3 (2011).

[46] Y. Chen and C. Suh. "Spectral MLE: Top-K Rank Aggregation from Pairwise Comparisons". In: *International Conference on Machine Learning*. 2015.

[47] F. Chu, Y. Wang, and C. Zaniolo. "An adaptive learning approach for noisy data streams". In: *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE. 2004, pp. 351–354.

[48] F. Chung and M. Radcliffe. "On the spectra of general random graphs". In: *The electronic journal of combinatorics* 18.1 (2011), P215.

[49] L. S. Collet. "Elimination scoring: An empirical evaluation". In: *Journal of Educational Measurement* 8.3 (1971), pp. 209–214.

[50] V. Conitzer. "Prediction markets, mechanism design, and cooperative game theory". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2009, pp. 101–108.

[51] C. H. Coombs. "On the use of objective examinations". In: *Educational and Psychological Measurement* 13.2 (1953), pp. 308–310.

[52] C. H. Coombs, J. E. Milholland, and F. B. Womer. "The assessment of partial knowledge". In: *Educational and Psychological Measurement* 16.1 (1956), pp. 13–37.

[53] A. H. Copeland. "A reasonable social welfare function". In: *University of Michigan Seminar on Applications of Mathematics to the social sciences*. 1951.

[54] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[55] K. Crowston. "Amazon mechanical turk: A research tool for organizations and information systems scholars". In: *Shaping the Future of ICT Research. Methods and Approaches*. Springer, 2012, pp. 210–221.

[56] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. "Aggregating crowdsourced binary ratings". In: *Conference on World Wide Web*. 2013, pp. 285–294.

[57] A. Dasgupta and A. Ghosh. "Crowdsourced judgement elicitation with endogenous proficiency". In: *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2013, pp. 319–330.

[58] M. A. Davenport and J. Romberg. "An overview of low-rank matrix recovery from incomplete observations". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), pp. 608–622.

[59] D. Davidson and J. Marschak. "Experimental tests of a stochastic decision theory". In: *Measurement: Definitions and theories* (1959), pp. 233–69.

[60] A. Dawid and A. Skene. “Maximum likelihood estimation of observer error-rates using the EM algorithm”. In: *Applied Statistics* (1979), pp. 20–28.

[61] O. Dekel, F. Fischer, and A. D. Procaccia. “Incentive compatible regression learning”. In: *ACM-SIAM symposium on Discrete algorithms*. 2008, pp. 884–893.

[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *IEEE Conference on Computer Vision and Pattern Recognition, 2009*. IEEE. 2009, pp. 248–255.

[63] P. Diaconis. “A generalization of spectral analysis with application to ranked data”. In: *The Annals of Statistics* 17.3 (1989), pp. 949–979.

[64] P. Diaconis and R. L. Graham. “Spearman’s footrule as a measure of disarray”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), pp. 262–268.

[65] W. Ding, P. Ishwar, and V. Saligrama. “A Topic Modeling Approach to Ranking”. In: *Conference on Artificial Intelligence and Statistics*. 2015.

[66] J.-P. Doignon, A. Pekeč, and M. Regenwetter. “The repeated insertion model for rankings: Missing link between two subset choice models”. In: *Psychometrika* 69.1 (2004), pp. 33–54.

[67] D. Donoho and M. Gavish. “Minimax risk of matrix denoising by singular value thresholding”. In: *The Annals of Statistics* 42.6 (2014), pp. 2413–2440.

[68] D. Donoho and V. Stodden. “When does non-negative matrix factorization give a correct decomposition into parts?” In: *Advances in neural information processing systems*. 2003.

[69] C. Eickhoff and A. de Vries. “How crowdsourcable is your task”. In: *Crowdsourcing for search and data mining*. 2011.

[70] P. Erdős and A. Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5 (1960), pp. 17–61.

[71] B. Eriksson. “Learning to Top-K search using pairwise comparisons”. In: *Conference on Artificial Intelligence and Statistics*. 2013.

[72] A. D. Ewing, K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, et al. “Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection”. In: *Nature methods* 12.7 (2015), pp. 623–630.

[73] R. Fagin, A. Lotem, and M. Naor. “Optimal aggregation algorithms for middleware”. In: *Journal of computer and system sciences* 66.4 (2003), pp. 614–656.

[74] J.-C. Falmagne and M. Regenwetter. “A random utility model for approval voting”. In: *Journal of Mathematical Psychology* 40.2 (1996), pp. 152–159.

[75] F. Fang, M. Stinchcombe, and A. Whinston. “Putting Your Money Where Your Mouth Is: A Betting Platform for Better Prediction”. In: *Review of Network Economics* 6.2 (2007).

[76] V. F. Farias, S. Jagabathula, and D. Shah. “A nonparametric approach to modeling choice with limited data”. In: *Management Science* 59.2 (2013), pp. 305–322.

[77] W. Feller. “Generalization of a probability limit theorem of Cramér”. In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 361–372.

[78] P. C. Fishburn. “Binary choice probabilities: on the varieties of stochastic transitivity”. In: *Journal of Mathematical psychology* 10.4 (1973), pp. 327–352.

[79] M. A. Fligner and J. S. Verducci. *Probability models and statistical analyses for ranking data*. Vol. 80. Springer, 1993.

[80] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. “CrowdDB: answering queries with crowdsourcing”. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. ACM. 2011, pp. 61–72.

[81] P. Fraternali, A. Castelletti, R Soncini-Sessa, C Vaca Ruiz, and A. Rizzoli. “Putting humans in the loop: Social computing for Water Resources Management”. In: *Environmental Modelling & Software* 37 (2012), pp. 68–77.

[82] U. Gadiraju, B. Fetahu, and R. Kawase. “Training Workers for Improving Performance in Crowdsourcing Microtasks”. In: *Design for Teaching and Learning in a Networked World*. 2015.

[83] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. “Understanding malicious behavior in crowdsourcing platforms: The case of online surveys”. In: *ACM Conference on Human Factors in Computing Systems*. 2015.

[84] C. Gao, Y. Lu, and D. Zhou. “Exact Exponent in Optimal Rates for Crowdsourcing”. In: *International Conference on Machine Learning (ICML)*. 2016.

[85] C. Gao and D. Zhou. “Minimax optimal convergence rates for estimating ground truth from crowdsourced labels”. In: *arXiv preprint arXiv:1310.5764* (2013).

[86] F. Gao and J. A. Wellner. “Entropy estimate for high-dimensional monotonic functions”. In: *Journal of Multivariate Analysis* 98.9 (2007), pp. 1751–1764.

[87] P Gay, C Lehan, J Moore, G Bracey, and N Gugliucci. “Cosmoquest: A cyberinfrastructure for crowdsourcing planetary surface mapping and more”. In: *Proceedings of the 2014 Lunar and Planetary Institute Science Conference*. Vol. 45. 2014, p. 2927.

[88] A. Ghosh, S. Kale, and P. McAfee. “Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content”. In: *ACM conference on Electronic commerce*. 2011.

[89] J. D. Gibbons, I. Olkin, and M. Sobel. “A subset selection technique for scoring items on a multiple choice test”. In: *Psychometrika* 44.3 (1979), pp. 259–270.

[90] E. N. Gilbert. “A comparison of signalling alphabets”. In: *Bell System Technical Journal* 31.3 (1952), pp. 504–522.

[91] N. Gillis. "Sparse and unique nonnegative matrix factorization through data preprocessing". In: *Journal of Machine Learning Research* 13.Nov (2012), pp. 3349–3386.

[92] N. Gillis. "The why and how of nonnegative matrix factorization". In: *Regularization, Optimization, Kernels, and Support Vector Machines* 12.257 (2014).

[93] R. J. Giuly, K.-Y. Kim, and M. H. Ellisman. "DP2: Distributed 3D image segmentation using micro-labor workforce". In: *Bioinformatics* 29.10 (2013), pp. 1359–1360.

[94] T. Gneiting and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[95] C. Gobinet, E. Perrin, and R. Huez. "Application of non-negative matrix factorization to fluorescence spectroscopy". In: *European Signal Processing Conference*. 2004.

[96] B. Good and A. Su. "Crowdsourcing for bioinformatics." In: *Bioinformatics (Oxford, England)* 29.16 (2013), p. 1925.

[97] T. L. Griffiths. "Manifesto for a new (computational) cognitive revolution". In: *Cognition* 135 (2015), pp. 21–23.

[98] B. Hajek, S. Oh, and J. Xu. "Minimax-optimal Inference from Partial Rankings". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1475–1483.

[99] S. Hanneke and L. Yang. "Negative results for active learning with convex losses". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2010, pp. 321–325.

[100] R. Heckel, N. B. Shah, K. Ramchandran, and M. J. Wainwright. "Active Ranking from Pairwise Comparisons and when Parametric Assumptions Don't Help". In: *arXiv preprint arXiv:1606.08842* (2016).

[101] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 82–97.

[102] C.-J. Ho, S. Jabbari, and J. W. Vaughan. "Adaptive task assignment for crowdsourced classification". In: *International Conference on Machine Learning (ICML)*. 2013, pp. 534–542.

[103] P. Horst. "The chance element in the multiple choice test item". In: *The Journal of General Psychology* 6.1 (1932), pp. 209–211.

[104] C.-J. Hsieh, N. Natarajan, and I. Dhillon. "PU learning for matrix completion". In: *International Conference on Machine Learning*. 2015, pp. 2445–2453.

[105] D. Hunter. "MM algorithms for generalized Bradley-Terry models". In: *Annals of Statistics* (2004), pp. 384–406.

[106] I. F. Ilyas, G. Beskales, and M. A. Soliman. "A survey of top-k query processing techniques in relational database systems". In: *ACM Computing Surveys* (2008).

[107] S. Jagabathula and D. Shah. "Inferring rankings under constrained sensing". In: *Advances in Neural Information Processing Systems*. 2008, pp. 753–760.

[108] S. Jagabathula, L. Subramanian, and A. Venkataraman. "Reputation-based Worker Filtering in Crowdsourcing". In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 2492–2500.

[109] P. Jain, P. Netrapalli, and S. Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 665–674.

[110] K. Jamieson, S. Katariya, A. Deshpande, and R. Nowak. "Sparse Dueling Bandits". In: *arXiv preprint arXiv:1502.00133* (2015).

[111] T. Jiang and A. Vardy. "Asymptotic improvement of the Gilbert-Varshamov bound on the size of binary codes". In: *IEEE Transactions on Information Theory* (2004).

[112] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. "Statistical ranking and combinatorial Hodge theory". In: *Mathematical Programming* 127.1 (2011), pp. 203–244.

[113] C. Jin, S. M. Kakade, and P. Netrapalli. "Provable efficient online matrix completion via non-convex stochastic gradient descent". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4520–4528.

[114] W. P. Jones and S. A. Loe. "Optimal Number of Questionnaire Response Categories More May Not Be Better". In: *SAGE Open* 3.2 (2013), p. 2158244013489691.

[115] V. Kamble, D. Marn, N. Shah, A. Parekh, and K. Ramachandran. "Truth Serums for Massively Crowdsourced Evaluation Tasks". In: *arXiv preprint arXiv:1507.07045* (2015).

[116] D. R. Karger, S. Oh, and D. Shah. "Budget-optimal crowdsourcing using low-rank matrix approximations". In: *49th Annual Allerton Conference on Communication, Control, and Computing*. 2011, pp. 284–291.

[117] D. R. Karger, S. Oh, and D. Shah. "Iterative learning for reliable crowdsourcing systems". In: *Advances in neural information processing systems*. 2011, pp. 1953–1961.

[118] E. Kaufmann and S. Kalyanakrishnan. "Information complexity in bandit subset selection". In: *Conference on Learning Theory*. 2013, pp. 228–251.

[119] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. "Crowdsourcing for book search evaluation: impact of HIT design on comparative system ranking". In: *ACM SIGIR conference on Research and development in Information Retrieval*. 2011, pp. 205–214.

[120] J. Kellett and K. Mott. "Presidential primaries: Measuring popular choice". In: *Polity* (1977), pp. 528–537.

[121] C. Kenyon-Mathieu and W. Schudy. "How to rank with few errors". In: *Symposium on Theory of computing (STOC)*. ACM. 2007, pp. 95–103.

[122] R. H. Keshavan, A. Montanari, and S. Oh. “Matrix completion from noisy entries”. In: *Journal of Machine Learning Research* 11.Jul (2010), pp. 2057–2078.

[123] E. Keuleers and D. A. Balota. *Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments.* 2015.

[124] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, et al. “Crystal structure of a monomeric retroviral protease solved by protein folding game players”. In: *Nature structural & molecular biology* 18.10 (2011), pp. 1175–1177.

[125] A. Khetan and S. Oh. “Reliable Crowdsourcing under the Generalized Dawid-Skene Model”. In: *arXiv preprint arXiv:1602.03481* (2016).

[126] B. Kimelfeld and Y. Sagiv. “Finding and approximating top-k answers in keyword proximity search”. In: *Symposium on Principles of database systems.* 2006.

[127] A. J. King, R. W. Gehl, D. Grossman, and J. D. Jensen. “Skin self-examinations and visual identification of atypical nevi: Comparing individual and crowdsourcing approaches”. In: *Cancer epidemiology* 37.6 (2013), pp. 979–984.

[128] T. Kitching, J Rhodes, C Heymans, R Massey, Q Liu, M Cobzarenco, et al. “Image analysis for cosmology: Shape measurement challenge review & results from the Mapping Dark Matter challenge”. In: *Astronomy and Computing* 10 (2015), pp. 9–21.

[129] T. Klein, E. Rio, et al. “Concentration around the mean for maxima of empirical processes”. In: *The Annals of Probability* 33.3 (2005), pp. 1060–1077.

[130] O. Klopp. “Noisy low-rank matrix completion with general sampling distribution”. In: *Bernoulli* 20.1 (2014), pp. 282–303.

[131] T. Kolokolnikov, B. Osting, and J. Von Brecht. “Algebraic connectivity of Erdös-Rényi graphs near the connectivity threshold”. 2014.

[132] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion”. In: *The Annals of Statistics* (2011), pp. 2302–2329.

[133] Y. Koren, R. Bell, and C. Volinsky. “Matrix factorization techniques for recommender systems”. In: *Computer* 42.8 (2009).

[134] A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman, et al. “Crowdsourcing in Computer Vision”. In: *Foundations and Trends® in Computer Graphics and Vision* 10.3 (2016), pp. 177–243.

[135] R. Kyng, A. Rao, and S. Sachdeva. “Fast, Provable Algorithms for Isotonic Regression in all L_p-norms”. In: *Advances in Neural Information Processing Systems.* 2015, pp. 2701–2709.

[136] S. P. Lalley. “Concentration inequalities”. In: *Lecture notes, University of Chicago* (2013).

[137] N. Lambert and Y. Shoham. “Eliciting truthful answers to multiple-choice questions”. In: *ACM conference on Electronic commerce*. 2009, pp. 109–118.

[138] E. Landhuis. “Science and Culture: Putting a game face on biomedical research”. In: *Proceedings of the National Academy of Sciences* 113.24 (2016), pp. 6577–6578.

[139] A. Lang and J. Rio-Ross. “Using Amazon Mechanical Turk to transcribe historical handwritten documents”. In: *The Code4Lib Journal* 15 (2011).

[140] J.-F. Laslier and K. Van der Straeten. “A live experiment on approval voting”. In: *Experimental Economics* 11.1 (2008), pp. 97–105.

[141] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. “Theorems on positive data: On the uniqueness of NMF”. In: *Computational intelligence and neuroscience* 2008 (2008).

[142] J. Lawson, R. J. Robinson-Vyas, J. P. McQuillan, A. Paterson, S. Christie, M. Kidza-Griffiths, et al. “Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays”. In: *British journal of cancer* 116.2 (2017), pp. 237–245.

[143] J. Le, A. Edmonds, V. Hester, and L. Biewald. “Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution”. In: *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 2010, pp. 21–26.

[144] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.

[145] D. D. Lee and H. S. Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791.

[146] E. W. Lee, C. P. Lim, R. K. Yuen, and S. Lo. “A hybrid neural network model for noisy data regression”. In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 34.2 (2004), pp. 951–960.

[147] J. Lee, W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, et al. “RNA design rules from a massive open laboratory”. In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2122–2127.

[148] J. H. Lee and X. Hu. “Generating ground truth for music mood classification using mechanical turk”. In: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. ACM. 2012, pp. 129–138.

[149] V. I. Levenshtein. “Upper-bound estimates for fixed-weight codes”. In: *Problemy Peredachi Informatsii* 7.4 (1971), pp. 3–12.

[150] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, et al. “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey”. In: *Monthly Notices of the Royal Astronomical Society* 389.3 (2008), pp. 1179–1189.

[151] Q. Liu, J. Peng, and A. T. Ihler. “Variational inference for crowdsourcing”. In: *Advances in Neural Information Processing Systems.* 2012, pp. 692–700.

[152] T. Lombrozo and B. Rehder. “Functions in biological kind classification”. In: *Cognitive psychology* 65.4 (2012), pp. 457–485.

[153] P. M. Long and R. A. Servedio. “Random classification noise defeats all convex potential boosters”. In: *Machine Learning* 78.3 (2010), pp. 287–304.

[154] R. D. Luce. “Individual choice behavior, a theoretical analysis”. In: *Bull. Amer. Math. Soc. 66 (1960), 259-260* (1960), pp. 0002–9904.

[155] M. A. Luengo-Oroz, A. Arranz, and J. Frean. “Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears”. In: *Journal of medical Internet research* 14.6 (2012).

[156] Y. Luo, N. B. Shah, J. Huang, and J. Walrand. “Parametric Prediction from Parametric Agents”. In: *Workshop on Social Computing and User Generated Content (SCUGC)* (2016).

[157] N. Manwani and P. Sastry. “Noise tolerance under risk minimization”. In: *IEEE Transactions on Cybernetics* 43.3 (2013), pp. 1146–1151.

[158] J. I. Marden. *Analyzing and modeling rank data.* CRC Press, 1996.

[159] A. Marley. “Aggregation theorems and the combination of probabilistic rank orders”. In: *Probability models and statistical analyses for ranking data.* Springer, 1993, pp. 216–240.

[160] J. Massó and M. Vorsatz. “Weighted approval voting”. In: *Economic Theory* 36.1 (2008), pp. 129–146.

[161] J. Matoušek and J. Vondrák. “The probabilistic method”. In: *Lecture Notes, Department of Applied Mathematics, Charles University, Prague* (2001).

[162] A. B. McCoy, A. Wright, D. Rogith, S. Fathiamini, A. J. Ottenbacher, and D. F. Sittig. “Development of a clinician reputation metric to identify appropriate problem-medication pairs in a crowdsourced knowledge base”. In: *Journal of biomedical informatics* 48 (2014), pp. 66–72.

[163] D. H. McLaughlin and R. D. Luce. “Stochastic transitivity and cancellation of preferences between bitter-sweet solutions”. In: *Psychonomic Science* 2.1-12 (1965), pp. 89–90.

[164] D. Mease, A. J. Wyner, and A. Buja. “Boosted classification trees and class probability/quantile estimation”. In: *The Journal of Machine Learning Research* 8 (2007), pp. 409–439.

[165] P Melchior, E Sheldon, A Drlica-Wagner, E. Rykoff, T. Abbott, F. Abdalla, et al. “Crowdsourcing quality control for Dark Energy Survey images”. In: *Astronomy and Computing* 16 (2016), pp. 99–108.

[166] A. Metwally, D. Agrawal, and A. El Abbadi. "Efficient computation of frequent and top-k elements in data streams". In: *Database Theory-ICDT*. 2005.

[167] A. N. Meyer, C. A. Longhurst, and H. Singh. "Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of CrowdMed". In: *Journal of medical Internet research* 18.1 (2016).

[168] S. Michel, P. Triantafillou, and G. Weikum. "Klee: A framework for distributed top-k query algorithms". In: *International conference on Very large data bases*. 2005.

[169] G. A. Miller. "The magical number seven, plus or minus two: some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81.

[170] J. D. Miller, M. Crowe, B. Weiss, J. L. Maples-Keller, and D. R. Lynam. "Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazons Mechanical Turk." In: *Personality Disorders: Theory, Research, and Treatment* 8.1 (2017), p. 26.

[171] N. Miller, P. Resnick, and R. Zeckhauser. "Eliciting informative feedback: The peer-prediction method". In: *Management Science* 51.9 (2005), pp. 1359–1373.

[172] I. Mitliagkas, A. Gopalan, C. Caramanis, and S. Vishwanath. "User rankings from comparisons: Learning permutations in high dimensions". In: *Allerton Conference on Communication, Control, and Computing*. 2011.

[173] J. M. Mortensen, M. A. Musen, and N. F. Noy. "Crowdsourcing the Verification of Relationships in Biomedical Ontologies". In: *AMIA Annual Symposium Proceedings*. Vol. 2013. American Medical Informatics Association. 2013, p. 1020.

[174] S. Mullainathan, J. Schwartzstein, and A. Shleifer. "Coarse Thinking and Persuasion". In: *The Quarterly journal of economics* 123.2 (2008), pp. 577–619.

[175] P. Nakkiran, N. B. Shah, and K. Rashmi. "Fundamental limits on communication for oblivious updates in storage networks". In: *2014 IEEE Global Communications Conference*. IEEE. 2014, pp. 2363–2368.

[176] P. Nakkiran, N. B. Shah, K. Rashmi, A. Sahai, and K. Ramchandran. "Optimal Oblivious Updates in Distributed Storage Networks". In: (2016).

[177] S. Negahban, S. Oh, and D. Shah. "Iterative ranking from pair-wise comparisons". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2474–2482.

[178] S. Negahban and M. J. Wainwright. "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise". In: *Journal of Machine Learning Research* 13.May (2012), pp. 1665–1697.

[179] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, et al. "Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography". In: *Radiology* 262.3 (2012), pp. 824–833.

[180] I. Norheim-Hagtun and P. Meier. "Crowdsourcing for crisis mapping in Haiti". In: *innovations* 5.4 (2010), pp. 81–89.

[181] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. "Platemate: crowdsourcing nutritional analysis from food photographs". In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM. 2011, pp. 1–12.

[182] D. or nothing". `http://wikipedia.org/wiki/Double_or_nothing`. Last accessed: July 31, 2014.

[183] O. Okolloh. "Ushahidi, or testimony: Web 2.0 tools for crowdsourcing crisis information". In: *Participatory learning and action* 59.1 (2009), pp. 65–70.

[184] R. I. Oliveira. "Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges". In: *arXiv preprint:0911.0600* (2009).

[185] G. Ottewell. "The arthmetic of voting". In: *In defence of variety* (1977).

[186] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. "Ranking and combining multiple predictors without labeled data". In: *Proceedings of the National Academy of Sciences* 111.4 (2014), pp. 1253–1258.

[187] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. "Tuned Models of Peer Assessment in MOOCs". In: *International Conference on Educational Data Mining*. 2013.

[188] D. Prelec. "A Bayesian truth serum for subjective data". In: *Science* 306.5695 (2004), pp. 462–466.

[189] A. D. Procaccia and N. Shah. "Is Approval Voting Optimal Given Approval Votes?" In: *NIPS*. 2015.

[190] G. Radanovic and B. Faltings. "A robust bayesian truth serum for non-binary signals". In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*. EPFL-CONF-197486. 2013, pp. 833–839.

[191] G. Radanovic, B. Faltings, and R. Jurca. "Incentives for Effort in Crowdsourcing using the Peer Truth Serum". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.4 (2016), p. 48.

[192] A. Rajkumar and S. Agarwal. "A statistical convergence perspective of algorithms for rank aggregation from pairwise data". In: *Proceedings of the 31st International Conference on Machine Learning*. 2014, pp. 118–126.

[193] A. Rajkumar, S. Ghoshal, L.-H. Lim, and S. Agarwal. "Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top". In: *International Conference on Machine Learning*. 2015.

[194] K. V. Rashmi, N. B. Shah, and P. V. Kumar. "Enabling Node Repair in Any Erasure Code for Distributed Storage". In: *Proc. IEEE International Symposium on Information Theory (ISIT)*. July 2011.

[195] K. V. Rashmi, N. B. Shah, and P. V. Kumar. "Optimal Exact-Regenerating Codes for the MSR and MBR Points via a Product-Matrix Construction". In: *IEEE Transactions on Information Theory* 57.8 (Aug. 2011), pp. 5227–5239.

[196] K. V. Rashmi, N. B. Shah, K. Ramchandran, and P. Kumar. "Regenerating Codes for Errors and Erasures in Distributed Storage". In: *Proc. International Symposium on Information Theory*. July 2012.

[197] K. V. Rashmi, N. B. Shah, and P. V. Kumar. *Distributed storage system and a method thereof*. US Patent App. 13/110,534. May 2011.

[198] K. V. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran. "Explicit and Optimal Exact-Regenerating Codes for the Minimum-Bandwidth Point in Distributed Storage". In: *Proc. IEEE ISIT*. Austin, June 2010.

[199] K. V. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran. "Explicit Construction of Optimal Exact Regenerating Codes for Distributed Storage". In: *Proc. Allerton Conf.* Sept. 2009.

[200] K. V. Rashmi, N. B. Shah, K. Ramchandran, and P. V. Kumar. "Information-theoretically Secure Erasure Codes for Distributed Storage". In: 2017 (to appear).

[201] K. Rashmi, P. Nakkiran, J. Wang, N. B. Shah, and K. Ramchandran. "Having Your Cake and Eating It Too: Jointly Optimal Erasure Codes for I/O, Storage, and Network-bandwidth." In: *FAST*. 2015, pp. 81–94.

[202] K. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K. Ramchandran. "A hitchhiker's guide to fast and efficient data reconstruction in erasure-coded data centers". In: *ACM SIGCOMM Computer Communication Review* 44.4 (2015), pp. 331–342.

[203] K. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K. Ramchandran. "A Solution to the Network Challenges of Data Recovery in Erasure-coded Distributed Storage Systems: A Study on the Facebook Warehouse Cluster." In: *HotStorage*. 2013.

[204] K. Rashmi, N. B. Shah, and K. Ramchandran. "A piggybacking design framework for read-and download-efficient distributed storage codes". In: *IEEE Transactions on Information Theory* (2017).

[205] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, et al. "Learning from crowds". In: *The Journal of Machine Learning Research* 99 (2010), pp. 1297–1322.

[206] B. Recht. "A simpler approach to matrix completion". In: *Journal of Machine Learning Research* 12.Dec (2011), pp. 3413–3430.

[207] M. Regenwetter and I. Tsetlin. "Approval voting and positional voting methods: Inference, relationship, examples". In: *Social Choice and Welfare* 22.3 (2004), pp. 539–566.

[208] K. Reynolds, A. Kontostathis, and L. Edwards. "Using machine learning to detect cyberbullying". In: *International Conference on Machine Learning and Applications and Workshops*. Vol. 2. 2011, pp. 241–244.

[209] T. Robertson, F. Wright, R. L. Dykstra, and T Robertson. *Order restricted statistical inference*. Vol. 229. Wiley New York, 1988.

[210] Y. Roth, F Pétavy, and J Céré. “The state of crowdsourcing in 2015”. In: *eYeka Analyst Report* (2015).

[211] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al. “Imagenet large scale visual recognition challenge”. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[212] T. L. Saaty and M. S. Ozdemir. “Why the magic number seven plus or minus two”. In: *Mathematical and Computer Modelling* 38.3 (2003), pp. 233–244.

[213] P.-M. Samson. “Concentration of measure inequalities for Markov chains and Φ-mixing processes”. In: *Annals of Probability* (2000), pp. 416–461.

[214] C. Sarasua, E. Simperl, and N. Noy. “Crowdmap: Crowdsourcing ontology alignment with microtasks”. In: *The Semantic Web–ISWC 2012* (2012), pp. 525–541.

[215] R. Sarver and A. Klapuri. “Application of nonnegative matrix factorization to signal-adaptive audio effects”. In: *Proc. DAFx*. 2011, pp. 249–252.

[216] L. J. Savage. “Elicitation of personal probabilities and expectations”. In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801.

[217] N. B. Shah, K. V. Rashmi, and P. V. Kumar. “Information-theoretically Secure Regenerating Codes for Distributed Storage”. In: *Proc. Globecom*. Houston, Dec. 2011.

[218] N. Shah, B. Tabiban, K. Muandet, I. Guyon, and U. von Luxberg. “Design and Analysis of the NIPS 2016 Review Process”. In: *arXiv preprint* (2017).

[219] N. Shah and D. Zhou. “No oops, you wont do it again: mechanisms for self-correction in crowdsourcing”. In: *International Conference on Machine Learning*. 2016.

[220] 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”. In: *Journal of Machine Learning Research* (2016).

[221] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. “Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues”. In: *IEEE Transactions on Information Theory* (2017).

[222] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “A Permutation-based Model for Crowd Labeling: Optimal Estimation and Robustness”. In: *arXiv preprint arXiv:1606.09632* (2016).

[223] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “Feeling the Bern: Adaptive Estimators for Bernoulli Probabilities of Pairwise Comparisons”. In: *International Symposium on Information Theory (ISIT)*. 2016.

[224] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “Permutation-rank matrix completion”. In: *arXiv preprint* (2017).

[225] N. B. Shah, J. K. Bradley, A. Parekh, M. Wainwright, and K. Ramchandran. "A Case for Ordinal Peer-evaluation in MOOCs". In: *NIPS Workshop on Data Driven Education*. Dec. 2013.

[226] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran. "Explicit codes uniformly reducing repair bandwidth in distributed storage". In: *National Conference on Communications (NCC)*. Chennai, Jan. 2010.

[227] N. B. Shah, K. V. Rashmi, and P. V. Kumar. "A Flexible Class of Regenerating Codes for Distributed Storage". In: *Proc. ISIT*. Austin, June 2010.

[228] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran. "Explicit Codes Minimizing Repair Bandwidth for Distributed Storage". In: *Proc. IEEE ITW*. Cairo, Jan. 2010.

[229] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran. "Interference Alignment in Regenerating Codes for Distributed Storage: Necessity and Code Constructions". In: *IEEE Transactions on Information Theory* 58.4 (Apr. 2012), pp. 2134–2158.

[230] N. B. Shah, K. V. Rashmi, and K. Ramchandran. "Secure Network Coding for Distributed Secret Sharing with Low Communication Cost". In: *Proc. IEEE International Symposium on Information Theory (ISIT)*. Istanbul, July 2013.

[231] N. B. Shah, K. Rashmi, P. V. Kumar, and K. Ramchandran. "Regenerating codes for distributed storage networks". In: *International Workshop on the Arithmetic of Finite Fields*. Springer. 2010, pp. 215–223.

[232] N. B. Shah, K. Rashmi, and K. Ramchandran. "Distributed secret dissemination across a network". In: *IEEE Journal of Selected Topics in Signal Processing* 9.7 (2015), pp. 1206–1216.

[233] N. B. Shah, K. Rashmi, and K. Ramchandran. "One extra bit of download ensures perfectly private information retrieval". In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 856–860.

[234] N. B. Shah and M. J. Wainwright. "Simple, robust and optimal ranking from pairwise comparisons". In: *arXiv preprint arXiv:1512.08949* (2015).

[235] N. B. Shah and D. Zhou. "Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing". In: *Journal of Machine Learning Research* 17 (2016), pp. 1–52.

[236] N. B. Shah and D. Zhou. "On the impossibility of convex inference in human computation". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[237] N. B. Shah, D. Zhou, and Y. Peres. "Approval Voting and Incentives in Crowdsourcing". In: *ICML*. 2015.

[238] R. Shiffrin and R. Nosofsky. "Seven plus or minus two: a commentary on capacity limitations." In: *Psychological review* 101.2 (1994), p. 357.

[239] V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. “Informed Truthfulness in Multi-Task Peer Prediction”. In: *arXiv preprint arXiv:1603.03151* (2016).

[240] H. Siddiqi. “Does Coarse Thinking Matter for Option Pricing? Evidence from an Experiment.” In: *IUP Journal of Behavioral Finance* 8.2 (2011).

[241] M. J. Silvapulle and P. K. Sen. *Constrained statistical inference: Order, inequality, and shape constraints*. Vol. 912. John Wiley & Sons, 2011.

[242] R. J. Smith and R. M. Merchant. “Harnessing the crowd to accelerate molecular medicine research”. In: *Trends in Molecular Medicine* 21.7 (2015), pp. 403–405.

[243] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2008, pp. 254–263.

[244] H. Soufiani, D. Parkes, and L. Xia. “Computing parametric ranking models via rank-breaking”. In: *ICML*. 2014.

[245] N. Srebro, N. Alon, and T. S. Jaakkola. “Generalization error bounds for collaborative prediction with low-rank matrices”. In: *Advances In Neural Information Processing Systems*. 2005.

[246] G. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. 1990.

[247] N. Stewart, G. D. Brown, and N. Chater. “Absolute identification by relative judgment.” In: *Psychological review* 112.4 (2005), p. 881.

[248] B. Szörényi, R. Busa-Fekete, A. Paul, and E. Hüllermeier. “Online rank elicitation for Plackett-Luce: A dueling bandits approach”. In: *nips*. 2015, pp. 604–612.

[249] K. Talwar. “The price of truth: Frugality in truthful mechanisms”. In: *Annual Symposium on Theoretical Aspects of Computer Science*. Springer. 2003, pp. 608–619.

[250] T. Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Society Providence, RI, 2012.

[251] F. J. Theis, K. Stadlthanner, and T. Tanaka. “First results on uniqueness of sparse non-negative matrix factorization”. In: *European Signal Processing Conference*. 2005.

[252] R. Thompson. “The behavior of eigenvalues and singular values under perturbations of restricted rank”. In: *Linear Algebra and its Applications* 13.1 (1976), pp. 69–78.

[253] L. L. Thurstone. “A law of comparative judgment”. In: *Psychological Review* 34.4 (1927), p. 273.

[254] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. 2008.

[255] A. Tversky. “Elimination by aspects: A theory of choice.” In: *Psychological review* 79.4 (1972), p. 281.

[256] A. Van Der Vaart and J. Wellner. “Weak Convergence”. In: *Weak Convergence and Empirical Processes*. Springer, 1996, pp. 16–28.

[257] A. Vandendorpe, N.-D. Ho, S. Vanduffel, and P. Van Dooren. “On the parameterization of the CreditRisk+ model for estimating credit portfolio risk”. In: *Insurance: Mathematics and Economics* 42.2 (2008), pp. 736–745.

[258] R. Varshamov. “Estimate of the number of signals in error correcting codes”. In: *Dokl. Akad. Nauk SSSR*. 1957.

[259] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. “Recaptcha: Human-based character recognition via web security measures”. In: *Science* 321.5895 (2008), pp. 1465–1468.

[260] J. Vuurens, A. P. de Vries, and C. Eickhoff. “How much spam can you take? An analysis of crowdsourcing results to increase accuracy”. In: *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*. 2011, pp. 21–26.

[261] P. Wais, S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, et al. “Towards building a high-quality workforce with Mechanical Turk”. In: *NIPS workshop on computational social science and the wisdom of crowds* (2010).

[262] S. C. Warby, S. L. Wendt, P. Welinder, E. G. Munk, O. Carrillo, H. B. Sorensen, et al. “Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods”. In: *Nature methods* 11.4 (2014), pp. 385–392.

[263] F. L. Wauthier, M. I. Jordan, and N. Jojic. “Efficient Ranking from Pairwise Comparisons”. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. 2013.

[264] R. J. Weber. “Comparison of voting systems”. In: *New Haven: Cowles Foundation Discussion paper A* 498 (1977).

[265] J. Whitehill, P. Ruvolo, T.-f. Wu, J. Bergsma, and J. Movellan. “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise”. In: *Advances in neural information processing systems*. 2009, pp. 2035–2043.

[266] P. Whitla. “Crowdsourcing and its application in marketing activities”. In: *Contemporary Management Research* 5.1 (2009).

[267] J. Wolfers and E. Zitzewitz. *Prediction markets*. Tech. rep. National Bureau of Economic Research, 2004.

[268] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. “Parsing clothing in fashion photographs”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 3570–3577.

[269] M. Yetisgen-Yildiz, I. Solti, F. Xia, and S. Halgrim. “Preliminary experience with Amazon’s Mechanical Turk for annotating medical named entities”. In: *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. 2010.

[270] M.-C. Yuen, I. King, and K.-S. Leung. “A survey of crowdsourcing systems”. In: *IEEE Inernational Conference on Social Computing*. 2011.

[271] M.-C. Yuen, I. King, and K.-S. Leung. “Task matching in crowdsourcing”. In: *IEEE International Conference on Cyber, Physical and Social Computing*. 2011, pp. 409–412.

[272] S.-Y. Yun, M. Lelarge, and A. Proutiere. “Streaming, Memory Limited Matrix Completion with Noise”. In: *arXiv preprint arXiv:1504.03156* (2015).

[273] J. Zhang, X. Wu, and V. S. Sheng. “Learning from crowdsourced labeled data: a survey”. In: *Artificial Intelligence Review* 46.4 (2016), pp. 543–576.

[274] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. “Spectral methods meet EM: A provably optimal algorithm for crowdsourcing”. In: *Advances in neural information processing systems*. 2014, pp. 1260–1268.

[275] D. Zhou, Q. Liu, J. C. Platt, C. Meek, and N. B. Shah. “Regularized minimax conditional entropy for crowdsourcing”. In: *arXiv preprint arXiv:1503.07240* (2015).

[276] D. Zhou, J. Platt, S. Basu, and Y. Mao. “Learning from the wisdom of crowds by minimax entropy”. In: *Advances in Neural Information Processing Systems 25*. 2012, pp. 2204–2212.

[277] Y. Zhou, X. Chen, and J. Li. “Optimal PAC multiple arm identification with applications to crowdsourcing”. In: *International Conference on Machine Learning (ICML)*. 2014, pp. 217–225.